Redbooks
ibm.com/redbooks

IBM SONAS Implementation Guide

Megan Gilge

Marcos Figueiredo Jr

Daesung Kim

Mary Lovelace

Bill Marshall

Gabor Penzes

Ravikumar Ramaswamy

Joe Roa

John Tarella

Michael Taylor

Shradha Nayak Thakare

IBM®

Redbooks

International Technical Support Organization

IBM SONAS Implementation Guide

June 2015

SG24-7962-01

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

Second Edition (June 2015)

This edition applies to IBM Scale Out Network Attached Storage Version 1.5.1 (product number 5639-SN1).

Contents

Notices

This information was developed for products and services offered in the U.S.A.

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, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: 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 states 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 supply in any way it believes appropriate without incurring any obligation to you.

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

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.

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 the names and addresses used by an actual business enterprise 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.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Active Cloud Engine®
AIX®
DS5000™
DS8000®
Easy Tier®
eServer™
Global Technology Services®
GPFS™
IBM®
IBM z™
ProtecTIER®
Real-time Compression™
Redbooks®
Redbooks (logo) ®
Storwize®
System Storage®
Tivoli®
XIV®
z/OS®

The following terms are trademarks of other companies:

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

Linear Tape-Open, LTO, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

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

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

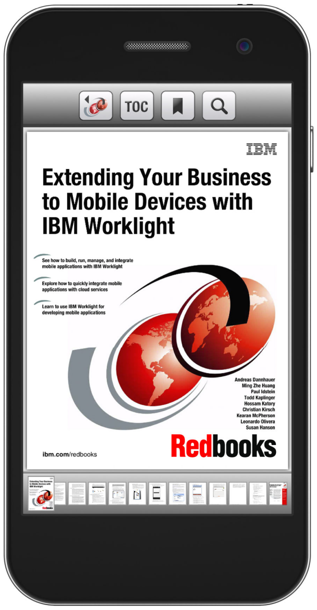
Other company, product, or service names may be trademarks or service marks of others.

Find and read thousands of IBM Redbooks publications

- ▸ Search, bookmark, save and organize favorites
- ▸ Get up-to-the-minute Redbooks news and announcements
- ▸ Link to the latest Redbooks blogs and videos

Get the latest version of the Redbooks Mobile App

iOS ◀ **Download Now** ▶ Android



IBM

Extending Your Business to Mobile Devices with IBM Worklight

See how to build, run, manage, and integrate mobile applications with IBM Worklight

Explore how to quickly integrate mobile applications with cloud services

Learn to use IBM Worklight for developing mobile applications

Andreas Dannhauer
Ming Zhe Huang
Paul Idstein
Todd Kaplinger
Hossam Katory
Christian Kirsch
Kearan McPherson
Leonardo Olivera
Susan Hanson

ibm.com/redbooks

Redbooks

Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!

ibm.com/Redbooks
About Redbooks → Business Partner Programs

Preface

IBM® Scale Out Network Attached Storage (SONAS) is a scale out network-attached storage offering that is designed to manage vast repositories of information in enterprise environments that require large capacities, high levels of performance, and high availability.

SONAS provides a range of reliable, scalable storage solutions for various storage requirements. These capabilities are achieved by using network access protocols such as Network File System (NFS), Common Internet File System (CIFS), Hypertext Transfer Protocol Secure (HTTPS), File Transfer Protocol (FTP), and Secure Copy Protocol (SCP). Using built-in RAID technologies, all data is well-protected with options to add more protection through mirroring, replication, snapshots, and backup. These storage systems are also characterized by simple management interfaces that make installation, administration, and troubleshooting uncomplicated and straightforward.

This IBM Redbooks® publication is the companion to *IBM SONAS Best Practices*, SG24-8051. It is intended for storage administrators who have ordered their SONAS solution and are ready to install, customize, and use it. It provides backup and availability scenarios information about configuration and troubleshooting. This book applies to IBM SONAS Version 1.5.5. It is useful for earlier releases of IBM SONAS as well.

Authors

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Tucson Center.

Megan Gilge is a Project Leader at the IBM International Technical Support Organization. Before joining the ITSO, she was an Information Developer in the IBM Semiconductor Solutions and User Technologies areas. Megan holds a bachelor degree in liberal arts from Michigan Technological University and a master degree in English from Saint Louis University.

Marcos Figueiredo Jr is a Storage Specialist in Brazil with four years of experience in network-attached storage (NAS) environments. He holds a degree in computer science from Universidade de Brasilia. His areas of expertise include IBM Tivoli® Storage Manager and storage infrastructure consulting, design, implementation services, open systems storage, and storage performance monitoring and tuning. Marcos is presently focusing on open systems, including NAS and virtualization.

Daesung Kim is a storage Top Gun for the IBM Global Technology Services® Organization in Korea. He provides second-level support as a Product Field Engineer for the Technical Support Group. He joined IBM in 2006 and worked for many years as an IBM System Support Representative (SSR) for industry customers. In his current role as a second-level support engineer in Seoul, Korea, he supports SONAS, midrange storage, and virtual tape hardware products. He has experience with the Linux, Windows, Sun Solaris, and IBM AIX® operating systems.

Mary Lovelace was a Consulting IT Specialist at the International Technical Support Organization. She has more than 20 years of experience with IBM in large systems, storage, and storage networking product education, system engineering and consultancy, and systems support. She has written many IBM Redbooks publications about Scale Out Network Attached Storage, IBM Tivoli Storage Productivity Center, Tivoli Storage Manager, and IBM z/OS® storage products.

Bill Marshall is a Linux Subject Matter Expert in Rochester, MN in Strategic Outsourcing, IBM Global Technology Services. He has a master degree in computer science from Iowa State University. Bill has expertise in the areas of Linux, file serving including Samba, and in-depth experience with Windows and Linux interoperability. Bill has worked on Linux since 2001 and is certified as an IBM Expert IT Specialist. Bill's areas of interest include automation and scripting, distributed file systems, and KVM virtualization.

Gabor Penzes is an IT Specialist and a certified SNIA Storage Specialist working for IBM STG Lab Services in Hungary. He has been working for IBM for six years, consulting on and implementing IBM eServer™ pSeries (AIX) and Storage projects with IBM customers in various industries. Gabor has also an extensive background in Information Security disciplines and has worked with UNIX based systems for over 10 years. Gabor graduated from the University of Pecs in Pecs, Hungary.

Ravikumar Ramaswamy is an Advisory Software Engineer with IBM India software labs in Pune, India. He holds a bachelor degree in computer engineering from the University of Mumbai. Ravikumar has 14 years of total experience in IT, most of which has been in storage and networking. He is working as a SONAS L3 support engineer and is involved in various customer engagements in GMU. Ravikumar also serves as a Lab advocate for SONAS customers. Before his current assignment, Ravikumar worked in IBM GPFS™ FVT for Linux on IBM z™ Systems and SONAS regression testing. He has experience with networking technologies, including network management and DNS/DHCP implementation.

Joa Roa was an IBM XIV® and SONAS Solutions Architect working for IBM STG from upstate New York. He has over 25 years of experience in UNIX server environments and enterprise-level data protection with extensive experience in real-world application of enterprise-level block and file storage. His career spans 14 years of technical leadership in the US Marine Corps and over 15 years in corporate America enterprise IT UNIX and storage systems. Joe holds a degree in electronics engineering and works primarily helping companies solve IT storage problems with improved storage solutions on advanced IBM storage platforms.

John Tarella is an Executive IT Specialist who works for IBM Global Services in Italy. He has 28 years of experience in storage and performance management on mainframe and distributed environments. He holds a degree in seismic structural engineering from Politecnico di Milano, Italy. His areas of expertise include IBM Tivoli Storage Manager and storage infrastructure consulting, design, implementation services, open systems storage, and storage performance monitoring and tuning. Now, he is working on storage infrastructures and data protection for IBM managed cloud environments. He has written extensively on z/OS DFSMS, IBM Tivoli Storage Manager, SANs, storage business continuity solutions, content management, ILM solutions, and SONAS. He also has an interest in Web 2.0 and social networking tools and methodologies.

Michael Taylor is a Storage Specialist in the IBM Systems and Technology Group in Tucson, Arizona. He holds a bachelor degree in management information systems from the University of Arizona. He has 15 years of experience with IBM in various roles. He has worked with a significant portion of the IBM Storage portfolio, including Linear Tape-Open (LTO) tape drives, IBM DS8000®, Virtual Tape Server (VTS), TS7740, 3584 tape library, 3592 tape drives with encryption, DCS3700, TS7650G IBM ProtecTIER® deduplication, IBM Storwize® V7000, Storwize V7000 Unified, and SONAS. He has most recently focused on SONAS with Gateway attached storage and a new assignment with Elastic Storage.

Shradha Nayak Thakare is a Staff Software Engineer working with IBM India Software Labs in Pune, India. She holds a Bachelor of Computer Science Engineering degree and has eight years of experience. She has been working in the storage domain since and has good expertise in Scale out File Service (SoFS) and SONAS. She works as a Level 3 developer for SONAS and also assists many customer engagements for SONAS. Shradha is interested in storage products and cloud storage. She is focusing on SONAS authentication and authorization and assisting customers to set them up correctly. Shradha is also interested in social media and social networking tools and methodologies.

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

Adam Childers
Mathias Dietz
Greg Kishi
Andreas Luengen
Thomas Luther
Christof Schmitt
Frederick Stock
Desiree Strom
Mark Taylor
IBM Systems & Technology Group

Sven Oehme
Renu Tewari
IBM Research

Thanks to the authors of the previous editions of this book.

Authors of the first edition, IBM SONAS Implementation Guide, published in December 2012, were:

Jichan Chong
Curtis Neal
Ravikumar Ramaswamy
Alexander Saupp
Mladen Vukoje

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 ITSO 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, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- 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

Installation planning

Scale Out Network Attached Storage (SONAS) is a sophisticated network-attached storage (NAS) solution that can be configured in many different ways. A clear understanding of how your environment works and how you intend to use this appliance is critical to a successful SONAS implementation.

There are a few major considerations for planning your SONAS cluster. This chapter emphasizes planning to help ensure the highest level of success in installing and integrating SONAS into your environment.

This chapter describes the following topics:

- Pre-installation planning review
- Gathering SONAS requirements
- Physical planning
- Installation plan
- Configuration plan
- Data protection considerations

1.1 Pre-installation planning review

This section provides a high-level review of the installation process because it can help you understand the types of information and planning that are pertinent to a successful installation in your environment. Figure 1-1 shows a high-level diagram of SONAS in a typical environment. Physically, SONAS is one or more IBM racks, depending on the size of your configuration. Each rack is either a Primary Rack or an Interface Expansion rack, which can be a combination of Interface and Storage Node racks. Logically, SONAS connects to the network and presents itself as one NAS. It is managed as a single entity and is supported by IBM as a single product with one product number.

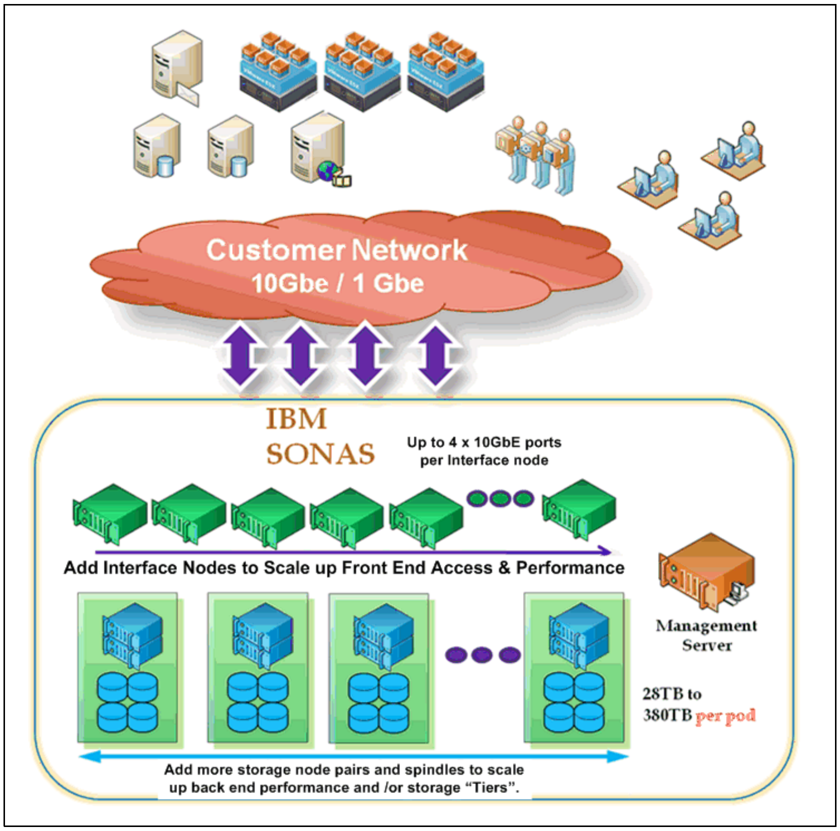

Customer Network
10Gbe / 1 Gbe
IBM
SONAS
Up to 4 x 10GbE ports
per Interface node
Add Interface Nodes to Scale up Front End Access & Performance
Management
Server
28TB to
380TB per pod
Add more storage node pairs and spindles to scale
up back end performance and /or storage "Tiers".

Figure 1-1 Overview of SONAS in a typical customer environment

The following list presents a high-level outline of the steps for connecting, installing, configuring, and integrating SONAS into your environment:

1. Physical:
   a. Physical shipping to your data center
   b. Locating and installation each SONAS rack
   c. Physically connecting the racks together (medium and large implementations)
   d. Connecting power

2. Installation:
   a. Connecting network ports to your redundant Ethernet switches
   b. Connecting the external storage (Gateway Solution) to the SONAS Storage nodes
      i. SONAS Gateway with XIV storage is supported through SAN attach only (customer supplied SAN). Direct attach is not supported.
      ii. SONAS Gateway with Storwize V7000 storage is supported as SAN Attach (customer supplied SAN) *or* direct attached.
      iii. SONAS Gateway with DCS3700 storage is supported as Direct Attached only. SAN connectivity with DCS3700 is not supported.
   c. Powering on external storage (SONAS Gateway)
   d. Creating and allocating LUNs from external storage (SONAS Gateway)
   e. Powering on the storage enclosures from redundant power distribution sources (SONAS Appliance (internal DDN storage)) from redundant power distribution sources
   f. Powering on the storage controllers from redundant power distribution sources (SONAS Appliance (internal DDN storage)) from redundant power distribution sources
   g. Powering on the Management node or Integrated Management nodes from redundant power distribution sources
   h. Installing the latest code from a DVD in the Primary Integrated Management node
   i. Running the first-time installation script
   j. Powering on the Storage nodes from redundant power distribution sources
   k. Powering on the Interface nodes from redundant power distribution sources
   l. Completing the first-time installation from redundant power distribution sources

> **Note:** As of SONAS V1.4.1, all SONAS configurations are sold as Gateways. The SONAS Appliance with internal DDN storage is no longer available for new installations. Existing Appliance configurations are fully supported but cannot be expanded with DDN storage. Instead, it is possible to expand storage capacity with DCS3700 storage. For more information, see Chapter 2, “Installation and configuration for a SONAS appliance with DDN” on page 61.

3. Configuration (basic and advanced configuration):
   a. Configuring for user administration
   b. Connecting to the network (VLANs, IP addresses, and DNS)
   c. Configuring for AD or LDAP authentication
   d. Adding external storage (SONAS Gateway) if not done above
   e. Identifying and setting up storage pools
   f. Creating internal file systems and file sets
   g. Setting up exports (shares)
   h. Configuring and testing high availability and remediation
   i. Connecting users and applications
   j. Configuring name space extensions with remote caching

4. Data protection:
   a. Configuring snapshot schedules
   b. Determining a backup method
   c. Connecting to backup services
   d. Configuring replication to another SONAS

The following sections describe the main steps for planning for a successful implementation of SONAS. These steps include the requirements gathering phase, and focus on the physical planning considerations (such as space, power consumption, and noise).

1.2 Gathering SONAS requirements

This section lists the prerequisite information that you need to install, configure, and integrate successfully SONAS into your environment. There are two major aspects to the information gathering phase. The first aspect is to gather the basic information that you need to install and configure successfully SONAS in your data center and information about how to connect it to your authentication domain. The second aspect is to help you gather sufficient information about your environment to help you configure SONAS to integrate into your environment, set up file systems, and plan for migration and relocation of data to the SONAS environment.

There are other aspects to completing your deployment, which include the configuration of backup and restore services, replication, and disaster recovery planning. These topics are described in later chapters in this book.

1.2.1 Information that is needed to complete the basic installation and configuration

Your sizing estimates can help ensure that your data center is prepared to receive and install the equipment (power, cooling, floor loading, and access) according to design. To successfully complete a basic installation and configuration of a SONAS appliance, you need the details that are listed in the following sections.

Physical installation

The physical installation information that is listed here is based on a SONAS installation with no Gateway devices:

- How many SONAS racks (frames) are to be installed per site.
- Data center location of the SONAS base rack.
- Number of Interface nodes.
- Number of 10 GbE optical ports per Interface node (goes to total Switch port count).
- Number of 1 GbE RJ45 ports per Interface node (goes to total Switch port count).
- Modem port (if USB-based call home was specified).
- Number of Storage Nodes (Gateway Solution)
- Maximum floor loading for Storage Racks is 1350 kg per rack (342 kg per m^2 (70 lb per ft^2). This value assumes a fully loaded rack (Storage Appliance with DDN storage).
- Power requirements of each rack (4 PDUs, with a minimum of two power circuits).
- BTU/hr rating of each rack (storage racks often have higher BTU/hr rating).

- For each additional frame, the distance from the RXA must be less than 50 meters (IB cable run).
- Proper spacing between racks (see 1.3, “Physical planning” on page 9).
- Clear understanding of rack dimensions, weight, clearance, and tilting limits during transportation from your loading dock to your computer room floor. Exceeding tilting limits per rack can void your warranty.

The SONAS installation setup software also prompts you to verify the locations and serial numbers of various components. This information helps to ensure that it can correctly map the location of the device to the graphical maps in the SONAS GUI Management Tool. The information that is provided here must correctly reflect your implementation to support clarity in ongoing maintenance of the installed product.

Software configuration

The software configuration information that is described in this chapter is based on a SONAS installation with no Gateway storage devices. The following information is required to successfully install and configure your SONAS appliance software:

- SONAS cluster name (Fully Qualified DNS Name (FQDN))
- A private IP address range that your network is not using. This private IP range is for the SONAS inter-component communication or internal data network.
- A range of “customer network” IP addresses to allocate to your SONAS cluster:
  - One IP address for the management console CLI/GUI
  - Management Console Default Gateway IP address
  - Management Console subnet Mask
  - Number of Management nodes
  - Customer Service IP address for the primary Management node
  - Customer Service IP address for the secondary Management node (if there are two Management nodes)
- VLAN Tag for the Service and Management Network
- DNS Servers
- DNS Search criteria and search path
- NTP Server IP address
- Time Zone
- Number of frames (racks) being installed
- Upper InfiniBand Switch Serial number (as read from the label)
- Lower InfiniBand Switch Serial number (as read from the label)
- Authentication method:
  - If Microsoft AD:
    - Domain name
    - IP address of the domain controller
    - Domain Administrator (or equivalent) name and password (or have an administrator ready to insert it at the time of installation)
    - Services For UNIX details

  - If LDAP:
    - Organizational unit
    - UserID / password (or equivalent) name and password (or have an administrator ready to insert it at the time of installation)
    - Kerberos Authentication Server IP address
    - Certificate details
  - SAMBA PDC or Windows NT 4 Domain:
    - Domain Name
    - User name and password of a Domain Account Administrator (or equivalent) name and password (or have an administrator ready to insert it at the time of installation)
    - IP address of the primary domain controller (PDC)

1.2.2 Information that is needed to complete the integration and implementation

To complete successfully the integration and implementation of a SONAS appliance, you need additional information. IBM has a questionnaire spreadsheet in Microsoft Excel format to help you document the details of your environment. This information can help you complete the integration and implementation of your SONAS appliance. Your IBM Account team provides you with SONAS questionnaires to assist with this process before you place your SONAS order. You can use the information for configuring your SONAS order and during the installation and configuration.

At a high level, you must have a clear understanding of the following factors:

- Which applications will be hosted on your new SONAS environment.
- How the data protection requirements are currently influenced by underlying storage.
- How all current application clients currently access each of the data stores and file systems and how those clients are currently authenticated.
- Answers to these questions directly apply to SONAS integration and data migration considerations):
  - Storage requirements.
  - The current data size historical growth curve of your data. Determine how much historical data you need to keep on disk, and what the average growth.
  - The migrated data size and anticipated growth curve of your data. Consider how much future historical data you need to keep on disk (input to SONAS sizing storage tiers, and ILM).
  - Consider the I/O workload profiles and patterns that affect performance or SLA requirements for current applications that are targeted for migration to SONAS (this information is critical to accurate sizing of SONAS cluster nodes, capacity, disk types, and spindle counts in a well-planned SONAS solution).
  - Consider how your applications access data on SONAS. Decide whether data access occurs directly through HTTP, FTP, SCP, or NFS, or whether the application depends on its host operating system to access that data on SONAS on its behalf with CIFS or NFS. Application data access determines file access protocols, certificates, DNS aliases, name space, and data arrangement considerations.
- Consider whether your current NAS appliances are placed behind any firewalls or network load balancers.

- Consider whether you expect SONAS to be placed behind any firewalls or network balancers. If so, you must ensure that the load balancing is transparent and the service availability tracking settings on the load balancer do not work counter to how SONAS load balances its IP addresses.
- If you are using Global redirectors, you must ensure that the redirection of clients to the SONAS on your disaster recovery (DR) site is properly coordinated with the completion of the recovery steps for SONAS replication and failover.
- You must establish your data protection plan to achieve your backup, recovery, and DR RPO and RTO commitments.
- Determine how your SONAS appliance will be used (inputs to setting expectations, migration method, capacity, and performance planning).
- For disaster recovery, determine the distance, data protection criteria, number of files, average file size and daily rate of change, and planned recovery time objective (RTO) and recovery point objective (RPO) of your replication requirements (inputs to sizing your solution at both sites for Interface nodes and capacity).
- Growth plans in terms of physical site expansion floor space, power, network, and cooling. This information helps you understand growth room for space, power, heating, and network ports.
- Consider how the appliance can be supported, both by IBM and by your operations team. It factors into ensuring a smooth and synergetic support experience.
- Consider your current hardware and software support arrangement with IBM and how notification of events is to be managed.

1.2.3 Collecting requirements from the questionnaires

As a pre-sales activity, your IBM Account team presents you with SONAS questionnaires that can help IBM understand all the important information and requirements of your particular solution. It is important to take the time that is required to respond as completely and accurately as possible to ensure that cluster design is optimal from a cost and sizing perspective. It also helps to ensure that proper resource involvement is engaged early and no surprises result at the time of installation. Later, as you choose to add more services to your scale out NAS solution, conduct a re-evaluation of cluster sizing with your IBM representative to ensure that the system is scaled appropriately before new services are migrated. The following section lists sample topics that the questionnaires cover.

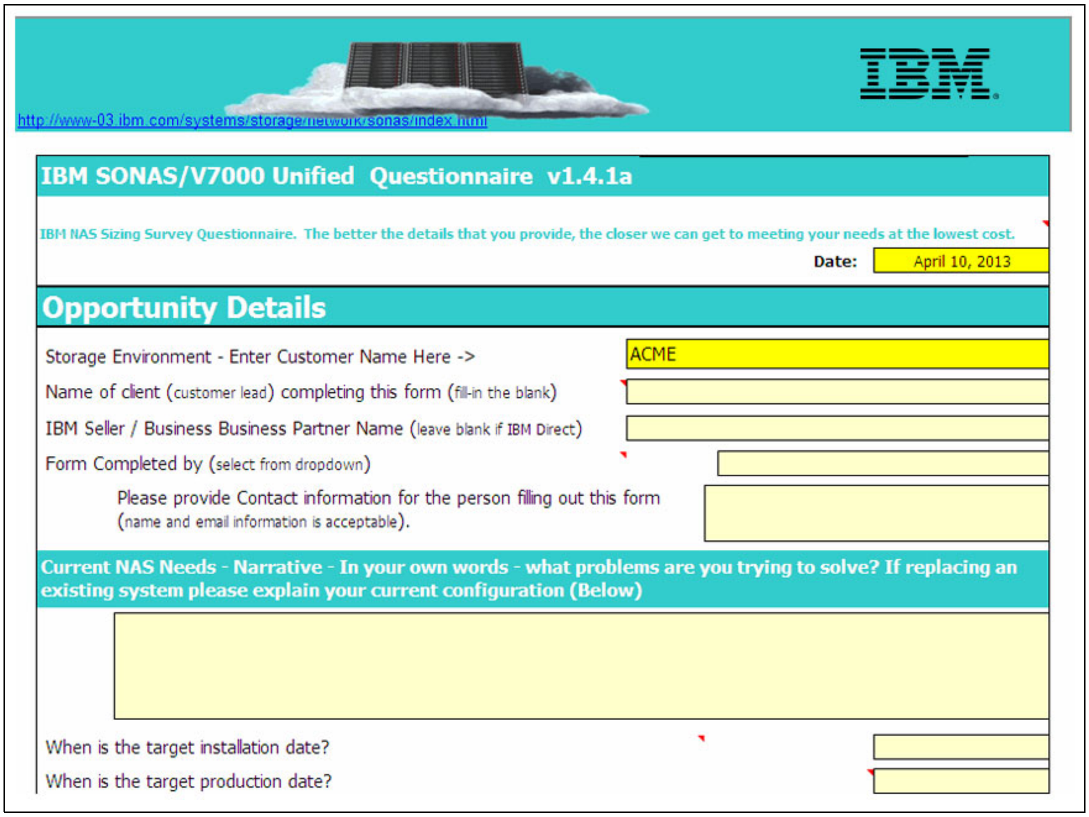
Sample SONAS solution preparation questionnaire

The purpose of the solution preparation questionnaire is to capture as much information that might help clarify the environment in which SONAS will be installed. The questionnaire also serves as a single place to capture current usage patterns within your environment, systems, and protocols on which you depend. It serves as a single point of reference for what must be done to facilitate a successful implementation post migration.

It is important to note that the SONAS appliance is an NAS solution. It serves file I/O over CIFS, NFS, FTP, HTTP, and SCP. It does not serve block I/O through Fibre Channel or iSCSI. Although there are many similarities between SONAS and other NAS products, there are also differences. The questionnaire makes it easier to qualify SONAS as an appropriate solution for your goal and to plan how the solution must be prepared, installed, implemented, and integrated into your environment.

To help accurate collect this data, IBM has several tools that are available in the form of *client questionnaires*. A version of this questionnaire is available as an MS Excel Spreadsheet called `IBM_SONAS_Requirements_Survey.xls`. Your IBM representative can provide access to the latest edition of this spreadsheet, in cooperation with the IBM SONAS development team.

Figure 1-2 shows some key sections of the questionnaire that focus on some of the pertinent solution design considerations. This chapter assumes that the necessary data was collected and that the solution was properly sized.



http://www-03.ibm.com/systems/storage/network/sonas/index.html

IBM SONAS/V7000 Unified Questionnaire v1.4.1a

IBM NAS Sizing Survey Questionnaire. The better the details that you provide, the closer we can get to meeting your needs at the lowest cost.

Date: April 10, 2013

Opportunity Details

Storage Environment - Enter Customer Name Here ->	ACME
Name of client (customer lead) completing this form (fill-in the blank)	
IBM Seller / Business Business Partner Name (leave blank if IBM Direct)	
Form Completed by (select from dropdown)	
Please provide Contact information for the person filling out this form (name and email information is acceptable).	

Current NAS Needs - Narrative - In your own words - what problems are you trying to solve? If replacing an existing system please explain your current configuration (Below)

When is the target installation date?	
When is the target production date?	

Figure 1-2 Screen capture of the SONAS Pre-Design Questionnaire

Questions from the NAS Migration Questionnaire

This section lists some questions from the NAS Migration Questionnaire. There are more questions that are listed in the questionnaire. This information helps you and IBM to ensure that you consider everything that is required for a successful migration and implementation.

The following list provides examples of the type of data that is collected:

- Customer Name
- Total number of NAS file servers to be migrated
- Total amount of TB to be migrated
- Total number of files (in millions)
- Total number of file systems
- Protocols In use:
  - CIFS
  - NFS
  - Multiprotocol
  - FTP
  - HTTP
  - Other

- NAS platforms to be migrated:
  - EMC Celerra
  - NetApp
  - Other
- Client environment:
  - Total number of clients
  - Total number of concurrent Clients
  - Client Networking (1 GbE, 10 GbE, or Mixed)
- Client platforms in use:
  - Windows
  - IBM AIX
  - Linux (all types)
  - Samba
  - Mac OSX
  - Solaris
- Network environment

1.3 Physical planning

This section describes the physical installation of your SONAS racks, including power, weight, and spacing considerations and distances from the main rack. It considers the three different configurations: single rack (small), 2 -7 racks (medium), and 8 - 15 racks (large).

This section first briefly reviews the rack types and the components they contain. For more information about the SONAS hardware, see the SONAS information in the IBM Knowledge Center found at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/landing/sonas_151_kc_welcome.html

1.3.1 Rack types

There are four basic rack types for SONAS configurations:

- Base rack (type RXA)

  RXA Gateway (Feature Code 9006, 9007, or 9008) has two 36-port InfiniBand switches, up to eight Interface nodes, and up to 10 Storage nodes (five Storage node pairs) for connectivity to XIV, Storwize V7000, DCS3700, and RPQ IBM DS8000 storage (see Chapter 3, “Installation and configuration for SONAS Gateway solutions” on page 97).

  > **Note:** Base Rack RXA1 (FC 9003), RXA2 (FC 9004), and RXA3 (FC 9005) are discontinued. The SONAS Appliance with internal DDN storage is no longer available for new installations. New installations are sold as SONAS Gateway solutions only. Existing Appliance customers with the internal DDN storage are supported and can expand their existing system as needed.

- Interface expansion rack RXC (type RXC)
- Interface and storage expansion rack RXC (type RXC with iRPQ)
- Storage expansion rack RXB (type RXB)

> **Note:** The RXB Storage expansion rack is available for existing SONAS Appliance (internal DDN storage) configurations only. It not available for SONAS Gateway configurations. For SONAS Gateway configurations, the RXC rack can be ordered as an expansion to the RXA base rack for added Interface and Storage node capacity.

Base rack RXA

The SONAS system always contains a base rack that contains the Management node (for versions before SONAS V1.2 when integrated Management nodes were not available), two 36-port InfiniBand switches, a keyboard, video, and mouse (KVM) unit, a minimum of two Interface nodes, and a minimum of two Storage nodes. Figure 1-3 shows the base SONAS rack that is based on Generation 2 hardware (with integrated Management nodes).

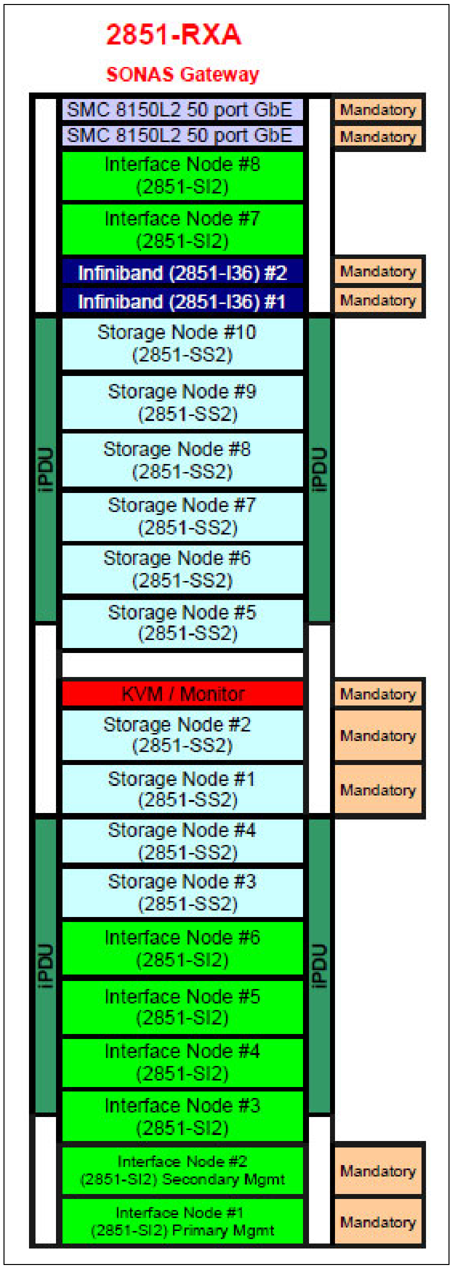

2851-RXA
SONAS Gateway
SMC 8150L2 50 port GbE
Mandatory
SMC 8150L2 50 port GbE
Mandatory
Interface Node #8 (2851-SI2)
Interface Node #7 (2851-SI2)
Infiniband (2851-I36) #2
Mandatory
Infiniband (2851-I36) #1
Mandatory
Storage Node #10 (2851-SS2)
Storage Node #9 (2851-SS2)
Storage Node #8 (2851-SS2)
Storage Node #7 (2851-SS2)
Storage Node #6 (2851-SS2)
Storage Node #5 (2851-SS2)
iPDU
iPDU
KVM / Monitor
Mandatory
Storage Node #2 (2851-SS2)
Mandatory
Storage Node #1 (2851-SS2)
Mandatory
Storage Node #4 (2851-SS2)
Storage Node #3 (2851-SS2)
Interface Node #6 (2851-SI2)
Interface Node #5 (2851-SI2)
Interface Node #4 (2851-SI2)
Interface Node #3 (2851-SI2)
iPDU
iPDU
Interface Node #2 (2851-SI2) Secondary Mgmt
Mandatory
Interface Node #1 (2851-SI2) Primary Mgmt
Mandatory

Figure 1-3 RXA rack configuration

SONAS base rack RXA feature code 9006, 9007, or 9008

This base rack has two Gigabit Ethernet (GbE) switches at the top of the rack, two 36-port InfiniBand switches, a KVM, at least two Interface nodes, and at least two Storage nodes.

In a Gateway solution, there is no physical storage with the rack. All storage must be ordered external to the SONAS base rack or expansion rack. In the base rack, you are limited to 2x36 InfiniBand ports. These switches cannot be expanded or exchanged. The base rack can be expanded to a total of 18 nodes (eight Interface nodes and 10 Storage nodes). You can add another interface expansion rack to the base rack (see “Interface expansion rack RXC” on page 12). The SONAS cluster can be scaled out (expanded) to a total of 34 nodes across the base rack and expansion rack, which consist of a combination of Interface nodes and Storage nodes.

Rack specifications:

- The rack must have two 50-port 10/100/1000 Ethernet switches for internal IP management network, two 36-port InfiniBand switches, and a KVM.
- The rack must have a minimum of two Interface nodes. The rest of the Interface node bays are expandable options (up to eight Interface nodes in this rack).
- The rack must have a minimum of two Storage nodes. The rest of the Storage node bays are expandable options (up to 10 Interface nodes in this rack).
- The first two Interface nodes take on the roles of Management nodes in Generation 2 hardware and beyond, and as such have a third hard disk drive (HDD) installed for log collection.

Interface expansion rack RXC

The IBM SONAS Interface expansion rack extends the number of Interface nodes or Storage nodes to an existing base rack by providing up to 20 more Interface nodes and Storage nodes. The total number of nodes cannot exceed 34 nodes between the base rack and expansion rack. The two 50-port 10/100/1000 Ethernet switches and at least one Interface node per Interface expansion rack are mandatory. Figure 1-4 shows an Interface expansion rack. There can be only one interface expansion rack in a SONAS configuration.

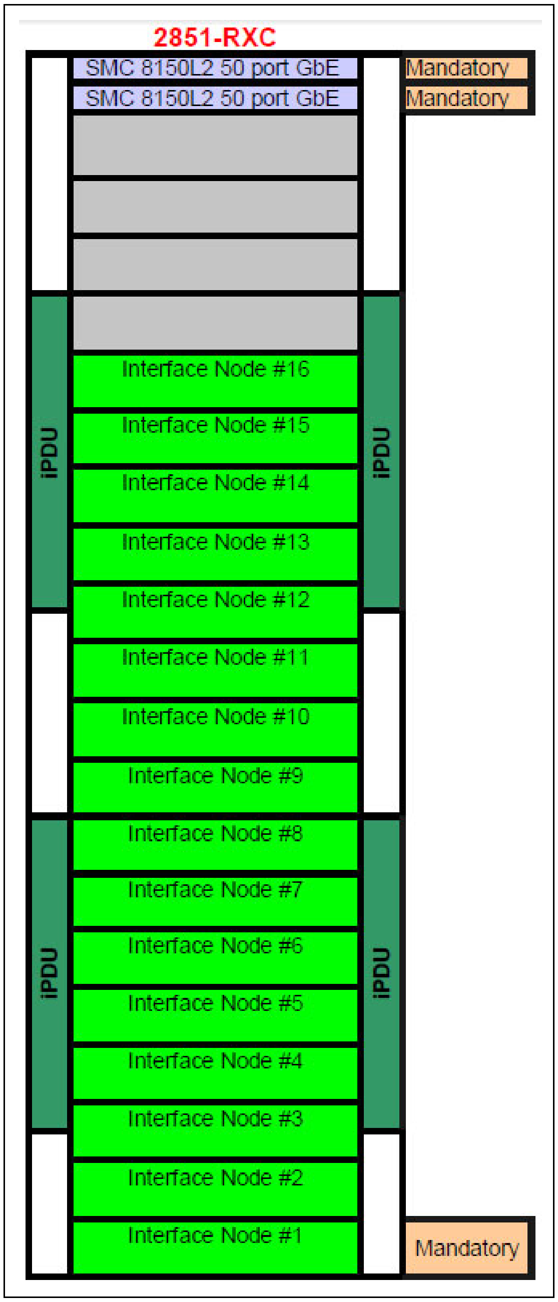


Figure 1-4 Interface expansion rack

Figure 1-5 on page 13 shows a SONAS Gateway interface and storage expansion rack.

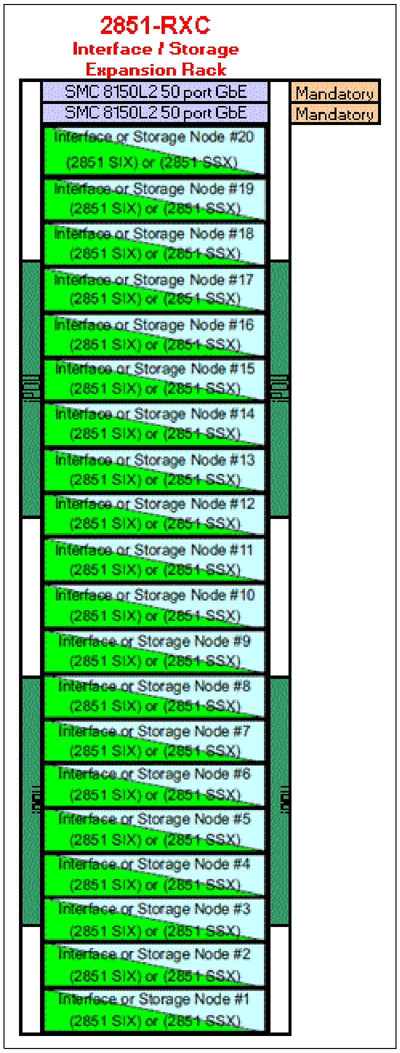


Figure 1-5 SONAS Gateway interface and storage expansion rack

Rack specifications:

- The rack must have two 50-port 10/100/1000 Ethernet switches for an internal IP management network.
- In SONAS Gateway configurations, the rack must have a minimum of one Interface node. The rest of the bays are expandable with a combination of Interface and Storage nodes. A SONAS Gateway configuration cannot exceed 34 nodes.
- In a SONAS Appliance configuration (Internal DDN Storage), a total of 20 more Interface nodes can be configured in the RXC rack. A SONAS appliance configuration cannot exceed 30 nodes (Interface *or* Storage)
- For SONAS V1.4 and later releases, the RXC cabinet allows for a mix of Interface and Storage nodes with Storage nodes installed in pairs only (other restrictions might apply).

Storage expansion rack RXB

The 2851-RXB storage expansion rack extends the storage capacity of an existing base rack by adding up to two more Storage pods per rack. Each Storage pod consists of two Storage nodes and up to two storage controllers and two storage expansion units. This configuration means that each Storage pod can hold up to 240 disks, which can be of different types in groups of 60 disks. Figure 1-6 shows the storage expansion rack.

> **Note:** The RXB Storage expansion rack is available for existing SONAS Appliance (internal DDN storage) configurations only. It is not available for SONAS Gateway configurations. For SONAS Gateway configurations, the RXC rack can be ordered as an expansion to the RXA base rack for added Interface and Storage node capacity.

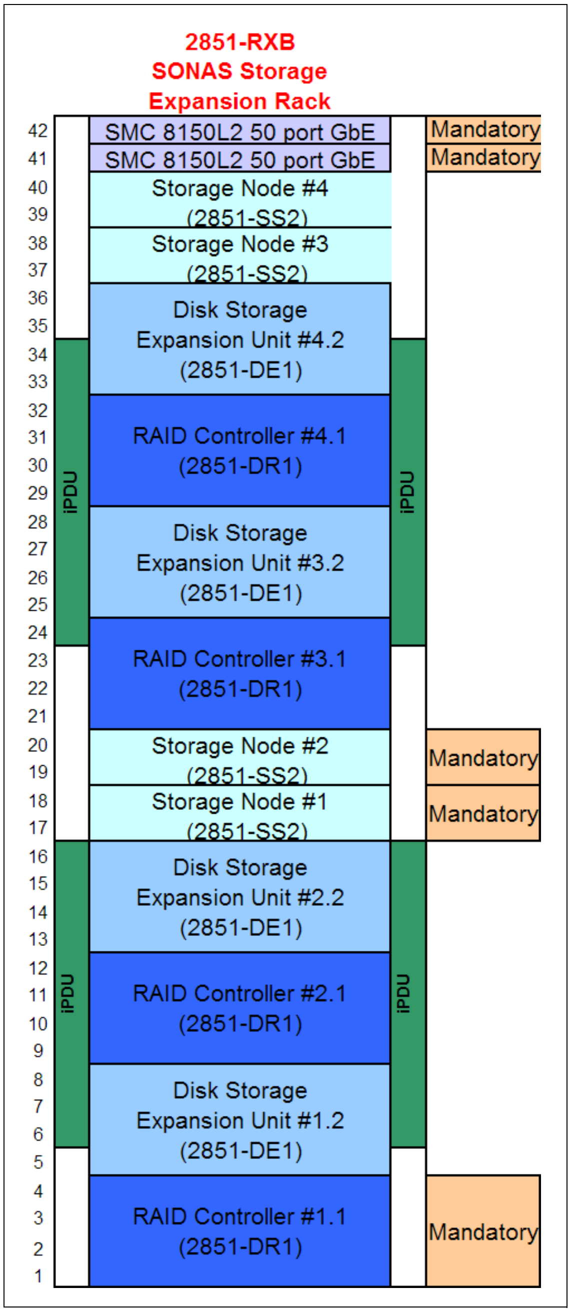


Figure 1-6 SONAS Storage Expansion Rack RXB

Rack density and weight considerations

One of the benefits of SONAS is its ability to scale to large capacities by using high-density storage. With the current deployment models, up to 480 disks per rack (288 TB if you are using 600 GB SAS or 1 PB if you are using 3 TB NL-SAS) can be deployed. It is half to one third of the typical space that is required in a data center to hold the same capacity. By consolidating more disk into a smaller space, you reduce the overall power consumption but increase the weight per square meter. This consolidation means that special consideration must be made for the floor loading of your data center and any restrictions that you might have on power distribution per rack space.

Although SONAS uses less power than might typically be required on less dense storage systems, it can still exceed any restrictions that your facility might have regarding power consumption per square meter or per tile. Therefore, your power constraints must also be considered. Additionally, it is worth noting that IBM RXA, RXB, and RXC racks are certified to meet the necessary weight, power, and environmental requirements.

Additional constraints affect the shipping logistics of the SONAS solution to your data center floor. They include loading dock, elevator, or shipping containers, and setup considerations, such as the use of raised floor or overhead rack cabling. For more information, see *IBM Scale Out Network Attached Storage Introduction and Planning Guide*, GA32-0716.

Review your configuration with the IBM team to determine your floor loading, power, cooling, and cabling distribution requirements against the hardware specifications for SONAS.

> **Note:** The RXB rack is for the SONAS Appliance (internal DDN storage) and does not support integration of Gateway storage.

1.3.2 Environment

The following requirements must be considered for the SONAS Appliance (non-Gateway) configurations and be applied to your data center location.

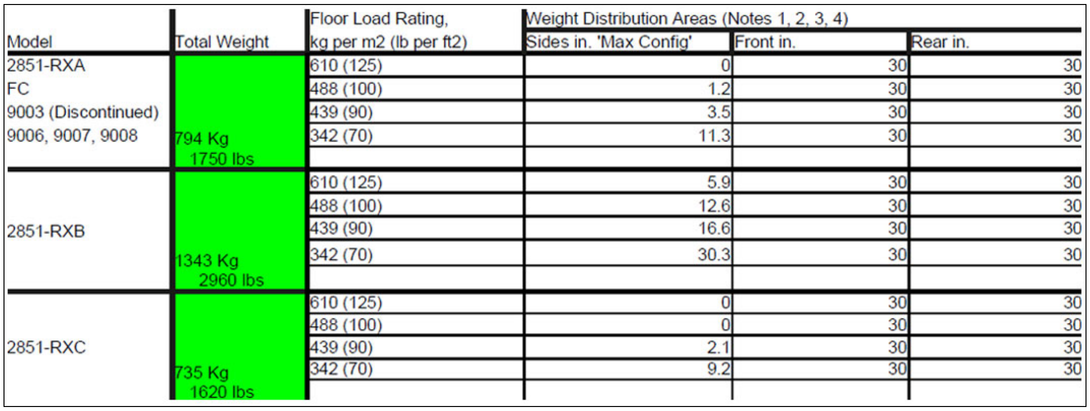
Floor load requirements

Your data center location must meet the floor load requirements.

To ensure that your location meets the floor load requirements and to determine the weight distribution area that is required for the floor load, complete the following steps:

1. Discover the floor load rating of the location where you plan to install the systems.
2. Determine whether the floor load rating of the location meets the minimum floor load rating that is used by IBM, which is 342 kg per m^2 or 70 lb per ft^2.

3. Using the table in Figure 1-7, complete the following steps for each system:
   a. Find the rows that are associated with the system model type.
   b. Locate the configuration row that corresponds with the floor load rating of the site.
   c. Identify the weight distribution area that is needed for that storage unit and floor load rating.



		Floor Load Rating,	Weight Distribution Areas (Notes 1, 2, 3, 4)		
Model	Total Weight	kg per m2 (lb per ft2)	Sides in. 'Max Config'	Front in.	Rear in.
2851-RXA FC 9003 (Discontinued) 9006, 9007, 9008	794 Kg 1750 lbs	610 (125)	0	30	30
		488 (100)	1.2	30	30
		439 (90)	3.5	30	30
		342 (70)	11.3	30	30
2851-RXB	1343 Kg 2960 lbs	610 (125)	5.9	30	30
		488 (100)	12.6	30	30
		439 (90)	16.6	30	30
		342 (70)	30.3	30	30
2851-RXC	735 Kg 1620 lbs	610 (125)	0	30	30
		488 (100)	0	30	30
		439 (90)	2.1	30	30
		342 (70)	9.2	30	30

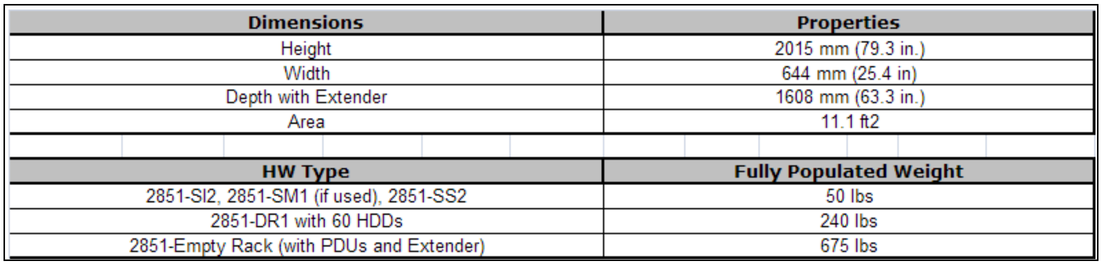
Figure 1-7 Weight distribution area per SONAS rack

The rack that is used for the SONAS program is the IBM Enterprise Class 42U rack. This rack addresses the special requirements of customers who want a tall enclosure to house the maximum amount of equipment in the smallest possible floor space.

The 42U rack is 2.0 m (79.3 inches) tall. For SONAS, IBM also offers an 8-inch rack extension kit for the enterprise rack, which makes the entire package of rack and extension 4 feet deep. The extension contains management brackets and a patch panel to simplify cabling requirements. The extension kit mounts to the rear door hinge points of the base enterprise rack.

> **Note:** By default, IBM does not support the unracking of SONAS hardware and reracking it into customer-provided racks (frames). Any such request requires special case authorization and a special quotation from the IBM SONAS Development Team leadership.

When you determine the appropriate weight distribution area by following the IBM specifications that are described in Figure 1-7, the weight distribution areas are wider than the actual rack, to avoid possible over loading and risking tilting of the storage racks. All values that are listed assume a fully loaded rack and represent an individual frame model only. Figure 1-8 shows the rack dimensions and individual node and storage controller weights.



Dimensions	**Properties**
Height	2015 mm (79.3 in.)
Width	644 mm (25.4 in)
Depth with Extender	1608 mm (63.3 in.)
Area	11.1 ft2
HW Type	**Fully Populated Weight**
2851-SI2, 2851-SM1 (if used), 2851-SS2	50 lbs
2851-DR1 with 60 HDDs	240 lbs
2851-Empty Rack (with PDUs and Extender)	675 lbs

Figure 1-8 Rack dimensions and individual weights

Space clearance

Assume that the sizing that you did led to a configuration with one Base rack and two Storage Expansion racks, which can be set up on the same row in your data center. You must ensure first that all SONAS racks have at least 762 mm (30 in.) of free space in front and in the back of them to account for air flow and temperature regulation requirements. Additionally, depending on the type of rack and the weight distribution areas, they also have 155 + 313 = 468 mm or 313 + 313 = 626 mm between them, as described in Figure 1-9.

Weight distribution:

- Weight distribution areas cannot overlap.
- Weight distribution areas are calculated for the maximum weight of the models.

Weight distribution areas are calculated for individual frame model types only. If you want to install multiple frames in the same row, or in the same lab area, the space between each frame is cumulative.

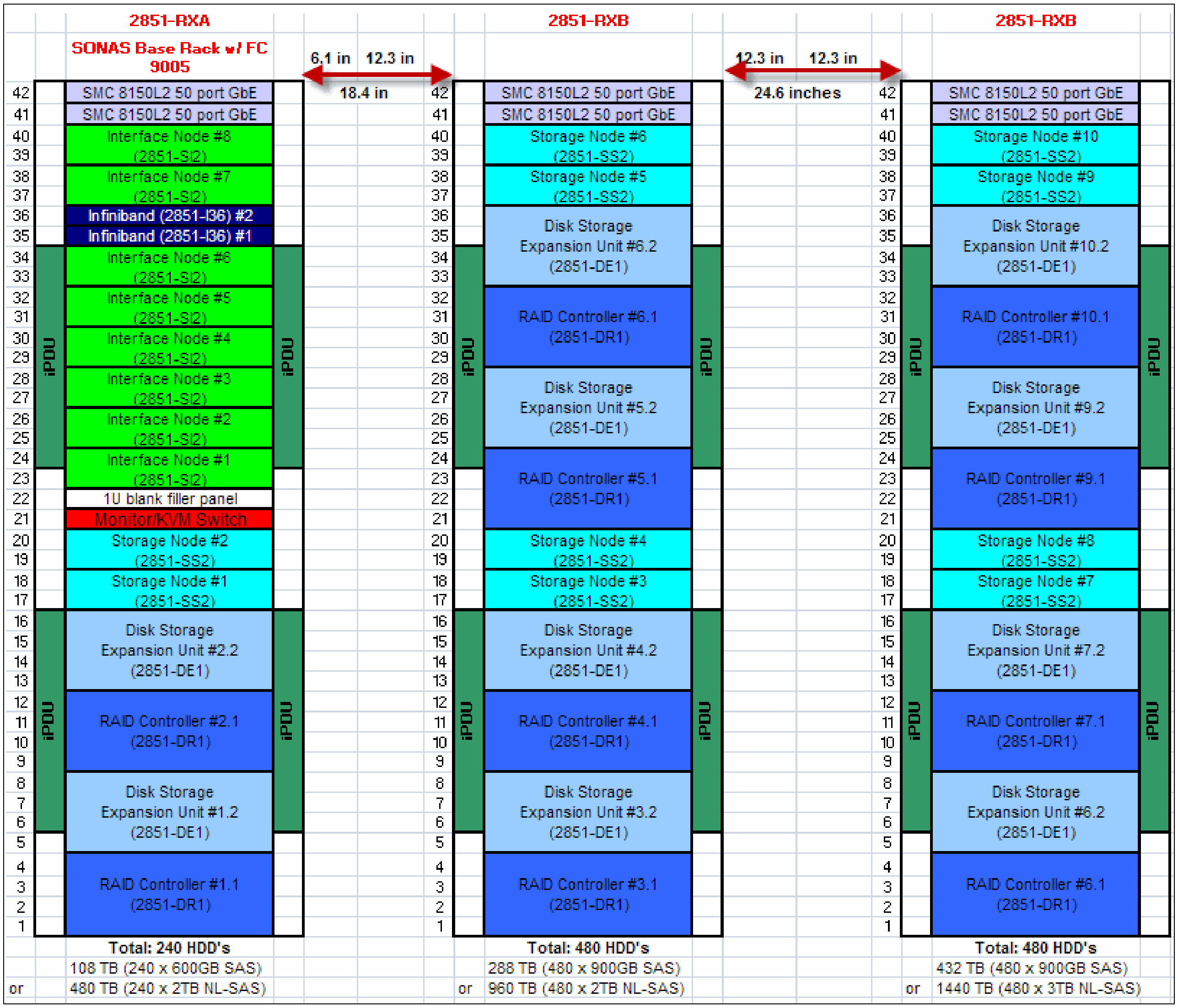

2851-RXA
SONAS Base Rack w/ FC 9005
6.1 in
12.3 in
18.4 in
2851-RXB
12.3 in
12.3 in
24.6 inches
2851-RXB
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Interface Node #8 (2851-SI2)
Interface Node #7 (2851-SI2)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Interface Node #6 (2851-SI2)
Interface Node #5 (2851-SI2)
Interface Node #4 (2851-SI2)
Interface Node #3 (2851-SI2)
Interface Node #2 (2851-SI2)
Interface Node #1 (2851-SI2)
1U blank filler panel
Monitor/KVM Switch
Storage Node #2 (2851-SS2)
Storage Node #1 (2851-SS2)
Disk Storage Expansion Unit #2.2 (2851-DE1)
RAID Controller #2.1 (2851-DR1)
Disk Storage Expansion Unit #1.2 (2851-DE1)
RAID Controller #1.1 (2851-DR1)
Total: 240 HDD's
108 TB (240 x 600GB SAS)
or 480 TB (240 x 2TB NL-SAS)
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Storage Node #6 (2851-SS2)
Storage Node #5 (2851-SS2)
Disk Storage Expansion Unit #6.2 (2851-DE1)
RAID Controller #6.1 (2851-DR1)
Disk Storage Expansion Unit #5.2 (2851-DE1)
RAID Controller #5.1 (2851-DR1)
Storage Node #4 (2851-SS2)
Storage Node #3 (2851-SS2)
Disk Storage Expansion Unit #4.2 (2851-DE1)
RAID Controller #4.1 (2851-DR1)
Disk Storage Expansion Unit #3.2 (2851-DE1)
RAID Controller #3.1 (2851-DR1)
Total: 480 HDD's
288 TB (480 x 900GB SAS)
or 960 TB (480 x 2TB NL-SAS)
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Storage Node #10 (2851-SS2)
Storage Node #9 (2851-SS2)
Disk Storage Expansion Unit #10.2 (2851-DE1)
RAID Controller #10.1 (2851-DR1)
Disk Storage Expansion Unit #9.2 (2851-DE1)
RAID Controller #9.1 (2851-DR1)
Storage Node #8 (2851-SS2)
Storage Node #7 (2851-SS2)
Disk Storage Expansion Unit #7.2 (2851-DE1)
RAID Controller #7.1 (2851-DR1)
Disk Storage Expansion Unit #6.2 (2851-DE1)
RAID Controller #6.1 (2851-DR1)
Total: 480 HDD's
432 TB (480 x 900GB SAS)
or 1440 TB (480 x 3TB NL-SAS)
iPDU

Figure 1-9 Floor loading and spacing example with different types of storage

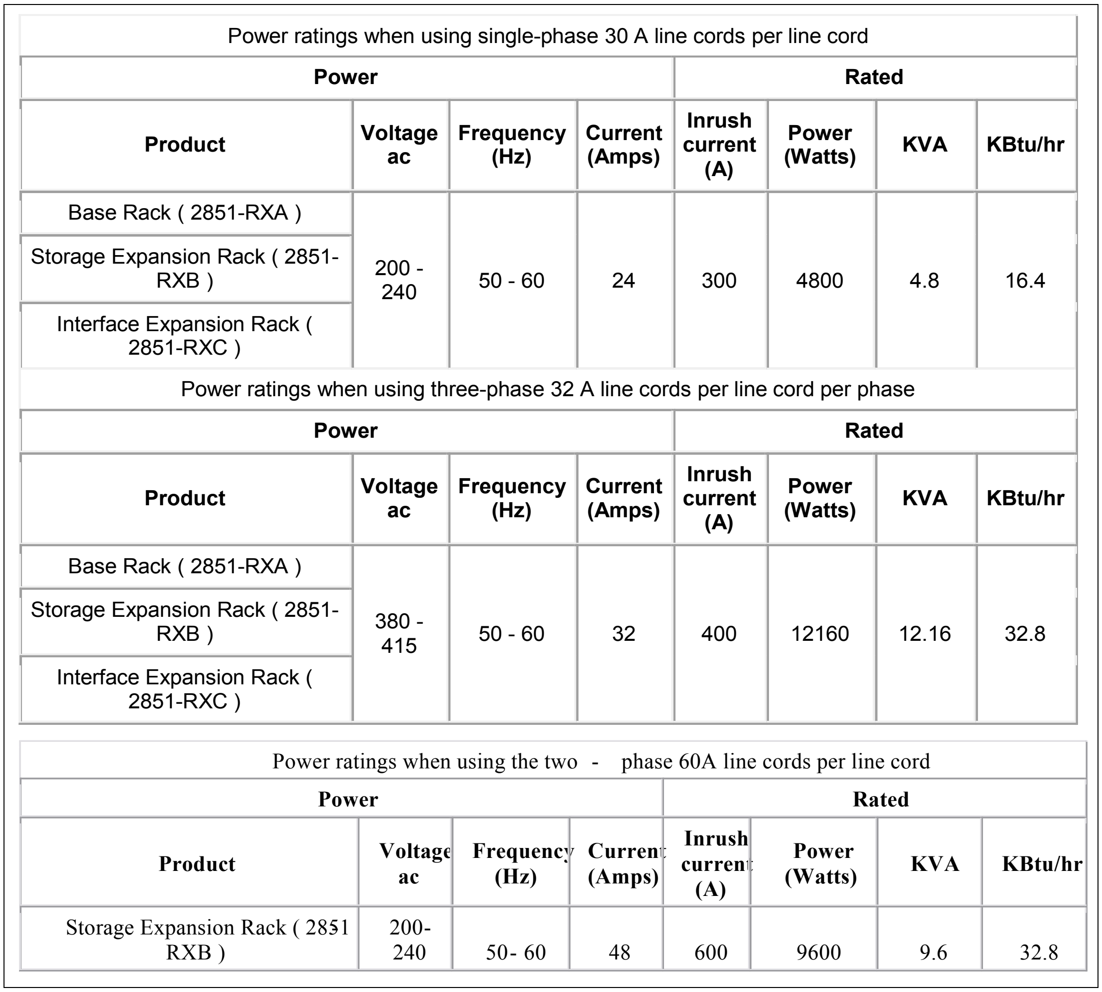
Weight distribution areas are calculated for individual frame model types only. If you want to install multiple frames in the same lab area, the space between each frame is cumulative. For example, perhaps you want to install a system that consists of a 2851-RXA with feature code 9005 and two 2851-RXB expansion frames in the same row. Your floor load rating is 100 pounds per square foot. You must place the RXA a minimum of 18.4 inches from the first RXB. The first RXB must be a minimum of 24.6 inches from the second RXB. These numbers are determined by the sum of the weight distribution areas for both of your frame model types.

> **Note:** RXA feature codes 9003, 9004, and 9005 are discontinued. RXB racks are available for existing SONAS Appliance (internal DDN storage). RXB is not available for SONAS Gateway installations. RXC can be configured with a mixture of Interface and Storage nodes. An RXA and RXC rack for Gateway storage can be configured with a total of 34 nodes between them.

SONAS power requirements

Ensure that your operating environment meets the proper ac power and voltage requirements.

Each of the power distribution units (PDUs) contains twelve 200 - 240 V ac outlets that provide power to the drawers and devices in the rack, as shown in Figure 1-10.



Power ratings when using single-phase 30 A line cords per line cord

Power				Rated			
Product	**Voltage ac**	**Frequency (Hz)**	**Current (Amps)**	**Inrush current (A)**	**Power (Watts)**	**KVA**	**KBtu/hr**
Base Rack (2851-RXA)	200 - 240	50 - 60	24	300	4800	4.8	16.4
Storage Expansion Rack (2851-RXB)							
Interface Expansion Rack (2851-RXC)							

Power ratings when using three-phase 32 A line cords per line cord per phase

Power				Rated			
Product	**Voltage ac**	**Frequency (Hz)**	**Current (Amps)**	**Inrush current (A)**	**Power (Watts)**	**KVA**	**KBtu/hr**
Base Rack (2851-RXA)	380 - 415	50 - 60	32	400	12160	12.16	32.8
Storage Expansion Rack (2851-RXB)							
Interface Expansion Rack (2851-RXC)							

Power ratings when using the two - phase 60A line cords per line cord

Power				Rated			
Product	**Voltage ac**	**Frequency (Hz)**	**Current (Amps)**	**Inrush current (A)**	**Power (Watts)**	**KVA**	**KBtu/hr**
Storage Expansion Rack (2851 RXB)	200- 240	50- 60	48	600	9600	9.6	32.8

Figure 1-10 SONAS frame power requirements

Each rack has either four intelligent PDUs (iPDUs) or four base PDUs. The iPDUs collect energy use information from energy-management components in IBM devices and report the data to the Active Energy Manager feature of IBM Systems Director, if it is installed on a customer server. IBM Systems Director can measure and monitor power consumption.

Each rack requires four power cords, or two feature codes. Each power cord feature is two cords.

Preliminary power consumption test results are for heavy usage scenarios, where each rack is fully populated, running input/output operations, and using 15 K RPM or 10 K RPM SAS drives that use more power than the nearline SAS drives.

The 6 A power cord feature (FC 9859) is offered for use on SONAS 2851- RXB racks when you are running one or more storage enclosures that are populated with either 480 x 600 GB or 480 x 900 GB SAS HDDs. Because this configuration draws more power, either the 60 A power cords or 32 A three-phase power cords are required.

These power options are listed in the IBM Scale Out Network Attached Storage (SONAS) V 1.5.1 product documentation in the IBM Knowledge Center found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/landing/sonas_151_kc_welcome.html

Power cords:

The choice of power cord is not a power issue. It is an issue with the rated power for the power cords. The standard rack power cords are 30 A, which is what customers traditionally have in their data centers. Power cords are industry (not IBM) rated at 80%, so 24 A is the maximum you can load through each cord to avoid spikes and surges that trip the data center’s circuit breakers.

SONAS uses two primary and two secondary power cords (four per rack). At 24 A per power cord, you have 48 A to power the entire rack. Assume that one circuit can handle the full load. At 200 V input, that is 9600 W.

A fully populated 480 drive RXB rack with 2 TB nearline SAS uses 9,522 W. If you fully populate the rack with SAS drives, it uses over 12,000 W, which cannot be physically handled by 24 A power cords.

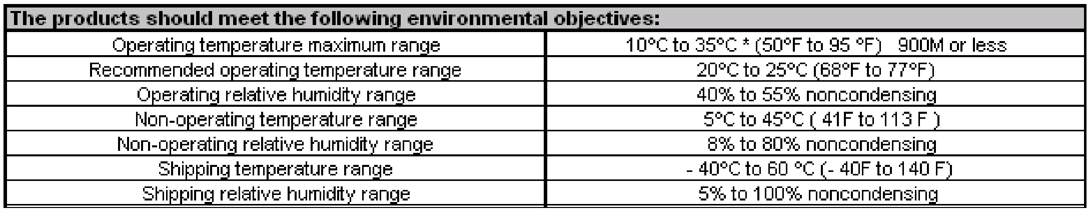
The PDUs in SONAS can handle 48 A per PDU, so the 24 A power cord does not stress the PDU to the maximum. Because SONAS has four PDUs, you can use a 60 A power cord, which is industry-rated at 80%, or 48 A, and use the same PDUs.

By using 60 A power cords, you effectively double the amount of power that the power cords can hold. This doubling lifts the drive restriction on RXB racks and SAS drives. The same situation applies to the 32 A three-phase cords, which are predominantly used in EMEA.

Heat and cooling

To optimize the cooling system of your SONAS storage solution, you can use a raised floor to increase air circulation in combination with perforated tiles. For more information about such a configuration, see *SONAS Introduction and Planning Guide,* GA32-0716.

For information about the temperature and humidity while the system is in use or shut down, see Figure 1-11.



The products should meet the following environmental objectives:	
Operating temperature maximum range	10°C to 35°C * (50°F to 95 °F) 900M or less
Recommended operating temperature range	20°C to 25°C (68°F to 77°F)
Operating relative humidity range	40% to 55% noncondensing
Non-operating temperature range	5°C to 45°C (41F to 113 F)
Non-operating relative humidity range	8% to 80% noncondensing
Shipping temperature range	- 40°C to 60 °C (- 40F to 140 F)
Shipping relative humidity range	5% to 100% noncondensing

Figure 1-11   Cooling measurements

Noise

Based on acoustics tests that were performed for a SONAS system, these values apply:

- 90 dB registered for a fully populated Base rack (2851-RXA) system
- Up to 93 dB in the worst scenario with a fully populated Storage Expansion Rack (2851-RXB)

The system operating acoustic noise specifications are Declared Sound Power Level, LwAd is less than 94 dBA at 1 m at 23 ºC.

However, you can reduce the audible sound level of the components that are installed in each rack by up to 6 dB with the acoustic doors feature (feature code 6249 for each SONAS rack).

1.4 Installation plan

This section describes the installation steps after the equipment is physically connected. Here you can gain an understanding of what other information you need to complete the installation phase of SONAS.

1.4.1 Installation workflow

This section begins with an overview of the installation work flow. The walk-through for installation is described in Chapter 2, “Installation and configuration for a SONAS appliance with DDN” on page 61 and Chapter 3, “Installation and configuration for SONAS Gateway solutions” on page 97. The installation phase takes you through the power-on and first-time configuration steps. In the configuration phase that follows, you integrate SONAS into your environment and set up authentication, file systems, exports, and backups.

For the installation phase to be completed, you or IBM personnel need physical access to the Management node (the first Interface node) of the SONAS appliance. Access is necessary to update the code and starting the software installation.

Here is the installation workflow:

- Physical installation and cabling.
- Power distribution.
- Power on storage enclosures (Appliance (DDN Storage)).
- Power on storage controllers (Appliance (DDN Storage)).
- Create and allocate volumes and LUNs to SONAS Storage nodes (Gateway Configurations).
- Power on the first node and run the initial DVD installation.

- Run the first-time installation script and enter the information that is listed in Table 1-1 through Table 1-4 on page 24.
- Power on Storage nodes.
- Power on Interface nodes.
- Verify the node position against the node discovery phase results.
- Complete the first-time installation script.
- Verify hardware wellness.
- Configure cluster services.
- Enable license.
- Configure the command-line interface (CLI).

The steps that are listed outline what you are asked to do in Chapter 3, “Installation and configuration for SONAS Gateway solutions” on page 97. Start or power on equipment *only* in the sequence that is specified. For a small configuration, the installation can take about 3 hours. For a medium configuration, the installation can take up to 5 hours. For a large configuration, the installation can take 8 - 15 hours, depending on the number of nodes.

1.4.2 Installation checklist questions

To complete the installation phase, you need the information that is listed in the tables in this section. This information is critical and is required by the tasks that are described later in this book.

- Management node configuration (Table 1-1)
- Quorum topology (Table 1-2 on page 23)
- CLI credentials (Table 1-3 on page 23)
- Node locations (Table 1-4 on page 24)

Table 1-1 Management node configuration

Question #	Field	Value	Notes
1	Cluster Name		The name of your IBM SONAS cluster. Example: *sonascluster* This name becomes the default NAS (computer) name that you register in DNS and in AD. Windows hosts use this name in their UNC path names when mapping drives.
2	Domain Name		Your network domain name. Example: `mydomain.com` The Cluster Name and Domain Name are typically used in combination. Example: `sonascluster.mydomain.com` The name that is used by HTTP, FTP, and SCP clients. NFS clients can also use this name, but because of NFSv3 limitations, it is better to use IP addresses when you map NFS clients to exports on SONAS.

Question #	Field	Value	Notes
3	Internal IP address Range	Specify 1, 2, or 3.	SONAS requires an IP address range for use within its internal data backbone. You must use a predetermined range, but the range must not conflict with your existing network configuration for your users and hosts. Here are the available IP address ranges: 1. 172.31.*.* 2. 192.168.*.* 3. 10.254.*.* If you are already using the first range, choose the second range. If you are using both the first and second ranges, choose the third range.
4	Management console IP address		This IP address is associated to the Management node. It must be on the public network and accessible by the storage administrator.
5	Management console gateway		The Gateway IP address of your Management console.
6	Management console subnet mask		The Subnet Mask of your Management console.
7	Management Console Service IP		An IP that is reserved to Management node functions in the Version 1.3 or above cluster configuration.
8	Management Console VLAN ID		Optional. A list of one or more Virtual LAN Identifiers. It is useful only if you have turned on VLAN Tagging on your Ethernet (or aggregated) ports. The VLAN ID must be 2 - 4095. If you do not use VLANs, leave this field blank.
9	Host name	mgmt001st001.	Your preassigned Management node host name.
10	Root Password		You can specify the password that you want to be set on the Management node for root access. By default, it is Passw0rd (where P is capitalized and 0 is zero).
11	NTP Server IP address		For certificate-based authentication (Microsoft AD and LDAP through Kerberos), this address must be in time sync with your authentication server. NTP is used to ensure time accuracy, and SONAS must be configured to use the same server as your authentication domain. A second NTP Server is best for redundancy. **Notes:** ▸ The Network Time Protocol (NTP) Servers) can be either local or on the internet. ▸ Only the Management node requires a connection to your NTP server, and it becomes the NTP server for the whole cluster.

Question #	Field	Value	Notes
12	Time Zone		A numeric index that refers to the time zone list that is provided in the installation script. Specify the number that corresponds to your location.
13	Number of frames that are being installed		Specify the total quantity of rack frames in this cluster (minimum = 1).

SONAS uses IBM GPFS as the internal clustered file system. In this file system, quorum nodes provide a mechanism to avoid split-brain scenarios if there is a disaster. You must specify an odd number for quorum nodes in the cluster. In Table 1-2, you provide the quorum topology of your SONAS system. During the installation process, SONAS recommends the quorum nodes. This value must be changed only for complex multi-rack solutions. The quorum nodes help ensure the integrity of the file system if there is a disaster.

Table 1-2 Quorum topology

Question #	Field	Value (defaults)	Notes
14	Quorum Storage nodes	strg001st001 strg002st001	1. Your first action is to select an odd number of Quorum Nodes. You can use both Interface and Storage nodes. Valid choices are 3, 5, or 7. 2. If your cluster is composed of more than a single frame, you must spread your quorum nodes across several frames. 3. After you build the appropriate topology, write the Interface node and Storage node numbers in the table.
15	Quorum Interface nodes	int001st001	

In Table 1-3, you provide CLI credentials. Your SONAS administrator uses these credentials to connect to the CLI or GUI to manage your entire SONAS Storage Solution. SONAS also supports role-based access control (RBAC) and you can create multiple administrator accounts with different responsibilities. For more information, see the SONAS IBM Knowledge Center, found at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/landing/sonas_151_kc_welcome.html

Table 1-3 CLI credentials

Question #	Field	Value (default)	Notes
16	CLI User ID	admin	Your SONAS administrator uses this ID for a GUI or CLI connection, for example, admin.
17	CLI Password	admin	The password corresponding to the User ID, for example, admin.

In Table 1-4, enter the locations of the SONAS nodes in your data center. The Rack number is the number of the rack that contains this node, and the position indicates position (U) where this node is installed in the rack. The Node Serial Number is the serial number of the node. The InfiniBand Port Number is the InfiniBand Switch port number where the node is connected. You do not have this information until the equipment is shipped. Capture this information *after* the equipment arrives and is installed. This information is for your future reference.

Node numbering begins with the lowest number rack and the lowest position node for its class. For example, for an RXA Feature Code 9006, 9007, or 9008, you number the Interface nodes as int001st001 at rack 1 slot 1, int002st001 at rack 1 position 3, and so on. When you finish counting the nodes in the first rack, proceed the next rack and continue with the node number, such as int009st001 at rack 2, slot location 1. The Storage nodes are numbered strg001st001 in rack 1 slot 17 and strg002st001 in rack 1 slot 19. The next Storage node, strg003st001, for example, is rack 1 slot 13 and strg004st001 is in rack 1 slot 15. The information for 17 is based on a lab environment.

Table 1-4 Node locations (based on cable audit after physical installation phase)

Question #	Node number	Rack number/ position	Node serial number	InfiniBand port number
18	Interface node #1	Rack 1/1	KXY780	IB 1-2 and IB2-2
19	Interface node #2			
	...			
20	Storage node #1			
21	Storage node #2			

Important: Record and verify the cabling between nodes and switches. Part of the SONAS cluster health requires that each node is connected to its respective ports.

1.5 Configuration plan

This section describes the configuration of your solution. This configuration includes how you might want to place your data, how you want your users and applications to access this data, the layout of Interface nodes in relation to your data and your network, Information Lifecycle Management considerations, name space considerations, and remote caching. The next section describes data protection.

1.5.1 Basic and advanced configurations

First, this section reviews the basic configuration components for SONAS. It presents a few tables that contain basic information that is pertinent to a successful SONAS configuration.

Basic configuration

The following list of tables contains questions that must be answered:

- Customer information and remote Call Home configuration
- DNS configuration
- NAT configuration

- Authentication methods
- Samba PDC - NT4 configuration
- NIS configuration
- File access protocol information
- Customer facing network configuration

Advanced configuration

Advanced configuration planning of your SONAS appliance helps you correctly size your SONAS appliances for capacity, performance, and reach. You must plan these aspects:

- Your internal file systems layout, which includes file sets, quotas, and exports (shares)
- Performance
- Information Lifecycle Management (ILM)
- Network integration
- Antivirus implementation
- Remote caching and name space mapping
- DNS and aliases

Next, basic configuration components are described. In Table 1-5, you provide some information about the remote configuration of your SONAS to enable the Call Home feature.

Table 1-5 Customer information and remote Call Home configuration

Question #	Field	Value	Notes
1	Company Name		Your company name or the name of the company who you want IBM support personnel to contact when they are responding to support issues and queries.
2	Address		The address where your SONAS appliance is physically. Include the room and X,Y location within the room. Example: Bldg. 123, Room 456, RXA:9003 xxx-yyy, 789 N Data Center Rd, City, State.
3	Customer Contact Phone Number		If there is a severe issue, this number is the primary contact for IBM service.
4	Off Shift Customer Contact Phone Number		This number is the alternative phone number.
5	IP Address of Proxy Server (for Call Home)		Optional. SONAS can trigger a call home over the IP network. It uses HTTPS protocol. In many data centers, access to the outside world is restricted except through a proxy server. If your environment requires it, you need those details here. If you do not provide them, the Call Home feature does not work.
6	Port of Proxy Server (for Call Home)		Optional. As before.
7	User ID for Proxy Server (for Call Home)		Optional. Provide the user ID of the proxy server if it is needed to access the internet for the Call Home feature.
8	Password for Proxy Server (for Call Home)		Optional. Provide the password of the proxy server if it is needed to access the internet for the Call Home feature.

Next, you provide details of your authentication method, as shown in Table 1-6. You must integrate your SONAS system into your existing authentication environment, which can be Active Directory (AD), Lightweight Directory Access Protocol (LDAP), NT4 PDC, or Network Information Service (NIS).

Table 1-6 Authentication methods (AD and LDAP)

Question #	Field	Value	Notes
1	Authentication Method	[ ] Microsoft Active Directory or [ ] LDAP	Here, you determine whether you require SONAS to use Microsoft AD or LDAP as the external authentication service. Only one authentication method can be used.
2	AD Server IP address		For an Active Directory configuration, you need to provide the IP address of the Active Directory Domain Controller. This address is the controller that is closest to the SONAS appliance. In some cases, clients run batch jobs to create users in AD and immediately map them to SONAS. It might be an issue in large environments with multiple Domain Controllers, as details of the newly created users might not be propagated in time to the Domain Controller that is being used by SONAS. You need to adjust your domain replication settings to work around it, or point SONAS to the same Domain Controller that you use for batch user creation.
3	AD Directory UserID		This user ID and the password are used to authenticate to the Active Directory server so that SONAS can add itself to the domain. You must ensure that this account temporarily has Domain Admin privileges, or that you have previously created the computer account by using the cluster name.
4	AD Password		The password that is associated with the user ID.
5	LDAP IP Address		For an LDAP configuration, you must provide the IP address of the remote LDAP server.
6	LDAP SSL Method	[ ] Off [ ] SSL (Secure Sockets Layer) [ ] TLS (Transport Layer Security)	For an LDAP configuration, you can choose to use an open (unencrypted) or secure (encrypted) communication between your SONAS cluster and the LDAP server. For secure communication, two methods can be used: SSL or TLS. When SSL or TLS is used, a security certificate file must be copied from your LDAP server to the IBM SONAS Management node.
7	LDAP Cluster Name		The Cluster Name that is specified in Table 1-1 (for example, *sonascluster*).
8	LDAP Domain Name		The Domain Name that is specified in Table 1-1 (for example, `mydomain.com`).

Question #	Field	Value	Notes
9	LDAP Suffix		The suffix, rootdn, and rootpw from the /etc/openldap/slapd.conf file on your LDAP server. This information can be found in /etc/openldap/slapd.conf on your LDAP server.
10	LDAP rootdn		
11	LDAP rootpw		
12	LDAP Certificate Path		If you choose the SSL or TSL method, you must provide the path on the IBM SONAS Management node where you intend to copy the Certificate file.

If you want to connect SONAS to a SAMBA or Windows NT 4 Domain, then you must prepare the information in Table 1-7.

Table 1-7 Authentication methods (Samba PDC - NT4)

Question #	Field	Value	Notes
1	NT4 Server IP Address		The IP address of the Samba PDC - NT4 Domain Controller on the customer's network.
2	NT4 Admin User		The Admin user ID for the customer's Samba PDC - NT4 Domain.
3	NT4 Admin Password		The password for the Admin user ID on the customer's Samba PDC - NT4 Domain.
4	NT4 Domain Name		The Domain Name for the customer's Samba PDC - NT4 Domain.
5	NT4 NetBIOS Name		The NetBIOS Name for the customer's Samba PDC - NT4 Domain.

Network Information Service (NIS) is also commonly called Yellow Pages. It is typically used by customers with UNIX servers by using the Network File System (NFS) protocol to access network-attached storage (NAS). In some environments, it is also used to manage user IDs when UNIX servers are used with Microsoft Active Directory as an authentication method.

Table 1-8 Authentication methods (NIS)

Question #	Field	Value	Notes
1	NIS mode	[ ] Basic - NIS is used (to provide NFS NetGroup support) in an environment without Active Directory(AD), LDAP, or Samba Primary Domain Controller (PDC). [ ] Extended - NIS is used (to provide NFS NetGroup support and to map UNIX IDs to Windows IDs) for an environment where Active Directory (AD) or Samba Primary Domain Controller (PDC) is used for authentication.	NIS is typically used for one of the following purposes: ► NIS can be used to provide NFS Netgroup support in an environment *without* AD, LDAP, or PDC. ► NIS can be used to provide NFS Netgroup support in an environment *with* AD or PDC. ► NIS can be used to provide NFS NetGroup support and map UNIX user IDs (which are numeric) to Windows user IDs (which are text strings), allowing UNIX servers to access network-attached storage devices that use Microsoft Active Directory or PDC to authenticate users.

Question #	Field	Value	Notes

77	Domain Map		If the NIS mode is Basic, leave this field blank.
			If the NIS mode is Extended, this field is optional. This field can be used to specify the mapping between AD domains and different NIS domains.
			When you specify a domain map, use a colon between the AD domain and the NIS domains. For example:
			`ad_domain:nis_domain1`
			If more than one NIS domain is specified, use a comma-separated list. For example:
			`ad_domain:nis_domain1,nis_domain2`
			To specify more than one AD domain, use a semicolon. For example:
			`ad_domain1:nis_domain1,nis_domain2;`
			`ad_domain2:nis_domain3,nis_domain4`

78	Server Map		This field must be used to specify the mapping between NIS servers and NIS domains.
			When you specify a server map, use a colon between the NIS server and the NIS domains. For example:
			`nis_server:nis_domain1`
			If more than one NIS domain is specified, use a comma-separated list. For example:
			`nis_server:nis_domain1,nis_domain2`
			To specify more than one NIS server, use a semicolon. For example:
			`nis_server1:nis_domain1,nis_domain2;`
			`nis_server2:nis_domain3,nis_domain4`

Question #	Field	Value	Notes

77	Domain Map		If the NIS mode is Basic, leave this field blank.
			If the NIS mode is Extended, this field is optional. This field can be used to specify the mapping between AD domains and different NIS domains.
			When you specify a domain map, use a colon between the AD domain and the NIS domains. For example:
			`ad_domain:nis_domain1`
			If more than one NIS domain is specified, use a comma-separated list. For example:
			`ad_domain:nis_domain1,nis_domain2`
			To specify more than one AD domain, use a semicolon. For example:
			`ad_domain1:nis_domain1,nis_domain2;`
			`ad_domain2:nis_domain3,nis_domain4`

78	Server Map		This field must be used to specify the mapping between NIS servers and NIS domains.
			When you specify a server map, use a colon between the NIS server and the NIS domains. For example:
			`nis_server:nis_domain1`
			If more than one NIS domain is specified, use a comma-separated list. For example:
			`nis_server:nis_domain1,nis_domain2`
			To specify more than one NIS server, use a semicolon. For example:
			`nis_server1:nis_domain1,nis_domain2;`
			`nis_server2:nis_domain3,nis_domain4`

Question #	Field	Value	Notes

79	User Map		If the NIS Mode is Basic, leave this field blank. This optional field can be used to specify the handling for a user who is not known to the NIS server. Only one rule can be specified for each AD or PDC domain.
			The handling is specified by using one of the following keywords:
			► DENY_ACCESS denies any user from the specified domain access if they do not have a mapping entry in the NIS. For example: `ad_domain1:DENY_ACCESS`
			► AUTO specifies that a new ID for the user is generated from the specific domain, which does not have an entry in the NIS. This ID is generated from a pre-specified ID range and is auto-incremented. The administrator must ensure that existing NIS IDs do not fall into this provided ID range. This mapping is kept in SONAS and NIS is not aware of this ID mapping. The ID range can be specified by using the ID Map User Range and ID Map Group Range options. For example: `ad_domain1:AUTO`
			► DEFAULT specifies that any user from the specified domain who does not have a mapping entry in the NIS server is mapped to a specified user (typically a guest user). For example: `ad_domain1:DEFAULT:ad_domain\guest`
			To specify rules for multiple AD or PDC domains, separate the rules with a semicolon. For example:
			`ad_domain1:DENY_ACCESS;`
			`ad_domain2:AUTO;`
			`ad_domain3:DEFAULT:ad_domain3\guest`

80	NIS Domain		This field must be used to specify the NIS Domain to be stored in the registry.
81	Use ID Map	[ ] Use ID Map - NIS is used to map UNIX IDs to Windows IDs for an environment where Active Directory (AD) or Samba Primary Domain Controller (PDC) is used for Authentication.	If the NIS Mode is Basic, leave this field blank. If you checked NIS - NFS NetGroup support *without User ID Mapping* in the Options field of Table 1-8 on page 27, leave this field blank. If you checked NIS - NFS NetGroup support *with User ID Mapping* in the Options field of Table 1-8 on page 27, then the Use ID Map field must be checked.
82	ID Map User Range		If the Use ID Map field is blank, leave this field blank. If the Use ID Map field is checked *and* at least one User Map rule is AUTO, then you must specify a User Range, a Group Range, or both. For example: `10000-20000` **Note:** The User Range values must be a minimum of 1024.

Question #	Field	Value	Notes
83	ID Map Group Range		If the Use ID Map field is blank, leave this field blank. If the Use ID Map field is checked *and* at least one User Map rule is AUTO, you must specify a User Range, a Group Range, or both. For example: `30000-40000`. **Note:** The Group Range values must be a minimum of 1024.

After your SONAS appliance is integrated into your existing environment, and the authentication method is configured, you can create exports (shares) to grant access to SONAS users. In Table 1-9, you can specify which file access protocols you want to enable in SONAS.

Table 1-9 File access protocol information

Question #	Field	Value	Notes
84	Protocols	[ ] CIFS (Common Internet File System) [ ] HTTP (Hypertext Transfer Protocol) [ ] FTP (File Transfer Protocol) [ ] SCP (Secure Copy Protocol) [ ] NFS (Network File System)	A list of file access protocols that the users can use to access the system. Check one or more options.
85	Owner		The owner of a folder or file set in SONAS that you designate for sharing. It can be a user name, or a combination of Domain\username. For example: `admin1` or `Domain1\admin1`

86	CIFS Options		If the Protocols row does not have CIFS checked, leave this field blank. If CIFS is checked, you can specify CIFS options. The options are a comma-separated key-value pair list. Here are the valid CIFS options:
			`browseable=yes`
			`comment="Place comment here”`
			For example:
			`-cifs browseable=yes,comment= "IBM SONAS"`
			In some environments, you might need your CIFS environment to mimic certain behavior of Microsoft file Servers, such as permissions inheritance. You can specify the CIFS options here to enable that feature.

Question #	Field	Value	Notes
87	NFS options	IP Address: Subnet Mask: CIDR Equivalent of the Subnet Mask: Access Options: [ ] ro or [ ] rw [ ] root_squash or [ ] no_root_squash [ ] async or [ ] sync.	If the Protocols row does not have NFS checked, leave this field blank. This row is useful as a template to be repeated for NFS exports you plan to use on SONAS. NFS options include a list of clients that are allowed to access the NFS export, and the type of access to be granted to each client. For more information about specifying a list of clients that are allowed to access the NFS shared drive, see note 1 in the following Note box. For more information about Access Options, see note 2 in the following Note box. Example: `-nfs "10.0.0.0/16 (rw,no_root_squash,async)"` For more information about using NIS for NetGroup support, see Note 3 in the following Note box.

Notes for Table 1-9 on page 31:

1. Clients are specified by a combination of IP address and subnet mask (in CIDR /XX format). For example, if you want to access the NFS shared drive from any client with an IP address of 10.0.*.* (10.0.0.0 - 10.0.255.255), then specify an IP address of 10.0.0.0 and a subnet mask of 255.255.0.0. Then, look up the CIDR equivalent, which in this case is 16. Specify the IP address, subnet mask, and the CIDR equivalent of the subnet mask.
2. The Access Options that you specify are granted to all clients that are specified by using the IP Address/Subnet Mask. Valid options are as follows:
   - **`ro`**: This flag causes the NFS mount to be read-only. This flag is enabled by default.
   - **`rw`**: This flag causes the NFS mount to be read/write.
   - **`root_squash`**: This flag denies superusers on the clients any special access rights. It is used to provide improved security against an unauthorized user with root access. This flag is enabled by default.
   - **`no_root_squash`**: This flag allows superusers on the clients to have superuser access to the exported directories.
   - **`async`**: This flag causes the system to provide completion status to a client as soon as all data to be written is in system memory, but before it is written to disk. This enhances performance, but can result in undetectable data loss.
   - **`data loss`**: This flag indicates whether the system fails before all data is written to disk. The flag is enabled in the default settings.
   - **`sync`**: This flag causes the system not to provide completion status to a client until all data is written to disk. It can reduce performance, but improves protection against undetected data loss.
3. If you are using NIS for NetGroup support, the NetGroups are defined in the `/etc/netgroup` file on the NIS server. They define network-wide groups that are used for permission checking when processing requests for remote mounts and remote logins. NIS NetGroups are supported for NFS clients. When you create an NFS export that is intended for use by NIS NetGroups, the NFS options must be in the format `"@<netgroup_name>(rw,root_squash)"`.

In Table 1-10 and Table 1-11, you provide information about your existing DNS and NAT configuration.

Table 1-10 DNS configuration

Question #	Field	Value	Notes
27	IP Address of Domain Name Services (DNS) Server(s)		Here, you need to provide the IP address of one or more Domain Name Services (DNS) Servers that you are using inside your network. **Note:** Your DNS server must refer to your cluster and eventually contain all the IP addresses for allocation to the Interface nodes that serve your customer network. You must configure DNS for round-robin as well.
28	Domain		The domain name of your cluster (such as `mycompany.com`). **Note:** This field is not required and can be left blank. If it is left blank, then no domain name is set for the cluster.

29	Search String(s)		A list of one or more domain names to be used when you are trying to resolve a short name. For example:
			`mycompany.com`
			`storage.mycompany.com`
			`servers.mycompany.com`)
			Note: This field is not required and can be left blank. If it is left blank, then no search string is set for the cluster.

In Table 1-11, you provide the NAT configuration information.

Table 1-11 NAT configuration

Question #	Field	Value	Notes
30	IP Address		This IP address is not a Data Path connection (it is not used to write files to or read files from the Interface nodes). It is used to provide a path from each Management node and Interface node to the customer network for Authentication/Authorization purposes. So, even if a node has its Data Path ports disabled (for example, an Interface node with a hardware problem can have its Data Path ports disabled and under control of the software), the node can still access external servers (such as Active Directory or LDAP servers) for authentication and authorization.
31	Subnet Mask		The Subnet Mask that is associated with the IP address.
32	CIDR Equivalent of the Subnet Mask		The CIDR (/XX) equivalent of the Subnet Mask that is specified.
33	Gateway		The Default Gateway that is associated with the IP address.

In Table 1-12, you provide details of the Interface subnet and network information.

Table 1-12 Customer facing network

Question #	Field	Value	Notes
49	Subnet		Basically, the public or user network. This network is used for communication between SONAS Interface nodes and your application servers and users. For example, if you have three Interface nodes on a single network, with IP addresses 9.11.136.101 - 9.11.136.103, then your subnet is 9.11.136.0, and your subnet mask is 255.255.255.0 (/24 in CIDR format). Every Interface node potentially has up to four 10 GbE ports and 6 1 GbE ports. You can create a number of interfaces that are called ethX0, ethX1, and so on. These interfaces can consist of one or more ports of the same type. Additionally, if your switch supports it, you can configure an interface to be an aggregated bond that uses LACP or Etherchannel protocols. It allows you to set up different networks to serve different purposes. For example, you might want to use two 10 GbE ports to present NAS to users, and use two 1 GbE ports for data replication or backups.
50	Subnet Mask		Subnet Mask that is associated with the Subnet listed.
51	CIDR equivalent of the Subnet Mask		Subnet Mask that is listed, and converted to CIDR format.
52	VLAN ID		Optional. A list of one or more Virtual LAN Identifiers. It is useful only if you have turned on VLAN Tagging on your Ethernet (or aggregated) ports. VLAN ID must be 2 - 4095. If you do not use VLANs, then leave this field blank. Do not use VLAN 1 when VLAN Tagging is turned on because it is often used by other vendors for VLAN management-related network I/O.
53	Group Name		The name that is assigned to a network group. It allows you to reference a set of Interface nodes by using a meaningful name instead of a list of IP addresses or host names. If you do not use network groups, leave this field blank.
54	Interface node number and host name	IP address Subnet/Subnetmask Gateway	(Repeat it for each Interface node.)

1.5.2 SONAS internal file system

The SONAS internal file system is built on the IBM General Parallel File System (GPFS). GPFS is a cluster file system. This section describes the architecture and the important components that must be considered before you build a file system on SONAS.

GPFS architecture

GPFS is a cluster file system, which means that it provides concurrent access to a single file system or set of file systems from multiple nodes in a defined GPFS cluster. This configuration enables high-performance access to this common set of data to support a scale-out solution or provide a high availability platform. SONAS is built on GPFS, which provides the underlying file systems. GPFS also provides concurrent access across all nodes in the SONAS cluster.

GPFS provides a global name space, shared file system access among GPFS clusters, simultaneous file access from multiple nodes, high recoverability and data availability through replication, the ability to make changes when a file system is mounted, and simplified administration even in large environments. SONAS GPFS file systems support up to 4 billion files per file system with a maximum of over several hundred million files in async replication. When you replicate hundreds of millions of files, it can take a long time to scan and process them in any environment. Consider this time before you assume a small RPO. For more information about the maximum number of file systems in a file system or independent file set, go to the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/adm_gpfs_limitations.html

Within each file system, files are written to disk as in other UNIX file systems, by using inodes, indirect blocks, and data blocks. Inodes and indirect blocks are considered metadata, as distinguished from data (file content).

File system overview

A file system consists of a set of disks that are used to store file metadata and data, and structures that are used by GPFS, including quota files and GPFS recovery logs. GPFS file systems have two basic file system data structures or building blocks: Data and metadata. See Figure 1-12.

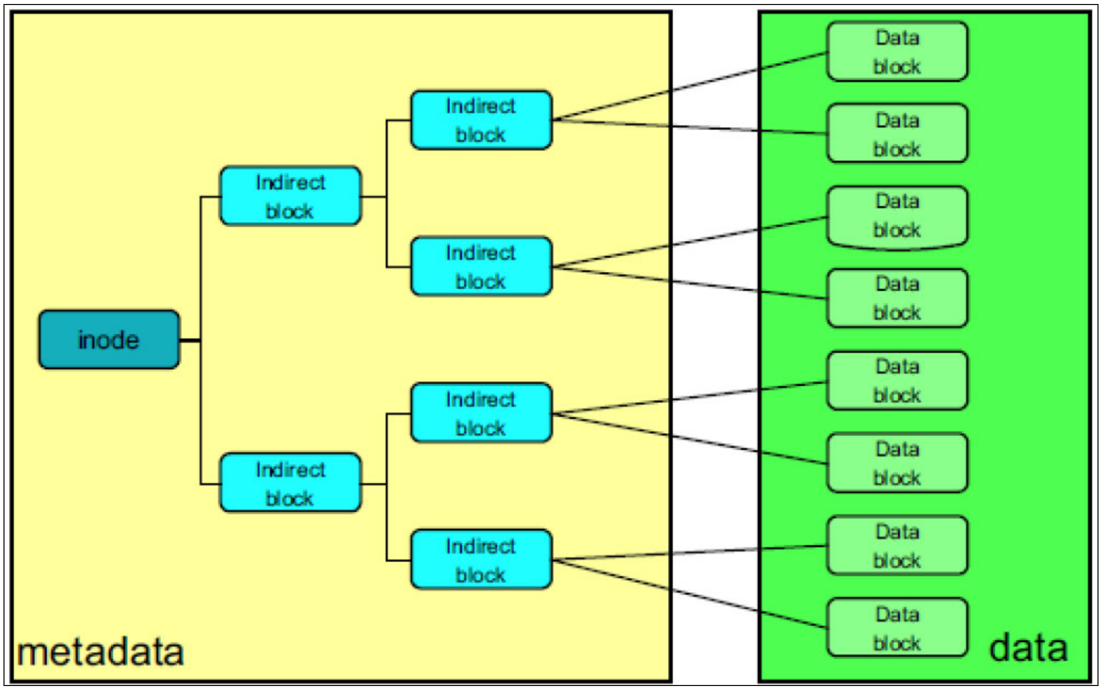


Figure 1-12 GPFS file system building blocks

Metadata

Metadata is a collection of data structures that contain information about file data. Inodes (Index-node), indirect blocks, and directories are all considered metadata. Metadata contains all the information that pertains to any file or directory, such as file name, file size, owner, access rights, creation, modified, and access times, and other extended attributes. Understand that every file system object, whether it is a regular file, device file, links, or directory, takes an inode. Most of the metadata in a GPFS file system is primarily in a special file that contains all of the file inodes, and user directory information. There are some other metadata files, but they are small and have little effect on the space that is required to store metadata. Metadata is always stored on disks in the system pool.

Inodes

Every file system file or object requires an inode. Each inode is 512 bytes and contains all the file information, including pointers to the data blocks, whether directly or indirectly (indirect blocks, which in turn point to data blocks), where the file data is stored. GPFS sets the maximum number of allowed inodes, and hence files, at file system creation time. It is based on a complex calculation, but for file systems larger than 8 TB, it defaults to 100 million files or inodes. Therefore, the maximum allowable files on an 8 TB or larger file system is 100 million files. However, this limit can be increased when the file system is created.

For file systems that are doing parallel file creation, if the total number of free inodes is 5% or less of the total number of inodes, there is the potential for slowdown in file system access. Take this possibility into consideration when you are changing your file system.

Directory blocks

Directories are files, and hence inodes that point to directory blocks. These directory blocks are part of the metadata and contain the information about how the file name space is organized. The name space is a tree: Each directory is a file whose contents map file names with regular inodes. The cost of directories is a function of usage and file naming structures. A directory costs at least the minimum file size (subblock size) for each directory and more if the number of entries is large. For a 256 KB block size file system, the minimum directory is 8 KB. The number of directory entries per directory block varies with customer usage and is largely based on the file name lengths. In a directory block, directory entries are allocated in chunks of 16 bytes and each entry has 12 bytes of overhead. So, names up to 4 bytes take one chunk (4 + 12 = 16) names 5 - 20 bytes take two chunks (20 + 12 = 32), and so on. If a directory block is filled, and each directory entry takes 32 bytes (assuming file names are 20 bytes or less long), an 8 K directory block can hold a little less than 240 entries before another 8 K is allocated to the directory block.

Metadata is relatively small compared to the data space. Most GPFS file systems mix data and metadata across all available disks and NSDs in the file system. Therefore, how much metadata space you need is not important because the data and metadata share all available disks. However, if you are using multiple storage pools or are planning to use high-performance storage, such as solid-state disks (SSDs) for metadata, you might need to better predict the amount of metadata space that is required. However, trying to predict the metadata size can be a complex task, especially if you do not know what kind of data or how many files and directories are being created.

Metadata sizing

The following formulas can be used to help determine the space to set aside for metadata. If metadata is being replicated (failover groups), then double all numbers. When doubled, half are in failure group 1 and the other half are in failure group 2.

- Expected Regular files and inodes:

  max-Files * 512 bytes = Inode Space

- Expected Directories blocks:

  max-Directories * SubBlockSize = Directory Block Space

  The preceding example assumes that all directory entries fit into a single directory subblock chunk. If the file names are longer than 20 characters, or a directory takes more than a single chunk, this space must be accounted for.

- If you use snapshots, use the following formula:

  inodeFileSize * #snapshots * PercentageChangedInodesBetweenSnapshots

- Each file with Extended Attributes (EAs) uses a 16 KiB block for the inode. For example, Tivoli Storage Manager hierarchical storage migration (HSM) uses EAs to store its data about migrated inodes.

  16 KiB * InodesWithEAs

  The challenge of computing possible metadata usage is to determine the directory occupancy for your system, snapshot usage, and inodes with Extended Attributes. However, a good number to use for metadata sizing is 5% of file system total size. Therefore, if you decide to create a 50 TB file system and put metadata on SSDs or fast SAS drives, set aside 2.5 TB (50 TB * 0.05) for metadata. If you are replicating metadata, double this number to 5 TB. It is always better to overprovision for metadata because it can severely affect performance.

Data

The data blocks are the blocks of file data. The data blocks are pointed to either by the inode directly or indirectly through a pointer to a block of direct pointers. The data block size is a function of the file system block size. When file systems are created, you can specify the block size of the file system. With GPFS, the block size is divided into subblocks, which are the smallest unit of allocation. This subblock size is 1/32 the size of the defined file system block size. Therefore, if a file system is defined with a block size of 256 K, the smallest subblock is 8 K. If the file system was defined with a block size of 1 M (1048756 bytes), the subblock size is 32 K.

Block size

One of the most important aspects of building a high-performing file system is the number of underlying LUNs (also known as NSDs) and the file system block size. SONAS supports block sizes of 256 K (default), 1 M and 4 M. These sizes are defined when the file system is created and cannot be changed without re-creating the file system. When you re-create a file system, you must first back up the data and then restore it after you re-create the file system. Therefore, carefully plan for what type of data will be written and read. Also, plan for the I/O patterns and advanced functions because advance functions can have a huge effect on metadata.

The block size determines the minimum disk space allocation unit because GPFS divides each block into 32 subblocks. Files smaller than one block in size are stored in fragments, which are made up of one or more subblocks. Large files are stored in a number of full blocks plus zero or more subblocks to hold the data at the end of the file. The block size is the largest contiguous amount of disk space that is allocated to a file and therefore the largest amount of data that can be accessed in a single I/O operation. The subblock is the smallest unit of disk space that can be allocated. For a block size of 256 KB, GPFS reads as much as 256 KB of data in a single I/O operation and small files can occupy as little as 8 KB of disk space. With a block size of 1 M, small files occupy as little as 32 KB of disk space (not counting the inode), but GPFS is unable to read more than 1 M in a single I/O operation.

The block size also determines the maximum size of a read or write request that the file system sends to the underlying disk driver. From a performance perspective, set the block size to match the application buffer size, the RAID stripe size, or a multiple of the RAID stripe size. If the block size does not match the RAID stripe size, performance can be severely degraded, especially for write operations. Therefore, a block of 256 K, 1 M, or 4 M determines how to configure the underlying RAID set on the Storwize V7000 and DCS3700 systems and the number of disks. The XIV storage is fixed at 1 M and cannot be changed.

In file systems with a mix of file sizes, a small block size has a large effect on performance when you are accessing large files. In this kind of system, use a block size of 256 KB (8 KB subblock). Even if only 1% of the files are large, the amount of space that is taken by the large files usually dominates the amount of space that is used on disk, and the waste in the subblock that is used for small files is insignificant. Larger block sizes of up to 1 MB are often a good choice when the dominant workload for this file system is large files that are accessed sequentially.

The effect of block size on file system performance largely depends on the application I/O pattern. A larger block size is often beneficial for large sequential read and write workloads. A smaller block size is likely to offer better performance for small file, small random read and write, and metadata-intensive workloads. The efficiency of many algorithms that rely on caching file data in a page pool depends more on the number of blocks that are cached instead of the absolute amount of data. For a page pool, a larger file system block size means that fewer blocks are cached. Therefore, when you create file systems with a block size that is larger than the default size of 256 KB, it is best that you increase the page pool size in proportion to the block size. Data is cached in Interface nodes memory, so it is important to correctly plan RAM memory size for Interface nodes.

File system LUNs (NSDs)

The number and underlying RAID configuration of the NSDs can have a great effect on the performance of the file system. Configure all disks to be the same (size and RAID configuration) in each storage pool. If multiple tiers (storage pools) are used, put the fastest disks in the system pool because it is where the metadata is stored.

The number of disks that are allocated to a file system should be an even number that is divisible by 4, such as 4, 8, or 12. This guideline is because of the number of I/O paths between the Storage node pair or pod and the storage platform. This guideline helps ensure that disks (and LUNs) are spread evenly across Storage nodes, Storage node HBA ports, and storage platform controllers. In most cases, it is better to have a least eight LUNs for each file system to maximize the I/O spread across all components, and increase the effective queue depth (I/O spread across more LUNs and more queues). Eight also divides well into all of the subblock sizes.

1.5.3 SONAS appliance (internal DDN storage)

A storage controller in SONAS is configured to support 32 KB chunk sizes. This value is preconfigured and cannot be changed. It means, for example, a default SONAS block size (256 KB) is divided by eight disks in each RAID array and written to eight different drives in 32 KB chunks each. Each RAID array in SONAS consists of 10 disks; eight are available for data because RAID 6 consists of 8+P+Q spare drives and RAID 5 consists of 8+P+spare drives.

In Figure 1-13 on page 41, you can see the basic components that make up a file system in SONAS. This information serves two primary purposes. First, you can see how a RAID 6 array from disks in a drawer in groups of 10 (eight Data + P + Q) is automatically set up. Only one LUN is created on each logical array. Because each enclosure has 60 disks, there are six LUNs per enclosure. Each Storage pod, therefore, can have up to 24 LUNs. These LUNs are presented to the Storage nodes in the pod and they in turn present these LUNs as NSDs to the Interface nodes. The Storage nodes can equally see each LUN. Automatic rules are used to predetermine which Storage node becomes the preferred node to present the NSD. The other node presents the alternative NSDs and is the standby for the first node’s NSDs.

Note: The examples in this section are from a SONAS Appliance (internal DDN disk storage subsystem). The NSDs in Gateways, which are supplied by the underlaying disk subsystem, might have different sizes and configurations. Carefully consider the underlaying LUN configuration for each Gateway disk subsystem configuration. The examples here can be applied to all configurations.

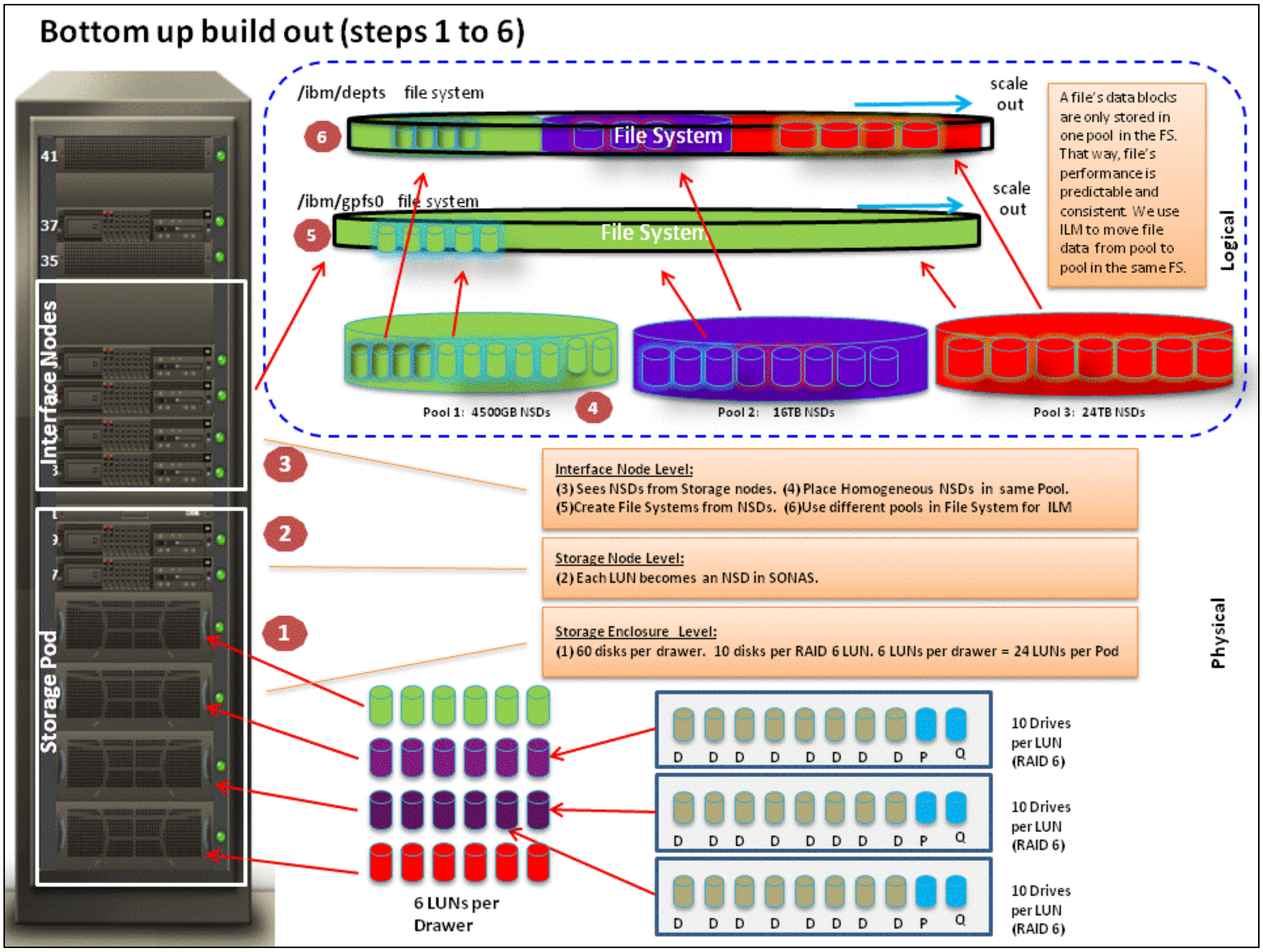
If you have storage enclosures with different types of disk (for example, 60 x 600 GB disks, 60 x 2 TB disks, and 120 x 3 TB disks), place the NSDs of the same type in different storage pools. In the example, you can have three storage pools called system (6 x 4.8 TB NSDs), silver (6x 16 TB NSDs), and bronze (12x 24 TB NSDs).

A file system is created at the Interface node level and can consist of any combination of NSDs from the different pools. It might appear that the file system performance is “lumpy”, as file systems are often assumed to be striped across all NSDs (or LUNs). In GPFS, a file system is striped only across NSDs from the same pool. It means that a file, when written, has all its data written only to the NSDs of one storage pool. Without an inline policy to store files, all new files are, by default, created on the “system” pool.

The block size also defines the largest single I/O that will be made to the underlying disks before moving on to the next disk (or NSD) of the file system. Therefore, if the block size is 256 K, the largest I/O is 256 K, at which time I/O proceeds to the next NSD and the next. Therefore, it is important to align the block size with the underlying disk (or NSD) RAID segment size.

Note: If multiple storage pools are used, place the fastest disks in the “System” pool because it is where all the metadata is stored.

Figure 1-13 shows building out a SONAS file system from the bottom up.



A file's data blocks are only stored in one pool in the FS. That way, file's performance is predictable and consistent. We use ILM to move file data from pool to pool in the same FS.

Interface Node Level:
(3) Sees NSDs from Storage nodes. (4) Place Homogeneous NSDs in same Pool.
(5)Create File Systems from NSDs. (6)Use different pools in File System for ILM

Storage Node Level:
(2) Each LUN becomes an NSD in SONAS.

Storage Enclosure Level:
(1) 60 disks per drawer. 10 disks per RAID 6 LUN. 6 LUNs per drawer = 24 LUNs per Pod

Figure 1-13 Build out a SONAS file system from the bottom up

Figure 1-14 is an example of how SONAS writes data to the underlying disk subsystem. In the example, the parity function is part of the controller functions, not SONAS.

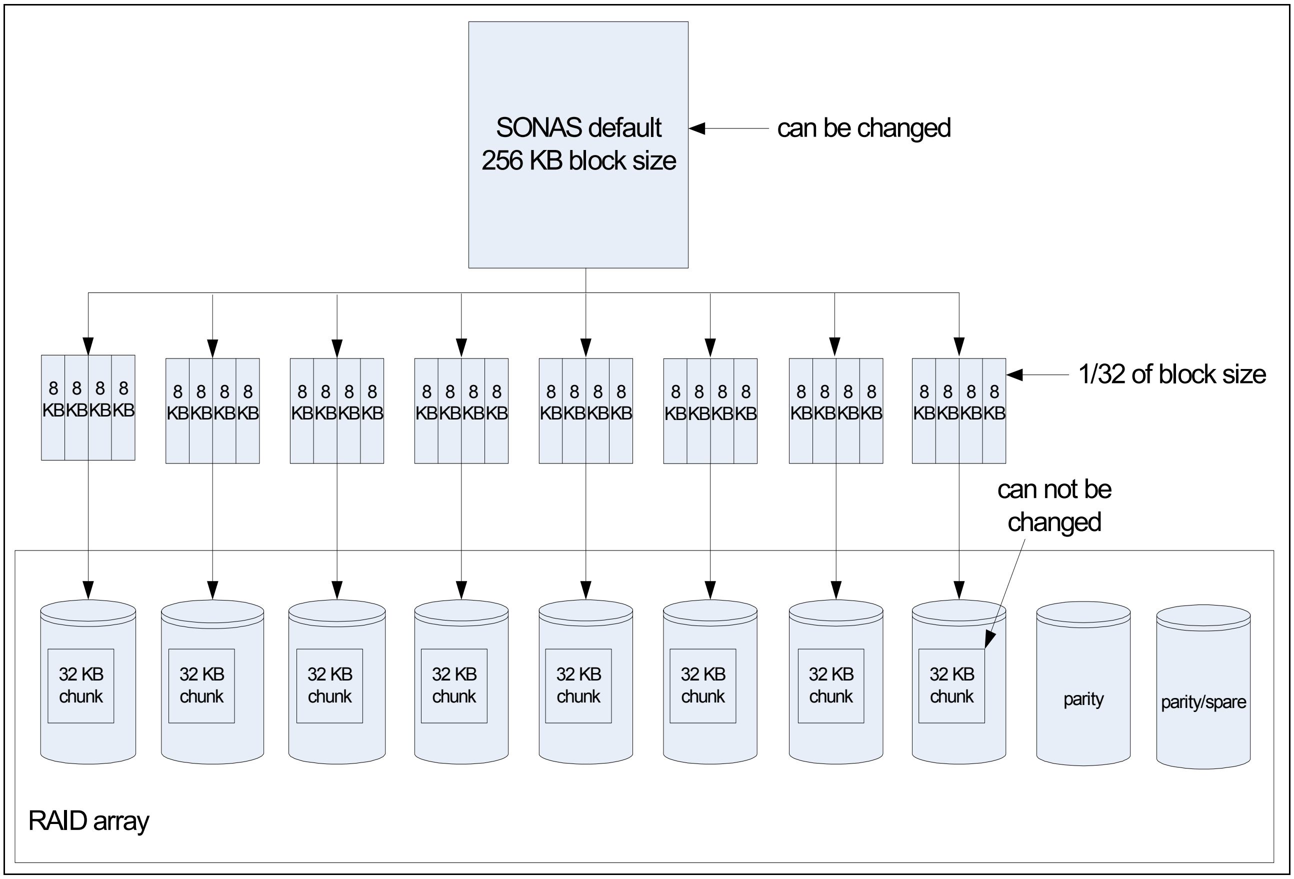


Figure 1-14 How SONAS writes data to disks

Setting up storage pools

A storage pool is a collection of disks with similar properties (such as an array of 900 GB SAS drives) that provide a specific quality of service (QoS) for specific use, such as to store all files for a particular application or a specific business division. Using storage pools, you can create tiers of storage by grouping storage devices based on performance, or reliability characteristics. For example, one pool can be an enterprise class storage system that hosts high-performance SAS disks, and another pool can consist of a set of economical nearline SAS disks. The storage pool is managed as a group, as the storage pool provides a means to partition the management of the file system’s storage.

There are two types of storage pools:

- System storage pool (exists by default)

  A system storage pool contains the system metadata (system and file attributes, directories, indirect blocks, symbolic links, policy file, configuration information, and metadata server state) that is accessible to all metadata servers in the cluster. Metadata cannot be moved out of the system storage pool. The system storage pool is allowed to store user data, and by default, it goes into the system storage pool unless the placement policy is activated.

  The system storage pool cannot be removed unless you are deleting the entire file system. Disks inside a system pool can be deleted if there is at least one disk that is assigned to system pool or enough disks with space to store existing metadata. The system storage pool contains metadata that is scanned for many operations, such as backup, replication, and antivirus work. Use the fastest and the most reliable disks for better performance of the whole SONAS file system and failure protection. There can be only one system storage pool per file system, and this pool is required.

- User-defined storage pool

  Up to seven other user-defined storage pools can be created per file system. The user storage pool does not contain metadata. It stores only data, so disks that are assigned to the additional (non-system) user storage pools can have only the usage type, “data only”.

A maximum of eight storage pools per file system can be created, including the required system storage pool. The storage pool is an attribute of each disk and is specified as a field in each disk descriptor when the file system is created or when disk is added to an existing file system.

SONAS offers internal storage pools and external storage pools. Internal storage pools are managed within SONAS. External storage pools are managed by an external application such as Tivoli Storage Manager. SONAS manages the movement of data to and from external storage pools. SONAS provides integrated automatic tiered storage (Integrated Lifecycle Management (ILM)), and provides an integrated global policy engine to enable centralized management of files and file sets in the one or multiple logical storage pools. This flexible arrangement allows file based movement down to a *per file* basis if needed.

1.5.4 Integrating SONAS into your network

This section describes how to integrate your new SONAS system into your existing network environment. This network integration requires a user authentication method to grant SONAS access, public and private networks, and an IP address load balancing mechanism configuration.

Authentication that uses AD or LDAP

You can use your existing authentication method environment to grant user access to SONAS. SONAS supports the following authentication method configurations:

- Microsoft Active Directory
- Lightweight Directory Access Protocol (LDAP)
- LDAP with MIT Kerberos
- SAMBA primary domain controller (PDC)

However, SONAS does not support running multiple authentication methods in parallel. The rule is only one type of authentication method at any time.

A user who attempts to access SONAS enters a user ID and password. The user ID and password are sent across the customer network to the remote authentication and authorization server, which compares the user ID and password to valid user ID and password combinations in its local database. If they match, the user is considered authenticated. The remote server sends a response back to IBM SONAS, confirming that the user was authenticated and providing authorization information.

Authentication is the process to identify a user, and authorization is the process to grant access to resources to the identified user.

A detailed description is provided in 3.8.3, "Authentication with Active Directory or Lightweight Directory Access Protocol" on page 229 for the AD or LDAP configuration. More briefly, authentication can be done with the **cfgad**, **cfgldap**, and **chkauth** commands.

Microsoft Active Directory

One method for user authentication is to communicate with a remote authentication and authorization server that is running Microsoft Active Directory software. The Active Directory software provides authentication and authorization services.

For the **cfgad** command, you must provide information such as the Active Directory Server IP address and cluster name. Basically, this information was required in Table 1-6 on page 26. Here you need answers to questions #35 - #37.

Run the following **cfgad** command:

```
cfgad -as <ActiveDirectoryServerIP> -c <clustername>.<domainname> -u <username> -p
<password>
```

Where:

- **<ActiveDirectoryServerIP>**: IP Address of the remote Active Directory server, as specified in Table 1-6 on page 26, question #35.
- **<clustername>**: Cluster Name, as specified in Table 1-1 on page 21, question #1.
- **<domainname>**: Domain Name, as specified in Table 1-1 on page 21, question #2.
- **<username>**: Active Directory User ID, as specified in Table 1-6 on page 26, question #36.
- **<password>**: Active Directory Password, as specified in Table 1-6 on page 26, question #37.

Here is an example:

```
cli cfgad -as 9.11.136.116 -c sonascluster.mydomain.com -u aduser -p adpassword
```

To check whether this cluster is now part of the Active Directory domain, run the following **chkauth** command:

```
cli chkauth -c <clustername>.<domainname> -t
```

Where:

- **<clustername>**: Cluster Name, as specified in Table 1-1 on page 21, question #1.
- **<domainname>**: Domain Name, as specified in Table 1-1 on page 21, question #2.

Here is an example:

```
cli chkauth -c sonascluster.mydomain.com -t
```

If the **cfgad** command is successful, in the output from the **chkauth** command, you see "`CHECK SECRETS OF SERVER SUCCEED`" or a similar message.

LDAP

Another method for user authentication is to communicate with a remote authentication and authorization server running Lightweight Directory Access Protocol (LDAP) software. The LDAP software provides authentication and authorization services.

For the **cfgldap** command, you must provide information such as the LDAP Server IP address and the cluster name. Basically, this information was required in Table 1-6 on page 26. Here you need answers to questions #38 - #44.

Run the following **cfgldap** command:

```
cfgldap -c <cluster name> -d <domain name> -lb <suffix> -ldn <rootdn> -lpw
<rootpw> -ls <ldap server> -ssl <ssl method> -v
```

Where:

- **<cluster name>**: Cluster Name, as specified in Table 1-6 on page 26, question #39
- **<domain name>**: Domain Name, as specified in Table 1-6 on page 26, question #40
- **<suffix>**: The suffix, as specified in Table 1-6 on page 26, question #41
- **<rootdn>**: The rootdn, as specified in Table 1-6 on page 26, question #42
- **<rootpw>**: The password for access to the remote LDAP server, as specified in Table 1-6 on page 26, question #43
- **<LDAP Server IP>**: IP address of the remote Active Directory server, as specified in Table 1-6 on page 26, question #38-0.
- **<ssl method>**: the SSL method, as specified in Table 1-6 on page 26, question #38

Here is an example:

```
cli cfgldap -c sonascluster -d mydomain.com -lb "dc=sonasldap,dc=com" -ldn
"cn=Manager,dc=sonasldap,dc=com" -lpw secret -ls 9.10.11.12 -ssl tls -v
```

To check whether this cluster is now part of the Active Directory domain, run **chkauth** .

Planning IP addresses

This section briefly describes the public and private IP addresses to prevent conflicts during SONAS use. For more information about these networks, see 2.5.2, “Understanding the IP addresses for internal networking” on page 70.

In Table 1-1 on page 21, Question #3, you were prompted for an available IP address range.

SONAS is composed of three different networks. One of these networks is the *public network,* which is used for SONAS users or administrators to access Interface nodes or Management nodes. The other two networks are the *private network,* or *management network,* which is used by the Management node to handle the whole cluster, and the *data network,* or *InfiniBand network,* on top of which the SONAS file system is built. These last two networks, private and data, are not used by SONAS users or administrators. But as they coexist on all nodes with the public network, ensure that you do not use the same network segments to avoid some IP conflicts.

There are only three choices for the *private network range.* The default setting for public IP addresses is the range 172.31.*.* However, you might already use this particular range in your existing environment, so the 192.168.*.* range might be more appropriate. Similarly, if you are using both the 172.31.*.* and 192.168.*.* ranges, then the range 10.254.*.* must be used as a private network instead.

To determine which IP address ranges are currently used in your data center location, ask your network administrators.

Data access and IP address balancing

This section describes the information that is required to set up the SONAS IP address balancing.

This IP balancing is handled both by the DNS and the CTDB layers. CTDB layer works, in coordination with the DNS, to provide SONAS users access to data.

Details about your DNS configuration are required. With this information, you can set up the connection between your DNS and your SONAS. For the data access through the client network, SONAS users mount exports by using CIFS, NFS, or FTP protocols.

Because the SONAS storage solution is designed to be a good candidate for cloud storage, accessing SONAS data must be as transparent as possible from a technical point of view. Basically, your SONAS users do not need to know or even understand how to access the data. They simply want to access it.

This process works because of an appropriate DNS configuration and the CTDB layer. First, the DNS is responsible for routing SONAS user requests to Interface nodes in a round-robin manner, which means that two consecutive requests can access data through two distinct Interface nodes.

In the tables in 1.4, “Installation plan” on page 20, the `sonascluster.mydomain.com` DNS host name is used as an example. For consistency considerations, the same name is used in the following schemes, which provide step-by-step descriptions of the DNS and CTDB mechanism in a basic environment. This environment is composed of three interfaces nodes, one DNS server, and two active clients. One client runs a Linux operating system and the other one runs a Windows operating system. Again, for consistency considerations, a Management node and Storage pods are also shown, even if they do not have any effect on the DNS or CTDB mechanism. The last FTP client is also here to remind you of the last protocol in use in SONAS.

The first SONAS user, running the Linux operating system, wants to mount an NFS share on their workstation and run **mount** with the `sonascluster.mydomain.com` DNS host name, as described in the upper left corner in Figure 1-15. This request is caught by the DNS server (step 1), which then looks inside its list of IP addresses and forwards the request to the appropriate Interface node (step 2). It happens in a round-robin way, and it sends an acknowledgment to the Linux SONAS user (step 3). The connection between the first SONAS user and one Interface node is then established, as you can see with the dashed arrow in Figure 1-15.

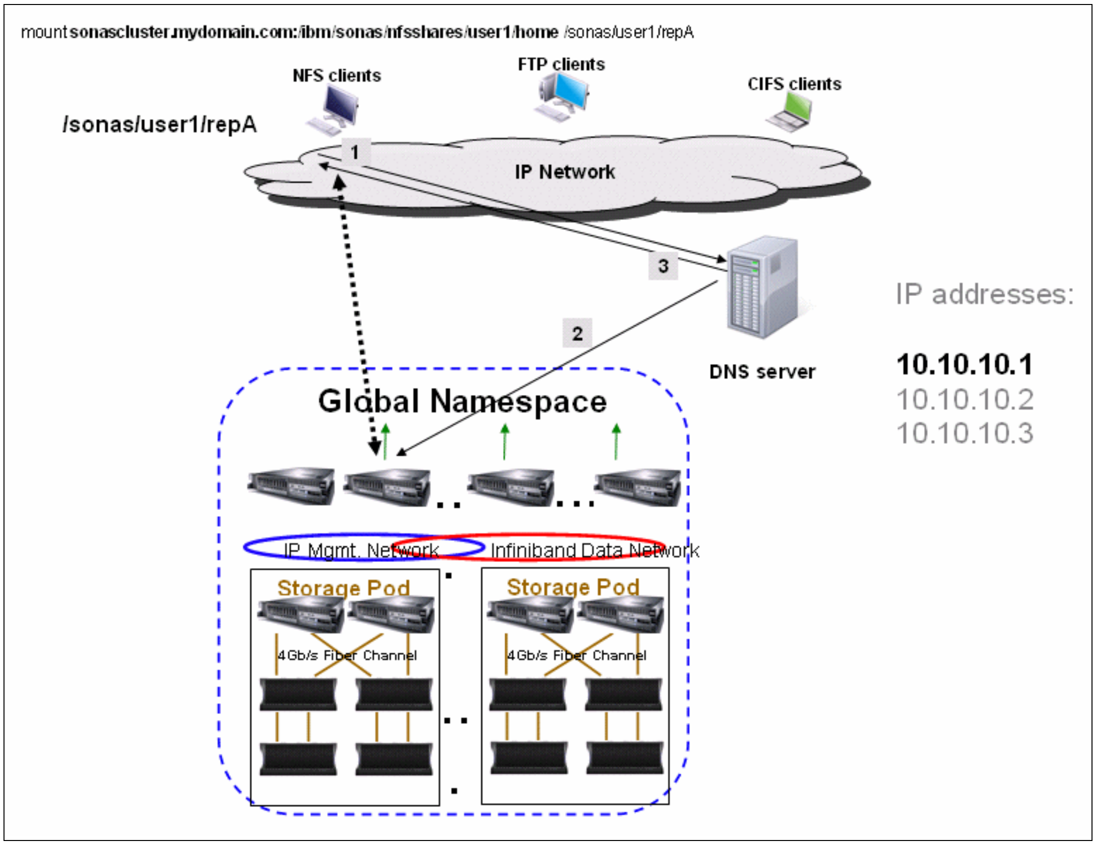


Figure 1-15 SONAS user accessing data with the NFS protocol

Now, assume that there is a second SONAS user who also needs to access data that is hosted on the SONAS storage solution with a CIFS protocol from a Windows notebook. That user runs **`net use`** (or uses the Map Network Drive tool) with the same `sonascluster.mydomain.com` DNS host name, as shown in Figure 1-16.

This second request is caught here again by the DNS server, which, in a round-robin way, assigns the next IP address to this second user. Then, steps 1 - 3 are repeated, as described in Figure 1-16. The final connection between the second SONAS user and the Interface node is then established. See the new dashed arrow on the right.

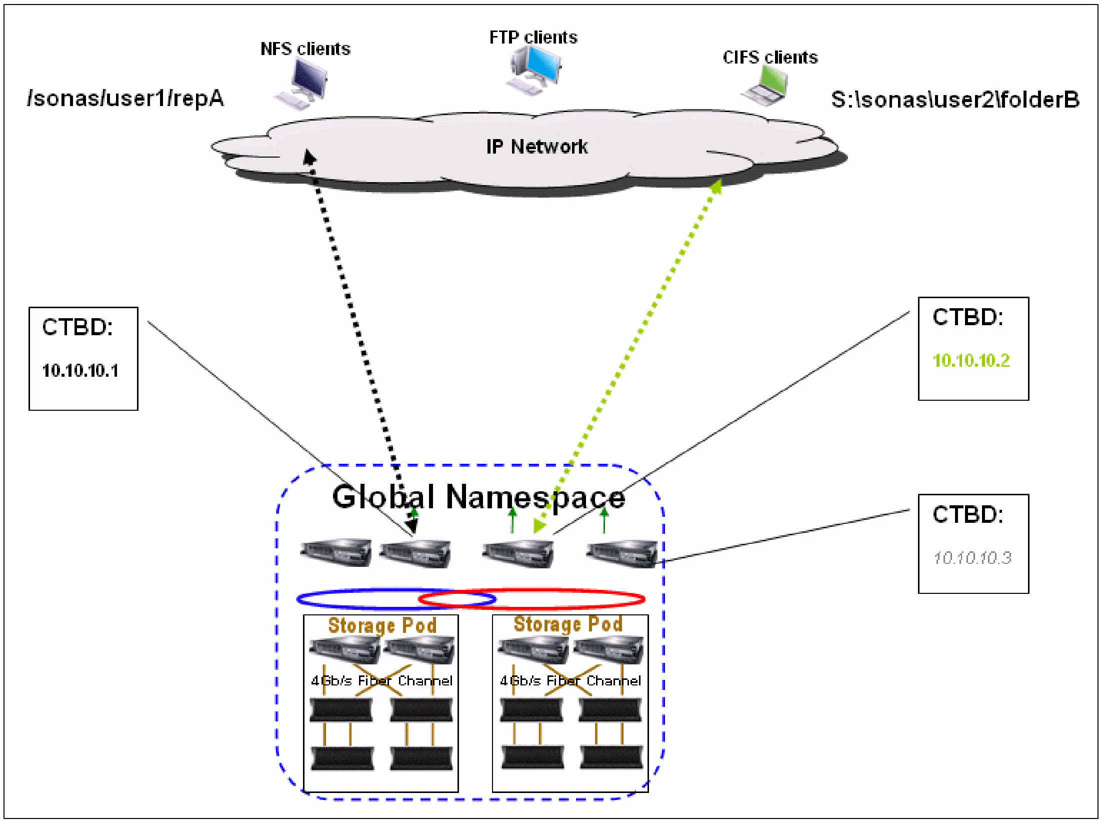


Figure 1-16 SONAS user accessing data with the CIFS protocol

Connections between SONAS users and Interface nodes remain active until shares are unmounted from SONAS users, or in case of Interface node failure.

If there is an Interface node failure, the IP address balancing is handled by the CTDB layer. The CTDB layer works with a table to handle Interface node failure. Briefly, this table is re-created as soon as a new event happens. An event can be an Interface node failure or recovery. Table entries are Interface nodes identifiers and public IP addresses.

In Figure 1-17, the SONAS is configured in such a way that the CTDB has a table with three Interface node identifiers and three public IP addresses for SONAS users.

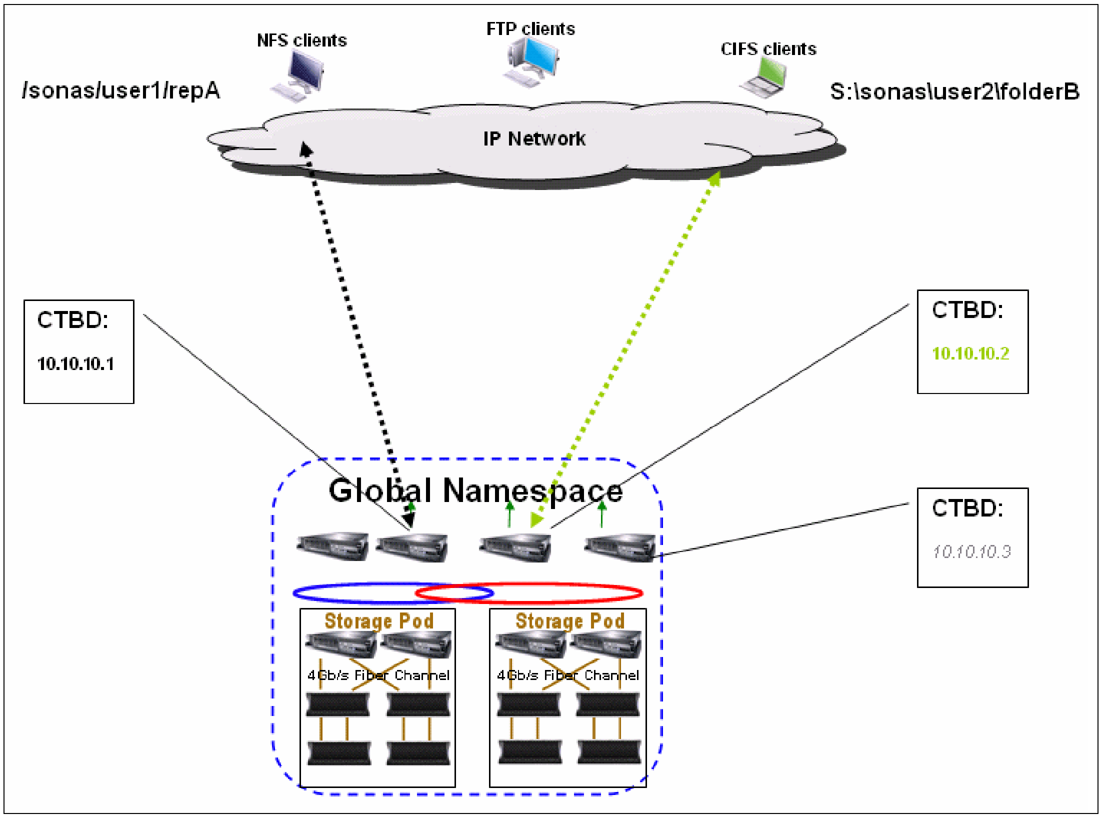


Figure 1-17 CTDB table with three Interface node identifiers and three IP addresses

In this example environment, there are three Interface nodes (#1, #2, and #3) and three IP addresses. The CTDB table is created with these entries:

- #1, #2, and #3
- 10.10.10.1, 10.10.10.2, and 10.10.10.3

From the CTDB point of view:

- #1 is responsible for 10.10.10.1.
- #2 is responsible for 10.10.10.2.
- #3 is responsible for 10.10.10.3.

With your two SONAS users that are connected, as shown in Figure 1-17, only the two first Interface nodes are used. The first Interface node is using the 10.10.10.1 IP address, and the second one is using 10.10.10.2, according to the CTDB table.

If there is failure of the first Interface node, which was in charge of the 10.10.10.1 IP address, this IP address 10.10.10.1 is handled by the last Interface node, as shown in Figure 1-18.

From the CTDB point of view, if there is a failure, you now have these responsibilities:

- #2 is responsible for 10.10.10.2.
- #3 is responsible for 10.10.10.3 and 10.10.10.1.

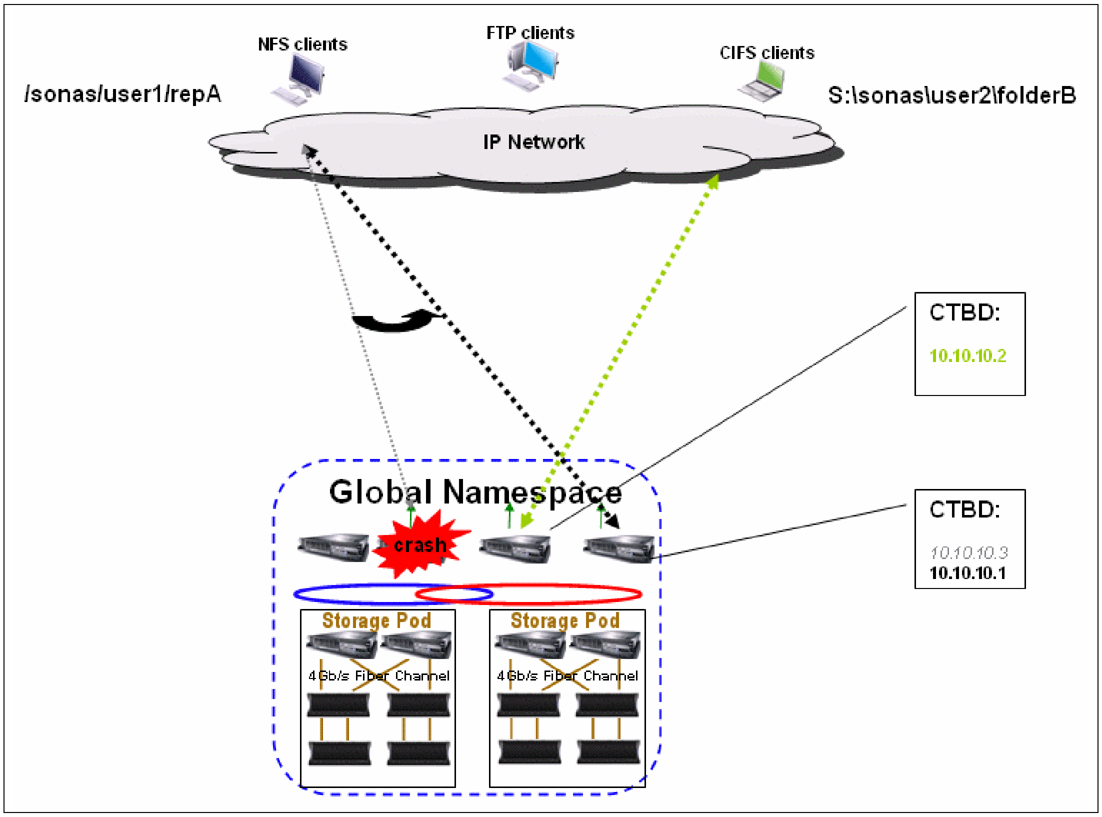


Figure 1-18 CTDB table with Interface node identifiers and IP mappings after a failure

As you can see in Figure 1-18, the first NFS SONAS user now has an active connection to the last Interface node.

It is basically the CTDB that is handling the IP address balancing. Your DNS is handling the round-robin method, and the CTDB is in charge of the IP failover.

However, in the previous example, there is a potential load balancing bottleneck in case of the failure of one Interface node. Indeed, assuming a third user is accessing the SONAS through the FTP protocol, as described in Figure 1-19, the connection is established with the last dashed arrow on the third Interface node. The first NFS user is still connected to the SONAS through the first Interface node, and the second CIFS user is connected to the SONAS through the second Interface node. The last FTP user is accessing the SONAS through the third Interface node (the DNS here again gave the next IP address).

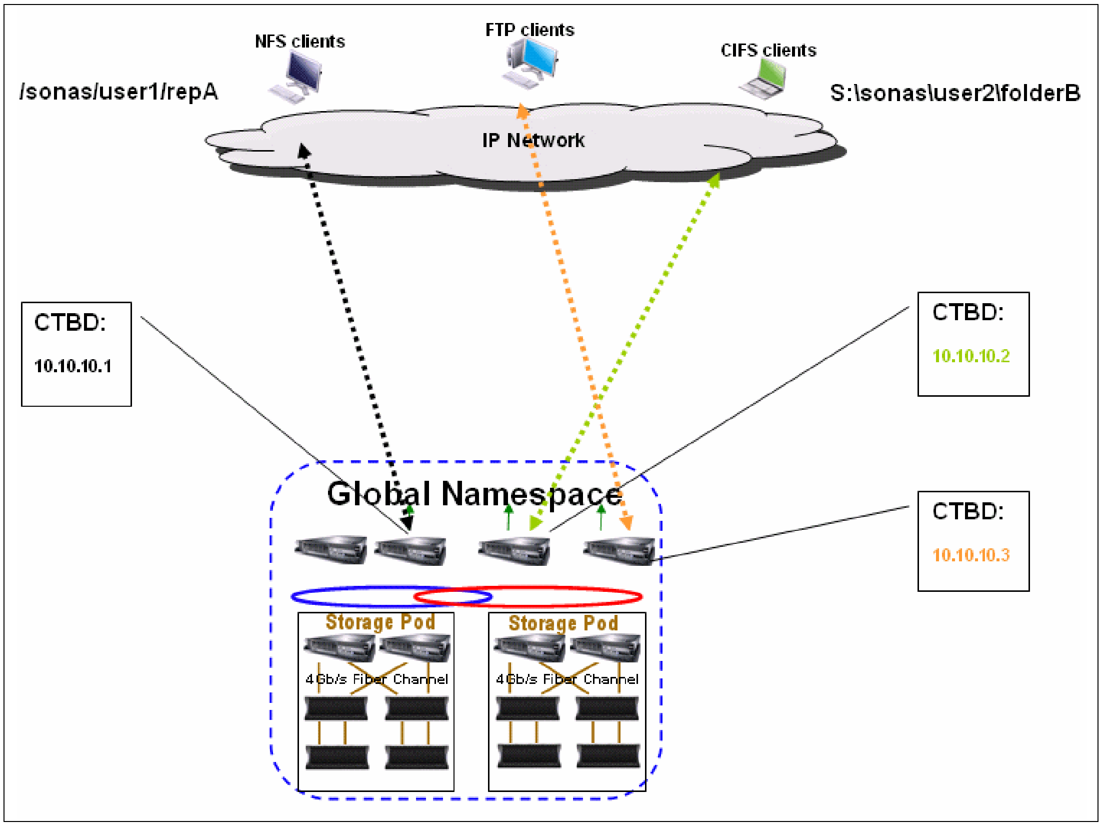


Figure 1-19 CTDB IP address balancing

You might notice that from here that all incoming users are related to Interface nodes #1, #2, or #3 in the same way because of the DNS round-robin configuration. For example, you might have four users who are connected to each Interface node, as described in Figure 1-20.

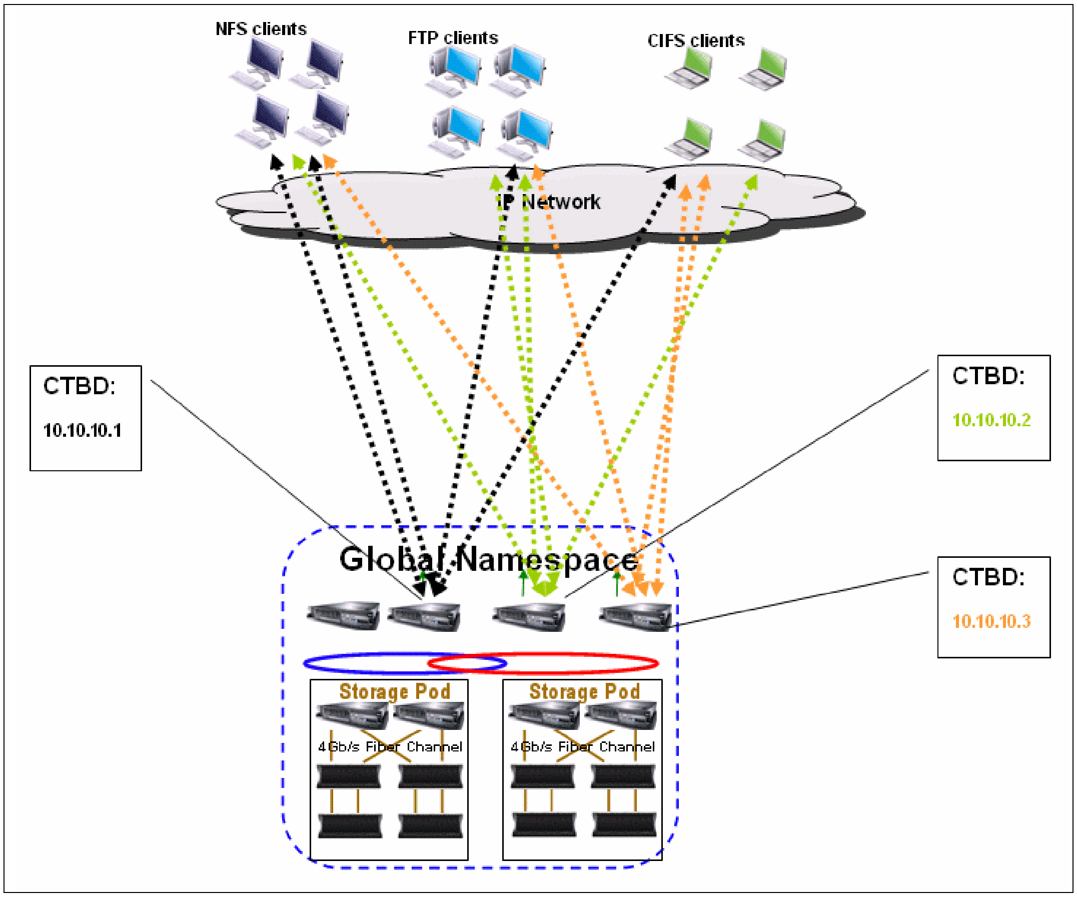


Figure 1-20 Interface node relationships showing CTDB round-robin assignment

The bottleneck that was mentioned earlier in this section appears if one Interface node fails. Indeed, the IP address that is handled by this failing Interface node migrates, as do all users and their workload, to another Interface node according to the CTDB table. You then have one Interface node that is handling a single IP address and four user workloads (second Interface node) and the third Interface node that is handling two IP addresses and eight user workloads as described in Figure 1-21.

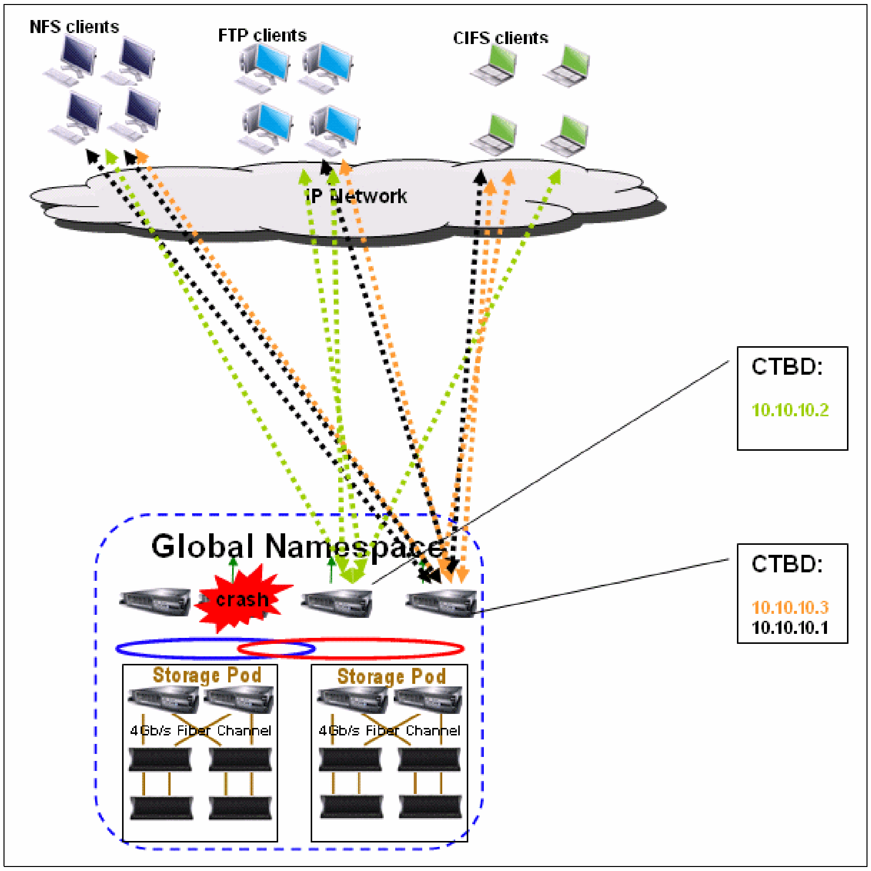


Figure 1-21   Interface node assignment and workload distribution according to the CTDB table

The original overall SONAS users workload was equally load balanced between the three Interface nodes, 33% of the workload each. After the Interface node crash and with the previous CTDB configuration, the workload is now 33% on the second Interface node and 66% on the third Interface node.

To avoid this situation, a simple configuration might be to create more IP addresses than Interface nodes that are available. Basically, in this example, six IP addresses, two per Interface node, might be more appropriate, as shown in Figure 1-22.

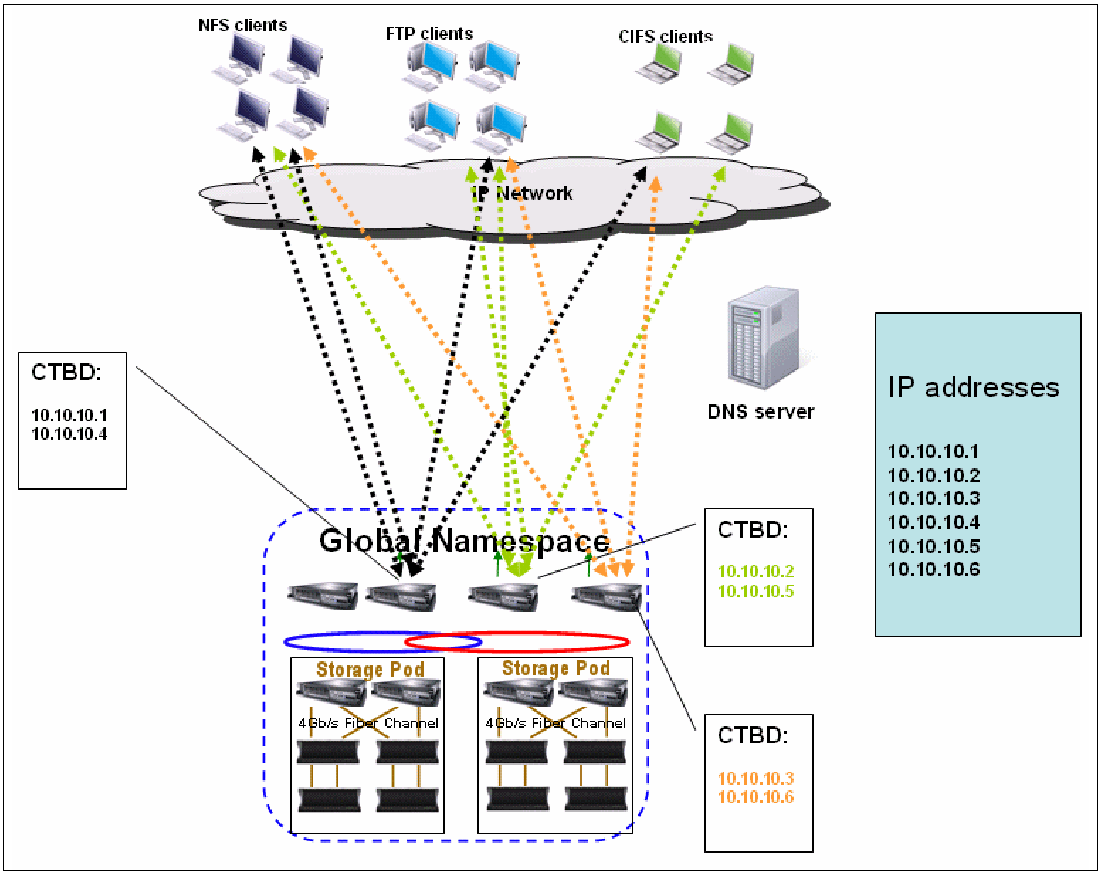


Figure 1-22 CTDB with more IP addresses than Interface nodes assigned

In that case, the original CTDB table is as follows:

- #1 is responsible for 10.10.10.1 and 10.10.10.4.
- #2 is responsible for 10.10.10.2 and 10.10.10.5.
- #3 is responsible for 10.10.10.3 and 10.10.10.6.

If there is a failure, the failing Interface node, previously in charge of two IP addresses, offloads its first IP address to the second Interface node and its second IP address to the third Interface node. Here is the new CTDB table:

- #2 is responsible for 10.10.10.1 and 10.10.10.2 and 10.10.10.5.
- #3 is responsible for 10.10.10.3 and 10.10.10.4 and 10.10.10.6.

The result is a 50-50% workload spread among the two remaining Interface nodes after the crash, as described in Figure 1-23.

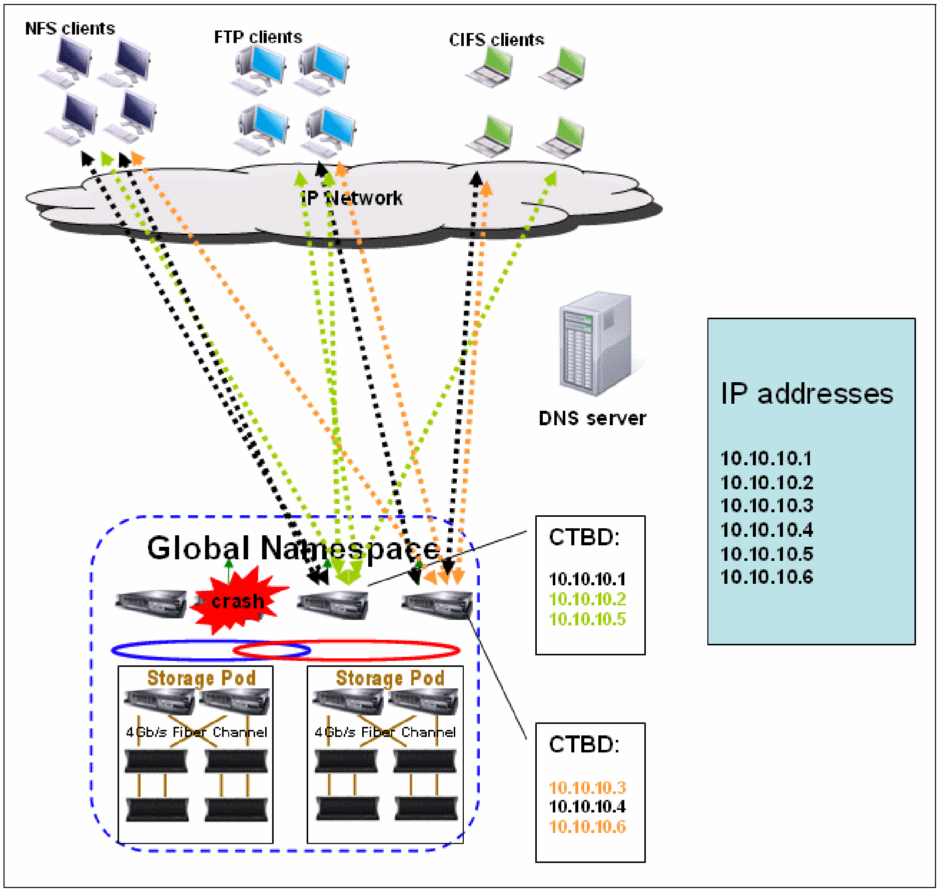


Figure 1-23 Even workload distribution after Interface node failure

After the first Interface node is back again, it is a new event. Here is the new CTDB table:

- #1 is responsible for 10.10.10.1 and 10.10.10.4.
- #2 is responsible for 10.10.10.2 and 10.10.10.5.
- #3 is responsible for 10.10.10.3 and 10.10.10.6.

The traffic can then be load balanced on the three Interface nodes again.

Share access

As described in “Data access and IP address balancing” on page 46, you must attach your SONAS system to your existing DNS and use a DNS round-robin configuration to load balance the user SONAS IP requests to all Interface nodes. (This is not *workload* load balancing.) But for any specific reason, you might want to use directly the IP address instead of the DNS host name.

For the CTDB layer, “Data access and IP address balancing” on page 46 shows you how to configure your IP Public network and CTDB to load balance the workload from one failed Interface node to the remaining nodes. The typical SONAS use is to map one SONAS user to a single Interface node to take advantage of the caching inside the Interface node. However, you might need to use the same CIFS share twice from the same SONAS user (through two drive letters), and then use two Interface nodes. However, do not do it with NFS shares. Because of the NFS design, the NFS protocol must send metadata to different NFS services that might be on two separate nodes in such a configuration.

So, it is preferable to map clients to a targeted IP address instead of a round-robin host name to ensure that it is not mounting multiple shares in the same cluster from different SONAS Interface nodes. It also is preferable to establish two public IPs per Interface node to help achieve node failover load balancing in the cluster design.

Managing access control lists

SONAS V1.5 supports the management of access control lists (ACLs) through the CLI and the GUI. For more information, see 7.5.6, "Managing access control lists" on page 559.

SONAS before Version 1.5 does not offer ACL management through the CLI and GUI. To manage ACLs in these versions, you must create a CIFS export so that the ACLs can be edited by using a Windows client.

1.6 Data protection considerations

It is important to consider data protection before installation because the preferred practice for your environment can affect sizing and even physical planning.

In most scenarios, organizations plan to extend their existing backup and recovery plans to include a SONAS appliance. It might include full or incremental backup regimes that they might have on their existing storage solutions. Because SONAS has the potential to host billions of files and grow to tens of petabytes, you might want to reassess your backup and recovery strategy.

If you already use Tivoli Storage Manager for backup, you can connect SONAS to your Tivoli Storage Manager backup server for backups. You must consider the volume of data that is being backed up and the potential storage and tape capacities that you require to host the data pools. Tivoli Storage Manager agents are already installed in SONAS and can communicate with SONAS to request a parallel GPFS scan of all files in the target file system and return a target list of files to be backed up. It means that the Tivoli Storage Manager agents can then back up the files without the agents needing to do a file scan first. Tivoli Storage Manager based backups can be run in parallel from one or more Interface nodes to maximize your backup speeds. Restores are done with proxies.

If you are using Network Data Management Protocol (NDMP)-based backup engines, such as Comm Vault, Symantec NetBackup, or Legato, you can configure SONAS for NDMP-based backups instead. However, NDMP-based backups are not compatible with Tivoli Storage Manager based backups. The options that you have for backups are mutually exclusive. As with Tivoli Storage Manager, your NDMP-based incremental backup triggers an internal parallel scan of all changed files and then they are made available to the NDMP server. NDMP-based backups are done over TCP/IP, and your backup server is responsible for streaming the returned data to tape.

Decide whether you want to add an independent data protection solution specifically for your SONAS solution or to use an existing data protection platform.

For backups, a SONAS solution supports backups by using Tivoli Storage Manager in native mode or with NDMP. Traditional Tivoli Storage Manager services use Tivoli Storage Manager agents that are already installed in each SONAS Interface node. These agents integrate with SONAS internals and can generate a list of changed files for backup at a rate of about 1 million files per minute. Tivoli Storage Manager then can do a targeted backup of the changed files. If you are not using Tivoli Storage Manager, you can use an NDMP-aware backup solution. The options are mutually exclusive.

If you use tape for HSM, SONAS includes a powerful tape tier capability of storage management by using HSM features, and simplified policies for migrating data from disk to tape. If you plan to use space management on tape-based file system space (with HSM), Tivoli Storage Manager is the only option because in that scenario NDMP is not supported.

Note: HSM functions depend on Tivoli Storage Manager agents, which implies that you cannot use NDMP-based backups if you must implement HSM.

There are some granularity benefits to using NDMP data protection. It can help you use full and incremental backups on an independent file set level instead of a full file system level. Also, NDMP backups are snapshot-based, so you can back up a point-in-time view of your data.

After you determine the data protection vendor and scheme to follow, your IBM sales account representative works with you to size your retention and copy plans against the applications and capacities that are targeted for SONAS to properly size the hardware and software requirements that accompany the SONAS solution.

Backing up PBs of data might be problematic at best. With a large environment, it might be better to use replication or separate data that requires backup in independent file sets.

Also, in some cases, when the cluster meets performance challenges (because of high usage), frequent backup and replication job processing reduces cluster performance during job list scanning and data backup cycles. In this case, you might need to reduce the replication and backup frequency until appropriate cluster growth is provisioned and enabled.

For more information about data protection, see Chapter 6, “Backup and recovery, availability, and resiliency functions” on page 367.

1.6.1 Async replication considerations

If you are going to use async replication with SONAS, you must consider that, by default, the async replication tool creates a local snapshot of the file tree that is being replicated (at the source cluster). It uses the snapshot as the source of the replication to the destination system. It is the preferred disaster recovery (DR) method because it creates a defined point-in-time image of the data that is being protected from a disaster. With average file sizes, SONAS scan engines can typically build replication task lists at a rate of about 1 million files per minute. SONAS then seeks the changed blocks and delivers only the delta (changes) that were made since the last replication cycle. Again, it is preferable to not plan replication cycles too aggressively because the process takes some performance cycles from the cluster when it builds the task list. It also takes some bandwidth from the Interface node and back-end storage during the period of transfer. It is preferable to adjust gradually frequency cycles until optimal configuration is achieved.

Snapshot handling

After the completion of async replication, the snapshot that is created in the source file system is removed. However, you must ensure sufficient storage at the replication source and destination for holding a replica of source file tree and associated snapshots. A snapshot is a space-efficient copy of a file system when the snapshot is initiated. The space that is occupied by the snapshot at the time of creation and before any files are written to the file system is a few kilobytes for control structures.

No additional space is required for data in a snapshot before the first write to the file system after the creation of the snapshot. As files are updated, the space that is used increases to reflect the main branch copy and also a copy for the snapshot. The cost of it is the actual size of the write rounded up to the size of a file system block for larger files or the size of a subblock for small files. In addition, there is a cost for more inode space and indirect block space to keep data pointers to both the main branch and snapshot copies of the data. This cost grows as more files in the snapshot differ from the main branch, but the growth is not linear because the unit of allocation for inodes is chunks in the inode file, which are the size of the file system subblock.

After the completion of the async replication, a snapshot of the file system that contains the replica target is created. The effect of snapshots on the SONAS capacity depends on the purpose of the snapshots. If the snapshots are used temporarily for creating an external backup and removed afterward, the effect is most likely not significant for configuration planning. In cases where the snapshots are taken frequently for replication or as backup to enable users to do an easy restore, the effect cannot be disregarded. However, the effect depends on the frequency with which the snapshot is taken, the length of time that each snapshot exists, and the number of the files in the file system that are changed by the users in addition to the size of the writes and changes.

Backup for disaster recovery purposes

The following list has key implications about using the HSM functions when file systems are backed up for disaster recovery purposes with the async replication engine:

- Source and destination primary storage capacity: The primary storage on the source and destination SONAS systems must be reasonably balanced in terms of capacity. HSM allows for the retention of more data than the primary storage capacity, and async replication is a file-based replication. So, planning must be done to ensure that the destination SONAS system has enough storage to hold the entire contents of the source data (both primary and secondary storage).
- HSM at destination: If the destination system uses HSM on the SONAS storage, consider having enough primary storage at the destination to ensure that the change delta can be replicated over into its primary storage as part of the DR process. If the movement of the data from the destination location’s primary to secondary storage is not fast enough, the replication process can outpace this movement, causing a performance bottleneck in completing the disaster recovery cycle.

  Therefore, the capacity of the destination system to move data to the secondary storage must be sufficiently configured to ensure that enough data was pre-migrated to the secondary storage to account for the next async replication cycle. Also, ensure that the amount of data to be replicated can be achieved without waiting for movement to secondary storage. For example, enough Tivoli Storage Manager managed tape drives must be allocated and operational, along with enough media.

Failure groups

SONAS allows you to organize your hardware into failure groups. A failure group is a set of disks that share a common point of failure that can cause them all to become simultaneously unavailable. SONAS software can provide RAID 1 mirroring at the software level. In this case, failure groups are defined that are duplicates of each other. They are defined to be stored on different disk subsystems. If a disk subsystem fails and cannot be accessed, the SONAS software automatically switches to the other half of the failure group. Expansion racks with storage pods can be moved away from each other for the length of the InfiniBand cables. The longest available cable is 50 m. It means that, for example, you are allowed to scratch the cluster and move two storage expansion racks to a distance of 50 m and create a mirror on a failure group level between these two racks.

With the failure of a single disk, if you did not specify multiple failure groups and replication of metadata, SONAS cannot continue because it cannot write logs or other critical metadata. If you specified multiple failure groups and replication of metadata, the failure of multiple disks in the same failure group puts you in the same position. In either of these situations, GPFS forcibly unmounts the file system. It is preferable to replicate at least the metadata between two storage pods, so you might consider creating two failure groups for two storage pods.

If two substorage devices are placed in each Storage node pair, consider replicating metadata only between both substorage devices to offer metadata protection for a failed substorage solution.

Redundancy

SONAS is designed to be a highly available storage solution. This high availability relies on hardware redundancy and software high availability with GPFS and CTDB.

As you plan to integrate SONAS into your own existing infrastructure, you must ensure that all external services and equipment are also highly available. Also, ensure that there is enough redundancy in your cluster solution to provide adequate performance and reliability for services if there is any component failure. Your SONAS needs an Active Directory Server (or LDAP) for authentication. Consider whether this authentication server, and the NTP and DNS servers, are redundant.

From a hardware point of view, consider whether you have redundant power and whether there are network switches for the public network.

1.6.2 Backup considerations

There are also preferred practices for backing up your storage. First, consider stopping your applications cleanly to ensure consistent data capture at the time that backups or snapshots are created. Next, take a snapshot and use it for the backup process when you restart your application.

1.7 Summary

This chapter describes the various activities and requirements to deploy a SONAS environment.

It is important to ensure that you are fully aware of all the power, cooling, size, weight, and restrictions of the final design. Also, consider any special delivery requirements for your site (for example, truck size limitations, or elevator size and weight restrictions).

IBM uses two Technical Delivery Assurance (TDA) reports that review in detail all the aspects of a SONAS sale. The first TDA is for the order (from a sales and a sales planning perspective), and is designed to ensure that your IBM representatives have a clear understanding of your requirements and that SONAS is can meet your expectations.

The second TDA covers the delivery and installation readiness aspects of a SONAS deployment. It is to ensure that the appliance can be properly and safely and placed in the appropriate location of your data center. It covers environmental considerations, such as weight, height, power, access to the floor space, and lift carrying capability.

The TDA process includes a full review of all the details of the order to include the previously mentioned subjects to prevent surprises at delivery. Ensure that you are a part of the TDA review process before sales commitment and that you are also a part of the pre-installation TDA. For more information about the TDA review process and order assurance, talk to your IBM sales representative.

2

Installation and configuration for a SONAS appliance with DDN

This chapter provides information about the basic installation and configuration of your SONAS appliance. The SONAS appliance is integrated with high-density DDN type disk array technology. The information in this chapter is specific to the SONAS appliance. SONAS Gateway storage installation is described in Chapter 3, “Installation and configuration for SONAS Gateway solutions” on page 97.

> **Note:** DDN storage is no longer included in new SONAS installations. All new systems use Gateway storage. For more information, see Chapter 3, “Installation and configuration for SONAS Gateway solutions” on page 97. This chapter provides reference information for existing DDN Appliance installations.
>
> Customers who currently have a SONAS appliance with DDN and require additional storage capacity must order a supported Gateway storage solution (DCS3700, Storwize V7000, or XIV). This configuration is referred to as “intermix”, is requested by RPQ, and requires a new RXC frame and storage pod. Different storage types cannot be intermixed in the same storage pod.

This chapter describes the following topics:

- Pre-installation
- Installation
- Post software installation
- Software configuration
- Sample environment
- Creating exports for data access
- Modifying access control lists to the shared export

2.1 Pre-installation

At this point, your SONAS purchase is completed and delivered to you. You are now ready to integrate your SONAS appliance into your existing environment. Complete the following steps:

1. Review the floor plan and pre-installation planning sheet to determine whether all information was provided.
2. If the pre-installation planning sheet is not complete, contact the Storage Administrator. This information is required through the rest of the installation, and the installation cannot start until the pre-installation planning sheet is complete.
3. The IBM authorized service provider does all the necessary preliminary planning work, which includes verifying the information in the planning worksheets to ensure that you understand the specific requirements, such as the physical or networking environments for the SONAS system.

2.2 Installation

Installation of a SONAS appliance requires both hardware and software installation.

When you are reading this chapter, remember SONAS V1.3 introduced the Integrated Management node feature. This feature allows the management functions of SONAS, which provides both the GUI and the command-line interface (CLI), to run on two of the Interface nodes in a SONAS system. It eliminates the need for a physically separate Management node, which was required in prior releases. The first Interface node in a SONAS system is designated as serving the role of the primary/active management server. The second Interface node in a SONAS system serves the role of the secondary (passive) management server.

2.2.1 Hardware installation

This section provides a high-level overview of the tasks that are needed to complete the SONAS hardware installation.

The IBM SONAS appliance that is shipped must be unpacked and moved to the necessary location. The appliance, when it is shipped from the IBM manufacturing unit, has all the connections to the nodes inside the rack already made. The internal connections are InfiniBand connections, which the nodes use to communicate with each other.

The IBM authorized service provider completes the following tasks:

1. Builds and assembles the hardware components into the final SONAS system.
2. Checks internal connections for any loose cables that might have occurred during transport.
3. Prepares connections for an expansion rack, if required.
4. Loads the disk drive modules into the storage drawer.
5. Powers on the storage controllers, storage expansions, KVM switch, and display module.
6. Powers on the Management role node.
7. Inserts the SONAS DVD into the first (bottom) Interface node. If older generation hardware is used, inserts the DVD into the Management role node (at the top of the rack).

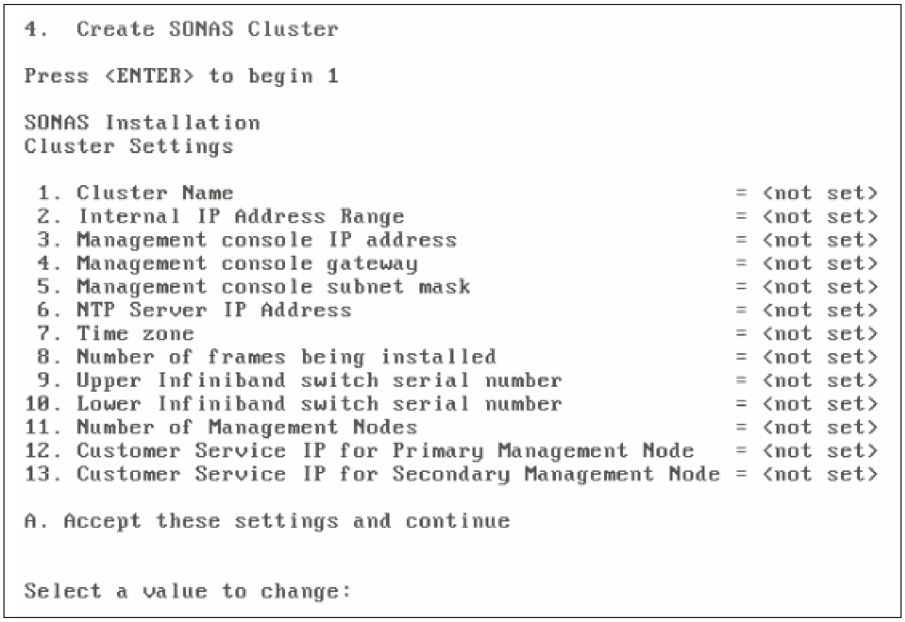
2.2.2 Software installation

This section provides a high-level overview of the tasks that are needed to complete the SONAS software installation.

After they complete the hardware installation, the IBM authorized service provider begins the software installation process. During this process, the IBM authorized service provider run the `first_time_install` script.

> **Note:** SONAS V1.5.1 introduces several first-time installation enhancements:
>
> - Options to specify the VLAN and external management adapter during the installation to reduce the requirements that are needed to run the **`chnwmgt`** command
> - Enhanced feedback and polling during the installation process to speed the process of nodes checking in and presenting additional information
> - Network connectivity to the Management node is permitted, which can speed the installation process

The initial steps require you to provide the configuration information that is shown in Figure 2-1. Provide this information before you start the installation. For the information that is needed, see the planning tables in Chapter 1, “Installation planning” on page 1.



```
4.  Create SONAS Cluster

Press <ENTER> to begin 1

SONAS Installation
Cluster Settings

 1. Cluster Name                                       = <not set>
 2. Internal IP Address Range                          = <not set>
 3. Management console IP address                      = <not set>
 4. Management console gateway                         = <not set>
 5. Management console subnet mask                     = <not set>
 6. NTP Server IP Address                              = <not set>
 7. Time zone                                          = <not set>
 8. Number of frames being installed                   = <not set>
 9. Upper Infiniband switch serial number              = <not set>
10. Lower Infiniband switch serial number              = <not set>
11. Number of Management Nodes                         = <not set>
12. Customer Service IP for Primary Management Node    = <not set>
13. Customer Service IP for Secondary Management Node = <not set>

A. Accept these settings and continue


Select a value to change:
```

Figure 2-1 Enter the basic SONAS configuration details

After you have input the required parameters, IBM authorized service provider does the installation and configuration of internal Ethernet and InfiniBand switches. In a multiple-rack environment, a script displays a message that prompts you to connect Ethernet switches from the second rack and other racks. When internal Ethernet and InfiniBand switches are installed, a script posts a note to power on Interface nodes and Storage nodes.

Nodes use the network boot option (TFTP). The ISO image is transferred onto the detected nodes. After the installation of the ISO image file, the configuration script attempts to detect any new nodes. Interface nodes might be detected faster than Storage nodes because they Interface nodes have a smaller software stack that is installed on them. The number of detected nodes is displayed. When all the nodes are detected, press Enter to stop node polling. Review the list of detected nodes. The script detects automatically the node's location, frame ID, and slot. Review this information and correct it if necessary. When all the information is entered correctly and confirmed, the software is configured.

2.2.3 Checking the health of the node hardware

The IBM authorized service provider runs a script that checks the health of the Management role nodes, Interface nodes, Storage nodes, InfiniBand switches, and Ethernet switches. This script ensures that the Management role node can communicate with the Interface nodes and the Storage nodes.

On the Management node, run **`cnrsmenu`**, press Enter, and then complete the following actions.

1. Select 2. System Checkout Menus.
2. Select 2. Refresh ALL Checks on All Nodes and Display Results and press Enter.

Review the result of the checks and verify that the checks have status of OK. For any problems that are reported by this command, see the *IBM Scale Out Network Attached Storage Troubleshooting Guide,* GA32-0717.

2.3 Post software installation

At the end of software installation, the SONAS system has a fully configured clustered file system (GPFS) with all the disks configured for the file system for use, the Management GUI Interface running, and a CLI Interface running.

A user admin is generated with a default password of *admin*. Add the cluster to the CLI Interface by running **`addcluster`**.

The SONAS appliance is now ready for further configuration.

2.4 Software configuration

The software configuration can either be done by IBM personnel as an extra service offering or by you as a system administrator of the SONAS appliance. It is carried out after the hardware and software installation. The pre-installation planning sheets in Chapter 1, "Installation planning" on page 1 require you to complete the administrative details and environment. These details include information about the network and authentication method that is used. The software configuration procedure uses a series of CLI commands.

Here is a high-level overview of the process:

1. Verify that the nodes are ready by checking the status of the nodes by running **`lsnode -c <clustername>`**. This command must display the roles of each node correctly and the GPFS status for each node must be active.
2. Configure the Cluster Manager (CTDB).

3. Create the failover group and file system (GPFS) by running **chdisk mkfs**.
4. Configure the DNS Server IP address and Domain by running **setnwdns**.
5. Configure the NAT gateway by running **mknwnatgateway**.
6. Integrate with an authentication server such as Active Domain (AD) or LDAP by running **cfgad** (and optional **cfgsfu**), **cfgldap**, or **cfgnt4**.
7. Configure the Data Path IP Addresses, Group, and attach IP address by running **mknw**, **mknwgroup**, and **attachnw**.
8. Connect a client workstation through a configured export to verify that the system is reachable remotely and that the interfaces work as expected.

2.5 Sample environment

This section provides example steps for installing and configuring a SONAS appliance.

Consider the following setup:

- Hardware considerations: The rack contains one Management node, two Interface nodes, two Storage nodes, switches, and InfiniBand connections.
- Software considerations: AD/LDAP is configured on an external server, and a file system and export information are available.

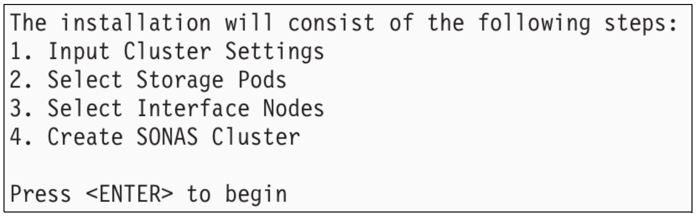
The cluster name in this example is `melbsonas.ad.mel.stg.ibm`.

2.5.1 Initial hardware installation

As mentioned in 2.2.1, “Hardware installation” on page 62, the IBM authorized service provider prepares the SONAS system. The racks are assembled. Nodes are interconnected so that they can communicate with each other. The software is installed on the nodes.

The cluster configuration data must be present in the pre-installation planning sheet. Run the `first_time_install` script, which configures the cluster.

In the following screen captures, you can see several steps that are captured during the installation procedure that is performed by the IBM authorized service provider. Figure 2-2 shows the options that are shown when the `first_time_install` script is run.

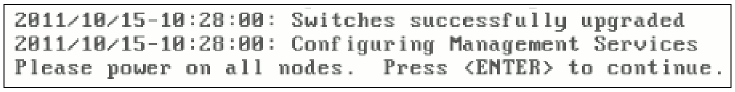


```
The installation will consist of the following steps:
1. Input Cluster Settings
2. Select Storage Pods
3. Select Interface Nodes
4. Create SONAS Cluster

Press <ENTER> to begin
```

Figure 2-2 Sample of first_time_install being run

As the script proceeds, it asks for the configuration parameters to configure the cluster. These details include the Management node IP, Internal IP Range, Root Password, Subnet IP Address, NTP Server IP, and more.

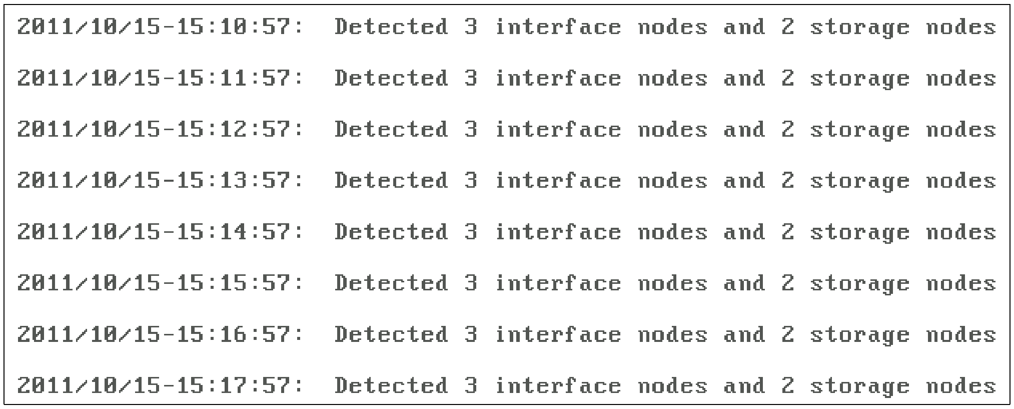
When the script finishes configuring internal Ethernet and InfiniBand switches, it prompts you to power on all nodes, as shown in Figure 2-3.



```
2011/10/15-10:28:00: Switches successfully upgraded
2011/10/15-10:28:00: Configuring Management Services
Please power on all nodes.  Press <ENTER> to continue.
```

Figure 2-3 Power on all nodes message during the running of the first_time_install script

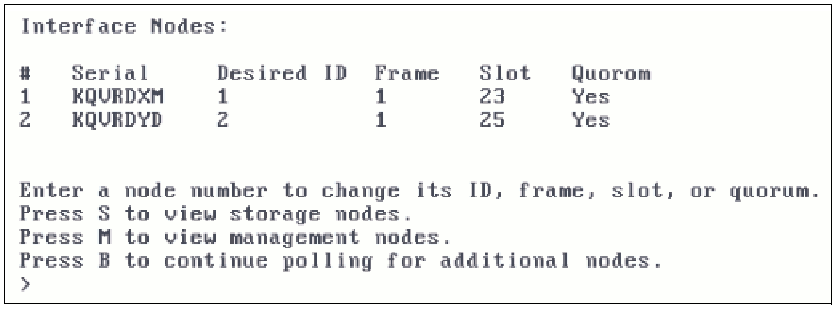
Figure 2-4 shows the detection of Interface nodes and Storage nodes when the nodes are powered on.



```
2011/10/15-15:10:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:11:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:12:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:13:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:14:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:15:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:16:57:  Detected 3 interface nodes and 2 storage nodes

2011/10/15-15:17:57:  Detected 3 interface nodes and 2 storage nodes
```

Figure 2-4 The script detects the Interface nodes and Storage nodes

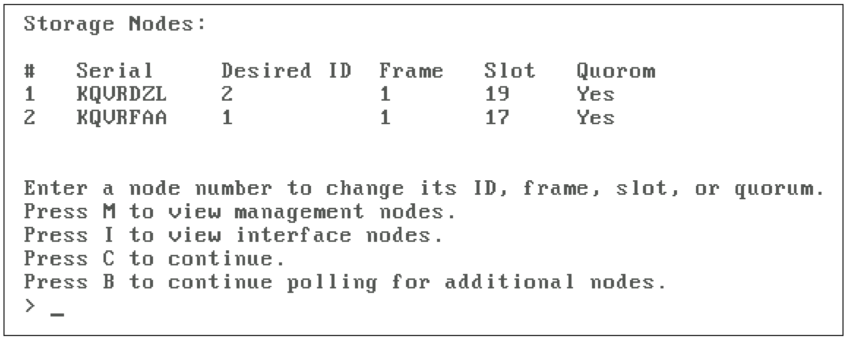
Figure 2-5 and Figure 2-6 show the assignment of ID for the Interface nodes and Storage nodes in the cluster, which includes the step to assign the nodes to be Quorum nodes or not.



```
Interface Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQVRDXM    1           1       23     Yes
2   KQVRDYD    2           1       25     Yes


Enter a node number to change its ID, frame, slot, or quorum.
Press S to view storage nodes.
Press M to view management nodes.
Press B to continue polling for additional nodes.
>
```

Figure 2-5 Identify the sequence of the Interface nodes and assigning quorum nodes

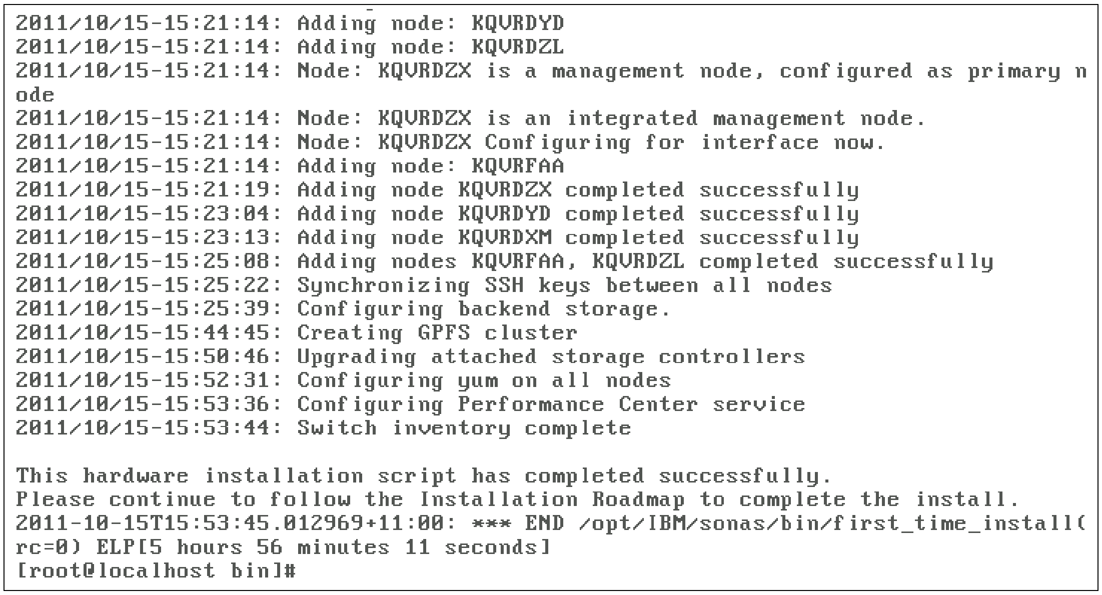
Figure 2-6 shows the Storage nodes.



```
Storage Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQVRDZL    2           1       19     Yes
2   KQVRFAA    1           1       17     Yes


Enter a node number to change its ID, frame, slot, or quorum.
Press M to view management nodes.
Press I to view interface nodes.
Press C to continue.
Press B to continue polling for additional nodes.
> _
```

Figure 2-6 Identify the sequence of Storage nodes and assigning quorum nodes

Figure 2-7 shows the configuration of the cluster where each of the Interface nodes and Storage nodes are added as a part of the cluster and the cluster nodes are prepared to communicate with each other. Figure 2-7 shows the end of the script, after which the cluster is successfully configured.



```
2011/10/15-15:21:14: Adding node: KQVRDYD
2011/10/15-15:21:14: Adding node: KQVRDZL
2011/10/15-15:21:14: Node: KQVRDZX is a management node, configured as primary n
ode
2011/10/15-15:21:14: Node: KQVRDZX is an integrated management node.
2011/10/15-15:21:14: Node: KQVRDZX Configuring for interface now.
2011/10/15-15:21:14: Adding node: KQVRFAA
2011/10/15-15:21:19: Adding node KQVRDZX completed successfully
2011/10/15-15:23:04: Adding node KQVRDYD completed successfully
2011/10/15-15:23:13: Adding node KQVRDXM completed successfully
2011/10/15-15:25:08: Adding nodes KQVRFAA, KQVRDZL completed successfully
2011/10/15-15:25:22: Synchronizing SSH keys between all nodes
2011/10/15-15:25:39: Configuring backend storage.
2011/10/15-15:44:45: Creating GPFS cluster
2011/10/15-15:50:46: Upgrading attached storage controllers
2011/10/15-15:52:31: Configuring yum on all nodes
2011/10/15-15:53:36: Configuring Performance Center service
2011/10/15-15:53:44: Switch inventory complete

This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.
2011-10-15T15:53:45.012969+11:00: *** END /opt/IBM/sonas/bin/first_time_install(
rc=0) ELP[5 hours 56 minutes 11 seconds]
[root@localhost bin]#
```

Figure 2-7 Cluster being created and the first_time_install script completes

The health of the system is then checked. The IBM authorized service provider then logs in to the Management node and runs the health check commands.

The **cnrsscheck** command is run or accessed by **cnrsmenu** to check the health of the Ethernet switches, InfiniBand switches, and the storage drawers. The **cnrsscheck** command also checks to see whether the roles for the nodes are assigned correctly and whether they can communicate with each other. Example 2-1 shows the command output for the sample cluster setup.

Example 2-1 Run cnrsscheck to check the overall health of the cluster

```
[root@Humboldt.mgmt001st001 ~]# cnrssccheck --nodes=all --checks=all
vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv  mgmt001st001   vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
================================================================================
Run checks on mgmt001st001
It might take a few minutes.

EthSwCheck ... OK
IbSwCheck ... OK
NodeCheck ... OK
================================================================================
IBM SONAS Checkout Version 1.00 executed on: 2010-04-21 23:07:57+00:00
Command syntax and parameters: /opt/IBM/sonas/bin/cnrsscdisplay --all
================================================================================
Host Name:         mgmt001st001
Check Status File: /opt/IBM/sonas/ras/config/rsSnScStatusComponent.xml
================================================================================
================================================================================
Summary of NON-OK Statuses:
  Warnings:   0
  Degrades:   0
  Failures:   0
  Offlines:   0
================================================================================
Ethernet Switch status:

Verify Ethernet Switch Configuration (Frame:1, Slot:41)               OK
Verify Ethernet Switch Hardware (Frame:1, Slot:41)                    OK
Verify Ethernet Switch Firmware (Frame:1, Slot:41)                    OK
Verify Ethernet Switch Link (Frame:1, Slot:41)                        OK
Verify Ethernet Switch Configuration (Frame:1, Slot:42)               OK
Verify Ethernet Switch Hardware (Frame:1, Slot:42)                    OK
Verify Ethernet Switch Firmware (Frame:1, Slot:42)                    OK
Verify Ethernet Switch Link (Frame:1, Slot:42)                        OK
================================================================================
InfiniBand Switch status:

Verify InfiniBand Switch Configuration (Frame:1, Slot:35)             OK
Verify InfiniBand Switch Hardware (Frame:1, Slot:35)                  OK
Verify InfiniBand Switch Firmware (Frame:1, Slot:35)                  OK
Verify InfiniBand Switch Link (Frame:1, Slot:35)                      OK
Verify InfiniBand Switch Configuration (Frame:1, Slot:36)             OK
Verify InfiniBand Switch Hardware (Frame:1, Slot:36)                  OK
Verify InfiniBand Switch Firmware (Frame:1, Slot:36)                  OK
Verify InfiniBand Switch Link (Frame:1, Slot:36)                      OK
================================================================================
Node status:

Verify Node General                                                   OK
================================================================================
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^   mgmt001st001   ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv  strg001st001   vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
================================================================================
Run checks on strg001st001
It might take a few minutes.

FcHbaCheck ... OK
DdnCheck ... OK
DdnLogCollector ... OK
================================================================================
IBM SONAS Checkout Version 1.00 executed on: 2010-04-21 23:10:46+00:00
Command syntax and parameters: /opt/IBM/sonas/bin/cnrsscdisplay --all
================================================================================
Host Name:         strg001st001
Check Status File: /opt/IBM/sonas/ras/config/rsSnScStatusComponent.xml
================================================================================
================================================================================
Summary of NON-OK Statuses:
  Warnings:   0
  Degrades:   0
  Failures:   0
  Offlines:   0
================================================================================
DDN Disk Enclosure status:
```

```
Verify Disk Enclosure Configuration (Frame:1, Slot:1)                 OK
Verify Disk in Disk Enclosure (Frame:1, Slot:1)                       OK
Verify Disk Enclosure Hardware (Frame:1, Slot:1)                      OK
Verify Disk Enclosure Firmware (Frame:1, Slot:1)                      OK
Verify Array in Disk Enclosure (Frame:1, Slot:1)                      OK
================================================================================
Fibre Channel HBA status:

Verify Fibre Channel HBA Configuration (Frame:1, Slot:17, Instance:0) OK
Verify Fibre Channel HBA Firmware (Frame:1, Slot:17, Instance:0)      OK
Verify Fibre Channel HBA Link (Frame:1, Slot:17, Instance:0)          OK
Verify Fibre Channel HBA Configuration (Frame:1, Slot:17, Instance:1) OK
Verify Fibre Channel HBA Firmware (Frame:1, Slot:17, Instance:1)      OK
Verify Fibre Channel HBA Link (Frame:1, Slot:17, Instance:1)          OK
================================================================================
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^  strg001st001   ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv  strg002st001   vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
================================================================================
Run checks on strg002st001
It might take a few minutes.

FcHbaCheck ... OK
DdnCheck ... OK
DdnLogCollector ... OK
================================================================================
IBM SONAS Checkout Version 1.00 executed on: 2010-04-21 23:13:26+00:00
Command syntax and parameters: /opt/IBM/sonas/bin/cnrsscdisplay --all
================================================================================
Host Name:         strg002st001
Check Status File: /opt/IBM/sonas/ras/config/rsSnScStatusComponent.xml
================================================================================
================================================================================
Summary of NON-OK Statuses:
  Warnings:   0
  Degrades:   0
  Failures:   0
  Offlines:   0
================================================================================
DDN Disk Enclosure status:

Verify Disk Enclosure Configuration (Frame:1, Slot:1)                 OK
Verify Disk in Disk Enclosure (Frame:1, Slot:1)                       OK
Verify Disk Enclosure Hardware (Frame:1, Slot:1)                      OK
Verify Disk Enclosure Firmware (Frame:1, Slot:1)                      OK
Verify Array in Disk Enclosure (Frame:1, Slot:1)                      OK
================================================================================
Fibre Channel HBA status:

Verify Fibre Channel HBA Configuration (Frame:1, Slot:19, Instance:0) OK
Verify Fibre Channel HBA Firmware (Frame:1, Slot:19, Instance:0)      OK
Verify Fibre Channel HBA Link (Frame:1, Slot:19, Instance:0)          OK
Verify Fibre Channel HBA Configuration (Frame:1, Slot:19, Instance:1) OK
Verify Fibre Channel HBA Firmware (Frame:1, Slot:19, Instance:1)      OK
Verify Fibre Channel HBA Link (Frame:1, Slot:19, Instance:1)          OK
================================================================================
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^  strg002st001   ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

At the end of the hardware installation, the cluster is created. The IBM authorized service provider then creates a CLI user, and adds the cluster to the GUI, as shown in Example 2-2.

Example 2-2 Create a CLI user by running mkuser

```
[SONAS]# mkuser -p Passw0rd cliuser
EFSSG0019I The user cliuser has been successfully created.

[root@furby.mgmt001st001 cli]# addcluster -h int001st001 -p Passw0rd
EFSSG0024I The cluster Furby.storage.tucson.ibm.com has been successfully added
```

You must enable the license, as shown in Example 2-3. Then, the cluster is ready for the rest of the software configuration.

Example 2-3 Enable license

```
[SONAS]# enablelicense
EFSSG0197I The license was enabled successfully!
```

2.5.2 Understanding the IP addresses for internal networking

For internal networking, there is a Management network and an InfiniBand network. The Management IP address and InfiniBand addresses, as described in “Planning IP addresses” on page 45, have the IP `172.31.*.*`. It is chosen from the three options that are available, as described in Chapter 1, “Installation planning” on page 1.

For this example, the `172.31.*.*` IP address range is selected. Although the first two parts of the IP address remain constant, the last two parts vary.

Management IP range

This network is used by the Management node to send management data to the Interface nodes and Storage nodes. It is a *private network* that is not reachable by outside clients. There is no data that is transferred in this network. Only management-related communication, such as commands or passing management-related information from the Management nodes to the Interface nodes and Storage nodes, is transferred. For more information, see the IBM SONAS IBM Knowledge Center, found at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/ovr_to_network_overview.html

In the example SONAS, the Management IP has a range of `172.31`.4.* for the Interface nodes, `172.31.8.*` for the Management node, and `172.31.6.*` for the Storage node. These ranges are set by the installation script when it creates the SONAS cluster.

Here, the first two parts of the IP address are constant. Then, depending on the Interface nodes, Management node, and Storage node, the management IP address is assigned as follows:

- Interface node: `172.31.4.*`
- Management node: `172.31.8.*`
- Storage node: `172.31.6.*`

Here, the last part of the IP address is incremented sequentially depending on the number of Interface nodes and Storage nodes.

InfiniBand IP range

This network range is used for data transfer between the Interface node and Storage node. Like the Management IP, it is a private network and is not reachable by outside clients, as described in Chapter 2, “Installation and configuration for a SONAS appliance with DDN” on page 61.

The InfiniBand IP has a range of `172.31.132.*` for the Interface nodes, `172.31.136.*` for the Management node, and `172.31.134.*` for the Storage node. These ranges are set by the installation script when it creates the SONAS cluster.

Here, the first two parts of the IP address are constant. Then, depending on the Interface nodes, Management node, and Storage node, the management IP address is assigned as follows:

- Interface node: `172.31.132.*`
- Management node: `172.31.136.*`
- Storage node: `172.31.134.*`

2.5.3 Accessing the GUI for the initial software installation

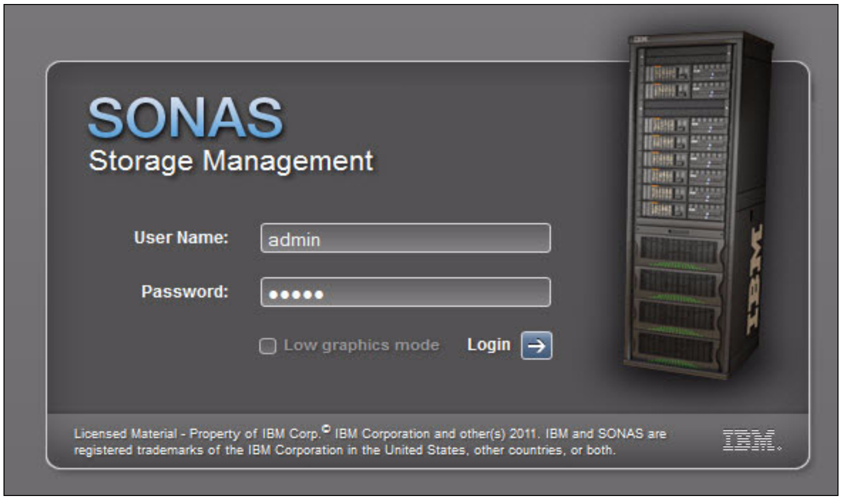
SONAS V1.3 allows the initial configuration to be done through the GUI, which eases the installation for IBM personnel or a system administrator. The information in this section explains the initial software configuration through a GUI. However, the CLI commands that can be used are mentioned as a reference. A few commands are available in the CLI only, in which case the instructions are CLI-based.

Logging in to the Management node

To begin, you must log in to the Management node by using the admin user ID:

```
https://your_management_ip:1081
```

Enter the admin password, as shown in Figure 2-8.


SONAS

Storage Management

User Name: admin

Password: ●●●●●

Licensed Material - Property of IBM Corp.© IBM Corporation and other(s) 2011. IBM and SONAS are registered trademarks of the IBM Corporation in the United States, other countries, or both.

Figure 2-8 SONAS login window

When you are logged in to the GUI, the menu on the left side can be used to navigate within the GUI. The selected menu is also shown at the top of the window. You can find the menus in the screen captures that follow. You can use this feature to easily identify the position of the selected function within the menu, as shown in Figure 2-9.

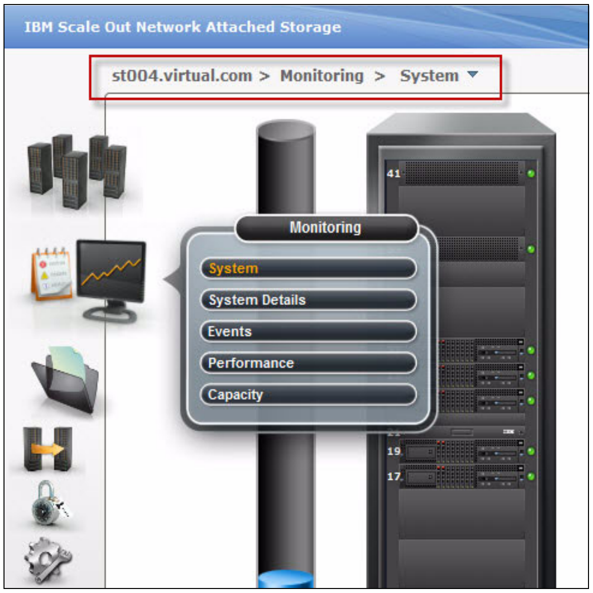

IBM Scale Out Network Attached Storage
st004.virtual.com > Monitoring > System
Monitoring
System
System Details
Events
Performance
Capacity
41
19
17

Figure 2-9 Menu navigation path

Verifying that the nodes are ready

Before you begin the configuration, make sure that the cluster was added to the Management interface and that the nodes are in the ready state by clicking **Monitoring** → **System**, as shown in Figure 2-10 on page 73. Confirm that all the physical nodes are visible and the status light is green, as shown in the highlighted section.

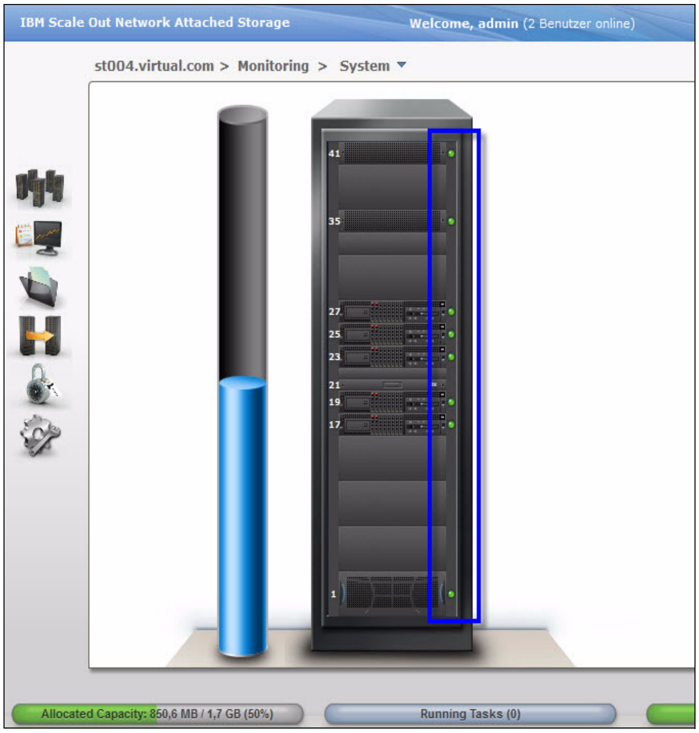

IBM Scale Out Network Attached Storage
Welcome, admin (2 Benutzer online)
st004.virtual.com > Monitoring > System
41
35
27
25
23
21
19
17
1
Allocated Capacity: 850,6 MB / 1,7 GB (50%)
Running Tasks (0)

Figure 2-10 Component status lights

If there are any yellow or red status lights, click them and you are redirected to the System Details pane, where additional information is shown in a text format. The equivalent CLI commands are shown in Example 2-4.

Example 2-4 Equivalent CLI commands for verifying that the nodes are ready

```
lsnode
```

GUI and CLI commands: For all examples in this chapter where the GUI is used, the equivalent CLI commands are provided. Although each of the GUI actions uses a CLI command, not all CLI commands are available in the GUI. Additionally, the GUI might wrap functions of multiple CLI commands together, meaning that a single GUI action might trigger multiple CLI commands. The progress section shows all of those CLI commands and their status in addition to the result. The audit log also provides the CLI commands that are used by the GUI.

2.5.4 Configuring the Cluster Manager

The Cluster Manager (CTDB) manages the SONAS cluster to a large extent. It is a part of the SONAS appliance and holds important configuration data for the cluster. The CTDB acts as the cluster manager for the SONAS appliance.

You configure the SONAS CTDB by running **cfgcluster** on the Management node. The command requires you to add a Public Cluster Name for the cluster, which is the name that is used to advertise the cluster to a neighboring network, such as a Windows client machine. This name is limited to 15 ASCII characters without any spaces or special characters, as shown in Example 2-5.

Example 2-5 Configure the Cluster Manager with the cfgcluster command

```
[SONAS]$ cfgcluster st004
Are you sure to initialize the cluster configuration ?
Do you really want to perform the operation (yes/no - default no):yes
(1/7) Prepare CIFS configuration
(2/7) Write CIFS configuration on public nodes
(3/7) Write cluster manager configuration on public nodes
(4/7) Import CIFS configuration into registry
(5/7) Write initial configuration for NFS,FTP,HTTP and SCP
(6/7) Restart cluster manager to activate new configuration
(7/7) Initializing registry defaults
EFSSG0114I Initialized cluster configuration successfully
EFSSG1000I The command completed successfully.
```

The command prompts you with the "`Do you really want to perform the operation?`" message. Type `yes` and press Enter to continue.

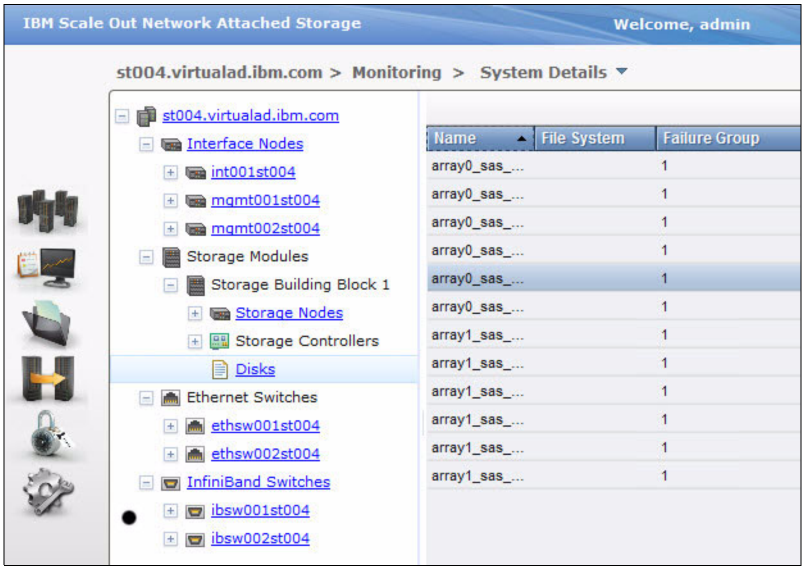
Verify that the cluster was configured by running **lscluster**. This command must display the CTDB cluster name that you used to configure the Cluster Manager. The output of the command is shown in Example 2-6. The public cluster name is `st004.virtual.com`.

Example 2-6 Verify the cluster details by running lscluster

```
[SONAS]$ lscluster
Cluster id           Name              Primary server Secondary server Profile
12402884792088706415 st004.virtual.com strg001st004   strg002st004     SONAS
EFSSG1000I The command completed successfully.
```

2.5.5 Listing all available disks

The available disks can be checked in the System Details pane of the GUI, as shown in Figure 2-11 on page 75.



IBM Scale Out Network Attached Storage Welcome, admin

st004.virtualad.ibm.com > Monitoring > System Details ▾

- st004.virtualad.ibm.com
  - Interface Nodes
    - int001st004
    - mgmt001st004
    - mgmt002st004
  - Storage Modules
    - Storage Building Block 1
      - Storage Nodes
      - Storage Controllers
      - Disks
  - Ethernet Switches
    - ethsw001st004
    - ethsw002st004
  - InfiniBand Switches
    - ibsw001st004
    - ibsw002st004

Name	File System	Failure Group
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array1_sas_...		1
array1_sas_...		1
array1_sas_...		1
array1_sas_...		1
array1_sas_...		1
array1_sas_...		1

Figure 2-11 List the available disks through the GUI

Example 2-7 shows the equivalent CLI commands.

Example 2-7 Equivalent CLI command for listing all available disks

```
lsdisk
```

2.5.6 Adding a second failure group

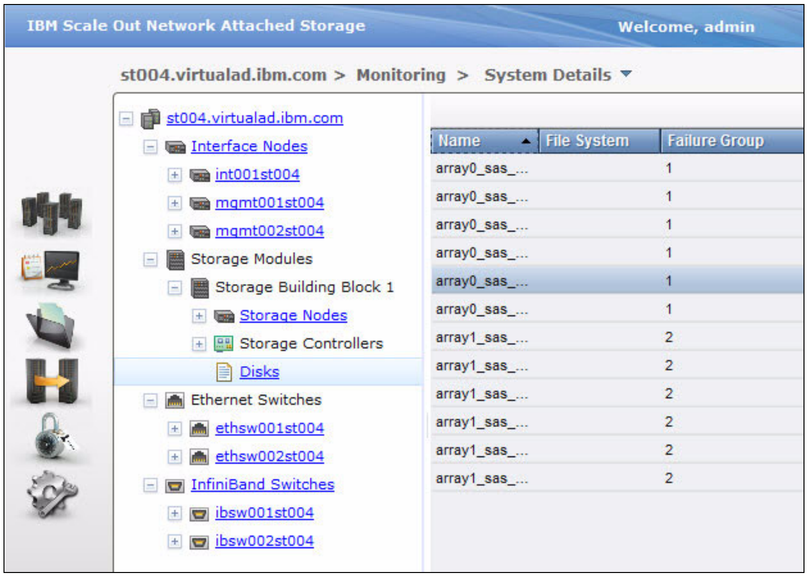
The failure group for all the disks is the default failure group, which is assigned at the time of cluster creation. To enable replication of data within the file system, there must be more than one failure group available.

In this example, a file system with replication enabled is created. Therefore, the failure group of disks that are part of the file system is modified to have another one. Run **chdisk** to modify the failure group property of the disk, as shown in Example 2-8.

Example 2-8 Change the failure group of disks by running chdisk

```
[SONAS]# chdisk
array1_sata_60001ff0732f8558c010001,array1_sata_60001ff0732f8578c030003,array1_sata_60001ff0732f
8598c050005,array1_sata_60001ff0732f85d8c090009,array1_sata_60001ff0732f85f8c0b000b,
array1_sata_60001ff0732f8608c0f000c --failuregroup 2
```

You can verify the changed failure groups by running `lsdisk` or by using the GUI, as shown in Figure 2-12.

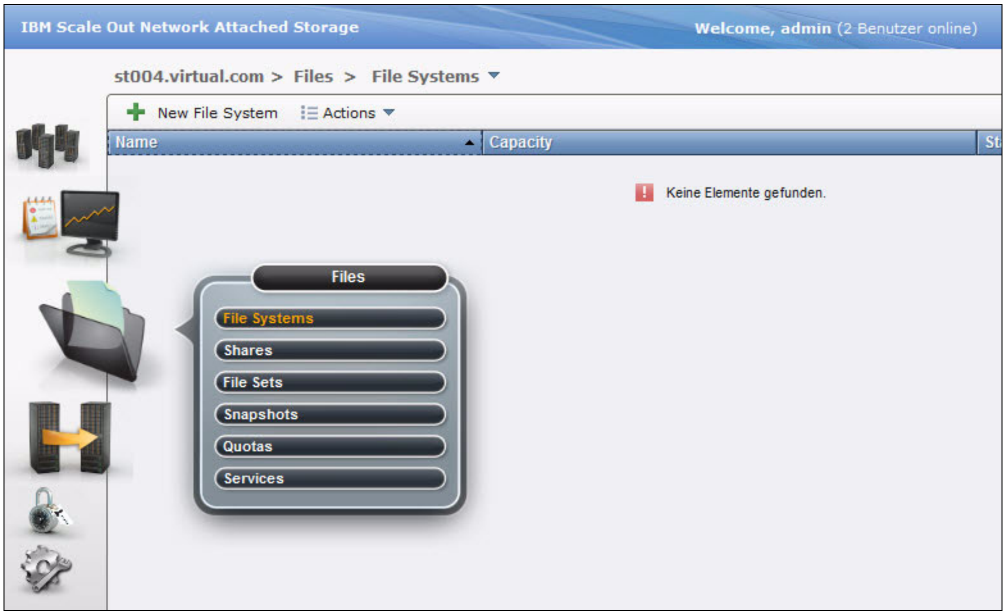


Name	File System	Failure Group
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array0_sas_...		1
array1_sas_...		2
array1_sas_...		2
array1_sas_...		2
array1_sas_...		2
array1_sas_...		2
array1_sas_...		2

Figure 2-12 Failure groups for disks

2.5.7 Creating the GPFS file system

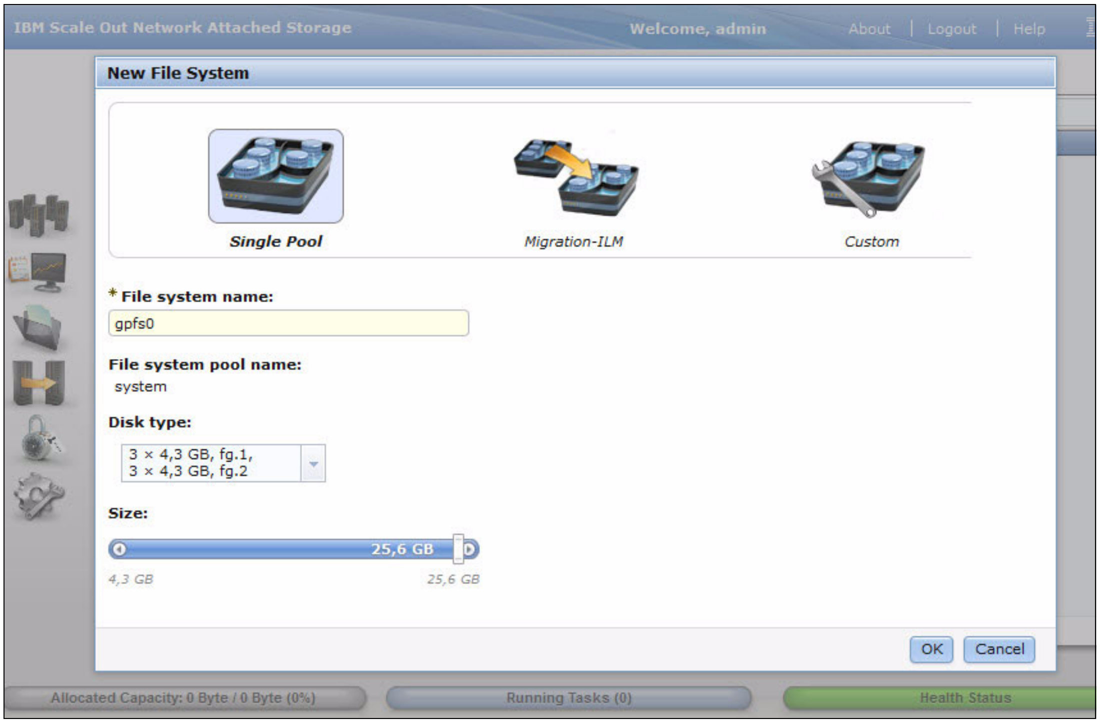
The underlying clustered file system that SONAS uses is the IBM GPFS file system. To verify existing file systems or to create one, click **Files** → **File Systems**. Figure 2-13 shows how to create the file system. Not all the available disks are used; in this example, two disks from failure group 1 and two disks from failure group 2 are used.


IBM Scale Out Network Attached Storage
Welcome, admin (2 Benutzer online)
st004.virtual.com > Files > File Systems
New File System
Actions
Name
Capacity
St
Files
File Systems
Shares
File Sets
Snapshots
Quotas
Services

Keine Elemente gefunden.

Figure 2-13 File Systems menu in the GUI

Figure 2-14 shows the New File System wizard for creating a file system within a single pool, which uses LUNs from a single disk tier. Two different failure groups with three available LUNs each are used for creation. As multiple disks are selected, the slider bar can be used to change the capacity, where the increments are single LUN sizes.



* File system name:
gpfs0

File system pool name:
system

Disk type:
3 × 4,3 GB, fg.1,
3 × 4,3 GB, fg.2

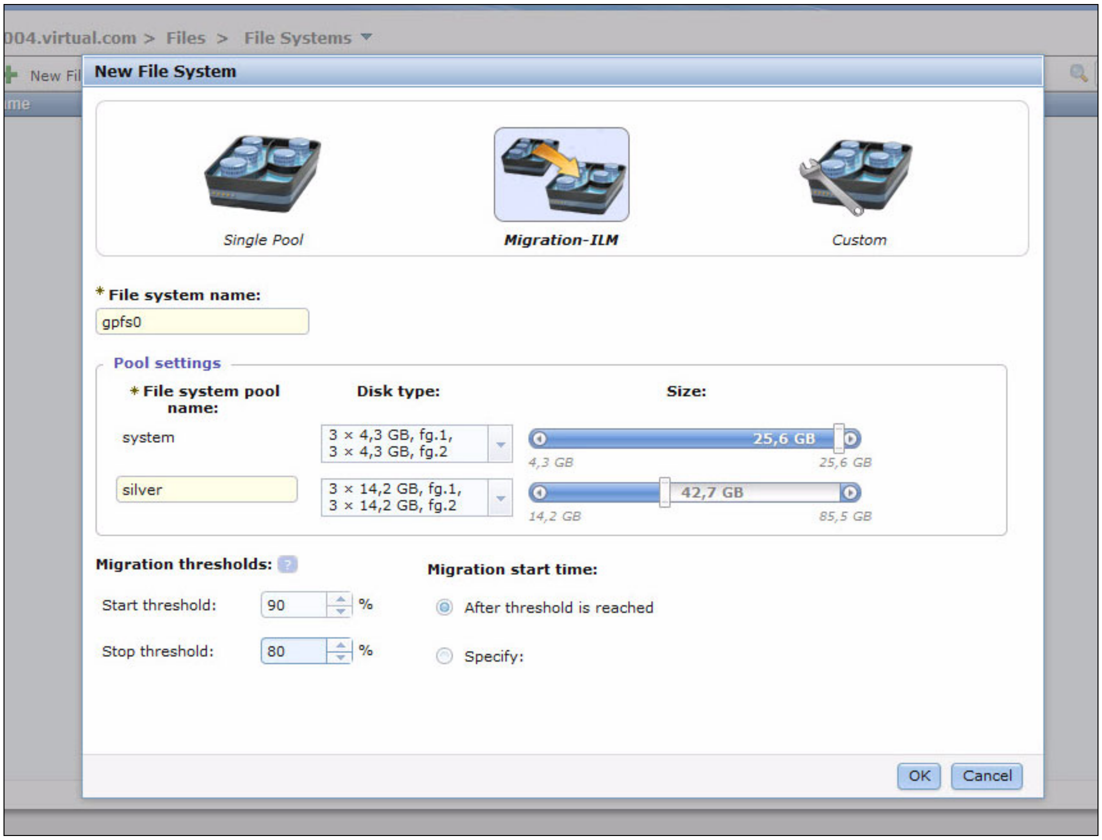
Size:
25,6 GB

Figure 2-14 New File System wizard

In SONAS V1.3, the earlier concept of a “master file system” is no longer used; the dependencies were removed.

To create a file system that uses ILM functions, click the **Migration ILM** tab and use a graphical, threshold-based migration rule. For more advanced policies and custom migration or placement rules, click the **Custom** tab, which provides a text editor.

In Figure 2-15, six LUNs in two failure groups from the performance tier are used as a primary pool. More LUNs are used for a *silver pool*, which is a migration target for data whenever 90% of the performance tier is used. 20% space is made available automatically after the threshold is exceeded.



New File System

* File system name: gpfs0

Pool settings

* File system pool name:	Disk type:	Size:
system	3 × 4,3 GB, fg.1, 3 × 4,3 GB, fg.2	25,6 GB (4,3 GB – 25,6 GB)
silver	3 × 14,2 GB, fg.1, 3 × 14,2 GB, fg.2	42,7 GB (14,2 GB – 85,5 GB)

Migration thresholds:

Start threshold: 90 %

Stop threshold: 80 %

Migration start time:

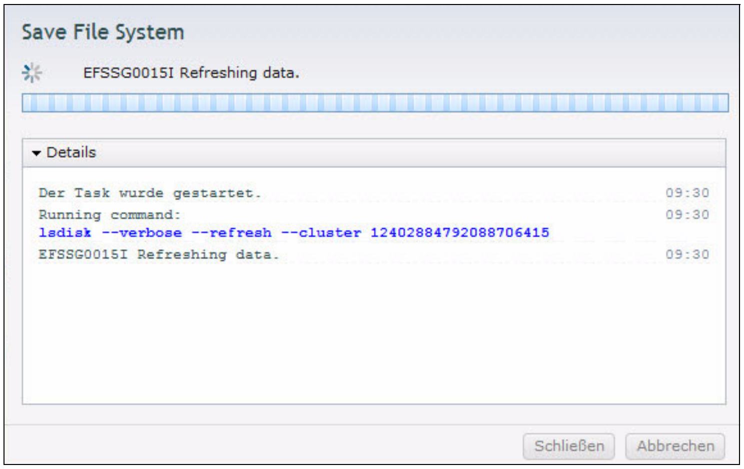
● After threshold is reached

○ Specify:

Figure 2-15   Add a second pool as a migration target

While the commands are run, the corresponding CLI commands are shown, in addition to the respective status output of those commands (see Figure 2-16). Using the CLI commands as a template, it is possible to use the GUI as a script generator for automation through the CLI.

Figure 2-16 shows file system creation tasks still in progress.



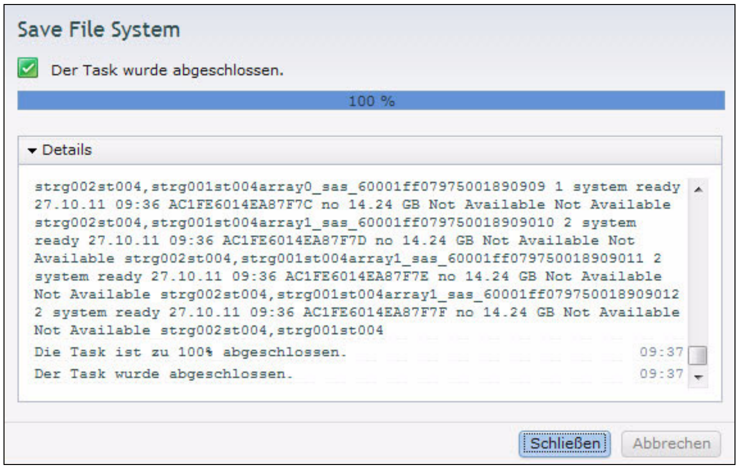
Save File System

EFSSG0015I Refreshing data.

```
Der Task wurde gestartet.                                     09:30
Running command:                                              09:30
lsdisk --verbose --refresh --cluster 12402884792088706415
EFSSG0015I Refreshing data.                                   09:30
```

Figure 2-16   Process of creating a file system

Figure 2-17 confirms the successful creation of the file system with all the individual steps completed.



Save File System

Der Task wurde abgeschlossen.

```
strg002st004,strg001st004array0_sas_60001ff07975001890909 1 system ready
27.10.11 09:36 AC1FE6014EA87F7C no 14.24 GB Not Available Not Available
strg002st004,strg001st004array1_sas_60001ff079750018909010 2 system
ready 27.10.11 09:36 AC1FE6014EA87F7D no 14.24 GB Not Available Not
Available strg002st004,strg001st004array1_sas_60001ff079750018909011 2
system ready 27.10.11 09:36 AC1FE6014EA87F7E no 14.24 GB Not Available
Not Available strg002st004,strg001st004array1_sas_60001ff079750018909012
2 system ready 27.10.11 09:36 AC1FE6014EA87F7F no 14.24 GB Not Available
Not Available strg002st004,strg001st004
Die Task ist zu 100% abgeschlossen.                            09:37
Der Task wurde abgeschlossen.                                  09:37
```

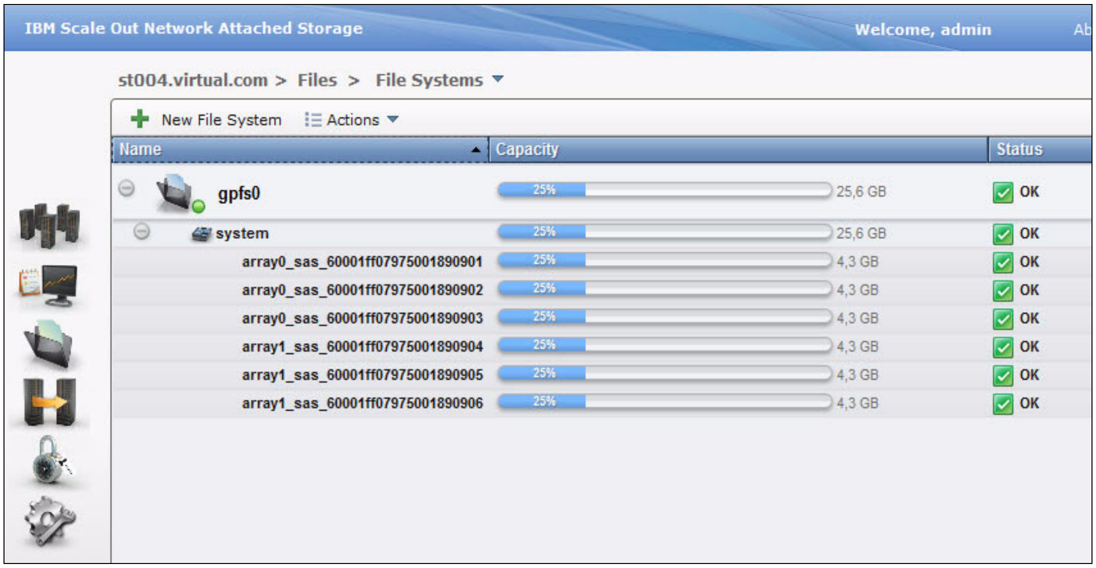
Figure 2-17 Create a New File System process finished

Example 2-9 shows the equivalent CLI commands.

Example 2-9 Equivalent CLI commands for creating the GPFS file system

```
lsdisk --verbose --refresh --cluster 12402884792088706415

chdisk
array0_sas_60001ff07975001890901,array1_sas_60001ff07975001890904,array0_sas_60001ff079750018909
02,array1_sas_60001ff07975001890905,array0_sas_60001ff07975001890903 --pool system --usagetype
dataAndMetadata --cluster 12402884792088706415

chdisk
array0_sas_60001ff07975001890907,array0_sas_60001ff07975001890908,array0_sas_60001ff079750018909
09 --pool silver --usagetype dataOnly --cluster 12402884792088706415

mkfs gpfs0 '/ibm/gpfs0' -b 256K -F
array0_sas_60001ff07975001890901,array1_sas_60001ff07975001890904,array0_sas_60001ff079750018909
02,array1_sas_60001ff07975001890905,array0_sas_60001ff07975001890903,array0_sas_60001ff079750018
90907,array0_sas_60001ff07975001890908,array0_sas_60001ff07975001890909 -j cluster --noverify -R
meta --logplacement striped --quota yes --cluster 12402884792088706415

mkpolicy gpfs0_generatedPolicy --rules 'RULE 'generatedMigrationRule' MIGRATE FROM POOL 'system'
THRESHOLD(90,80) TO POOL 'silver'RULE 'default' SET POOL 'system'' --cluster 'st004.virtual.com'

chkpolicy '/ibm/gpfs0' --policyNames gpfs0_generatedPolicy --cluster 12402884792088706415

setpolicy gpfs0 --policyNames gpfs0_generatedPolicy --cluster 12402884792088706415
```

To reconfirm the successful creation and review the layout of the newly created file system, click **File** → **File Systems**, as shown in Figure 2-18.



IBM Scale Out Network Attached Storage
Welcome, admin
Ab
st004.virtual.com > Files > File Systems
New File System
Actions

Name	Capacity		Status
gpfs0	25%	25,6 GB	✓ OK
system	25%	25,6 GB	✓ OK
array0_sas_60001ff07975001890901	25%	4,3 GB	✓ OK
array0_sas_60001ff07975001890902	25%	4,3 GB	✓ OK
array0_sas_60001ff07975001890903	25%	4,3 GB	✓ OK
array1_sas_60001ff07975001890904	25%	4,3 GB	✓ OK
array1_sas_60001ff07975001890905	25%	4,3 GB	✓ OK
array1_sas_60001ff07975001890906	25%	4,3 GB	✓ OK

Figure 2-18 View the newly created file system with disks

The equivalent CLI commands are shown in Example 2-10.

Example 2-10 Equivalent CLI commands for reviewing the layout of the newly created file system

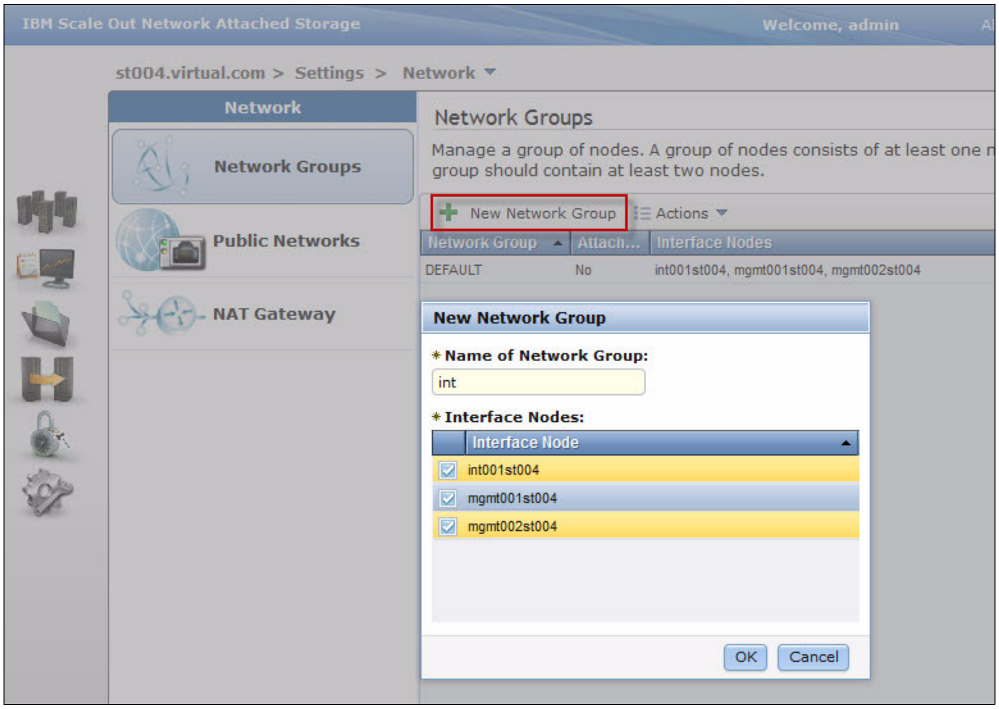
```
lsfs
lsdisk
lspolicy  -A
lspolicy  -P gpfs0_generatedPolicy
```

2.5.8 Configuring the Data Path IP address group

SONAS provides a flexible way of attaching multiple networks to the cluster. Depending on your network architecture, you might decide to spend time planning for it. You might also choose to accept the default settings if your configuration does not require special configuration of bonds, VLANs, or multiple networks.

Within SONAS, you can group a subset of Interface nodes into a *network group* and attach one or multiple networks to each network group. This group can be used to grant access to independent groups in addition to separating tenants on a VLAN or source network level.

The first step is to create a network group. In Figure 2-19, a single group, including all Interface nodes and client traffic serving nodes that host the SONAS GUI and CLI, is created.



Network Groups

Manage a group of nodes. A group of nodes consists of at least one r group should contain at least two nodes.

Network Group	Attach...	Interface Nodes
DEFAULT	No	int001st004, mgmt001st004, mgmt002st004

New Network Group

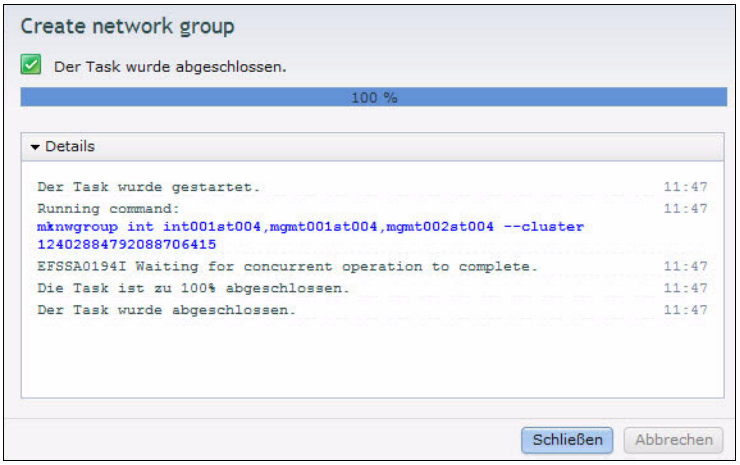
* Name of Network Group: int

* Interface Nodes:

	Interface Node
☑	int001st004
☑	mgmt001st004
☑	mgmt002st004

Figure 2-19 Create a network group

Figure 2-20 shows the CLI commands that are run in the background to complete the GUI task.



☑ Der Task wurde abgeschlossen.

```
Der Task wurde gestartet.                                              11:47
Running command:                                                       11:47
mknwgroup int int001st004,mgmt001st004,mgmt002st004 --cluster
12402884792088706415
EFSSA0194I Waiting for concurrent operation to complete.              11:47
Die Task ist zu 100% abgeschlossen.                                    11:47
Der Task wurde abgeschlossen.                                          11:47
```

Figure 2-20 CLI commands that are run behind the GUI to create the network group

The equivalent CLI commands to create the network group are shown in Example 2-11.

Example 2-11 Configure Data Path IP Group by running mknwgroup

```
mknwgroup int int001st004, mgmt001st004, mgmt002st004
lsnwgroup -r
```

2.5.9 Configuring Data Path IP addresses

Now that the network group is created, the next step is to create a network and choose which IP addresses from that network range SONAS must use for client access. It is a preferred practice to choose two Data Path IP addresses per Interface node to allow a 50:50 split to two other Interface nodes in case the node is set to maintenance mode. For example, this transfer might occur during a concurrent update (see Figure 2-21).

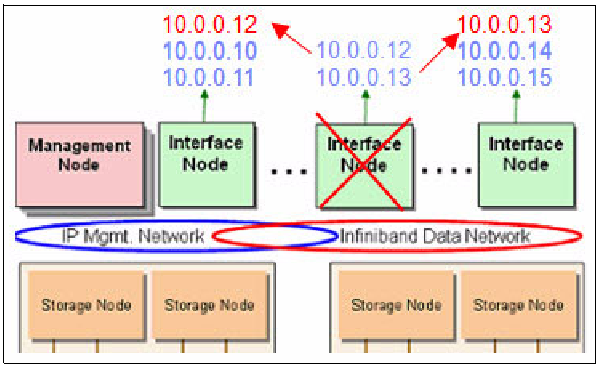
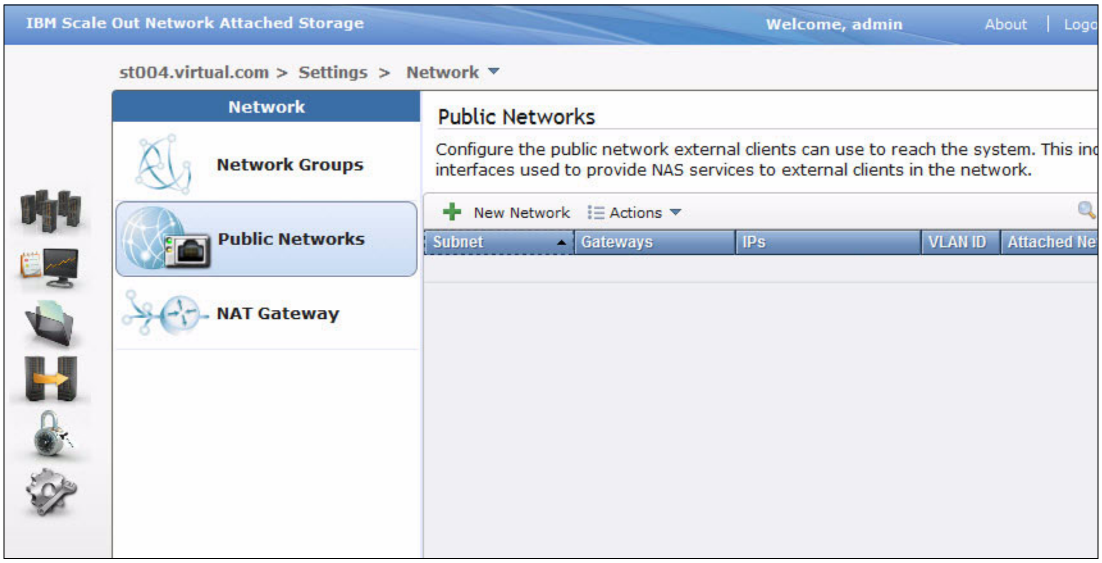


Figure 2-21 Failover - two Data Path IPs per Interface node allow a balanced failover

Creating a network

To create a network, click **Settings** → **Network** and click the **Public Networks** tab, as shown in Figure 2-22.



Public Networks

Configure the public network external clients can use to reach the system. This inc interfaces used to provide NAS services to external clients in the network.

Figure 2-22 No public networks are configured initially

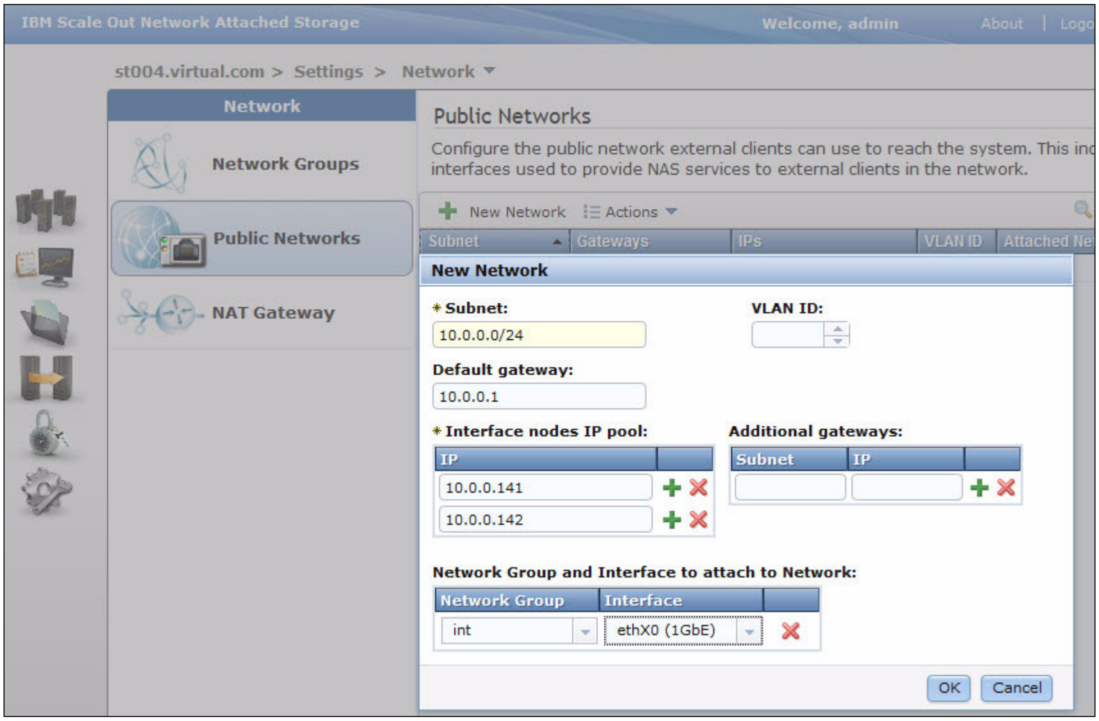
When you are creating a network, you define the subnet, which is unique and therefore is used as the identifier for this specific network configuration within SONAS afterward. A default gateway and optional specific routes can be configured for each network.

The configuration of IP addresses to be used by SONAS is accomplished by completing the Interface nodes IP Pool fields, as shown in Figure 2-23 on page 83. Ensure that you have at least one IP per Interface node; otherwise, it does not serve any traffic. Two IP addresses per Interface node provide better balancing if there is a failover.

VLAN configuration

Optionally, a VLAN can also be configured. For example, a customer might have networks in three different VLANs and plan to attach all three of them to SONAS. In that case, ensure that the network switch to which SONAS connects is set up correctly. There are several options for configuring the VLAN.

- Option A: SONAS connects to a switch port that sees all three VLANs.
- Option B: A SONAS Interface node with six ports (two onboard plus a quad-port expansion card) has three bonds (ethX0, ethX1, and ethX2) defined. Each bond connects to a different port on the switch with different VLANs configured on the switch level. In this setup, you must ensure that you assign the network to the adapter that connects to the corresponding port with the matching VLAN configuration on the switch.



Public Networks

Configure the public network external clients can use to reach the system. This inc interfaces used to provide NAS services to external clients in the network.

New Network

*Subnet: 10.0.0.0/24

VLAN ID:

Default gateway: 10.0.0.1

*Interface nodes IP pool:

IP
10.0.0.141
10.0.0.142

Additional gateways:

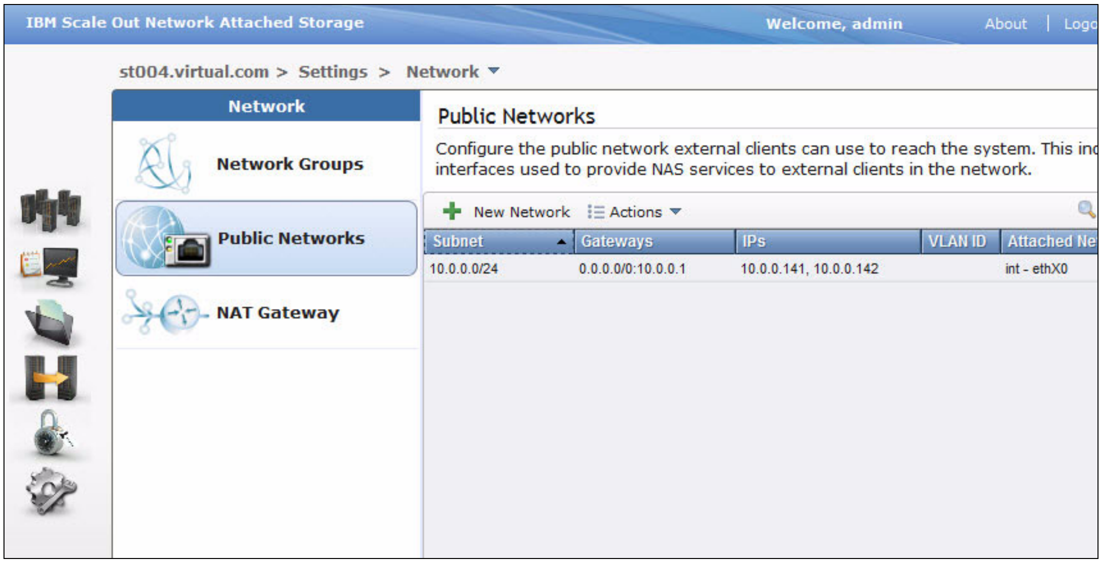
Subnet	IP

Network Group and Interface to attach to Network:

Network Group	Interface
int	ethX0 (1GbE)

Figure 2-23   Initial network configuration

Figure 2-24 shows the Public Networks pane of the SONAS GUI with the configured public network.



Public Networks

Configure the public network external clients can use to reach the system. This inc interfaces used to provide NAS services to external clients in the network.

Subnet	Gateways	IPs	VLAN ID	Attached Ne
10.0.0.0/24	0.0.0.0/0:10.0.0.1	10.0.0.141, 10.0.0.142		int - ethX0

Figure 2-24 Configured public network

Example 2-12 shows the equivalent CLI commands for configuring the public network that is shown in Figure 2-23 on page 83.

Example 2-12 Configure the Data Path IP by running mknw

```
mknw 10.0.0.0/24 0.0.0.0/0:10.0.0.1 ··add 10.0.0.141,10.0.0.142
attachnw 10.0.0.0/24 ethX0 -g int
```

2.5.10 Configuring the DNS Server IP addresses and domains

The SONAS appliance must be configured with the IP address of the Domain Name Services (DNS) servers and the domains. These IP addresses are also called the public IP addresses, which are accessible on your network. Only the management role node and the Interface nodes are accessible on your network.

Figure 2-25 shows a redundant set of DNS server and multiple search domains for name resolution.

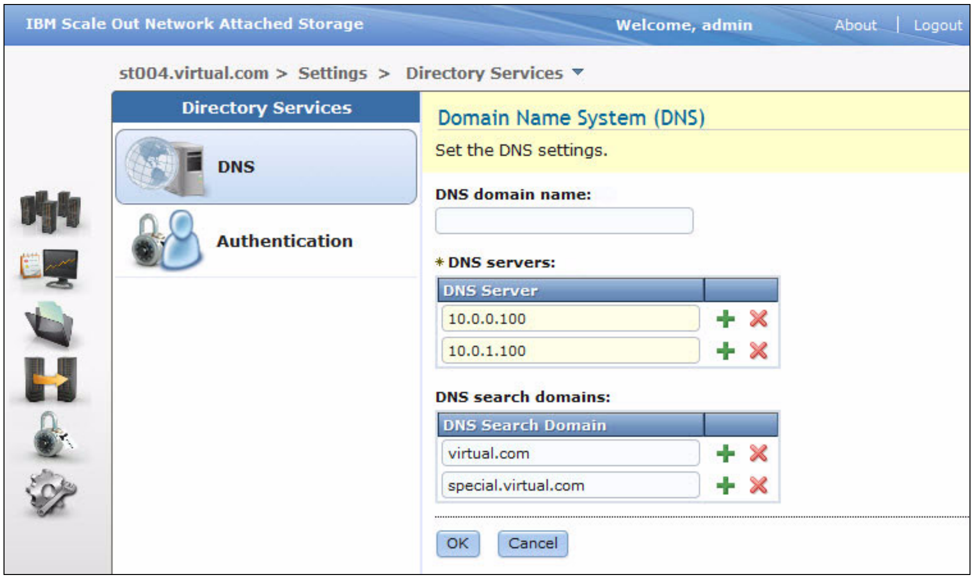


Figure 2-25 Add DNS servers

The equivalent CLI commands to configure the DNS are shown in Example 2-13.

Example 2-13 Configure DNS with only a DNS server IP

```
setnwdns 10.0.0.100,10.0.1.100 --domain --search 'virtual.com,special.virtual.com' --cluster
12402884792088706415
```

2.5.11 Configuring the Network Address Translation Gateway

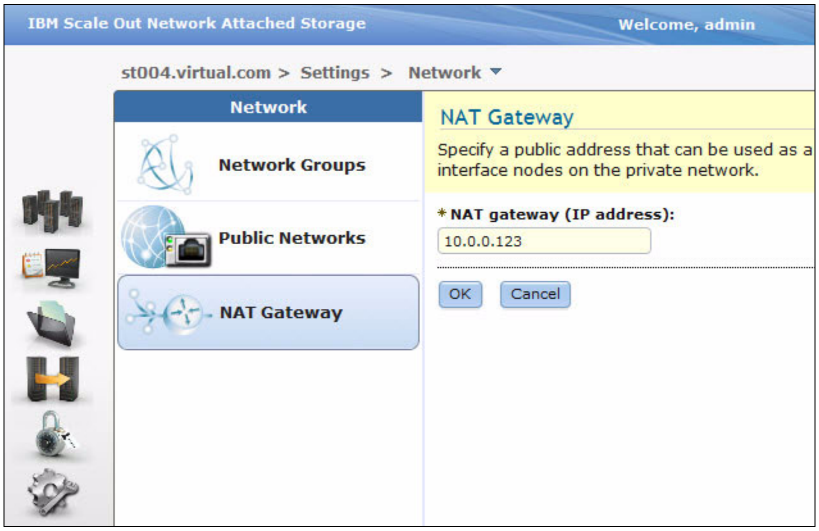
Network Address Translation (NAT) is a technique that is used with the network routers. The SONAS appliance has its Interface nodes that communication with each other by using a private IP address. This network is not accessible by your network.

The public IP addresses are the addresses through which you can access the Management role node and the Interface nodes. Hence, the Management role node and Interface nodes have a private IP address for internal communication and a public IP address for external communication. NAT allows a single IP address on your network or public IP address to be used to access the Management role node and Interface nodes on their private network IP addresses.

The network router converts the IP address and port on your network to a corresponding IP address and port on the private network. This IP address is not a data path connection and is not used for reading or writing files. It is used to provide a path from the Management role node and Interface nodes to your network for the authorization and authentication process.

If possible, put a NAT Gateway in your cluster to provide an alternative path for Management node communication to your Interface nodes in case external communication is not available and communication still requires a path for completing management communication tasks (this configuration is sometimes helpful for upgrades).

Figure 2-26 shows the SONAS GUI window to configure the NAT gateway.



Specify a public address that can be used as a
interface nodes on the private network.

* NAT gateway (IP address): 10.0.0.123

Figure 2-26 Configure the NAT gateway

Example 2-14 shows the equivalent **mknwnatgateway** commands to configure a NAT Gateway.

Example 2-14 CLI commands to configure a NAT Gateway

```
mknwnatgateway '10.0.0.123' --cluster 12402884792088706415
mknwnatgateway 10.0.0.123/23 ethX0 10.0.0.1 172.31.128.0/17
mgmt001st001,int001st001,int002st001,int003st001,int004st001,int005st001,int006st001
```

As you can see in the second command-line example, it is possible to define specific routing and hosting for the NAT Gateway. It might be required if Interface nodes are connected to different networks or VLANs, the Management node is on a separate network range, or clients are attached by private networks that do not have access to NTP and to the Active Directory server.

In this scenario, the public NAT gateway IP is `10.0.0.123`, the Interface is `ethX0`, the default gateway is `10.0.0.1`, the private network IP address is `172.31.128.0/17`, and the nodes that are specified are Management nodes and the six Interface nodes, as shown in Example 2-15. All the Management and Interface nodes talk to the outside word on their public IP through the NAT Gateway.

Example 2-15 Verify that the NAT Gateway is successfully configured by running lsnwnatgateway

```
[SONAS]# lsnwnatgateway
Public IP       Public interface Default gateway Private network Nodes
9.11.137.246/23 ethX0            9.11.136.1      172.31.128.0/17,
172.31.136.2,172.31.132.1,172.31.132.2,172.31.132.3,172.31.132.4,172.31.132.5,172.31.132.6
```

2.5.12 Configuring authentication: Active Directory and Lightweight Directory Access Protocol

SONAS requires that the users who access the appliance must be authorized and authenticated. You can choose to use Active Directory (AD) or LDAP for authentication and authorization. SONAS supports both authentication methods and has equivalent GUI and CLI commands for the respective configurations.

When users access SONAS that is configured with AD, they are required to enter their user ID and password. This user ID and password pair is sent across the network to the remote authentication/authorization server, which compares the user ID and password (hash) to the valid user ID and password combinations in the database. If they match, the user is considered to be authenticated. The remote server then sends a response to SONAS confirming that the user was successfully authenticated.

For NFS V3, there is no such authorization; access management is done by the IP or host name of the connecting NFS client. After this initial authorization, all user IDs (UIDs) that are passed from the NFS client to the NFS Server are trusted as *valid UIDs*. A more secure configuration can use Kerberos tickets for each user, which are granted from a central Kerberos ticket, passed with the user request, and accepted by the SONAS appliance.

Terminology:

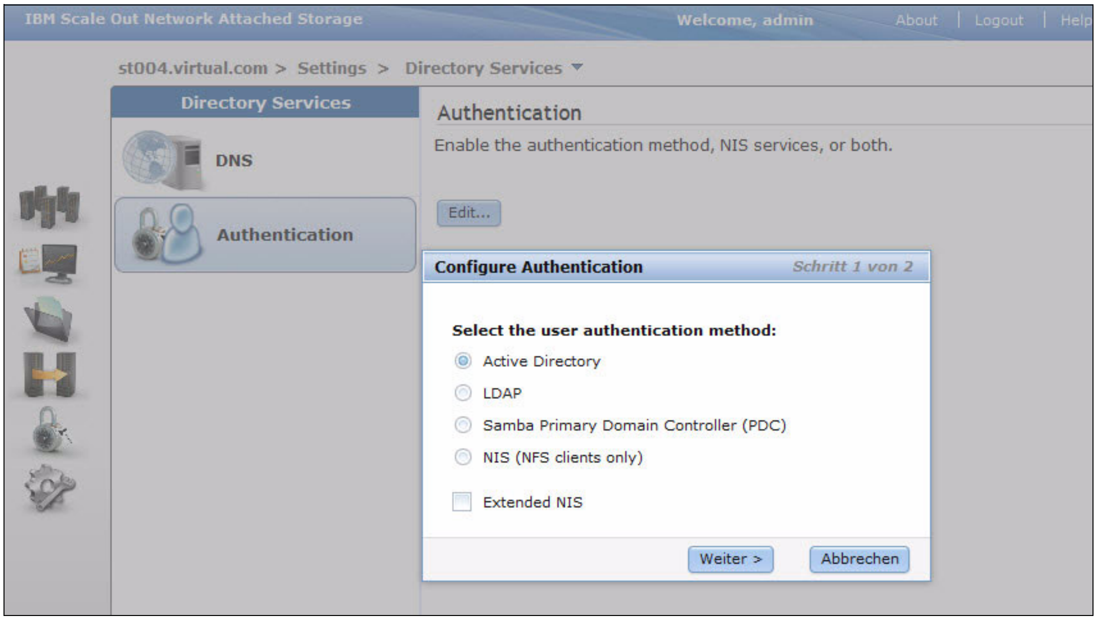
1. *Authentication* is the process of verifying the identity of the user. Users confirm that they are indeed the users they are claiming to be. It is typically accomplished by verifying the user ID and password.
2. *Authorization* is the process of determining whether the users are allowed to access. The users might have permissions to access certain files but might not have permissions to access others. It is typically done by ACLs.

The following sections describe the configuration of Active Directory Server (AD) and LDAP in detail. You choose one of the authentication methods.

Configuring for Active Directory

Authentication can be configured or reviewed in the GUI by clicking **Settings** → **Directory Services** → **Authentication**, as shown in Figure 2-27. You can use the CLI command to do more health checks on the authentication configuration and make user- or group-specific lookup queries.

DNS resolution must be present to look up the AD server based on domain and host name.



Enable the authentication method, NIS services, or both.

Configure Authentication — Schritt 1 von 2

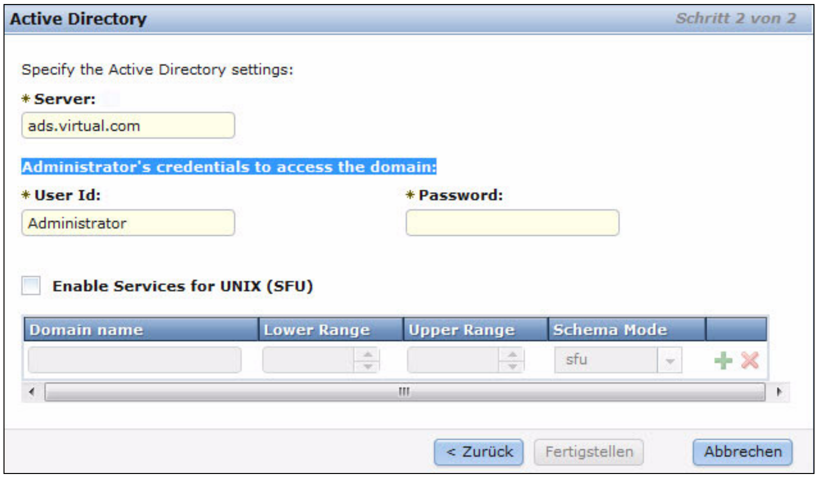
Select the user authentication method:

● Active Directory
○ LDAP
○ Samba Primary Domain Controller (PDC)
○ NIS (NFS clients only)

□ Extended NIS

Figure 2-27 Configure authentication

The AD configuration in Figure 2-28 uses a domain server named `ads.virtual.com`. The administrative user name and password are required for the initial configuration. They are not stored, but used to join the trusted domain and create a computer account for SONAS, which is a requirement to verify user and password combinations.



Active Directory

Specify the Active Directory settings:

* Server: ads.virtual.com

Administrator's credentials to access the domain:

* User Id: Administrator

* Password:

□ Enable Services for UNIX (SFU)

Domain name	Lower Range	Upper Range	Schema Mode
			sfu

Figure 2-28 Configure an authentication method by using Active Directory

The **`cfgad`** command that is shown in Example 2-16 is the equivalent command that is used in the SONAS GUI to create the Active Directory.

Example 2-16 Configure Active Directory by running cfgad

```
cfgad -as 10.0.0.100 -c sonas.virtual.com -u Administrator -p password
```

Verify that the cluster is now part of the AD domain by running **`chkauth`**, as shown in Example 2-17.

Example 2-17 Verify that the Active Directory server was successfully configured

```
[SONAS]$ chkauth -c sonas.virtual.com -t
Command_Output_Data             UID GID Home_Directory Template_Shell
CHECK SECRETS OF SERVER SUCCEED
```

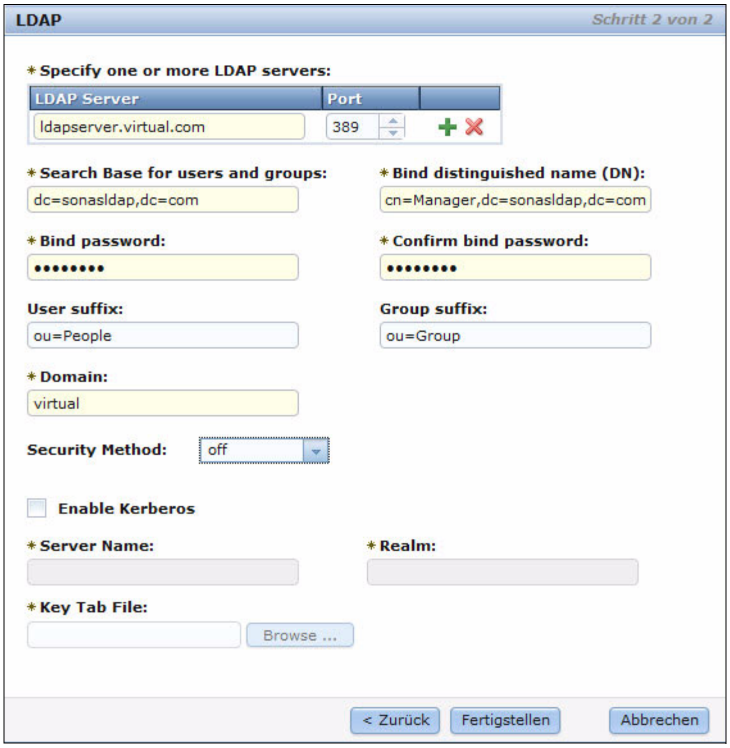
Configuring for Lightweight Directory Access Protocol

You can run **`cfgldap`** to configure the LDAP server. After the configuration, you must check whether it was successful by running **`chkauth`**. See Example 2-17 for the command use.

Figure 2-29 on page 89 shows the parameters that are used in the GUI for the LDAP configuration:

- The LDAP server IP is ldapserver.
- The suffix is dc=sonasldap,dc=com.
- rootdn is cn=manager, dc=sonasldap, dc=com.
- The password is secret.
- The SSL method is tls.

You can get this information from your LDAP administrator. It is found in the `/etc/eopn/ldap/slapd.conf` file on the LDAP server. This information is also in the pre-installation planning sheet that is found in Chapter 1, “Installation planning” on page 1.



LDAP

Schritt 2 von 2

* Specify one or more LDAP servers:

LDAP Server	Port
ldapserver.virtual.com	389

* Search Base for users and groups: dc=sonasldap,dc=com

* Bind distinguished name (DN): cn=Manager,dc=sonasldap,dc=com

* Bind password: ••••••••

* Confirm bind password: ••••••••

User suffix: ou=People

Group suffix: ou=Group

* Domain: virtual

Security Method: off

□ Enable Kerberos

* Server Name:

* Realm:

* Key Tab File:

Figure 2-29 Configure the authentication method by using LDAP

Example 2-18 shows the **cfgldap** command that corresponds to the LDAP configuration process through the GUI. The parameters are similar to the ones that are used in the GUI:

- The LDAP server (ls) is `ldapserver`.
- The suffix (lb) is “`dc=sonasldap,dc=com`”.
- rootdn (ldn) is “`cn=manager, dc=sonasldap, dc=com`”.
- The password (lpw) is `secret`.
- The SSL method (ssl) is `tls`.

Example 2-18 LDAP authentication configuration by running cfgldap

```
[SONAS]# cfgldap -c sonas.virtual.com -d virtual.com -lb “dc=sonasldap,dc=com” -ldn
“cn=Manager,dc=sonasldap,dc=com” -lpw secret -ls ldapserver -ssl tls -v
```

Verify that the cluster is now part of the LDAP server by running **chkauth**, as shown in Example 2-19.

Example 2-19 Verify that the LDAP server was successfully configured

```
[SONAS]# chkauth -c sonas.virtual.com -t
Command_Output_Data            UID GID Home_Directory Template_Shell
CHECK SECRETS OF SERVER SUCCEED
```

2.6 Creating exports for data access

SONAS allows clients to access the data that is stored on a file system by using protocols such as CIFS, NFS, FTP, HTTPS, SCP, and SFTP. Data exports (also referred to as *shares*) are a subtree within the SONAS global namespace that are shared and can be accessed by clients. Exports can be created by using the SONAS GUI or by running **`mkexport`**.

When creating an export, you must enter the Sharename and a Directory Path for the export. The Directory Path is the path where the directory that is to be accessed by the clients is. This directory is seen by the clients with its Sharename. The Sharename is used by the clients to mount the export or access the export.

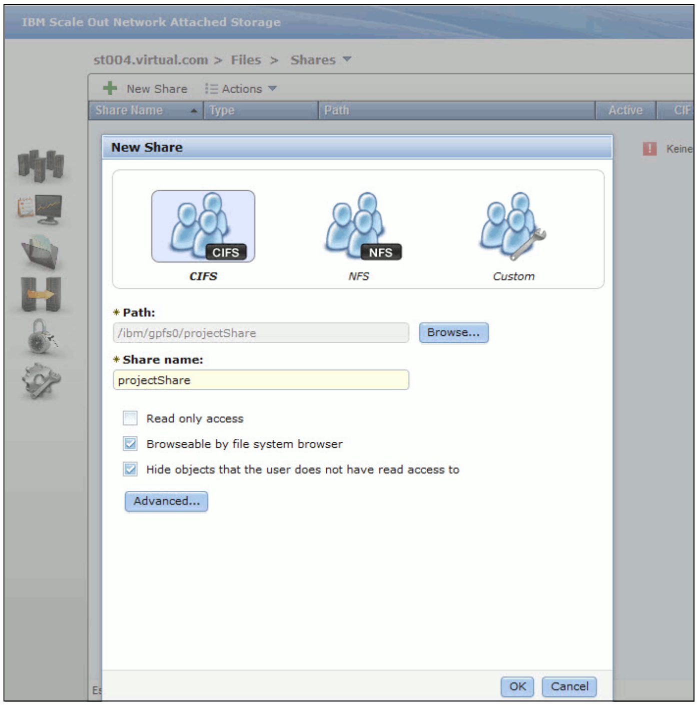
Depending on the protocol for which you want to configure your export, you must pass the respective parameters. As an administrator, you must provide all these details:

1. FTP has no parameter.
2. NFS requires you to pass certain parameters, for example:
   - Client/IP/Subnet/Mask: The clients that can access the NFS share. "*" implies all clients can access the share.
   - `ro` or `rw`: Depending on whether the export must have read-only access or read/write access.
   - `root_Squash`: A security mechanism that maps (client) root to anonymous. It is enabled by default. You can set it to `no_root_squash`, which is not recommended other than for testing.
   - async: It is enabled by default. You can set it to sync if required.
3. CIFS requires you to have certain parameters, for example:
   - browsable: Can the (Windows) client see the share? The default is Yes.
   - comment: You can write any comment for the CIFS export.

In the SONAS V1.3 or later GUI, you have three options to create a share:

- Create a CIFS only share.
- Create an NFS only share.
- Create a custom share and select multiple protocols for the same export.

Figure 2-30 shows the New Share GUI window that is used to create a CIFS only export.



* Path: /ibm/gpfs0/projectShare

* Share name: projectShare

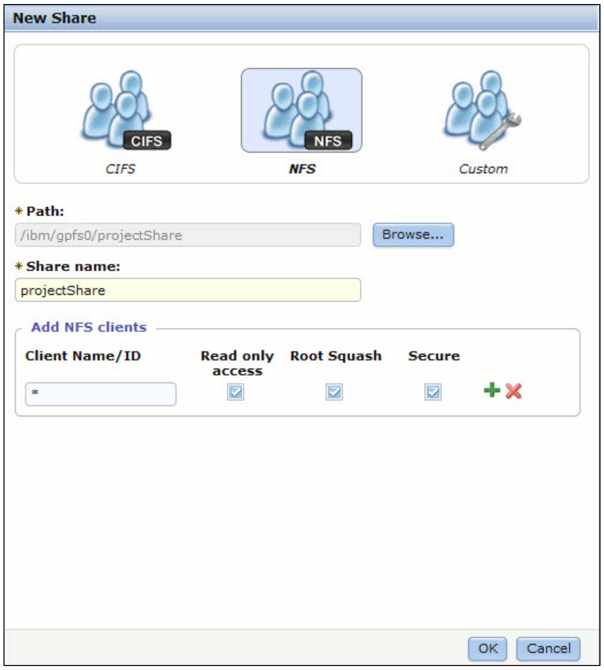
□ Read only access

☑ Browseable by file system browser

☑ Hide objects that the user does not have read access to

Figure 2-30 Create a CIFS only export with the SONAS GUI

Figure 2-31 is the New Share GUI window that is used to create an NFS export.



* Path: /ibm/gpfs0/projectShare

* Share name: projectShare

Add NFS clients

Client Name/ID	Read only access	Root Squash	Secure
*	☑	☑	☑

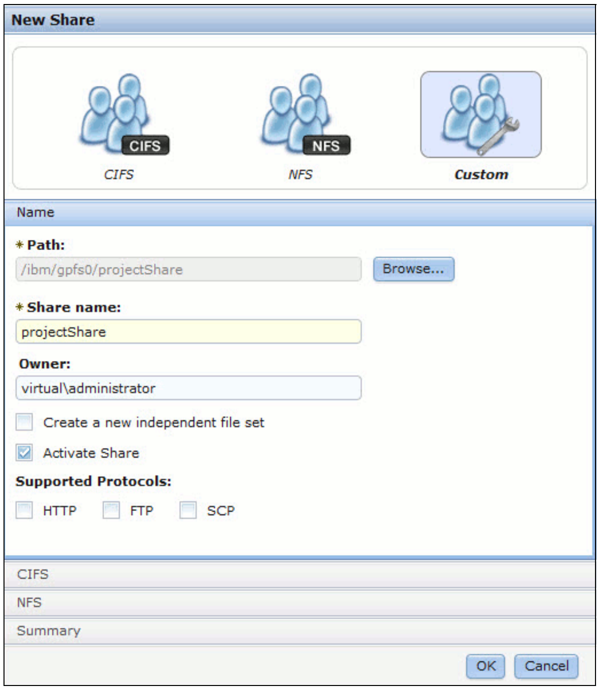
Figure 2-31 Creation of an NFS only export

During the creation of an export, it is essential to pick a directory owner. For example, it is a preferred practice to set the domain administrator as the owner of the root directory of the newly created export. That way, after the export is available, the domain administrator can connect to SONAS and set the permissions of the directory by using Windows Explorer. The domain administrator might add domain users with read-only access while adding a project group with read/write access. All of this work can be done with standard Windows tools after the export is created. The only prerequisite is to have an Owner for the share that is part of the Active Directory domain.

Within the Custom share option (see Figure 2-32), it is also possible to create an inactive share (templates) for later use. The ability to create an independent file set for the new share is important. Independent file sets allow creation of snapshots on a finer granularity. Previously, only file system level snapshots were available. Additionally, it is possible to set individual user-, group-, or file-set-level quotas when you create the independent file set. Alternatively, it is possible to select a dependent file set that was created earlier when creating an export, which ensures that there are no snapshots for the file set itself but that the export is included in the parents file system snapshot.

CIFS and NFS have detail tabs in the Custom share GUI window that allow further customization. HTTPS, FTP, and SCP do not require any configuration in the GUI.

Figure 2-32 shows the New Share window.



* Path: /ibm/gpfs0/projectShare

* Share name: projectShare

Owner: virtual\administrator

□ Create a new independent file set

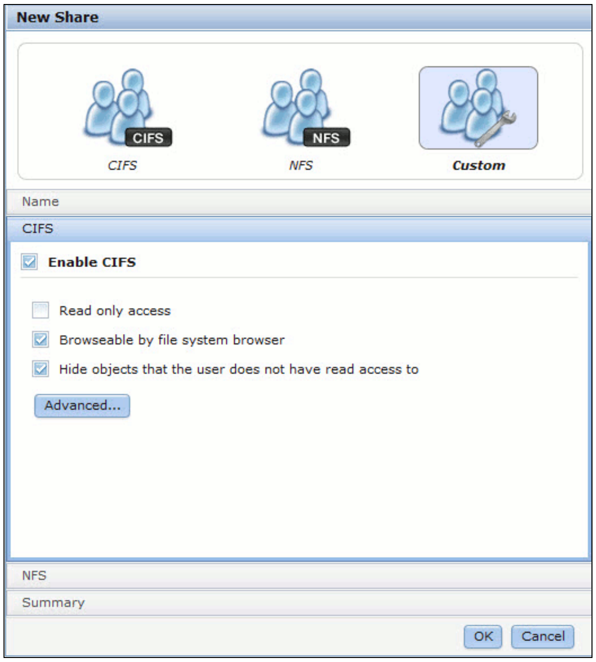
☑ Activate Share

Supported Protocols:

□ HTTP □ FTP □ SCP

Figure 2-32 Create custom shares to allow multiprotocol access to the same data

Figure 2-33 shows the Custom export CIFS information in the New Share GUI window.



☑ **Enable CIFS**

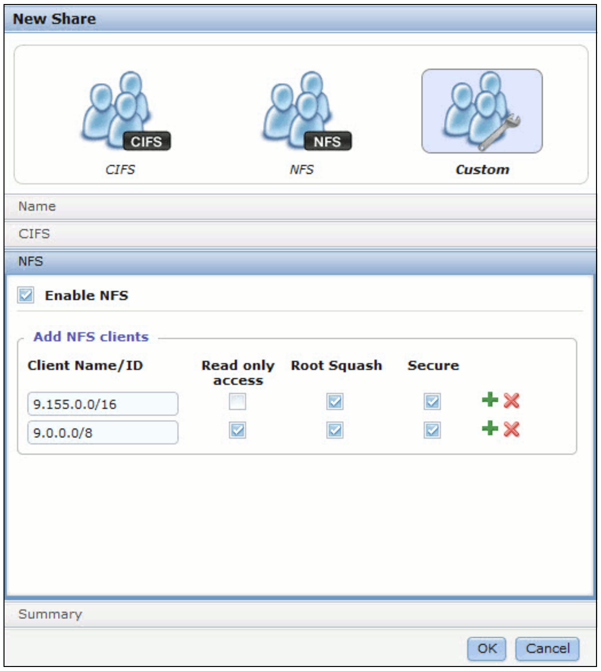
□ Read only access

☑ Browseable by file system browser

☑ Hide objects that the user does not have read access to

Figure 2-33 CIFS tab of the Custom share

Figure 2-34 shows the Custom export NFS information in the New Share GUI window.



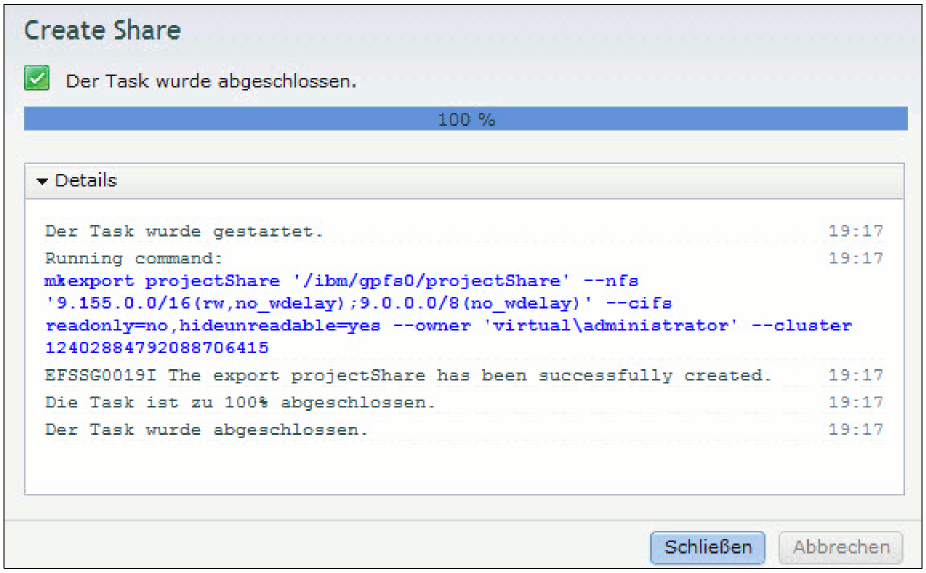
☑ **Enable NFS**

Add NFS clients

Client Name/ID	Read only access	Root Squash	Secure	
9.155.0.0/16	□	☑	☑	+ ✗
9.0.0.0/8	☑	☑	☑	+ ✗

Figure 2-34 The NFS window allows setting the NFS client names or IP ranges

Figure 2-35 shows the creation status and successful completion window of an NFS and CIFS share.



```
Der Task wurde gestartet.                                                   19:17
Running command:                                                            19:17
mkexport projectShare '/ibm/gpfs0/projectShare' --nfs
'9.155.0.0/16(rw,no_wdelay);9.0.0.0/8(no_wdelay)' --cifs
readonly=no,hideunreadable=yes --owner 'virtual\administrator' --cluster
12402884792088706415
EFSSG0019I The export projectShare has been successfully created.           19:17
Die Task ist zu 100% abgeschlossen.                                         19:17
Der Task wurde abgeschlossen.                                               19:17
```

Figure 2-35 Successfully create a share summary window

Example 2-20 shows the equivalent CLI commands to create the shares from the GUI.

Example 2-20 Create a data export by running mkexport

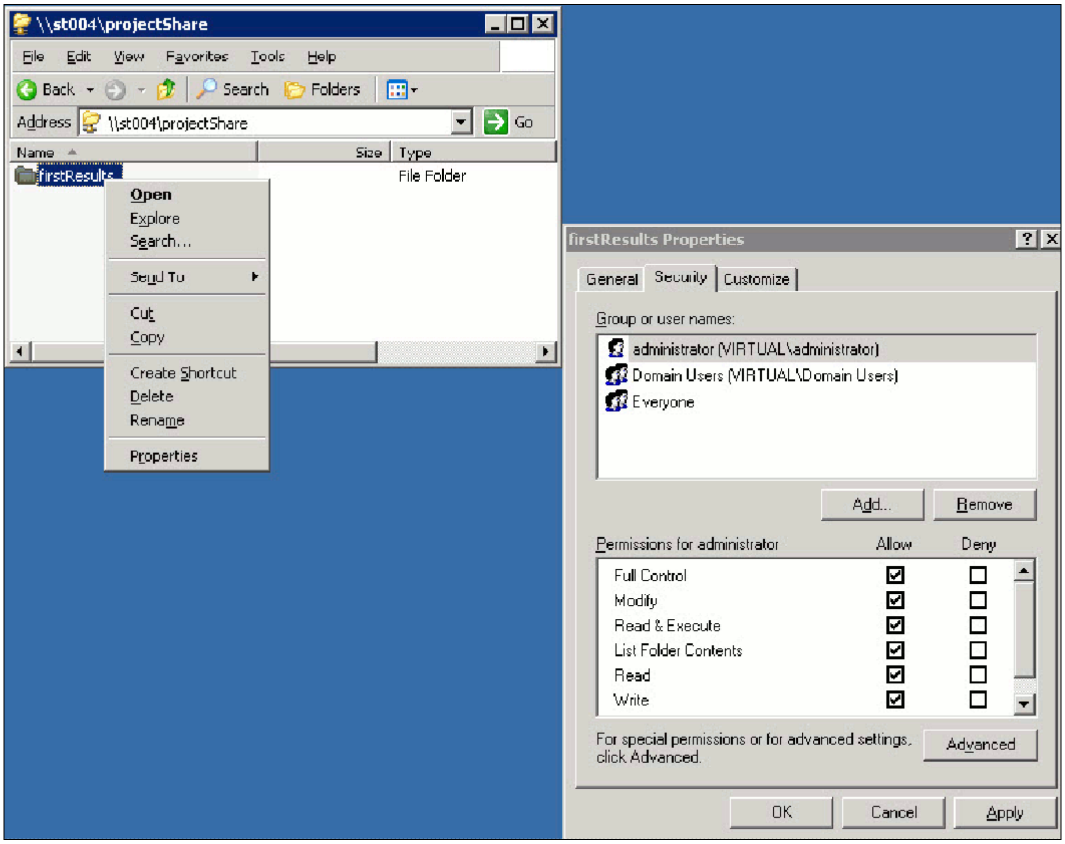
```
[SONAS]# mkexport projectShare '/ibm/gpfs0/projectShare' --nfs
'9.155.0.0/16(rw,no_wdelay);9.0.0.0/8(no_wdelay)' --cifs readonly=no,hideunreadable=yes --owner
'virtual\administrator' --cluster 12402884792088706415
[SONAS]# lsexport -v
```

2.7 Modifying access control lists to the shared export

Access control lists (ACLs) are used to specify the authority a user or group must have to access a file, directory, or file system. A user or group can be granted read-only access to files in a directory, and given full (create/write/read/execute) access to files in another directory. Only a user who was granted authorization in the ACLs can access files on the IBM SONAS appliance.

For pure NFS environments, the **chmod** and **chown** commands can be used to change permissions and access to files. For Windows environments, or for mixed environments with both NFS and CIFS clients, ACL management must be done only from Windows clients. Running **chmod** from NFS clients overwrites the extended ACLs that contain permissions for the Windows clients.

The ACLs can be set by a CIFS client when you provide a domain user as the owner during the setup of the export, as shown in Figure 2-36.



Group or user names:

- administrator (VIRTUAL\administrator)
- Domain Users (VIRTUAL\Domain Users)
- Everyone

Permissions for administrator	Allow	Deny
Full Control	☑	□
Modify	☑	□
Read & Execute	☑	□
List Folder Contents	☑	□
Read	☑	□
Write	☑	□

For special permissions or for advanced settings, click Advanced.

Figure 2-36 Verification and modification of the ACLs by MS Windows Explorer

Alternatively, with SONAS V1.5.1, the user can set ACLs through the CLI or GUI. For more information, see 7.5.6, “Managing access control lists” on page 559.

3

Installation and configuration for SONAS Gateway solutions

This chapter provides information about the basic installation and configuration of the SONAS Gateway solution.

This chapter describes the following topics:

- SONAS Gateway overview
- Pre-installation tasks
- Installation overview
- SONAS Gateway configuration with XIV storage
- SONAS Gateway configuration with Storwize V7000 storage
- SONAS Gateway configuration with DCS3700 storage
- Other considerations
- SONAS integration into your network
- Attaching SONAS to customer applications

3.1 SONAS Gateway overview

The SONAS Gateway system is a SONAS with all the features and functions of a SONAS Appliance (with Internal DDN Storage (Feature Codes 9003, 9004, and 9005), but it uses IBM external storage, such as the XIV, Storwize V7000, and DCS3700 systems. The feature codes (FCs) for XIV, Storwize V7000, and DCS3700 systems are 9006, 9007, and 9008. With all Gateway solutions, the storage is considered external and is therefore a separate line item when it is ordered.

The SONAS Gateway configurations are shipped in pre-wired racks that are made up of internal switching components along with SONAS Interface nodes (with integrated management services), and Storage nodes. The solution is prepared for integration with IBM storage solutions of various types to help you use the highest scale, performance, and flexibility from synergistically aligned storage. When combined, these capabilities offer a robust unified storage solution to solve real business problems.

SONAS Gateways can be attached to XIV, Storwize V7000, or DCS3700 storage frames of varying disk types and capacities. These additional advantages include enhanced ease of use, reliability, and performance, and they reduce the total cost of ownership (TCO).

SONAS provides the flexibility to “intermix” different storage vendors behind the same SONAS installation through RPQ. For example, a customer that has DDN and needs capacity expansion can purchase a Storwize V7000, XIV, or DCS3700 system. The different supported storage vendors must be isolated in their own dedicated storage pod because they cannot be attached to the same storage pod. There are some strategies that should be considered when intermixing storage types in a SONAS environment. These strategies are described in 3.1.1, “File system overview” on page 99.

New SONAS installations can also provide a flexible solution that has multiple storage types behind a single SONAS installation. Table 3-1 lists the possible intermix solutions.

Table 3-1 Intermix storage that is supported behind a SONAS installation

Storage type	DDN	XIV	Storwize V7000	DCS3700
DDN	Yes	RPQ	RPQ	Yes
XIV	RPQ	Yes	RPQ	RPQ
Storwize V7000	RPQ	RPQ	Yes	RPQ
DCS3700	Not applicable[a]	RPQ	RPQ	Yes

[a] You cannot start with DCS3700 storage and add DDN or DR1.

Note: As of SONAS V1.4.1, all SONAS configurations are sold as Gateways. The SONAS Appliance with internal DDN storage is no longer available for new installations. Existing Appliance configurations continue to be fully supported. DDN storage and RXB racks were discontinued on December 6, 2013. More capacity expansion for existing Appliance configurations can be on any type of supported SONAS Gateway storage (XIV, Storwize V7000, and DCS3700)

For SONAS Gateway solutions (XIV, Storwize V7000, and DCS3700), all disks, types, and capacities that are supported by the underlying storage product are supported by the SONAS Gateway solution.

3.1.1 File system overview

The SONAS file systems are built on the IBM General Parallel File System (GPFS). GPFS is a cluster file system that provides concurrent access to a single file system or file set from multiple nodes. These clustered nodes are all interconnected by a redundant InfiniBand network. The SONAS solution enables high performance access to a common set of data to support a scale-out solution and provide a high availability platform.

GPFS provides a global namespace, shared file system access among GPFS clusters, simultaneous file access from multiple nodes, high recoverability, and data availability through replication, the ability to make changes while a file system is mounted, and simplified administration even in large environments.

File system block size

One of the most important aspects of building a high-performance file system is the number of underlying LUNs (NSDs) and the file system block size. It is important to note the relationship between the number of NSDs and block size. This chapter describes this relationship because it is important to understand it when you build file systems, whether they are on XIV, Storwize V7000, or DCS3700 storage frames.

SONAS supports block sizes of 256 KB (default), 1 M, and 4 M. These sizes are defined when the file system is created and cannot be changed without re-creating the file system. This process requires you to back up the data and restore it after you re-create the file system. Therefore, carefully plan what type of data will be written and read and the I/O patterns and advanced functions because advance functions can have a huge impact on metadata.

The block size determines the minimum disk space allocation unit because GPFS divides each block into 32 subblocks. Files that are smaller than one block in size are stored in fragments, which are made up of one or more subblocks. Large files are stored in a number of full blocks plus zero or more subblocks to hold the data at the end of the file. The block size is the largest contiguous amount of disk space that is allocated to a file and therefore the largest amount of data that can be accessed in a single I/O operation. The subblock is the smallest unit of disk space that can be allocated. For a block size of 256 KB, GPFS reads as much as 256 KB of data in a single I/O operation and small files can occupy as little as 8 KB of disk space (256 KB / 32). For a block size of 1 M, small files occupy as little as 32 KB of disk space (not counting the inode), but GPFS reads or write no more than 1 M in a single I/O operation.

The block size also determines the maximum size of a read or write request that the file system sends to the underlying disk driver. From a performance perspective, set the block size to match the application buffer size, the RAID stripe size, or a multiple of the RAID stripe size. If the block size does not match the RAID stripe size, performance can be severely degraded, especially for write operations. Therefore, a block of 256 KB, 1 M, or 4 M determines how to configure the underlying RAID set on the Storwize V7000 and DCS3700 storage and the number of disks. The XIV storage is fixed at 1 M and cannot be changed. Therefore, use a 1 M file system block size for XIV storage.

In file systems with a mix of variance in file sizes, a small block size has a large effect on performance when you access large files. In this kind of system, use a block size of 256 KB (8 KB subblock). Even if only 1% of the files are large, the amount of space that is taken by the large files usually dominates the amount of space that is used on disk. The waste in the subblock that is used for small files is insignificant. Larger block sizes up to 1 MB are often a good choice when the performance of large files that are accessed sequentially are the dominant workload for this file system.

The effect of block size on file system performance largely depends on the application I/O pattern. A larger block size is often beneficial for large sequential read and write workloads. A smaller block size is likely to offer better performance for small file, small random read and write, and metadata-intensive workloads. The efficiency of many algorithms that rely on caching file data in a page pool depends more on the number of blocks that are cached than the absolute amount of data. For a page pool, a larger file system block size means that fewer blocks are cached. Therefore, when you create file systems with a block size larger than the default value of 256 KB, it is preferable that you increase the page pool size in proportion to the block size. Data is cached in the Interface node memory, so it is important to correctly plan the RAM memory size in Interface nodes.

Number of LUNs (NSDs)

The number of disks that are allocated to any file system should be an even number that is divisible by 4, such as 4, 8, 12, 16, ... This guideline is because of the number of I/O paths between the Storage node pair or pod and the storage platform. This setting ensures that disks (and LUNs) are spread evenly across Storage nodes, Storage node HBA ports, and storage platform controllers. It is better to have a minimum of eight LUNs for each file system to maximize the I/O spread and to increase effective queue depth (I/O spread across more LUNs, hence more queues). Eight also divides well into all of the subblock sizes.

Intermixing storage in file system considerations

Intermixing storage types in a single SONAS installation provides a high level of flexibility to fit the needs of the business. Consider the following three approaches for intermixed storage:

- Separate GPFS file systems for each storage vendor
  - Advantage: Complete separation from a performance and availability perspective
  - Disadvantages
    - Requires creation and management of separate GPFS file systems and NAS exports.
    - Storage space in each storage system is not shared within a single file system.
- Separate file system pools for each storage vendor type in the same file system
  - Advantages
    - Ability to use existing GPFS file systems and NAS exports.
    - Separation of storage in different disk storage systems to different file system pools
    - File placement policies can be used to direct specific files to specific file system pools.
    - ILM policies can be used to perform file system pool migration to migrate files from DDN file system pools to another storage vendor file system pools.
    - This approach provides for the eventual migration of files and directories from DDN storage to another vendor storage without having to copy the data from one file system to another file system and redirect NAS clients to different NAS exports.
  - Disadvantages
    - Storage space in the different file system pools must be monitored and managed separately.
    - Requires a new file placement policy to steer specific files to specific storage pools, migration policies to move files between storage pools, or both.

- Different storage vendor storage intermixed in the same GPFS file system pools in the same file system
  - Advantages
    - The storage space in the file system pool can be managed as a single entity even though it contains both DDN storage and DCS3700 storage.
    - This approach provides for the eventual migration of files and directories from DDN storage to DCS3700 storage without having to copy the data from one file system to another file system and redirect NAS clients to different NAS exports.
  - Disadvantages
    - No separation of storage from a performance and availability perspective.
    - An individual file or directory can be spread across both the DDN storage system and DCS3700 storage.
    - If the same file system and the same pool are used, it is best to match NSD size, RAID configuration, and drive capacity and speed.

3.2 Pre-installation tasks

After your SONAS purchase is completed and delivered to you, complete the following tasks to integrate your SONAS appliance into your existing environment:

1. Review the floor plan and pre-installation planning sheet to determine whether all the information is provided.
2. If the pre-installation planning sheet is not complete, contact the Storage Administrator. This information is required for the rest of the installation, and the installation cannot start until the pre-installation planning sheet is complete.
3. The IBM authorized service provider does all of the necessary preliminary planning work. This work includes verifying the information in the planning worksheets to ensure that you are aware of the specific requirements, such as the physical environment or the networking environment for the SONAS system. For more information about planning, see Chapter 1, “Installation planning” on page 1, and the SONAS planning information in the IBM Knowledge Center, found at the the following website:

   http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/planning.html

3.2.1 SONAS rack

This section describes how to install the SONAS rack.

Rack overview

The SONAS base rack is an RXA. It consists of the following minimum components:

- Two Interface nodes
- Two Storage nodes
- Two InfiniBand switches
- Two Ethernet switches
- Keyboard, video, and mouse console (KVM)

Figure 3-1 shows the rack layout.

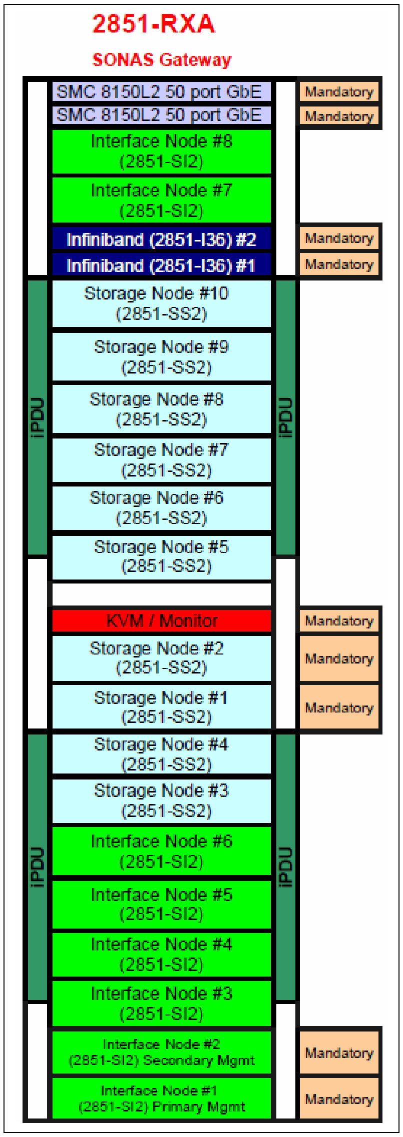

2851-RXA
SONAS Gateway
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Interface Node #8 (2851-SI2)
Interface Node #7 (2851-SI2)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Storage Node #10 (2851-SS2)
Storage Node #9 (2851-SS2)
Storage Node #8 (2851-SS2)
Storage Node #7 (2851-SS2)
Storage Node #6 (2851-SS2)
Storage Node #5 (2851-SS2)
KVM / Monitor
Storage Node #2 (2851-SS2)
Storage Node #1 (2851-SS2)
Storage Node #4 (2851-SS2)
Storage Node #3 (2851-SS2)
Interface Node #6 (2851-SI2)
Interface Node #5 (2851-SI2)
Interface Node #4 (2851-SI2)
Interface Node #3 (2851-SI2)
Interface Node #2 (2851-SI2) Secondary Mgmt
Interface Node #1 (2851-SI2) Primary Mgmt
iPDU
iPDU
iPDU
iPDU
Mandatory
Mandatory
Mandatory
Mandatory
Mandatory
Mandatory
Mandatory
Mandatory
Mandatory

Figure 3-1 SONAS Gateway RXA base rack

Rack cabling

One of the first tasks is to *verify the internal SONAS rack cabling (point-to-point)*. A color-coded cabling guide and label kit is included in all SONAS rack shipments. Ensure that each cable is correctly labeled at both ends and verify each connection for location and firm connection before you power on the SONAS frame. An example of the cabling guide is shown in Figure 3-2 on page 103.

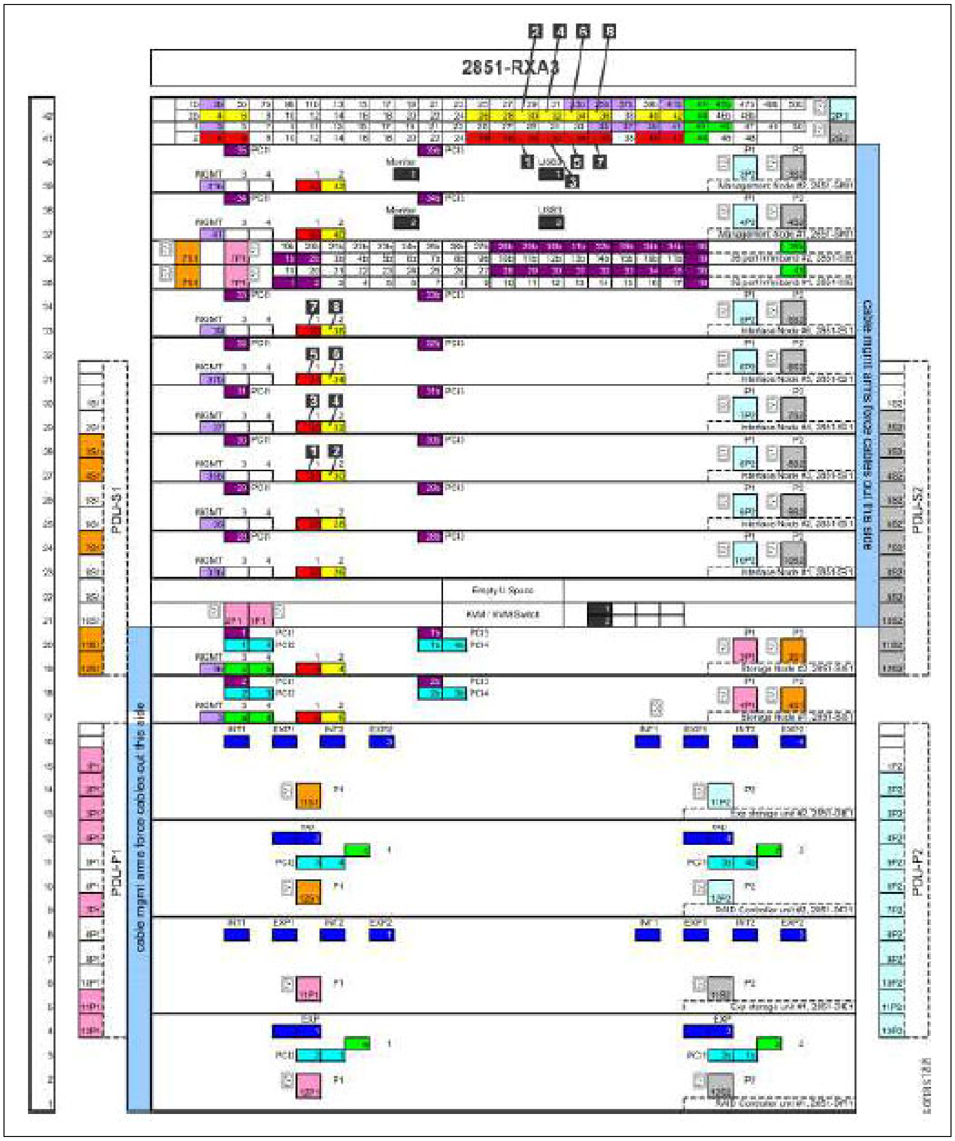

2851-RXA3
Cable mgmt arms force cables out this side
cable mgmt arms force cables out this side
PDU-S1
PDU-S2
PDU-P1
PDU-P2
Empty U Space
KVM / KVM Switch

Figure 3-2 A page of the color-coded cable guide from the SONAS installation guide

Whether you are installing a SONAS Gateway with XIV, DCS3700, Storwize V7000, or DS8000 series (RPQ Only) storage, the storage subsystem type does not influence the SONAS Gateway frame configuration. It does not matter which type of storage you plan to use in your Gateway because the storage type does not affect the positioning or configuration of your SONAS frame hardware. A Gateway of any type requires a 2851-RXA configuration begin. To expand to more Interface or Storage nodes to support the Gateway expansion, the only option is to purchase a 2851-RXC Interface node expansion frame, which also allows for placement of Storage node pairs for Gateway expansion applications.

Connecting SONAS rack-mounted KVM connections

As a minimum, the primary Management node and secondary Management node must be connected to the SONAS rack-installed KVM.

For more information about this process, see *IBM SONAS Installation Guide,* GA32-0715. It is set up by the IBM authorized installer at the time of the initial installation.

All SONAS hardware is tested in manufacturing and shipped with the *Manufacturing Cleanup* process (all systems are correctly set to pre-installation defaults). This process reduces the count on failed components or cabling issues in the field. However, by powering on the Management node and running **`get_version`** from the CLI, you can view the version of software that is installed on the Management node by manufacturing.

3.2.2 Validating SONAS server PCI card installations

In some cases, specific orders might require subtle changes to the base default configuration. It is important to ensure that these options are installed in the proper PCI slots, and that proper cable seating is confirmed, before you run the `first_time_install` script or commit to a configuration. If these cards are incorrectly installed, the SONAS configuration is likely to fail. However, to ensure success, reliability, and ongoing support, it is an important preparatory installation check.

For example, check the following things:

- Number of memory expansion modules
- Type and quantity of Ethernet adapters
- All cable connections (Ethernet, InfiniBand, power)
- All cable labels
- Interface nodes
- Storage nodes

3.3 Installation overview

This section gives an overview of the installation process. Most of the installation process is independent of the back-end storage, whether it is XIV, Storwize V7000, or DCS3700 storage. Later sections show each storage platform and its integration in to the SONAS Gateway.

IBM Scale Out Network Attached Storage - Installation Guide for IBM SONAS Gateway: Attaching IBM SONAS to IBM XIV Storage System, IBM Storwize V7000, or IBM DCS3700 or IBM SONAS Storage 2, GA32-2223 includes all the details for a complete installation. It supplements the information in this chapter and provides more background and explanation. You can download it from the IBM Knowledge Center at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_r_publications.html?lang=en

> **Note:** SONAS V1.5.1 includes the following first-time installation enhancements:
>
> - Options to specify the VLAN and external management adapter during installation to reduce requirements to run the **`chnwmgt`** command
> - Enhanced feedback and polling during the installation process to speed the process for nodes to check in and present additional information
> - Network connectivity to the Management node is permitted, which can speed the installation process

This list provides a simple overview of the SONAS Gateway installation process. The remainder of this chapter provides a detailed installation process.

1. Verify power, cooling, access, and installation readiness.
2. Position the SONAS rack.

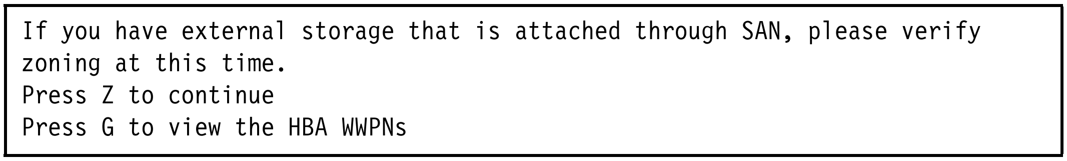
3. Ensure that external Fibre Channel fabric switches are safely rack-mounted to external customer racks.
4. Verify the internal SONAS rack cabling (point-to-point) by using the included color cabling charts.
5. Connect the SONAS rack-mounted KVM connections.
6. If the Management node is not at the correct code level, ensure that the SONAS rack was prepared with the manufacturing cleanup process with IBM Service validation (all systems are properly unconfigured and ready for reinstallation).
7. Validate that the SONAS Server PCI card installations, including 10 GbE, are in the appropriate slots and that all cable labels are accurately placed on both ends of all cables.
8. Capture all frame, slot, and serial number placement identification information.
9. Write down the WWPNs from each Storage node and use the list for switch zoning (XIV and Storwize V7000), direct-attached DCS3700 storage, host LUN allocation, and back-end Storage node configuration. The `first_time_install` script prompts you when the Storage nodes are powered on and started far enough to be able to query and complete this work.
10. Ensure that the storage system (XIV, Storwize V7000, or DCS3700) is properly configured for provisioning the SONAS cluster, nodes, and node WWPN ports. Make sure that you verified the WWPNs.
11. Ensure that the storage system has a correct “SONAS” Storage Pool (thick and regular pool) and that correctly sized volumes for the storage platform are configured and mapped to the SONAS cluster (two hosts in a cluster that do not map directly to each host).
12. Insert the latest GA or currently recommended SONAS Release DVD into the Management node (while other nodes are shut down), and restart the node to install the operating system and SONAS software on the Management node. For SONAS V1.3 and later releases, the Primary Management node service is run on Interface node 1 (int001st001), which is the bottom Interface node in the frame.

    For earlier releases, or when you have an independent Management node, install code from the Management node in slot 37.
13. While the Management node is loading with the code level that is required, verify and configure the SAN zoning (XIV and Storwize V7000) between the storage system and the Storage nodes. For DCS3700 storage, verify the proper Fibre Channel direct connectivity between the DCS3700 and the Storage nodes. Also, verify from the Storage platform that all required LUNs are created and mapped to the Storage nodes.
14. From the Management node, before you run the `first_time_install` script, run the following command:

    ```
    /opt/IBM/sonas/bin/mfg/cfg_gateway_rpq
    ```

    Running this pre-installation script sets a flag for the installation, which tells the installation process to configure the SONAS cluster without configuring internal (that is, DDN) storage. It allows the storage configuration process to remain a separate, manual configuration process.
15. Run the `first_time_install` script and follow the prompts while using the Installation Planning Sheet, the *IBM SONAS Installation Guide*, GA32-0715, and *IBM SONAS Configuration Guide*, GA32-0718 to create the initial cluster.
16. When the SONAS `first_time_install` script prompts you to do so, power on and discover all the Interface and Storage nodes (confirm the node instance ID, rack position, frame number, and quorum device specification).

17. When the `first_time_install` prompts you to do so, zone Storage node WWPNs to external storage ports if necessary. Add WWPNs to external storage host groups and logical drive mappings. The user prompt for zoning looks like Figure 3-3.



```
If you have external storage that is attached through SAN, please verify
zoning at this time.
Press Z to continue
Press G to view the HBA WWPNs
```

Figure 3-3 User prompt for zoning

18. Enable licensing (for all nodes).

19. List and verify the connected LUNs and disks for use by GPFS.

20. If the expected devices are not present, or storage LUNs and disks were not properly allocated to the Storage nodes before running the `first_time_install` script, complete one of the following actions:

    – For SONAS V1.4.1 and later, you can run **`mkdisk --luns`** to rescan all Storage nodes for new storage. This command can be run as the admin or root user.

    – An alternative method of rescanning disks (for all SONAS versions) is to run the following command:

    ```
    /opt/IBM/sonas/bin/cnaddstorage
    ```

    When it is run, there is an option for which Storage node pair to scan. Each Storage node pair (pod) must be scanned. If multiple storage pods exist, repeat this process on all of the attached storage pods.

21. If multiple alike platforms (XIV, Storwize V7000, or DCS3700) exist in each storage pod, create the second failure group and use all the NSDs from the second frame in failure group 2. Each storage platform provides LUN protection and reliability through some form of RAID. However, metadata replication is a preferred practice because the metadata is critical to the survival of the file system. You can also replicate the data, but it might not be preferred because of performance and economic (it doubles the space requirements and price) reasons.

    Make sure that the failure groups use a balanced number of disks that have balanced second Storage node preferences by running **`mmlsnsd –l`** to ensure a load-balanced configuration.

22. Configure the cluster.

23. Install the Network Bonding, NAT Gateway, Network Group, DNS, AD, LDAP, and Round Robin configurations for access and authentication.

24. Create the primary “file system”, dependent and independent file set, and initial exports. Before you create any file system, consider the block size and block allocation map type. Because these values are defined when the file system is created, they cannot be changed without deleting and re-creating the file system. This process requires you to back up and restore the data.

25. Disk Pool, Usage Type, and Failure Group. These parameters are assigned to the disks and NSDs and must be defined before the disks are associated with a file system. A workaround is to remove a disk from the file system, change the parameters, and readd it back into the file system. This process assumes that there is enough space on the remaining disks for the used capacity

26. Test DNS and NTP access to the cluster.

27. Test authentication.

28. Set up and test the Call Home service.

29. Test access to all the required protocols.

30. Verify and validate all pertinent log files.

31. Verify status queries and the GUI status.

32. Verify and validate associated event logs.

33. Clear miscellaneous initial installation errors and information warnings.

Depending on your storage type, continue with one of the following sections:

- 3.4, “SONAS Gateway configuration with XIV storage” on page 107
- 3.5, “SONAS Gateway configuration with Storwize V7000 storage” on page 138
- 3.6, “SONAS Gateway configuration with DCS3700 storage” on page 174

3.4 SONAS Gateway configuration with XIV storage

The SONAS Gateway configuration with XIV (Gen2 or Gen3) is listed as a feature code option when you order the SONAS gateway. XIV gateways are FC 9006. Much like the Storwize V7000 (FC 9007), SONAS with XIV is configured as a multi-system, grid-based storage solution that offers customers great flexibility in NAS with an IBM solution. SONAS is built from hardware and software components. They offer moderate entry points with true vertical and horizontal modular scaling of capacity and performance to meet the requirements of the most demanding of customer use cases.

Figure 3-4 shows a SONAS configuration with an XIV Gateway.

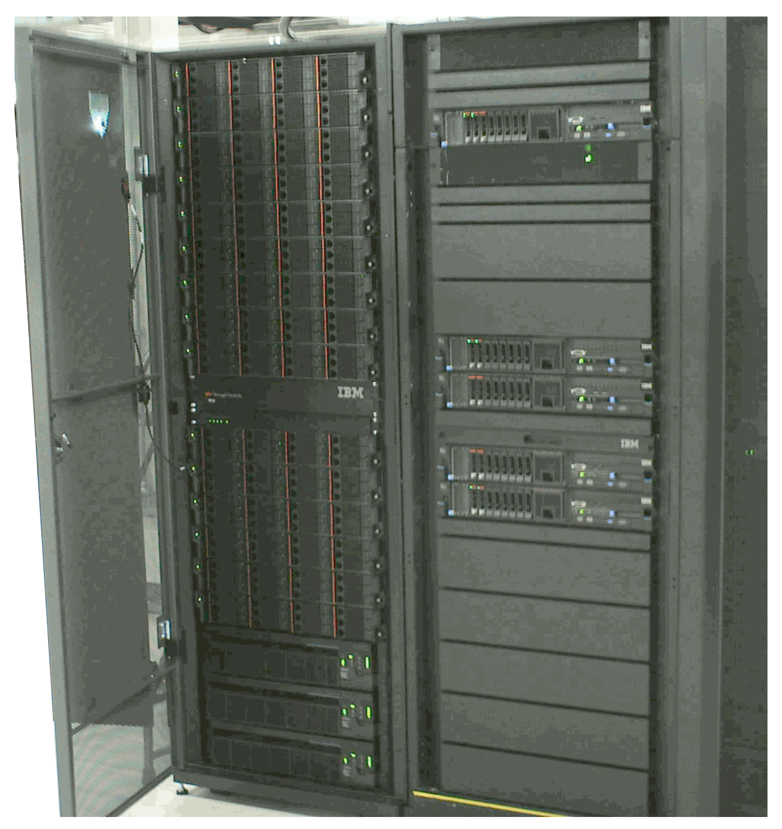


Figure 3-4 Small configuration of a SONAS Gateway with one XIV

3.4.1 Overview of XIV storage

XIV storage consists of frames of disks that are grouped in sets of 12 x 1 TB, 2 TB, 3 TB, or 4 TB SAS drives (Gen3), or 1 TB or 2 TB SATA hard disks (XIV Gen2) per module, with 6 - 15 modules. XIV storage can be purchased for use with SONAS in full 15 Module Frames or in frames of 6, 9, 10, 11, 12, 13, or 14 modules (called Partial Populated XIV Frames). The smallest supported XIV population is nine modules to support a production environment. Six module XIVs are suitable for proof of concepts (POCs) or demonstration purposes only.

Note: All hard disk drives (HDD) types and sizes that are supported by the XIV are supported when attached to an IBM SONAS Gateway system.

The partially populated XIV frames can also be purchased with the Capacity on Demand (CoD) options. You can use the CoD option to pre-populate more modules of disks in the frame for anticipated growth (you can pay for expanded capacity when you use it). You can also use the CoD options to use more processor and memory from the installed modules while not yet using the available capacity. It is often a popular choice for rapid-growth environments, or for when the effect of snapshot management and retention plans on cluster capacity is not known.

XIV storage is an easy to learn storage platform with tier 1 reliability that is based on special use (by design) of tier 2 components. It provides true tier 1 performance at tier 2 pricing.

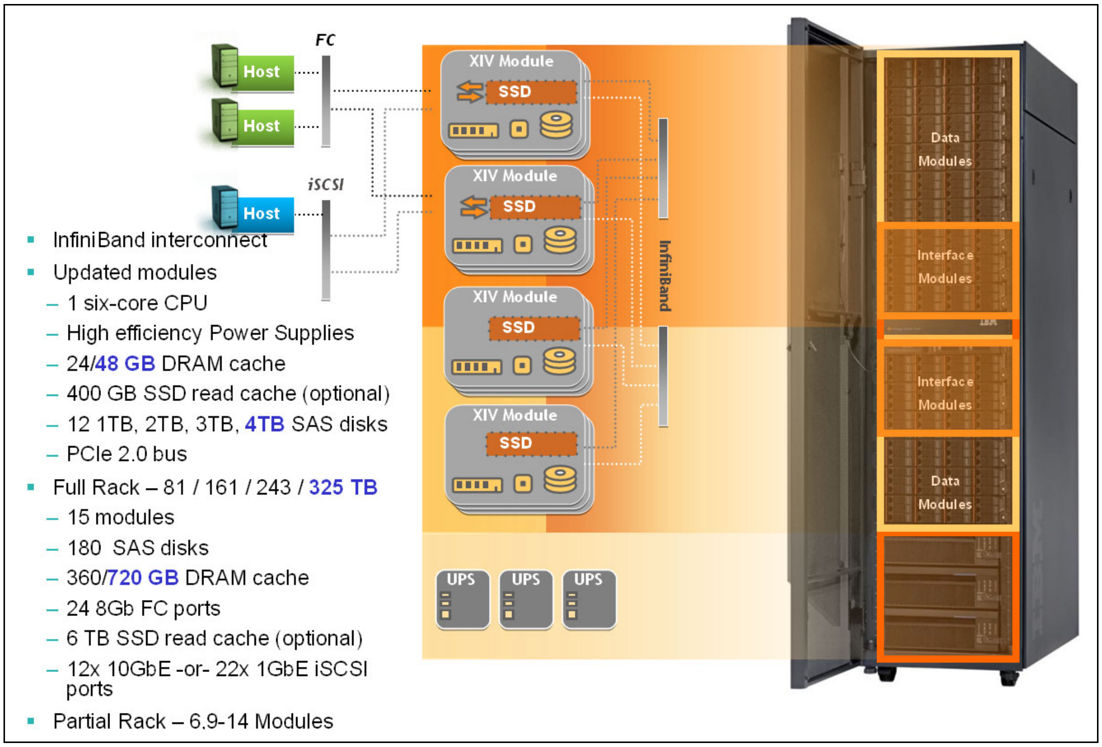
The Gateway solution allows for both block and file client host provisioning capabilities by SONAS or directly to the storage through the fabric.

Here is a simplified explanation of XIV storage. XIV storage comes preconfigured in a data chunk (1 MB) mirrored data format of a proprietary RAID strategy, which unique to XIV. It has a large distributed cache and virtualized hot spare capacity, which allow every spindle in the frame to share every workload across every module and every spindle. This function helps keep an equilibrium of activity balanced across all the resources in the frame.

XIV storage and its grid-based design are optimally suited for SONAS and the SONAS grid-based cluster design. Together, they offer the highest reliability and sustainable performance of all of the SONAS solutions. It is easy to install and, if the layout is balanced, it never allows a hot spot in the SONAS cluster storage, even during component failures.

It also offers the rapid rebuilding for failed drives and disk modules for a frame-based storage environment, largely because of its virtualized hot spare capacity and full array stripe design.

Figure 3-5 shows XIV Gen3 Model 214 component details.



- InfiniBand interconnect
- Updated modules
  - 1 six-core CPU
  - High efficiency Power Supplies
  - 24/**48 GB** DRAM cache
  - 400 GB SSD read cache (optional)
  - 12 1TB, 2TB, 3TB, **4TB** SAS disks
  - PCIe 2.0 bus
- Full Rack – 81 / 161 / 243 / **325 TB**
  - 15 modules
  - 180 SAS disks
  - 360/**720 GB** DRAM cache
  - 24 8Gb FC ports
  - 6 TB SSD read cache (optional)
  - 12x 10GbE -or- 22x 1GbE iSCSI ports
- Partial Rack – 6,9-14 Modules

Figure 3-5 XIV Gen3 Model 214 details

Although XIV Gen2 is no longer available for sale, it will continue to be supported until 2019. XIV Gen2 and XIV Gen3 are both supported in the SONAS Gateway configuration. However, when both types are used, they must be installed in separate storage pods for optimal use.

One or two XIV subsystems can be used behind each SONAS storage pod.

3.4.2 SONAS with XIV: Maximum optimal capacity solution

This configuration is the maximum configuration that is available with the SONAS XIV Gateway solution. It consists of a combination of Interface and Storage nodes for a total of 34 nodes. The following configuration provides 6.5 PB of usable SONAS storage:

- One SONAS Base Rack (2851-RXA)
  - Eight Interface nodes
  - Ten Storage nodes
- One Interface node expansion frame (2851-RXC)
  - Six Interface nodes
  - Ten Storage nodes
- Twenty 15-module XIV frames (Gen3) with 4 TB drives (325 TB). Total usable solution capacity = 6.5 PB.
- Three-hundred and twenty customer-provided Fibre Channel ports (12 ports per XIV (total of 240 XIV ports), four ports per SONAS Storage node (total 80 Storage node ports)).
  - One-hundred sixty ports for Fabric-A
  - One-hundred sixty ports for Fabric-B

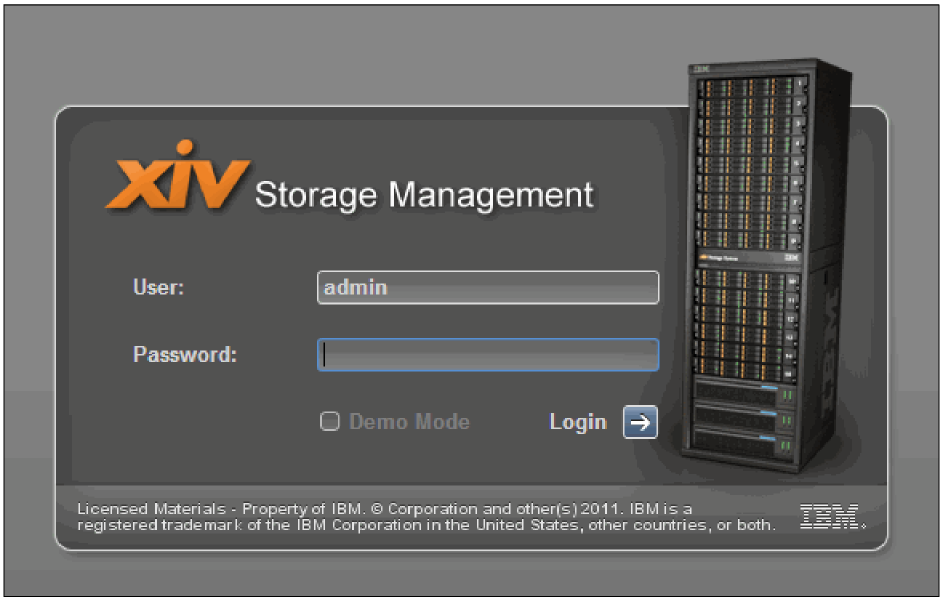
3.4.3 Detailed configuration and installation information for SONAS with XIV

This section describes the XIV Gateway installation tasks.

Ensuring that the XIV storage is correctly configured

The next installation step is to ensure that the XIV storage is properly configured (power, cooling, Ethernet, and Call Home). This process is typically done by an IBM authorized installer.

Figure 3-6 shows the XIV GUI login window.



User: admin

Password:

Licensed Materials - Property of IBM. © Corporation and other(s) 2011. IBM is a registered trademark of the IBM Corporation in the United States, other countries, or both.

Figure 3-6 XIV GUI - login

Figure 3-7 shows the GUI view of multiple managed XIV frames.

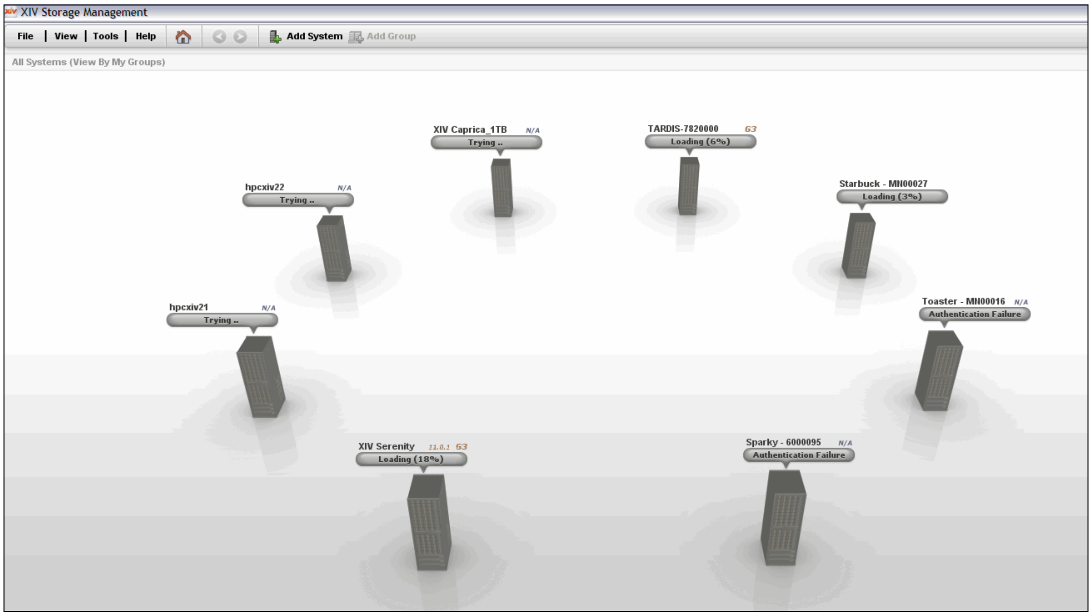


Figure 3-7 XIV GUI - view of multiple managed XIV frames from one GUI

Creating proper zoning to a SONAS Storage node from XIV

Now is a good time to prepare the XIV configuration. To begin, you must get the WWPNs from the stickers on the back of the SONAS Storage nodes. To confirm the WWPNs, run the following command from the Management node after the cluster is configured with the `first_time_install` script:

```
/opt/IBM/sonas/bin/cn_get_wwpns
```

Two ports from node 1 (one from each of the 8 Gb FC PCI cards) are zoned to FC switch 1, and the remaining ports from each card are connected to FC Switch 2 (see Figure 3-8).

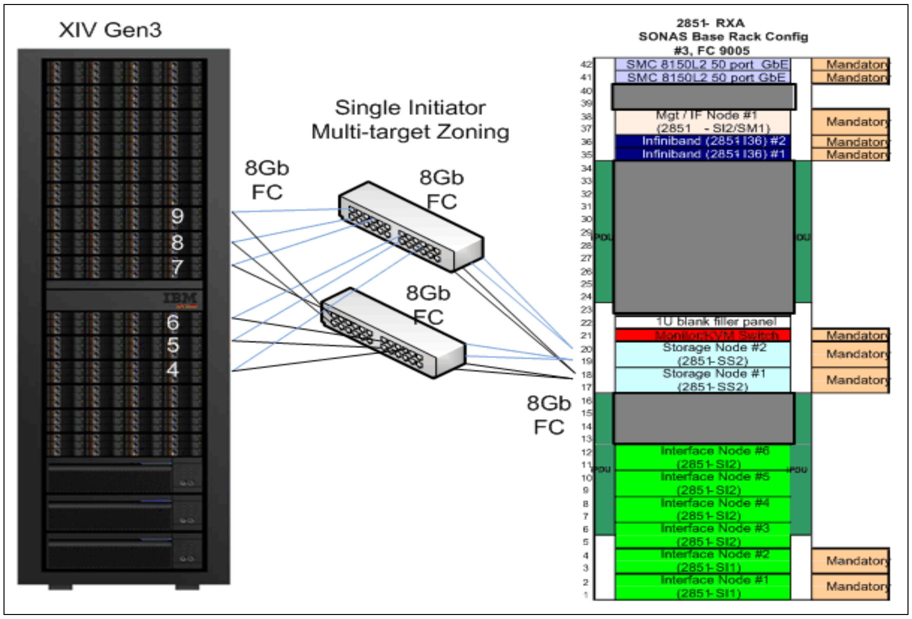
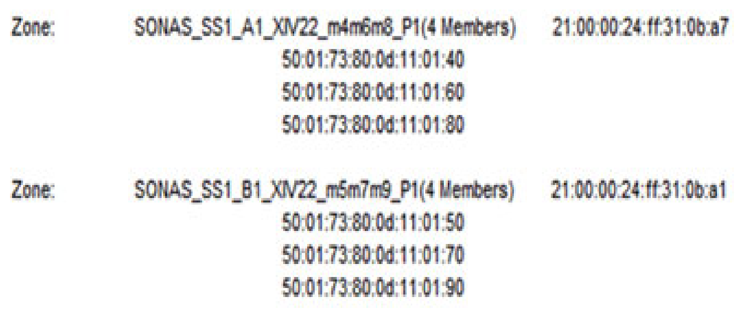


Figure 3-8 XIV to SONAS Storage node port zoning (port 1 and 3 from all XIV modules)

If you are connecting a single XIV frame, ensure that one port from each XIV Interface Module is connected to each of the two FC switches (in a full-frame XIV that is six ports to switch 1 and six ports to switch 2). Make sure that they are all set as targets on the XIV side. Connect XIV port-1 of each module to switch 1 and XIV port-3 of each module to switch 2. If you are zoning port-4 on the XIV subsystem, it must be changed from default “initiator” port to a “target” port.

On the SAN switch, single-initiator zoning is required. Each host port is zoned to three XIV Interface Module ports. Host port 1 is zoned to even-numbered XIV modules (4, 6, and 8), and Host port 2 is zoned to odd-numbered XIV modules (5 and 7), as shown in Figure 3-9.



```
Zone:     SONAS_SS1_A1_XIV22_m4m6m8_P1(4 Members)     21:00:00:24:ff:31:0b:a7
                              50:01:73:80:0d:11:01:40
                              50:01:73:80:0d:11:01:60
                              50:01:73:80:0d:11:01:80

Zone:     SONAS_SS1_B1_XIV22_m5m7m9_P1(4 Members)     21:00:00:24:ff:31:0b:a1
                              50:01:73:80:0d:11:01:50
                              50:01:73:80:0d:11:01:70
                              50:01:73:80:0d:11:01:90
```

Figure 3-9 SONAS to XIV zone sample - single initiator, multi-target

For more information about connectivity and zoning requirements, see the SONAS V1.5.1 information in the IBM Knowledge Center, found at the followng website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_svc_hardwareinstallationguide05.html

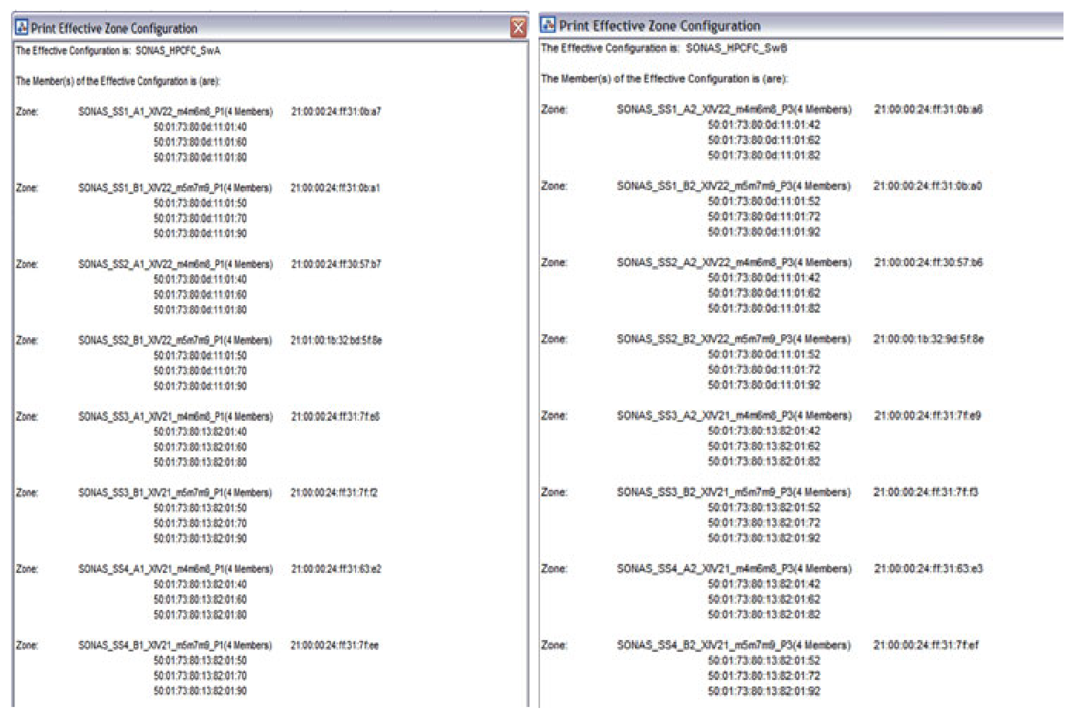
Also, see *IBM Scale Out Network Attached Storage - Installation Guide for IBM SONAS Gateway: Attaching IBM SONAS to IBM XIV, IBM Storwize V7000, or IBM DCS3700 or IBM SONAS Storage 2*, GA32-2223 at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_r_publications.html?lang=en

For simplicity, pertinent information is also provided here. A preferred practice for XIV solutions is to provide 12 paths per LUN for a full XIV frame configuration.

Single-initiator, multi-target zoning is a preferred practice. Each host port zones to three XIV module target ports. WWN and port zoning are both supported. If port zoning is used, field-based HBA replacements (fix on fail) do not require zoning changes at the switches (both methods require meticulous validation on zoning, port settings, and cable labels). Either zoning method is supported.

Spread your host ports across the switch ASICs as evenly as possible, and follow switch manufacturer guidelines for Red Hat cluster type clients. Figure 3-10 shows a zoning configuration of a single SONAS initiator to three XIV targets with 12 paths per LUN per Storage node.


Print Effective Zone Configuration

```
The Effective Configuration is:  SONAS_HPCFC_SwA

The Member(s) of the Effective Configuration is (are):

Zone:        SONAS_SS1_A1_XIV22_m4m6m8_P1(4 Members)     21:00:00:24:ff:31:0b:a7
                      50:01:73:80:0d:11:01:40
                      50:01:73:80:0d:11:01:60
                      50:01:73:80:0d:11:01:80

Zone:        SONAS_SS1_B1_XIV22_m5m7m9_P1(4 Members)     21:00:00:24:ff:31:0b:a1
                      50:01:73:80:0d:11:01:50
                      50:01:73:80:0d:11:01:70
                      50:01:73:80:0d:11:01:90

Zone:        SONAS_SS2_A1_XIV22_m4m6m8_P1(4 Members)     21:00:00:24:ff:30:57:b7
                      50:01:73:80:0d:11:01:40
                      50:01:73:80:0d:11:01:60
                      50:01:73:80:0d:11:01:80

Zone:        SONAS_SS2_B1_XIV22_m5m7m9_P1(4 Members)     21:01:00:1b:32:bd:5f:8e
                      50:01:73:80:0d:11:01:50
                      50:01:73:80:0d:11:01:70
                      50:01:73:80:0d:11:01:90

Zone:        SONAS_SS3_A1_XIV21_m4m6m8_P1(4 Members)     21:00:00:24:ff:31:7f:e8
                      50:01:73:80:13:82:01:40
                      50:01:73:80:13:82:01:60
                      50:01:73:80:13:82:01:80

Zone:        SONAS_SS3_B1_XIV21_m5m7m9_P1(4 Members)     21:00:00:24:ff:31:7f:f2
                      50:01:73:80:13:82:01:50
                      50:01:73:80:13:82:01:70
                      50:01:73:80:13:82:01:90

Zone:        SONAS_SS4_A1_XIV21_m4m6m8_P1(4 Members)     21:00:00:24:ff:31:63:e2
                      50:01:73:80:13:82:01:40
                      50:01:73:80:13:82:01:60
                      50:01:73:80:13:82:01:80

Zone:        SONAS_SS4_B1_XIV21_m5m7m9_P1(4 Members)     21:00:00:24:ff:31:7f:ee
                      50:01:73:80:13:82:01:50
                      50:01:73:80:13:82:01:70
                      50:01:73:80:13:82:01:90
```

Print Effective Zone Configuration

```
The Effective Configuration is:  SONAS_HPCFC_SwB

The Member(s) of the Effective Configuration is (are):

Zone:        SONAS_SS1_A2_XIV22_m4m6m8_P3(4 Members)     21:00:00:24:ff:31:0b:a6
                      50:01:73:80:0d:11:01:42
                      50:01:73:80:0d:11:01:62
                      50:01:73:80:0d:11:01:82

Zone:        SONAS_SS1_B2_XIV22_m5m7m9_P3(4 Members)     21:00:00:24:ff:31:0b:a0
                      50:01:73:80:0d:11:01:52
                      50:01:73:80:0d:11:01:72
                      50:01:73:80:0d:11:01:92

Zone:        SONAS_SS2_A2_XIV22_m4m6m8_P3(4 Members)     21:00:00:24:ff:30:57:b6
                      50:01:73:80:0d:11:01:42
                      50:01:73:80:0d:11:01:62
                      50:01:73:80:0d:11:01:82

Zone:        SONAS_SS2_B2_XIV22_m5m7m9_P3(4 Members)     21:00:00:1b:32:9d:5f:8e
                      50:01:73:80:0d:11:01:52
                      50:01:73:80:0d:11:01:72
                      50:01:73:80:0d:11:01:92

Zone:        SONAS_SS3_A2_XIV21_m4m6m8_P3(4 Members)     21:00:00:24:ff:31:7f:e9
                      50:01:73:80:13:82:01:42
                      50:01:73:80:13:82:01:62
                      50:01:73:80:13:82:01:82

Zone:        SONAS_SS3_B2_XIV21_m5m7m9_P3(4 Members)     21:00:00:24:ff:31:7f:f3
                      50:01:73:80:13:82:01:52
                      50:01:73:80:13:82:01:72
                      50:01:73:80:13:82:01:92

Zone:        SONAS_SS4_A2_XIV21_m4m6m8_P3(4 Members)     21:00:00:24:ff:31:63:e3
                      50:01:73:80:13:82:01:42
                      50:01:73:80:13:82:01:62
                      50:01:73:80:13:82:01:82

Zone:        SONAS_SS4_B2_XIV21_m5m7m9_P3(4 Members)     21:00:00:24:ff:31:7f:ef
                      50:01:73:80:13:82:01:52
                      50:01:73:80:13:82:01:72
                      50:01:73:80:13:82:01:92
```

Figure 3-10   Brocade zoning - a single SONAS initiator to three XIV targets (12 paths per LUN)

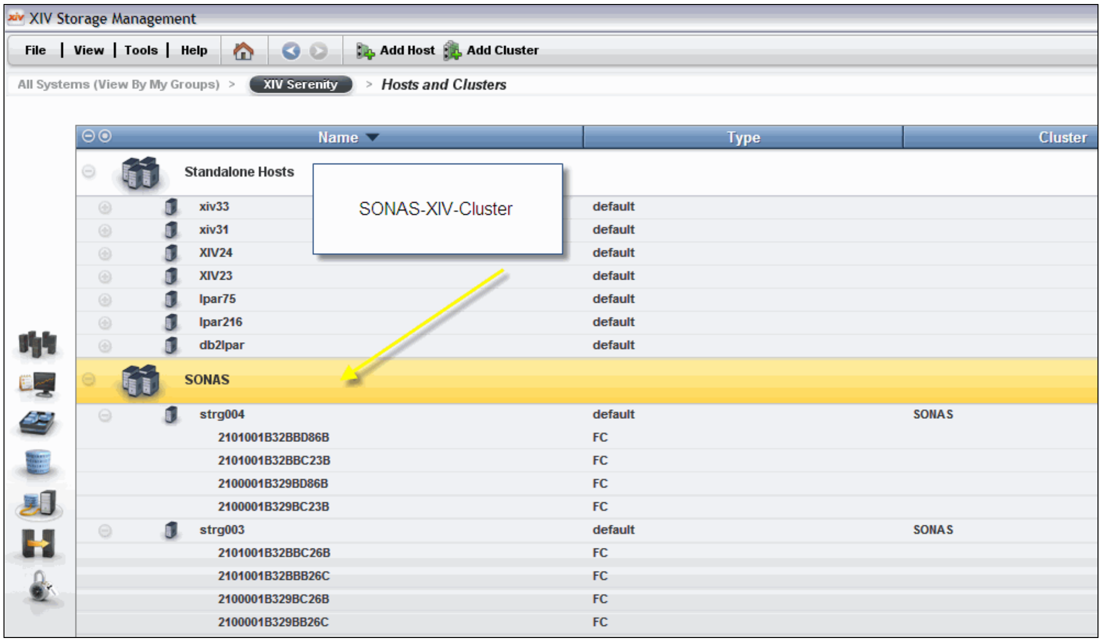
Defining SONAS to XIV storage

In the XIV GUI, create the following configuration from the Host and Clusters Menu:

- One SONAS cluster
- Two SONAS hosts (Storage nodes) per XIV
- Four WWPNs per SONAS host (ports)

It is easier to zone the SONAS Storage nodes in the fabric and to pick the WWPNs in the XIV GUI when the Storage nodes are powered on and the nodes are started. However, it can be done before or after the cluster software `first_time_install` script is run, if the RPQ flag is set.

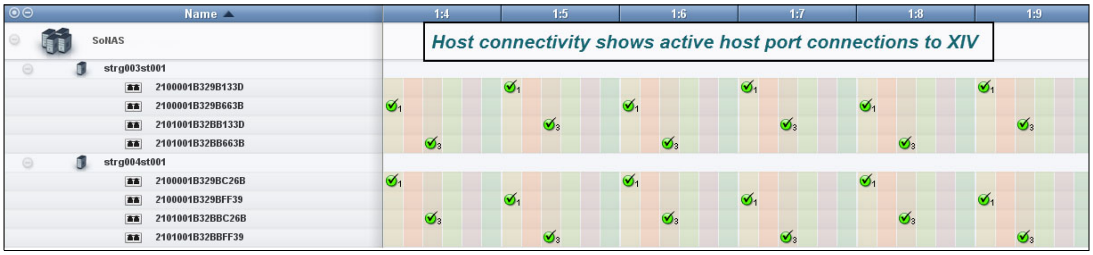
Figure 3-11 shows the XIV GUI with the SONAS Storage Cluster, nodes, and host ports configured.



Name	Type	Cluster
Standalone Hosts		
xiv33	default	
xiv31	default	
XIV24	default	
XIV23	default	
lpar75	default	
lpar216	default	
db2lpar	default	
SONAS		
strg004	default	SONAS
2101001B32BBD86B	FC	
2101001B32BBC23B	FC	
2100001B329BD86B	FC	
2100001B329BC23B	FC	
strg003	default	SONAS
2101001B32BBC26B	FC	
2101001B32BBB26C	FC	
2100001B329BC26B	FC	
2100001B329BB26C	FC	

Figure 3-11 XIV GUI - XIV cluster, hosts (SONAS Storage nodes), and host ports

Verify the host connectivity from the XIV Host Connectivity GUI window (see Figure 3-12). The window shows 12 paths per Storage node.



Host connectivity shows active host port connections to XIV

Name
SoNAS
strg003st001
2100001B329B133D
2100001B329B663B
2101001B32BB133D
2101001B32BB663B
strg004st001
2100001B329BC26B
2100001B329BFF39
2101001B32BBC26B
2101001B32BBFF39

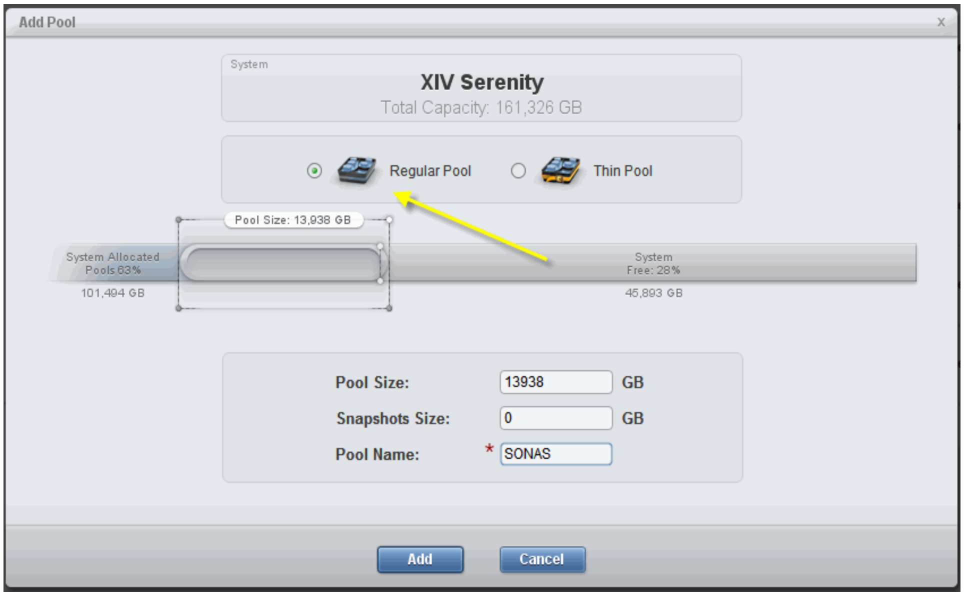
Figure 3-12 XIV GUI - XIV Host Connectivity - shows SONAS port to XIV module connections

Ensuring that the XIV storage has a proper SONAS storage pool and volumes

Here is an outline of the steps to create volumes from the XIV storage:

1. Configure one SONAS Storage Pool (thick or regular pool).
2. Ensure that all volumes follow the same size and name convention (sequentially numbered).
3. Set the Size = 4009 GB (4 TB) for Gen2 XIV (Gen3 sizes are slightly larger).
4. Set the Name = SONAS_(1, 2, 3, 4, and so on).

Figure 3-13 shows the XIV Add Pool GUI window.



● Regular Pool ○ Thin Pool

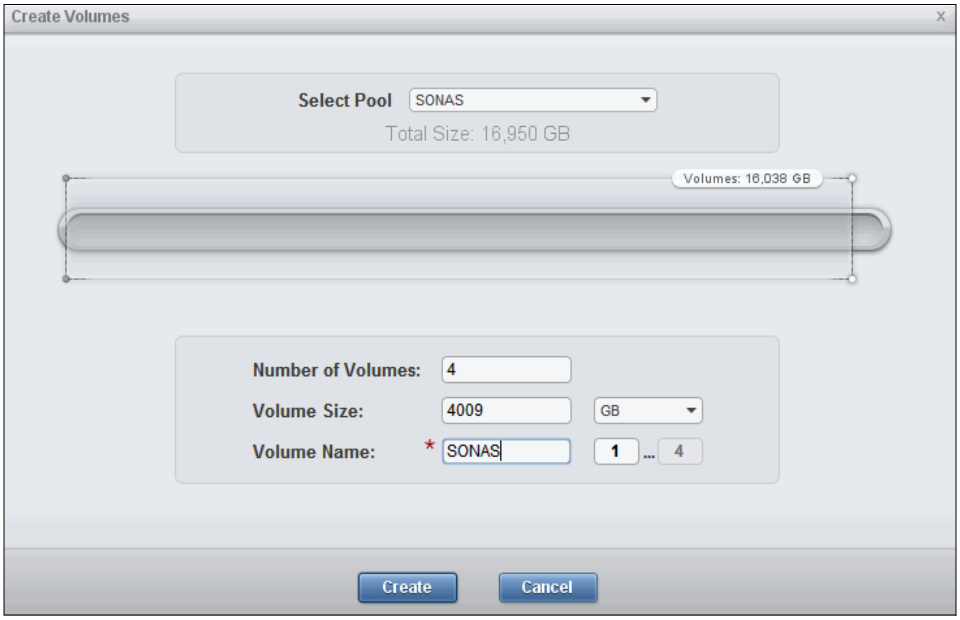
Pool Size: 13938 GB

Snapshots Size: 0 GB

Pool Name: * SONAS

Figure 3-13 XIV GUI- Storage pool configuration - Thick (Regular), 0 snapshots, SONAS naming

Figure 3-14 on page 115 shows the Create Volumes window in the XIV GUI to create volumes from the SONAS pool.



Select Pool: SONAS

Total Size: 16,950 GB

Volumes: 16,038 GB

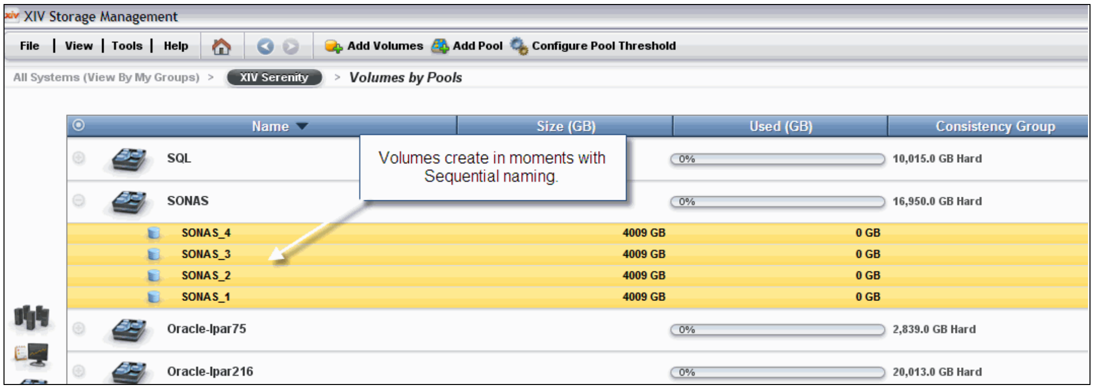
Number of Volumes: 4

Volume Size: 4009 GB

Volume Name: * SONAS 1 ... 4

Figure 3-14   XIV GUI - Volume definitions - 4009 GB (4 TB), SONAS naming (sequential #)

Figure 3-15 shows the volumes that are created from the SONAS volume. Note the sequential number naming of the volumes.



XIV Storage Management
File | View | Tools | Help
Add Volumes
Add Pool
Configure Pool Threshold
All Systems (View By My Groups) > XIV Serenity > Volumes by Pools

Name	Size (GB)	Used (GB)	Consistency Group
SQL		0%	10,015.0 GB Hard
SONAS		0%	16,950.0 GB Hard
SONAS_4	4009 GB	0 GB	
SONAS_3	4009 GB	0 GB	
SONAS_2	4009 GB	0 GB	
SONAS_1	4009 GB	0 GB	
Oracle-lpar75		0%	2,839.0 GB Hard
Oracle-lpar216		0%	20,013.0 GB Hard

Volumes create in moments with Sequential naming.

Figure 3-15   XIV GUI - XIV volume creation

Mapping the SONAS volumes to the SONAS cluster

After you create the volumes, you must map them to the SONAS cluster, as shown in Figure 3-16. Consider the following guidelines:

- Use standard Volume Map conventions (size and names).
- All volumes map to the cluster (not to the hosts within the cluster).
- Do not force mapping to LUN0 in XIV.

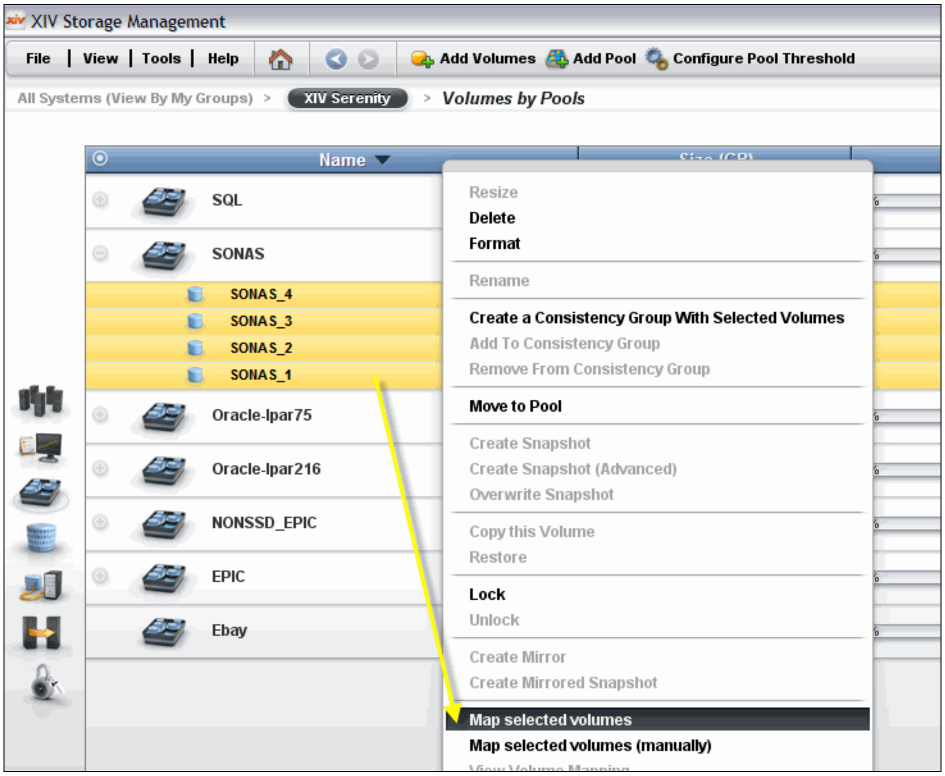


Figure 3-16 XIV GUI - XIV Volume mapping to a SONAS cluster (not hosts)

All XIV work is done and you are ready to proceed with SONAS code load and cluster installation and configuration.

Loading SONAS code and installing the cluster

To load the SONAS code and install the cluster, complete the following steps:

1. Insert the latest GA SONAS Release DVD into the Management node and power it on. This process is done while all other nodes (Interface and Storage nodes) remain off.
2. Install the operating system and software on the Management node (int001st001 in SONAS V1.3 and later).

   The SONAS software is obtained and installed by the IBM Service Representative (IBM PFE).

   If the installer resets the nodes to manufacturing default settings with the `manufacturing_cleanup` script, all nodes are shut down again. Only the Management node is powered on to begin the new SONAS code installation.

> **Tip:** Installation takes approximately 35 - 45 minutes to complete. The system restarts several times before completion. Initial installation is done when the DVD is ejected and the Management node allows a system login.

3. When the operating system is installed on the primary Management node, the installer runs the following command to set the Gateway flag on the installation:

   ```
   /opt/IBM/sonas/bin/mfg/cfg_gateway_rpq
   ```

   This flag tells the SONAS installation that Integrated storage and storage configuration must be run separately from the code installation.

   After it is run, a file is with the following name is created:

   ```
   /opt/IBM/sonas/etc/mfg_gateway_rpq
   ```

   Removing the file removes the flag.

4. After the pod is initially configured with Gateway storage, a different process (explained in 3.4.4, “Adding XIV LUNs to an existing SONAS configuration” on page 133, 3.5.6, “Adding Storwize V7000 LUNs to an existing SONAS configuration” on page 170, and 3.6.4, “Adding DCS3700 LUNs to an existing SONAS configuration” on page 214) is used to add storage to it.

After this installation task is finished, the IBM authorized service provider begins the cluster installation process. During this process, the `first_time_install` script is run.

Running the first_time_install script

To run the `first_time_install` script, complete the following steps:

1. Log in to the SONAS Management node (mgmt001st001) as root.
2. Run the following command to change the directory to `/opt/IBM/sonas/bin`:

   ```
   cd /opt/IBM/sonas/bin
   ```

3. Run the following command to verify the version and release of SONAS:

   ```
   /opt/IBM/sonas/bin/get_version
   ```

   Validate that the version that is installed is the version that you want.

- Run the `first_time_install` script and follow the prompts by using the *IBM SONAS Installation Guide*, GA32-0715 (with real data) and the *SONAS Introduction and Planning Guide*, GA32-0716 to create the initial cluster.

  Run the `first_time_install` script by running the following command and follow the prompts:

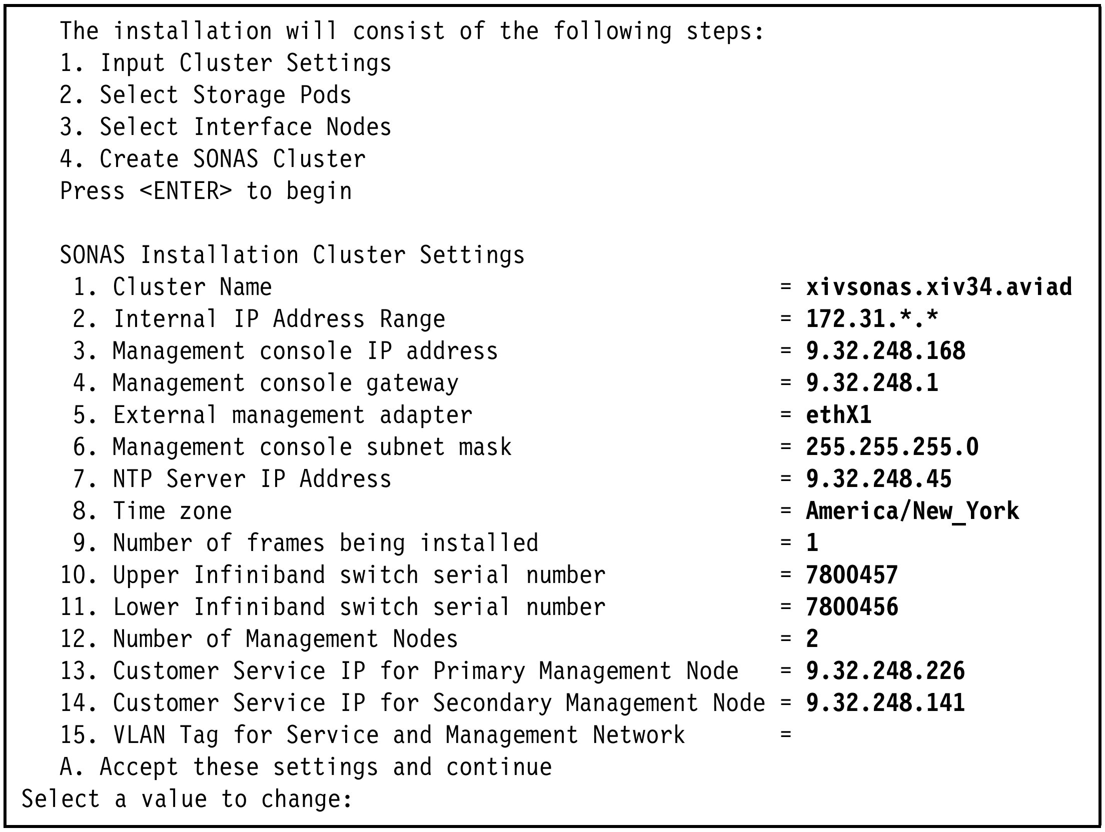
  ```
  /opt/IBM/sonas/bin/first_time_install
  ```

Defining cluster parameters

You must use the pre-installation planning worksheet to answer the questions that are related to first-time configuration of the cluster and Management node information. The initial steps require you to provide the configuration information that is shown in Figure 3-17. The customer provides this information before the installation is started. See the planning tables in Chapter 1, “Installation planning” on page 1 for the information that is needed.

After all the parameters are entered and verified, enter “A” and the prompt to continue.

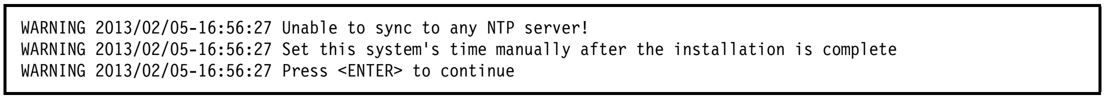
Figure 3-17 shows a `first_time_install` script for SONAS V1.5.1. Versions before Version 1.5.1 have *only* 13 cluster setting options.



```
   The installation will consist of the following steps:
   1. Input Cluster Settings
   2. Select Storage Pods
   3. Select Interface Nodes
   4. Create SONAS Cluster
   Press <ENTER> to begin

   SONAS Installation Cluster Settings
    1. Cluster Name                                     = xivsonas.xiv34.aviad
    2. Internal IP Address Range                        = 172.31.*.*
    3. Management console IP address                    = 9.32.248.168
    4. Management console gateway                       = 9.32.248.1
    5. External management adapter                      = ethX1
    6. Management console subnet mask                   = 255.255.255.0
    7. NTP Server IP Address                            = 9.32.248.45
    8. Time zone                                        = America/New_York
    9. Number of frames being installed                 = 1
   10. Upper Infiniband switch serial number            = 7800457
   11. Lower Infiniband switch serial number            = 7800456
   12. Number of Management Nodes                       = 2
   13. Customer Service IP for Primary Management Node   = 9.32.248.226
   14. Customer Service IP for Secondary Management Node = 9.32.248.141
   15. VLAN Tag for Service and Management Network      =
   A. Accept these settings and continue
Select a value to change:
```

Figure 3-17 Management node data from the first_time_install script

As the script progresses, in some cases, especially where the management port is not connected to a network because it will share the Public Network Ethernet ports, you receive a “Warning syncing NTP” message. SONAS V1.5.1 `first_time_install` updates are designed to reduce its occurrence. You might not receive this message. If you do receive it, it is acceptable. Press Enter to continue, as shown in Figure 3-18.



```
WARNING 2013/02/05-16:56:27 Unable to sync to any NTP server!
WARNING 2013/02/05-16:56:27 Set this system's time manually after the installation is complete
WARNING 2013/02/05-16:56:27 Press <ENTER> to continue
```

Figure 3-18 NTP sync warning

Powering on and discovering each node

To power on and discover the SONAS nodes, complete the following steps:

1. As the SONAS `first_time_install` script progresses, the script prompts you to **“**`power on all nodes`**”**. After you power on all nodes, you can press Enter from the `first_time_install` script prompt.

   Nodes use the Preboot eXecution Environment (PXE, also known as Pre-Execution Environment) to boot and load the OS image. The ISO image is transferred onto the detected nodes. A list of recognized nodes appears and is frequently updated. Nodes appear in the list as they are recognized and configured (Figure 3-19). This process might take 60 minutes or longer, depending on the number of SONAS nodes in the configuration, to recognize and complete the SONAS code load on each node.

2. When all Interface nodes and Storage nodes are discovered (and the quantity numbers keep repeating with no change), press Enter to continue the configuration.

```
Detected 0 Interface nodes and 0 Storage nodes
Detected 1 Interface nodes and 0 Storage nodes
Detected 2 Interface nodes and 0 Storage nodes
Detected 3 Interface nodes and 1 Storage node
Detected 3 Interface nodes and 2 Storage nodes
Detected 3 Interface nodes and 2 Storage nodes
```

Figure 3-19 Sample output for three Interface nodes and two Storage nodes

Verifying device IDs, rack and slot locations, and quorums

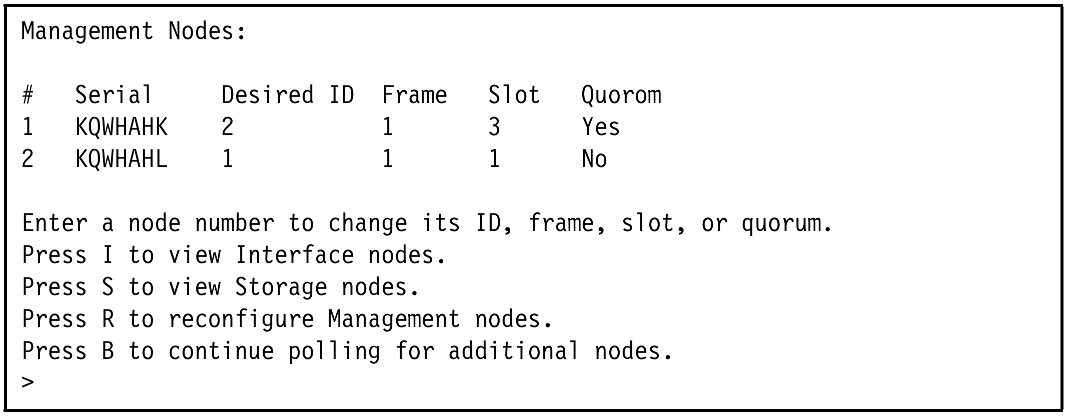
Correct rack cabling and configuration are key to a correctly working SONAS system and proper GUI representation. Accurate cabling in the internal SONAS frame (Ethernet and InfiniBand) switches ensure proper frame location identification and configuration of these nodes. However, locations can sometimes require redefinition in the `first_time_install` process. If it is not done correctly, the GUI might not properly align the nodes in the health center frame model.

You can adjust and keep the configuration when the accurate assignments are confirmed for Interface nodes (I), Management nodes (M) (Figure 3-20), and Storage nodes (S) (Figure 3-21).

Complete the following steps:

1. After all rack locations and Quorum statuses are verified, select "C" to continue to configure the cluster.

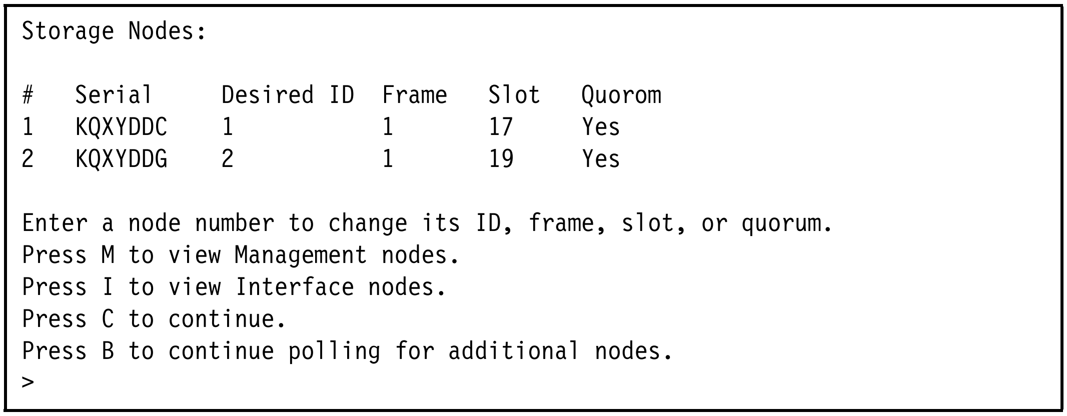
   Figure 3-20 shows the Management nodes.



```
Management Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQWHAHK    2           1       3      Yes
2   KQWHAHL    1           1       1      No

Enter a node number to change its ID, frame, slot, or quorum.
Press I to view Interface nodes.
Press S to view Storage nodes.
Press R to reconfigure Management nodes.
Press B to continue polling for additional nodes.
>
```

Figure 3-20 Management Nodes configuration

Figure 3-21 shows the Storage nodes.



```
Storage Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQXYDDC    1           1       17     Yes
2   KQXYDDG    2           1       19     Yes

Enter a node number to change its ID, frame, slot, or quorum.
Press M to view Management nodes.
Press I to view Interface nodes.
Press C to continue.
Press B to continue polling for additional nodes.
>
```

Figure 3-21 Storage Nodes configuration

Configuring the node instances, location, including serial numbers and quorum states, is key to a correctly working SONAS system. Typically, the first instance of a node type is in the lowest point of the rack of its node type and the sequence goes up as you move up the rack. For example, the bottom Interface node is int001st001 and the next one up is int002st001. The bottom Ethernet switch is switch 1, and the next one up is switch 2. The bottom InfiniBand switch is switch 1, and the next one up is switch 2.

This configuration changes for the SONAS Storage nodes in Gateway configurations. The first and second Storage nodes are above Storage node 3 and 4. The rest of the Storage nodes are above Storage nodes 1 and 2. You can adjust and keep the configuration when the accurate assignments are confirmed for Interface and Storage nodes (S), then select "C" to continue to configure the cluster.

2. Select the line item number to change the configuration (Device ID/instance, Frame, slot, and Quorum state).

   – Desired ID: This selection is the instance of the node type. Therefore, the first Management node instance has a desired ID of 1 and the second ID has a desired ID of 2.

     In Figure 3-22 on page 122, ID 1 is in the lowest slot of the rack. The bottom Interface node in the frame is int001st001 (slot 1 = node 1). The second Interface node from the bottom of the frame is int002st001 (slot 3= node 2).

     Storage nodes have a different configuration, as shown in Figure 3-22 on page 122. Storage node 1 and 2 are above Storage nodes 3 and 4.

   – Frame Number: The frame number relates to the frame instance of SONAS. If this frame is one of the SONAS frames, use 1. If you have a base rack and an expansion frame, then the base rack is 1 and the first expansion frame is 2.

   – Slot number (rack slot number): Use the lower U-number indicator of the slot for the device (in the frame).

   – Quorum: There are three sizes of *quorum node configuration* for SONAS installations (small = 3, medium = 5, and large = 7). If a cluster is designed as small (2 - 6 Interface nodes), you need three quorum nodes. If the cluster is medium sized (6 - 10 Interface nodes), set five quorum nodes. If the cluster is large (over 10 Interface nodes), set seven quorum nodes.

Figure 3-22 shows a SONAS Gateway RXA rack layout configuration and the node naming conventions.

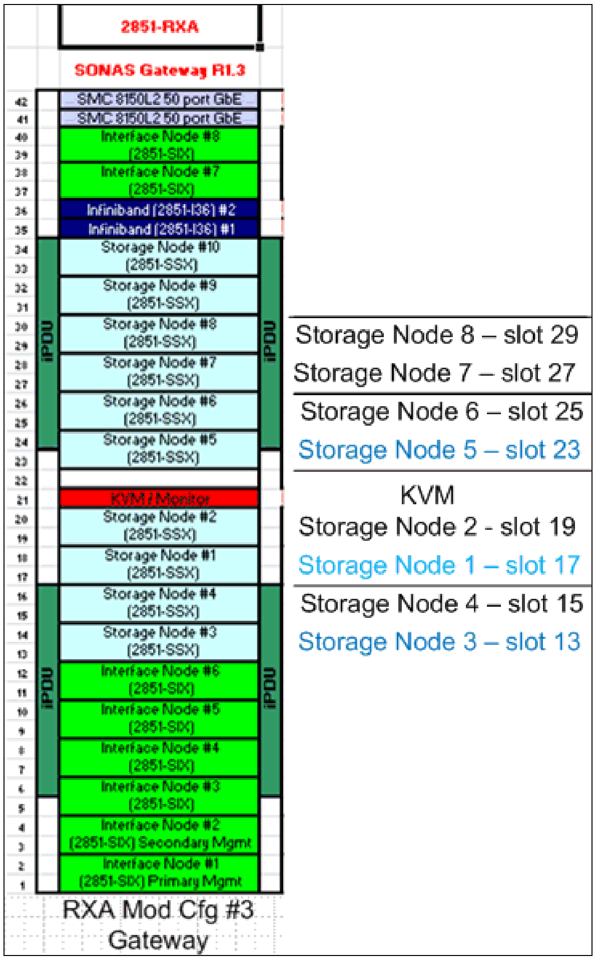

2851-RXA
SONAS Gateway R1.3
42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Interface Node #8 (2851-SIX)
Interface Node #7 (2851-SIX)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Storage Node #10 (2851-SSX)
Storage Node #9 (2851-SSX)
Storage Node #8 (2851-SSX)
Storage Node #7 (2851-SSX)
Storage Node #6 (2851-SSX)
Storage Node #5 (2851-SSX)
KVM / Monitor
Storage Node #2 (2851-SSX)
Storage Node #1 (2851-SSX)
Storage Node #4 (2851-SSX)
Storage Node #3 (2851-SSX)
Interface Node #6 (2851-SIX)
Interface Node #5 (2851-SIX)
Interface Node #4 (2851-SIX)
Interface Node #3 (2851-SIX)
Interface Node #2 (2851-SIX) Secondary Mgmt
Interface Node #1 (2851-SIX) Primary Mgmt
iPDU
iPDU
iPDU
iPDU
RXA Mod Cfg #3 Gateway
Storage Node 8 – slot 29
Storage Node 7 – slot 27
Storage Node 6 – slot 25
Storage Node 5 – slot 23
KVM
Storage Node 2 - slot 19
Storage Node 1 – slot 17
Storage Node 4 – slot 15
Storage Node 3 – slot 13

Figure 3-22 SONAS RXA - slot location offset for Storage nodes

> **Note:** In SONAS V.3.2 and later, the management functions are combined with the first two Interface nodes. Therefore, the first two Interface nodes (bottom two nodes of the rack) are called mgmt001st001 and mgmt002st001. In a minimal configuration (that is, a configuration with two Interface nodes), int001st001 or int002st001 do not exist.

3. After all Interface and Storage nodes are listed properly and verified, type “B” to poll again or type “C” to continue with the cluster configuration. You can enter “C” to continue only from the Storage node configuration screen (see Figure 3-21 on page 120). Option “C” can be accessed only through the Storage Node listing (“S”).

Configuring the cluster and completing the first_time_install script

The `first_time_install` script continues by creating the SONAS cluster and adding each node to the cluster. This process takes 30 - 40 minutes, or longer, depending on size of the configuration.

Figure 3-23 on page 123 shows the end of the script after which the cluster is successfully configured. If it does not end with a success statement, open a PMR and immediately escalate to support for immediate assistance.

```
2013/08/01-21:28:16: Configuring a SONAS gateway.
2013/08/01-21:29:17: Creating GPFS cluster
2013/08/01-21:29:53: Configuring GPFS Cluster settings:
2013/08/01-21:31:55: Configuring GPFS node settings on node: KQWHAHK
2013/08/01-21:32:03: Configuring GPFS node settings on node: KQWHAHL
2013/08/01-21:32:11: Configuring GPFS node settings on node: KQXYDDC
2013/08/01-21:32:16: Configuring GPFS node settings on node: KQXYDDG
2013/08/01-21:33:54: Configuring multipath on node:KQXYDDC
2013/08/01-21:34:30: Configuring multipath on node:KQXYDDG
2013/08/01-21:34:33: Validating the NSDs before continuing, this can take up to 20 minutes.
2013/08/01-21:35:59: Skipping storage subsystem upgrade
2013/08/01-21:36:00: Synchronizing the SONAS repository with the secondary Management node
2013/08/01-21:37:53: Configuring yum on all nodes
2013/08/01-21:40:22: Configuring Performance Center service
2013/08/01-21:40:32: Starting system health monitoring

2013/08/01-21:41:48: Switch inventory complete

This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.

2013-08-01T21:41:49.914308-04:00: *** END /opt/IBM/sonas/bin/first_time_install(rc=0) ELP[1 hours 13 minutes 44
seconds]
```

Figure 3-23 Cluster being created and the first_time_install script completes

In some cases, especially where the management network shares Ethernet with the public network, the `first_time_install` script ends with an NTP Sync error. SONAS V1.5.1 `first_time_install` updates are designed to reduce the occurrence of this error. It is okay to receive it because NTP correctly syncs when the management network is properly configured (see Figure 3-24).

```
This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.

1 error was found, please repair it before continuing.

This is the deferred error:
ERROR---set_mgmt Code 107/0AFE: Unable to set the system's timezone. Unable to sync to any NTP server ---ERROR
2013-02-07T12:33:56.279583-05:00: *** END /opt/IBM/sonas/bin/first_time_install(rc=0) ELP[19 hours 2 minutes 40
seconds]
```

Figure 3-24 The first_time_install script ends with an error

Postinstallation procedures after the first_time_install script finishes

The GPFS cluster is installed, configured, and running. The installer does postinstallation procedures.

Enabling licensing

Enable the license by running the following command:

```
/opt/IBM/sonas/bin/enablelicense --accept
```

The license can also be accepted the first time that the GUI is used.

Adding the GPFS cluster to SONAS management

Before any other management functions can be used, the GPFS cluster must be added to the SONAS Management subsystem. Run the following command:

```
cli addcluster -h mgmt001st001 -p Passw0rd
```

Configuring the SONAS management port

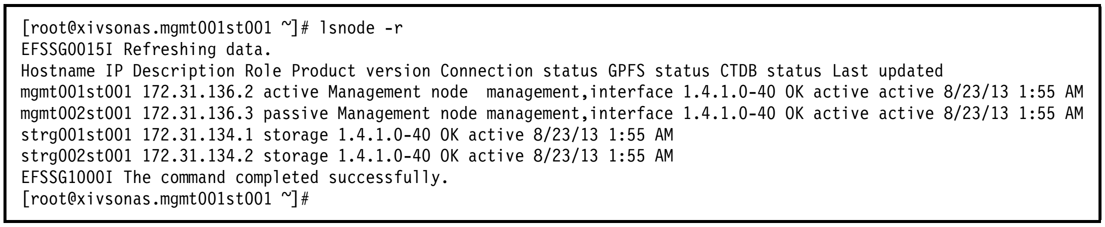
In many configurations, the management port is configured to use the same Ethernet ports as the public network (10 GbE). In these configurations, the management port is not connected to the network. Therefore, the management subsystem must be configured to use the 10 GbE network ports. SONAS V1.5.1 `first_time_install` enhancements allow the user to identify the external management adapter and prevent the need to run the **`chnwmgt`** command later. After a Version 1.5.1 installation, the **`chnwmgt`** command must be run only if there is if no external management adapter value as input at installation time or if an incorrect value is entered and must be changed later. Run the following command to configure the management ports on the 10 GbE network (if needed):

```
chnwmgt --interface ethX1
```

The SONAS GUI is available for use.

Verifying the node list

Verify the node list by running **`lsnode -r`**, as shown in Figure 3-25.



```
[root@xivsonas.mgmt001st001 ~]# lsnode -r
EFSSG0015I Refreshing data.
Hostname IP Description Role Product version Connection status GPFS status CTDB status Last updated
mgmt001st001 172.31.136.2 active Management node  management,interface 1.4.1.0-40 OK active active 8/23/13 1:55 AM
mgmt002st001 172.31.136.3 passive Management node management,interface 1.4.1.0-40 OK active active 8/23/13 1:55 AM
strg001st001 172.31.134.1 storage 1.4.1.0-40 OK active 8/23/13 1:55 AM
strg002st001 172.31.134.2 storage 1.4.1.0-40 OK active 8/23/13 1:55 AM
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-25 Sample lsnode output that shows configured nodes

Health check

It is time to check the system health.

The IBM authorized service provider logs in to the Management node and runs the health check commands. The following command is run to check the SONAS system overall health and all components (Ethernet Switches, InfiniBand Switches, and nodes):

```
cnrsscheck --nodes=all --checks=all
```

The command also checks whether the nodes have correctly assigned roles and whether they are able to communicate with each other.

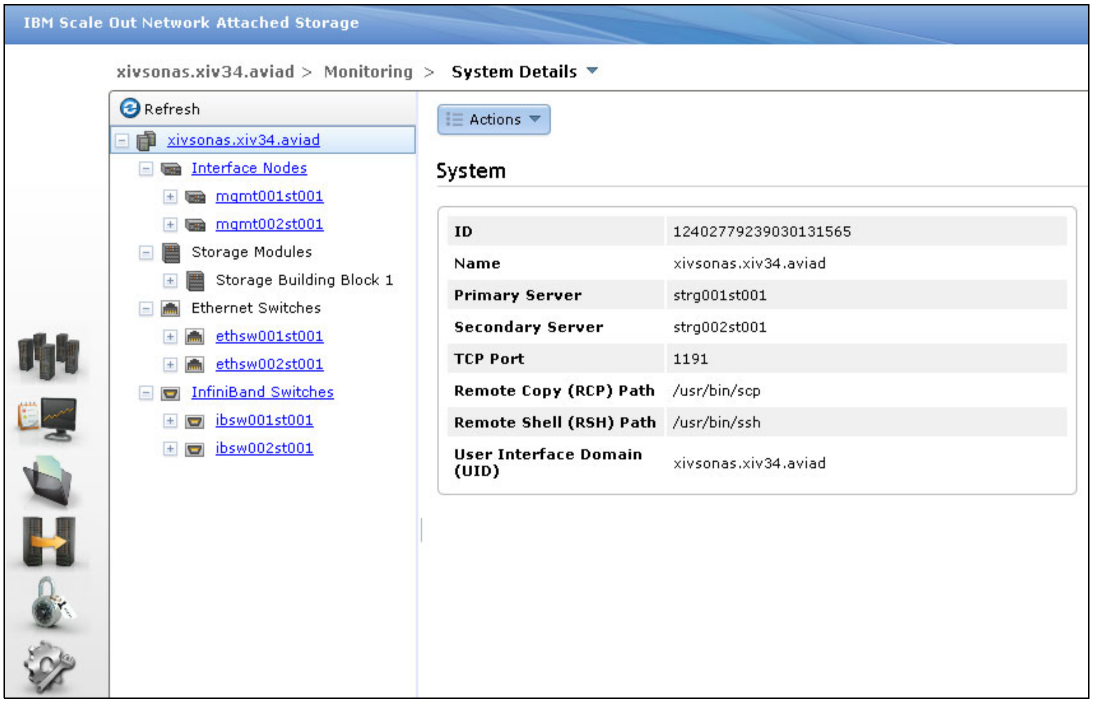
Figure 3-26 shows the command output for a sample cluster configuration.

```
==============================================================================================
                          Health summary for each node
----------------------------------------------------------------------------------------------
 Node name     - Target node name of summary
 Fatal         - If 1, indicates that fatal error occurred during check
 Warnings/Degrades/Failures/Offlines/Informational
               - Number of each NON-OK health

    Node name      |  Fatal  |   Warnings    Degrades    Failures    Offlines    Informational
  ---------------+---------+-------------------------------------------------------------------
    mgmt001st001   |    0    |       0           0           0           0             0
    mgmt002st001   |    0    |       0           0           0           0             0
    strg001st001   |    0    |       0           0           0           0             0
    strg002st001   |    0    |       0           0           0           0             0
----------------------------------------------------------------------------------------------
    IB Switch      |    0    |       0           0           0           0             0
    Ethernet/SMC   |    0    |       0           0           0           0             0
----------------------------------------------------------------------------------------------
Please check detailed status above if you can see error in table above.
If there is Fatal error, please login to target node and use 'cnrsscdisplay'
command for more details.
================================================================================
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-26 Sample output from the cnrsscheck command summary

Clearing miscellaneous installation errors

During the installation process, some miscellaneous errors, or information warnings, might be generated as nodes are loaded and rebooted and ports go on and off. Review and clear these errors before you run the health check. You can do this task in the GUI and the system monitoring window. Ensure that all components are green. For details, see Figure 3-27.



System

ID	12402779239030131565
Name	xivsonas.xiv34.aviad
Primary Server	strg001st001
Secondary Server	strg002st001
TCP Port	1191
Remote Copy (RCP) Path	/usr/bin/scp
Remote Shell (RSH) Path	/usr/bin/ssh
User Interface Domain (UID)	xivsonas.xiv34.aviad

Figure 3-27 SONAS Monitor System window

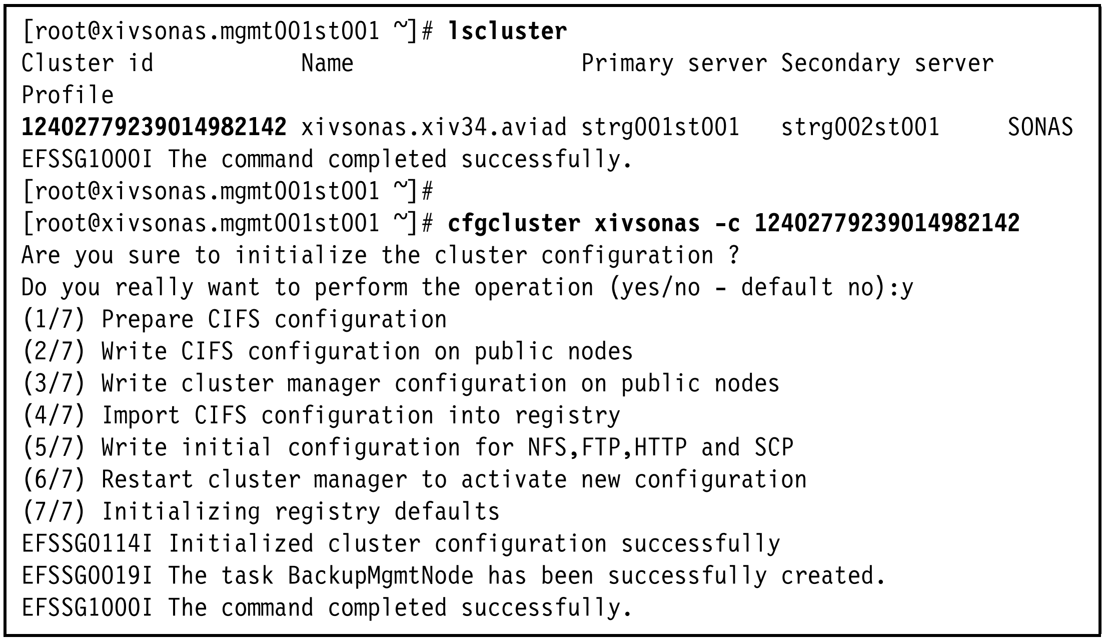
Completing the SONAS cluster configuration

Before you create file systems, file sets, or exports, you must complete the SONAS cluster configuration.

To complete the configuration, run the following commands:

- **`lscluster`**, which confirms proper cluster configuration and lists the unique cluster ID. This ID is required in the following command.
- **`cfgcluster`**, which creates the initial configuration for all supported protocols (HTTPS, NFS, CIFS, FTP, and SCP). Here is a list of its functions:
  - Prepare the CIFS configuration.
  - Distribute the CIFS configuration.
  - Distribute the CTDB configuration.
  - Import the CIFS configuration into the registry.
  - Write the initial configuration for NFS, FTP, HTTP, and SCP.
  - Restart CTDB to activate the new configuration.

  See the example in Figure 3-28.



```
[root@xivsonas.mgmt001st001 ~]# lscluster
Cluster id          Name                 Primary server Secondary server
Profile
12402779239014982142 xivsonas.xiv34.aviad strg001st001   strg002st001     SONAS
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
[root@xivsonas.mgmt001st001 ~]# cfgcluster xivsonas -c 12402779239014982142
Are you sure to initialize the cluster configuration ?
Do you really want to perform the operation (yes/no - default no):y
(1/7) Prepare CIFS configuration
(2/7) Write CIFS configuration on public nodes
(3/7) Write cluster manager configuration on public nodes
(4/7) Import CIFS configuration into registry
(5/7) Write initial configuration for NFS,FTP,HTTP and SCP
(6/7) Restart cluster manager to activate new configuration
(7/7) Initializing registry defaults
EFSSG0114I Initialized cluster configuration successfully
EFSSG0019I The task BackupMgmtNode has been successfully created.
EFSSG1000I The command completed successfully.
```

Figure 3-28 cfgcluster example

Upon completion, the cluster is now configured and operational.

However, more configuration is required for cluster networking and authentication. Because this configuration it is not directly related to the underlying storage, it is presented later in this chapter and is independent of the Gateway solutions. Read through the chapter to understand the network installation process and preferred practices for your cluster installation.

Starting with 3.8, “SONAS integration into your network” on page 218, this chapter provides detailed information about general SONAS and GPFS storage configuration information that is relative to any back-end storage. It also provides information about networking and other considerations that are not back-end storage specific. Read the entire chapter to understand all the installation considerations before you plan how to integrate SONAS into your environment installation.

Rescanning Storage nodes

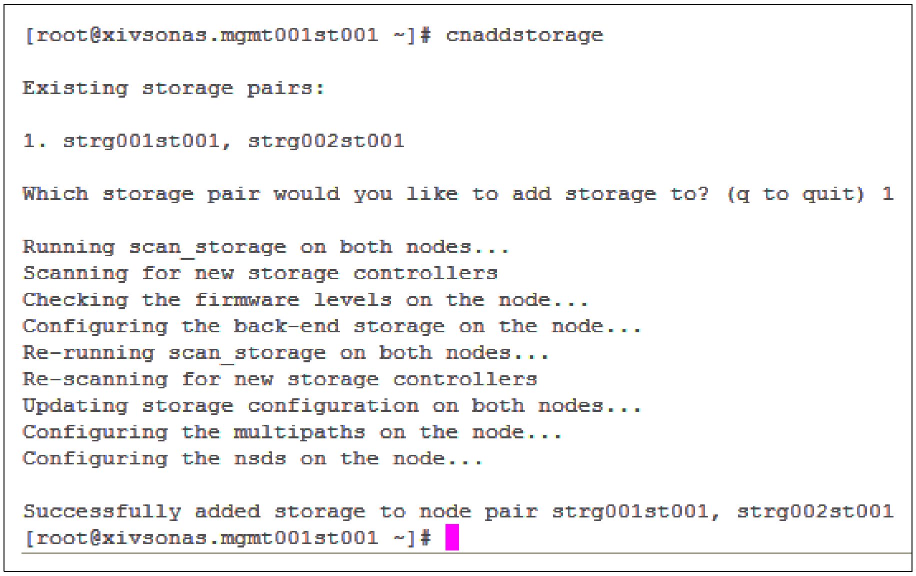
If the storage was not configured and allocated to the cluster before you ran the `first_time_install` script, run **`mkdisk --luns`** (SONAS V1.4.1 and above) as the admin or root user. This command rescans all available Storage nodes for newly allocated LUNs (see Figure 3-29).

```
[xivsonas.xiv34.aviad]$ mkdisk --luns
(1/3) Scanning for new devices
(2/3) Creating NSDs
(3/3) Adding to database
Successfully created disks:
XIV6000095_SONAS_11
XIV6000095_SONAS_12
EFSSG1000I The command completed successfully.
[xivsonas.xiv34.aviad]$
```

Figure 3-29 Example of mkdisk output

An alternative method is to run **`cnaddstorage`** as the root user. This command presents a short list of Storage node pairs from which to select.

Select the Storage node pair to which you added storage. SONAS automatically discovers new volumes, and creates the multipath devices and the associated NSDs. The command must be run for each Storage node pair or pod. This process might take up to six minutes to complete (see Figure 3-30).



```
[root@xivsonas.mgmt001st001 ~]# cnaddstorage

Existing storage pairs:

1. strg001st001, strg002st001

Which storage pair would you like to add storage to? (q to quit) 1

Running scan_storage on both nodes...
Scanning for new storage controllers
Checking the firmware levels on the node...
Configuring the back-end storage on the node...
Re-running scan_storage on both nodes...
Re-scanning for new storage controllers
Updating storage configuration on both nodes...
Configuring the multipaths on the node...
Configuring the nsds on the node...

Successfully added storage to node pair strg001st001, strg002st001
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-30 Sample output from the cnaddstorage command

When LUNs and disks are rescanned, the installer continues the installation process.

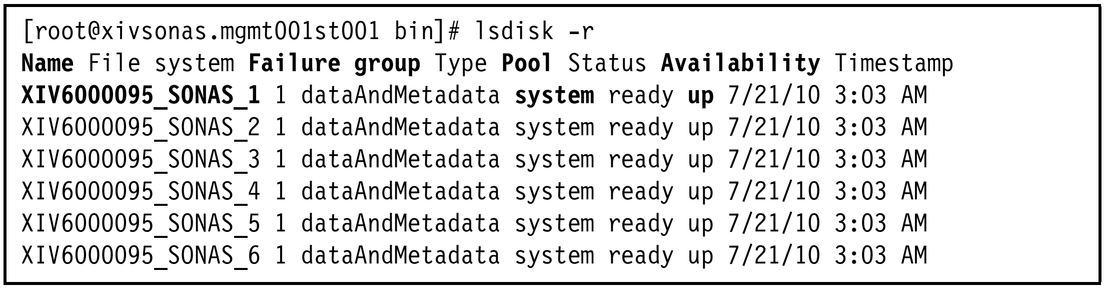
> **Tip:** If you are reinstalling SONAS, previous NSD labels must be revoked before you reinstall LUN devices. Either reformat or run **`dd if=/dev/zero of=/dev/mapper/`*`device-name`*** against the device for about 5 seconds to clear the label.

Verifying the disk configuration

From the Management node prompt, run the following command:

```
lsdisk -r
```

The output is shown in Figure 3-31.



```
[root@xivsonas.mgmt001st001 bin]# lsdisk -r
Name File system Failure group Type Pool Status Availability Timestamp
XIV6000095_SONAS_1 1 dataAndMetadata system ready up 7/21/10 3:03 AM
XIV6000095_SONAS_2 1 dataAndMetadata system ready up 7/21/10 3:03 AM
XIV6000095_SONAS_3 1 dataAndMetadata system ready up 7/21/10 3:03 AM
XIV6000095_SONAS_4 1 dataAndMetadata system ready up 7/21/10 3:03 AM
XIV6000095_SONAS_5 1 dataAndMetadata system ready up 7/21/10 3:03 AM
XIV6000095_SONAS_6 1 dataAndMetadata system ready up 7/21/10 3:03 AM
```

Figure 3-31 The lsdisk report of XIV volumes (XIV, XIVserialNumber, and volume_name_instance)

All the XIV disks from the pods are configured, and are visible from the Management node. By default, they are added to the GPFS “system” pool in failure group 1, and the Storage node preference is assigned in alternating LUN (called NSDs in GPFS) fashion.

Interpreting the lsdisk command output

When you review the output from the **lsdisk** command, you can see information about disk devices that you might want to consider before you start configuring your file systems.

The following list provides the output that is shown in Figure 3-31, starting from left to right across the top:

- The device list begins with *Name*, which refers to the NSD name. The NSD name is the GPFS “Network Storage Device” label:
  - For XIV based storage, the first three characters refer to the storage type (XIV).
  - The next seven characters refer to the XIV serial number (it is a unique identifier for all XIV frames). In this case, the XIV frame serial number is “6000095”.
  - The next set of characters refers to the Volume names that are assigned in the NSDs during volume creation (for XIV storage). In this example, the volumes were created with sequential numbering and the volume name SONAS. The XIV automatically appends the numbers after the volumes when you create a batch of like-sized volumes with the name SONAS. Again, this example created SONAS_1, SONAS_2, SONAS_3, and SONAS_4. The resulting NSD name is “XIV6000095_SONAS_1”.
- The next area highlights a preferred practice for *XIV volume naming*.

  When you create NSDs with XIV storage, use simple volume names, such as SONAS, and allow XIV to number them sequentially, always using the same size volume. With this practice, even if you create volumes from 10 different XIVs, the volume list is always listed by serial number and volume number, and they remain simple to manage logically.
- The next piece of information from the **lsdisk** list is *File system*.

  If the listed NSD is not yet assigned to a file system, this column remains blank for that device. When the device is assigned to a file system, the file system name (such as “gpfs0”) is listed in this column.

- The next column in the list is *Failure group*. The NSD failure group is a value that GPFS allows you to assign to any NSD to logically manage groups of disk devices in that file system. For example, if you have two XIV frames that provision storage into a single storage pod, you can, and typically should, assign the first XIV to failure group 1, and the second XIV to failure group 2. By doing so, you can define different replication between these groups.
- *Type* definition refers to the usage type of data for which you use that NSD. For example, you can use the disk for dataOnly, metadataOnly, or metadataAndData.
- The *Pool* definition refers to the category of disk for that disk type. If you have SAS and Near Line SAS disks and you choose to have them both in the same file system, this situation is possible by having them defined in different “pools”. The file system places data on the “system pool” by default. You can then migrate data to a different pool by a structured ILM policy.

  A preferred practice is to place your fastest tier of disk storage in the system pool and migrate to lower tiers from there. It ensures metadata placement on the highest speed disk technology.
- *Status* displays the readiness of the NSD for use.

  Disk status has five possible values, three of which are transitional:

  – Ready: Normal status.
  – Suspended: Indicates that data is to be migrated off this disk.
  – Being emptied: Transitional status in effect while a disk deletion is pending.
  – Replacing: Transitional status in effect for an old disk while replacement is pending.
  – Replacement: Transitional status in effect for a new disk while replacement is pending.

  GPFS allocates space only on disks with a status of ready or replacement.
- *Availability* refers to the following possible values:

  – Up: The disk is available to GPFS for normal read and write operations.

  – Down: No read and write operations can be done on the disk.

  – Recovering: An intermediate state for disks that are powering on. During this process, GPFS verifies and corrects data. Read operations can be done while a disk is in this state, but write operations cannot.

  – Unrecovered: Not all disks were successfully brought up. There are cases where multiple disks must be recovered simultaneously to bring the storage to a consistent state. The most obvious cases involve replication where both copies of some pointers to data on this disk are unavailable. Unrecovered means that the disk is physically available, but the prerequisites for running recovery are not satisfied.

  – Timestamp: Refers to the date and time created.

> **Note:** Disk Pool, Usage Type, and Failure Groups are assigned to the disks and NSDs and must be defined before the disks are associated with a file system. These parameters can be changed only by removing the disk from the file system, changing the parameters, and readding the disks to the file system. This process assumes that there is enough space on the remaining disks in the file system to maintain the used capacity.

Creating the file system

Before you create your file system, make sure that the volumes are evenly balanced across the Storage nodes. Balance LUNs and NSDs across all Storage nodes. When SONAS imports NSDs from the underlying storage subsystem, it assigns a Storage node preference to each NSD in alternating fashion.

With XIV solutions, it is simple because there is one RAID, size, disk type, and so on.

You can view the balance by running **mmlsnsd**.

The Storage node preference is important to performance because it defines the Storage node that tries to manage all I/O to that NSD at any particular time. The I/O is managed only by the second node in the Storage node preference if the first node fails to serve the I/O. This failover behavior is automatic. Figure 3-32 shows that for NSD “XIV6000095_SONAS_1,” the Storage node strg001st001 is the Storage node preference, and strg002st001 is the backup node for I/O that is sent to the NSD. This behavior can be confirmed in the XIV statistics monitor.

As you can see in Figure 3-32, the Storage node preference for strg001st001 is alternating sequentially by volume name. It makes file system assignment simple. If you provision a file system with consecutively named NSDs, the file system is automatically balanced across each Storage node if there are an even number of NSDs assigned to the file system.

```
[root@xivsonas.mgmt001st001 bin]# mmlsnsd -L
File system Disk name NSD servers
---------------------------------------------------------------------------
gpfs0 XIV6000095_SONAS_1 strg001st001,strg002st001
gpfs0 XIV6000095_SONAS_2 strg002st001,strg001st001
gpfs1 XIV6000095_SONAS_3 strg001st001,strg002st001
gpfs1 XIV6000095_SONAS_4 strg002st001,strg001st001
gpfs2 XIV6000095_SONAS_5 strg001st001,strg002st001
gpfs2 XIV6000095_SONAS_6 strg002st001,strg001st001
```

Figure 3-32 Output from the mmlsnsd command that shows the Storage node preference on the NSDs

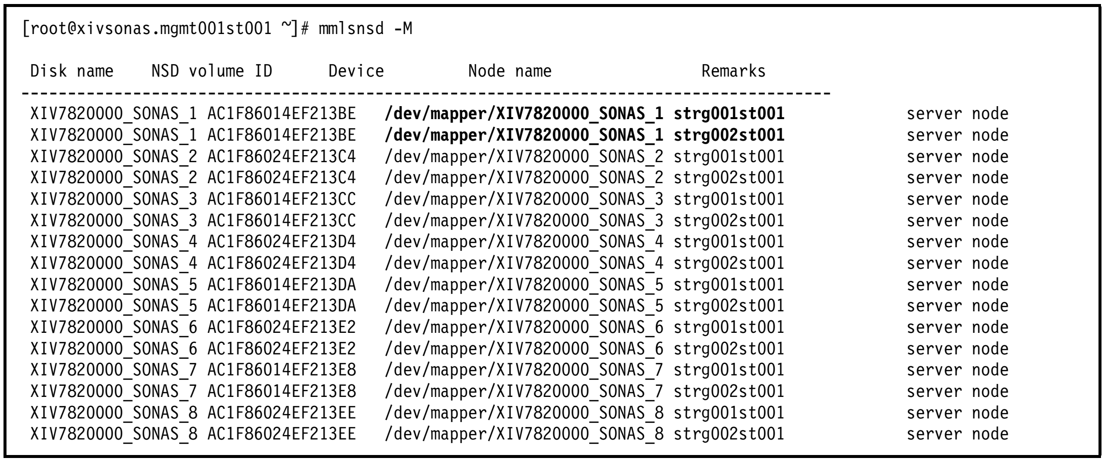
It is even more important when multiple frames are deployed. If you want the highest achievable performance, spread the NSD resources evenly across all Storage nodes. The following command output can help illustrate this concept.

In the case that is shown in Figure 3-33, the two XIVs are in the same Storage node pair. The different XIVs are clearly identifiable from the serial number indicator in the NSD name. When you create the file system, ensure that you type in the NSD devices across the Storage node and XIV frame preference.

```
[root@xivsonas.mgmt001st001 bin]# mmlsnsd
File system Disk name NSD servers
---------------------------------------------------------------------------
gpfs0 XIV6000095_SONAS_1 strg001st001,strg002st001
gpfs0 XIV6000095_SONAS_2 strg002st001,strg001st001
gpfs0 XIV7802005_SONAS_1 strg001st001,strg002st001
gpfs0 XIV7802005_SONAS_2 strg002st001,strg001st001
gpfs2 XIV6000095_SONAS_3 strg001st001,strg002st001
gpfs2 XIV7802005_SONAS_3 strg002st001,strg001st001
```

Figure 3-33 Multi XIV output from the mmlsnsd command that shows the Storage node preference

Additionally, it might be helpful to review the post configuration device management on the Storage node pairs. It can be done by running **mmlsnsd -m**, as shown in Figure 3-34.



```
[root@xivsonas.mgmt001st001 ~]# mmlsnsd -M

 Disk name    NSD volume ID      Device         Node name                Remarks
---------------------------------------------------------------------------------------
 XIV7820000_SONAS_1 AC1F86014EF213BE   /dev/mapper/XIV7820000_SONAS_1 strg001st001            server node
 XIV7820000_SONAS_1 AC1F86014EF213BE   /dev/mapper/XIV7820000_SONAS_1 strg002st001            server node
 XIV7820000_SONAS_2 AC1F86024EF213C4   /dev/mapper/XIV7820000_SONAS_2 strg001st001            server node
 XIV7820000_SONAS_2 AC1F86024EF213C4   /dev/mapper/XIV7820000_SONAS_2 strg002st001            server node
 XIV7820000_SONAS_3 AC1F86014EF213CC   /dev/mapper/XIV7820000_SONAS_3 strg001st001            server node
 XIV7820000_SONAS_3 AC1F86014EF213CC   /dev/mapper/XIV7820000_SONAS_3 strg002st001            server node
 XIV7820000_SONAS_4 AC1F86024EF213D4   /dev/mapper/XIV7820000_SONAS_4 strg001st001            server node
 XIV7820000_SONAS_4 AC1F86024EF213D4   /dev/mapper/XIV7820000_SONAS_4 strg002st001            server node
 XIV7820000_SONAS_5 AC1F86014EF213DA   /dev/mapper/XIV7820000_SONAS_5 strg001st001            server node
 XIV7820000_SONAS_5 AC1F86014EF213DA   /dev/mapper/XIV7820000_SONAS_5 strg002st001            server node
 XIV7820000_SONAS_6 AC1F86024EF213E2   /dev/mapper/XIV7820000_SONAS_6 strg001st001            server node
 XIV7820000_SONAS_6 AC1F86024EF213E2   /dev/mapper/XIV7820000_SONAS_6 strg002st001            server node
 XIV7820000_SONAS_7 AC1F86014EF213E8   /dev/mapper/XIV7820000_SONAS_7 strg001st001            server node
 XIV7820000_SONAS_7 AC1F86014EF213E8   /dev/mapper/XIV7820000_SONAS_7 strg002st001            server node
 XIV7820000_SONAS_8 AC1F86024EF213EE   /dev/mapper/XIV7820000_SONAS_8 strg001st001            server node
 XIV7820000_SONAS_8 AC1F86024EF213EE   /dev/mapper/XIV7820000_SONAS_8 strg002st001            server node
```

Figure 3-34 Sample CLI output for the mmlsnsd -M command

As shown in Figure 3-34, the **mmlsnsd** command must show device trees as the same for each NSD device on each of the Storage node pairs. In rare cases where they do not match, the issue must be corrected before you put those devices into an active file system. In many cases, a Storage node restart can fix issues.

Example 3-1 shows a **mkfs** command for evenly distributing the NSDs and Storage node preference during the creation of the gpfs0 file system when the two XIVs are connected to the same Storage node pair. It is a preferred practice consideration for two XIVs in the same pod.

Example 3-1 A mkfs command to distribute evenly the NSDs and Storage node preference

```
cli mkfs gpfs0 /ibm/gpfs0 -F
XIV6000095_SONAS_1,XIV6000095_SONAS_2,XIV7802005_SONAS_1,XIV7802005_SONAS_2 -R
meta -j cluster”
```

For now, consider the **-j** option in a GPFS **mkfs** command (for the cluster allocation type).

Data allocation types

GPFS offers two basic data and block allocation types: Scatter and Cluster. *Scatter* disperses I/O randomly across all NSDs, and *Cluster* lays it down in contiguous stripes across the NSDs.

Well-defined XIV NSDs typically perform better when the Cluster allocation type is used in file system creation for highly sequential workloads, and possibly scatter is best for small file averages in highly random workloads. Current data also suggests that the preferred file system block allocation, from a performance standpoint, is 256 KB or 1 M. The 256 KB size is the SONAS and GPFS default setting and preferred when average file sizes are less that 256 KB. However, 1 M aligns with XIV back-end RAID chunks and works best when average file sizes are near 1 MB or higher.

Tip: Some use cases (which are not yet fully analyzed) might serve I/O faster in larger block sizes against XIV storage. Avoid using blocks larger than 1 M for file systems unless specific preproduction testing concludes that it improves performance and usage.

Note: The block size and block allocation map type are defined when the file system is created and cannot be changed without deleting and re-creating the file system. This process requires you to back up and restore the data.

SONAS file system subblocks are 1/32 of the block size. So, the subblock size of a 256 KB file system block size is 8 KB, and the subblock size of a 1 M file system block size is 32 KB. The minimum subblock allocation is the subblock size that is based on the defined block size for that specific file system and any file set within that file system (dependent or independent).

The command in Example 3-1 on page 131 stripes the file system across the first NSD from the first XIV, then the second NSD of the same XIV, before it goes to the first NSD of the first XIV, then the second NSD of the second XIV. It stripes in alternating patterns across the Storage node pairs. If you do not use three NSDs, it is easy to see that one Storage node is processing more work than the other. It might not be a problem if there are many NSDs in a file system; make sure that you understand the example before you commit your configuration.

When each XIV is in a separate Storage node pair, the preferred practice configuration is slightly different. Figure 3-35 illustrates that difference.

```
[root@xivsonas.mgmt001st001 bin]# mmlsnsd -L
File system Disk name NSD servers
---------------------------------------------------------------------------
gpfs0 XIV6000095_SONAS_1 strg001st001,strg002st001
gpfs0 XIV7802005_SONAS_1 strg003st001,strg004st001
gpfs0 XIV6000095_SONAS_2 strg002st001,strg001st001
gpfs0 XIV7802005_SONAS_2 strg004st001,strg003st001
gpfs0 XIV6000095_SONAS_3 strg001st001,strg002st001
gpfs0 XIV7802005_SONAS_3 strg003st001,strg004st001
gpfs0 XIV6000095_SONAS_4 strg002st001,strg001st001
gpfs0 XIV7802005_SONAS_4 strg004st001,strg003st001
```

Figure 3-35 Multi XIV output from the mmlsnsd command that shows a two-pod node preference

As shown in Figure 3-35, the configuration is complex at first, depending on the scale of the solution. However, after the file system is created, there is nothing else to do. You can set it and forget it. Figure 3-36 shows a closer view of the **mkfs** command for this configuration.

```
cli mkfs gpfs0 /ibm/gpfs0 -F
XIV6000095_SONAS_1,XIV6000095_SONAS_1,XIV7802005_SONAS_1,XIV6000095_SONAS_2,XIV
7802005_SONAS_2,XIV6000095_SONAS_3,XIV7802005_SONAS_3,XIV6000095_SONAS_4,XIV780
2005_SONAS_4 -R meta -j cluster
```

Figure 3-36 A mkfs command to stripe across NSD and Storage node pairs

This example stripes evenly across NSDs and Storage node pairs to provide the highest level of dispersion in the file system stripe that is possible. It is a preferred practice consideration in NSD mapping for file system creation. If you are using *scatter* as your GPFS allocation type, the layout does not provide the planned balance, so it becomes less valuable. Plan the GPFS allocation type carefully.

Replicating metadata

In Figure 3-36 on page 132, **`-R meta`** is specified. It indicates that metadata is replicated across the NSD Failure groups. In fact, **`meta`** is the default replication setting. In this case, you do not need to specify it if you want to use this setting. However, it is assumed that you placed both XIVs into separate failure groups. There is little value in having more than two failure groups. However, when you have more than one XIV subsystem, this configuration allows you to protect the status of all metadata if you lose an XIV subsystem. If you have 10 XIVs subsystems, you can set up five XIV subsystems in Failure group 1 and the other five XIV subsystems in Failure group 2. It is a preferred practice consideration for reliability. However, replication reduces performance. It is acceptable to create your XIV based file systems with **`-R none`** (replication set to none).

The options for failure group replication are as follows:

```
-R { none | meta | all }
```

- **`none`**, which means no replication at all.
- **`meta`**, which indicates that the file system metadata is synchronously mirrored across two failure groups.
- **`all`**, which indicates that the file system data and metadata is synchronously mirrored across two failure groups.

The file system in this scenario is created in a logical, balanced fashion. However, you want to add NSDs to the file system.

Completing the SONAS installation

Because cluster networking and authentication are not dependent on the underlying storage, these topics are described later in the chapter.

Read the entire chapter to understand network installation process and preferred practices for your cluster installation. Starting with 3.8, “SONAS integration into your network” on page 218, it covers detailed information about general SONAS and GPFS storage configuration information relative to any back-end storage, along with networking and other considerations that are not back-end storage specific. Read through the remainder of this chapter in its entirety to capture all installation considerations before you plan for integrating SONAS into your environment.

3.4.4 Adding XIV LUNs to an existing SONAS configuration

As an example, assume that the file system is created in a logical, balanced way, and you want to add NSDs to the file system. This section provides an overview of this process followed by detailed instructions with screen captures.

Creating and selecting the volumes to be added

On the XIV subsystems, create the volumes (4 TB or 4002 GB) from the SONAS storage pool to add to the SONAS cluster as follows:

1. Expand the SONAS storage pool (right-click and resize) if necessary and create the volumes from the expended XIV storage pool. Be sure to use the same volume size and name convention and continue the numerical sequence from the last or previous volume that you created for SONAS.
2. From the Volumes list, select the new volumes for the SONAS solution, right-click, and map the selected volumes to the SONAS cluster.

Summary of adding NSDs to SONAS

The process to add NSDs to SONAS is as follows:

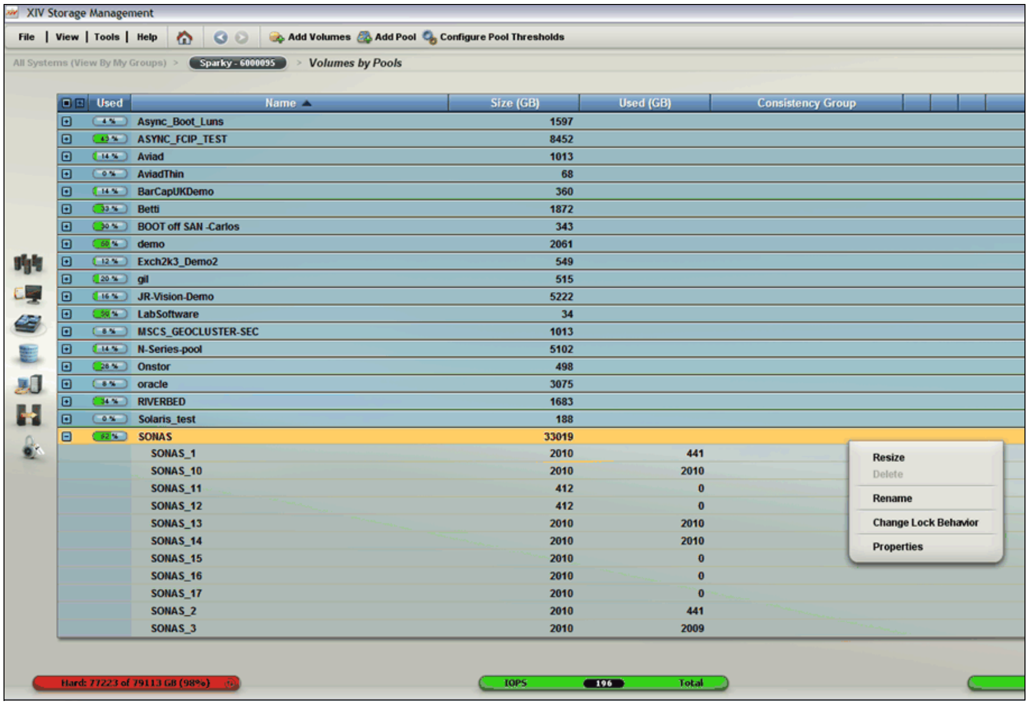
1. When you add disks, or there is a need to rescan for new storage, run **`mkdisk --luns`** (SONAS V1.4.1 and later) as the admin or root user. This command rescans all available Storage nodes for newly allocated LUNs.

   An alternative method is to run **`cnaddstorage`** as the root user. This command presents a short list of Storage node pairs from which to select.

   To use this method, select the Storage node pair to which you added storage. SONAS automatically discovers new volumes, and creates the multipath devices and the associated NSDs. The command must be run for each Storage node pair or pod. This process takes up to six minutes to complete.
2. On completion, you can view the NSDs by running **`lsdisk -r`**. They can be used for file system creation or expansion.

Detailed XIV work

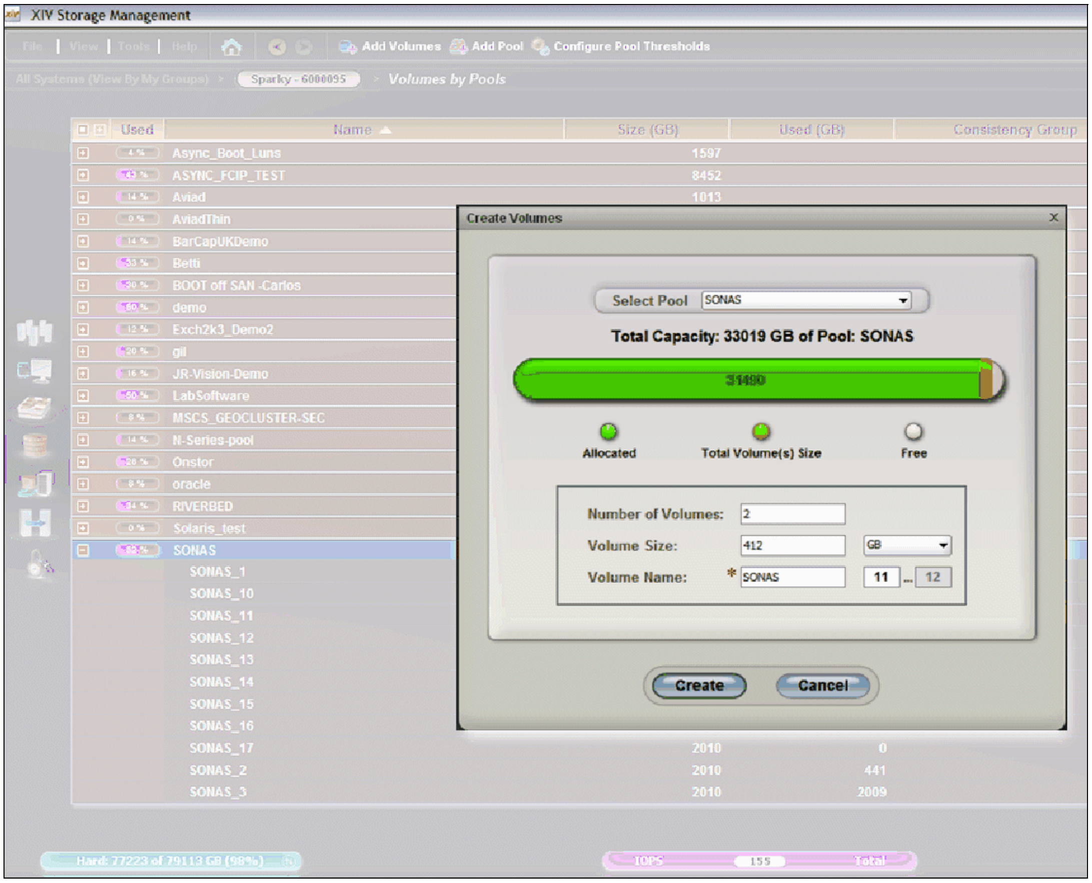
Ensure that your SONAS Storage Pool (on the XIV) has enough capacity for the volumes that you want. Figure 3-37 shows the GUI display of the Volume List.


XIV Storage Management
File | View | Tools | Help
Add Volumes
Add Pool
Configure Pool Thresholds
All Systems (View By My Groups) > Sparky - 6000095 > Volumes by Pools
Resize
Delete
Rename
Change Lock Behavior
Properties
Hard: 77223 of 79113 GB (98%)
IOPS
196
Total

Used	Name ▲	Size (GB)	Used (GB)	Consistency Group
4 %	Async_Boot_Luns	1597		
43 %	ASYNC_FCIP_TEST	8452		
14 %	Aviad	1013		
0 %	AviadThin	68		
14 %	BarCapUKDemo	360		
33 %	Betti	1872		
30 %	BOOT off SAN -Carlos	343		
60 %	demo	2061		
12 %	Exch2k3_Demo2	549		
20 %	gil	515		
16 %	JR-Vision-Demo	5222		
50 %	LabSoftware	34		
8 %	MSCS_GEOCLUSTER-SEC	1013		
14 %	N-Series-pool	5102		
28 %	Onstor	498		
8 %	oracle	3075		
34 %	RIVERBED	1683		
0 %	Solaris_test	188		
62 %	SONAS	33019		
	SONAS_1	2010	441	
	SONAS_10	2010	2010	
	SONAS_11	412	0	
	SONAS_12	412	0	
	SONAS_13	2010	2010	
	SONAS_14	2010	2010	
	SONAS_15	2010	0	
	SONAS_16	2010	0	
	SONAS_17	2010	0	
	SONAS_2	2010	441	
	SONAS_3	2010	2009	

Figure 3-37 XIV GUI - volume list from the SONAS storage pool

The next step is to size, name, and create the volumes, as shown in Figure 3-38.


XIV Storage Management
File | View | Tools | Help
Add Volumes
Add Pool
Configure Pool Thresholds
All Systems (View By My Groups) > Sparky - 6000095 > Volumes by Pools
Create Volumes
Allocated
Total Volume(s) Size
Free
Create
Cancel
Hard: 77223 of 79113 GB (98%)
IOPS
Total

Select Pool: SONAS

Total Capacity: 33019 GB of Pool: SONAS

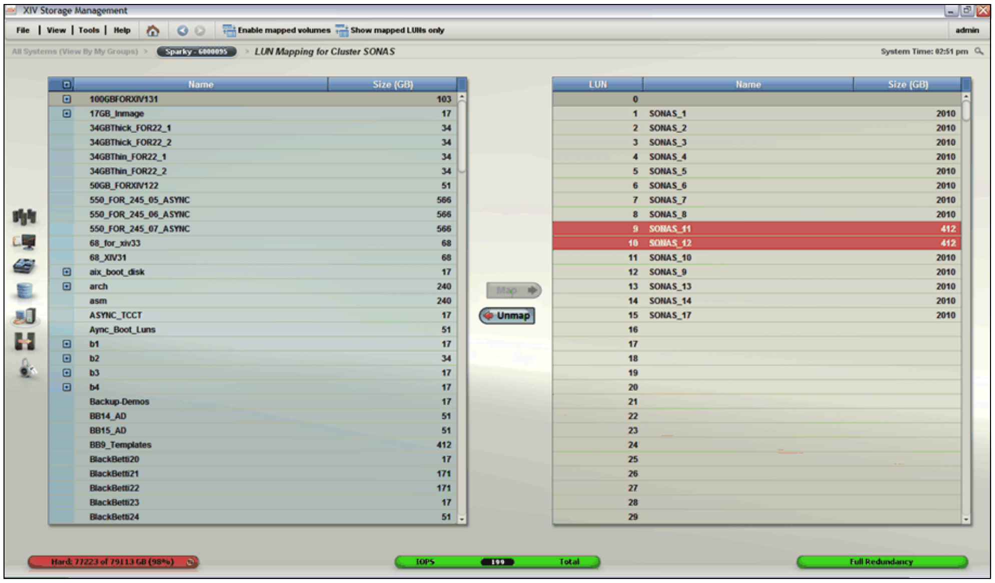
Number of Volumes: 2

Volume Size: 412 GB

Volume Name: * SONAS 11 ... 12

Figure 3-38 Select the number, size, and name of the volumes to add

Next, map the XIV volumes to the SONAS cluster, as shown in Figure 3-39.


XIV Storage Management
File | View | Tools | Help
Enable mapped volumes
Show mapped LUNs only
admin
All Systems (View By My Groups) > Sparky - 6000095 > LUN Mapping for Cluster SONAS
System Time: 02:51 pm
Map
Unmap
Hard: 77223 of 79113 GB (98%)
IOPS
199
Total
Full Redundancy

Name	Size (GB)
100GBFORXIV131	103
17GB_Inmage	17
34GBThick_FOR22_1	34
34GBThick_FOR22_2	34
34GBThin_FOR22_1	34
34GBThin_FOR22_2	34
50GB_FORXIV122	51
550_FOR_245_05_ASYNC	566
550_FOR_245_06_ASYNC	566
550_FOR_245_07_ASYNC	566
68_for_xiv33	68
68_XIV31	68
aix_boot_disk	17
arch	240
asm	240
ASYNC_TCCT	17
Aync_Boot_Luns	51
b1	17
b2	34
b3	17
b4	17
Backup-Demos	17
BB14_AD	51
BB15_AD	51
BB9_Templates	412
BlackBetti20	17
BlackBetti21	171
BlackBetti22	171
BlackBetti23	17
BlackBetti24	51

LUN	Name	Size (GB)
0		
1	SONAS_1	2010
2	SONAS_2	2010
3	SONAS_3	2010
4	SONAS_4	2010
5	SONAS_5	2010
6	SONAS_6	2010
7	SONAS_7	2010
8	SONAS_8	2010
9	SONAS_11	412
10	SONAS_12	412
11	SONAS_10	2010
12	SONAS_9	2010
13	SONAS_13	2010
14	SONAS_14	2010
15	SONAS_17	2010
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		
27		
28		
29		

Figure 3-39 XIV GUI - map the volume to the SONAS cluster

The XIV work is complete.

Detailed instructions for SONAS side configuration

To configure the SONAS, complete the following steps:

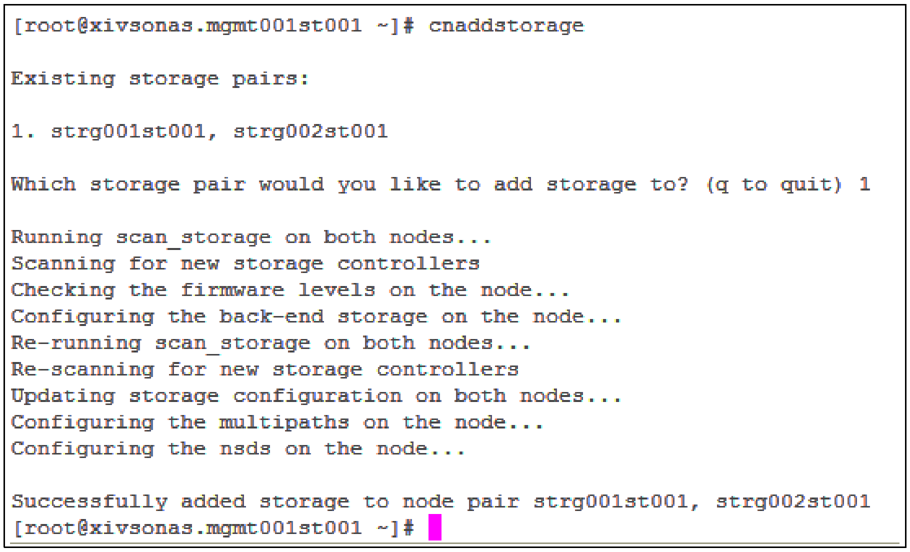
1. From the Management node in the SONAS cluster, run **`mkdisk --luns`** as the administrator user (see Figure 3-40). This command rescans for new storage on all available Storage nodes.

```
[xivsonas.xiv34.aviad]$ mkdisk --luns
(1/3) Scanning for new devices
(2/3) Creating NSDs
(3/3) Adding to database
Successfully created disks:
XIV6000095_SONAS_11
XIV6000095_SONAS_12
EFSSG1000I The command completed successfully.
[xivsonas.xiv34.aviad]$
```

Figure 3-40 The mkdisk command

An alternative method is to run **`cnaddstorage`** as the root user. This command presents a short list of Storage node pairs from which to select.

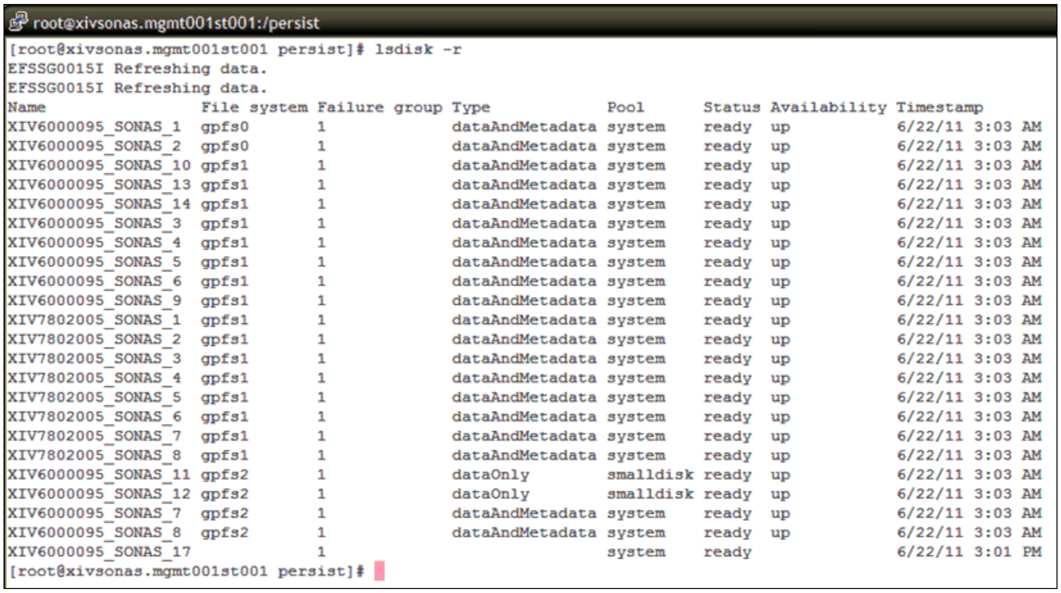
2. Select the Storage node pair to which you added storage. SONAS automatically discovers new volumes and creates the multipath devices and the associated NSDs. The command must be run for each Storage node pair or pod. This process can take up to six minutes to complete (see Figure 3-41).



```
[root@xivsonas.mgmt001st001 ~]# cnaddstorage

Existing storage pairs:

1. strg001st001, strg002st001

Which storage pair would you like to add storage to? (q to quit) 1

Running scan_storage on both nodes...
Scanning for new storage controllers
Checking the firmware levels on the node...
Configuring the back-end storage on the node...
Re-running scan_storage on both nodes...
Re-scanning for new storage controllers
Updating storage configuration on both nodes...
Configuring the multipaths on the node...
Configuring the nsds on the node...

Successfully added storage to node pair strg001st001, strg002st001
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-41 CLI - the cnaddstorage command

3. On completion, look for the devices by running **lsdisk -r** (see Figure 3-42).



```
[root@xivsonas.mgmt001st001 persist]# lsdisk -r
EFSSG0015I Refreshing data.
EFSSG0015I Refreshing data.
Name                File system Failure group Type            Pool      Status Availability Timestamp
XIV6000095_SONAS_1  gpfs0       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_2  gpfs0       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_10 gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_13 gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_14 gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_3  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_4  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_5  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_6  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_9  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_1  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_2  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_3  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_4  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_5  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_6  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_7  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_8  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_11 gpfs2       1             dataOnly        smalldisk ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_12 gpfs2       1             dataOnly        smalldisk ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_7  gpfs2       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_8  gpfs2       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_17             1                             system    ready               6/22/11 3:01 PM
[root@xivsonas.mgmt001st001 persist]#
```

Figure 3-42 CLI - output from lsdisk -r command sample

4. The additional disks/NSDs can now be added to an existing file system for more capacity, or be used to create one. Run **chdisk** to alter the disk parameters to the wanted values.

3.5 SONAS Gateway configuration with Storwize V7000 storage

The SONAS Gateway configuration with Storwize V7000 is listed as a feature code option when you order the SONAS gateway. Storwize V7000 Gateways are FC 9007. Much like the XIV storage (FC 9006), you can stack one or two Storwize V7000 controllers behind each SONAS storage pod. The Storwize V7000 storage with a SONAS Gateway offers the highest flexibility in storage configuration possibilities within SONAS solution offerings. You have the greatest flexibility with the number and types of drives that you can assemble in the SONAS cluster when you are using Storwize V7000 storage.

The highest-performance flexibility comes from the SONAS Storwize V7000 Gateway solution:

- This solution offers the greatest flexibility in disk speeds, types, and even drive size options in the IBM storage catalog. It allows for flexible storage tiering with the storage pod between SSD, SAS, and Near Line SAS storage options.
- The Storwize V7000 solution is also currently the only solution that accommodates support for an SSD tier of SONAS storage. It is a preferred practice solution for supporting the highest speed of device technology for metadata placement for GPFS (the underlying file system). XIV can be purchased with SSD, but it is not defined as a tier.

Note: SONAS V1.5.1 includes the following updates for Storwize V7000 storage:

- Support for Storwize V7000 Gen 2 2076-524 model, which can now include up to 20 expansion enclosures
- The ability to “intermix” Storwize V7000 Gen 1 with Gen 2 behind same or separate storage pods
- Ability to “intermix” Storwize V7000 (Gen 1 or Gen 2) with DDN appliance systems behind separate storage pods

3.5.1 Storwize V7000 software considerations

Storwize V7000 storage offers dependable data availability with over five nines availability (99.999%) with an average annual downtime of less than 5 minutes. The Storwize V7000 is packed with software features.

The following software features are supported for SONAS:

- Next Generation GUI
- CLI (use it for logical disk configuration to prevent selecting GUI defaults for some values)

The following software features are not supported for SONAS:

- IBM Easy Tier®
- IBM Real-time Compression™
- Storage virtualization (currently only supported on internal Storwize V7000 disks)
- Thin provision
- Clustering

3.5.2 Storwize V7000 hardware considerations

The Storwize V7000 consists of up to two controller-based enclosures and up to 18 expansion enclosures per SONAS storage pod. It can scale up to 480 disks and 720 TB raw capacity, per storage pod.

Enclosures available: Varieties

Storwize V7000 Gen 1 Enclosures are available in six varieties:

- Type 2076-112 = Control enclosure + 12 3.5" drives
- Type 2076-124 = Control enclosure + 24 2.5" drives
- Type 2076-212 = Expansion enclosure + 12 3.5" drives
- Type 2076-224 = Expansion enclosure + 24 2.5" drives
- Type 2076-312 = Control enclosure + 12 3.5" drives + 10 Gb iSCSI
- Type 2076-324 = Control enclosure + 24 2.5" drives + 10 Gb iSCSI

Storwize V7000 Gen 2 Enclosures are available in the following three varieties:

- Type 2076-524 = Control enclosure + 24 2.5" drives SFF Only
- Type 2076-12F = Storwize V7000 LFF Expansion Enclosure Model 12F
- Type 2076-24F = Storwize V7000 SFF Expansion Enclosure Model 24F

Storwize V7000 Gen 2 2076-524 Feature codes

This section lists the feature codes for the Storwize V7000 Gen 2 2076-524.

The following feature code is required: AHB1 8 Gb FC Adapter Pair (it is required, but not included in the default configuration)

The following feature codes are not required:

- AHB5 10 Gb Ethernet Adapter Pair
- AHC1 Compression Accelerator (Real-time Compression is not supported behind SONAS)
- AHCB Cache Upgrade (Cache upgrade is only supported for Real-time Compression, which is not supported)

Storwize V7000 offers both LFF and SFF 12 Gb SAS expansion enclosure models. The Storwize V7000 LFF Expansion Enclosure Model 12F supports up to twelve 3.5-inch drives, and the Storwize V7000 SFF Expansion Enclosure Model 24F supports up to twenty-four 2.5-inch drives. High-performance disk drives, high-capacity nearline disk drives, and flash (solid-state) drives are supported. Drives of the same form factor can be intermixed within an enclosure and LFF and SFF expansion enclosures can be intermixed within a Storwize V7000 system.

The controller-based enclosures contain two independent control units (referred to as canisters), which are based on SAN Volume Controller technology. They are clustered by an internal network, which uses the SAN Volume Controller clustering mechanism. Each enclosure also includes two PSUs. The control enclosure PSUs each house a battery pack that retains cached data if there is a power failure. The enclosures (shown in Figure 3-43) are interconnected with wide SAS cables (4 x 6 Gbps).

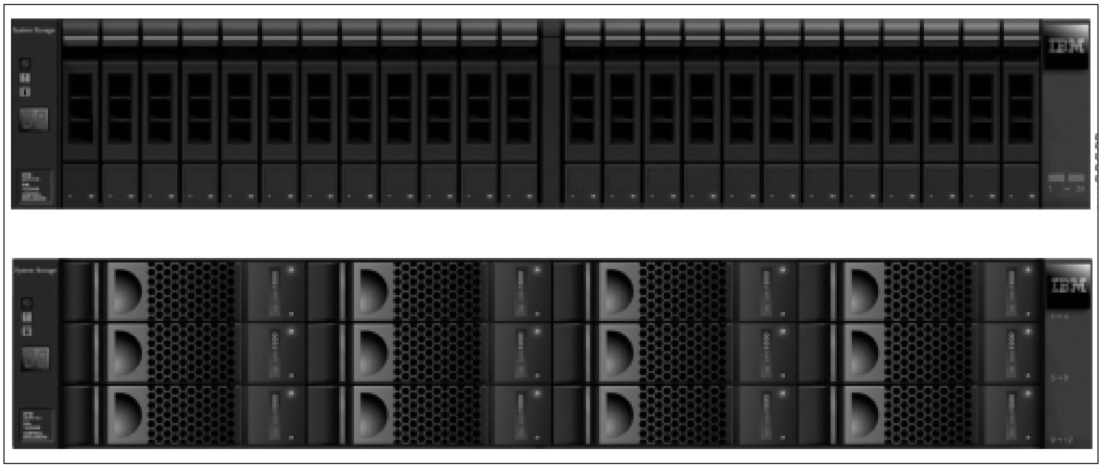


Figure 3-43 Two Storwize V7000 enclosure drive types (24 x 2.5" and 12 x 3.5" drives)

Figures 3-42 and 3-43 illustrate how the node canisters in the Storwize V7000 Gen 1 and Gen 2 differ and which Storage node cables go to which ports on the Storwize V7000 technologies.

Figure 3-44 shows a SONAS Storage node to Storwize V7000 Gen 1 cabling diagram.

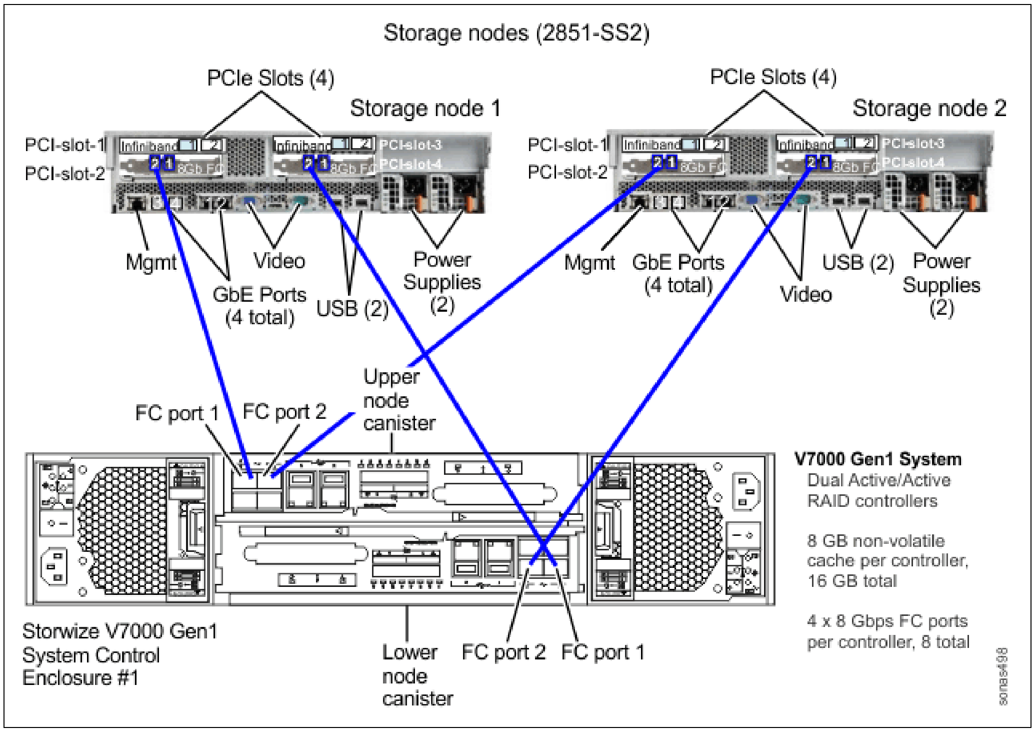


Figure 3-44 SONAS Storage node to Storwize V7000 Gen 1 cabling diagram shows node canisters on top and bottom

Figure 3-45 shows a SONAS Storage node to Storwize V7000 Gen 2 cabling diagram.

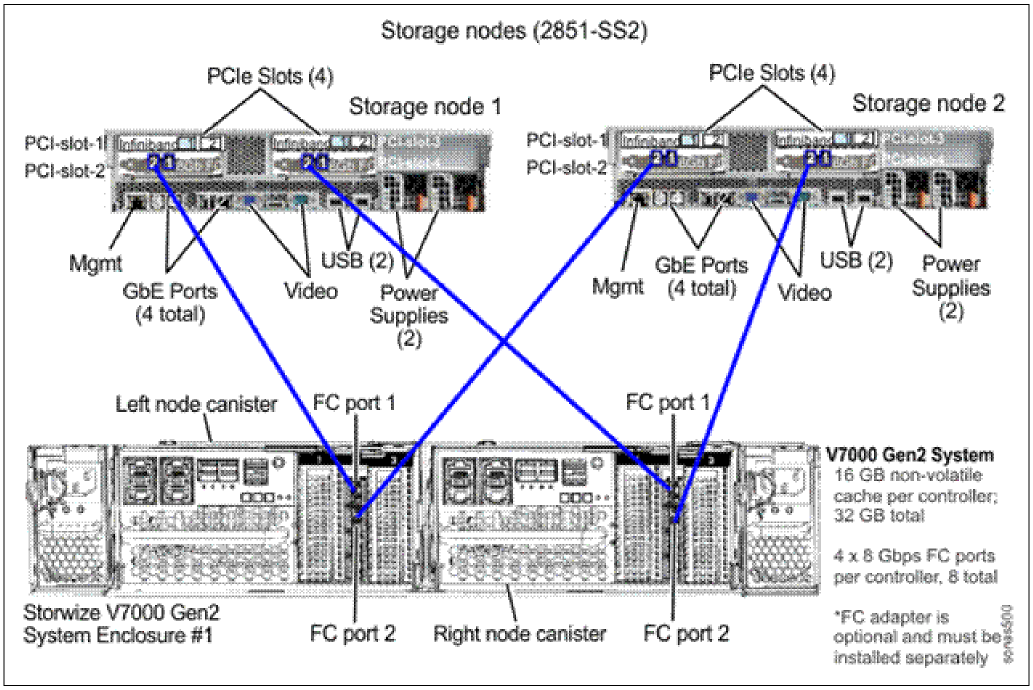


Figure 3-45 SONAS Storage node to Storwize V7000 Gen 2 cabling diagram shows node canisters side by side

Supported drives

The following 3.5-inch and 2.5-inch drives are supported in the Storwize V7000 enclosure:

- 3.5-inch disk drives: 2 TB, 3 TB, and 4 TB 7.2k Near-Line SAS disks
- 2.5-inch disk drives:
  - 146 GB, 300 GB, and 600 GB 15k SAS disks
  - 300 GB, 600 GB, 900 GB, and 1.2 TB 10k SAS disks
  - 200 GB, 400 GB, and 800 GB E-MLC (enterprise-grade multilevel cell) solid-state drive (SSD)
  - 1 TB 7.2k Near-Line SAS disk

Note: All hard disk drives (HDD) types and sizes and all SSDs supported by Storwize V7000 are supported when attached to an IBM SONAS Gateway system. For the latest supported drive sizes and types, see the Storwize V7000 documentation.

Available RAID levels

The following RAID levels are available:

- RAID 0 (Not supported under SONAS)
- RAID 1 (Not supported under SONAS)
- RAID 5 (striping, can survive one drive fault)
- RAID 6 (striping, can survive two drive faults)
- RAID 10 (RAID 0 on top of RAID 1)

RAID 0 arrays stripe data across the drives. The system supports RAID 0 arrays with just one member, which is similar to traditional JBOD attach. RAID 0 arrays have no redundancy, so they do not support hot spare takeover or immediate exchange. A RAID 0 array can be formed by 1 - 8 drives.

RAID 1 arrays stripe data over mirrored pairs of drives. A RAID 1 array mirrored pair is rebuilt independently. A RAID 1 array can be formed by two drives only.

RAID 5 arrays stripe data over the member drives with one parity strip on every stripe. RAID 5 arrays have single redundancy. The parity algorithm means that an array can tolerate no more than one member drive failure. A RAID 5 array can be formed by 3 - 16 drives.

RAID 6 arrays stripe data over the member drives with two parity stripes (known as the P-parity and the Q-parity) on every stripe. The two parity strips are calculated by using different algorithms, which give the array double redundancy. A RAID 6 array can be formed by 5 - 16 drives.

RAID 10 arrays have single redundancy. Although they can tolerate one failure from every mirrored pair, they cannot tolerate two-disk failures. One member out of every pair can be rebuilding or missing at the same time. A RAID 10 array can be formed by 2 - 16 drives.

Note: Storwize V7000 storage with the SONAS Gateway supports *only* RAID 5, 6, and 10.

3.5.3 SONAS with Storwize V7000 in the maximum optimal configuration (34-node example)

This example is the maximum configuration that is available with the SONAS Storwize V7000 Gateway solution. It consists of a combination of Interface and Storage nodes for a total of 34 nodes. This configuration provides 9.6 PB of raw SONAS storage:

- One SONAS Base Rack (2851-RXA)
  - Eight Interface nodes
  - Ten Storage nodes
- One Interface node expansion frame (2851-RXC)
  - Six Interface nodes
  - Ten Storage nodes
- Twenty Storwize V7000 controllers, each with nine Expansion enclosures with 12 x 4 TB drives in each enclosure.
  - Total raw capacity = 9.6 PB
  - Storwize V7000 Gen 2 supports up to 20 expansion enclosures, effectively doubling the raw capacity of a Gen 1 system

> **Note:** Due to multiple RAID configurations, only raw storage is listed.

- If SAN-attached: 160 Customer provided Fibre Channel Ports (four ports per Storwize V7000 (total of 80 Storwize V7000 ports), four ports per SONAS Storage node (total of 80 Storage nodes ports)). Fibre Channel switches must be in customer-supplied racks.
  - 80 ports on Fabric-A
  - 80 ports on Fabric-B
- If direct fiber connected, all available ports are available and cabled to the corresponding Storwize V7000 control enclosures. No extra SAN infrastructure required.

3.5.4 CIFS on Storwize V7000

The CIFS protocol is supported on Storwize V7000 when Storwize V7000 firmware V7.2 or higher is loaded. If the Storwize V7000 subsystem runs a firmware version earlier than Version 7.2, follow the instructions in “Creating a SCORE request to disable FNR for Storwize V7000 hardware with firmware versions earlier than Version 7.2” to request the proper Fast Node Restart (FNR) disablement with the SCORE process. (This situation rarely occurs with SONAS V1.4 or higher.)

Creating a SCORE request to disable FNR for Storwize V7000 hardware with firmware versions earlier than Version 7.2

This is process is not recommended. If possible, install Storwize V7000 firmware V7.2 or higher, which handles FNR internally and no additional code or settings are required.

If you continue to use firmware earlier than Version 7.2, ask your account team to create a SCORE request against Storwize V7000. The request must state that the account requires access to disable FNR to allow CIFS support for the IBM SONAS Gateway. The account team must also the level of Storwize V7000 software the customer currently has in the SCORE request, if known, or that the hardware is newly shipped from manufacturing. Always include a statement in the request to indicate how new the Storwize V7000 is.

After the request is approved, the account team receives instructions for how to download the FNR disabling package with a website link, user name, and password to use to access the package.

Tip: The account team must use the standard Storwize V7000 process to install the FNR disabling software, as documented in the Storwize V7000 published documentation.

After the FNR disable code is installed, FNR is disabled. No further action is required to disable FNR, or to verify whether it is disabled, unless new Storwize V7000 firmware is loaded. If new firmware is loaded, the SCORE request process, including creating a new SCORE statement, must be repeated.

The extended I/O timeouts that are described in the following section do not apply for IBM SONAS V1.3.1.x-xx with Storwize V7000 internal drives and FNR disabled.

Note: FNR is a feature of SAN Volume Controller and Storwize V7000 systems that is designed to maximize availability at the expense of longer recovery times for block hosts if there is a SAN Volume Controller/Storwize V7000 software failure condition (for example, code assertion). It functions by holding open an I/O to a failing node and controller canister long enough for the node and controller canister to do a “fast restart” and return online.

In this case, FNR for CIFS is not fast enough to be fully contained within the CIFS protocol timeout window of all CIFS hosts. Therefore, the FNR feature must be disabled to reliably send data over CIFS. After FNR is disabled, if a Storwize V7000 software failure occurs, an immediate multipath failover to the second controller canister occurs.

SONAS I/O timeouts with Storwize V7000 and external storage

SAN Volume Controller and Storwize V7000 together does not provide a guarantee for I/O blocking time because of a range of different vendors and IBM attached storage. Overall timeout causes I/O to complete, possibly with bad status, within six minutes. Users might have six minute or longer response times on NFS and CIFS connections. The CIFS protocol is less tolerable of storage timeouts than NFS and typically time out in 60 seconds. If a timeout does occur, I/O must be manually tried again if the host application does not automatically try again.

The SAN Volume Controller subsystem sends emergency Call Home requests on any occurrences greater than six minutes and IBM service must investigate the root cause. If these I/Os are GPFS metadata I/Os, multiple users might be affected. Assume that all SAN Volume Controller and Storwize V7000 L2 representatives can diagnose and repair this problem without recourse to IBM SONAS or GPFS development.

The failover timeouts must be set to accommodate these I/Os. Assume that the loss of any SONAS Interface node requires six minutes.

3.5.5 Detailed installation and configuration instructions for SONAS with Storwize V7000 storage

Storwize V7000 storage consists of a frame of disks that are grouped behind a dual set of clustered controllers.

Each Gen 1 controller can provide 8 GB or 16 GB of cache and Gen 2 can provide 64 TB of cache and manage 12 x 3.5” drives or 24 x 2.5” drives in the controller or in a Storwize V7000 expansion array. The Storwize V7000 arrays are linked together through redundant SAS cables. The Storwize V7000 Gen 1 controller can manage up to nine expansion arrays in each instance, with a minimum capacity of 120 drives (with 12 drive arrays) and a maximum capacity of 240 drives (with 24 drive arrays) per Storwize V7000. The Storwize V7000 Gen 2 controller can manage up to 20 expansion arrays in each instance, with a minimum capacity of 252 LFF drives (with 12 drive arrays) and a maximum capacity of 504 SFF drives (with 24 drive arrays) per Storwize V7000. The controller can have a mix, in between, depending on the disk size factor.

There are many disk configurations, including SSD, SAS, and Nearline SAS drives. In fact, the SONAS with Storwize V7000 Gateway solution offers the highest flexibility of storage types in the SONAS solution catalog.

Like the XIV solution, the SONAS frames can be partially populated. In fact, within the frame, even the arrays can be partially populated. For example, you can use the speed of SSD for use as the “metadata” container in the GPFS “system pool”. You might populate a Storwize V7000 controller or expansion frame with 8, 10, or 12 SSDs in a mirror (RAID 10) + a hot spare configuration and receive exceptional performance from it for metadata processing speed acceleration support. This configuration can improve I/Os per second to and from metadata intense operations, such as, backup, restore, async replication, and antivirus scanning.

Important: Consider all aspects of sizing for metadata. Sizing, performance, and reliability are important to the success of the configuration. Typically, eight, 10, or 12 SSDs are a minimum configuration for small file systems (analysis not considered). Also, with a RAID 10 mirrored configuration, it is commonly considered a preferred practice to allocate a hot spare for the RAID groups to reduce the likelihood of second drive failure in a mirrored set.

The Storwize V7000 subsystem itself offers virtualization of storage for both Internal and external storage. However, for the SONAS configuration, internal storage in the Storwize V7000 subsystem is the only configuration that is supported. Thick-provisioned volumes are required.

The SONAS Gateway with Storwize V7000 supports RAID 5, 6, and 10, and almost any disk type that Storwize V7000 supports is also supported for use in SONAS.

Optimal configuration of Storwize V7000 storage is more complicated than other storage types behind SONAS. It is thoroughly explained in 3.6, “SONAS Gateway configuration with DCS3700 storage” on page 174.

Ensuring that the Storwize V7000 storage is properly configured

Ensure that the Storwize V7000 storage is properly configured for power, cooling, Ethernet, and call home. This process is typically done by an IBM Certified Engineer.

Defining SONAS to Storwize V7000 storage

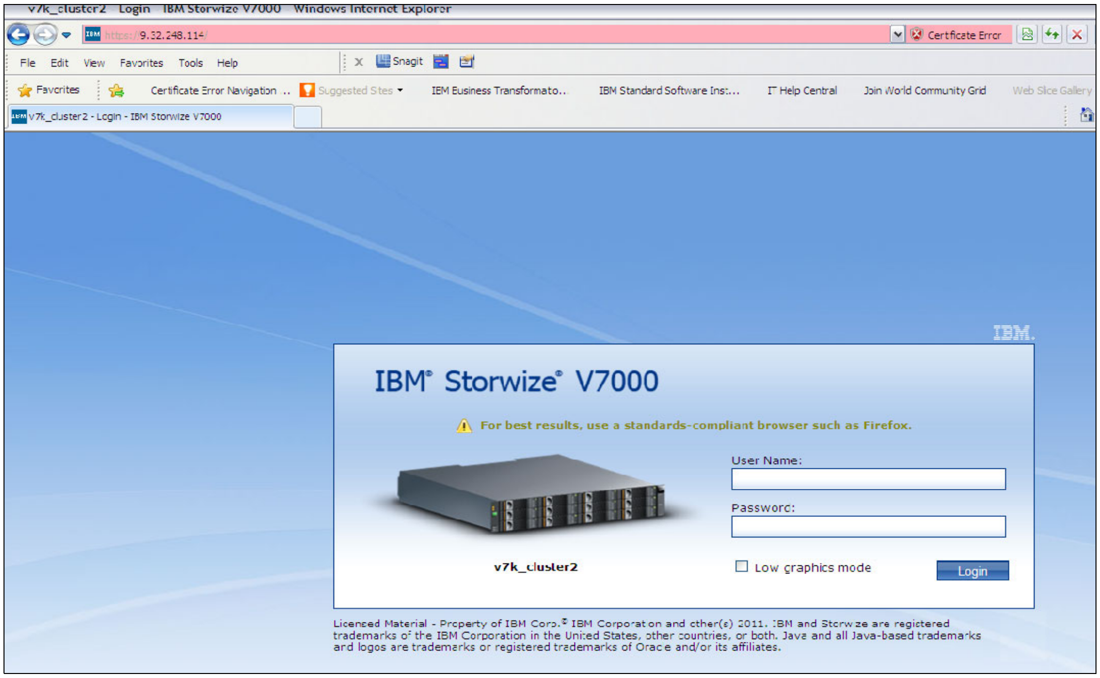
In the Storwize V7000 GUI, create the following things from the menu:

- One SONAS cluster
- Two SONAS hosts (Storage nodes) per Storwize V7000 subsystem
- Four WWPNs per SONAS host (ports)

It is easier to zone the SONAS Storage nodes in the fabric and to pick the WWPNs in Storwize V7000 GUI when the Storage nodes are powered on and the nodes are booted. That said, it can be done before or after the cluster software `first_time_install` script is run if the RPQ flag is set.

> **Note:** As of SONAS V1.4.1 The SONAS gateway solution with a Storwize V7000 storage configuration is supported through direct connect or SAN attach. This situation will not be reversed for clients currently on SAN, but new installs will have the opportunity to select the type of attachment of their choosing.

Figure 3-46 shows the Storwize V7000 GUI login.


v7k_cluster2 - Login - IBM Storwize V7000 - Windows Internet Explorer
https://9.32.248.114/
Certificate Error
File
Edit
View
Favorites
Tools
Help
Snagit
Favorites
Certificate Error Navigation ...
Suggested Sites
IBM Business Transformato...
IBM Standard Software Ins:...
IT Help Central
Join World Community Grid
Web Slice Gallery
v7k_cluster2 - Login - IBM Storwize V7000
IBM.
IBM® Storwize® V7000
User Name:
Password:
v7k_cluster2
□ Low graphics mode
Login

For best results, use a standards-compliant browser such as Firefox.

Licensed Material - Property of IBM Corp.© IBM Corporation and other(s) 2011. IBM and Storwize are registered trademarks of the IBM Corporation in the United States, other countries, or both. Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Figure 3-46 IBM Storwize V7000 GUI login

Figure 3-47 shows the Storwize V7000 GUI view of internal managed Storwize V7000 disks.

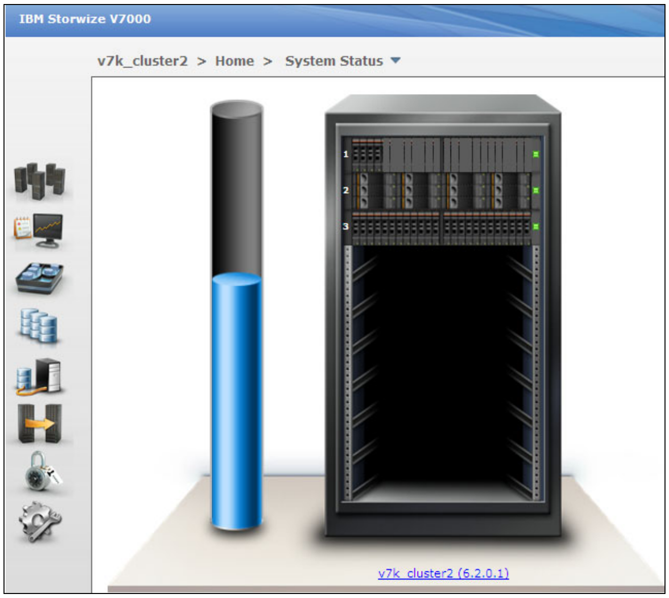


Figure 3-47 Storwize V7000 GUI - view of Internal managed Storwize V7000 disks

Understanding the Storwize V7000 storage configuration process

This section provides a high-level overview of the Storwize V7000 configuration process. More detailed instructions are provided later in this chapter. The following steps are required:

1. Position your hardware in the data center and ensure that all pre-installation requirements are reviewed and met.

2. Distribute the enclosure balance across the Storwize V7000 frames and controllers to establish the high distribution balance across the controllers.

   For example: You have two controllers and eight expansion enclosures with four drawers of SAS drives and four drawers of NL SAS drives. Balance the placement of those enclosures by putting two drawers of each type on each controller.

3. Install the Storwize V7000 hardware, update the software/firmware to a supported version (Version 6.3 or later), and configure Call Home support.

   Storwize V7000 installation follows the Storwize V7000 Installation Guide (found at `http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_r_publications.html?lang=en`). This process is not described in detail in this book. For more information, see the Storwize V7000 information in the IBM Knowledge Center, found at the following website:

   `http://www-01.ibm.com/support/knowledgecenter/ST3FR7/welcome?lang=en`

4. Connect your Ethernet cables to the Storwize V7000 and install or reset the access IP addresses, user name, and password by using the thumb drives and following the detailed process that is specified in the Storwize V7000 Installation Guide. Again, it is important to understand that it is managed separately from the SONAS.

5. Balance the disk technology placement within the Storwize V7000 frames, and create your MDisk groups from internal disks. Then, create your storage pools, balanced across the MDisk groups. Then, your volumes are created and balanced across your Storwize V7000 storage pools and mapped to both Storage node hosts. Figure 3-48 on page 148 shows this work flow.

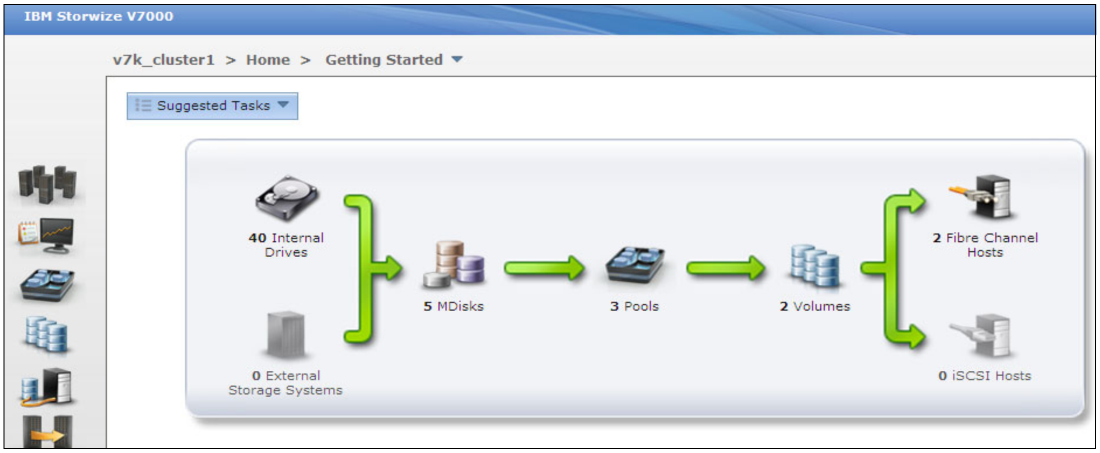


Figure 3-48 Storwize V7000 - Storage configuration work flow

6. Configure Storwize V7000 storage for SONAS with only “Internal Disk” (external disk is not supported for Storwize V7000 Gateway solutions in SONAS V1.3 or later).

   If possible, use the Storwize V7000 CLI to do the following logical configuration (and bypass the GUI default values):

   – Turn off Easy Tier.

   – Set a 128 KB stripe size.

   – Configure sequential versus striped vDisks. Eight data drives are good match for GPFS (eight data drives * 128 strip size = 1024 or a GPFS block size of 1 M), but 9+P+Q or 7+P+Q are not ideal. For optimal performance, avoid these configurations.

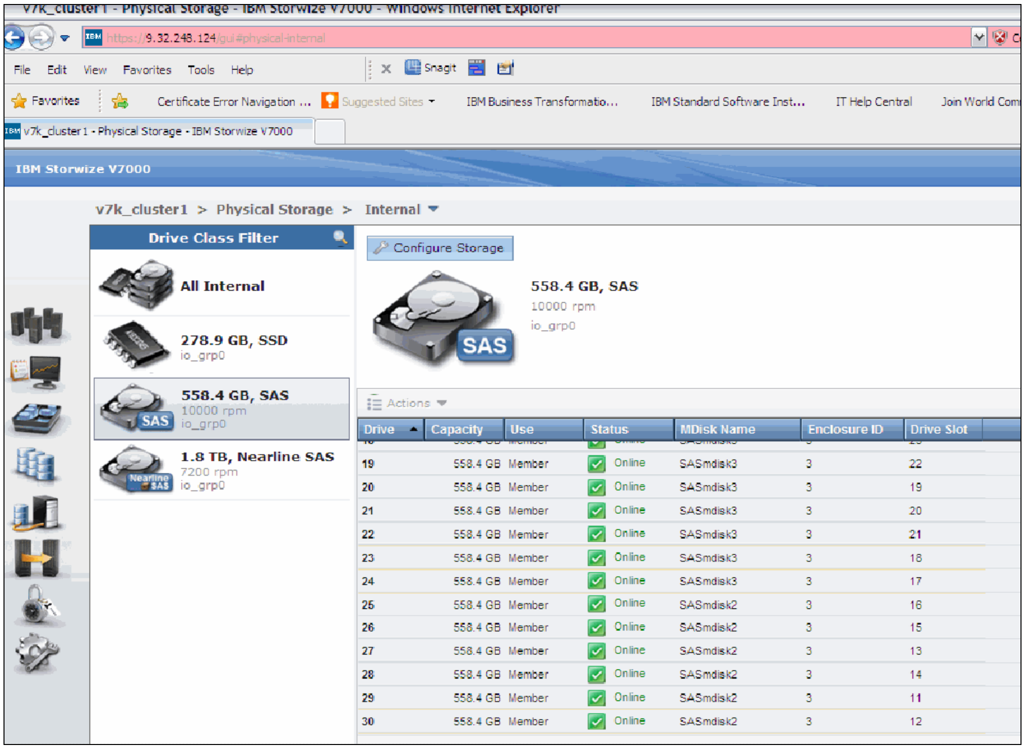
7. Remember, Easy Tier data migration from the Storwize V7000 subsystem is not supported in SONAS V1.3 or later.

8. Configure RAID groups. The “internal disk” is grouped by array or module (enclosure) of up to 12 or 24 drives, depending on the disk slot size (disk “slot size” = 2.5 in. or 3.5 in.). These disks are grouped into RAID groups (hot spares can be defined in those groups).

   – In many cases, you might choose to use RAID6 groups with no hot spares and fix failed drives on first drive failure, instead of choosing to spare them out and replace them later. In this case, a preferred practice is to use three RAID 6 groups of eight drives each in a 24 disk array.

   – In the default settings, the GUI MDisk creation stripes in 256 KB chunk sizes. When you crease these devices with CLI commands, you can change the chunk sizes. However, the non-default chunk size of 128 KB offers the optimal configuration for SONAS block size of 1 M (a solid configuration for most sequential dominant workloads).

Figure 3-49 shows internal physical storage from the Storwize V7000 GUI.



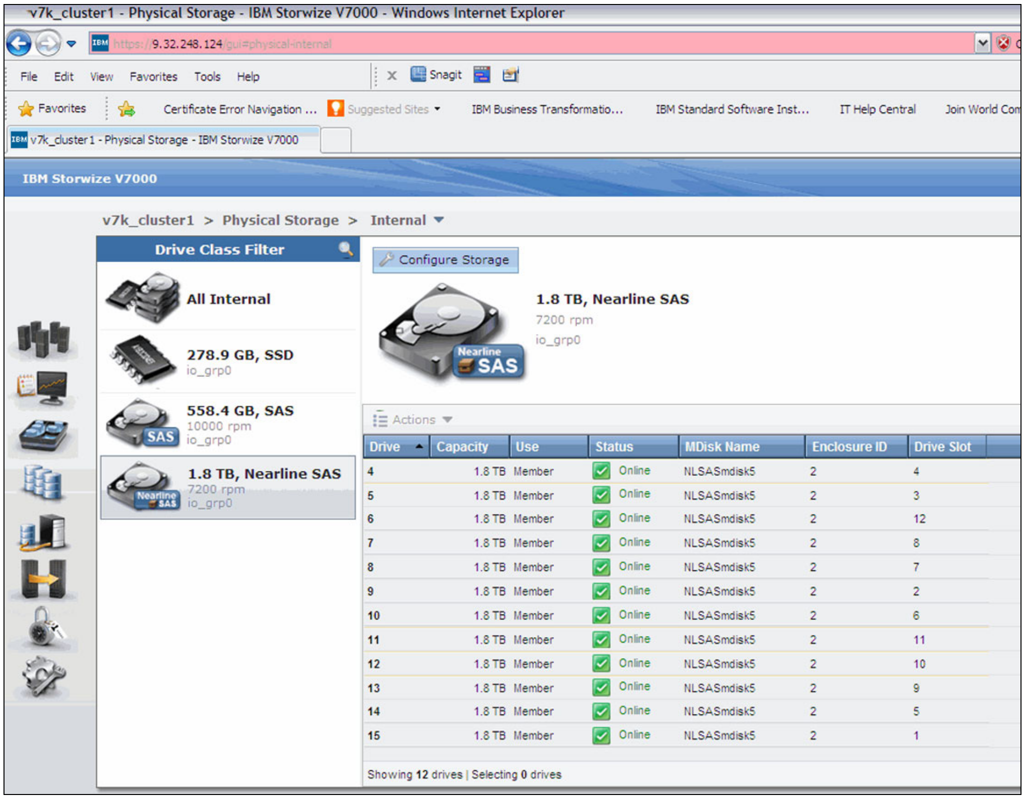
v7k_cluster1 - Physical Storage - IBM Storwize V7000 - Windows Internet Explorer
https://9.32.248.124/gui#physical-internal
File Edit View Favorites Tools Help
Favorites
Certificate Error Navigation ...
Suggested Sites
IBM Business Transformatio...
IBM Standard Software Inst...
IT Help Central
IBM Storwize V7000
v7k_cluster1 > Physical Storage > Internal
Drive Class Filter
All Internal
278.9 GB, SSD
io_grp0
558.4 GB, SAS
10000 rpm
io_grp0
1.8 TB, Nearline SAS
7200 rpm
io_grp0
Configure Storage
SAS
Actions

558.4 GB, SAS
10000 rpm
io_grp0

Drive	Capacity	Use	Status	MDisk Name	Enclosure ID	Drive Slot
19	558.4 GB	Member	✓ Online	SASmdisk3	3	22
20	558.4 GB	Member	✓ Online	SASmdisk3	3	19
21	558.4 GB	Member	✓ Online	SASmdisk3	3	20
22	558.4 GB	Member	✓ Online	SASmdisk3	3	21
23	558.4 GB	Member	✓ Online	SASmdisk3	3	18
24	558.4 GB	Member	✓ Online	SASmdisk3	3	17
25	558.4 GB	Member	✓ Online	SASmdisk2	3	16
26	558.4 GB	Member	✓ Online	SASmdisk2	3	15
27	558.4 GB	Member	✓ Online	SASmdisk2	3	13
28	558.4 GB	Member	✓ Online	SASmdisk2	3	14
29	558.4 GB	Member	✓ Online	SASmdisk2	3	11
30	558.4 GB	Member	✓ Online	SASmdisk2	3	12

Figure 3-49 Storwize V7000 GUI - image of 24 SAS drives per enclosure

– In Figure 3-50, you can see that there are three different disk types in the Storwize V7000. By selecting the internal disk view, you can see the enclosures and begin building your MDisk RAID groups.

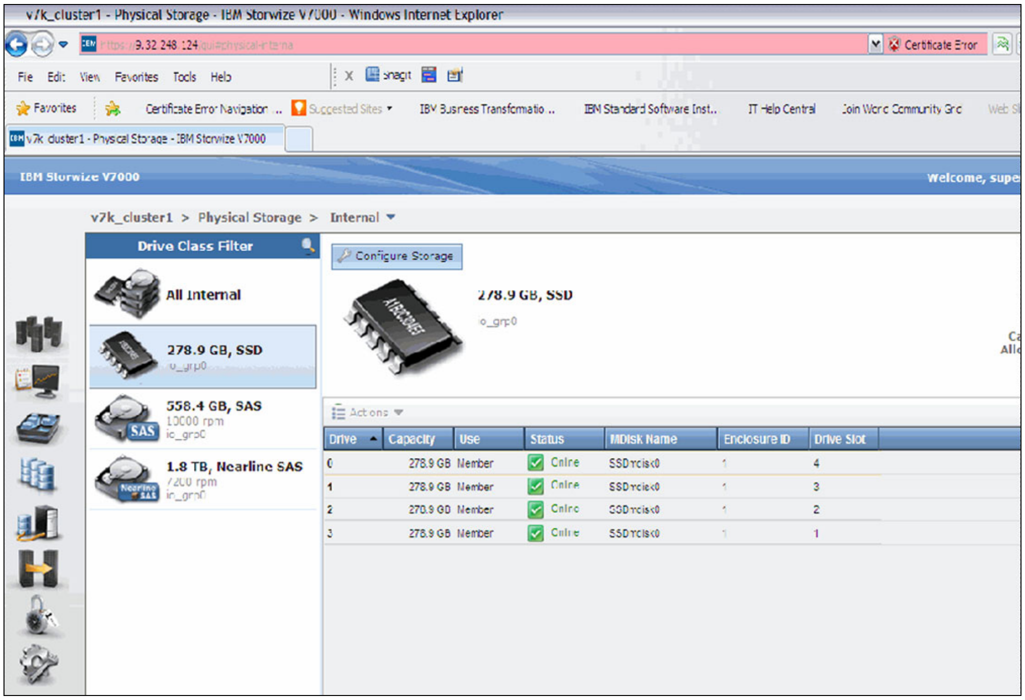

v7k_cluster1 - Physical Storage - IBM Storwize V7000 - Windows Internet Explorer
https://9.32.248.124/gui#physical-internal
File Edit View Favorites Tools Help
Snagit
Favorites
Certificate Error Navigation ...
Suggested Sites
IBM Business Transformatio...
IBM Standard Software Inst...
IT Help Central
Join World Com
v7k_cluster1 - Physical Storage - IBM Storwize V7000
IBM Storwize V7000
v7k_cluster1 > Physical Storage > Internal
Drive Class Filter
All Internal
278.9 GB, SSD
io_grp0
558.4 GB, SAS
10000 rpm
io_grp0
1.8 TB, Nearline SAS
7200 rpm
io_grp0
Configure Storage
1.8 TB, Nearline SAS
7200 rpm
io_grp0
Nearline SAS
Actions

Drive	Capacity	Use	Status	MDisk Name	Enclosure ID	Drive Slot
4	1.8 TB	Member	✓ Online	NLSASmdisk5	2	4
5	1.8 TB	Member	✓ Online	NLSASmdisk5	2	3
6	1.8 TB	Member	✓ Online	NLSASmdisk5	2	12
7	1.8 TB	Member	✓ Online	NLSASmdisk5	2	8
8	1.8 TB	Member	✓ Online	NLSASmdisk5	2	7
9	1.8 TB	Member	✓ Online	NLSASmdisk5	2	2
10	1.8 TB	Member	✓ Online	NLSASmdisk5	2	6
11	1.8 TB	Member	✓ Online	NLSASmdisk5	2	11
12	1.8 TB	Member	✓ Online	NLSASmdisk5	2	10
13	1.8 TB	Member	✓ Online	NLSASmdisk5	2	9
14	1.8 TB	Member	✓ Online	NLSASmdisk5	2	5
15	1.8 TB	Member	✓ Online	NLSASmdisk5	2	1

Showing 12 drives | Selecting 0 drives

Figure 3-50 Storwize V7000 GUI - image of twelve 2 TB NLSAS drives per enclosure

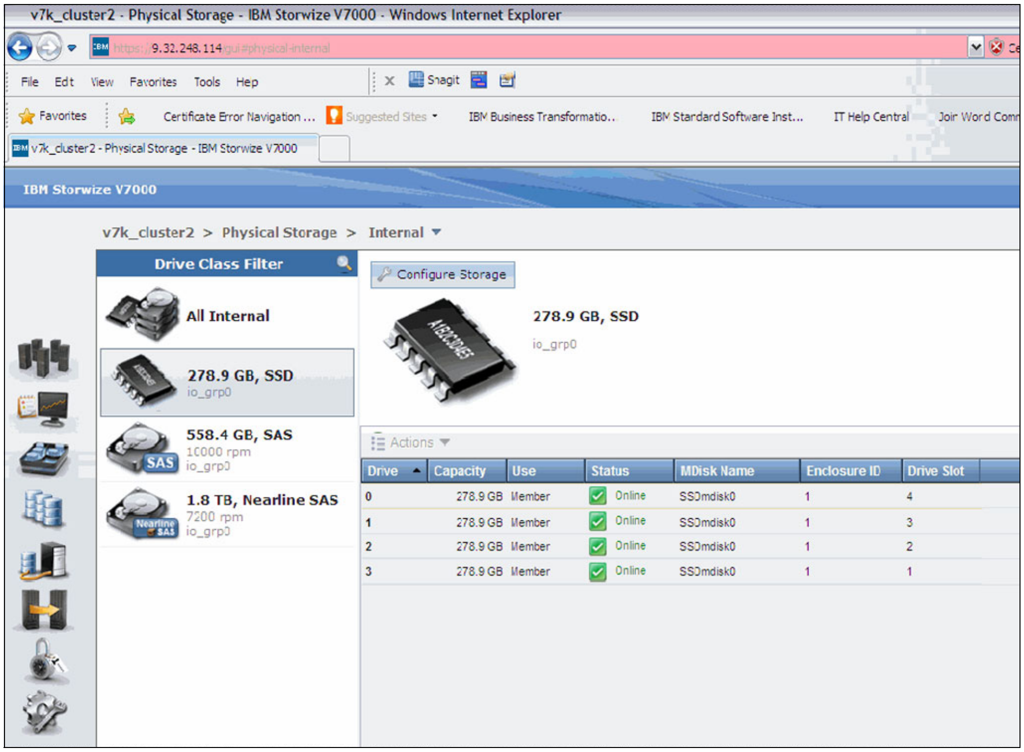
– Figure 3-51 on page 151 shows an enclosure with only a few SSD drives in it.


v7k_cluster1 - Physical Storage - IBM Storwize V7000 - Windows Internet Explorer
https://9.32.248.124/gui#physical-internal
Certificate Error
File Edit View Favorites Tools Help
Favorites
v7k_cluster1 - Physical Storage - IBM Storwize V7000
IBM Storwize V7000
Welcome, supe
v7k_cluster1 > Physical Storage > Internal
Drive Class Filter
All Internal
278.9 GB, SSD
io_grp0
558.4 GB, SAS
10000 rpm
io_grp0
1.8 TB, Nearline SAS
7200 rpm
io_grp0
Configure Storage
278.9 GB, SSD
io_grp0
Actions

Drive	Capacity	Use	Status	MDisk Name	Enclosure ID	Drive Slot
0	278.9 GB	Member	✓ Online	SSDmdisk0	1	4
1	278.9 GB	Member	✓ Online	SSDmdisk0	1	3
2	278.9 GB	Member	✓ Online	SSDmdisk0	1	2
3	278.9 GB	Member	✓ Online	SSDmdisk0	1	1

Figure 3-51 Storwize V7000 GUI - image of a partially populated Storwize V7000 enclosure

– When you create your MDisk, pools, and volumes, a preferred practice is to prefix the unit names with the type and or size of the disk technology. For example, you might use something like SASmdisk1, SASmdisk2, SASpool, SASvol1, or NLSAS2TBvol1, NLSAS3TBvol1, as shown in Figure 3-52.


v7k_cluster2 - Physical Storage - IBM Storwize V7000 - Windows Internet Explorer
https://9.32.248.114/gui#physical-internal
File Edit View Favorites Tools Help
Favorites
v7k_cluster2 - Physical Storage - IBM Storwize V7000
IBM Storwize V7000
v7k_cluster2 > Physical Storage > Internal
Drive Class Filter
All Internal
278.9 GB, SSD
io_grp0
558.4 GB, SAS
10000 rpm
io_grp0
1.8 TB, Nearline SAS
7200 rpm
io_grp0
Configure Storage
278.9 GB, SSD
io_grp0
Actions

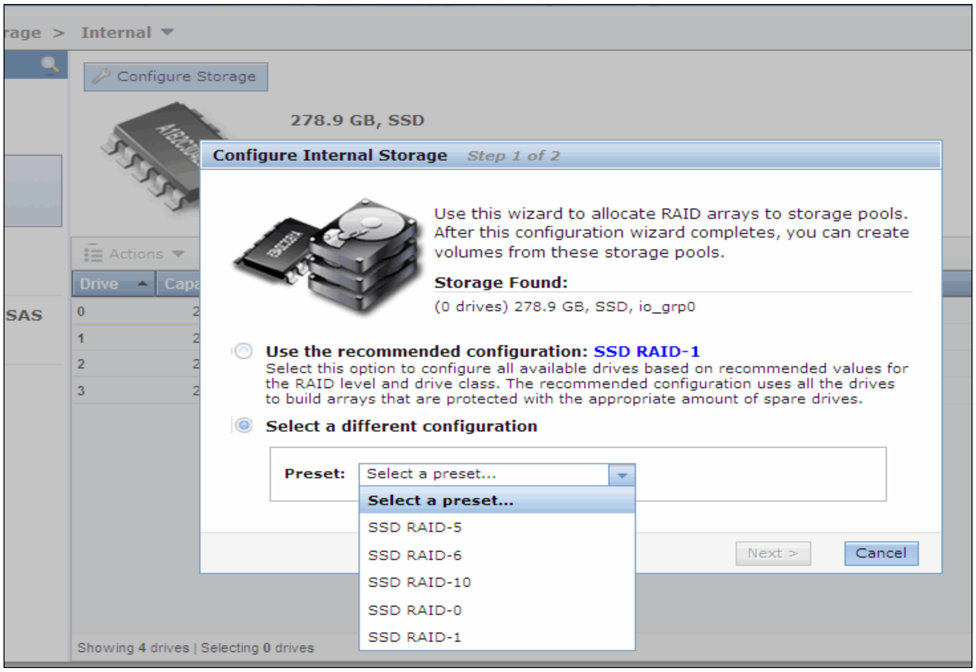
Drive	Capacity	Use	Status	MDisk Name	Enclosure ID	Drive Slot
0	278.9 GB	Member	✓ Online	SSDmdisk0	1	4
1	278.9 GB	Member	✓ Online	SSDmdisk0	1	3
2	278.9 GB	Member	✓ Online	SSDmdisk0	1	2
3	278.9 GB	Member	✓ Online	SSDmdisk0	1	1

Figure 3-52 Storwize V7000 GUI- image of storage that is preparing for MDisk group creation

9. After the MDisk groups are created, you can create a storage pool. A storage pool is a logical pool of available capacity that spans a set of like MDisk groups.

   – When you list the drives in the enclosure and click **Configure Storage**, you are given the option to use a recommended configuration or select a different configuration.

   – If you choose to select a different configuration, a drop-down list of supported RAID types for that disk set is shown (see Figure 3-53). You can choose to assign or not use hot spare devices in the configuration. You can simplify the choice by allowing the system to automatically configure it based on whether you are most interested in “capacity” or “performance”.



Configure Internal Storage *Step 1 of 2*

Use this wizard to allocate RAID arrays to storage pools. After this configuration wizard completes, you can create volumes from these storage pools.

Storage Found:

(0 drives) 278.9 GB, SSD, io_grp0

○ **Use the recommended configuration: SSD RAID-1**
Select this option to configure all available drives based on recommended values for the RAID level and drive class. The recommended configuration uses all the drives to build arrays that are protected with the appropriate amount of spare drives.

● **Select a different configuration**

Preset: Select a preset...

Figure 3-53 Storwize V7000 GUI - MDisk RAID set creation sequence

For example, you have 24 SAS drives in an enclosure and choose to create RAID 6 with no hot spares in eight drive groups (six drives for data and two drives for parity). You have three MDisk groups in that enclosure. Now, consider that you have three enclosures configured this way. You now have nine MDisk groups that are the same. You can then create a storage pool that spans those nine MDisk groups (within those three enclosures) and the use that pool for creating the SAS-based volumes.

   – After the MDisk groups are all configured, you can create storage pools across all like MDisk groups.

   – Then, create volumes from those storage pools. When you create volumes, it is a preferred practice to ensure that all volume sizes from any type of disk pool are of equal size, and provisioned in even quantities.

For example, you can provision an even number of 1 TB volumes from your NLSAS_pool to your Storage nodes. You can provision an even number of 275 GB volumes from your SSD_Pool. Using an even number of volumes helps ensure that you have balanced Storage node workload behavior. If you use seven volumes, you can have Storage node 1 working harder (with four Storage node preferred volumes) than Storage node 2 (with only three Storage node preferred volumes).

- After the volumes are provisioned, you can map those volumes to both Storage nodes in the pod. After all the volumes are provisioned and mapped to the SONAS cluster, you are finished with the Storwize V7000 GUI (until you later decide to add more storage to SONAS or when there is a problem in the Storwize V7000 hardware or software stack.
- Select all the volumes that you want to map and add them to the hostmap list and click **Apply** for each of the Storage nodes.

All of the Storwize V7000 work is done. You are ready to proceed with SONAS code load.

SONAS code load and cluster installation

Complete the following steps:

1. Insert the latest GA SONAS Release DVD into the Management node and power on. This process is done while all other nodes (Interface and Storage nodes) remain off. Install the OS and software on the Management node (int001st001 for SONAS V1.3 and later).

   The SONAS software is obtained and installed by the IBM Service Representative (IBM PFE).

   If the installer resets the nodes to manufacturing defaults by running the `manufacturing_cleanup` script, all nodes are shut down again, and only the Management node is powered on to begin the new SONAS code installation.

   > **Tip:** Installation takes approximately 35 - 45 minutes to complete, and it restarts several times until it completes. The initial installation is done when the DVD is ejected and the Management node allows a system login.

2. After the operating system is installed on the primary Management node, the installer runs the following command to set the Gateway flag on the installation:

   ```
   /opt/IBM/sonas/bin/mfg/cfg_gateway_rpq
   ```

   This flag tells the SONAS installation that Integrated storage and storage configuration must be run separately from the code installation.

   After it is run, a file set is called by running the following command:

   ```
   /opt/IBM/sonas/etc/mfg_gateway_rpq
   ```

   Removing the file removes the flag.

   After the pod is initially configured with Gateway storage to add storage to it, a different process (which is explained later in this chapter) is used.

   After completing this installation, the IBM authorized service provider begins the cluster installation process. During this process, the `first_time_install` script is run.

First-time installation process

To begin the first-time installation process, complete the following steps:

1. Log in to the SONAS Management node (mgmt001st001) as root.
2. Change to the /opt/IBM/sonas/bin by running the following command:

   ```
   cd /opt/IBM/sonas/bin
   ```

3. Verify the version of SONAS by running the following command:

   ```
   /opt/IBM/sonas/bin/get_version
   ```

   Validate that the version that is installed is the version that you want.

4. Run the `first_time_install` script and follow the prompts (for information about creating the initial cluster, see the *IBM SONAS Installation Guide*, GA32-0715 (with real data) and the *SONAS Introduction and Planning Guide*, GA32-0716) by running the following command:

```
/opt/IBM/sonas/bin/first_time_install
```

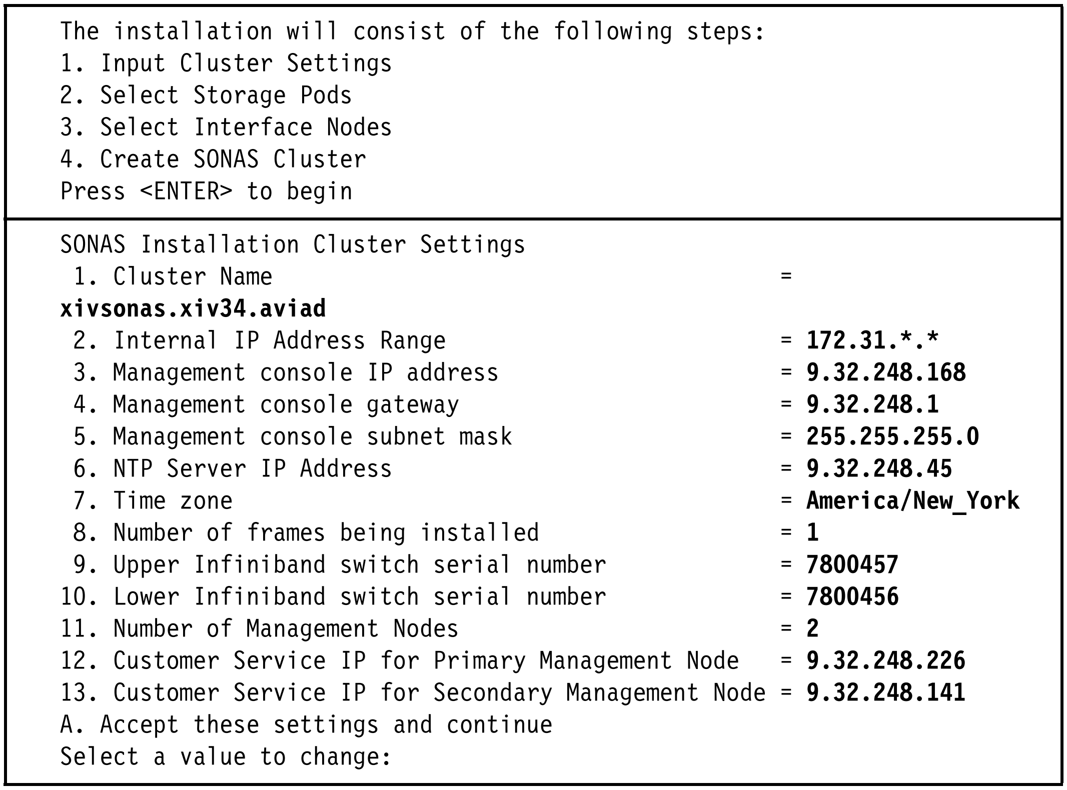
Defining the cluster parameters

To define the cluster parameters, complete the following steps:

1. Use the installation planning worksheet to answer the questions that are related to first-time configuration of the cluster and Management node information.

   The initial steps require you to provide the configuration information. See the example in Figure 3-54. The client provides this information before the installation is started. For the information that is needed, see the planning tables in Chapter 1, “Installation planning” on page 1.

2. After all the parameters are entered and verified, enter “A” at the prompt to continue. Figure 3-53 is from a `first_time_install` script for code that is earlier than the Version 1.5.1 code. SONAS V1.5.1 code includes two more settings, and provides a total of 15 cluster configuration settings.

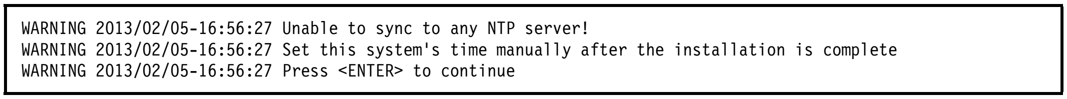


```
The installation will consist of the following steps:
1. Input Cluster Settings
2. Select Storage Pods
3. Select Interface Nodes
4. Create SONAS Cluster
Press <ENTER> to begin
```

```
SONAS Installation Cluster Settings
 1. Cluster Name                                       =
xivsonas.xiv34.aviad
 2. Internal IP Address Range                          = 172.31.*.*
 3. Management console IP address                      = 9.32.248.168
 4. Management console gateway                         = 9.32.248.1
 5. Management console subnet mask                     = 255.255.255.0
 6. NTP Server IP Address                              = 9.32.248.45
 7. Time zone                                          = America/New_York
 8. Number of frames being installed                   = 1
 9. Upper Infiniband switch serial number              = 7800457
10. Lower Infiniband switch serial number              = 7800456
11. Number of Management Nodes                         = 2
12. Customer Service IP for Primary Management Node   = 9.32.248.226
13. Customer Service IP for Secondary Management Node = 9.32.248.141
A. Accept these settings and continue
Select a value to change:
```

Figure 3-54 Example Management node data from the “first_time_install” script

As the script progresses, in some cases, especially where the management port is not connected to a network because the port shares the public network Ethernet ports, you might receive a warning message about NTP sync. The SONAS V1.5.1 `first_time_install` script updates are designed to reduce the occurrence of this message, so it might not occur. It is acceptable to receive this message. If you receive it, press Enter to continue, as shown in Figure 3-55 on page 155.



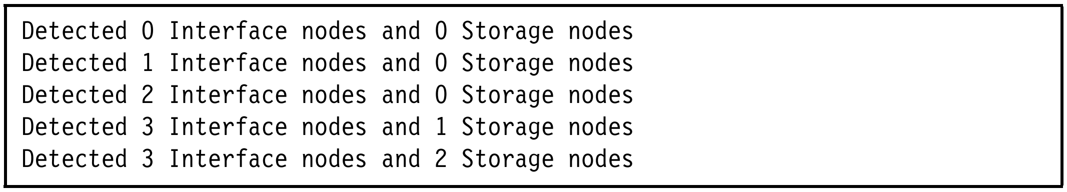
```
WARNING 2013/02/05-16:56:27 Unable to sync to any NTP server!
WARNING 2013/02/05-16:56:27 Set this system's time manually after the installation is complete
WARNING 2013/02/05-16:56:27 Press <ENTER> to continue
```

Figure 3-55 NTP sync warning

Powering on and discovering each node

To power on and discover each node, complete the following steps:

1. As the SONAS `first_time_install` script progresses, the script prompts you to **“**`power on all nodes`**”**. After you power on all nodes, you can press Enter from the `first_time_install` script prompt.
2. Nodes use the Preboot eXecution Environment (PXE, also known as Pre-Execution Environment) to boot and load the operating system image. The ISO image is transferred onto the detected nodes. A list of recognized nodes appears and is updated frequently. Nodes will show in the list as each is recognized and configured (Figure 3-56). It might take 60 minutes or longer, depending on the number of SONAS nodes in the configuration, to recognize and complete the SONAS code load on each node.
3. When all Interface nodes and Storage nodes are discovered (and quantity numbers keep repeating with no change), press Enter to continue the configuration.



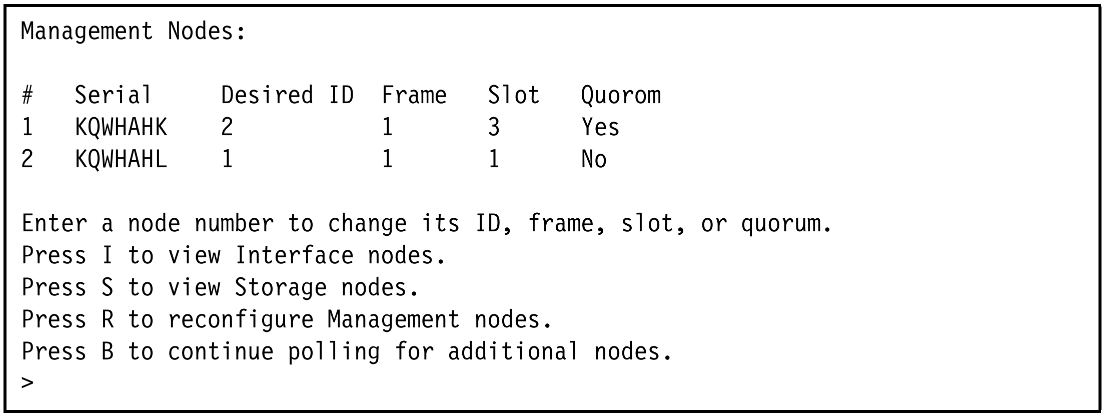
```
Detected 0 Interface nodes and 0 Storage nodes
Detected 1 Interface nodes and 0 Storage nodes
Detected 2 Interface nodes and 0 Storage nodes
Detected 3 Interface nodes and 1 Storage nodes
Detected 3 Interface nodes and 2 Storage nodes
```

Figure 3-56 Sample output for three Interface and two Storage nodes

Verifying device IDs, rack and slot locations, and quorums

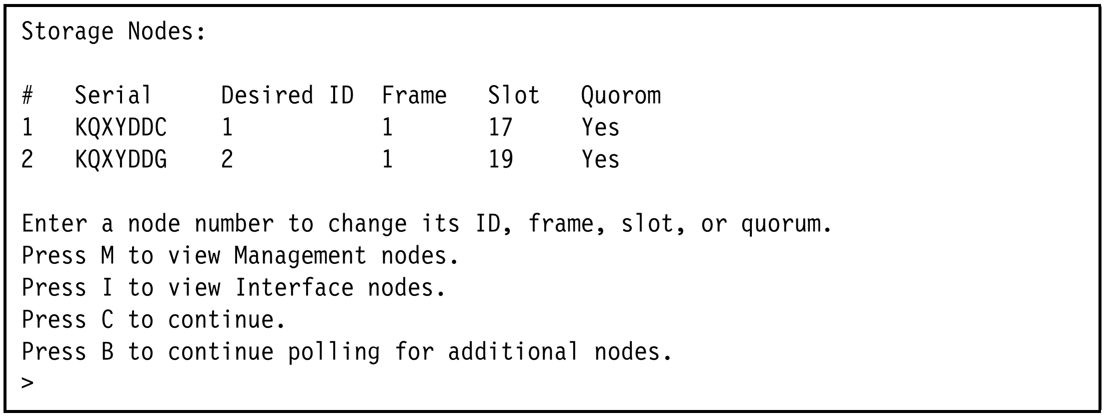
Proper rack cabling and configuration are key to a correctly working SONAS system and proper GUI representation. Accurate cabling in the internal SONAS frame (Ethernet and InfiniBand) switches ensure proper frame location identification and configuration of these nodes. However, locations can sometimes require redefinition in the `first_time_install` process. If it is not done correctly, the GUI might not properly align the nodes in the health center frame model.

You can adjust and keep the configuration when you confirm accurate assignments for Interface nodes (I), Management nodes (M) (Figure 3-57), and Storage nodes (S) (Figure 3-58). When all rack locations and Quorum statuses are verified, select “C” to continue to configure the cluster.



```
Management Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQWHAHK    2           1       3      Yes
2   KQWHAHL    1           1       1      No

Enter a node number to change its ID, frame, slot, or quorum.
Press I to view Interface nodes.
Press S to view Storage nodes.
Press R to reconfigure Management nodes.
Press B to continue polling for additional nodes.
>
```

Figure 3-57 Management Nodes configuration



```
Storage Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQXYDDC    1           1       17     Yes
2   KQXYDDG    2           1       19     Yes

Enter a node number to change its ID, frame, slot, or quorum.
Press M to view Management nodes.
Press I to view Interface nodes.
Press C to continue.
Press B to continue polling for additional nodes.
>
```

Figure 3-58 Storage Nodes configuration

Configuring the node instances, location, including serial numbers, and quorum states is key to a correctly working SONAS system. Typically, the first instance of a node type is in the lowest point of the rack of its node type and the instances sequence up as you move up the rack. For example, the bottom Interface node is int001st001, and the next one up is int002st001. The bottom Ethernet switch is switch 1, and the next one up is switch 2. The bottom InfiniBand switch is switch 1, and the next one up is switch 2. This configuration changes a bit for the SONAS Storage nodes in Gateway configurations. The first and second Storage nodes are above Storage nodes 3 and 4. The rest of the Storage nodes are above Storage nodes 1 and 2. You can adjust and keep the configuration when accurate assignments are confirmed for Interface and Storage nodes (S), then select “C” to continue to configure the cluster.

Select the line item number to change the configuration (Device ID/instance, Frame, slot, and Quorum state):

Desired ID This ID is the instance of the node type. Therefore, the first Management node instance has a desired ID of 1 and the second ID has a desired ID of 2. In Figure 3-59 on page 158 note that ID 1 is in the lowest slot of the rack. The bottom Interface node in the frame is int001st001 (slot 1 = node 1). The second Interface node from the bottom of the frame is int002st001 (slot 3= node 2). Also, Storage nodes have a different configuration, as noted in Figure 3-59 on page 158. Storage node 1 and 2 are above Storage nodes 3 and 4.

Frame Number The frame number relates to the frame instance of SONAS: If this frame is frame one of one SONAS frame, use 1. If you have a base rack and an expansion frame, the base rack is 1 and the first expansion frame is 2.

Slot number (This is also known as the rack slot number). Use the lower U-number indicator of the slot for the device (in the frame).

Quorum There are three sizes of *quorum node configuration* for SONAS installation (small = 3, medium = 5, and large = 7). If a cluster is small (2 - 6 Interface nodes), you need three quorum nodes. If the cluster is medium (6 - 10 Interface nodes), set five quorum nodes. If the cluster is large (over 10 Interface nodes), set seven quorum nodes.

Figure 3-59 shows a SONAS Gateway RXA rack layout configuration and the node-naming conventions.

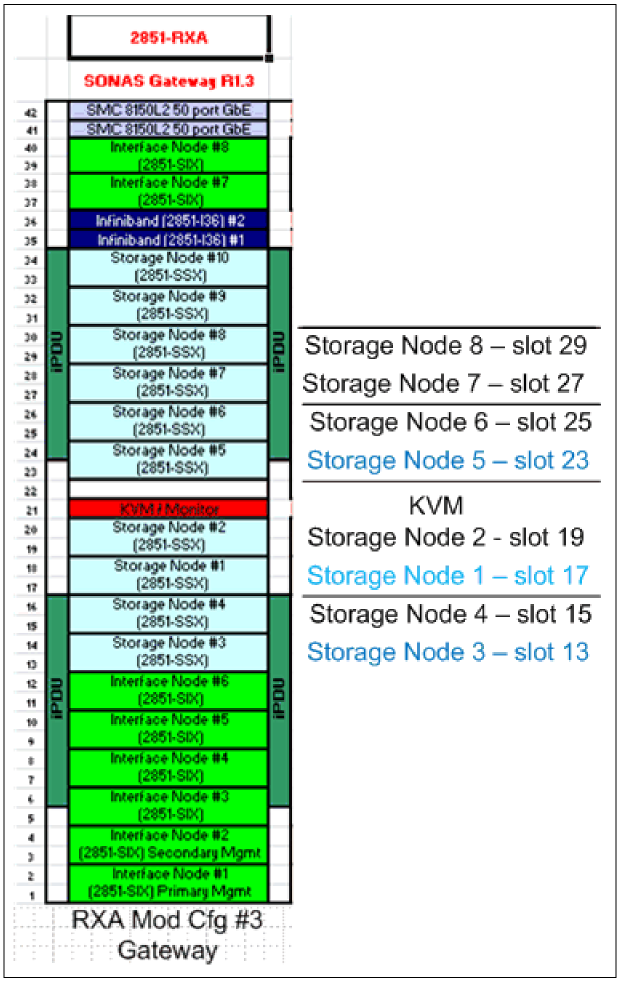

2851-RXA
SONAS Gateway R1.3
42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Interface Node #8 (2851-SIX)
Interface Node #7 (2851-SIX)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Storage Node #10 (2851-SSX)
Storage Node #9 (2851-SSX)
Storage Node #8 (2851-SSX)
Storage Node #7 (2851-SSX)
Storage Node #6 (2851-SSX)
Storage Node #5 (2851-SSX)
KVM / Monitor
Storage Node #2 (2851-SSX)
Storage Node #1 (2851-SSX)
Storage Node #4 (2851-SSX)
Storage Node #3 (2851-SSX)
Interface Node #6 (2851-SIX)
Interface Node #5 (2851-SIX)
Interface Node #4 (2851-SIX)
Interface Node #3 (2851-SIX)
Interface Node #2 (2851-SIX) Secondary Mgmt
Interface Node #1 (2851-SIX) Primary Mgmt
iPDU
iPDU
iPDU
iPDU
RXA Mod Cfg #3
Gateway
Storage Node 8 – slot 29
Storage Node 7 – slot 27
Storage Node 6 – slot 25
Storage Node 5 – slot 23
KVM
Storage Node 2 - slot 19
Storage Node 1 – slot 17
Storage Node 4 – slot 15
Storage Node 3 – slot 13

Figure 3-59 SONAS RXA - slot location offset for Storage nodes

Note: In SONAS V1.3.2 and later, the management functions are combined with the first two Interface nodes. Therefore, the first two Interface nodes (the bottom two nodes of the rack) are called mgmt001st001 and mgmt002st001. In a minimal configuration (that is, a two-Interface-node configuration), int001st001 or int002st001 do not exist.

After all Interface and Storage nodes are listed correctly and verified, type “B” to poll again or type “C” to continue with the cluster configuration. You can enter “C” to continue *only* from the Storage node configuration screen (see Figure 3-58 on page 156).

The first_time_install script configures the cluster and completes

The `first_time_install` script now continues by creating the SONAS cluster and adding each node to the cluster. This process takes another 30 - 40 minutes or longer depending on size of configuration. Figure 3-60 on page 159 shows the end of the script, after which the cluster is successfully configured. If it does not end with a success statement, open a PMR and immediately escalate to support for immediate assistance.

```
2013/08/01-21:28:16: Configuring a SONAS gateway.
2013/08/01-21:29:17: Creating GPFS cluster
2013/08/01-21:29:53: Configuring GPFS Cluster settings:
2013/08/01-21:31:55: Configuring GPFS node settings on node: KQWHAHK
2013/08/01-21:32:03: Configuring GPFS node settings on node: KQWHAHL
2013/08/01-21:32:11: Configuring GPFS node settings on node: KQXYDDC
2013/08/01-21:32:16: Configuring GPFS node settings on node: KQXYDDG
2013/08/01-21:33:54: Configuring multipath on node:KQXYDDC
2013/08/01-21:34:30: Configuring multipath on node:KQXYDDG
2013/08/01-21:34:33: Validating the NSDs before continuing, this can take up to 20 minutes.
2013/08/01-21:35:59: Skipping storage subsystem upgrade
2013/08/01-21:36:00: Synchronizing the SONAS repository with the secondary Management node
2013/08/01-21:37:53: Configuring yum on all nodes
2013/08/01-21:40:22: Configuring Performance Center service
2013/08/01-21:40:32: Starting system health monitoring

2013/08/01-21:41:48: Switch inventory complete

This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.

2013-08-01T21:41:49.914308-04:00: *** END /opt/IBM/sonas/bin/first_time_install(rc=0) ELP[1 hours 13 minutes 44
seconds]
```

Figure 3-60 Cluster being created and first_time_install script completes

In some cases, especially where the management network shares Ethernet ports with the public network, the `first_time_install` script ends with an NTP sync error. The SONAS V1.5.1 `first_time_install` script updates reduce the occurrence of this message, so it might not occur. It is okay to receive this message. NTP correctly syncs when the management network is correctly configured, as shown in Figure 3-61.

```
This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.

1 error was found, please repair it before continuing.

This is the deferred error:
ERROR---set_mgmt Code 107/0AFE: Unable to set the system's timezone. Unable to sync to any NTP server ---ERROR
2013-02-07T12:33:56.279583-05:00: *** END /opt/IBM/sonas/bin/first_time_install(rc=0) ELP[19 hours 2 minutes 40
seconds]
```

Figure 3-61 first_time_install script ends with error

Post first-time installation procedures

The GPFS cluster is installed, configured, and running. From here, the installer continues with postinstallation procedures.

Enabling licensing

To enable the license, run the following command:

```
/opt/IBM/sonas/bin/enablelicense --accept
```

The license can also be accepted the first time that you use the GUI.

Adding the GPFS cluster to the SONAS management subsystem

Before any other management functions can be used, the GPFS Cluster must be added to the SONAS management subsystem by running the following command:

```
cli addcluster -h mgmt001st001 -p Passw0rd
```

SONAS management port

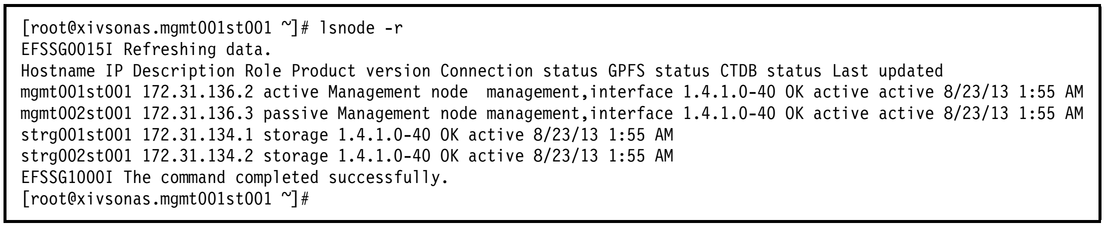
In many configurations, the management port is configured to use the same Ethernet ports as the public network (10 GbE). In these configurations, the management port is not connected to the network. Therefore, the management subsystem must be configured to use the 10 GbE network ports. The SONAS V1.5.1 `first_time_install` script enhancements allow the user to identify the external management adapter and prevent the need to run the **`chnwmgt`** command later. After a SONAS V1.5.1. installation, the **`chnwmgt`** command must be run *only* if no external management adapter value is input at the time of installation or if an incorrect value was entered and must be changed later. To configure the management ports on the 10 GbE network (if needed), run the following command:

```
chnwmgt --interface ethX1
```

The SONAS GUI is available for use.

Verifying the node list

Verify the node list by running the **`lsnode -r`** command, as shown in Figure 3-62.



```
[root@xivsonas.mgmt001st001 ~]# lsnode -r
EFSSG0015I Refreshing data.
Hostname IP Description Role Product version Connection status GPFS status CTDB status Last updated
mgmt001st001 172.31.136.2 active Management node  management,interface 1.4.1.0-40 OK active active 8/23/13 1:55 AM
mgmt002st001 172.31.136.3 passive Management node management,interface 1.4.1.0-40 OK active active 8/23/13 1:55 AM
strg001st001 172.31.134.1 storage 1.4.1.0-40 OK active 8/23/13 1:55 AM
strg002st001 172.31.134.2 storage 1.4.1.0-40 OK active 8/23/13 1:55 AM
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-62 lsnode output that shows configured nodes

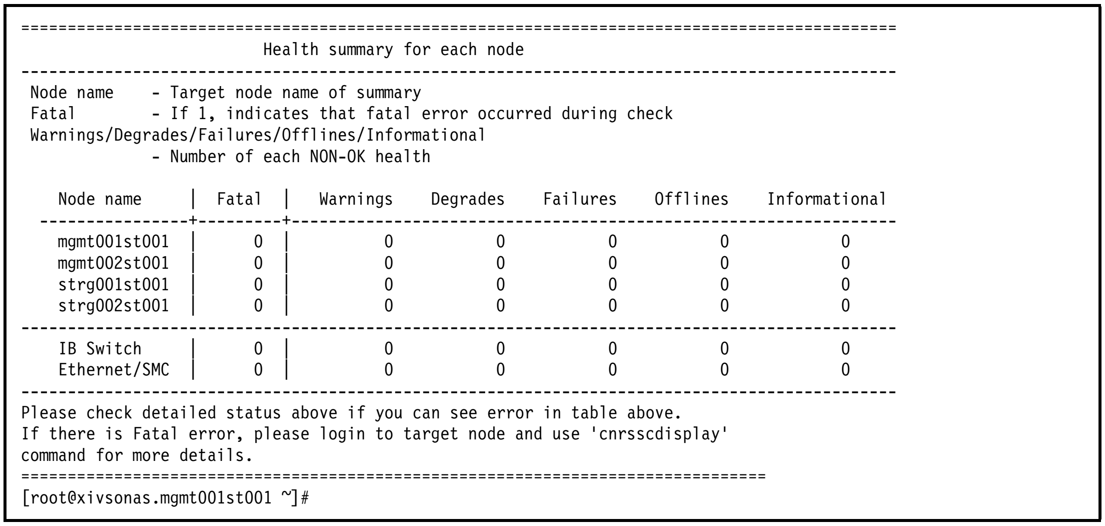
Health check

It is time to check the system health.

The IBM authorized service provider logs in to the Management node and runs the health check commands. The following command is run to check the SONAS system overall health and all components (Ethernet Switches, InfiniBand Switches, and nodes):

```
cnrsscheck --nodes=all --checks=all
```

The command also checks whether the nodes have correctly assigned roles and whether the nodes can communicate with each other. Figure 3-63 shows the command output for a sample cluster configuration.

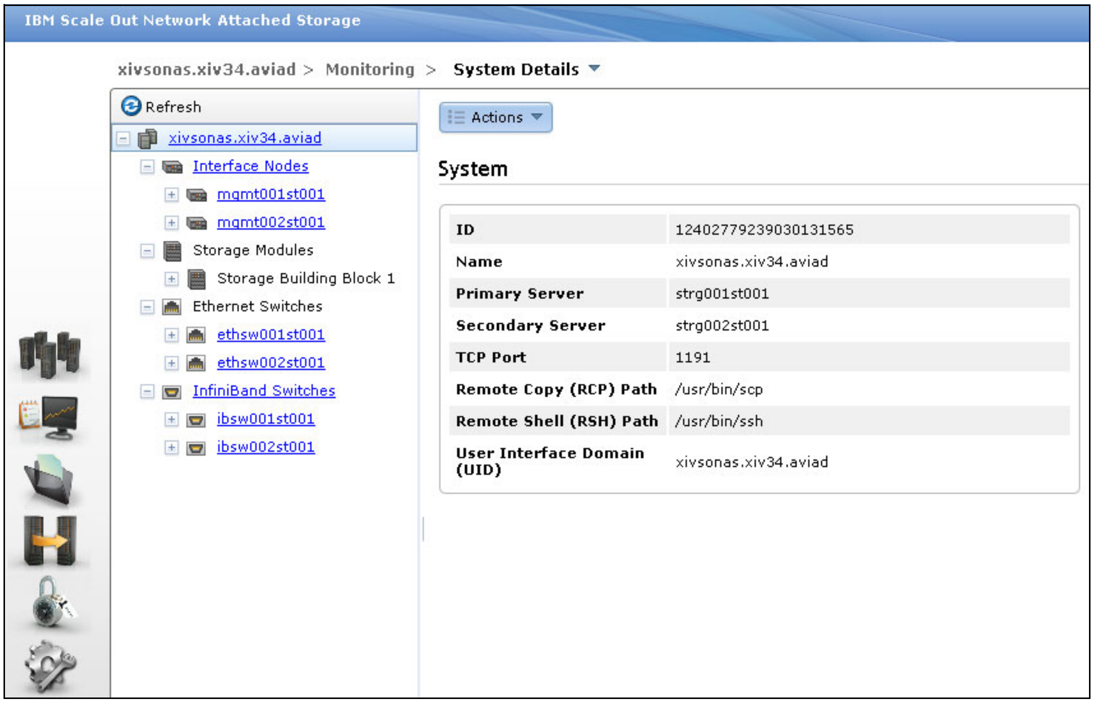


```
==============================================================================================
                          Health summary for each node
----------------------------------------------------------------------------------------------
 Node name     - Target node name of summary
 Fatal         - If 1, indicates that fatal error occurred during check
 Warnings/Degrades/Failures/Offlines/Informational
               - Number of each NON-OK health

    Node name      | Fatal  |   Warnings    Degrades    Failures    Offlines    Informational
  ---------------+--------+-----------------------------------------------------------------
    mgmt001st001  |    0   |      0           0           0           0            0
    mgmt002st001  |    0   |      0           0           0           0            0
    strg001st001  |    0   |      0           0           0           0            0
    strg002st001  |    0   |      0           0           0           0            0
----------------------------------------------------------------------------------------------
    IB Switch     |    0   |      0           0           0           0            0
    Ethernet/SMC  |    0   |      0           0           0           0            0
----------------------------------------------------------------------------------------------
Please check detailed status above if you can see error in table above.
If there is Fatal error, please login to target node and use 'cnrsscdisplay'
command for more details.
================================================================================
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-63   Sample output from the cnrsscheck summary

Clearing miscellaneous installation errors

During the installation process, some miscellaneous errors, or information warnings, might be generated as nodes are loaded and rebooted and ports go up and down. Review and clear errors and warnings before you run the health check. Use the GUI and the system monitoring screen. Ensure that all components are green. For more details, see Figure 3-64.



ID	12402779239030131565
Name	xivsonas.xiv34.aviad
Primary Server	strg001st001
Secondary Server	strg002st001
TCP Port	1191
Remote Copy (RCP) Path	/usr/bin/scp
Remote Shell (RSH) Path	/usr/bin/ssh
User Interface Domain (UID)	xivsonas.xiv34.aviad

Figure 3-64   SONAS Monitoring System Details window

Completing the SONAS cluster configuration

Before you create file systems, file sets, or exports, you must complete the SONAS cluster configuration by running the following commands:

- **`lscluster`** confirms the correct cluster configuration and lists the unique cluster ID that is required for the **`cfgcluster`** command.
- **`cfgcluster`** creates the initial configuration for all supported protocols (HTTPS, NFS, CIFS, FTP, and SCP), as shown in Figure 3-65. Here is list of the functions:
  – Prepare the CIFS configuration.
  – Distribute the CIFS configuration.
  – Distribute the CTDB configuration.
  – Import the CIFS configuration into the registry.
  – Write the initial configuration for NFS, FTP, HTTP, and SCP.
  – Restart CTDB to activate the new configuration.

```
[root@xivsonas.mgmt001st001 ~]# lscluster
Cluster id          Name                 Primary server Secondary server
Profile
12402779239014982142 xivsonas.xiv34.aviad strg001st001   strg002st001     SONAS
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
[root@xivsonas.mgmt001st001 ~]# cfgcluster xivsonas -c 12402779239014982142
Are you sure to initialize the cluster configuration ?
Do you really want to perform the operation (yes/no - default no):y
(1/7) Prepare CIFS configuration
(2/7) Write CIFS configuration on public nodes
(3/7) Write cluster manager configuration on public nodes
(4/7) Import CIFS configuration into registry
(5/7) Write initial configuration for NFS,FTP,HTTP and SCP
(6/7) Restart cluster manager to activate new configuration
(7/7) Initializing registry defaults
EFSSG0114I Initialized cluster configuration successfully
EFSSG0019I The task BackupMgmtNode has been successfully created.
EFSSG1000I The command completed successfully.
```

Figure 3-65 Example cfgcluster command

Upon completion, the cluster is fully configured and operational. However, more configuration is required for cluster networking and authentication. Because it is not dependent on the underlying storage type, it is described later in this chapter, independent of the Gateway solutions. Read through the chapter to understand network installation process and preferred practices for your cluster installation.

Starting with 3.8, “SONAS integration into your network” on page 218, this chapter provides detailed information about general SONAS and GPFS storage configuration information that applies to all back-end storage types. It also describes networking and other considerations that are not back-end storage specific. Read through the remainder of this chapter in its entirety to understand all installation considerations before you plan for installing your SONAS into your environment.

Rescanning Storage nodes

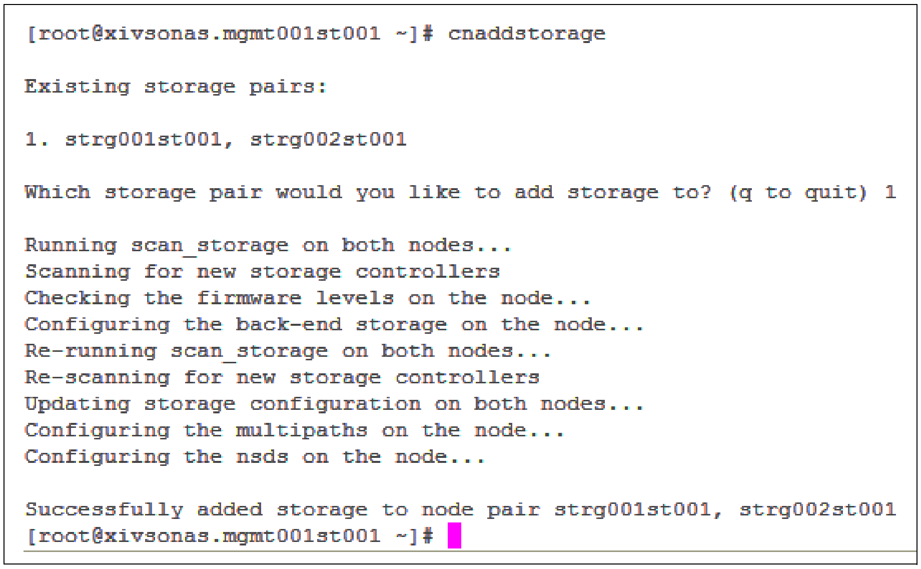
If the storage was not configured and allocated to the cluster before running the `first_time_install` script, run **`mkdisk --luns`** (SONAS V1.4.1 and later) as the administrator or root user. This command rescans all available Storage nodes for newly allocated LUNs (see Figure 3-66).

```
[root@xivsonas.mgmt001st001 ~]# mkdisk --luns
(1/3) Scanning for new devices
(2/3) Creating NSDs
(3/3) Adding to database
Successfully created disks:
SVC36005076802808162580000000000026_SVC36005076802808162580000000000026_mpath10
SVC36005076802808162640000000000027_SVC36005076802808162640000000000027_mpath11
SVC36005076802808162580000000000028_SVC36005076802808162580000000000028_mpath12
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-66 Example output from the mkdisk command

An alternative method is to run **`cnaddstorage`** as the root user. This command presents a short list of Storage node pairs from which to select.

When you run this command, select the Storage node pair to which you added storage. SONAS automatically discovers new volumes, and creates the multipath devices and the associated NSDs. This command must be run for a Storage node pair or pod. The process takes about six minutes to complete (see Figure 3-67).



```
[root@xivsonas.mgmt001st001 ~]# cnaddstorage

Existing storage pairs:

1. strg001st001, strg002st001

Which storage pair would you like to add storage to? (q to quit) 1

Running scan_storage on both nodes...
Scanning for new storage controllers
Checking the firmware levels on the node...
Configuring the back-end storage on the node...
Re-running scan_storage on both nodes...
Re-scanning for new storage controllers
Updating storage configuration on both nodes...
Configuring the multipaths on the node...
Configuring the nsds on the node...

Successfully added storage to node pair strg001st001, strg002st001
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-67 Example cnaddstorage output

After the LUNs and disks are rescanned, the installer continues the installation process.

> **Tip:** If you are reinstalling for any reason, previous NSD labels must be revoked before the reinstallation of LUN devices, either by reformatting or by running the **`dd if=/dev/zero of=/dev/mapper/`*`device-name`*** command against the device for about 5 seconds to clear the label.

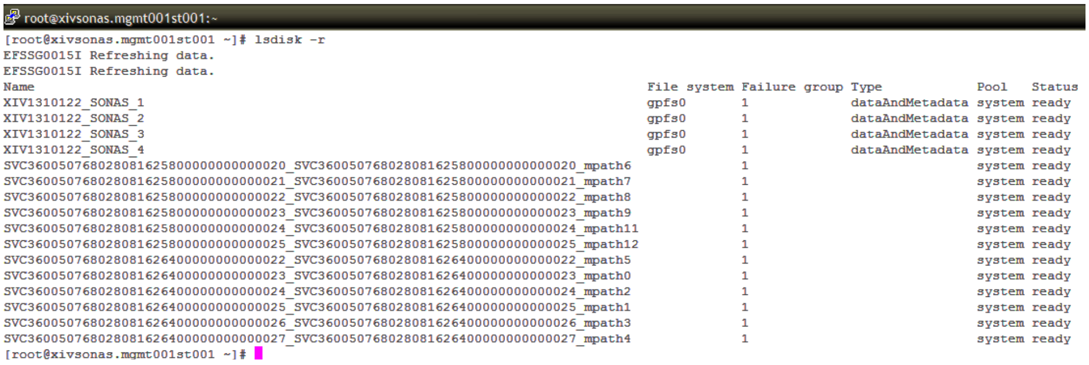
Verifying the disk configuration

To verify the disk configuration, run the following command at the Management node prompt:

```
lsdisk -r
```

The output is shown in Figure 3-68.

lsdisk -r



root@xivsonas.mgmt001st001:~

```
[root@xivsonas.mgmt001st001 ~]# lsdisk -r
EFSSG0015I Refreshing data.
EFSSG0015I Refreshing data.
Name                                                                              File system Failure group Type            Pool   Status
XIV1310122_SONAS_1                                                                gpfs0       1             dataAndMetadata system ready
XIV1310122_SONAS_2                                                                gpfs0       1             dataAndMetadata system ready
XIV1310122_SONAS_3                                                                gpfs0       1             dataAndMetadata system ready
XIV1310122_SONAS_4                                                                gpfs0       1             dataAndMetadata system ready
SVC360050768028081625800000000000020_SVC360050768028081625800000000000020_mpath6              1                             system ready
SVC360050768028081625800000000000021_SVC360050768028081625800000000000021_mpath7              1                             system ready
SVC360050768028081625800000000000022_SVC360050768028081625800000000000022_mpath8              1                             system ready
SVC360050768028081625800000000000023_SVC360050768028081625800000000000023_mpath9              1                             system ready
SVC360050768028081625800000000000024_SVC360050768028081625800000000000024_mpath11             1                             system ready
SVC360050768028081625800000000000025_SVC360050768028081625800000000000025_mpath12             1                             system ready
SVC360050768028081626400000000000022_SVC360050768028081626400000000000022_mpath5              1                             system ready
SVC360050768028081626400000000000023_SVC360050768028081626400000000000023_mpath0              1                             system ready
SVC360050768028081626400000000000024_SVC360050768028081626400000000000024_mpath2              1                             system ready
SVC360050768028081626400000000000025_SVC360050768028081626400000000000025_mpath1              1                             system ready
SVC360050768028081626400000000000026_SVC360050768028081626400000000000026_mpath3              1                             system ready
SVC360050768028081626400000000000027_SVC360050768028081626400000000000027_mpath4              1                             system ready
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-68 Example lsdisk report for Storwize V7000 volumes (2145, serialNumber, volume_name_instance)

All the Storwize V7000 disks from the pods are configured and are visible from the Management node. By default, they are added to the GPFS “system” pool in failure group “1”, and the Storage node preference is assigned in alternating LUN (called NSDs in GPFS) fashion.

Interpreting the lsdisk command output

The output of the **lsdisk** command provides information about the disk devices that you need to consider before you configure the file systems.

The following list describes the output that is shown in Figure 3-68 from left to right across the top:

- The device list begins with *Name*. This value refers to the NSD name. The NSD name is the GPFS “Network Storage Device” label. Storwize V7000 NSD names are cumbersome in comparison to their XIV counterparts.

  In the case of Storwize V7000 based storage, the first three characters refer to the storage type (SVC); this indicator is the same for Storwize V7000 and SAN Volume Controller.

  The next 36 characters refer to the Storwize V7000 serial number plus the Storwize V7000 volume ID. It is a unique identifier for all Storwize V7000 frames plus the volume ID.

  The next set of characters refers to a repeat of the first 39 characters. They are then followed by the OS /dev/mpath device number. There are approximately 87 characters in total. For Storwize V7000 storage, volume names are assigned in the NSDs during volume creation. For this example, the volumes were created with a sequential hexadecimal numbering pattern (that you cannot control) and the volume name, which indicates the drive type and volume number (SASvol1).

  The simplified volume name is not used on the SONAS side, so it is important to create a spreadsheet or some form of tracking for each volume identifier from each Storwize V7000 frame that you can later use to match the volumes that you want in SONAS. Otherwise, they all look similar and it is difficult to distinguish 1 TB NSD from the 2 TB NSD and SAS from NLSAS or SSD.

When you create NSDs with Storwize V7000 storage, use simple volume names, such as SASvol1 and NLSAS_2TB_Vol1, and allow the Storwize V7000 subsystem to number them sequentially. Always use the same size volume per Storwize V7000 storage pool. With this practice, even if you create volumes from 10 different Storwize V7000 subsystems, the volume list always lists by serial number and volume number and remains fairly simple to manage logically. For the Storwize V7000 subsystem, a preferred practice is to track meticulously information in a spreadsheet.

- The next piece of information from the `lsdisk` list is *File system*.

  If the listed NSD is not yet assigned to a file system, this column remains blank for that device. When the device is assigned to a file system, the file system name (such as "gpfs0") is listed in this column.

- The next column in the list is *Failure group*. The NSD failure group is a value that GPFS allows you to assign to any NSD to manage logically groups of disk devices in that file system. For example, if you have two Storwize V7000 frames to provide storage in a single storage pod, you can assign the first Storwize V7000 to failure group 1, and the second Storwize V7000 to failure group 2. This assignment allows you to define different replication between these groups.

- That leads us to the *Type* definition. The type refers to the usage type of data for which you can use that NSD. For example, you can use the disk for dataOnly, for metadataOnly, or for metadataAndData.

- The *Pool* definition refers to the category of disk for that disk type. If you have SAS and Near Line SAS in a file system and you choose to have them both in the same file system, it is possible by having them defined in different "pools". The file system places data on the "system pool" by default. You can then migrate data to a different pool by a structured ILM policy.

  A preferred practice is to place your fastest tier of disk storage in the system pool, and migrate to lower tiers from there. It ensures metadata placement on high-speed disk technology.

- *Status* displays the readiness of the NSD for use. Disk status has five possible values, three of which are transitional:

  - Ready: Normal status.
  - Suspended: Indicates that data will be migrated off this disk.
  - Being emptied: Transitional status in effect while a disk deletion is pending.
  - Replacing: Transitional status in effect for an old disk while replacement is pending.
  - Replacement: Transitional status in effect for a new disk while replacement is pending.

  GPFS allocates space only on disks with a status of Ready or Replacement.

- *Availability* refers to the following possible values:

  - Up: The disk is available to GPFS for normal read and write operations.
  - Down: No read and write operations can be used on the disk.
  - Recovering: An intermediate state for disks that are starting, during which GPFS verifies and corrects data. The read operations can be done while a disk is in this state, but write operations cannot.

- – Unrecovered: Not all disks were successfully brought up. There are cases where multiple disks must be recovered simultaneously to bring the storage to a consistent state. The most obvious cases involve replication where both copies of some pointers to data on this disk are unavailable. Unrecovered means that the disk is physically available, but the prerequisites for running recovery are not satisfied.
- – Timestamp: Refers to the date and time that the NSD is created.

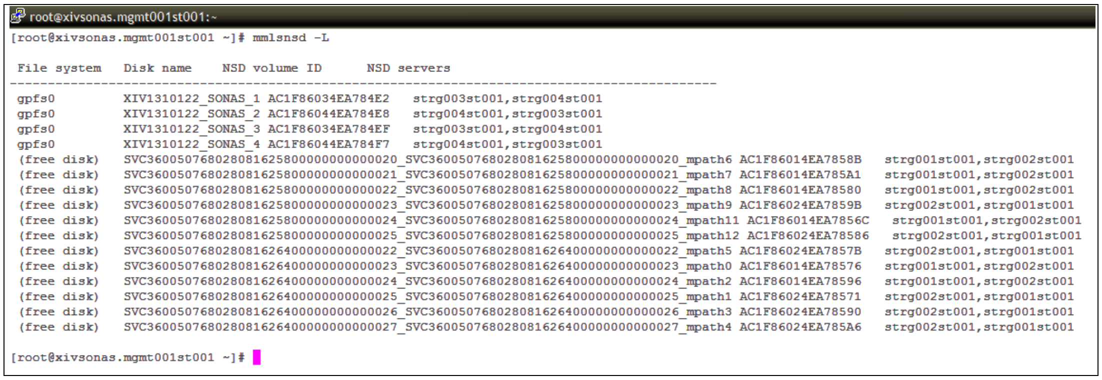
Note: Disk Pool, Usage Type, and Failure Groups are assigned to the disks/NSDs and must be defined before the disks are associated with a file system. These parameters can be changed only by removing the disk from the file system, changing the parameters, and reading the disk to the file system. This process assumes that there is enough space on the remaining disks in the file system to maintain the used capacity.

Creating the file system

Before you create your file system, ensure that the volumes you plan to use can evenly balance the work across the Storage nodes. It is important to create your file system with balance. When SONAS imports NSDs from the underlying storage subsystem, it assigns a Storage node preference to each NSD in alternating fashion.

With Storwize V7000, DCS3700, and DS8000 storage, it not as simple as with XIV storage because there is more than one RAID, size, and disk type.

To analyze the balance, run **`mmlsnsd -L`**. Select the disk types and NSD selections that go into the “system” pool versus other tier type pools and what file system structure you intend to use for them (see Figure 3-69).



```
[root@xivsonas.mgmt001st001 ~]# mmlsnsd -L

 File system   Disk name    NSD volume ID      NSD servers
---------------------------------------------------------------------------------------------
 gpfs0         XIV1310122_SONAS_1 AC1F86034EA784E2   strg003st001,strg004st001
 gpfs0         XIV1310122_SONAS_2 AC1F86044EA784E8   strg004st001,strg003st001
 gpfs0         XIV1310122_SONAS_3 AC1F86034EA784EF   strg003st001,strg004st001
 gpfs0         XIV1310122_SONAS_4 AC1F86044EA784F7   strg004st001,strg003st001
 (free disk)   SVC360050768028081625800000000000020_SVC360050768028081625800000000000020_mpath6 AC1F86014EA7858B   strg001st001,strg002st001
 (free disk)   SVC360050768028081625800000000000021_SVC360050768028081625800000000000021_mpath7 AC1F86014EA785A1   strg001st001,strg002st001
 (free disk)   SVC360050768028081625800000000000022_SVC360050768028081625800000000000022_mpath8 AC1F86014EA78580   strg001st001,strg002st001
 (free disk)   SVC360050768028081625800000000000023_SVC360050768028081625800000000000023_mpath9 AC1F86024EA7859B   strg002st001,strg001st001
 (free disk)   SVC360050768028081625800000000000024_SVC360050768028081625800000000000024_mpath11 AC1F86014EA7856C   strg001st001,strg002st001
 (free disk)   SVC360050768028081625800000000000025_SVC360050768028081625800000000000025_mpath12 AC1F86024EA78586   strg002st001,strg001st001
 (free disk)   SVC360050768028081626400000000000022_SVC360050768028081626400000000000022_mpath5 AC1F86024EA7857B   strg002st001,strg001st001
 (free disk)   SVC360050768028081626400000000000023_SVC360050768028081626400000000000023_mpath0 AC1F86014EA78576   strg001st001,strg002st001
 (free disk)   SVC360050768028081626400000000000024_SVC360050768028081626400000000000024_mpath2 AC1F86014EA78596   strg001st001,strg002st001
 (free disk)   SVC360050768028081626400000000000025_SVC360050768028081626400000000000025_mpath1 AC1F86024EA78571   strg002st001,strg001st001
 (free disk)   SVC360050768028081626400000000000026_SVC360050768028081626400000000000026_mpath3 AC1F86024EA78590   strg002st001,strg001st001
 (free disk)   SVC360050768028081626400000000000027_SVC360050768028081626400000000000027_mpath4 AC1F86024EA785A6   strg002st001,strg001st001

[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-69   Storwize V7000 CLI view of mmlsnsd -L output disk list with Storage node preference

The Storage node preference is important to performance because it defines the Storage node that tries to manage all I/O to that NSD at any particular time. The I/O is managed only by the second node in the Storage node preference if the first node fails to serve the I/O. This failover behavior is automatic.

Figure 3-69 shows that for NSD “`SVC360050768028081625800000000000020_SVC360050768028081625800000000000020_mpath6`”, the Storage node strg001st001 is the Storage node that is preferred, and strg002st001 is the backup node for I/O that is going to the NSD.

As shown in Figure 3-70, the Storage node preference for strg001st001 is alternating sequentially by volume name. It must be correctly mapped for SONAS file system creation. If you provision a file system with consecutively named NSDs, the file system is automatically balanced across each Storage node when there is an even number of NSDs assigned to the file system.

```
[root@xivsonas.mgmt001st001 bin]# mmlsnsd -L
File system Disk name NSD servers
---------------------------------------------------------------------------
gpfs0
SVC36005076802808162580000000000000020_SVC36005076802808162580000000000000020_mpa
th6 strg001st001,strg002st001
gpfs0
SVC36005076802808162580000000000000023_SVC36005076802808162580000000000000023_mpa
th9 strg002st001,strg001st001
gpfs0
SVC36005076802808162640000000000000022_SVC36005076802808162640000000000000022_mpa
th5 strg001st001,strg002st001
gpfs0
SVC36005076802808162640000000000000023_SVC36005076802808162640000000000000023_mpa
th0 strg002st001,strg001st001
```

Figure 3-70 Output from the mmlsnsd command that shows Storage node preference on XIV NSDs

It is even more important when multiple frames are deployed. If you want the highest achievable performance, spread out the NSD resources evenly across all Storage nodes. The following command output that is shown in Figure 3-71 can help illustrate this concept.

In Figure 3-70, the two Storwize V7000 subsystems are in the same Storage node pair. The different Storwize V7000 subsystems are clearly identifiable from the serial number indicator in the NSD Name. When you create the file system, you *must* type in the NSD devices across the Storage node and Storwize V7000 frame preference.

Figure 3-71 shows an example **mkfs** command for evenly distributing the NSDs and Storage node preference in the creation of the gpfs0 file system. The two Storwize V7000 subsystems are connected to the same Storage node pair (a preferred practice consideration for two XIV subsystems in the same Pod).

```
cli mkfs gpfs0 /ibm/gpfs0 -F
SVC36005076802808162580000000000000020_SVC36005076802808162580000000000000020_mpa
th6,SVC36005076802808162640000000000000023_SVC36005076802808162640000000000000023
_mpath0,SVC36005076802808162580000000000000023_SVC360050768028081625800000000000
0023_mpath9,SVC36005076802808162640000000000000022_SVC360050768028081626400000000
00000022_mpath5 -R meta -j cluster”
```

Figure 3-71 Example output for the mkfs gpfs0 /ibm/gpfs0 -F output command

This example uses the **-j** option of GPFS **mkfs** (for cluster allocation type).

Data allocation types

GPFS offers two basic data allocation types: Scatter and Cluster.

Scatter allocation type disperses I/O randomly across all NSDs. The Cluster allocation type lays it down contiguously across all striped NSDs.

Our testing shows that well-defined Storwize V7000 NSD striping performs better when the Cluster allocation type is used in file system creation for most large file sequential workloads. Current data also suggests that the best file system block allocation, from a performance stand point, is 1 M (non-default) for most large file sequential type workloads. In contrast, typically for small file random I/O patterns, block sizes of 256 KB and Scatter allocation types tend to serve with better performance. For this reason, most users might benefit from testing with simulated production workloads before they decide which type to choose.

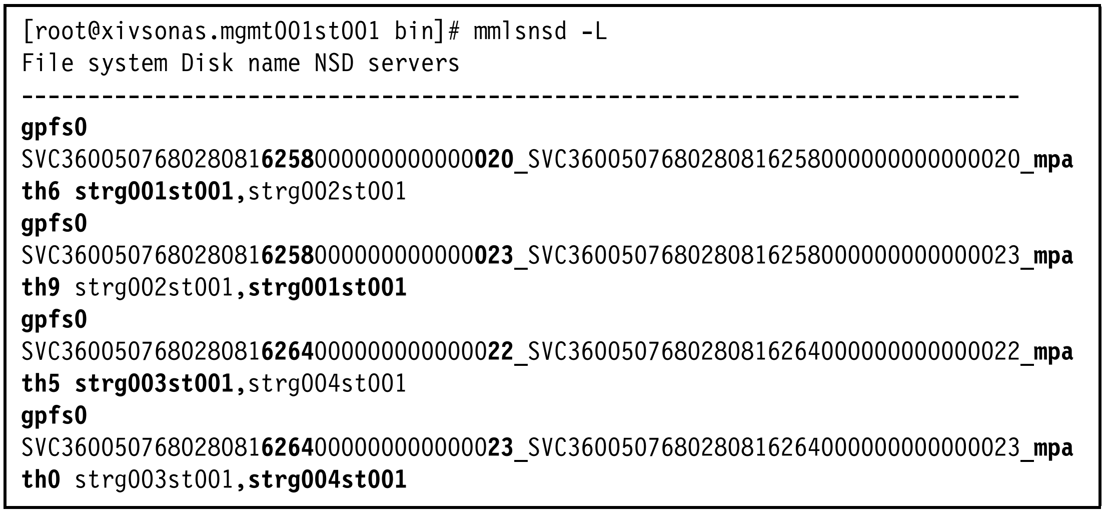
Tip: In some cases, it might be faster to serve I/O in larger block sizes.

Note: The block size and block allocation map type are defined when the file system is created. They cannot be changed without deleting and re-creating the file system. This process requires you to back up and restore the data.

SONAS file system subblocks are 1/32 of the block size. So, the subblock size of a 256 KB file system block size is 8 K, and the subblock size of a 1 M file system block size is 32 K. The minimum subblock allocation is the subblock size that is based on the defined block size for that specific file system and any file set within that file system (dependent or independent).

As shown in Figure 3-71 on page 167, the **mkfs** command stripes the file system across the first NSD from the first Storwize V7000, then the second NSD of the next Storwize V7000. Then, it jumps back to the first NSD of the first Storwize V7000 subsystem and then the second NSD of the second Storwize V7000 subsystem. This process stripes in alternating patterns across the Storage node pairs. If you use three NSDs, it is easy to see that one Storage node is processing more work than the other. It might not be a problem in the case of many NSDs within a file system, but it is important understand this scenario before you finalize your configuration.

When each Storwize V7000 subsystem is in a separate Storage node pair, the preferred practice configuration is slightly different. Figure 3-72 illustrates that difference.



```
[root@xivsonas.mgmt001st001 bin]# mmlsnsd -L
File system Disk name NSD servers
----------------------------------------------------------------------------
gpfs0
SVC3600507680280816258000000000000020_SVC3600507680280816258000000000000020_mpa
th6 strg001st001,strg002st001
gpfs0
SVC3600507680280816258000000000000023_SVC3600507680280816258000000000000023_mpa
th9 strg002st001,strg001st001
gpfs0
SVC3600507680280816264000000000000022_SVC3600507680280816264000000000000022_mpa
th5 strg003st001,strg004st001
gpfs0
SVC3600507680280816264000000000000023_SVC3600507680280816264000000000000023_mpa
th0 strg003st001,strg004st001
```

Figure 3-72 Multi XIV output from the mmlsnsd command, which shows a two-pod node preference

As shown in Figure 3-72 on page 168, the configuration is initially more complex, depending on the scale of your solution. However, after the file system is created, there is nothing else to do. You can set it and forget it. Figure 3-73 provides a closer look at a **mkfs** command that you can use for this configuration.

```
cli mkfs gpfs0 /ibm/gpfs0 -F
SVC360050768028081625800000000000020_SVC360050768028081625800000000000020_mpath6,
SVC360050768028081626400000000000022_SVC360050768028081626400000000000022_mpath5,
SVC360050768028081625800000000000023_SVC360050768028081625800000000000023_mpath9,
SVC360050768028081626400000000000023_SVC360050768028081626400000000000023_mpath0
-R meta -j cluster
```

Figure 3-73 CLI mkfs gpfs0 /ibm/gpfs0 -F command

As shown in Figure 3-73, data is striped evenly across NSDs and Storage node pairs to provide the highest level of dispersal in the file system stripe as possible. It is a preferred practice for NSD mapping for file system creation. Remember, if you are using Scatter as your GPFS allocation type, the layout does not provide the planned balance and it becomes less valuable. Therefore, plan it carefully.

Replicating metadata

In the sample **mkfs** command in Figure 3-73, the **-R meta** option is used, which specifies to replicate metadata across the NSD failure groups. Replication relates to synchronous replication of GPFS data across NSDs in separately defined failure groups. In fact, **meta** is the default replication setting. In this case, you do not need to specify it (if you placed both Storwize V7000 subsystems into separate failure groups). There is no value in having more than two failure groups. However, when you have more than one Storwize V7000 subsystem, you can use two failure groups to protect the status of all metadata if you lose a Storwize V7000 subsystem. If you have ten Storwize V7000 subsystems, you can set up five Storwize V7000 subsystems in Failure group 1 and the other five in Failure group 2. It is a preferred practice consideration for reliability.

Replication can reduce risk. It is acceptable to create your Storwize V7000 based file systems with **-R none** (replication set to none), when back-end storage is securely protected with sound data protection practices. However, without replication in failure groups, the GPFS file system sees one copy of data and metadata (regardless of how many copies exist in the back-end).

The options for Failure group Replication are as follows:

```
-R { none | meta | all }
```

- **none**, which means no replication at all.
- **meta**, which indicates that the file system metadata is synchronously mirrored across two failure groups.
- **all**, which indicates that the file system data and metadata is synchronously mirrored across two failure groups.

Completing the SONAS installation

Because cluster networking and authentication do not depend on the specific type of underlying storage, they are described in this chapter separately from the Gateway solutions. Read through the chapter to understand the network installation process and preferred practices for your cluster installation.

Sections 3.8, “SONAS integration into your network” on page 218 and later provide detailed information about general SONAS and GPFS storage configuration information that is relative to any back-end storage, in addition to networking and other considerations that are not back-end storage specific. Read through the remainder of this chapter in its *entirety* to understand all installation considerations before you plan to integrate SONAS into your environment.

3.5.6 Adding Storwize V7000 LUNs to an existing SONAS configuration

As an example, assume that the file system is created in a logical, balanced fashion, and that you want to add NSDs to the file system. This section provides an overview of this process, then provides details with screen captures.

Creating and selecting the volumes to be added

On the Storwize V7000s subsystems, to create the volumes that you want to add to the SONAS cluster, complete the following steps:

1. Expand the SONAS storage pools, if necessary, and create the volumes from the expanded Storwize V7000 storage pool. Be sure to use the same volume size and name convention and continue the numerical sequence from the last or previous volume that was created for SONAS.
2. From the Volumes list, select the new volumes for the SONAS solution, right-click, and map the selected volumes to the SONAS cluster.

Summary of adding NSDs to SONAS

This section summarizes the process of adding NSDs to SONAS.

- When you add disks or there is a need to rescan for new storage, run **`mkdisk --luns`** (SONAS V1.4.1 and later) as the admin or root user. This command rescans all available Storage nodes for newly allocated LUNs.

  An alternative method is to run **`cnaddstorage`** as the root user. This command presents a short list of Storage node pairs from which to select.

  To use this process, select the Storage node pair to which you added storage. SONAS automatically discovers new volumes and creates the multipath devices and the associated NSDs. The command must be run for each Storage node pair or pod where new storage was added. This process takes up to six minutes to complete.

- When the **`cnaddstorage`** command finishes, you can view all the devices by running **`lsdisk -r`**. The new devices can now be used to increase the size of an existing file system or create one.

Detailed Storwize V7000 work

From the Storwize V7000 GUI (see Figure 3-74), ensure that your SONAS storage pool has enough capacity for the volumes that you want.

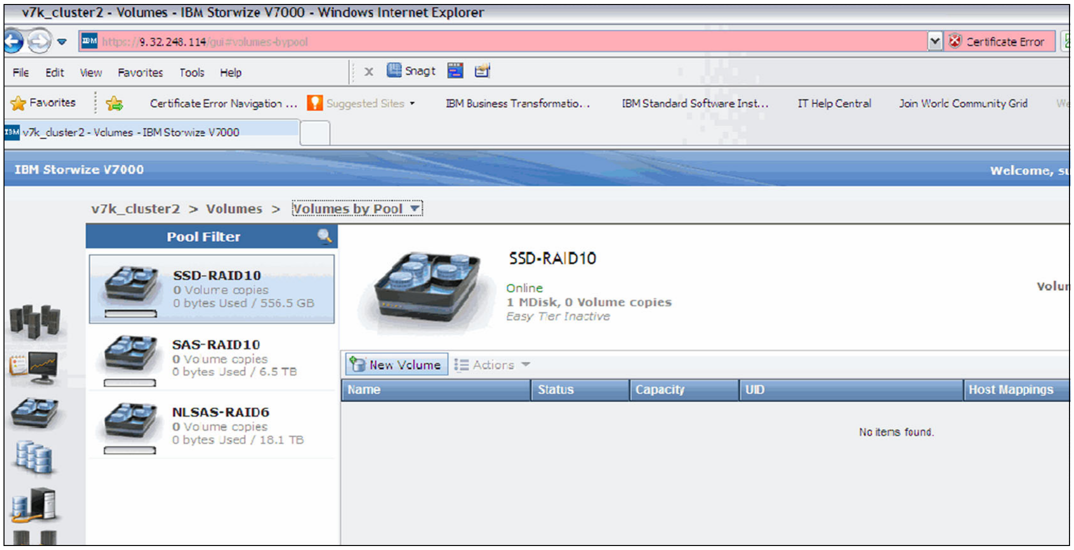
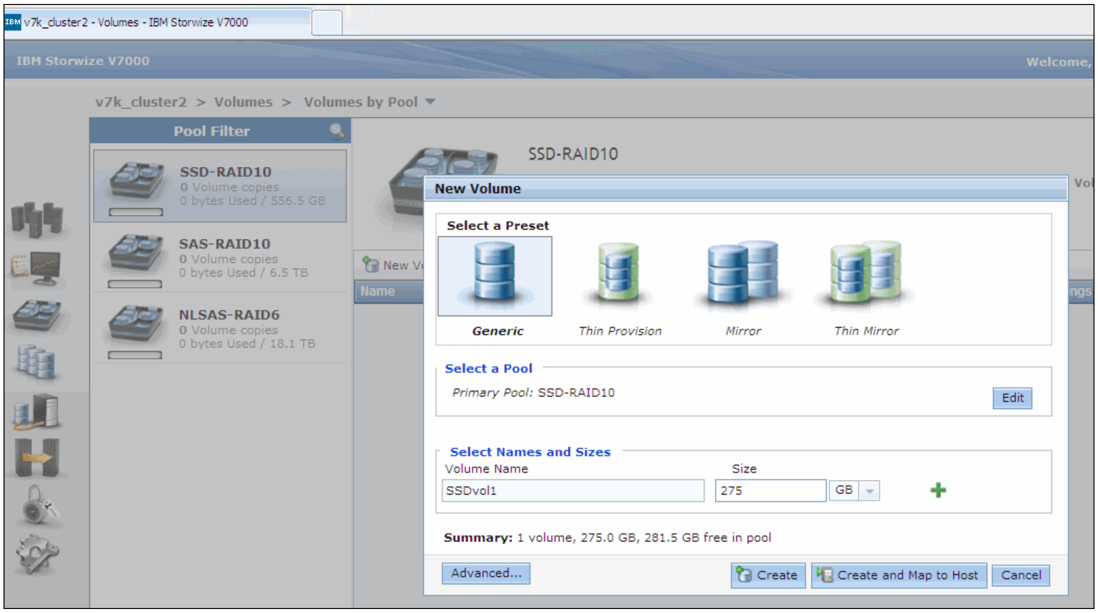


Figure 3-74 Storwize V7000 GUI - volume list from the storage pool

The next step is to size, name, and create the volumes, as shown in Figure 3-75.



Primary Pool: SSD-RAID10

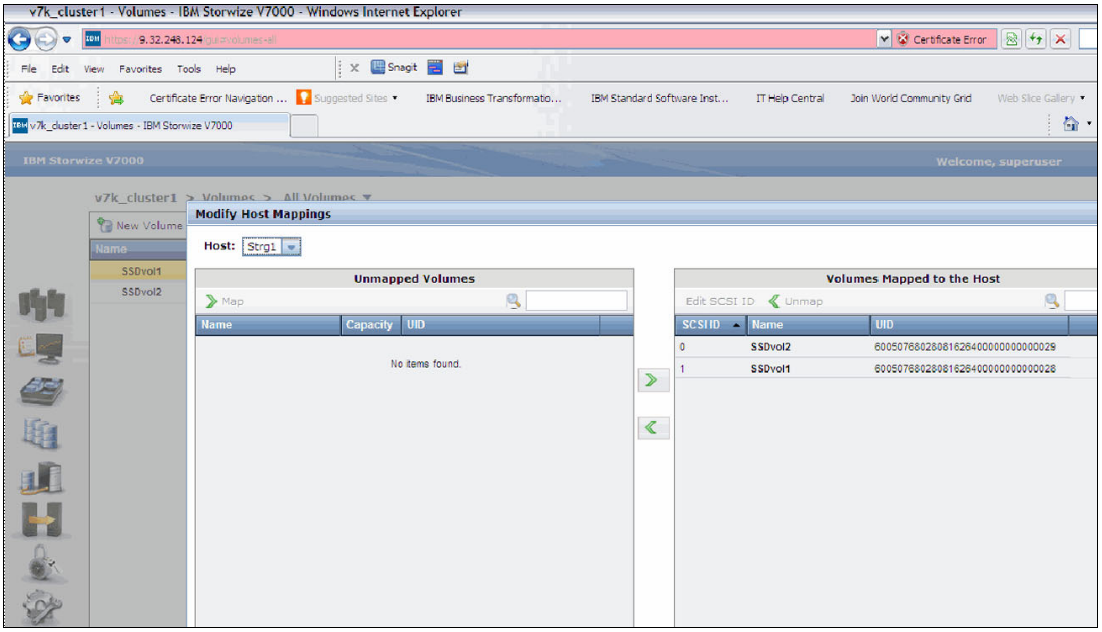
Volume Name: SSDvol1

Size: 275 GB

Summary: 1 volume, 275.0 GB, 281.5 GB free in pool

Figure 3-75 Select the number, size, and name of the volumes to add

The next step is to map the Storwize V7000 volumes to the SONAS cluster, as shown in Figure 3-76.



v7k_cluster1 - Volumes - IBM Storwize V7000 - Windows Internet Explorer
Certificate Error
File Edit View Favorites Tools Help
Snagit
Favorites
Certificate Error Navigation ...
Suggested Sites
IBM Business Transformatio...
IBM Standard Software Inst...
IT Help Central
Join World Community Grid
Web Slice Gallery
v7k_cluster1 - Volumes - IBM Storwize V7000
IBM Storwize V7000
Welcome, superuser
v7k_cluster1 > Volumes > All Volumes
New Volume
Name
SSDvol1
SSDvol2
Modify Host Mappings
Map
Edit SCSI ID
Unmap

Host: Strg1

Unmapped Volumes

Name	Capacity	UID

No items found.

Volumes Mapped to the Host

SCSI ID	Name	UID
0	SSDvol2	60050768028081626400000000000029
1	SSDvol1	60050768028081626400000000000028

Figure 3-76 Storwize V7000 GUI - map the volume to the SONAS cluster

The Storwize V7000 work is complete.

Detailed SONAS side work

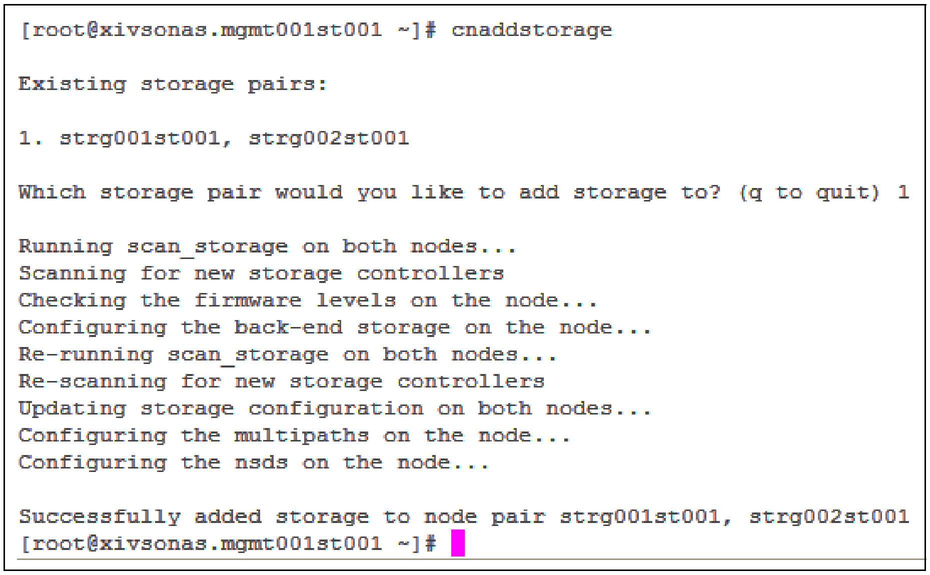
From the Management node in the SONAS cluster, run **mkdisk --luns** command as the admin user, as shown in Figure 3-77. This command rescans all available Storage nodes for storage.

```
[root@xivsonas.mgmt001st001 ~]# mkdisk --luns
(1/3) Scanning for new devices
(2/3) Creating NSDs
(3/3) Adding to database
Successfully created disks:
SVC360050768028081625800000000000026_SVC360050768028081625800000000000026_mpath10
SVC360050768028081626400000000000027_SVC360050768028081626400000000000027_mpath11
SVC360050768028081625800000000000028_SVC360050768028081625800000000000028_mpath12
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-77 Example mkdisk command

An alternative method is to run **cnaddstorage** as the root user. This command presents a short list of Storage node pairs from which to select.

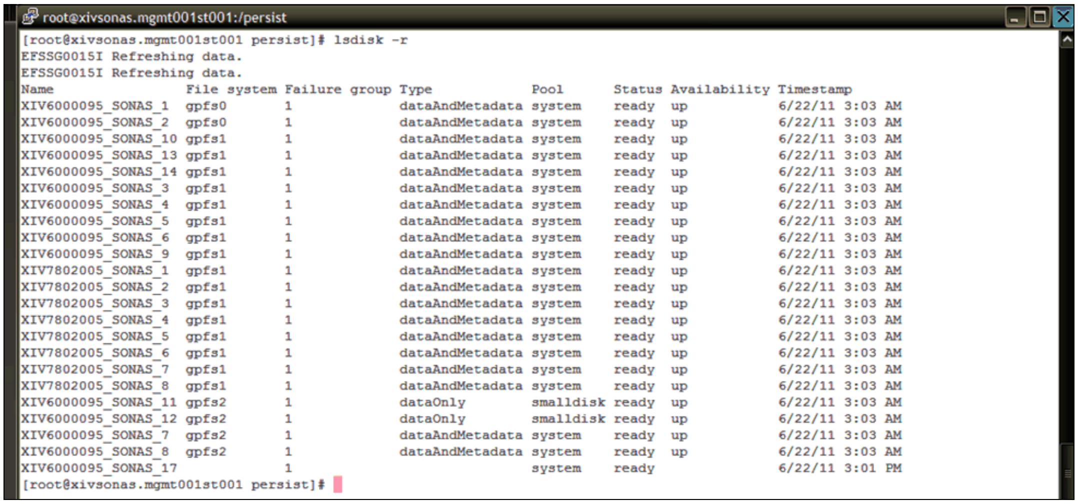
1. Select the Storage node pair to which you added storage. SONAS automatically discovers new volumes, and creates the multipath devices and the associated NSDs. The command must be run for each Storage node pair or pod. This process can take up to six minutes to complete (see Figure 3-78 on page 173).



```
[root@xivsonas.mgmt001st001 ~]# cnaddstorage

Existing storage pairs:

1. strg001st001, strg002st001

Which storage pair would you like to add storage to? (q to quit) 1

Running scan_storage on both nodes...
Scanning for new storage controllers
Checking the firmware levels on the node...
Configuring the back-end storage on the node...
Re-running scan_storage on both nodes...
Re-scanning for new storage controllers
Updating storage configuration on both nodes...
Configuring the multipaths on the node...
Configuring the nsds on the node...

Successfully added storage to node pair strg001st001, strg002st001
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-78   CLI cnaddstorage command sample

2. It takes a few minutes to discover the new volumes, create the multipath devices, and create the NSDs. Then, on completion, look for the devices by running **`lsdisk -r`**, as shown in Figure 3-79.



```
[root@xivsonas.mgmt001st001 persist]# lsdisk -r
EFSSG0015I Refreshing data.
EFSSG0015I Refreshing data.
Name                File system Failure group Type            Pool      Status Availability Timestamp
XIV6000095_SONAS_1  gpfs0       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_2  gpfs0       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_10 gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_13 gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_14 gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_3  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_4  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_5  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_6  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_9  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_1  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_2  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_3  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_4  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_5  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_6  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_7  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV7802005_SONAS_8  gpfs1       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_11 gpfs2       1             dataOnly        smalldisk ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_12 gpfs2       1             dataOnly        smalldisk ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_7  gpfs2       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_8  gpfs2       1             dataAndMetadata system    ready  up           6/22/11 3:03 AM
XIV6000095_SONAS_17             1                             system    ready               6/22/11 3:01 PM
[root@xivsonas.mgmt001st001 persist]#
```

Figure 3-79   Example output from lsdisk -r

3. The additional disks and NSDs can now be added to an existing file system for more capacity, or be used to create one. Run **`chdisk`** to alter the disk parameters to the wanted values.

3.6 SONAS Gateway configuration with DCS3700 storage

The SONAS Gateway configuration with DCS3700 (2851-DR2/DE2) storage is listed as a feature code option when you order the SONAS Gateway. DCS3700 Gateways are FC 9008. IBM offers the DCS3700 (1818-80C/E) outside of SONAS. These configurations are supported if the minimum configurations are met. Much like the gateway solutions for the XIV (FC 9006) and Storwize V7000 (FC 9007) subsystems, you can stack one or two DCS3700 controllers behind each SONAS storage pod.

The DCS3700 storage with a SONAS Gateway offers high flexibility in storage configuration, plus high-density storage in a final footprint with the inclusion of SSD for high performance within a SONAS storage solution. These features mean that you have the greatest flexibility, performance, and small footprint storage capabilities with the number and types of drives that you can assemble within the SONAS cluster and storage frames when you use DCS3700 storage.

SONAS DCS3700 Gateway solutions can provide high-performance flexibility:

- This solution offers great flexibility in disk speeds, types, and even drive size options in the IBM storage catalog. It allows for flexible storage tiering within the storage pod between SSD, SAS, and Near Line SAS storage options.
- The DCS3700 solution is also one of two possible solutions that accommodates support for an SSD tier of SONAS storage. It is a preferred practice solution for supporting the highest speed of device technology for metadata placement for GPFS and high-speed metadata scan rates (in the underlying file system where SSD is used).

3.6.1 Overview of DCS3700 storage

The DCS3700 subsystem consists of up to two controller enclosures per SONAS storage pod. Up to five expansion enclosures can be attached per controller, depending on the controller type.

- 1818-80C / 2851-DR2 Base Controller (Snowmass): Up to two expansion enclosures can be attached per controller enclosure. It can scale up to180 disks (60 per enclosure) and 720 TB (4 TB disk) raw capacity per controller (1.4 PB per SONAS storage pod).
- 1818-80C / 2851-DR2 with Performance Controller (Pikes Peak) (FC 3100): Up to five expansion enclosures can be attached per controller enclosure. It can scale up to 360 disks (60 per enclosure) and 1.4 PB (4 TB disk) raw capacity per controller (2.8 PB per SONAS storage pod).

Overview of enclosures

Throughout this section, there are references to 1818-80C, 1818-80E, 2851-DR2, and 2851-DE2. Here are descriptions:

- 1818-80C: This enclosure is a DCS3700 controller enclosure that is ordered separately from the SONAS Gateway solution. The 1818-80C is not directly linked to the SONAS Gateway solution, but is supported if the minimum configuration and code levels (NVRAM, Firmware) are met.
- 1818-80E: This enclosure is a DCS3700 expansion enclosure that is ordered separately from the SONAS Gateway solution. The 1818-80E is not directly linked to the SONAS Gateway solution, but is supported if the minimum configuration and code levels (ESM) are met.

- 2851-DR2: This enclosure is a DCS3700 controller enclosure that is ordered with or specifically for a SONAS Gateway. Though ordered separately, the 2851-DR2 is directly linked to the SONAS Gateway, is shipped with the proper code levels, and is ready to be integrated into a SONAS Gateway solution.
- 2851-DE2: This enclosure is a DCS3700 expansion enclosure that is ordered with or specifically for a SONAS Gateway. Though ordered separately, the 2851-DE2 is directly linked to the SONAS Gateway, is shipped with the proper code levels, and is ready to be integrated into a SONAS Gateway solution.

Note: Though the 1818 and 2851 models are both supported by a SONAS Gateway, the 1818 and 2851 models are *not* compatible with each other and are not interchangeable. A 2851-DE2 cannot be used as an expansion enclosure for a 1818-80C enclosure. Likewise, a 1818-80E cannot be used as an expansion enclosure to 2851-DR2.

Note: Supported firmware levels are 7.86.40.0 or higher. Each code level comes with its own NVRAM level. Be sure to use the proper NVRAM for your DCS3700 model (1818 or 2851). Fix Central does *not* supply the NVRAM or firmware for the 2851. For more information, see your local IBM representative.

DCS3700 (1818-80C and 2851-DR2) controller enclosure

The controller enclosures contain two independent controller canister units, which are clustered by an internal network (upper A and lower B). The controller units are referred to as *canisters*. Each enclosure also includes two power supply units (PSUs). The controller enclosure PSUs each house a battery pack that retains cached data if there is a power failure. The controllers contain the storage subsystem control logic, interface ports, and LEDs. The controllers install from the rear of the storage enclosure. Controller A is installed in storage bridge bay slot A (upper-middle) and controller B is installed in storage bridge bay slot B (lower-middle). All connections to the hosts and the expansion enclosures are made through the controllers.

Each controller or expansion enclosure contains five trays of up to 12 drives each, as shown in Figure 3-80.

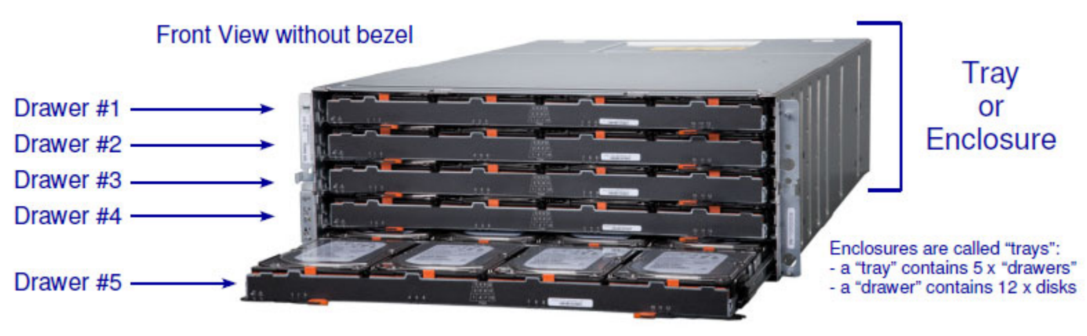

Front View without bezel
Drawer #1
Drawer #2
Drawer #3
Drawer #4
Drawer #5
Tray
or
Enclosure
Enclosures are called “trays”:
- a “tray” contains 5 x “drawers”
- a “drawer” contains 12 x disks

Figure 3-80 DCS3700 controller details

Figure 3-81 shows a rear view of the DCS3700 controller enclosure.

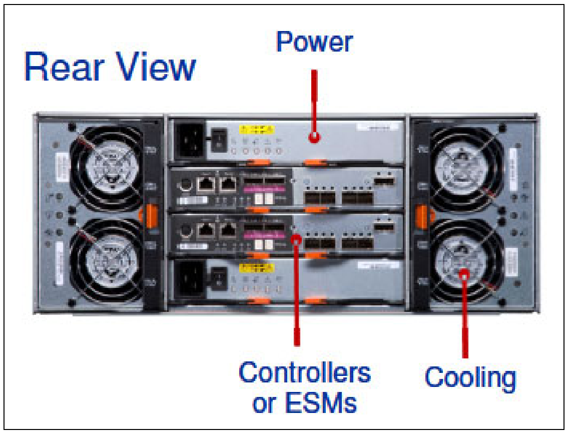


Figure 3-81 Rear view of the DCS3700 controller enclosure (note the upper and lower controller canisters)

Figure 3-82 shows the rear of the DCS3700 (1818-80C) controller canister.

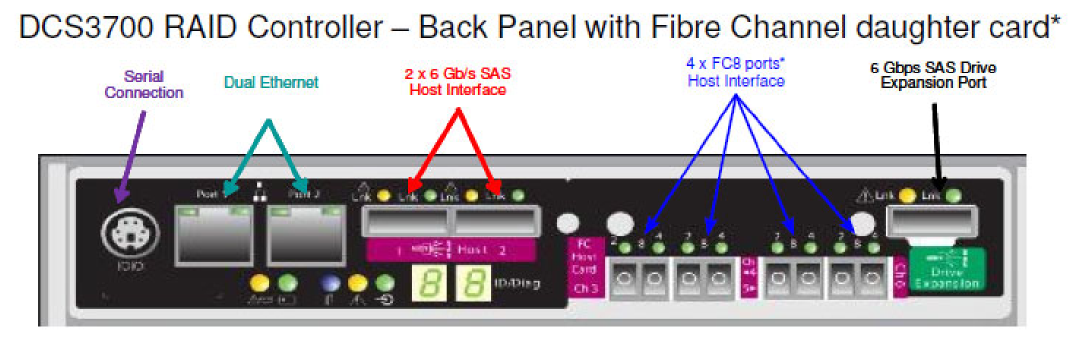

DCS3700 RAID Controller – Back Panel with Fibre Channel daughter card*
Serial Connection
Dual Ethernet
2 x 6 Gb/s SAS Host Interface
4 x FC8 ports* Host Interface
6 Gbps SAS Drive Expansion Port
Host 1 2
ID/Diag
FC Host Card Ch 3
Drive Expansion

Figure 3-82 Rear view of the DCS3700 (1818-80C) controller canister

Figure 3-83 shows a high-level view of the components of the DCS3700 chassis.

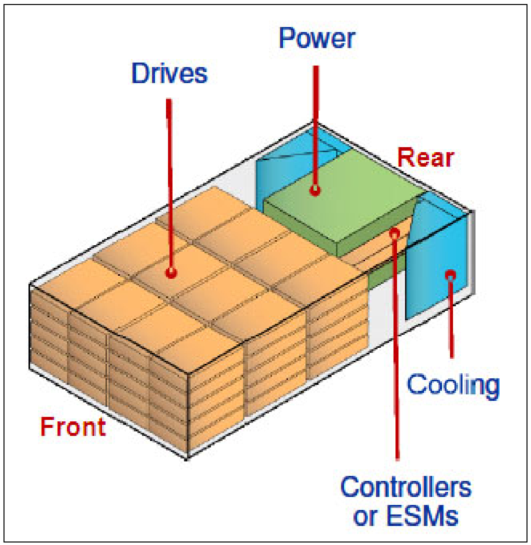


Figure 3-83 High-level view of the components of the DCS3700 chassis

Options for DCS3700 controller enclosures

This section lists the options for DCS3700 controller enclosures.

Base Controller (Snowmass)

The Base Controller (Snowmass) has the following cache options:

- 4 GB (Base Cache) (FC 3000)
- 8 GB (FC 3001)

The Performance Module Controller (Pikes Peak) includes two Performance Module Controller canisters, and two FC host port connections per module (with a blank slot to add a host interface card (HIC) with more ports). One SAS port provides connections to the expansions and drives. Eight partitions are included in the base.

The Base Controller has the following cache options:

- 12 GB (Base Cache) (FC 3110)
- 24 GB (FC 3111)
- 48 GB (FC 3112)

DCS3700 (1818-80E / 2851-DE2) expansion enclosure

The expansion unit (see Figure 3-84) consists of two environmental services modules (ESMs), 60 drive bays, and dual AC power supplies and cooling units. The expansion enclosures are connected to the DCS3700 controller enclosure with two SAS cables for redundancy in a daisy-chain fashion.

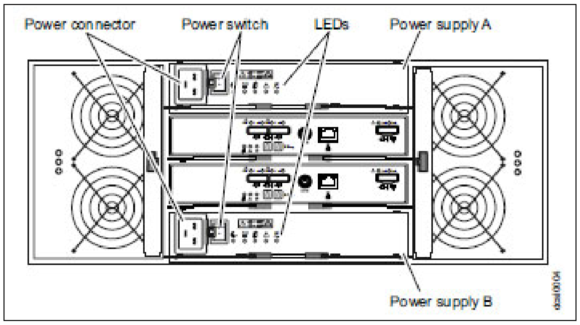


Figure 3-84 Rear view of a DCS3700 expansion enclosure

Enclosure interconnectivity

The enclosures (shown in Figure 3-85) are interconnected with wide SAS cables (2 x 6 Gbps) in a daisy-chain fashion. The SAS cables must be correctly connected as shown. five expansions per controller chain), see the DCS3700 documentation.

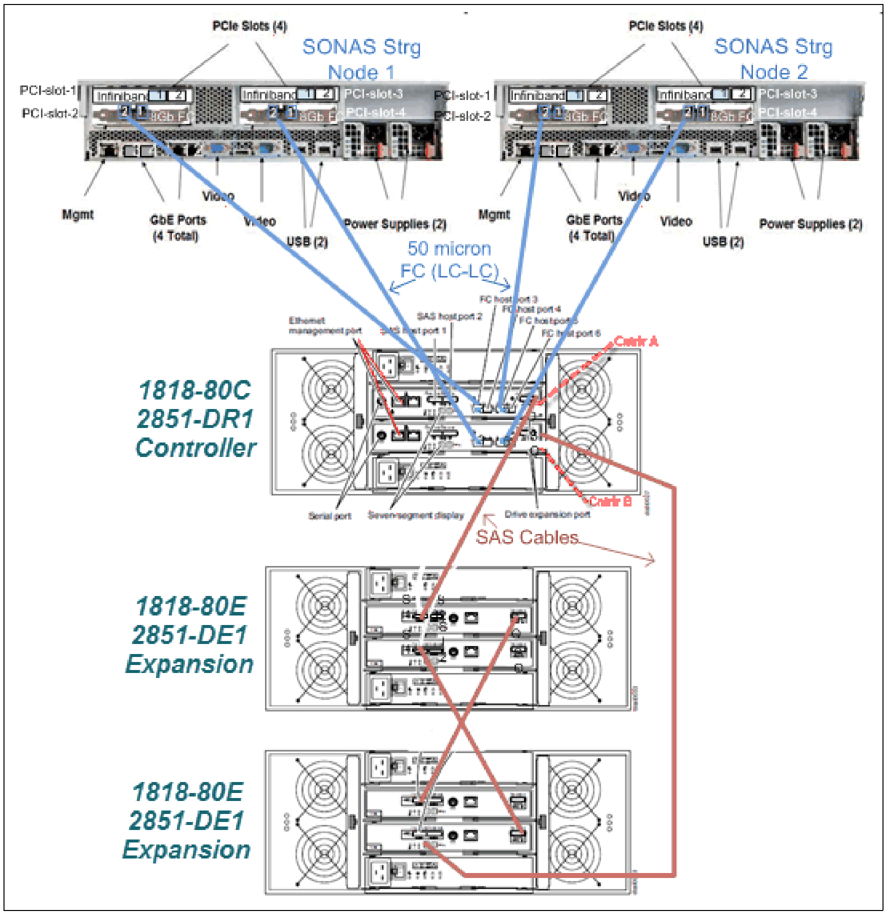


Figure 3-85 One controller with two expansion enclosures in a SONAS pod

Supported drives

Various 3.5-inch and 2.5-inch drives are supported in the DCS3700 enclosure:

- 3.5-inch disk drives: 2 TB, 3 TB, and 4 TB 7.2 K RPM Near-Line SAS disk
- 2.5-inch disk drives:
  - 300 GB 15 K RPM SAS disk1
  - 600 GB, 900 GB, 10 K RPM SAS disk
  - 200 GB, 400 GB, 800 GB E-MLC (enterprise-grade multilevel cell) solid-state drive (SSD)

Note: All hard disk drives (HDD) types and sizes and all SSDs that are supported by DCS3700 storage are supported when they are attached to an IBM SONAS Gateway system. For more information about the current drive supported types, see the IBM Knowledge Center at the following website:

`http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/dcs_configdcs37kstorage.html?lang=en`

Available RAID levels

Here is a list of the available RAID levels for DCS3700 storage:

- RAID 0 (Not supported for SONAS)
- RAID 1 (Not supported for SONAS)
- RAID 3 (Not supported for SONAS)
- RAID 5 (Block-level striping with striped parity)
- RAID 6 (Same as RAID 5, except with double parity; can survive two drive failures)
- RAID 10 (Striped mirror: high performance with redundancy)
- RAID DDP (Dynamic Disk Pools or just Disk Pools, is a new RAID level that is based on RAID 6.)

> **Note:** RAID levels 0, 1, and 3, are *not* supported by SONAS. Use RAID 5 *only* with SSDs, for which high density with redundancy is required. Use RAID 10 only for SSD. RAID 10 is automatically configured by choosing RAID 1 with four or more disks in the array configuration. It is not supported to use fewer than four disks in a RAID 1 array.

RAID details

RAID 6, disk pools, or both are the preferred RAID types for use with SONAS. Because Dynamic Disk Pool (DDP) is a new feature, more details are provided here:

- RAID 6: RAID 6 arrays stripe data over the member drives with two parity stripes (known as the P-parity and the Q-parity) on every stripe. The two parity strips are calculated by using a different algorithm, which gives the array double redundancy. A RAID 6 array can be formed by 5 - 16 drives.
- Disk pools: The DDP feature is a new way to deliver RAID protection and consistent performance. A disk pool provides the overall capacity that is needed to create one or more logical drives. A disk pool is similar to an array, with the following differences:
  - The data in a disk pool is stored randomly on all of the drives in the disk pool, unlike data in an array, which is stored on the same set of drives.
  - You do not specify a RAID level for a disk pool.
  - A disk pool does not use hot spare drives.
  - A disk pool allows as many drives to be grouped.
  - Each disk pool must have a minimum of 11 drives.
  - Although there is no limit to the maximum number of drives that can comprise a disk pool, the disk pool cannot contain more drives than the maximum limit for each storage subsystem.

LUNs that are created from disk pools share all disks within the pool. The advantage of disk pools is that rebuild times are greatly reduced, and the effect of a drive rebuild on any LUN is greatly reduced. Consider disk pools when you configure DCS3700 storage for SONAS. The drives in each disk pool must be of the same type, size, and speed. SSDs cannot be in a disk pool. If a control or expansion enclosure contains only 10 or fewer drives of identical type, size, and rotational speed, use a traditional RAID configuration for those drives.

- DDP benefits:
  - Easy to create: It is easy to create a disk pool by using the storage management software. To create a disk pool, you select the drives from a list of eligible drive candidates. After a disk pool is created, you create logical drives. When you create disk pool logical drives, the only attribute you must specify is the logical drive capacity.
  - Reduced hot spots: In a disk pool, the hot spots are minimized, as compared to an array because of the random manner in which the data is spread across many drives. The reduction of hot spots in the disk pool improves the performance of the storage subsystem.
  - Faster reconstruction of data: Disk pools do not use hot spare drives for data protection like an array does. Disk pools use spare capacity within each drive that comprises the disk pool. With disk pools, the reconstruction of data is much faster than arrays because the spare capacity in all of the drives that comprise the disk pool is used. It is not limited to one disk. Additionally, the data to reconstruct after a drive failure is reduced because the data is spread randomly across more drives in a disk pool than an array. Faster reconstruction of data in a disk pool also reduces the risk of more drive failures during a reconstruction operation. Unlike arrays, the period for which the disk pool is exposed to multiple drive failures during a reconstruction operation is reduced.
  - Consistent performance: DDPs deliver and maintain exceptional performance under all conditions, whether optimal or under a drive failure or rebuild. DDP minimizes the performance impact of a drive failure in multiple dimensions. By distributing parity information and spare capacity throughout the disk pool, DDP can use every drive in the pool for the intensive process of rebuilding a failed drive. This dynamic rebuild process can return the system to optimal condition faster than traditional RAID.
- Considerations:
  - Dynamic Segment Sizing (DSS) is not supported for disk pools. The segment size is set to 128 KB. Because the segment size is 128 KB, strongly consider a GPFS block size of 1 M.

    > **Note:** Initial testing has shown that a GPFS block size of 1 M has better performance with disk pools than a block size 256 KB.

  - You cannot change the RAID level of a disk pool. The storage management software automatically configures disk pools as RAID 6.
  - You cannot export a disk pool from a storage subsystem or import the disk pool to a different storage subsystem.

- All drive types (Fibre Channel, SATA, and SAS) in a disk pool must be the same type. SSD drives are not supported.
- If you revert to an earlier version of the controller firmware for a storage subsystem that is configured with a disk pool (and that version does not support disk pools), the logical drives are lost and the drives are treated as unaffiliated with a disk pool.

Disk pool availability

Higher availability of disk pools, concerning tray loss, can be achieved when at least one expansion enclosure is present and the number of drives within a pool on the same tray is limited to two. For disk pools, this configuration is made by creating a number of pools that is proportional to the number of enclosures, such that there are no more than two drives that are required per tray, and allowing the DCS3700 SM GUI to select the drives across enclosures. Thus, for example, with enclosures of identical drives, pools create pools according to the guidelines in Table 3-2.

Table 3-2 Disk pool example

Number of expansions in string	Total drives in each pool	Total number of disk pools
1	20	6
2	30	6
3	40	6
4	50	6
5	60	6

Disk pools do not have spare disks, but spare and preservation capacity, which is the amount of reserved spare space within the pool for rebuilding the data that is stored on failed disk drives. The preservation capacity is measured in terms of “drives” worth of spare capacity (one drive, two drives, three drives, and so on) and increases as the number of physical disk drives in the pool increases. Table 3-3 lists the default preservation drive count based on the number of physical disk drives in the pool. Ideally, use the default preservation drive count.

Table 3-3 Disk pool preservation capacity size

Number of drives in the disk pool	Default preservation drive count
11	1
12 - 31	2
32 - 63	3
64 - 27	4
128 - 191	6
192 - 256	7
256 or more	8

Setting aside enough reserve capacity to rebuild an entire tray of drives can provide the highest level of protection and control. By following the pool sizing for high availability with one or more enclosures, if you set aside two or more drives of preservation capacity, you can enable a “tray stop” procedure if you know in advance that a tray must be repaired. Unlike for traditional RAID groups, all drives are still used.

If you need to service a tray, the active drives within the tray can be stopped one at a time, during off hours if necessary. You can rebuild the pools one at a time with minimal performance impact and risk of data loss. After all of the drives in a tray are stopped, the service action can be done. After the service action, when the drives are brought online, the data in the pools is gradually rebalanced to return the pools to their original state.

3.6.2 Maximum DCS3700 configuration

This example is the maximum configuration that is available with the SONAS DCS3700 Gateway solution. It consists of a combination of Interface and Storage nodes for a total of 34 nodes. This configuration provides 14.4 or 28.8 PB of raw SONAS storage, depending on DCS3700 controller. It includes the following components:

- One SONAS Base Rack (2851-RXA)
  - Eight Interface nodes
  - Ten Storage nodes.
- One SONAS Interface Expansion Rack (2851-RXC)
  - Six Interface nodes
  - Ten Storage nodes
- Twenty DCS3700 Controllers (Snowmass), two expansions enclosures each with 60 X 4 TB drives
  - Total raw capacity = 240 TB per DCS3700 enclosure.
  - Total raw capacity = 720 TB per DCS3700 controller with two expansion enclosures.
  - Total raw capacity = 1.4 PB per storage pod or pair
  - Total raw capacity = 14.4 PB total solution (20 Storage nodes and 20 DCS3700 controllers)
- Twenty DCS3700 Performance Controllers (Pikes Peak) with five expansion enclosures with 60 x 4 TB disks drives in each.
  - Total raw capacity = 240 TB per DCS3700 enclosure.
  - Total raw capacity = 1.4 PB per DCS3700 controller with five expansion enclosures.
  - Total raw capacity = 2.8 PB per storage pod or pair
  - Total raw capacity = 28.8 PB total solution (20 Storage nodes and 20 DCS3700 controllers)

Note: The RXC Interface Expansion Frame as of SONAS V1.4.1 allows Gateway expansion by allowing the integration of Storage node pairs in the frame to extend the cluster to up to 20 storage pods (with up to 20 DCS3700 subsystems and up to 100 expansion enclosures).

Note: In large configurations, Fibre Channel and SAS link throughput must be considered. For example, on a pair of 8 Gb Fibre Channel links, the theoretical maximum throughput is ~1.6 GBps. SAS links between enclosures is limited to 6 GBps per link or 1.2 GBps per pair. As storage capacity increases, so do the throughput and bandwidth requirements. More capacity is available behind a limited number of ports and data paths. In some cases, IOPS and throughput are more important than storage capacity.

3.6.3 Installation and configuration information for SONAS with DCS3700 storage

DCS3700 storage consists of an enclosure (cabinet) of disks that are grouped behind a dual set of clustered controllers.

Each controller enclosure manages 60 x 3.5-inch drives or 60 x 2.5-inch drives stacked horizontally (12 per tray in the five trays of each controller). The expansion enclosure does not have the controllers, but can be configured with 60 drives in the same configuration. The DCS3700 enclosures are linked together through redundant SAS cables. The DCS3700 base controller can provide 4 GB or 8 GB of cache and manage up to two expansion arrays in each instance, with a minimum capacity of 20 drives (in a single 4U controller type array) and a maximum capacity of 180 drives per DCS3700 system. The DCS3700 Performance Controller (FC 3100) can provide 12 GB, 24 GB, or 48 GB of cache and manage up to five expansion enclosures for a total of 300 disks drives.

The DCS3700 can manage multiple types and sizes of disks, including SSD, SAS, or Nearline SAS drives. The DCS3700 allows the different drive technologies be placed in the same enclosures to create a more customizable configuration.

Like the XIV solution, the enclosures can be partially populated. For example, you can use the speed of SSD to use it as the “metadata” container in the GPFS “system pool”. You might populate a DCS3700 controller frame with 20 or more SSDs in a mirror (RAID 10), plus a hot spare in the configuration, and yield exceptional performance from it for metadata processing speed acceleration support. This configuration can improve IOPS to and from metadata-intense operations, such as backup, restore, async replication, and antivirus scanning.

Important: Consider all aspects of performance for metadata. It is more complicated than choosing 20 SSDs because correct sizing, performance, and reliability considerations are important to the success of each configuration. A preferred practice is to consider 20 as a minimum for small file systems (full analysis not considered). Also, for RAID 10 mirrored configurations, it is commonly considered preferred practice to allocate a global hot spare for the RAID groups to reduce the likelihood of a second drive failure within mirror sets.

Also, consider spreading SSD drives between two DCS3700 control enclosures to increase performance versus placing all SSD drives in one enclosure.

The SONAS Gateway with DCS3700 storage supports RAID 6, RAID DDP for all non-SSD disks, and RAID 10 and RAID 5 for SSD. Any disk type that DCS3700 storage supports is also supported for use in SONAS.

Optimal configuration of DCS3700 storage is more complicated than for other storage types behind SONAS.

Ensuring that the DCS3700 storage is properly configured

Ensure that the DCS3700 storage is properly configured in terms of power, cooling, Ethernet, and Call Home. This process is typically done by an IBM Certified Engineer.

Cabling DCS3700 controllers to Storage nodes

The DCS3700 controllers must be properly cabled to maintain a correctly working SONAS. Fibre Channel Direct Connect is the only supported method for connecting the DCS3700 controller to the SONAS Storage nodes. Fibre Channel SAN connectivity is not supported. Figure 3-86 shows the correct connectivity.

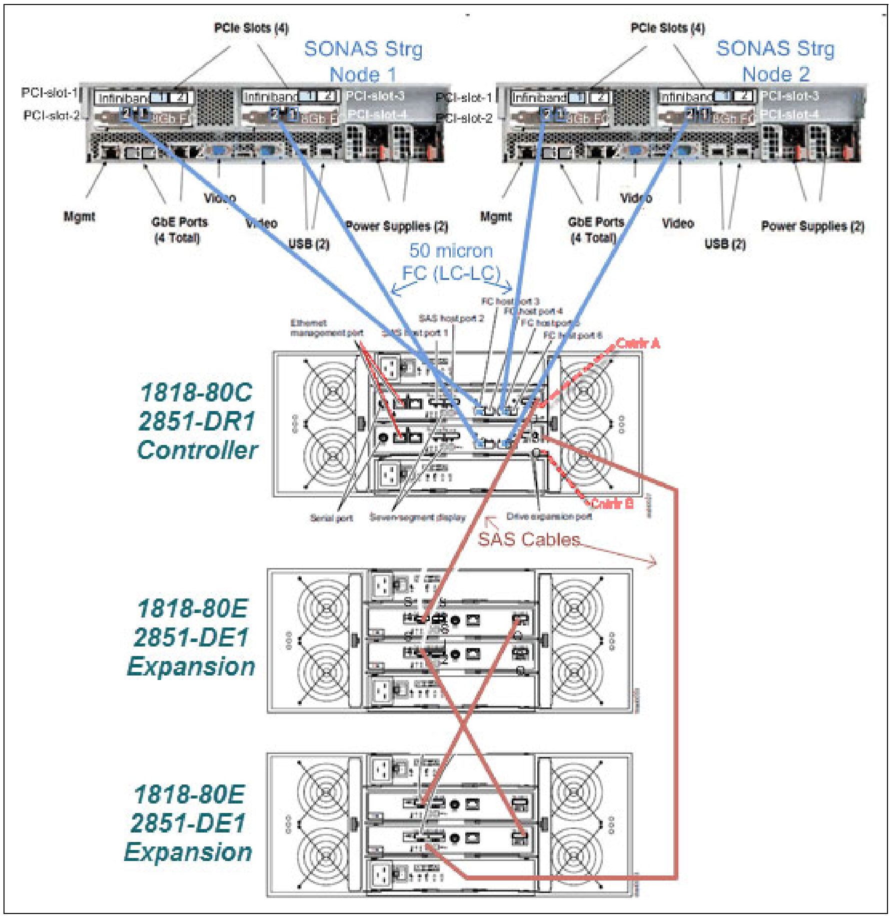


Figure 3-86 Single controller with two expansion enclosures in a SONAS pod

The DCS3700 is wired by using LC to LC Fibre Channel cables.

The only supported method of connecting SONAS to DCS3700 is with direct connection from the storage controllers to the active FC ports in the Storage node pair by using LC to LC cables, as shown in Figure 3-87.

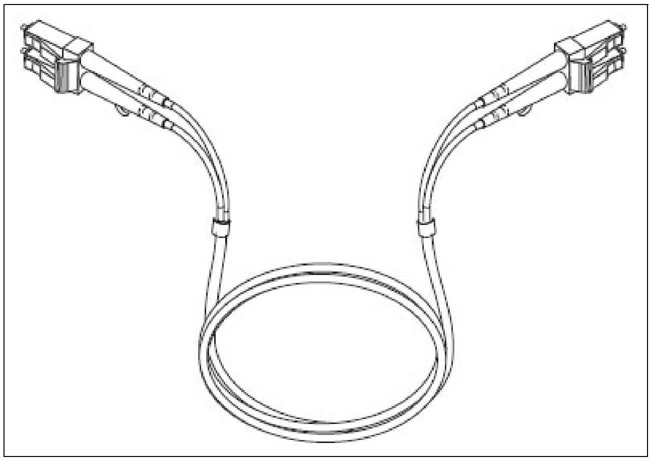


Figure 3-87   LC to LC cable diagram

The first DCS3700 controller is connected to port 2 on each HBA of each Storage node (as shown in Figure 3-86 on page 184), and the second DCS3700 controller is connected to port 1 on each HBA of each Storage node, as shown in Figure 3-88.

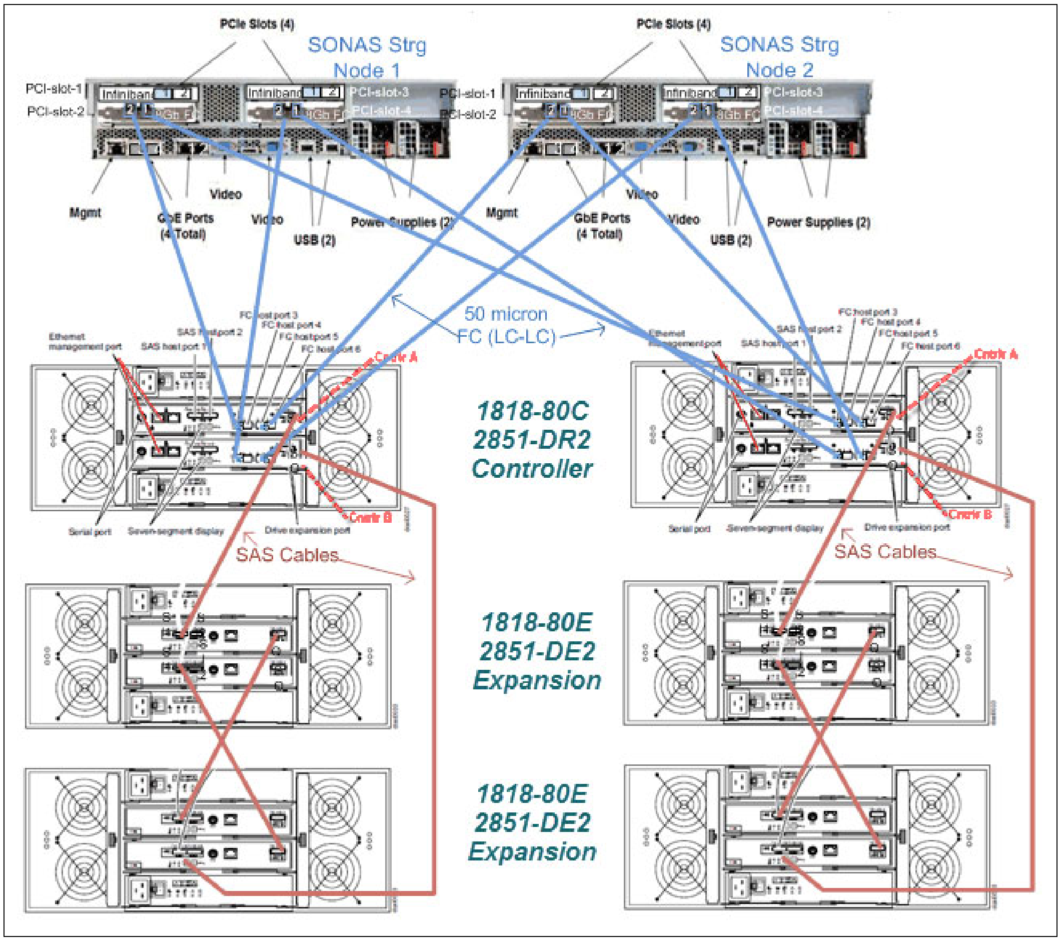


Figure 3-88   Two controllers with two expansion enclosures behind each controller in a SONAS pod

The cables are wired to port 2 of each SONAS Storage node for the first DCS3700 controller, and then to port 1 of each HBA when the second DCS3700 controller is added to the pod. With a single DCS3700 controller in the pod, you have one inactive FC port on each HBA in each of the two Storage nodes. However, a 16 Gb bandwidth is available per DCS3700 per Storage node in this design.

DCS3700 management host software

The DCS3700 is installed separately from the SONAS storage solution. As for all Gateway solutions, when you install and configure your DCS3700 storage, follow the process and recommendations of *IBM System Storage DCS3700 Installation, User's, and Maintenance Guide*, GA32-0959-02.

This chapter presents information that is specific to the SONAS gateway solution with the DCS3700 subsystem.

The DCS3700 subsystem is managed through a designated Management host that has the DS Storage Management Software Suite installed. The Management Host is critical for configuring the Storage, RAID Groups, volumes, and host mappings, but also for passing storage alerts.

The DCS3700 subsystem does not function as a Call Home storage device without the aid of a designated management host. For this reason, configure a primary and secondary management host.

Notes:

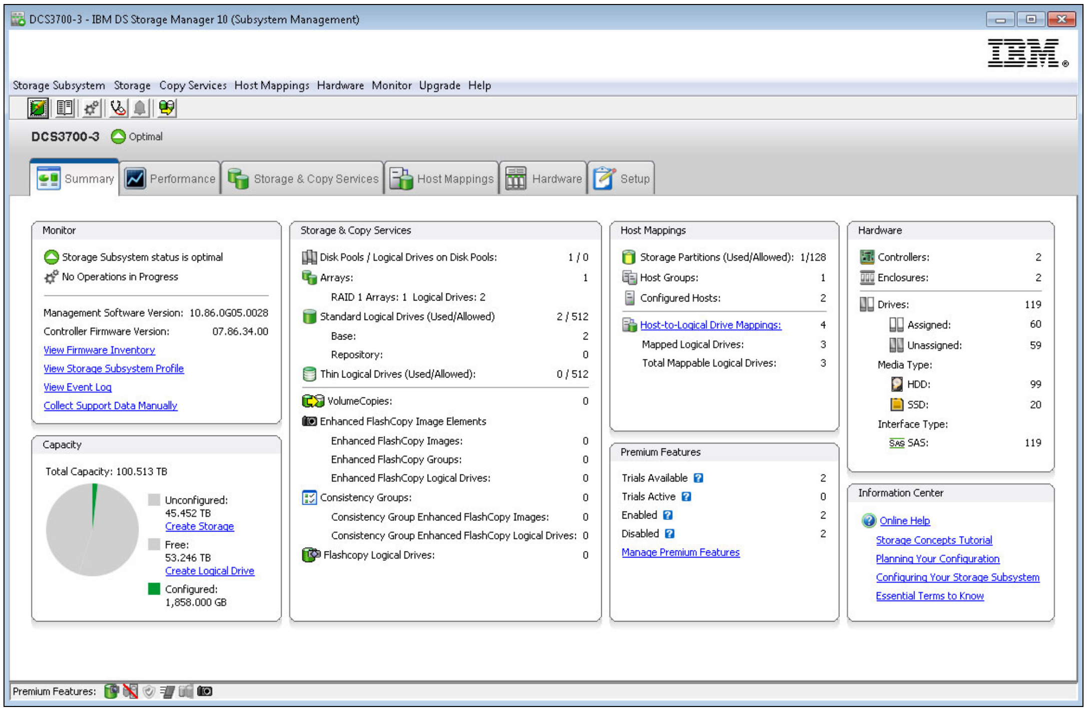
- You can monitor only storage subsystems that are within the management domain of the storage-management software.
- If you have not installed the DS Storage Manager Event Monitor service as part of the storage-management software installation, the Storage Manager Enterprise Management window must remain open. (If you close the window, you will not receive any alert notifications from the managed storage subsystems.)

For the applicable operating system instructions for installing the Storage Manager software, see the Enterprise Management online help and the *DS Storage Manager 10 Installation and Host Support Guide*. The guide is in the Documentation folder on the IBM Support Software DVD.

To download the latest version of the Storage Manager software, controller firmware, NVSRAM firmware, and the latest ESM firmware, go to the following website:

https://www.ibm.com/support/entry/portal/support

Figure 3-88 on page 185 shows the DCS3700 GUI view from the designated Windows server.


DCS3700-3 - IBM DS Storage Manager 10 (Subsystem Management)
IBM
Storage Subsystem Storage Copy Services Host Mappings Hardware Monitor Upgrade Help
Summary
Performance
Storage & Copy Services
Host Mappings
Hardware
Setup

DCS3700-3 Optimal

Monitor

Storage Subsystem status is optimal
No Operations in Progress

Management Software Version: 10.86.0G05.0028
Controller Firmware Version: 07.86.34.00
View Firmware Inventory
View Storage Subsystem Profile
View Event Log
Collect Support Data Manually

Capacity

Total Capacity: 100.513 TB
Unconfigured: 45.452 TB Create Storage
Free: 53.246 TB Create Logical Drive
Configured: 1,858.000 GB

Storage & Copy Services

Disk Pools / Logical Drives on Disk Pools: 1 / 0
Arrays: 1
RAID 1 Arrays: 1 Logical Drives: 2
Standard Logical Drives (Used/Allowed) 2 / 512
Base: 2
Repository: 0
Thin Logical Drives (Used/Allowed): 0 / 512
VolumeCopies: 0
Enhanced FlashCopy Image Elements
Enhanced FlashCopy Images: 0
Enhanced FlashCopy Groups: 0
Enhanced FlashCopy Logical Drives: 0
Consistency Groups: 0
Consistency Group Enhanced FlashCopy Images: 0
Consistency Group Enhanced FlashCopy Logical Drives: 0
Flashcopy Logical Drives: 0

Host Mappings

Storage Partitions (Used/Allowed): 1/128
Host Groups: 1
Configured Hosts: 2
Host-to-Logical Drive Mappings: 4
Mapped Logical Drives: 3
Total Mappable Logical Drives: 3

Premium Features

Trials Available 2
Trials Active 0
Enabled 2
Disabled 2
Manage Premium Features

Hardware

Controllers: 2
Enclosures: 2
Drives: 119
Assigned: 60
Unassigned: 59
Media Type:
HDD: 99
SSD: 20
Interface Type:
SAS: 119

Information Center

Online Help
Storage Concepts Tutorial
Planning Your Configuration
Configuring Your Storage Subsystem
Essential Terms to Know

Premium Features:

Figure 3-89 DCS3700 GUI view from the designated Windows server

> **Tip:** In some cases where an initial configuration `first_time_install` script is processed, additional configuration can be helpful. Before running the `first_time_install` script, spend time before the installation to update the firmware on the DCS3700 subsystems, add the Storage node hosts and SONAS cluster to the DCS3700 configuration, and create and assign the storage to the cluster.

Understanding the DCS3700 storage configuration process

This section provides a high-level overview of the configuration process. The process includes the following tasks:

- Position your hardware in the data center and ensure that all pre-installation requirements are reviewed and met.
- Distribute the enclosure balance across the DCS3700 frames and controllers to establish the high distribution balance across the controllers.

  For example, you have two controllers and four expansion enclosures with SAS in the controller enclosure and NLSAS drives in each of the expansion enclosures below the controller enclosures. Balance the placement of those enclosures across the pod.

- Install the DCS3700 hardware, update the software/firmware to Version 07.86.40 or later, and set up Call Home support in the DS Manager host for each DCS3700 subsystem.

  The DCS3700 installation follows the DCS3700 Installation Guide (found at `http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_r_publications.html?lang=en`). This process is not described in this book. For more information, see the DCS3700 information in the IBM Knowledge Center, found at the following website:

  `http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_r_publications.html?lang=en`

- Connect two Ethernet cables to each DCS3700 controller enclosure (one to the first port on each controller) and prepare or reset the access IP addresses by following the detailed process that is specified in the DCS3700 Installation Guide. The enclosure is managed separately from the SONAS.

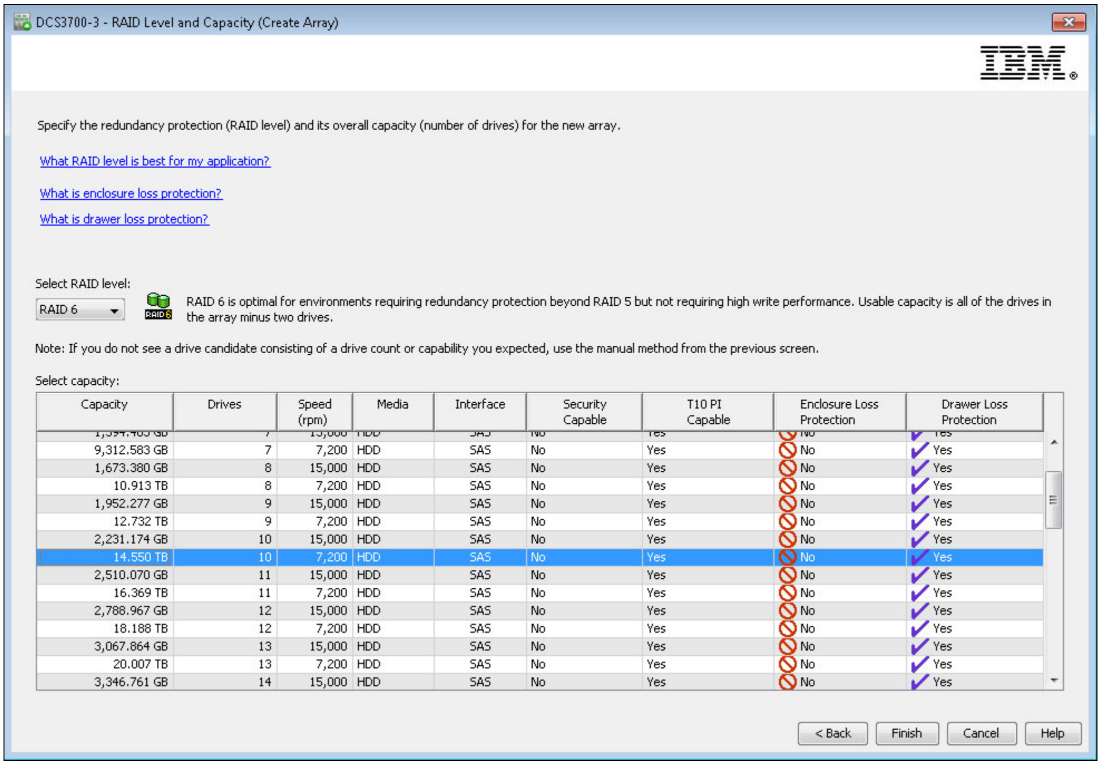
> **Note:** When you create DCS3700 LUNs and logical drives, make the number of drives divisible by 4, such as 4, 8, or 12. This division ensures that the LUNs are spread evenly across all controllers and canisters and Storage nodes. As LUNs are created, they are evenly allocated across both controllers and canisters in a DCS3700 frame.

Creating disk array groups

To create disk array groups, complete the following steps:

1. Create array groups that are evenly spread I/O across all five trays of each enclosure in groups of 10 drives.

   To simplify creation, use the Array Group wizard in the DCS3700 Manager GUI. The GUI tool automatically chooses the disks and layouts for maximum efficiency based on disk type and array size chosen. Create eight evenly sized array groups. This configuration means that you can create eight evenly sized LUNs (one from each array group). For details, see Figure 3-90.



DCS3700-3 - RAID Level and Capacity (Create Array)

Specify the redundancy protection (RAID level) and its overall capacity (number of drives) for the new array.

What RAID level is best for my application?

What is enclosure loss protection?

What is drawer loss protection?

Select RAID level: RAID 6

RAID 6 is optimal for environments requiring redundancy protection beyond RAID 5 but not requiring high write performance. Usable capacity is all of the drives in the array minus two drives.

Note: If you do not see a drive candidate consisting of a drive count or capability you expected, use the manual method from the previous screen.

Select capacity:

Capacity	Drives	Speed (rpm)	Media	Interface	Security Capable	T10 PI Capable	Enclosure Loss Protection	Drawer Loss Protection
1,394.403 GB	7	15,000	HDD	SAS	No	Yes	No	Yes
9,312.583 GB	7	7,200	HDD	SAS	No	Yes	No	Yes
1,673.380 GB	8	15,000	HDD	SAS	No	Yes	No	Yes
10.913 TB	8	7,200	HDD	SAS	No	Yes	No	Yes
1,952.277 GB	9	15,000	HDD	SAS	No	Yes	No	Yes
12.732 TB	9	7,200	HDD	SAS	No	Yes	No	Yes
2,231.174 GB	10	15,000	HDD	SAS	No	Yes	No	Yes
14.550 TB	10	7,200	HDD	SAS	No	Yes	No	Yes
2,510.070 GB	11	15,000	HDD	SAS	No	Yes	No	Yes
16.369 TB	11	7,200	HDD	SAS	No	Yes	No	Yes
2,788.967 GB	12	15,000	HDD	SAS	No	Yes	No	Yes
18.188 TB	12	7,200	HDD	SAS	No	Yes	No	Yes
3,067.864 GB	13	15,000	HDD	SAS	No	Yes	No	Yes
20.007 TB	13	7,200	HDD	SAS	No	Yes	No	Yes
3,346.761 GB	14	15,000	HDD	SAS	No	Yes	No	Yes

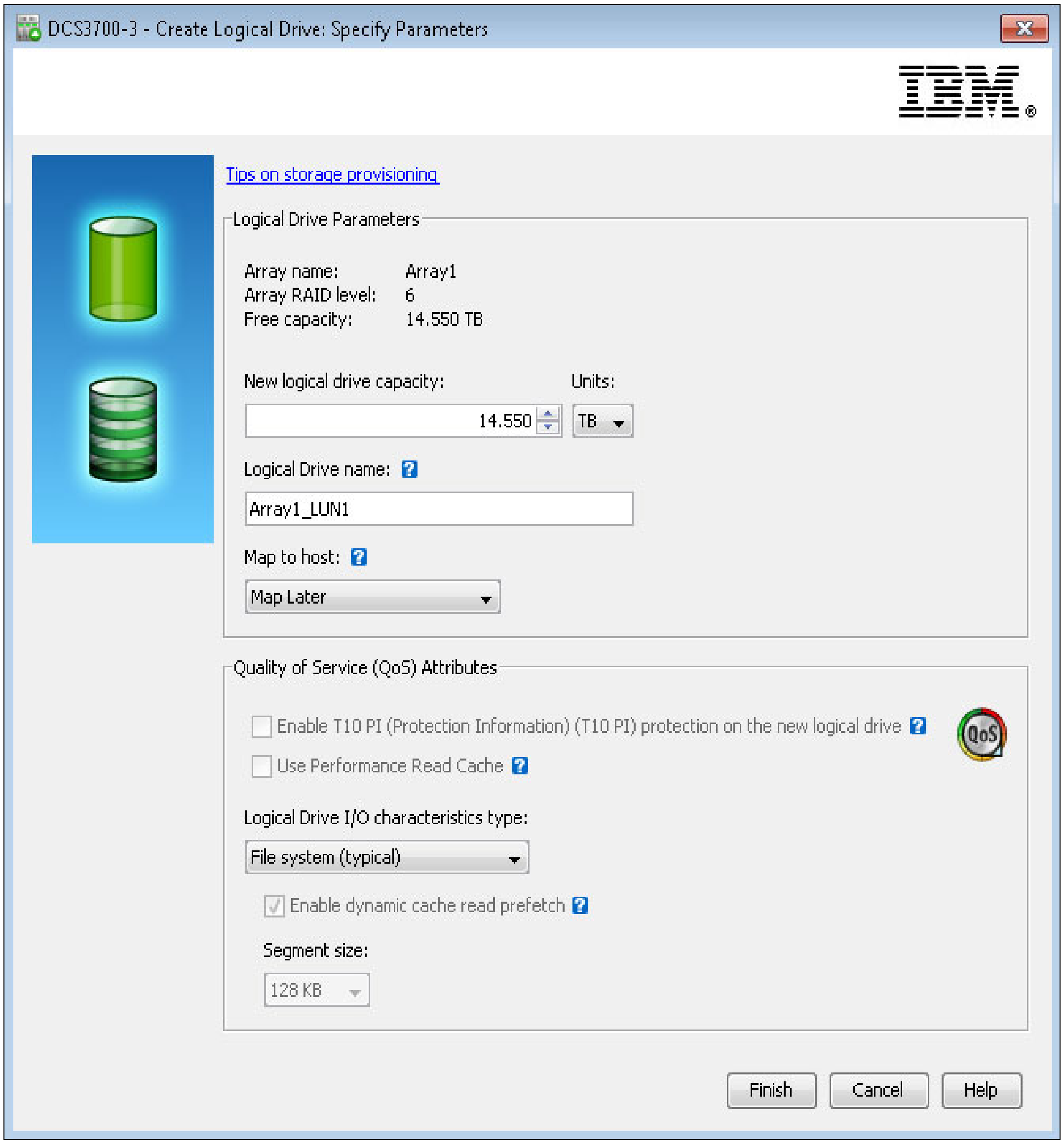
Figure 3-90   Create 10-disk RAID 6 arrays with 2 TB disks

2. Create one volume from each array that uses the entire space of the array and map them to the defined cluster, which contains both Storage node hosts. When you create the logical volume and LUN, you can choose the segment size. Determine the segment size based on the GPFS file system block size and the number of NSDs in the file system. Here are some examples:

   – If the GPFS file system is expected to hold mostly small files, set the file system block size to 256 KB. In this example, to spread evenly I/O across all controllers and all Storage nodes, it is preferable to configure eight LUNs, one from each array group with a segment size of 32 KB. 32 KB * 8 LUNs = 256 KB, which matches the block size of the file system.

   – If the GPFS file system is expected to hold larger files, set the file system block size to 1 MB. In this example, to spread evenly I/O across all controllers and all Storage nodes, it is preferable to configure eight LUNs, one from each array group with a segment size of 128 KB. 128 KB * 8 LUNs = 1024 KB, which matches the block size of the file system.

   For more details, see Figure 3-91.



DCS3700-3 - Create Logical Drive: Specify Parameters

Logical Drive Parameters

Array name: Array1
Array RAID level: 6
Free capacity: 14.550 TB

New logical drive capacity: 14.550
Units: TB

Logical Drive name: Array1_LUN1

Map to host: Map Later

Quality of Service (QoS) Attributes

□ Enable T10 PI (Protection Information) (T10 PI) protection on the new logical drive

□ Use Performance Read Cache

Logical Drive I/O characteristics type: File system (typical)

☑ Enable dynamic cache read prefetch

Segment size: 128 KB

Figure 3-91 Create a LUN for Array Group 1, where the segment size is 128 KB

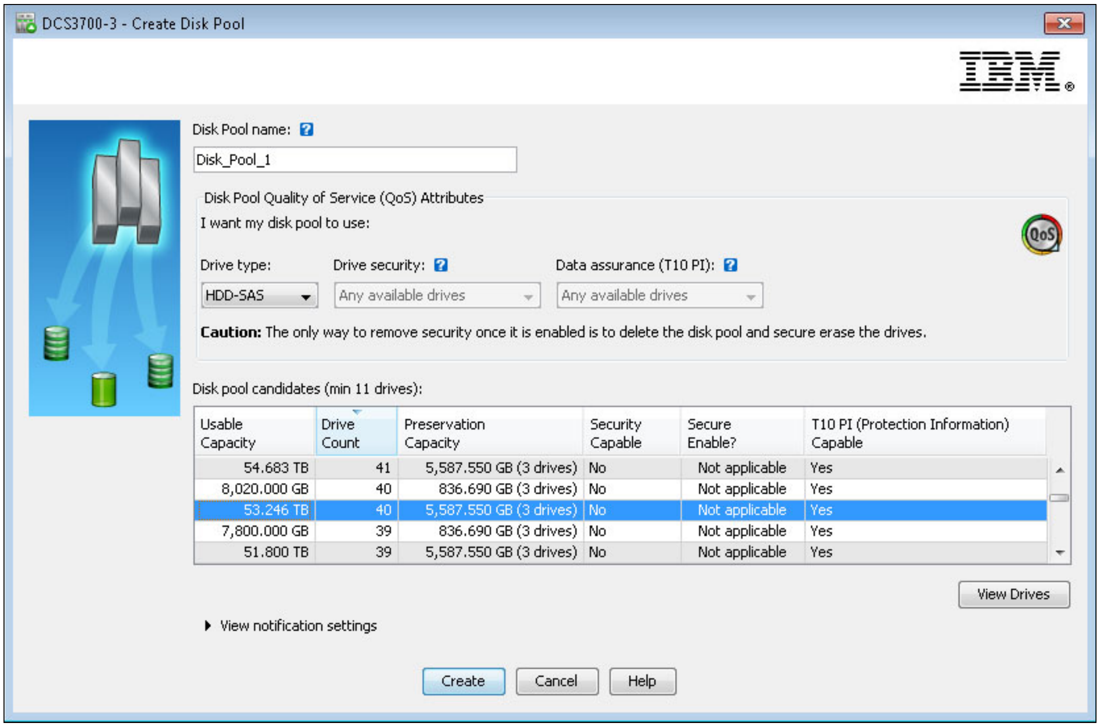
Note: For any GPFS file system, allocate LUNs and NSDs in a number that is divisible by 4. For example, create 4, 8, 12, or so on array groups and allocate one LUN from each array group. This configuration ensures that LUNs are spread evenly across all controllers and Storage nodes.

Creating disk pools

Disk pools are a new feature in SONAS V1.4.1 and offer advantages over regular RAID array groups (see “RAID details” on page 179.)

Use the Disk Pool wizard feature of the DCS3700 DS Manager GUI to create disk pools. This feature ensures proper layout to ensure the most efficient layout and provide the option to choose drawer and enclosure protection. When you create disk pools, choose a number of disks that is divisible by 10 because the underlying RAID protection of a disk pool LUN is a series of 4 GB RAID 6 stripes (8 + 2P).

For example, create a disk pool that uses 40 2-TB disks. This number is divisible by 10 and I/O is spread evenly across 20 disks in a two-enclosure configuration, as shown in Figure 3-92.



Disk Pool name: Disk_Pool_1

Disk Pool Quality of Service (QoS) Attributes

I want my disk pool to use:

Drive type: HDD-SAS

Drive security: Any available drives

Data assurance (T10 PI): Any available drives

Caution: The only way to remove security once it is enabled is to delete the disk pool and secure erase the drives.

Disk pool candidates (min 11 drives):

Usable Capacity	Drive Count	Preservation Capacity	Security Capable	Secure Enable?	T10 PI (Protection Information) Capable
54.683 TB	41	5,587.550 GB (3 drives)	No	Not applicable	Yes
8,020.000 GB	40	836.690 GB (3 drives)	No	Not applicable	Yes
53.246 TB	40	5,587.550 GB (3 drives)	No	Not applicable	Yes
7,800.000 GB	39	836.690 GB (3 drives)	No	Not applicable	Yes
51.800 TB	39	5,587.550 GB (3 drives)	No	Not applicable	Yes

Figure 3-92 Create a disk pool with 40 2-TB disks

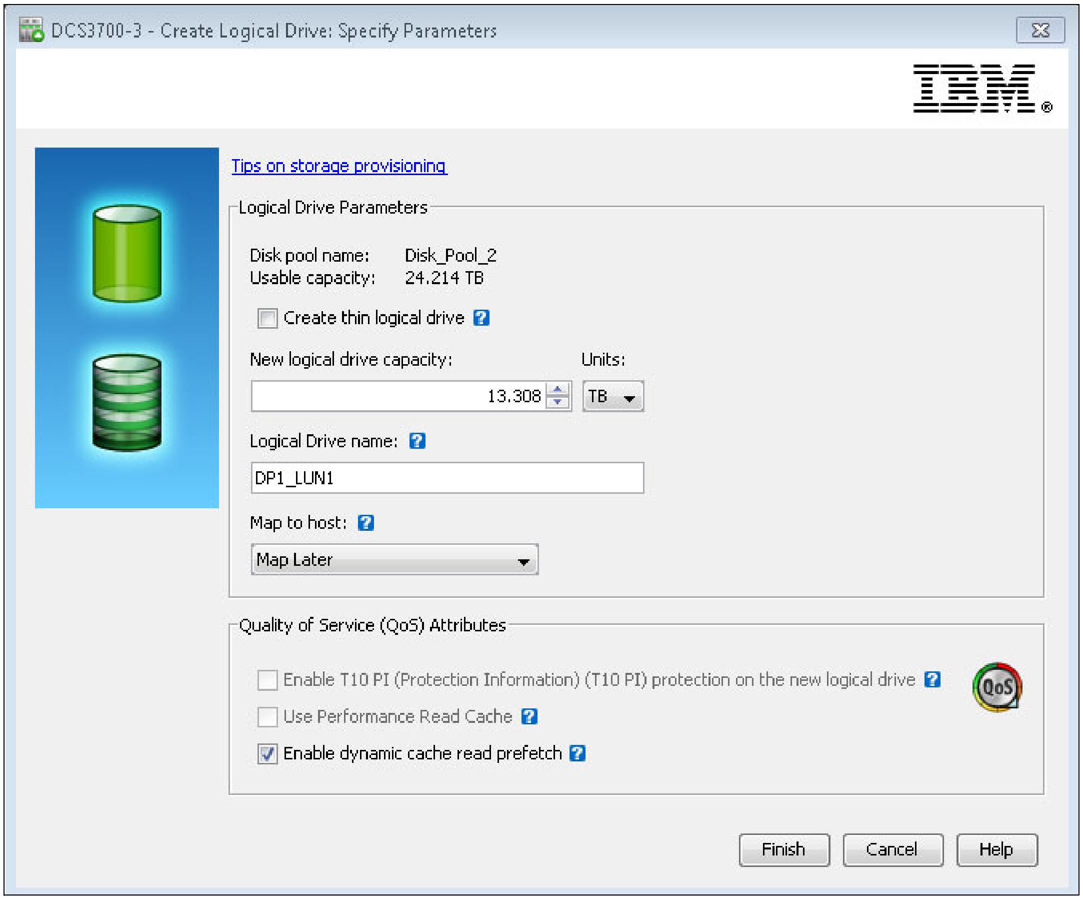
Configure disk pools in groups of 10 (11 drives are required because one drive must be set aside as a hot spare) disks. Use this configuration because LUNs are created as RAID 6 devices (8 + P + Q). LUNs that are created from a disk pool are configured as RAID 6 LUNs in 4 GB chunks with a segment size of 128 KB (the segment size cannot be changed). Make the number of equally sized LUNs the number of disks in the disk pool divisible by 10 (for example, 40 / 10 = 4). Also, allocate LUNs to SONAS in fours (4, 8, 12, and so on) to evenly distribute the LUNs across Storage nodes and storage controllers. Avoid using fewer than eight LUNs. LUNs are created in 4 GB chunks. Therefore, make the LUN size evenly divisible by 4 GB to maximize used capacity of the disk pool.

To determine the most space-efficient DDP LUN size, complete the following steps:

1. Look at the free space (GUI) when the pool is created (for example, 53.246 TB).
2. Convert this value to gigabytes (53.246 * 1024 = 54,523.904).
3. Divide the result by 4 (for a 4 GB chunk) and drop the remainder (54,523.904 / 4 = 13,630).
4. Divide the result by the number of LUNs and drop the remainder (13,630 / 4 = 3,407).
5. Multiply the result by 4 (a 4 GB chunk), for example, 3,407 * 4 = 13,628 GB.
6. Create four LUNs that are 13,628 GB *or* 13.308 TB.

Note: When you are using disk pools, set the GPFS file system block size to 1 M because the disk pool segment size is 128 KB. Initial testing has shown significant performance improvement when you use a 1 M versus a 256 K file system block size for disk pools.

Figure 3-93 shows the creation of a LUN in Disk Pool 1, where the size is 13308 TB.



DCS3700-3 - Create Logical Drive: Specify Parameters

Tips on storage provisioning

Logical Drive Parameters

Disk pool name: Disk_Pool_2
Usable capacity: 24.214 TB

□ Create thin logical drive

New logical drive capacity: 13.308
Units: TB

Logical Drive name: DP1_LUN1

Map to host: Map Later

Quality of Service (QoS) Attributes

□ Enable T10 PI (Protection Information) (T10 PI) protection on the new logical drive
□ Use Performance Read Cache
☑ Enable dynamic cache read prefetch

Figure 3-93 Create a LUN in Disk Pool 1, where the size is 13308 TB

Note: For any GPFS file system, allocate LUNs and NSDs in a number divisible by 4. Therefore, create 4, 8, 12, or so on, array groups and allocate one LUN from each array group. This configuration ensures that LUNs are spread evenly across all controllers and Storage nodes. Allocate a minimum of eight 8 LUNs per file system.

Updating the DCS3700 firmware

Carefully schedule DCS3700 firmware upgrades during periods of lower SONAS activity and cluster use.

Systems that are running SONAS firmware earlier than Version 1.5.1, correct DCS3700 firmware, and have NVSARAM and RDAC DCS3700 host types can have periods (up to 10 minutes) of low cluster throughput, which might negatively affect cluster usage. If possible, quiesce client I/O and advanced functions (backups, replication, snapshot schedules, and so on). Unmounting the file system is the only way to ensure that it is not affected by the DCS3700 firmware upgrade.

Systems running SONAS V1.5.1 firmware, proper DCS3700 firmware, and NVSRAM and ALUA DCS3700 host types can greatly reduce the impact to the cluster, which allows concurrent operations to continue. Carefully plan the DCS3700 firmware upgrade because it can strongly affect the system because logical drives fail over to non-preferred paths while a controller goes offline for the updates.

Here is a brief overview of the firmware update process for the DCS3700 subsystem.

> **Important:** This list is an overview of highlights and is in no way meant to replace the DCS3700 Installation Guide instructions (see `http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_r_publications.html?lang=en`). If you are unsure of the process, contact your local IBM account representative.

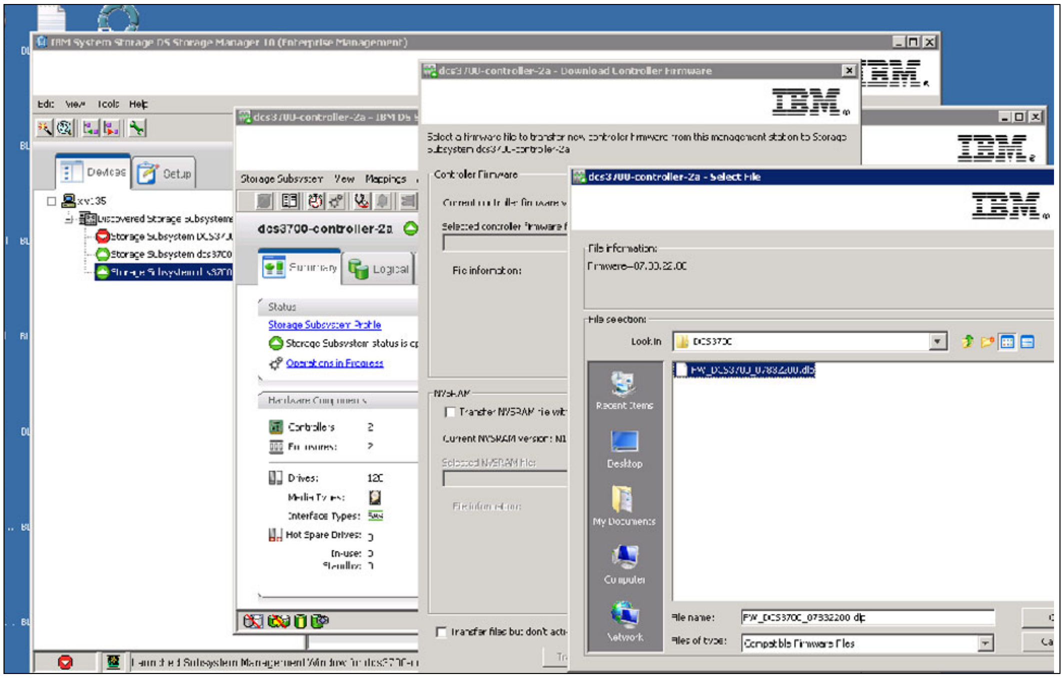
1. Place the DCS3700 firmware, NVSRAM, and ESM code in a directory on the DS Management Host. This example shows code placement in the following directory:

   ```
   C:\Desktop\DCS3700-FW\ibm_fw_ds3k_07832200_anyos_anycpu\Controller_Code_07832200_DS8K\
   ```

   > **Note:** Before you complete a DCS3700 firmware update, review and validate all events from the DCS3700 Storage Manager. Clear all events before you run the update because the update process fails if there are uncleared events.
   >
   > It is a preferred practice to confirm that there is a good cluster health state in the SONAS GUI and to ensure that there are no existing errors that can have a negative effect on the DCS3700 firmware upgrade.

2. Use the DS Storage Manager GUI to update the firmware. The figures in this step show highlights of the process.

   Figure 3-94 on page 193 shows the firmware code selection.



IBM System Storage DS Storage Manager 10 (Enterprise Management)
Edit View Tools Help
Devices
Setup
dcs3700-controller-2a - IBM DS
Storage Subsystem View Mappings
dcs3700-controller-2a
Summary
Logical
dcs3700-controller-2a - Download Controller Firmware
dcs3700-controller-2a - Select File
Recent Items
Desktop
My Documents
Computer
Network

Select a firmware file to transfer new controller firmware from this management station to Storage subsystem dcs3700-controller-2a

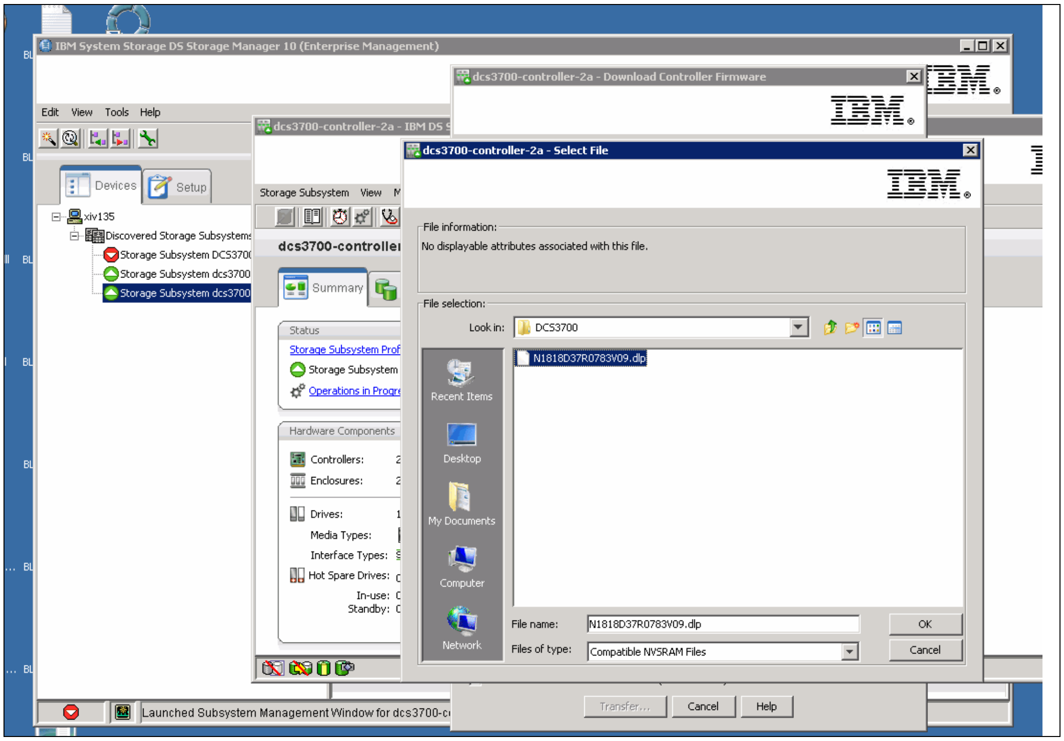
File information:
Firmware=07.83.22.00

File selection:
Look In: DCS3700
FW_DCS3700_07832200.dlp

File name: FW_DCS3700_07832200.dlp
Files of type: Compatible Firmware Files

Figure 3-94 DCS3700 FW update - firmware code selection

Figure 3-95 shows the NVSRAM code selection.



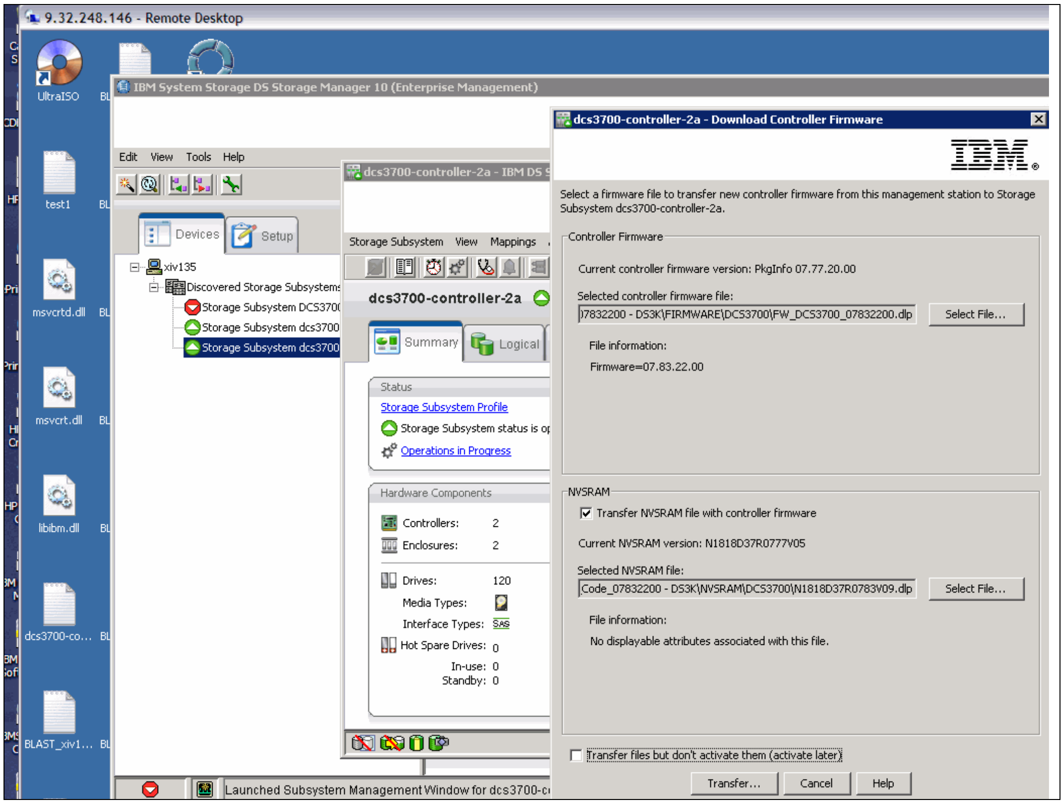
File information:
No displayable attributes associated with this file.

File selection:
Look in: DCS3700
N1818D37R0783V09.dlp

File name: N1818D37R0783V09.dlp
Files of type: Compatible NVSRAM Files

Figure 3-95 DCS3700 FW update - NVSRAM code selection

Figure 3-96 shows that the firmware and NVSRAM are selected.


9.32.248.146 - Remote Desktop
UltraISO
test1
msvcrtd.dll
msvcrt.dll
libibm.dll
dcs3700-co...
BLAST_xiv1...
IBM System Storage DS Storage Manager 10 (Enterprise Management)
Edit View Tools Help
Devices
Setup
xiv135
Discovered Storage Subsystems
Storage Subsystem DCS3700
Storage Subsystem dcs3700
Storage Subsystem dcs3700
dcs3700-controller-2a - IBM DS
Storage Subsystem View Mappings
dcs3700-controller-2a
Summary
Logical
Status
Storage Subsystem Profile
Storage Subsystem status is op
Operations in Progress
Hardware Components
Launched Subsystem Management Window for dcs3700-c

Controllers: 2
Enclosures: 2
Drives: 120
Media Types:
Interface Types: SAS
Hot Spare Drives: 0
In-use: 0
Standby: 0

dcs3700-controller-2a - Download Controller Firmware

Select a firmware file to transfer new controller firmware from this management station to Storage Subsystem dcs3700-controller-2a.

Controller Firmware

Current controller firmware version: PkgInfo 07.77.20.00

Selected controller firmware file:)7832200 - DS3K\FIRMWARE\DCS3700\FW_DCS3700_07832200.dlp [Select File...]

File information:

Firmware=07.83.22.00

NVSRAM

☑ Transfer NVSRAM file with controller firmware

Current NVSRAM version: N1818D37R0777V05

Selected NVSRAM file: Code_07832200 - DS3K\NVSRAM\DCS3700\N1818D37R0783V09.dlp [Select File...]

File information:

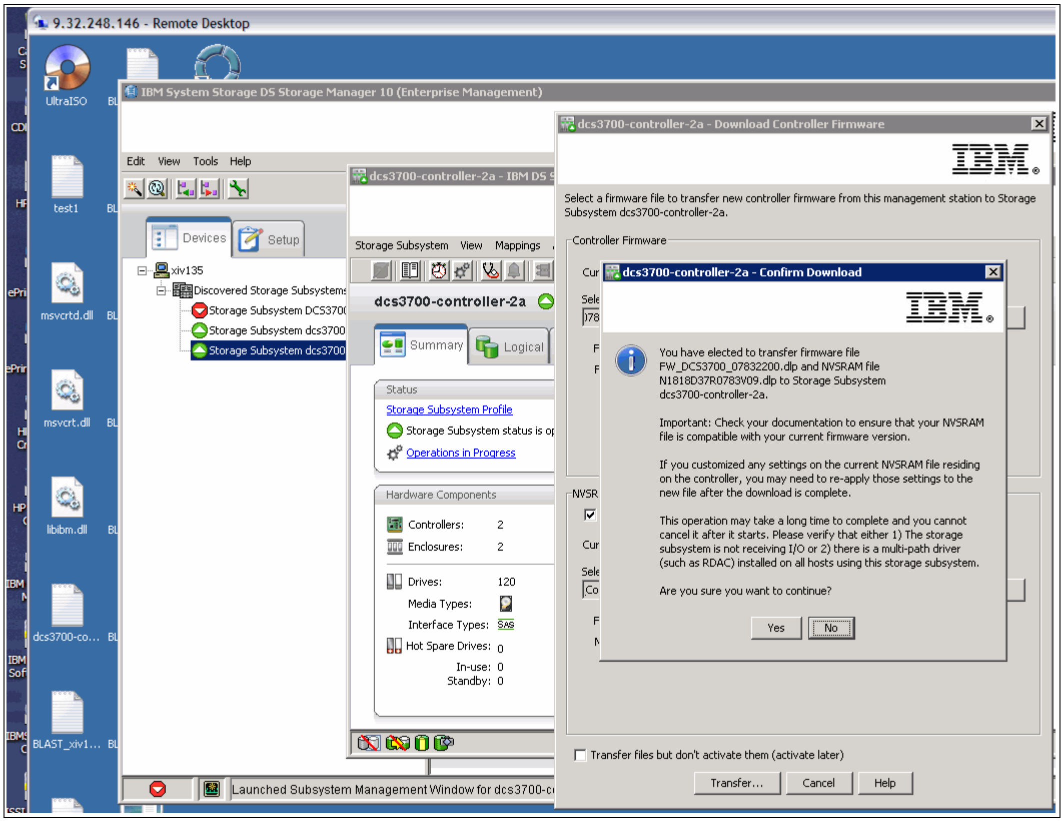
No displayable attributes associated with this file.

☐ Transfer files but don't activate them (activate later)

[Transfer...] [Cancel] [Help]

Figure 3-96 DCS3700 FW update - firmware and NVSRAM code selection

Figure 3-97 on page 195 shows the confirmation to continue message for the firmware update.


9.32.248.146 - Remote Desktop
IBM System Storage DS Storage Manager 10 (Enterprise Management)
dcs3700-controller-2a - Download Controller Firmware
dcs3700-controller-2a - Confirm Download
Devices
Setup
Summary
Logical
Yes
No
Transfer...
Cancel
Help

Select a firmware file to transfer new controller firmware from this management station to Storage Subsystem dcs3700-controller-2a.

You have elected to transfer firmware file FW_DCS3700_07832200.dlp and NVSRAM file N1818D37R0783V09.dlp to Storage Subsystem dcs3700-controller-2a.

Important: Check your documentation to ensure that your NVSRAM file is compatible with your current firmware version.

If you customized any settings on the current NVSRAM file residing on the controller, you may need to re-apply those settings to the new file after the download is complete.

This operation may take a long time to complete and you cannot cancel it after it starts. Please verify that either 1) The storage subsystem is not receiving I/O or 2) there is a multi-path driver (such as RDAC) installed on all hosts using this storage subsystem.

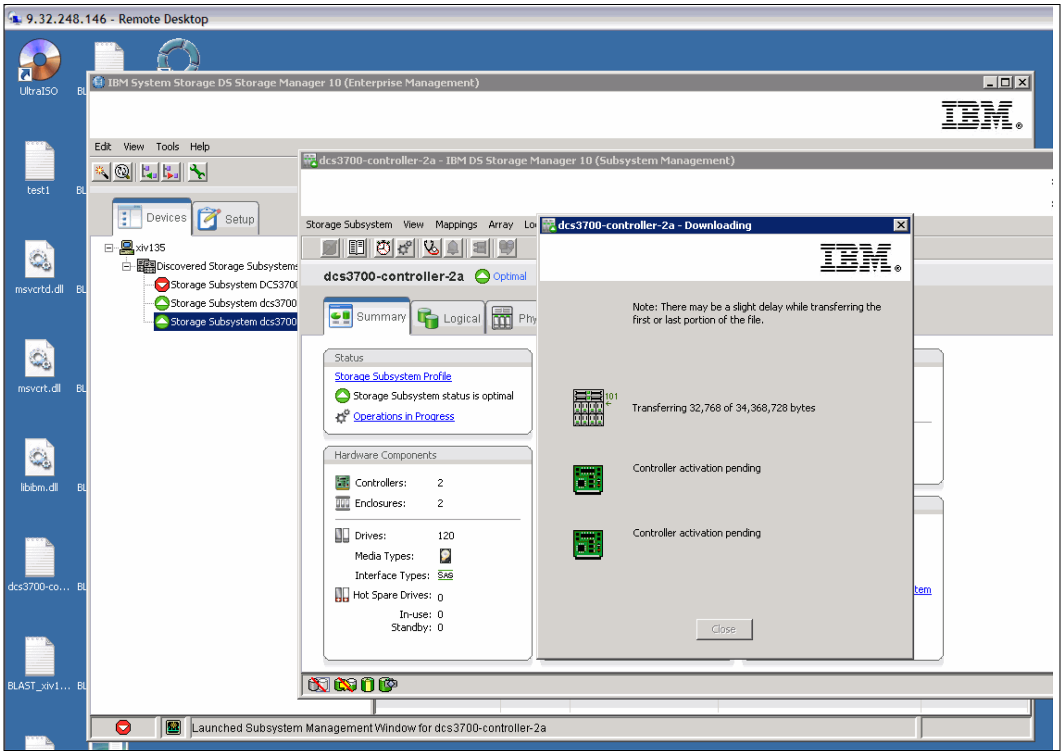
Are you sure you want to continue?

☐ Transfer files but don't activate them (activate later)

Launched Subsystem Management Window for dcs3700-c

Figure 3-97 Confirmation to continue message for the firmware update

Figure 3-98 shows the firmware update progress.


9.32.248.146 - Remote Desktop
IBM System Storage DS Storage Manager 10 (Enterprise Management)
dcs3700-controller-2a - IBM DS Storage Manager 10 (Subsystem Management)
dcs3700-controller-2a - Downloading
Devices
Setup
Summary
Logical
Close

Storage Subsystem status is optimal

Note: There may be a slight delay while transferring the first or last portion of the file.

Transferring 32,768 of 34,368,728 bytes

Controller activation pending

Controller activation pending

Launched Subsystem Management Window for dcs3700-controller-2a

Figure 3-98 Firmware update progress - takes approximately 5 minutes to complete

3. When the firmware update is confirmed, you can add the cluster and hosts.

Defining SONAS to the DCS3700 storage

To define SONAS to the DCS3700 storage, complete the following steps:

1. In the DCS3700 GUI, create the following items by using the menu:

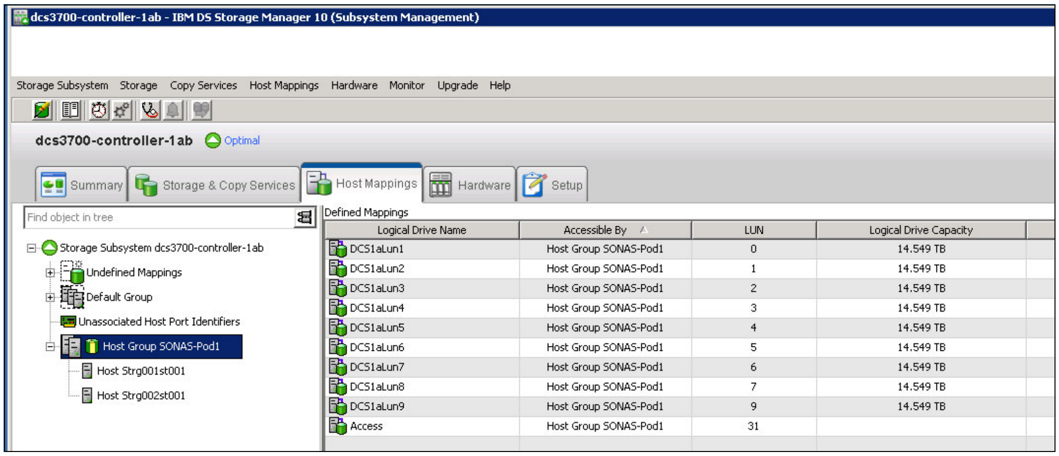
   – One SONAS cluster host group
   – Two SONAS hosts (Storage nodes) per DCS3700
   – Four WWPNs per SONAS host (ports)

   For SONAS installations that are earlier than Version 1.5.1, set the host type to Linux (MPP/RDAC).

   For SONAS installations that are at or later than Version 1.5.1, set the host type to LNXALUA.

   It is easier to add the SONAS Storage nodes (hosts) and their WWPNs in the DCS3700 GUI when the Storage nodes are powered on and the nodes are booted. The `first_time_install` script prompts you to zone and add the Storage node WWPNs to the external storage host groups because the installation process can determine when the Storage nodes are ready to be zoned. Ensure that the **`rpq_gateway_cfg`** flag is set before you run the SONAS `first_time_install` script, as described later in this section and in the SONAS Gateway installation instructions.

   Figure 3-99 shows an example of the host-defined and mapping window.



Logical Drive Name	Accessible By	LUN	Logical Drive Capacity
DCS1aLun1	Host Group SONAS-Pod1	0	14.549 TB
DCS1aLun2	Host Group SONAS-Pod1	1	14.549 TB
DCS1aLun3	Host Group SONAS-Pod1	2	14.549 TB
DCS1aLun4	Host Group SONAS-Pod1	3	14.549 TB
DCS1aLun5	Host Group SONAS-Pod1	4	14.549 TB
DCS1aLun6	Host Group SONAS-Pod1	5	14.549 TB
DCS1aLun7	Host Group SONAS-Pod1	6	14.549 TB
DCS1aLun8	Host Group SONAS-Pod1	7	14.549 TB
DCS1aLun9	Host Group SONAS-Pod1	9	14.549 TB
Access	Host Group SONAS-Pod1	31	

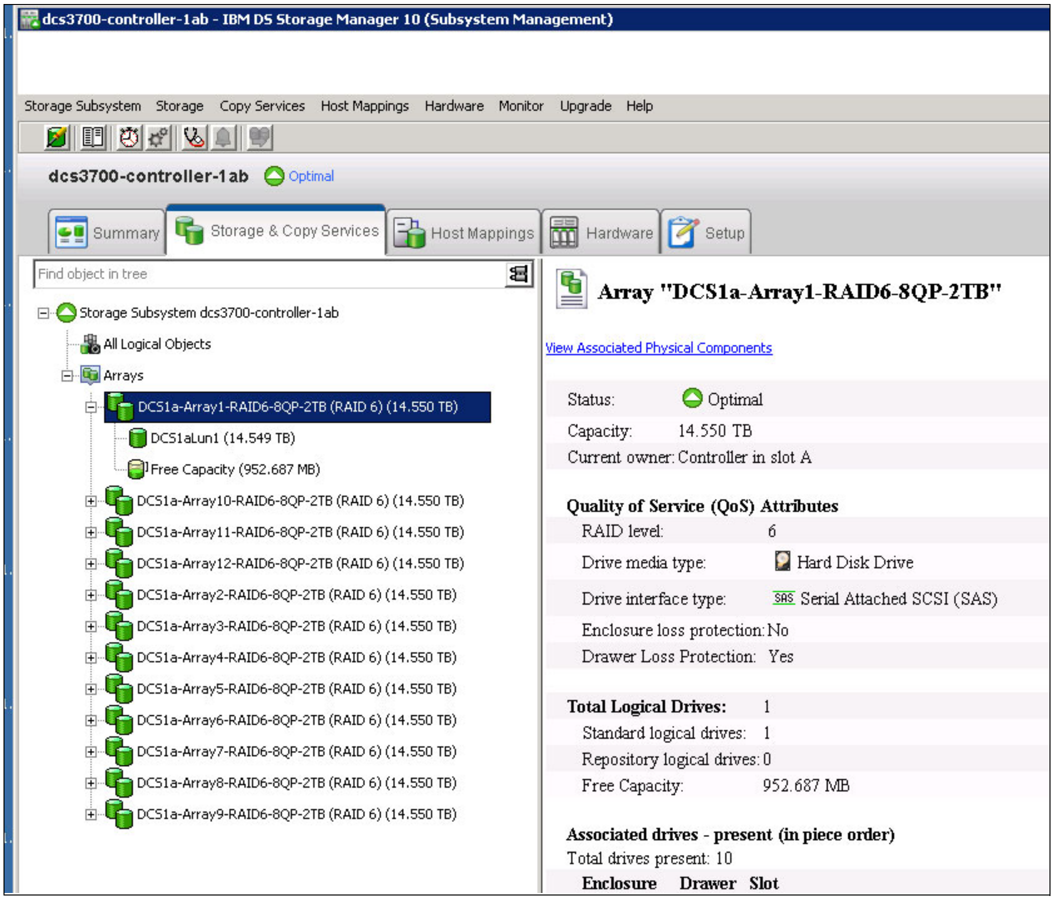
Figure 3-99 DCS3700 cluster definition with the SONAS Storage node pair

2. When the hosts are added, you can configure arrays and disk pools.

   A DCS3700 array configuration is a selected group of drives that are combined in a RAID definition to provision a volume to the SONAS Storage node cluster.

   For SONAS, you can build arrays from 10 drive groups that are set up in RAID 6 8 + P + Q with no hot spares. (RAID DDP or disk pools are a solution that can dramatically improve rebuild times on failed DCS3700 drives).

   Figure 3-100 on page 197 shows a DCS3700 RAID 6 array definition.



dcs3700-controller-1ab Optimal

- Storage Subsystem dcs3700-controller-1ab
  - All Logical Objects
  - Arrays
    - DCS1a-Array1-RAID6-8QP-2TB (RAID 6) (14.550 TB)
      - DCS1aLun1 (14.549 TB)
      - Free Capacity (952.687 MB)
    - DCS1a-Array10-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array11-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array12-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array2-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array3-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array4-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array5-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array6-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array7-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array8-RAID6-8QP-2TB (RAID 6) (14.550 TB)
    - DCS1a-Array9-RAID6-8QP-2TB (RAID 6) (14.550 TB)

Array "DCS1a-Array1-RAID6-8QP-2TB"

Status: Optimal
Capacity: 14.550 TB
Current owner: Controller in slot A

Quality of Service (QoS) Attributes
RAID level: 6
Drive media type: Hard Disk Drive
Drive interface type: SAS Serial Attached SCSI (SAS)
Enclosure loss protection: No
Drawer Loss Protection: Yes

Total Logical Drives: 1
Standard logical drives: 1
Repository logical drives: 0
Free Capacity: 952.687 MB

Associated drives - present (in piece order)
Total drives present: 10
Enclosure Drawer Slot

Figure 3-100 DCS3700 RAID 6 array definition

3. After the arrays, disk pools, or both are configured, you can create the volumes or logical drives.

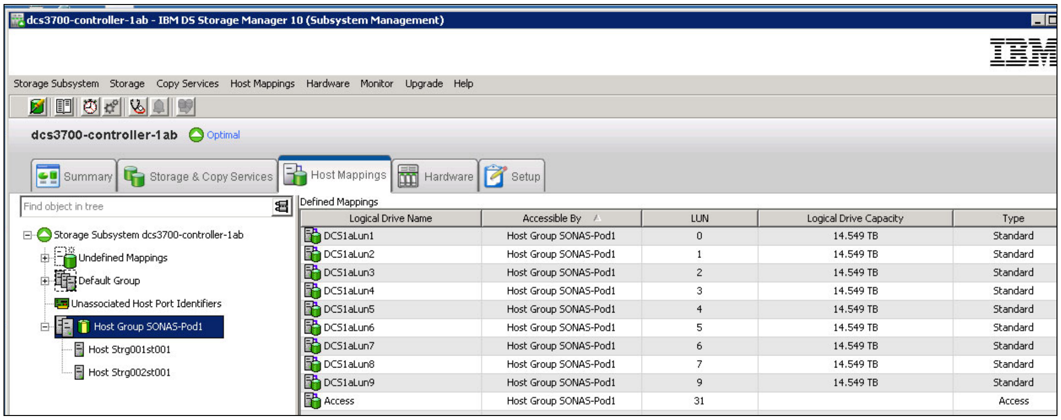
   The DCS3700 arrays are created with a 128 KB stripe width to align with a 1 MB file system block size definition. Provision one logical drive per array (six volumes per DCS3700 enclosure) unless you are using SSD drive technology for metadata usage; in this case, use 16 logical drives.

   Each logical drive is managed by a preferred DCS3700 controller, and I/O is served by a preferred GPFS Storage node (NSD server). For example, if there are 12 logical drives, the SONAS spreads six disks onto strg001 as the primary NSD server and the other six disks have strg002 as the primary server. (This configuration might be strg003 and strg004. It depends on the Storage node names for the referenced disks). After it is determined which six disks have which Storage node as their primary server, it is a preferred practice that three of the six disks have the primary owner of DCS3700 controller A and the other three have the primary owner DCS3700 controller B. This configuration ensures that the disks are properly balanced across the Storage nodes and the external storage controllers for optimal distribution and performance.

4. After the volumes are configured, you can map the volumes to the predefined host group, which holds the hosts definitions that are created for each SONAS Storage node.

   The volumes are mapped to the cluster (not the individual nodes), and they are mapped one at a time in the DCS3700 GUI.

Each DCS3700 maps an access volume to LUN ID 31 for appropriate communication to the cluster, as shown in Figure 3-101.



Defined Mappings

Logical Drive Name	Accessible By	LUN	Logical Drive Capacity	Type
DCS1aLun1	Host Group SONAS-Pod1	0	14.549 TB	Standard
DCS1aLun2	Host Group SONAS-Pod1	1	14.549 TB	Standard
DCS1aLun3	Host Group SONAS-Pod1	2	14.549 TB	Standard
DCS1aLun4	Host Group SONAS-Pod1	3	14.549 TB	Standard
DCS1aLun5	Host Group SONAS-Pod1	4	14.549 TB	Standard
DCS1aLun6	Host Group SONAS-Pod1	5	14.549 TB	Standard
DCS1aLun7	Host Group SONAS-Pod1	6	14.549 TB	Standard
DCS1aLun8	Host Group SONAS-Pod1	7	14.549 TB	Standard
DCS1aLun9	Host Group SONAS-Pod1	9	14.549 TB	Standard
Access	Host Group SONAS-Pod1	31		Access

Figure 3-101 DCS3700 volume mapping

All of the DCS3700 work is done. You are ready to proceed with SONAS code load.

Loading SONAS and installing the cluster

The SONAS software is obtained and installed by the IBM Service Representative (IBM PFE).

To load the SONAS code and install the cluster, complete the following steps:

1. Insert the latest GA SONAS release DVD into the Management node and power on the node. This process is done while all other nodes (Interface and Storage nodes) remain off. Install the OS and software on the Management node (int001st001 in SONAS V1.3 and later).

   If the installer resets the nodes to manufacturing defaults with the `manufacturing_cleanup` script, all nodes are shut down again, and only the Management node is powered on to begin the new SONAS code installation.

   > **Tip:** Installation takes approximately 35 - 45 minutes to complete, and it restarts several times until complete. Initial installation is done when the DVD is ejected and the Management node allows a system login.

2. When the operating system is installed on the primary Management node, the installer runs the following command to set the Gateway flag on the installation:

   ```
   /opt/IBM/sonas/bin/mfg/cfg_gateway_rpq
   ```

   This flag tells the SONAS installation that integrated storage and storage configuration must be run separately from the code installation.

   After the configuration is run, the following file is created:

   ```
   /opt/IBM/sonas/etc/mfg_gateway_rpq
   ```

   Removing the file removes the flag.

   After the pod is initially configured with Gateway storage to add new storage to it, it follows a different process, which is explained in 3.4.4, “Adding XIV LUNs to an existing SONAS configuration” on page 133, 3.5.4, “CIFS on Storwize V7000” on page 143, and 3.6.4, “Adding DCS3700 LUNs to an existing SONAS configuration” on page 214.

After you complete the installation, the IBM authorized service provider begins the cluster installation process. During this process, the `first_time_install` script runs.

First-time installation process

To complete the first-time installation process, complete the following steps:

1. Log in to the SONAS Management node (mgmt001st001) as the root user.
2. Change the directory to `/opt/IBM/sonas/bin` by running the following command:

   ```
   cd /opt/IBM/sonas/bin
   ```

3. Verify the version and release of SONAS by running the following command:

   ```
   /opt/IBM/sonas/bin/get_version
   ```

   Validate that the version that is installed is the version that you want.

4. Run the `first_time_install` script and follow the prompts. For information about creating the initial cluster, see the *IBM SONAS Installation Guide*, GA32-0715 (with real data) and the *SONAS Introduction and Planning Guide*, GA32-0716.

   Run the `first_time_install` script and follow the prompts by running the following command:

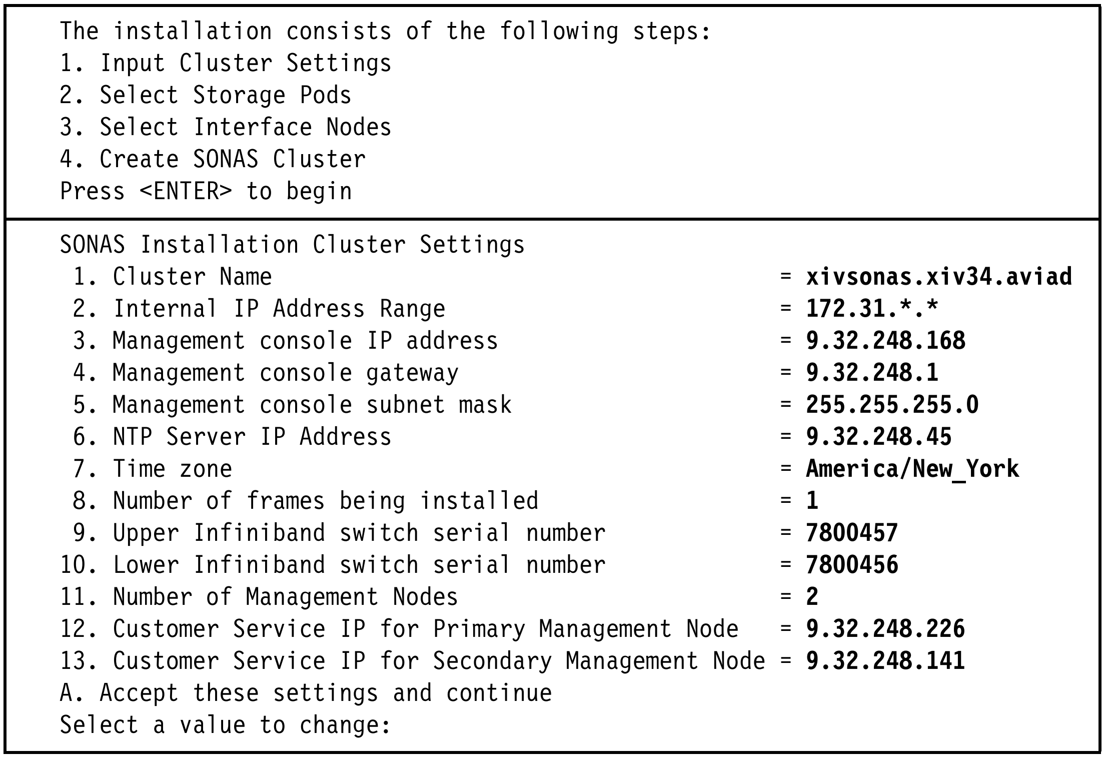
   ```
   /opt/IBM/sonas/bin/first_time_install
   ```

Defining the cluster parameters

To define the cluster parameters, complete the following steps:

1. Use the installation planning worksheet to answer the questions that are related to the first-time configuration of the cluster and Management node information. The initial steps require you to provide the configuration information, as shown in Figure 3-102 on page 200. The client provides this information before you start the installation. For the information that is needed, see the planning tables in Chapter 1, “Installation planning” on page 1.

2. When all the parameters are entered and verified, enter “A” at the prompt to continue. Figure 3-102 shows the output from a `first_time_install` script for software earlier than Version 1.5.1. The SONAS V1.5.1 software includes two more settings, which provide 15 cluster configuration settings.

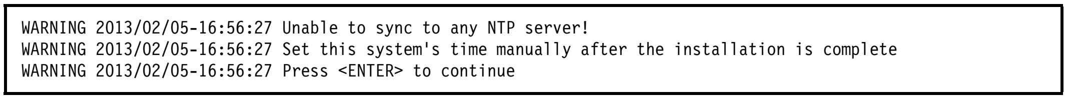


```
The installation consists of the following steps:
1. Input Cluster Settings
2. Select Storage Pods
3. Select Interface Nodes
4. Create SONAS Cluster
Press <ENTER> to begin
```

```
SONAS Installation Cluster Settings
 1. Cluster Name                                    = xivsonas.xiv34.aviad
 2. Internal IP Address Range                       = 172.31.*.*
 3. Management console IP address                   = 9.32.248.168
 4. Management console gateway                      = 9.32.248.1
 5. Management console subnet mask                  = 255.255.255.0
 6. NTP Server IP Address                           = 9.32.248.45
 7. Time zone                                       = America/New_York
 8. Number of frames being installed                = 1
 9. Upper Infiniband switch serial number           = 7800457
10. Lower Infiniband switch serial number           = 7800456
11. Number of Management Nodes                      = 2
12. Customer Service IP for Primary Management Node   = 9.32.248.226
13. Customer Service IP for Secondary Management Node = 9.32.248.141
A. Accept these settings and continue
Select a value to change:
```

Figure 3-102 Example Management node data from the first_time_install script

As the script progresses, in some cases, especially where the management port is not connected to a network because it shares the public network Ethernet ports, you receive a “Warning Syncing NTP” message. SONAS V1.5.1 `first_time_install` updates are designed to reduce its occurrence. You might not receive this message. If you do receive it, it is acceptable. Press Enter to continue, as shown in Figure 3-103.



```
WARNING 2013/02/05-16:56:27 Unable to sync to any NTP server!
WARNING 2013/02/05-16:56:27 Set this system's time manually after the installation is complete
WARNING 2013/02/05-16:56:27 Press <ENTER> to continue
```

Figure 3-103 NTP sync warning

Powering on and discovering each node

To power on and discover each node, complete the following steps:

1. As the SONAS `first_time_install` script progresses, the script prompts you to “`power on all nodes`”. After powering on all nodes, you can press Enter from the `first_time_install` script prompt.

2. Nodes use the Preboot Execution Environment (PXE, also known as Pre-Execution Environment) to boot and load the operating system image. The ISO image is transferred onto the detected nodes. A list of recognized nodes appears and is frequently updated. Nodes appear in the list as each one is recognized and configured (Figure 3-104). It might take 60 minutes or longer, depending on the number of SONAS nodes in the configuration, to recognize and complete the SONAS code load on each node.

```
Detected 0 Interface nodes and 0 Storage nodes
Detected 1 Interface nodes and 0 Storage nodes
Detected 2 Interface nodes and 0 Storage nodes
Detected 3 Interface nodes and 1 Storage nodes
Detected 3 Interface nodes and 2 Storage nodes
```

Figure 3-104 Sample output for three Interface and two Storage nodes

3. When all Interface nodes and Storage nodes are discovered (and quantity numbers keep repeating with no change), press Enter to continue the configuration.

Verifying device IDs, rack and slot locations, and quorums

Proper rack cabling and configuration are key to a healthy SONAS system and proper GUI representation. Accurate cabling in the internal SONAS frame (Ethernet and InfiniBand) switches ensure proper frame location identification and configuration of these nodes. However, locations can sometimes require redefinition in the `first_time_install` process. If it is not done correctly, the GUI might not properly align the nodes in the health center frame model.

You can adjust and keep the configuration when the accurate assignments are confirmed for Interface nodes (I), Management nodes (M) (Figure 3-105), and Storage nodes (S) (Figure 3-106 on page 202).

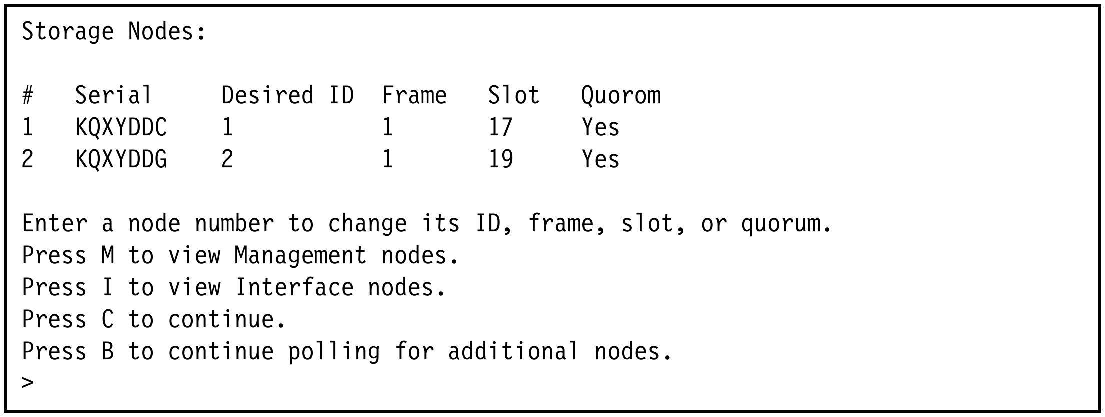
After you verify all rack locations and the quorum status, select “C” to continue to configure the cluster.

Figure 3-105 shows the Management node configuration.

```
Management Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQWHAHK    2           1       3      Yes
2   KQWHAHL    1           1       1      No

Enter a node number to change its ID, frame, slot, or quorum.
Press I to view Interface nodes.
Press S to view Storage nodes.
Press R to reconfigure Management nodes.
Press B to continue polling for additional nodes.
>
```

Figure 3-105 Management node configuration

Figure 3-106 shows the Storage node configuration.



```
Storage Nodes:

#   Serial     Desired ID  Frame   Slot   Quorom
1   KQXYDDC    1           1       17     Yes
2   KQXYDDG    2           1       19     Yes

Enter a node number to change its ID, frame, slot, or quorum.
Press M to view Management nodes.
Press I to view Interface nodes.
Press C to continue.
Press B to continue polling for additional nodes.
>
```

Figure 3-106 Storage node configuration

Configuring the node instances, location, serial numbers, and quorum states is key to a correctly working SONAS system. Typically, the first instance of a node type is in the lowest point of the rack of its node type and they go in sequence while moving up the rack. For example, the bottom Interface node is int001st001, then the next one up is int002st001. The bottom Ethernet switch is switch 1, and the next one up is switch 2. The bottom InfiniBand switch is switch 1, and the next one up is switch 2. This process changes for the SONAS Storage nodes in Gateway configurations. The first and second Storage nodes are above Storage nodes 3 and 4. The rest of the Storage nodes are above Storage nodes 1 and 2.

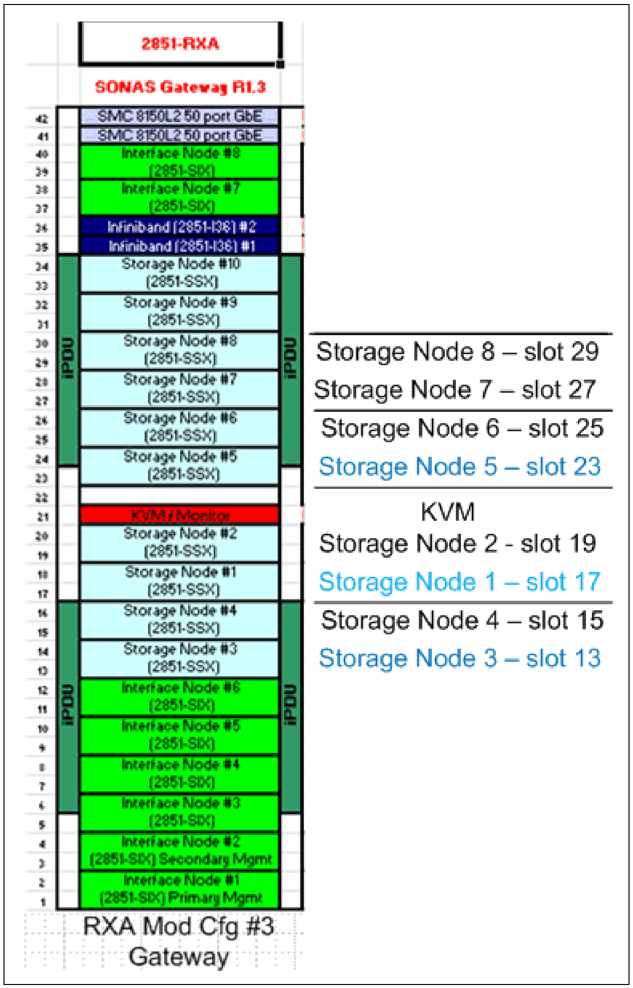
You can adjust and keep the configuration when the accurate assignments are confirmed for Interface and Storage nodes (S), select “C” to continue to configure the cluster.

Select the line item number to change the configuration (Device ID/instance, Frame, slot, and Quorum state):

- Desired ID: This ID is the instance of the node type. Therefore, the first Management node instance has a desired ID of 1 and the second ID has a desired ID of 2. In Figure 3-107 on page 203, ID 1 is in the lowest slot of the rack. The bottom Interface node in the frame is int001st001 (slot 1 = node 1). The second Interface node from the bottom of the frame is int002st001 (slot 3= node 2). Also, Storage nodes have a little different configuration, as shown in Figure 3-107 on page 203. Storage node 1 and 2 are above Storage nodes 3 and 4.
- Frame Number: The frame number relates to the frame instance of SONAS. If this frame is frame one of one SONAS frame, use 1. If you have a base rack and an expansion frame, then the base rack is 1, and the first expansion frame is 2.
- Slot number (rack slot number): Use the lower U-number indicator of the slot for the device (in the frame).
- Quorum: There are three sizes of *quorum node configuration* for SONAS installation (small = 3, medium = 5, and large = 7). If a cluster is designed small (2 - 6 Interface nodes), you need three quorum nodes. If the cluster is medium (6 - 10 Interface nodes), set five quorum nodes. If the cluster is large (over 10 Interface nodes), set seven quorum nodes.

Figure 3-107 shows a SONAS Gateway RXA rack layout configuration and the node naming conventions.

> **Note:** Figure 3-59 on page 158 shows a SONAS Gateway RXA rack layout configuration and the node naming conventions.



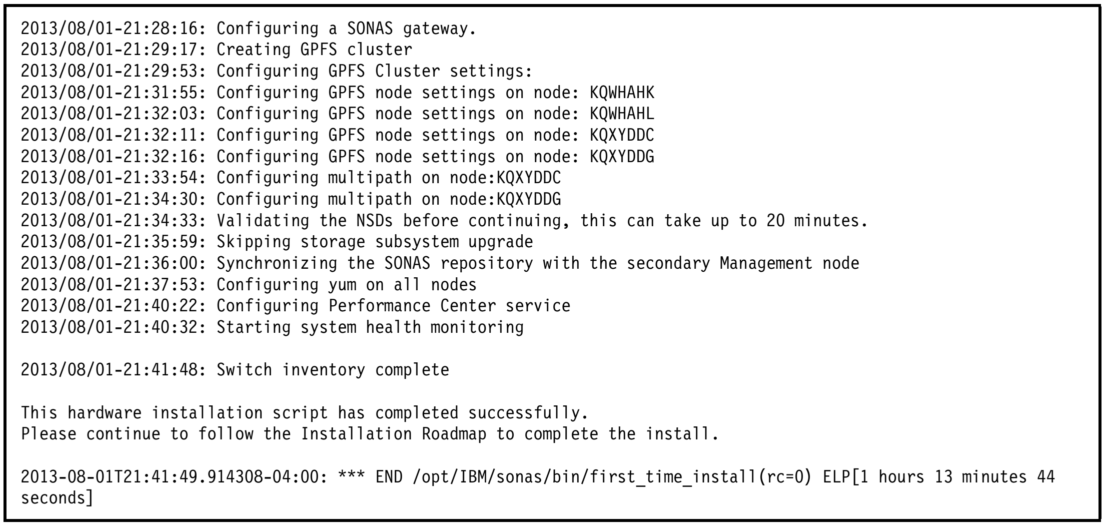
2851-RXA
SONAS Gateway R1.3
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Interface Node #8 (2851-SIX)
Interface Node #7 (2851-SIX)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Storage Node #10 (2851-SSX)
Storage Node #9 (2851-SSX)
Storage Node #8 (2851-SSX)
Storage Node #7 (2851-SSX)
Storage Node #6 (2851-SSX)
Storage Node #5 (2851-SSX)
KVM / Monitor
Storage Node #2 (2851-SSX)
Storage Node #1 (2851-SSX)
Storage Node #4 (2851-SSX)
Storage Node #3 (2851-SSX)
Interface Node #6 (2851-SIX)
Interface Node #5 (2851-SIX)
Interface Node #4 (2851-SIX)
Interface Node #3 (2851-SIX)
Interface Node #2 (2851-SIX) Secondary Mgmt
Interface Node #1 (2851-SIX) Primary Mgmt
iPDU
iPDU
iPDU
iPDU
Storage Node 8 – slot 29
Storage Node 7 – slot 27
Storage Node 6 – slot 25
Storage Node 5 – slot 23
KVM
Storage Node 2 - slot 19
Storage Node 1 – slot 17
Storage Node 4 – slot 15
Storage Node 3 – slot 13
RXA Mod Cfg #3
Gateway

Figure 3-107 SONAS RXA - slot location offset for Storage nodes

After all the Interface and Storage nodes are listed properly and verified, type “B” to poll again or type “C” to continue with the cluster configuration. You can enter “C” to continue *only* from the Storage node configuration screen (See Figure 3-106 on page 202).

The first_time_install script configures the cluster and completes

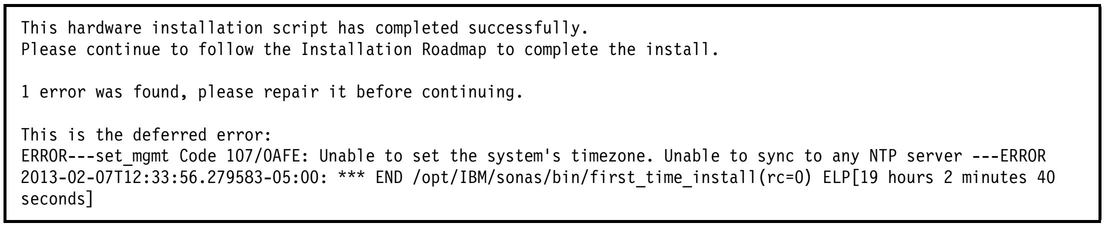
The `first_time_install` script now continues by creating the SONAS cluster and adding each node to the cluster. This process takes another 30 - 40 minutes or longer, depending on the size of the configuration. Figure 3-108 shows the end of the script, after which the cluster is successfully configured. If the script does not end with a success statement, open a PMR and escalate it to support for immediate assistance.



```
2013/08/01-21:28:16: Configuring a SONAS gateway.
2013/08/01-21:29:17: Creating GPFS cluster
2013/08/01-21:29:53: Configuring GPFS Cluster settings:
2013/08/01-21:31:55: Configuring GPFS node settings on node: KQWHAHK
2013/08/01-21:32:03: Configuring GPFS node settings on node: KQWHAHL
2013/08/01-21:32:11: Configuring GPFS node settings on node: KQXYDDC
2013/08/01-21:32:16: Configuring GPFS node settings on node: KQXYDDG
2013/08/01-21:33:54: Configuring multipath on node:KQXYDDC
2013/08/01-21:34:30: Configuring multipath on node:KQXYDDG
2013/08/01-21:34:33: Validating the NSDs before continuing, this can take up to 20 minutes.
2013/08/01-21:35:59: Skipping storage subsystem upgrade
2013/08/01-21:36:00: Synchronizing the SONAS repository with the secondary Management node
2013/08/01-21:37:53: Configuring yum on all nodes
2013/08/01-21:40:22: Configuring Performance Center service
2013/08/01-21:40:32: Starting system health monitoring

2013/08/01-21:41:48: Switch inventory complete

This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.

2013-08-01T21:41:49.914308-04:00: *** END /opt/IBM/sonas/bin/first_time_install(rc=0) ELP[1 hours 13 minutes 44
seconds]
```

Figure 3-108   Cluster being created and the first_time_install script completes

In some cases, especially where the management network shares Ethernet ports with the public network, the `first_time_install` script ends with an NTP sync error. SONAS V1.5.1 `first_time_install` updates are designed to reduce the occurrence of this message and it might not be noticed. However, it is acceptable to receive this message. NTP properly syncs when the management network is properly configured (see Figure 3-109.)



```
This hardware installation script has completed successfully.
Please continue to follow the Installation Roadmap to complete the install.

1 error was found, please repair it before continuing.

This is the deferred error:
ERROR---set_mgmt Code 107/0AFE: Unable to set the system's timezone. Unable to sync to any NTP server ---ERROR
2013-02-07T12:33:56.279583-05:00: *** END /opt/IBM/sonas/bin/first_time_install(rc=0) ELP[19 hours 2 minutes 40
seconds]
```

Figure 3-109   The first_time_install script ends with an NTP error

Post-first-time-installation procedures

The GPFS cluster is installed, configured, and is running. From here, the installer continues with postinstallation procedures.

Enabling licensing

Enable the license by running the following command:

```
/opt/IBM/sonas/bin/enablelicense --accept
```

The license can also be accepted the first time that you access GUI.

Adding the GPFS cluster to SONAS Management

Before any other management functions can be done, the GPFS cluster must be added to the SONAS Management subsystem by running the following command:

```
cli addcluster -h mgmt001st001 -p Passw0rd
```

Configuring the SONAS management port

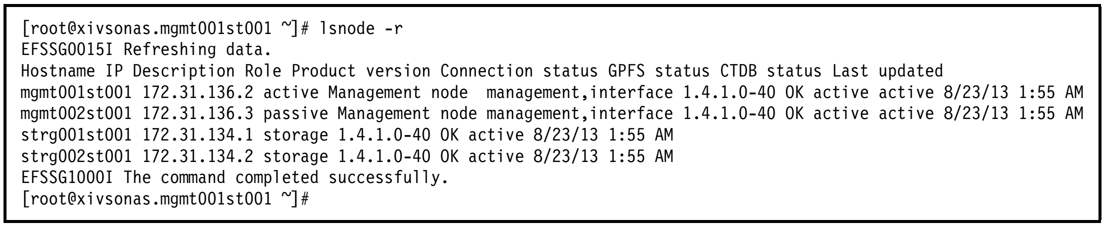
In many configurations, the management port is configured to use the same Ethernet ports as the public network (10 GbE). In these configurations, the management port is not connected to the network. Therefore, the management subsystem must be configured to use the 10 GbE network ports. The SONAS V1.5.1 `first_time_install` script enhancements allow you to identify the external management adapter and do not require you to run **`chnwmgt`** later. After a SONAS V1.5.1. installation, the **`chnwmgt`** command must be run *only* if there is no external management adapter value as input at the time of installation or if an incorrect value is entered and must be changed later. If you need to configure the management ports on the 10 GbE network, run the following command:

```
chnwmgt --interface ethX1
```

The SONAS GUI is available for use.

Verifying the node list

Verify the node list by running **`lsnode -r`**, as shown in Figure 3-110.



```
[root@xivsonas.mgmt001st001 ~]# lsnode -r
EFSSG0015I Refreshing data.
Hostname IP Description Role Product version Connection status GPFS status CTDB status Last updated
mgmt001st001 172.31.136.2 active Management node  management,interface 1.4.1.0-40 OK active active 8/23/13 1:55 AM
mgmt002st001 172.31.136.3 passive Management node management,interface 1.4.1.0-40 OK active active 8/23/13 1:55 AM
strg001st001 172.31.134.1 storage 1.4.1.0-40 OK active 8/23/13 1:55 AM
strg002st001 172.31.134.2 storage 1.4.1.0-40 OK active 8/23/13 1:55 AM
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

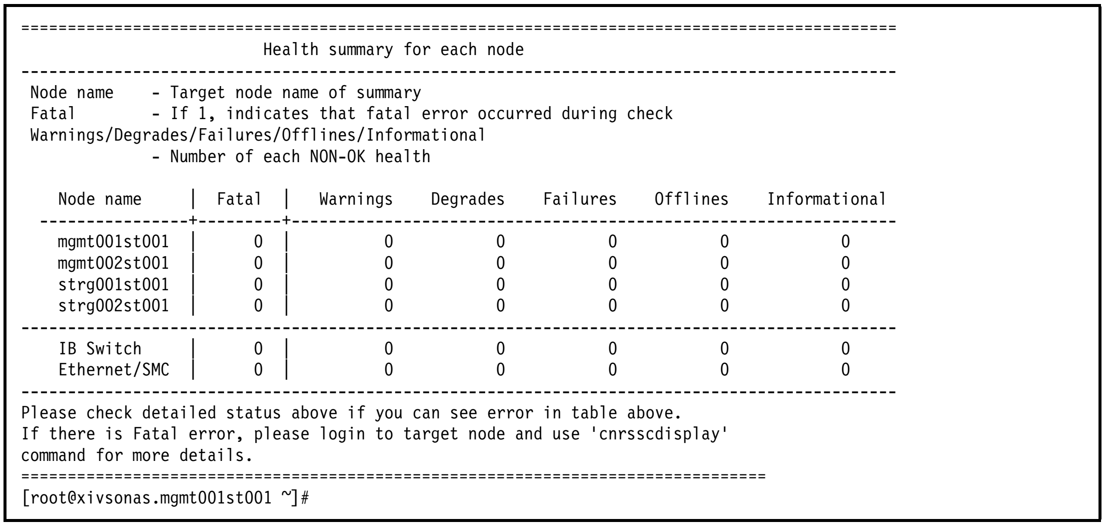
Figure 3-110 Sample lsnode output that shows configured nodes

Checking the system health

It is time to check the system health. The IBM authorized service provider logs in to the Management node and runs the health check commands. The following command checks the SONAS system overall health and all components (Ethernet switches, InfiniBand switches, and nodes):

```
cnrsscheck --nodes=all --checks=all
```

The command also checks to see whether the nodes have correctly assigned roles and whether they can communicate with each other. Figure 3-111 shows the command output for the sample cluster configuration.

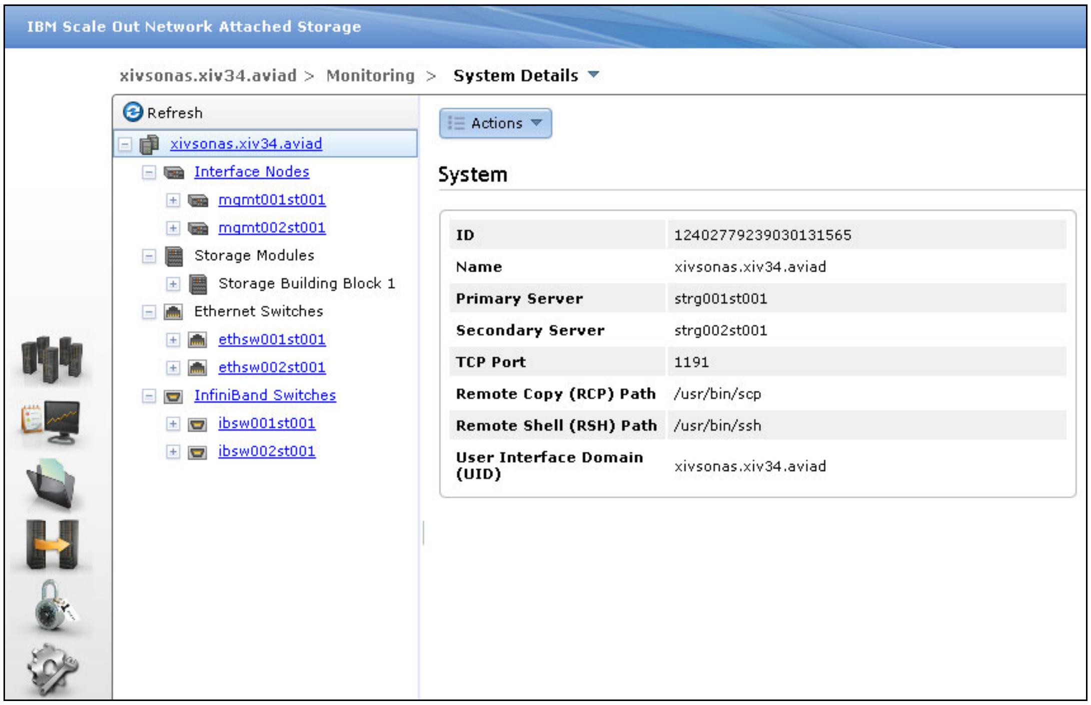


```
==============================================================================================
                          Health summary for each node
----------------------------------------------------------------------------------------------
 Node name     - Target node name of summary
 Fatal         - If 1, indicates that fatal error occurred during check
 Warnings/Degrades/Failures/Offlines/Informational
               - Number of each NON-OK health

    Node name      |  Fatal  |  Warnings    Degrades    Failures    Offlines    Informational
  ----------------+---------+-------------------------------------------------------------------
    mgmt001st001  |     0   |       0           0           0           0            0
    mgmt002st001  |     0   |       0           0           0           0            0
    strg001st001  |     0   |       0           0           0           0            0
    strg002st001  |     0   |       0           0           0           0            0
----------------------------------------------------------------------------------------------
    IB Switch     |     0   |       0           0           0           0            0
    Ethernet/SMC  |     0   |       0           0           0           0            0
----------------------------------------------------------------------------------------------
Please check detailed status above if you can see error in table above.
If there is Fatal error, please login to target node and use 'cnrsscdisplay'
command for more details.
================================================================================
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-111   Sample output from the cnrsscheck summary

Clearing miscellaneous installation errors

During the installation process, some miscellaneous errors or information warnings might be generated as nodes are loaded, rebooted, and ports go up and down. Use the GUI and the system monitoring window to review and clear these errors before you run the health check. Ensure that all components are green (see Figure 3-112).



System

ID	12402779239030131565
Name	xivsonas.xiv34.aviad
Primary Server	strg001st001
Secondary Server	strg002st001
TCP Port	1191
Remote Copy (RCP) Path	/usr/bin/scp
Remote Shell (RSH) Path	/usr/bin/ssh
User Interface Domain (UID)	xivsonas.xiv34.aviad

Figure 3-112   SONAS Monitoring System Details window

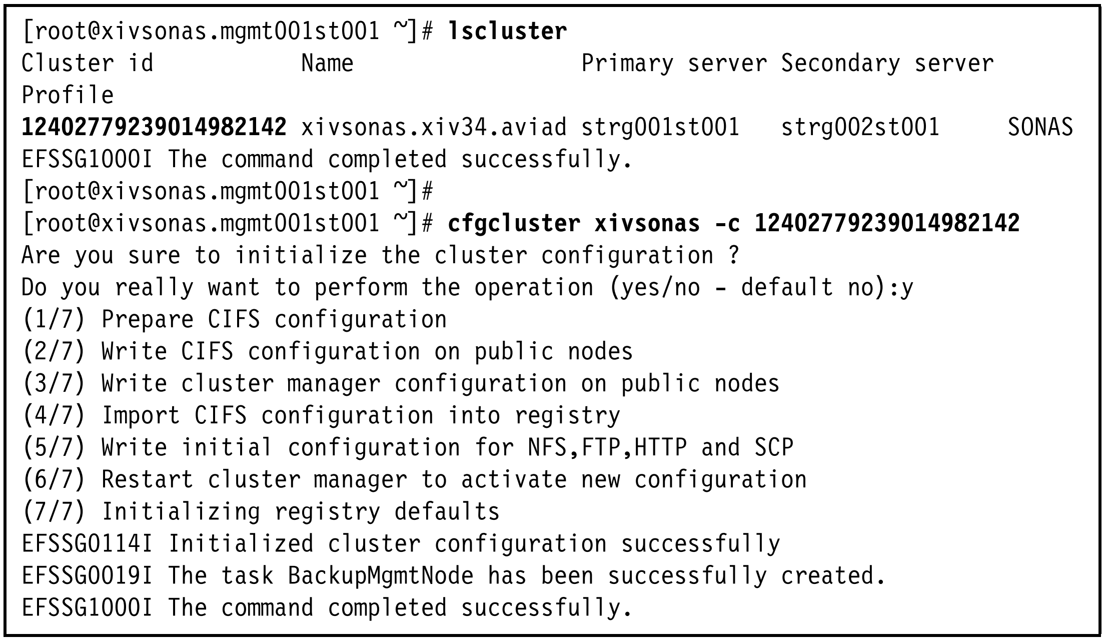
Completing the SONAS cluster configuration

Before you create a file system, file sets, or exports, you must complete the SONAS cluster configuration.

To complete the cluster configuration, complete the following steps:

1. Run `lscluster` to confirm proper cluster configuration and list the unique cluster ID, which is required for the `cfgcluster` command.
2. Run `cfgcluster` to create the initial configuration for all supported protocols (HTTPS, NFS, CIFS, FTP, and SCP). Here is a list of the command's functions:
   - Prepare the CIFS configuration.
   - Distribute the CIFS configuration.
   - Distribute the CTDB configuration.
   - Import the CIFS configuration into the registry.
   - Write the initial configuration for NFS, FTP, HTTP, and SCP.
   - Restart CTDB to activate a new configuration.

   Figure 3-113 shows an example of this command.



```
[root@xivsonas.mgmt001st001 ~]# lscluster
Cluster id          Name                 Primary server Secondary server
Profile
12402779239014982142 xivsonas.xiv34.aviad strg001st001   strg002st001     SONAS
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
[root@xivsonas.mgmt001st001 ~]# cfgcluster xivsonas -c 12402779239014982142
Are you sure to initialize the cluster configuration ?
Do you really want to perform the operation (yes/no - default no):y
(1/7) Prepare CIFS configuration
(2/7) Write CIFS configuration on public nodes
(3/7) Write cluster manager configuration on public nodes
(4/7) Import CIFS configuration into registry
(5/7) Write initial configuration for NFS,FTP,HTTP and SCP
(6/7) Restart cluster manager to activate new configuration
(7/7) Initializing registry defaults
EFSSG0114I Initialized cluster configuration successfully
EFSSG0019I The task BackupMgmtNode has been successfully created.
EFSSG1000I The command completed successfully.
```

Figure 3-113 Example cfgcluster command

Upon completion, the cluster is now fully configured and operational. However, more configuration is required for cluster networking and authentication. Because these functions are not dependent on the type of underlying storage, they are described later in this chapter, independent of the Gateway solutions. Read through the chapter to understand network installation process and preferred practices for your cluster installation.

Sections 3.8, "SONAS integration into your network" on page 218 and later provide detailed information about general SONAS and GPFS storage configuration information that applies to any back-end storage, along with networking and other considerations that are not back-end storage specific. Read the remainder of this chapter in its *entirety* to understand all installation considerations before you plan to install SONAS into your environment.

Rescanning Storage nodes

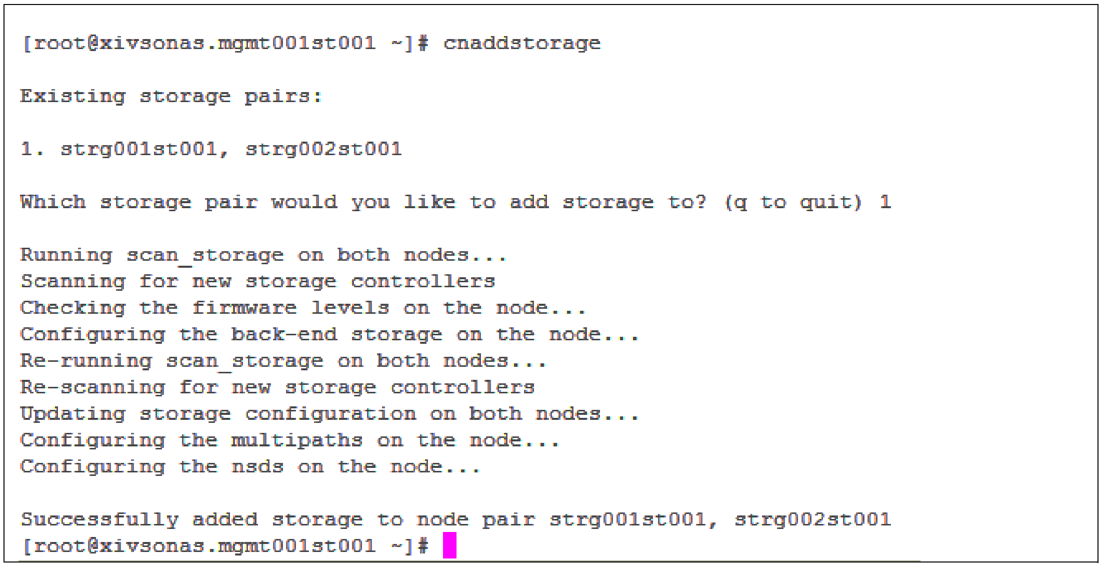
If the storage was not configured and allocated to the cluster before the `first_time_install` script ran, run **`mkdisk --luns`** (SONAS V1.4.1 and later) as the administrator or root user. This command rescans all available Storage nodes for newly allocated LUNs (see Figure 3-114).

```
[root@xivsonas.mgmt001st001 ~]# mkdisk --luns
(1/3) Scanning for new devices
(2/3) Creating NSDs
(3/3) Adding to database
Successfully created disks:
DCS3700_360080e50002ee7bc00000c5551b73a8c
DCS3700_360080e50002ea78c00000cfd51b73794
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-114 Rescan for disks

An alternative method is to run **`cnaddstorage`** as the root user. This command presents a short list of Storage node pairs from which to select.

If you choose this method, select the Storage node pair to which you added storage. SONAS automatically discovers new volumes and creates the multipath devices and the associated NSDs. The command must be run for each Storage node pair or pod. This process might take up to six minutes to complete. For details, see Figure 3-115.



```
[root@xivsonas.mgmt001st001 ~]# cnaddstorage

Existing storage pairs:

1. strg001st001, strg002st001

Which storage pair would you like to add storage to? (q to quit) 1

Running scan_storage on both nodes...
Scanning for new storage controllers
Checking the firmware levels on the node...
Configuring the back-end storage on the node...
Re-running scan_storage on both nodes...
Re-scanning for new storage controllers
Updating storage configuration on both nodes...
Configuring the multipaths on the node...
Configuring the nsds on the node...

Successfully added storage to node pair strg001st001, strg002st001
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-115 CLI cnaddstorage command

After the LUNs and disks are rescanned, the installer continues the installation process.

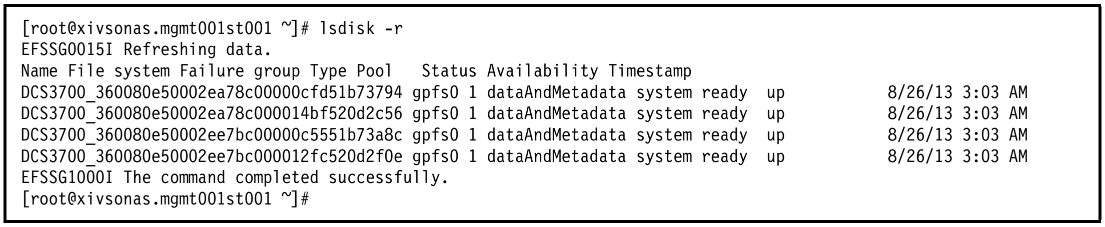
> **Tip:** If you perform a reinstallation for any reason, previous NSD labels must be revoked before reinstalling LUN devices, either by reformatting or by running the **`dd if=/dev/zero of=/dev/mapper/`*`device-name`*** command against the device for about 5 seconds to clear the label.

Verifying the disk configuration

From the Management node prompt, run the following command:

```
lsdisk -r
```

The output is shown in Figure 3-116.



```
[root@xivsonas.mgmt001st001 ~]# lsdisk -r
EFSSG0015I Refreshing data.
Name File system Failure group Type Pool   Status Availability Timestamp
DCS3700_360080e50002ea78c00000cfd51b73794 gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ea78c000014bf520d2c56 gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ee7bc00000c5551b73a8c gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ee7bc000012fc520d2f0e gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-116   lsdisk example output

All the DCS3700 disks from the pods are configured and are visible from the Management node. By default, they are added to the GPFS “system” pool in Failure group 1, and the Storage node preference is assigned in alternating LUN (called NSDs in GPFS) fashion.

Interpreting the lsdisk command output

The **lsdisk** command displays information about disk devices that you must consider before you configure the file systems.

Here is a description of the output that is shown in Figure 3-116 from left to right across the top:

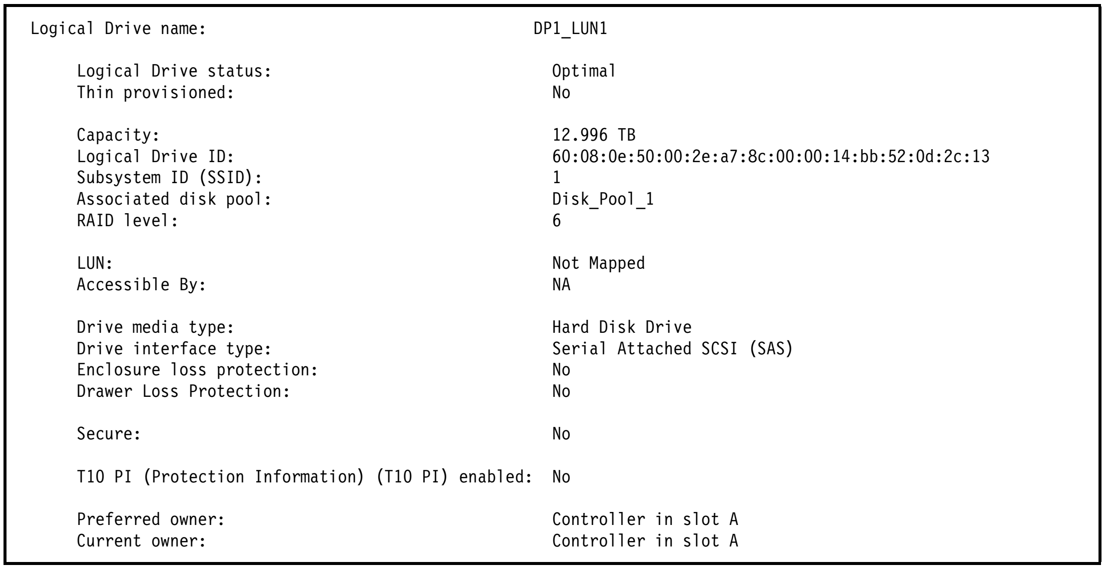
- The device list begins with *Name*. This name refers to the NSD name. The NSD name is the GPFS “Network Storage Device” label. DCS3700 NSD names are complicated in comparison to their XIV counterparts.

  For DCS3700 based storage, the first three characters refer to the storage type (1818); this indicator is the same for the DCS3700.

  The next set of characters refer to the DCS3700 volume ID or serial number, which is unique worldwide. Unlike the Storwize V7000, you cannot map a piece of the unique serial number to a particular DCS3700 frame.

  The simplified volume name is not used on the SONAS side, so it is important to create a spreadsheet or some form of tracking for each volume identifier from each DCS3700 frame. You can use this tracking method to match the volumes in SONAS. Otherwise, they all look similar, and it can be difficult to distinguish the 1 TB NSD from the 2 TB NSD, and SAS from NLSAS or SSD.

Figure 3-117 shows the listing from the DCS3700 subsystem. The following data points are critical: Logical Volume Name (Logical Drive name), LUN size (Capacity), unique Identifier (Logical Drive ID), Disk Pool (Associated disk pool), and preferred controller (Preferred Owner). This data is captured for every DCS3700 disk and controller in the SONAS configuration. This data is used later when you create file systems.



```
Logical Drive name:                              DP1_LUN1

     Logical Drive status:                         Optimal
     Thin provisioned:                             No

     Capacity:                                     12.996 TB
     Logical Drive ID:                             60:08:0e:50:00:2e:a7:8c:00:00:14:bb:52:0d:2c:13
     Subsystem ID (SSID):                          1
     Associated disk pool:                         Disk_Pool_1
     RAID level:                                   6

     LUN:                                          Not Mapped
     Accessible By:                                NA

     Drive media type:                             Hard Disk Drive
     Drive interface type:                         Serial Attached SCSI (SAS)
     Enclosure loss protection:                    No
     Drawer Loss Protection:                       No

     Secure:                                       No

     T10 PI (Protection Information) (T10 PI) enabled:  No

     Preferred owner:                              Controller in slot A
     Current owner:                                Controller in slot A
```

Figure 3-117   DCS3700 CLI output

- The next piece of information from the lsdisk list is *File system*.

  If the listed NSD is not yet assigned to a file system, this column remains blank for that device. When the device is assigned to a file system, the file system name (such as "gpfs0") is listed in this column.

- The next column in the list is *Failure group*. The NSD failure group is a value that GPFS allows you to assign to any NSD to logically manage groups of disk devices in that file system. For example, if you have two DCS3700 frames that are provisioning storage into a single storage pod, you can assign the first DCS3700 to Failure group 1, and the second DCS3700 to Failure group 2. You can use this assignment to define different replication between these groups.
- That next item is the *Type* definition. The type refers to the usage type of data for which you can use that NSD. For example, you can use the disk for dataOnly, for metadataOnly, or for metadataAndData.
- The Pool definition refers to the category of disk for that disk type. If you have SAS and Near Line SAS in a file system and you choose to have them both in the same file system, you can do so by having them defined in different pools. The file system places data on the system pool by default. You can then migrate data to a different pool by using a structured ILM policy.

  A preferred practice is to place your fastest tier of disk storage in the system pool, and migrate to lower tiers from there. This practice ensures metadata placement on high-speed disk technology.

- *Status* displays the readiness of the NSD for use. Disk status has five possible values, three of which are transitional:
  - Ready: Normal status.
  - Suspended: Indicates that data will be migrated off this disk.
  - Being emptied: Transitional status in effect while a disk deletion is pending.

  - Replacing: Transitional status in effect for an old disk while replacement is pending.
  - Replacement: Transitional status in effect for a new disk while replacement is pending.

  GPFS allocates space only on disks with a status of ready or replacement.

- *Availability* refers to the following possible values:
  - Up: The disk is available to GPFS for normal read and write operations.
  - Down: No read and write operations can be done on the disk.
  - Recovering: An intermediate state for disks that are coming up, during which GPFS verifies and corrects data. The read operations can be done while a disk is in this state, but write operations cannot.
  - Unrecovered: Not all disks were successfully brought up. There are cases where multiple disks must be recovered simultaneously to bring the storage to a consistent state. The most obvious cases involve replication where both copies of some pointers to data on this disk are unavailable. Unrecovered means that the disk is physically available, but the prerequisites for running recovery are not satisfied.
  - Timestamp: Refers to the creation date and time.

Note: Disk Pool, Usage Type, and Failure Groups are assigned to the disks and NSDs and must be defined before the disks are associated with a file system. These parameters can be changed only by removing the disk from the file system, changing the parameters, and reading them to the file system. This process assumes that there is enough space on the remaining disks in the file system to maintain the used capacity. It might require a restriping of data to the disk or disks that are added back to the file system.

Creating the file system

Before you create your file system, make sure that the volumes are evenly balanced across the Storage nodes. It is important that the LUNs and NSDs are balanced across all Storage nodes. When SONAS imports NSDs from the underlying storage subsystem, it assigns a Storage node preference to each NSD in alternating fashion.

With XIV solutions, it is simple because there is one RAID level, size, disk, type, and so on. These attributes can be viewed by running **`mmlsnsd`**.

The Storage node preference is important to performance because it defines the Storage node that tries to manage all I/O to that NSD at any particular time. The I/O is managed only by the second node in the Storage node preference if the first node fails to serve the I/O. This failover behavior is automatic. Figure 3-118 on page 212 shows that for NSD "`DCS3700_360080e50002ea78c00000cfd51b73794`", the Storage node strg001st001 is the Storage node preference, and strg002st001 is the backup node for I/O that is destined for that NSD. This behavior is easily confirmed in the XIV statistics monitor.

As you can see in Figure 3-118, the Storage node preference for strg001st001 alternates sequentially by volume name, which makes file system assignment simple. If you provision a file system with consecutively named NSDs, the file system automatically balances across each Storage node if there are an even number of NSDs that are assigned to the file system.

```
[root@xivsonas.mgmt001st001 ~]# mmlsnsd

 File system   Disk name    NSD servers
---------------------------------------------------------------------------
 gpfs0 DCS3700_360080e50002ea78c00000cfd51b73794 strg001st001,strg002st001
 gpfs0 DCS3700_360080e50002ee7bc00000c5551b73a8c strg002st001,strg001st001
 gpfs0 DCS3700_360080e50002ea78c000014bf520d2c56 strg001st001,strg002st001
 gpfs0 DCS3700_360080e50002ee7bc000012fc520d2f0e strg002st001,strg001st001

[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-118 Example of mmlsnsd output that shows Storage node preference

DCS3700 preferred path controllers

Unlike the XIV or Storwize V7000 subsystems, which are active/active multi-pathing storage platforms, the DCS3700 subsystem is an active/passive (preferred path) storage platform. This configuration means that each device and LUN has a preferred path and does not provide load balancing between the canisters and controllers. Upon a path or controller failure, the non-preferred controller or canister takes over the ownership of that LUN. Because of the preferred path methodology, it is important to spread the I/O across not only all Storage nodes, but also all controllers and canisters in a frame. It is important to alternate I/O across both Storage nodes and to alternate controllers and canisters. Map each NSD to a DCS3700 LUN with the NSD naming conventions (LUN serial number) and the LUN's preferred path. As described in "Interpreting the lsdisk command output" on page 164, it is a preferred practice to create a spread sheet or another tracking mechanism for the controller and canister preferred path with the preferred Storage node.

Note: The preferred path is set upon LUN creation. After the Storage nodes are booted and the NSDs are initially configured, do not try to manually switch the preferred controller.

A best-case example is that all LUNs, whose preferred path is canister A, have strg001st001 as the preferred I/O owner. The LUNs whose preferred path is canister B have strg002st001 as the preferred I/O owner. However, this configuration is not guaranteed.

It is even more important when multiple frames are deployed. If you want the highest achievable performance, spread NSD resources evenly across all Storage nodes. Figure 3-119 shows an example of a **mkfs** command for evenly distributing the NSDs and Storage node preference in the creation of the gpfs0 file system.

```
mkfs gpfs0 -b 1M -F
DCS3700_360080e50002ea78c00000cfd51b73794,DCS3700_360080e50002ee7bc00000c5551b7
3a8c,DCS3700_360080e50002ea78c000014bf520d2c56,DCS3700_360080e50002ee7bc000012f
c520d2f0e -R meta -j cluster”
```

Figure 3-119 Sample mkfs gpfs0 /ibm/gpfs0 -F command output

For now, consider the **-j** option of the GPFS **mkfs** command (for cluster allocation type).

Data allocation types

GPFS offers two basic data allocation types: Scatter and Cluster.

The Scatter allocation type disperses I/O randomly across all NSDs, and the Cluster allocation type lays it down contiguously across all striped NSDs.

Testing has shown that well-defined NSD striping performs when the Cluster allocation type is used for file system creation for most large file sequential workloads. Current data also suggests that the preferred file system block allocation, from a performance stand point, is 1 M (non-default) for most large file sequential type workloads. In contrast, typically for small file random I/O patterns, block sizes of 256 KB and Scatter allocation types tend to have better performance. For this reason, most clients might benefit from testing with simulated production workloads before selecting a type.

> **Tip:** Some use cases (not yet analyzed) might serve I/O faster in larger block sizes.

As shown in Figure 3-119 on page 212, the **mkfs** command stripes the file system across the first NSD whose preferred I/O node is strg001, the second NSD's preferred I/O node is strg002, and so on. If you are working with an odd number of NSDs, it is easy to see that one Storage node processes more work than the other. It might not be a problem for configurations with many NSDs in a file system. However, be sure to understand this scenario before you commit your configuration.

The mkfs command -R option

The **mkfs** sample command in Figure 3-119 on page 212 includes the **-R meta** option, which specifies the replication of metadata across the NSD failure groups. Replication relates to synchronous replication of GPFS data across NSDs in separately defined failure groups. The **meta** option is the default replication setting, so in this case you do not need to specify it if you want to use it. This statement assumes that both Storwize V7000 subsystems are in separate Failure groups.

There is no value in having more than two failure groups. However, when you have more than one Storwize V7000 subsystem, having two failure groups allows you to protect the status of all metadata if you lose a Storwize V7000 subsystem. If you might have 10 Storwize V7000 subsystems, you can set up five Storwize V7000 subsystems in Failure group 1 and the other five in Failure group 2. It is a preferred practice for reliability.

Replication can reduce performance. It is acceptable to create your Storwize V7000 based file systems with **-R none** (replication set to none) when the back-end storage is securely protected with sound data protection practices. However, without replication in failure groups, the GPFS file system sees one copy of data and metadata (regardless of how many copies exist in the back-end storage).

The options for failure group replication are as follows:

```
-R { none | meta | all }
```

- **none**, which means no replication at all.
- **meta**, which indicates that the file system metadata is synchronously mirrored across two failure groups.
- **all**, which indicates that the file system data and metadata is synchronously mirrored across two failure groups.

3.6.4 Adding DCS3700 LUNs to an existing SONAS configuration

As an example, assume that the file system is created in a logically balanced fashion, and you want to add NSDs to the file system. This section provides an overview for this process, then provides details with screen captures. The following tasks are required:

1. From the DCS3700 subsystem, create the volumes from the DCS3700 storage arrays (one volume per array) or by using disk pools.
2. Map the new SONAS volumes to the SONAS Storage nodes (the hosts).
3. From the Management node in the SONAS cluster, run **`mkdisk --luns`** as the admin user (see Figure 3-120).

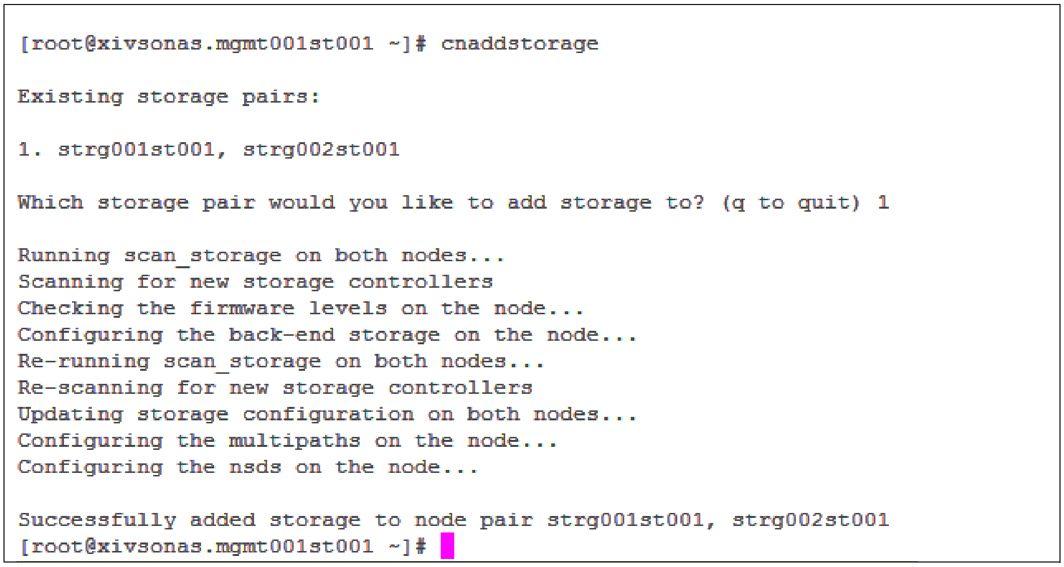
```
[root@xivsonas.mgmt001st001 ~]# mkdisk --luns
(1/3) Scanning for new devices
(2/3) Creating NSDs
(3/3) Adding to database
Successfully created disks:
DCS3700_360080e50002ea78c00000cfd51b74569
DCS3700_360080e50002ea78c000014bf520d2acb
DCS3700_360080e50002ee7bc00000c5551b792ea
DCS3700_360080e50002ee7bc000012fc520d192a
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-120 The mkdisk command

An alternative method is to run **`cnaddstorage`** as the root user, as shown in Figure 3-121.

Select the Storage node pair to which you are adding volumes.

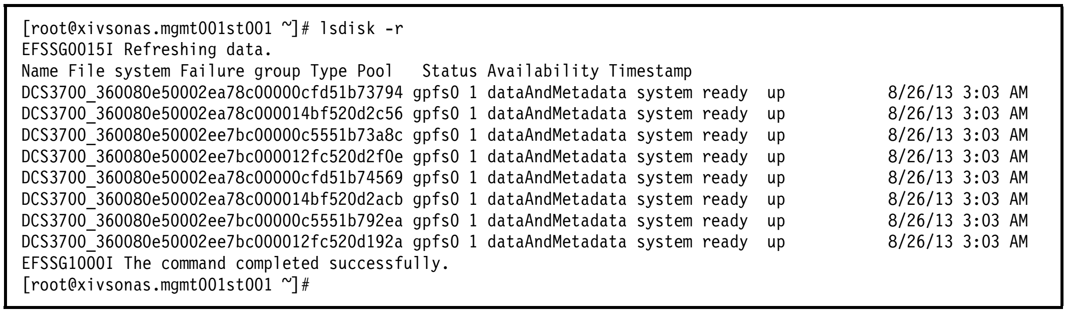
SONAS discovers newly attached storage, creates multipath devices from them, and creates NSDs from those multipath devices. This process takes about six minutes to complete.



```
[root@xivsonas.mgmt001st001 ~]# cnaddstorage

Existing storage pairs:

1. strg001st001, strg002st001

Which storage pair would you like to add storage to? (q to quit) 1

Running scan_storage on both nodes...
Scanning for new storage controllers
Checking the firmware levels on the node...
Configuring the back-end storage on the node...
Re-running scan_storage on both nodes...
Re-scanning for new storage controllers
Updating storage configuration on both nodes...
Configuring the multipaths on the node...
Configuring the nsds on the node...

Successfully added storage to node pair strg001st001, strg002st001
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-121 The cnaddstorage command

4. On the Management node, run **lsdisk -r** to confirm your newly added devices and verify that they are ready to be added to file systems, as shown in Figure 3-122.



```
[root@xivsonas.mgmt001st001 ~]# lsdisk -r
EFSSG0015I Refreshing data.
Name File system Failure group Type Pool   Status Availability Timestamp
DCS3700_360080e50002ea78c00000cfd51b73794 gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ea78c000014bf520d2c56 gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ee7bc00000c5551b73a8c gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ee7bc000012fc520d2f0e gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ea78c00000cfd51b74569 gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ea78c000014bf520d2acb gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ee7bc00000c5551b792ea gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
DCS3700_360080e50002ee7bc000012fc520d192a gpfs0 1 dataAndMetadata system ready  up          8/26/13 3:03 AM
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 3-122 The lsdisk output

5. The additional disks and NSDs can now be added to an existing file system for more capacity, or they can be used to create a file system. To alter the disk parameters, run **chdisk**.

3.7 Other considerations

This section describes configuration that is not directly related to the underlying storage platform but must be considered when you configure SONAS.

3.7.1 Split-brain and quorum devices

The IBM SONAS system quorum is a group of nodes that determine the health of IBM SONAS system components and decide when to discontinue an unhealthy node. There are no default quorum nodes; you must specify which nodes have this role. A valid quorum is one half plus one of the explicitly defined quorum nodes. The benefit of this quorum-based technique is that it enforces consistent operation in a distributed system. If a node is determined by the quorum members to be unhealthy, the quorum can vote to discontinue the unhealthy node.

A quorum must exist for a node to be able to mount or access a file system. This requirement prevents any failing node from writing corrupted data to the system. If there is a quorum loss, the IBM SONAS system unmounts the file systems on all the nodes and attempts to reestablish a quorum and initiate a file system recovery.

The number of quorum nodes is initially defined when the GPFS cluster is defined and set up during the first-time installation process. Quorum devices must be defined in odd numbers with a minimum of three nodes. There are three sizes of *quorum node configuration* for SONAS installation (small = 3, medium = 5, and large = 7). If a cluster is designed as small (2 - 6 Interface nodes), you need three quorum nodes. If the cluster is medium (6 - 10 Interface nodes), set five quorum nodes. If the cluster is large (more than 10 Interface nodes), set seven quorum nodes. For a small configuration, where three quorums are defined, if one node fails that is defined as a quorum node, add another node as a quorum node. This proactive approach ensures that in the unlikely event that another quorum node fails, the file systems stay mounted and active.

3.7.2 File system overhead and characteristics

There are two classes of file system overhead in the SONAS file system. One is the basic overhead of a file system and the overhead that is required to manage an amount of storage. This overhead includes disk headers, basic file system structures, and allocation maps for disk blocks. The second type of overhead is the space that is required to support user usage of the file system. This space includes user directories plus the inodes and indirect blocks for files and potential files.

Both classes of metadata are replicated in a SONAS system for fault tolerance. The system overhead depends on the number of LUNs and the size of the LUNs that are assigned to a file system. It is typically about a few hundred megabytes or less per file system. The metadata in support of usage can be far higher, but is largely a function of usage. The cost of directories is a function of usage and file naming structures. A directory costs at least the minimum file size for each directory and more if the number of entries is large. For a 256 KB block size file system, the minimum directory is 8 KB. The number of directory entries per directory block varies with customer usage. For example, if the average directory contains 10 entries, the cost of a directory is 800 bytes. This number can be doubled for metadata replication.

The cost of inodes is a function of how the file system is configured. By default, SONAS is configured with 50 M inodes that are preallocated and a maximum allowed inodes value of 100 M. By default, an inode requires 512 bytes of storage. The default settings require 50 GB of storage for inodes (512 * 50 M * 2 for replication). If the user has 50 M files with an average directory that holds 10 files, the cost for directories is about 80 GB. Higher-density directories require less space for the same number of files. There might also be a requirement for space for indirect blocks for larger files. These two categories add to overhead for a file system, with other minor usages such as recovery logs or message logs.

3.7.3 SONAS failure groups

SONAS allows you to organize your hardware into failure groups. A failure group is a set of disks that share a common point of failure that can cause them all to become simultaneously unavailable. The SONAS software can provide synchronous RAID 1 mirroring at the software level. In this case, failure groups are defined that are duplicates of each other, defined to be on different disk subsystems. If a disk subsystem fails and cannot be accessed, the SONAS software automatically switches to the other half of the failure group. Expansion racks with storage pods can be moved away from each other up to the length of the InfiniBand cables. Currently, the longest cable that is available is 50 m. This length means that, for example, you can scratch the cluster, move two storage expansion racks to a distance of 50 m, and create a mirror on a failure group level between these two racks.

With the failure of a single NSD, if you have not specified multiple failure groups and replication of metadata, SONAS cannot continue because it cannot write logs or other critical metadata. If you specify multiple failure groups and replication of metadata, the failure of multiple disks in the same failure group creates the same scenario. In either of these situations, GPFS forcibly unmounts the file system. It is a preferred practice for high reliability to replicate at least metadata between two storage pods.

3.7.4 Setting up SONAS storage pools

A storage pool is a collection of disks with similar properties, which provides a specific quality of service (QoS0 for specific use, such as to store all files for a particular application or a specific line of business. Using storage pools, you can create tiers of storage by grouping storage devices based on performance or reliability characteristics. For example, one pool can be an enterprise class storage system that hosts high-performance SAS disks, and another pool can consist of a set of economical Nearline SAS disks. The storage pool is managed together as a group, as the storage pool provide a means to partition the management of the file system's storage. There are two types of storage pools:

- System storage pool (exists by default):

  A storage pool that contains the system metadata (system and file attributes, directories' indirect blocks, symbolic links, policy file, configuration information, and metadata server state) that is accessible to all metadata servers in the cluster. Metadata cannot be moved out of the system storage pool. The system storage pool is allowed to store user data as well, and by default it goes into the system storage pool unless a placement policy is activated. The system storage pool cannot be removed without deleting the entire file system. Disks inside a system pool can be deleted if there is at least one disk that is assigned to system pool or enough disks with space to store existing metadata. System storage pool contains metadata, so use the fastest and the most reliable disks for reasons such as better performance of whole SONAS file system and failure protection. There can be only one system pool per file system, and the pool is required.

- User storage pool:

  Up to seven user storage pools can be created per file system. User storage pools do not contain metadata. They store only data. Therefore, disks that are assigned to a user storage pool can be only of usage type "data only".

A maximum of eight storage pools per file system can be created, including the required system storage pool. A storage pool is an attribute of each disk and is specified as a field in each disk descriptor when the file system is created or when the disk is added to an existing file system.

SONAS offers internal storage pools and external storage pools. Internal storage pools are managed within SONAS. External storage pools are managed by an external application such as Tivoli Storage Manager. SONAS manages the movement of data to and from external storage pools. SONAS provides integrated automatic tiered storage (Integrated Lifecycle Management (ILM)), and provides an integrated global policy engine to enable centralized management of files and file sets in the one or multiple logical storage pools. For more information about storage pools, see the SONAS information in the IBM Knowledge Center.

3.7.5 Integrated Management Module

Each node (IBM System x) has its own Integrated Management Module (IMM) for remote "lights out" management. SONAS monitoring uses IMM to monitor and manage many aspects of the nodes. The IMM can be used for limited use in emergency situations, such as powering off or powering on individual nodes if required. Be careful when you use the IMM because some commands can render the server unusable.

The default user ID is USERID and the default password is PASSW0RD (the number zero, not the letter O). The IMM can be reached by using telnet from the Management nodes. Telnet to the node host name plus “IMM”, such as strg001st001imm to telnet to Storage node1’s IMM. The cluster’s host names are stored in the standard `/etc/hosts` file on the Management nodes, and the IP addresses are determined and set in the hosts file during the first-time installation process. Example 3-2 is an example of powering on a node.

Example 3-2 Power on a node

```
[root@xivsonas.mgmt001st001 ~]# telnet strg003st001imm
Trying 172.31.7.3...
Connected to strg003st001imm (172.31.7.3).
Escape character is '^]'.
Welcome to the server management network terminal!
login : USERID
Password:
Legacy CLI Authorization
system>power on
Ok
system>power state
Power: On
State: OS booted
system>system>exit
Connection closed by foreign host.
[root@xivsonas.mgmt001st001 ~]#
```

> **Note:** You might not be able to telnet to all nodes, especially new nodes that can have the factory default IP addresses and nodes that have not gone through the MES process.

3.8 SONAS integration into your network

This section describes how to integrate your new SONAS system into your existing network environment. The full network integration requires a user authentication method to grant SONAS access, planning public and private networks, and configuration of an IP address load-balancing mechanism.

Before cluster data access authentication can be configured, the cluster management and external IP communication must be configured.

There are a few differences in how the networking is configured based on whether the solution has an Independent Management node service or Integrated Management nodes.

If you have an Independent Management node configuration (a heritage configuration solution from SONAS V1.1 deployments), another configuration might be necessary. It is common that the Management node uses 1 GbE for Management network configuration (access to the GUI and the CLI), in addition to cross-cluster configuration management for replication between clusters, and so on. The Management node automatically configures the networking on the ethX0 ports and the external networking works immediately.

In the most current releases of SONAS, clients choose to use Integrated Management nodes (for built-in Management node role redundancy availability) and have multiple 10 GbE networking connections for IBM SONAS solutions to provide the highest possible performance.

In the second configuration, do not use 1 GbE and the 10 GbE ports on the same systems. In this case, connect the 10 GbE ports (only) on the two Interface nodes that are designated as Management nodes. Move the management services over to the ethX1 ports by running **`chnwmgt`** from the console for the initial installation. With the default settings, the SONAS installation uses ethX*0* for Management node IP configurations.

Note: The **`chnwmgt --interface ethX1 --force`** command moves all the management service communication configurations to the ethX1 interface.

The management and external IP communications can then be serviced on the same designated ports.

Note: Correctly configure the network port devices (bonding) on the network devices before external communication is bound to those ports. The default bonding configuration is Active/Passive (auto failover) between the subordinate ports in the network bond group (typically port 1 and 2 of each network card). Network bonding can be changed later by detaching netgroups from those ports, making your changes, then reattaching the netgroups.

For more information about SONAS networking, see the SONAS information in the IBM Knowledge Center, found at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/pln_t_network.html

3.8.1 Planning IP addresses

This section briefly describes the public and private IP addresses to help prevent conflicts during SONAS use. For more information about these networks, both private and public, see 2.5.2, “Understanding the IP addresses for internal networking” on page 70.

In Table 1-1 on page 21, Question #3, you are prompted for an available IP address range.

SONAS is composed of three different networks. One of these networks is the public network, which is used for SONAS users or administrators to access Interface nodes or Management nodes. The other two networks are the private network (management network), which is used by the Management node to handle the whole cluster, and the data network (InfiniBand network), on which the SONAS file system is built. These two networks, private and data, are not used by SONAS users or the administrator. However, because they coexist on all nodes with the public network, ensure that you do not use the same values, which can cause IP conflicts.

There are only three choices for the private network range. The default setting for public IP addresses is the range 172.31.*.*. If you already use this range in your existing environment, the 192.168.*.* range might be more appropriate. Similarly, if you are using both the 172.31.*.* and 192.168.*.* ranges, the range 10.254.*.* must be used as the private network instead.

To determine the IP address ranges that are currently used on your data center location, contact your network administrators.

3.8.2 Data access and IP address balancing

This section provides information that is required to set up SONAS IP address balancing. This IP balancing is handled both by the DNS and the CTDB layers.

This section describes how the CTDB layer works, in coordination with the DNS, to provide SONAS users access to data.

As described in Chapter 1, “Installation planning” on page 1, some details for your DNS configuration are required. With this information, you can set up the connection between your DNS and your SONAS. For the data access through the client network, SONAS users must mount exports by using CIFS, NFS, or FTP protocols.

Because the SONAS storage solution is also designed for cloud storage, accessing SONAS data must be as transparent as possible from a technical point of view. SONAS users do not have to know or even understand how to access the data; they need only to access it.

This process works because of an appropriate DNS configuration and the CTDB layer. First, the DNS is responsible for routing SONAS user requests to Interface nodes in a round-robin manner. This configuration means that two consecutive connect requests can access data through two distinct Interface nodes.

In the tables in Chapter 1, “Installation planning” on page 1, the `sonascluster.mydomain.com` DNS host name example is used. For consistency considerations, the same name is used in the following schemes, which provide step-by-step descriptions of the DNS and CTDB mechanism in a basic environment. This environment is composed of three Interface nodes, one DNS server, and two active clients. One client runs a Linux operating system and the other one runs a Windows operating system. Again, for consistency considerations, a Management node and storage pods are also represented, even if they do not affect the DNS and CTDB mechanism. The last FTP client is here to provide a reminder of the last protocol in use in SONAS.

The first SONAS user, running the Linux operating system, wants to mount an NFS share on their workstation. To mount the share, run **mount** with the `sonascluster.mydomain.com` DNS host name, as described in the upper left corner in Figure 3-123. This request is caught by the DNS server (step 1), which then looks inside its list of IP addresses and forwards the request to the appropriate Interface node (step 2). It happens in a round-robin way, and it sends an acknowledgment to the Linux SONAS user (step 3). The connection between the first SONAS user and one Interface node is then established, as shown by the dashed arrow in Figure 3-123.

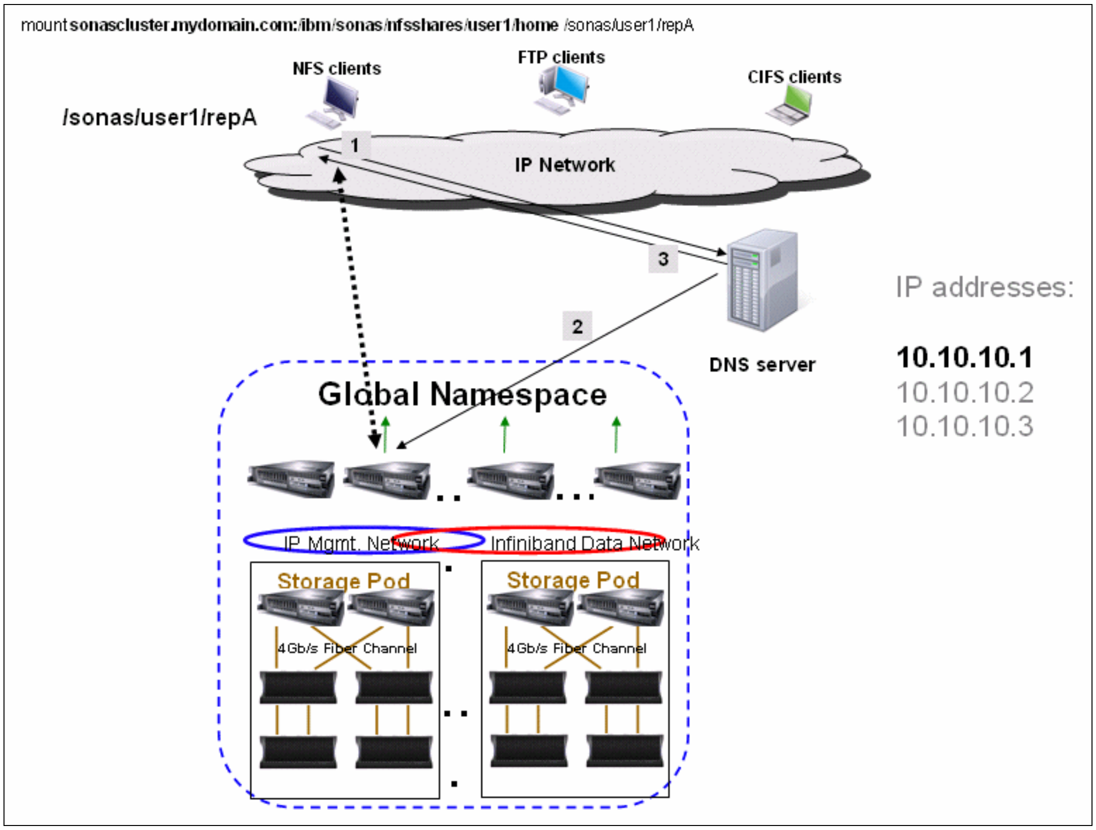
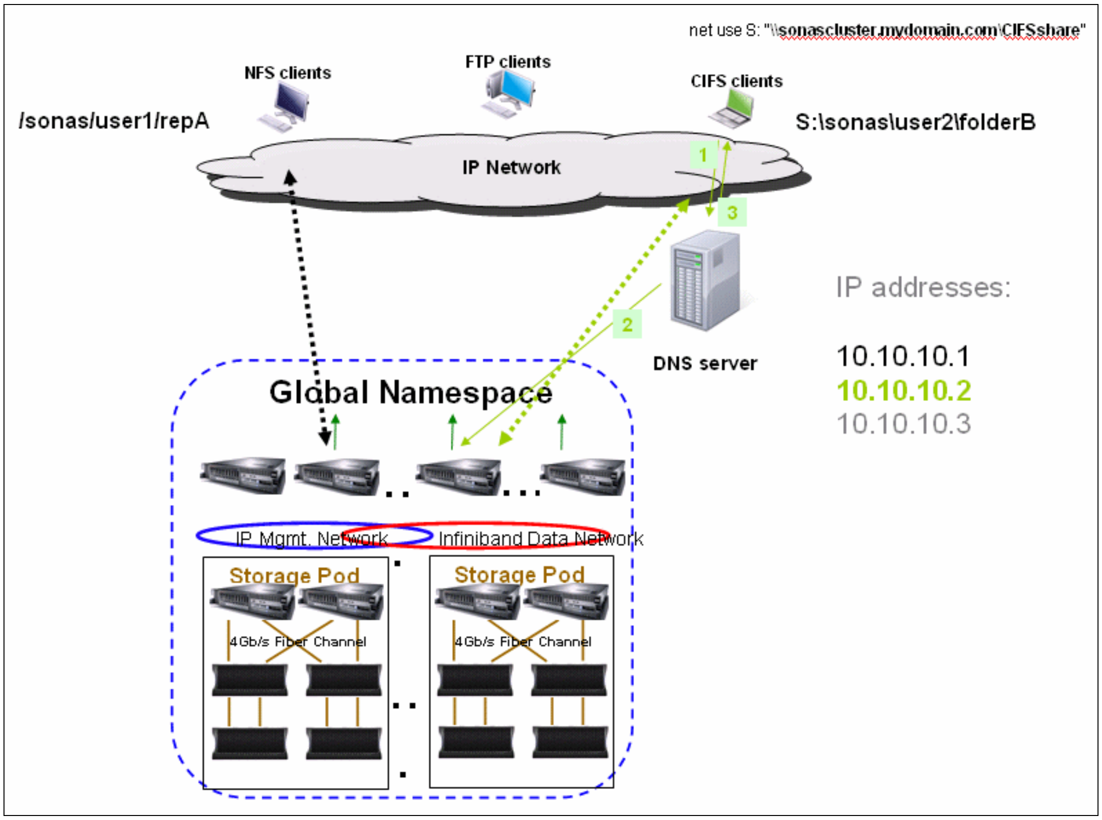


Figure 3-123 A SONAS user is accessing data with the NFS protocol

Now, assume that a second SONAS user also needs to access data that is hosted on the SONAS storage solution by using a CIFS protocol from a Windows notebook. That user runs **`net use`** (or uses the Map Network Drive tool) and uses the same `sonascluster.mydomain.com` DNS host name, as shown in Figure 3-124.

This second request is caught here again by the DNS server, which, in a round-robin way, assigns the next IP address to this second user. Next, steps 1 - 3 are repeated, as described in Figure 3-124. The final connection between the second SONAS user and the Interface node is then established, as shown by the new dashed arrow on the right.



```
net use S: "\\sonascluster.mydomain.com\CIFSshare"
```

Figure 3-124 A SONAS user is accessing data with the CIFS protocol

Connections between SONAS users and Interface nodes remain active until shares are unmounted from SONAS users or an Interface node failure occurs.

For an Interface node failure, the IP address balancing is handled by the CTDB layer, which works with a table. Briefly, this table is re-created when a new event happens. An event can be an Interface node failure or recovery. Table entries are Interface node identifiers and public IP addresses.

In Figure 3-125, the SONAS is configured in such a way that the CTDB has a table with three Interface node identifiers and three public IP addresses for SONAS users.

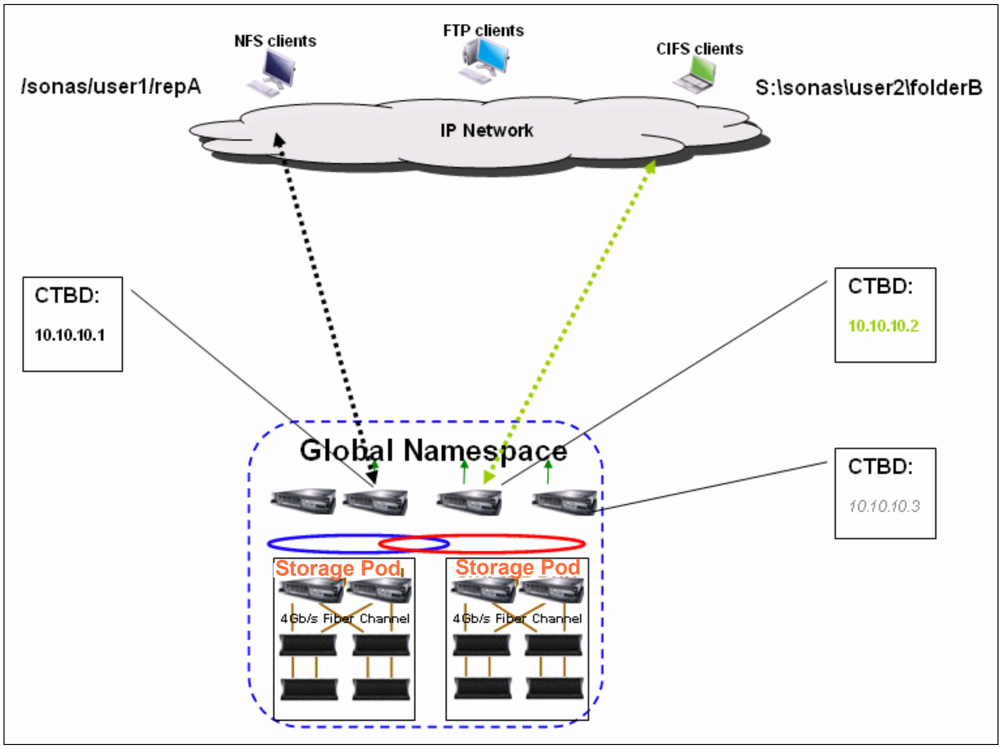


Figure 3-125 CTDB table with three Interface node identifiers and three IP addresses

The example environment has three Interface nodes (#1, #2, and #3) and three IP addresses. The CTDB table is created with these entries:

- #1, #2, and #3
- 10.10.10.1, 10.10.10.2, and 10.10.10.3

For CTDB, the following things are true:

- #1 is responsible for 10.10.10.1.
- #2 is responsible for 10.10.10.2.
- #3 is responsible for 10.10.10.3.

When both SONAS users are connected, as shown in Figure 3-125, only the first two Interface nodes are used. The first Interface node uses the 10.10.10.1 IP address and the second one uses 10.10.10.2, according to the CTDB table.

If the first Interface node (which is in charge of the 10.10.10.1 IP address) fails, this IP address is then handled by the last Interface node, as shown in Figure 3-126.

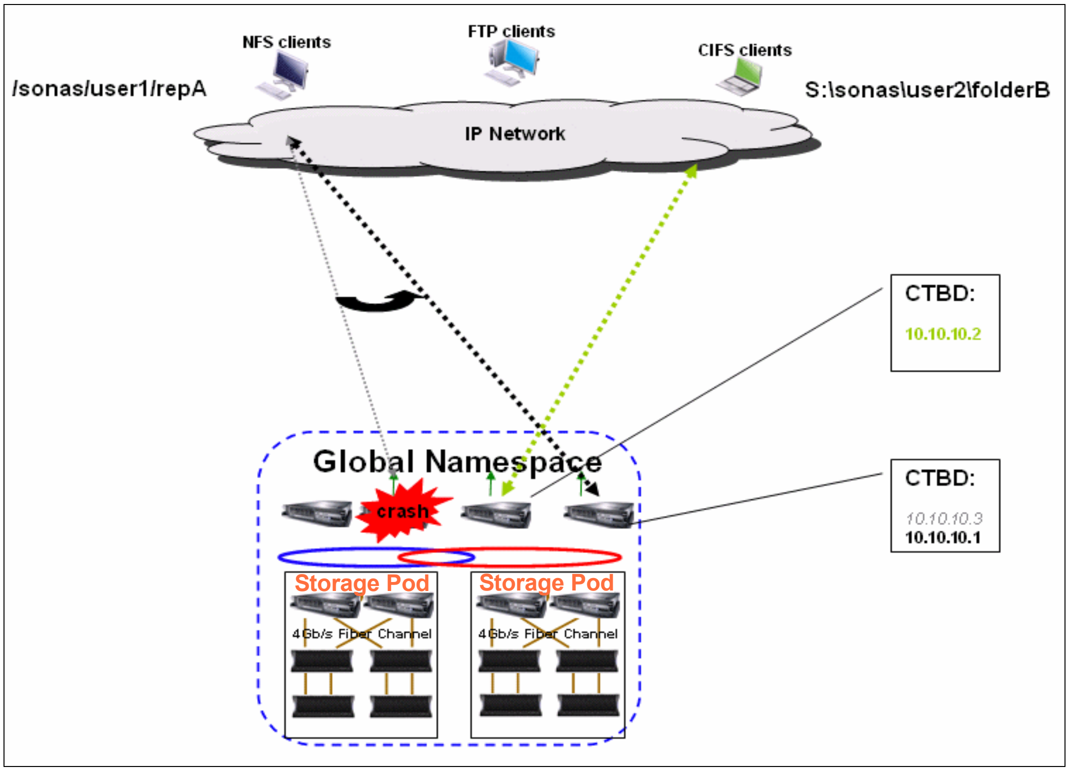


Figure 3-126 CTDB table with Interface node identifiers and IP mappings after failure

For CTDB, because of the failure, the configuration is changed:

- #2 is responsible for 10.10.10.2.
- #3 is responsible for 10.10.10.3 and 10.10.10.1.

As you can see in Figure 3-126, the first NFS SONAS user now has an active connection to the last Interface node. It is basically how the CTDB handles the IP address balancing. Your DNS is handling the round-robin method while the CTDB is in charge of the IP failover.

However, in the previous example, there is a potential load-balancing bottleneck if one Interface node fails. If a third user accesses the SONAS with the FTP protocol, as described in Figure 3-127, the connection is established as shown by the last dashed arrow on the third Interface node. The first NFS user is still connected to the SONAS through the first Interface node. The second CIFS user is connected to the SONAS through the second Interface node. The last FTP user is accessing the SONAS through the third Interface node (the DNS here again gave the next IP address).

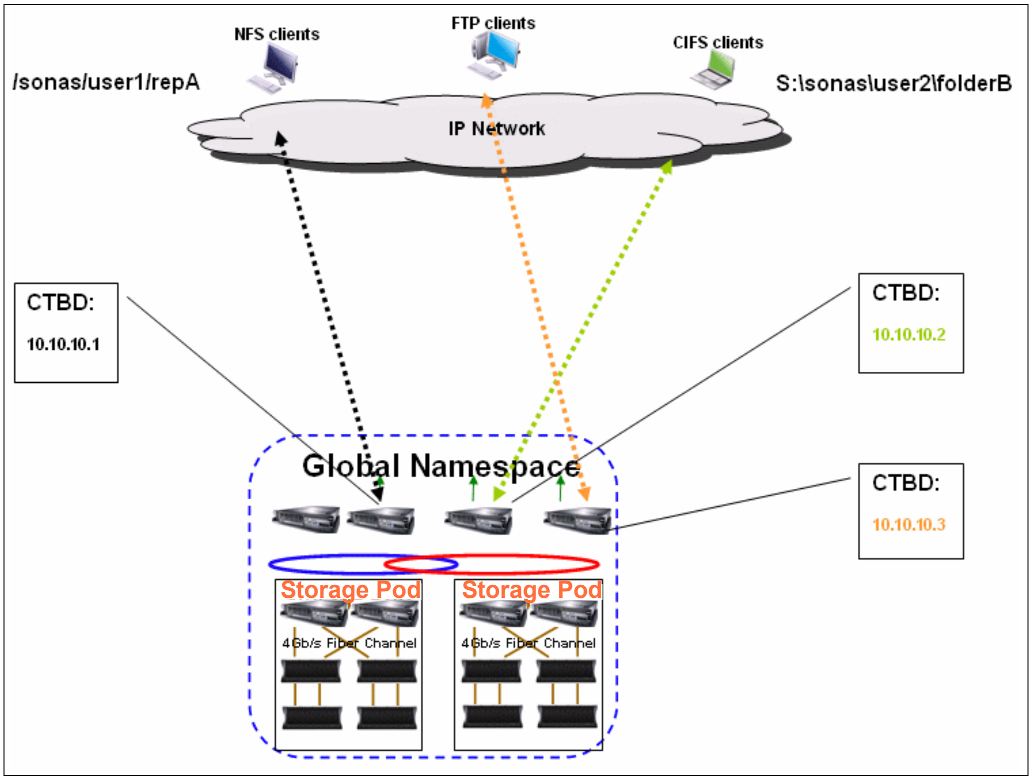


Figure 3-127 CTDB IP address balancing

You might notice that, from here, all incoming users are related to Interface nodes #1, #2, or #3 in the same way because of the DNS round-robin configuration. As an example, you might have four users who are connected to each Interface node, as described in Figure 3-128.

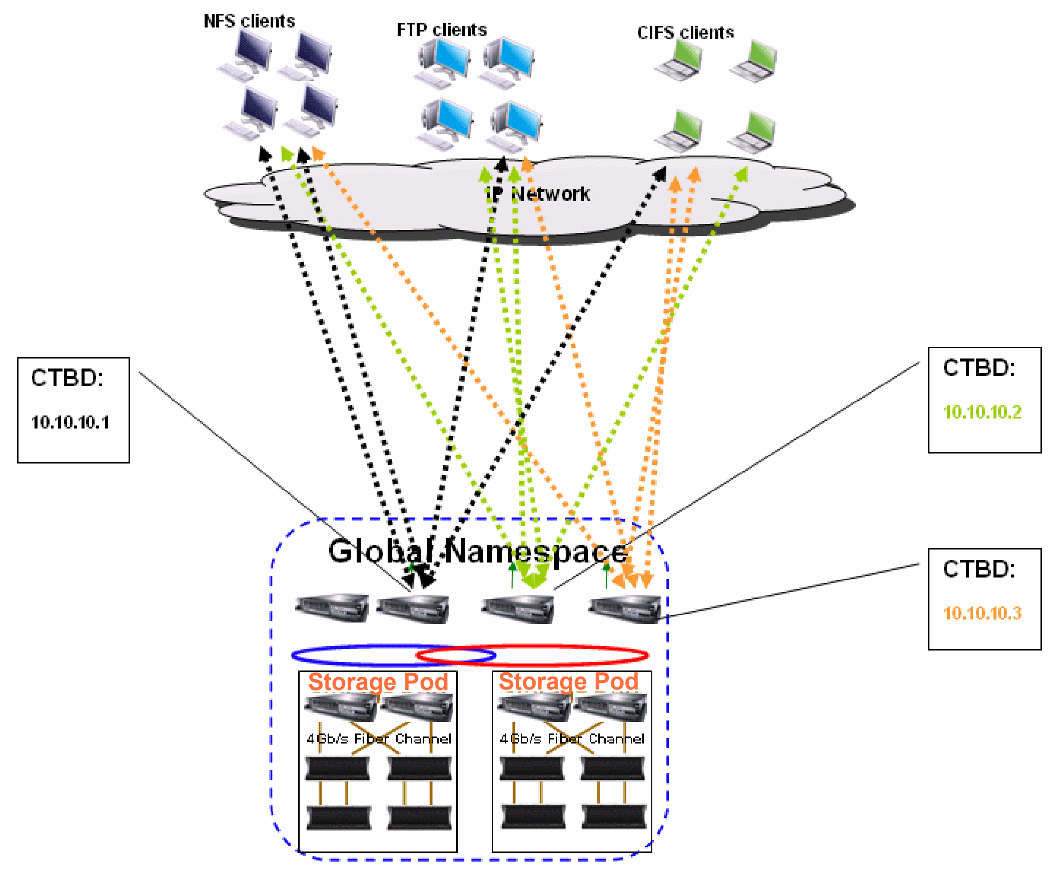


Figure 3-128 Interface node relationships that show CTDB round-robin assignment

The bottleneck that is described earlier in this section occurs if one Interface node fails. Indeed, the IP address that is handled by this failing Interface node migrates, as do all users and their workload, to another Interface node according to the CTDB table. You then have one Interface node that handles a single IP address and four user workloads (second Interface node), and the third Interface node that handles two IP addresses and eight user workloads, as shown in Figure 3-129.

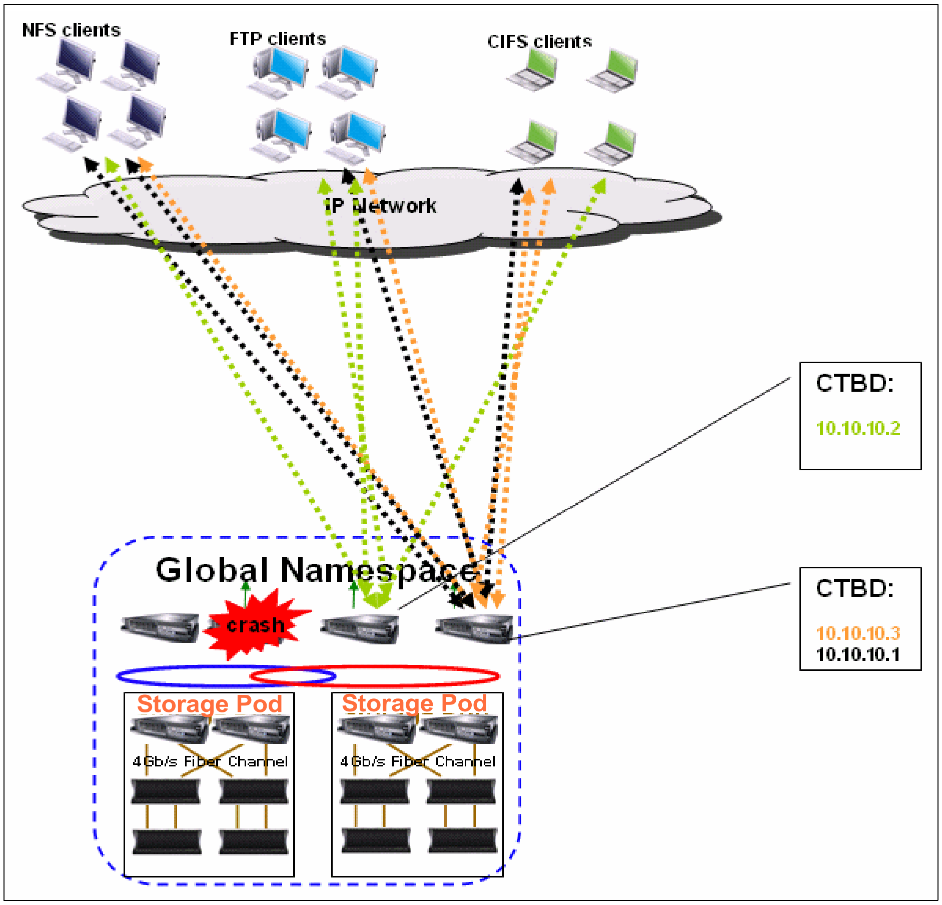


Figure 3-129 Interface node assignment and workload distribution according to the CTDB table

The original overall SONAS users workload was equally load balanced between the three Interface nodes, which each had 33% of the workload. After the Interface node crash, and with the previous CTDB configuration, the workload is now 33% on the second Interface node and 66% on the third Interface node.

To avoid this situation, a simple configuration might be to create more IP addresses than available Interface nodes. Basically, in this example, six IP addresses, two per Interface node, might be more appropriate, as shown in Figure 3-130.

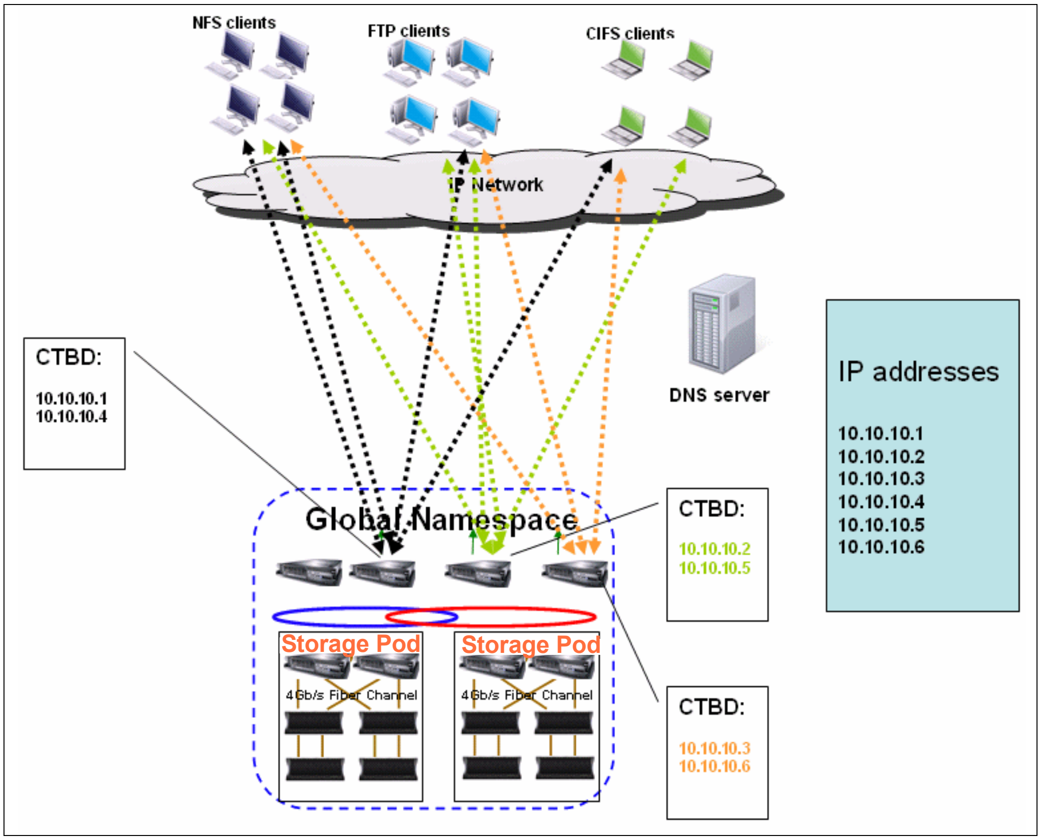


Figure 3-130   CTDB with more IP addresses than Interface nodes assigned

In this case, the original CTDB table is as follows:

- #1 is responsible for 10.10.10.1 and 10.10.10.4.
- #2 is responsible for 10.10.10.2 and 10.10.10.5.
- #3 is responsible for 10.10.10.3 and 10.10.10.6.

If a failure occurs, the failing Interface node (previously in charge of two IP addresses) offloads its first IP address to the second Interface node and its second IP address to the third Interface node. Here is the new CTDB table:

- #2 is responsible for 10.10.10.1, 10.10.10.2, and 10.10.10.5.
- #3 is responsible for 10.10.10.3, 10.10.10.4, and 10.10.10.6.

The results of this situation are a 50-50% workload that is spread across the two remaining Interface nodes after the crash, as shown in Figure 3-131.

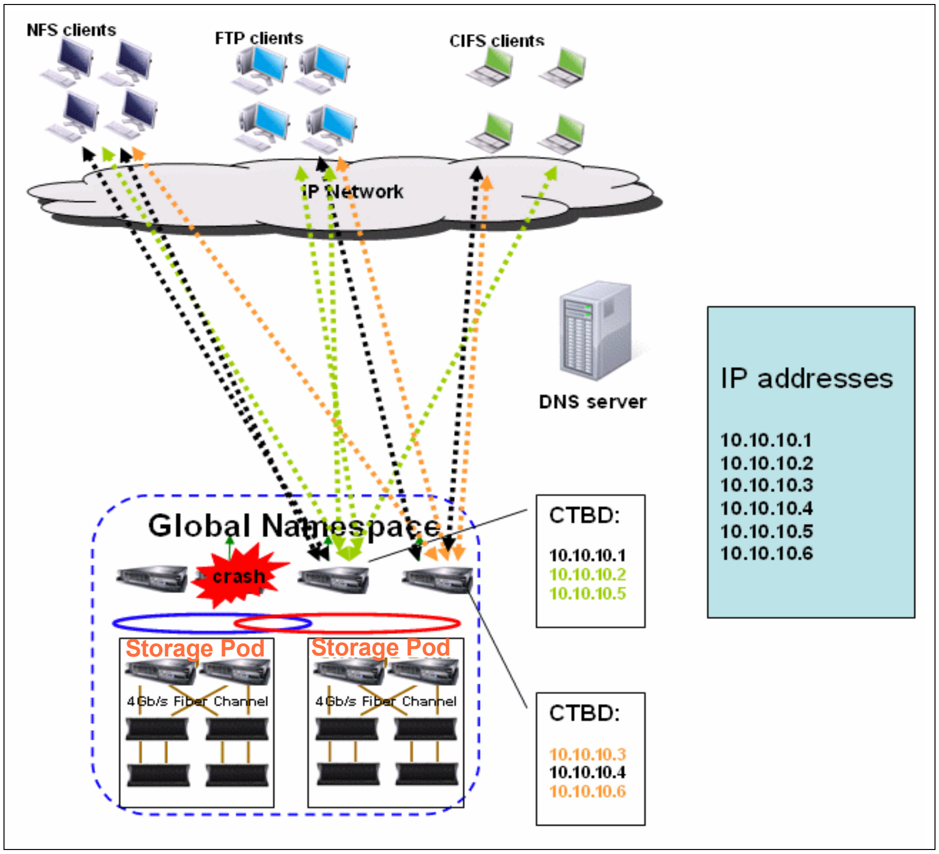


Figure 3-131 Even workload distribution after an Interface node failure

After the first Interface node is back, it is a new event, and the new CTDB table is as follows:

- #1 is responsible for 10.10.10.1 and 10.10.10.4.
- #2 is responsible for 10.10.10.2 and 10.10.10.5.
- #3 is responsible for 10.10.10.3 and 10.10.10.6.

This configuration means that the traffic is load balanced on the three Interface nodes again.

3.8.3 Authentication with Active Directory or Lightweight Directory Access Protocol

You can use your existing authentication method environment to grant user access to SONAS. SONAS supports the following authentication method configurations:

- Microsoft Active Directory (AD) and MS AD + Service for UNIX
- Lightweight Directory Access Protocol (LDAP)
- LDAP with MIT Kerberos
- SAMBA primary domain controller (PDC).

However, SONAS does not support multiple authentication methods that run in parallel. The rule is only one type of authentication method for a cluster and its replication partners.

When a user attempts to access SONAS, they enter a user ID and password. The user ID and password are sent across the customer's network to the remote authentication and authorization server, which compares the user ID and password to valid user ID and password combinations in its local database. If they match, the user is considered to be authenticated. The remote server sends a response to SONAS, which confirms that the user is authenticated and provides the authorization information.

Authentication is the process to identify a user, and authorization is the process to grant access to resources to the identified user.

AD or LDAP configuration can be configured with the **cfgad**, **cfgldap**, and **chkauth** commands.

Microsoft Active Directory

One method for user authentication is to communicate with a remote authentication and authorization server that is running Active Directory software. The Active Directory software provides authentication and authorization services.

To run **cfgad**, you must provide information such as the Active Directory Server IP address and cluster name. This information is collected in the tables in 1.2, “Gathering SONAS requirements” on page 4. The answers to questions #35 - #37 are required for this configuration.

Run the **cfgad** command, as shown in Example 3-3.

Example 3-3 Example of the cfgad command

```
cfgad -as <ActiveDirectoryServerIP> -c <clustername>.<domainname> -u <username> -p
<password>
```

- **<ActiveDirectoryServerIP>**: IP address of the remote Active Directory server, as specified in Table 1-6 on page 26, question #35.
- **<clustername>:** Cluster name, as specified in Table 1-1 on page 21, question #1.
- **<domainname>:** Domain name, as specified in Table 1-1 on page 21, question #2.
- **<username>:** Active Directory user ID, as specified in Figure 1-6 on page 14, question #3.
- **<password>:** Active Directory password, as specified in Figure 1-6 on page 14, question #4.

Example 3-4 shows a sample **cfgad** command with the parameters based on an example lab environment.

Example 3-4 A sample cfgad command that uses parameters based on a sample lab environment

```
cli cfgad -as 9.11.136.116 -c sonascluster.mydomain.com -u aduser -p adpassword
```

To check whether this cluster is now part the Active Directory domain, run **chkauth**, as shown in Example 3-5.

Example 3-5 Sample chkauth command

```
cli chkauth -c <clustername>.<domainname> -t
cli chkauth -c sonascluster.mydomain.com -t
```

- **<clustername>:** Cluster Name, as specified in Figure 1-1 on page 2.
- **<domainname>:** Domain Name, as specified in Figure 1-1 on page 2.

If the **cfgad** command is successful, the output from the **chkauth** command shows "CHECK SECRETS OF SERVER SUCCEED" or a similar message.

LDAP

Another method for user authentication is to communicate with a remote authentication and authorization server that is running LDAP software. The LDAP software provides authentication and authorization services.

To run **cfgldap**, you must provide information the LDAP Server IP address and the cluster name. This information is in Table 1-6 on page 26.

Run **cfgldap**, as shown in Example 3-6.

Example 3-6 Sample cfgldap command

```
cfgldap -c <cluster name> -d <domain name> -lb <suffix> -ldn <rootdn> -lpw
<rootpw> -ls <ldap server> -ssl <ssl method> -v
```

- **<cluster name>:** Cluster name, as specified in Table 1-6 on page 26, Question #7.
- **<domain name>:** Domain name, as specified in Table 1-6 on page 26, Question #8.
- **<suffix>:** The suffix, as specified in Table 1-6 on page 26, Question #9.
- **<rootdn>:** The rootdn, as specified in Table 1-6 on page 26, Question #10.
- **<rootpw>:** The password for access to the remote LDAP server, as specified in Table 1-6 on page 26, Question #11.
- **<LDAP Server IP>**: IP address of the remote Active Directory server, as specified in Table 1-6 on page 26, Question #5.
- **<ssl method>:** SSL method, as specified in Table 1-6 on page 26, Question #6.

Example 3-7 shows the **cfgldap** command with parameters in a lab environment.

Example 3-7 Example cfgldap command with parameters based on a lab environment

```
cfgldap -c sonascluster -d mydomain.com -lb "dc=sonasldap,dc=com" -ldn
"cn=Manager,dc=sonasldap,dc=com" -lpw secret -ls 9.10.11.12 -ssl tls -v
```

To check whether this cluster is now part the Active Directory domain, run the **chkauth** command that is described in Example 3-5 on page 230.

3.9 Attaching SONAS to customer applications

This section is a summary of what you need to keep in mind before you integrate your SONAS into your existing infrastructure and use it.

3.9.1 Redundancy

SONAS is designed as a high availability storage solution. It relies on hardware redundancy and software high availability with GPFS and CTDB. As you plan to integrate SONAS into your existing infrastructure, you must ensure that all external services or equipment is also highly available. For example, SONAS requires an Active Directory server (or LDAP) for authentication. Therefore, you must ensure that this authentication server is redundant.

Likewise, you must consider your NTP and DNS servers. You must have redundant power for your hardware. Also, consider the network switches for the public network.

3.9.2 Access to shares

As described in 3.8.2, “Data access and IP address balancing” on page 220, configuration is required if you want to use directly the IP address instead of the DNS host name. You must attach your SONAS system to your existing DNS and use a DNS round-robin configuration to provide loa.d balancing for user SONAS IP requests to all Interface nodes. This configuration does not provide workload balancing.

For the CTDB layer, 3.8.2, “Data access and IP address balancing” on page 220 also shows you how to configure your IP public network and CTDB to load balance the workload from one failed Interface node to the remaining ones. The typical SONAS use is to map one SONAS user to a single Interface node to use the caching inside Interface node. You might need to use the same CIFS share twice from the same SONAS user (through two drive letters), and then use two Interface nodes. However, do not do use this configuration with NFS shares. Because of the NFS design, the NFS protocol needs to send metadata to different NFS services that might be on two separate nodes in such a configuration.

3.9.3 Migration considerations

If you plan to migrate your existing environment and business applications to a SONAS storage solution, be aware that NAS storage is not always the most appropriate option. If your business application writes or reads data from a locally attached solution (DAS), you might increase the latency on a storage base solution. Similarly, if your application makes many writes, even small ones, on a locally attached solution, it can quickly overload your network switches.

A workaround for these requirements is to first use caching on the client side to reduce the higher bandwidth impact on performance, and to combine I/O requests on the client side to reduce I/O size. You can also modify your application to be more tolerant in case of packet loss or timeout expiration because of the IP protocol, and make it try again.

3.9.4 Backup considerations

There are also preferred practices for backing up your storage. First, stop your application cleanly process to have consistent data, then take a snapshot and use it for backup processes while you restart your application.

4

Authentication

This chapter describes authentication methods and SONAS integration.

This chapter describes the following topics:

- Basic authentication concepts
- Basic authorization concepts
- Configuring SONAS authentication
- Configuring SONAS with Active Directory
- Configuring SONAS with Active Directory + SFU
- Configuring Active Directory with NT4 and PDC
- Configuring SONAS with LDAP
- Configuring SONAS with Network Information Service
- Configuring SONAS with local authentication
- Listing the authentication method that is configured on SONAS
- Checking the authentication setting for the cluster
- Cleaning up authentication
- Working with the ID map cache
- Working with SONAS authorization

4.1 Basic authentication concepts

This section describes authentication concepts.

4.1.1 Overview of authentication

The objective of authentication is to verify the claimed identity of users and components. As the first process, authentication provides a way of identifying a user, typically by having the user enter a valid user name and valid password before access is granted. Typically, the process of authentication is based on each user having a unique set of criteria for gaining access. The authentication server compares a user's authentication credentials with user credentials that are stored in a database. If the credentials match, the user is granted access to the resource. If the credentials are at variance, authentication fails and resource access is denied.

For more information, see the Authentication basic topics information in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_authentication_basic_concepts.html

4.1.2 Mapping of user IDs, group IDs, and security identifiers

SONAS stores all user data on GPFS file systems that use user IDs (UIDs) and group IDs (GIDs) for access control. For CIFS access, SONAS must map security identifiers (SIDs) to UIDs or GIDs to enforce access control. NFS clients send the UID and GID of a user that requests access to a file. SONAS uses the Linux default access control mechanism by comparing the received UID and GID with the UIDs and GIDs that are stored in GPFS.

The UIDs and GIDs that are used by the NFS clients must match the UIDs and UIDs that are stored inside GPFS.

For HTTP, SFTP, and SCP access, SONAS requires users to authenticate with a user name. SONAS must map the user name to one UID and one or more GIDs for GPFS access control.

Figure 4-1 on page 235 shows how the authentication works in SONAS with the ID mapping. SONAS maps the user names and groups to the user ID and group ID across all nodes.

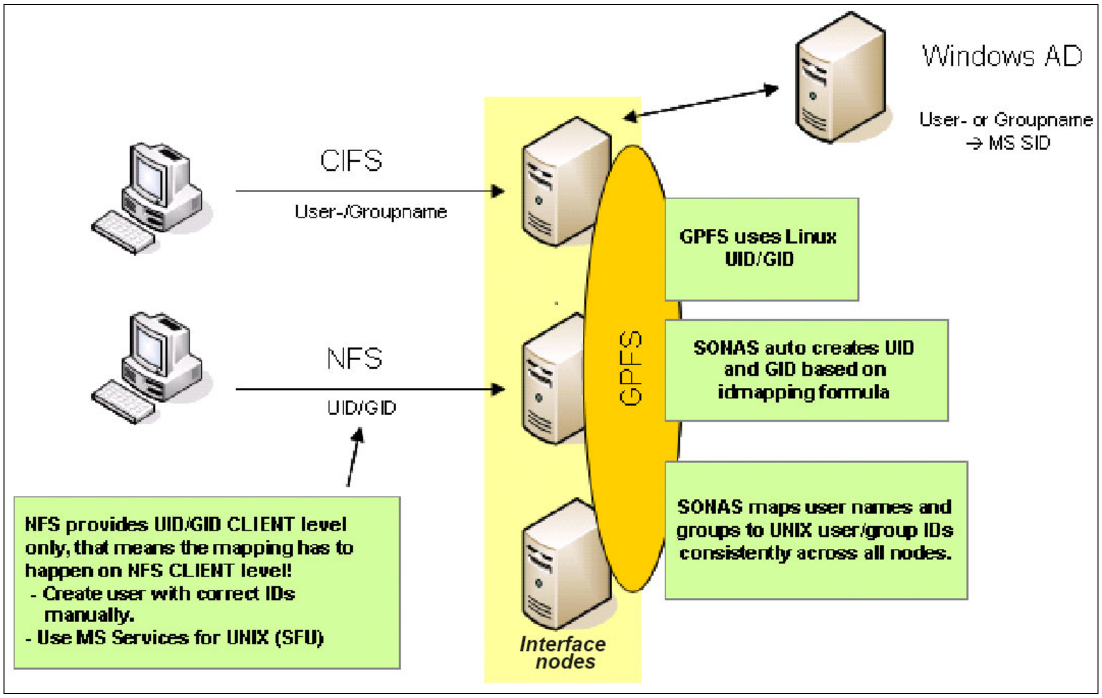

Windows AD
User- or Groupname
→ MS SID
CIFS
User-/Groupname
NFS
UID/GID
GPFS
GPFS uses Linux UID/GID
SONAS auto creates UID and GID based on idmapping formula
SONAS maps user names and groups to UNIX user/group IDs consistently across all nodes.
NFS provides UID/GID CLIENT level only, that means the mapping has to happen on NFS CLIENT level!
- Create user with correct IDs manually.
- Use MS Services for UNIX (SFU)
Interface nodes

Figure 4-1 SONAS authentication and ID mapping

When a CIFS client that uses Microsoft Windows connects to SONAS, it first contacts the Microsoft Active Directory (AD) to check for a user name and password combination. The UID and GID pair is automatically created by using the formula for calculating the ID map. For more information, see 4.4, “Configuring SONAS with Active Directory” on page 239.

For NFS access from UNIX clients, the UID is provided by the UNIX client. For mixed access from Windows and UNIX, AD with SFU or LDAP can be used.

For SFU, the ID mapping is stored in SFU. For LDAP, the ID mapping is stored on the external LDAP server.

4.2 Basic authorization concepts

This section describes basic concepts of authorization.

The objective of *authorization* is to grant or deny an already authenticated identity (SONAS user, SONAS administrator, or IBM service personnel) access to resources (read a file that is stored on SONAS or run a privileged command).

The objective of *access control* is to assure that only authenticated and authorized identities get access to certain resources. Access control must prevent unauthorized access (for example, reject a SONAS CLI command or reject read access to a SONAS user for a file that is stored on SONAS and is owned by another user).

Generally, an *access control list (ACL)* is a list of permissions that is attached to a resource. An ACL describes which identities are allowed to access the resource (for example, read, write, and execute). ACLs are the built-in access control mechanism of the UNIX and Windows operating systems. SONAS uses the Linux built-in ACL mechanism for access control to files that are stored on GPFS.

For more information, see the “Managing authorization and access control lists” topic in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_authorization_topic_welcome.html

4.2.1 GPFS NFSv4 ACLs

There are a broad range of ACL formats that differ in syntax and semantics. The ACL format that is defined by NFSv4 is also called NFSv4 ACL. GPFS ACLs implement the NFSv4 style ACL format, which is sometimes referred as GPFS NFSv4 ACL. SONAS stores all user files in GPFS. The GPFS NFSv4 ACLs are used for access control of files that are stored on SONAS.

For more information, see the “ACL permissions required to work on files and directories” topic in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_authorization_permission.html

4.2.2 POSIX bits

The POSIX bits of a file are a different way to specify access permissions to files. UNIX file systems provide options to specify the owner and the group of a file. You can use the POSIX bits of a file to configure access control for the owner, the group, and for all other users to read, update, or run the file. POSIX bits are less flexible than ACLs.

> **Important:** Changing the POSIX bits of a GPFS file system starts a modification of its GPFS NFSv4 ACL. Because SONAS uses GPFS NFSv4 ACLs for access control, SONAS administrators and IBM service personnel should *never change the POSIX bits of files that are stored on GPFS.*

For more information, see the “ACL permissions required to work on files and directories” topic in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_authorization_permission.html

4.2.3 ACL mapping

GPFS NFSv4 ACLs and CIFS ACLs are not compatible. For example, CIFS supports unlimited nested groups, which are not fully supported by GPFS NFSv4 ACLs. SONAS maps CIFS ACLs on a “best can do” basis to GPFS NFSv4 ACLs, which results in some limitations. In this aspect, SONAS is not fully compatible with CIFS.

For more information, see the “Authorization limitations” topic in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/adm_authorization_limitations.html

4.3 Configuring SONAS authentication

To enable read and write access to directories and files on the IBM SONAS system, you must configure the IBM SONAS environment for user authentication. Only one user authentication method, and only one instance of that method, can be supported at any time.

SONAS supports different authentication methods and therefore has different ID mapping mechanisms:

- Central ID mapping

  In this method, the ID mappings are stored in an external server. These servers are in most cases the authentication server itself.

  - LDAP Server: UIDs and GIDs for the users and groups are stored on the LDAP server as the central location.
  - SFU: UIDs and GIDs are stored in SFU on the AD. This method requires the SFU schema extension or RFC2307.
  - Network Information Service (NIS): UIDs and GIDs are stored in the NIS server.

  With central ID mapping, both Linux clients and Windows clients can access the user and group information and therefore access data on SONAS.

- Internal ID mapping

  In this method, ID mappings are stored inside SONAS. When you use AD, without SFU or NIS or with SambaPDC/NT4, the ID mappings are created and stored on SONAS.

  For SONAS Version 1.5.1 and later, with support for a local authentication server, the ID mappings are created and stored locally within SONAS. The local ID mapping method uses a reserved ID range and allocates UIDs or GIDs on first-come, first-served incremental basis, or as assigned by the administrator, users and groups are created.

  A simple algorithm is used to calculate the ID mappings. The SID for user and groups on AD is converted to a relevant UID and GID. This ID mapping is distributed across the cluster by CTDB.

  UID/GID are not calculated by auto increment. These values are generated based on the following formula:

  ID = RANGE_LOWER_VALUE + RANGE_SIZE * DOMAIN_NUMBER + RID

  Where:

  - RANGE_LOWER_VALUE and RANGE_SIZE are specified on SONAS when you run the **cfgad** command.
  - SID is fetched from the authentication source.
  - DOMAIN_NUMBER is assigned to each of the domains that SONAS recognizes.

These values include the domain with which SONAS is configured and all the other domains that are in trust with this domain. The domain numbers are assigned sequentially as and when there is access from that domain.

The DOMAIN_NUMBER for the same domain can be different on different SONAS systems, which implies that the same user/group has different a UID/GIDs on different SONAS clusters.

For SONAS V1.5.1.0, there are changes to the way that automatic internal ID maps are generated that supports multiple ranges to be allocated to domains and also provides support to maintain a consistent ID map across different clusters. The ID mapping that is generated, meaning the UID and GID for every Windows user and group, is created as follows:

ID = RANGE_LOWER_VALUE + (RANGE_SIZE * DOMAIN_NUMBER) + RID - (MULTIPLIER*RANGE_SIZE)

- RANGE_LOWER_VALUE, RANGE_SIZE: These are specified on SONAS when you run the **`cfgad`** command. For more information, see the man page for **`cfgad`** in the IBM Knowledge Center at the following website:

  http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/manpages/cfgad.html
- RID: Fetched from the authentication source.
- DOMAIN_NUMBER:= Assigned to each domain that SONAS recognizes and starts with 0.
- MULTIPLIER: Generated internally by SONAS.

The default value for range is 10000000 - 299999999, and the default value for rangesize is 1000000. Thus, with default values, 290 domains each of size 1000000 can be mapped. The lowerID of the range must be at least 1000 and the rangesize must be at least 2000.

For more information and a support matrix for SONAS authentication and ID mapping, see the “Managing authentication and ID mapping” topic in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_auth_srv_topic_welcome.html

4.3.1 Configuring SONAS authentication with the SONAS GUI

You can use the SONAS GUI in addition to the CLI to configure authentication for SONAS V1.4.1 and later.

4.3.2 SONAS configuration with multiple instances of the same directory service

When you configure authentication with AD, SONAS does not support configuring the cluster to multiple AD servers. Only configuration against a single AD server is allowed.

However, configuration that includes support for multiple AD servers (multiple AD servers means different AD servers that serve distinct domains) is possible when the AD servers are in two-way trust with each other. Thus, configuring against the single AD server makes all the domains (and hence support multiple AD servers) who are in trust with the SONAS visible.

When you configure SONAS with AD plus SFU, enabling SFU for a trusted domain requires a two-way trust between the principal and the trusted domain.

For LDAP servers, multiple LDAP servers can be configured if they are replicas of the same master LDAP server, or are any LDAP host with the same schema, and contain data that is imported from the same LDAP Data Interchange Format (LDIF) file.

4.3.3 Authentication server location

The authentication server is external to SONAS and requires proper connectivity to and from the SONAS. The authentication server must be configured separately. The SONAS GUI or CLI does not provide any way to configure or manage the external authentication server. This is true even for the Kerberos KDC server. For Version 1.5.1 and later, SONAS supports a local authentication server that is internally hosted in the SONAS cluster.

4.3.4 Supported data access protocols

SONAS provides server-side authentication configuration for various protocols, which includes CIFS, FTP, SCP, NFS, and HTTP. For NFSv3, only the protocol configuration is done. However, for Kerberos, a few configuration steps for NFSv3 must be done on SONAS nodes. Because authentication happens on the NFSv3 client side, authentication must be configured on the client side.

For SONAS V1.5.1 and later, NFSv4 (the IBM user space implementation of the NFS version 4 protocol user space) is also supported with some restrictions. For more information about NFSv4, see *IBM SONAS Best Practices*, SG24-8051.

4.3.5 Nodes that are configured for SONAS authentication

Only the SONAS Interface and Management nodes are configured for authentication by users. Back-end nodes (Storage nodes) are not part of this configuration.

> **Attention:** The time must be synchronized for all SONAS nodes and the authentication server (such as the Kerberos KDC server). Authentication does not work if the time is not synchronized. The authentication configuration does not ensure synchronization, so it must be manually configured.

4.4 Configuring SONAS with Active Directory

To use AD, SONAS must be configured and joined to the AD domain by using the **`cfgad`** command (This command automatically creates the required computer account in AD). An administrator account with privileges to join the machine is required to run the **`cfgad`** command.

The cluster name that is specified during installation (by using the **`cfgcluster`** command) is used as the machine account name. You can check the cluster name by using the **`lscluster`** command. If a machine account by this name exists on AD, SONAS uses that machine account.

SONAS determines the preferred domain controller from the AD Server name and uses it for authentication.

The **`cfgad --preferredDC`** option forces SONAS to use a dedicated preferred domain controller. If you omit the **`--preferredDC`** option, the SONAS system locates all available domain controllers automatically.

4.4.1 Choosing SONAS with Active Directory

Here are considerations for using SONAS with AD:

- You use AD to store user information and user passwords.
- You do not use SFU.
- You do not plan to use Storwize V7000 Unified remote replication.

4.4.2 ID mapping methods that are available for SONAS with AD

The following ID mapping methods are available for SONAS with AD:

- Microsoft AD with internal automatic ID mapping.

> **Note:** For SONAS V1.5.1 and later, the internal automatic ID mapping component is used to export the ID maps and import the map in another cluster, which eliminates the requirement for an external ID mapping server to support features such as asynchronous replication and remote caching.

- Microsoft AD with SFU
- Microsoft AD with NIS
- Local ID mapping with a local authentication server with SONAS V1.5.1

4.4.3 Prerequisites for configuring SONAS with Active Directory

The **cfgad** command requires an administrative account with privileges to join a computer. This user can join SONAS as a machine account in the AD domain. SONAS does not store the provided AD administrator ID and password. A temporary AD administrator ID is sufficient and it can be removed after the **cfgad** command successfully completes.

For more information, see the “Authentication limitations” topic in the IBM Knowledge Center, found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/adm_authentication_limitations.html

4.4.4 Understanding ID mapping and range for Active Directory

ID mapping specifies the **range** and **rangesize** of the UID/GID pool. The default value for **range** is 10000000 - 299999999 and the default value for **rangesize** is 1000000. With default values, 290 domains each of size 1000000 can be mapped, based on the following formula:

Number of Domains = <lowerID of the range> - <higherID of the range>:<rangesize>

The lowerID of the range must be at least 1000 and the **rangesize** value must be at least 2000.

The **rangesize** (number of IDs per domain) defines the available number of UIDs/GIDs per domain. When a user or group is defined in AD, it is identified by an SID that includes a component that is called a Relative Identifier (RID).

If the RID of any user is greater than the **rangesize**, then that user cannot access SONAS exports. So, select **rangesize** to allow the highest possible RID of users and groups. The RID value depends on the number of users and groups. Consider the planned growth of users or groups when you select a **rangesize** value.

> **SONAS V1.5.1 and later:** With SONAS V1.5.1 and later, the issue of the RID being greater than **rangesize** no longer exists because the autorid algorithm supports multiple ranges for domains. For more information, see 4.6.3, “Deterministic ID mapping support in SONAS V1.5.1” on page 261.

The following section gives an example of how to choose a **rangesize** value:

1. Find the highest RID that has been assigned until now. To determine the highest RID, find the '*rIDNextRID*' attribute in AD.

   One way to find this value is to use the **dcdiag** command on the command prompt of the operating system that is hosting AD. Run the command as follows:

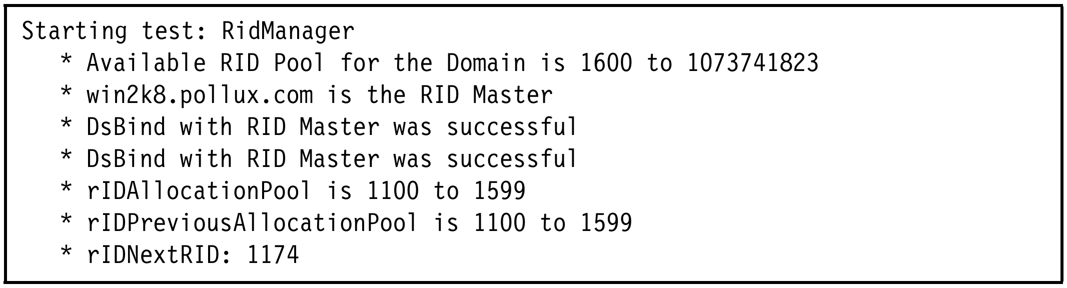
   ```
   "dcdiag /s:{IP of system hosting AD}  /v /test:ridmanager"
   ```

   See Example 4-1.

   Example 4-1 Command to calculate the rangesize value

   ```
   C:\Program Files\Support Tools>C:\Program Files\Support Tools>dcdiag
   /s:10.0.0.123 /v /test:ridmanager
   ```

   The output is shown in Figure 4-2. The highest value for RID, the rIDNextRID, is 1174.



```
Starting test: RidManager
   * Available RID Pool for the Domain is 1600 to 1073741823
   * win2k8.pollux.com is the RID Master
   * DsBind with RID Master was successful
   * DsBind with RID Master was successful
   * rIDAllocationPool is 1100 to 1599
   * rIDPreviousAllocationPool is 1100 to 1599
   * rIDNextRID: 1174
```

   Figure 4-2 Command output that displays the largest RID that was used until now

   Another method to find the *rIDNextRID* value is to run an LDAP query on the following DN Path:

   ```
   CN=Rid Set,Cn=computername,ou=domain controllers,DC=domain,DC=COM
   ```

   If there is more than one domain controller that serves the AD domain, use the highest RID among the domain controllers. If there is more than one domain, use the highest RID among the domains.

2. Determine the expected number of users that will be added to the current number of users.

3. Add the highest RID from step 1 to number of users from step 2. This step forms your **rangesize** value.

Important: Choose the **`rangesize`** value carefully because it cannot be changed after the first ranges for domains are defined.

After you configure IBM SONAS with AD authentication, only the **`higherID`** of the range of **`idMapConfig`** parameter can be increased. No other changes to **`idMapConfig`** parameters are permitted.

4.4.5 Configuring Active Directory with the GUI

This section describes how to configure AD by using the GUI.

Starting the Directory Services configuration in the GUI

To start configuring authentication with the GUI, complete the following steps:

1. Click **Settings** → **Directory Services**. See Figure 4-3.

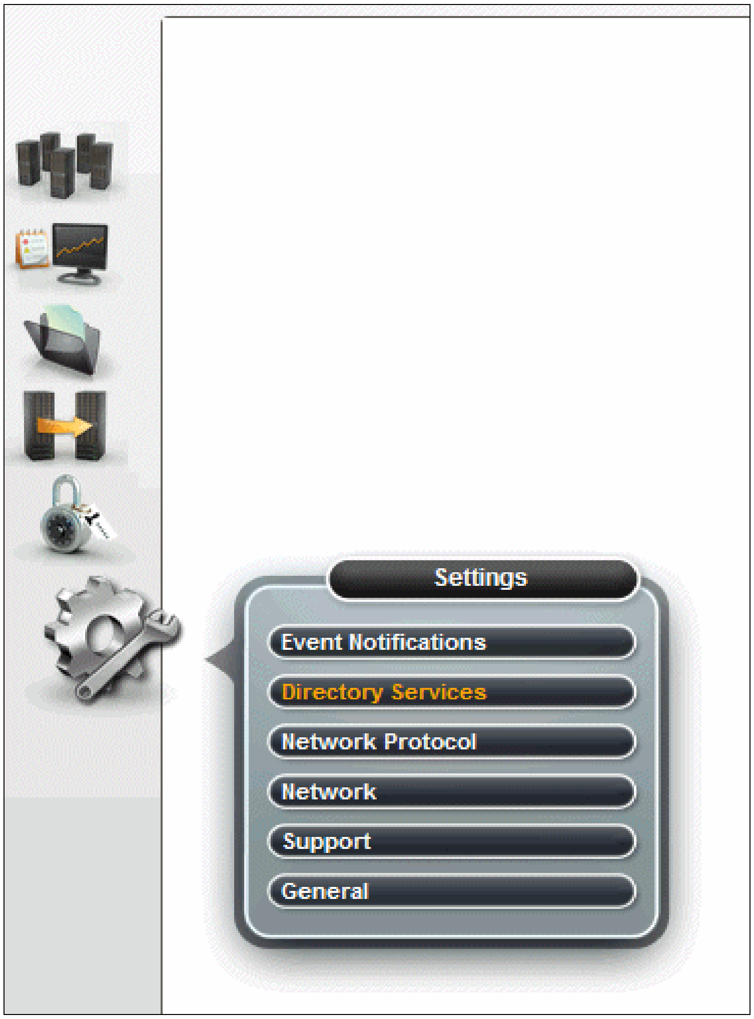
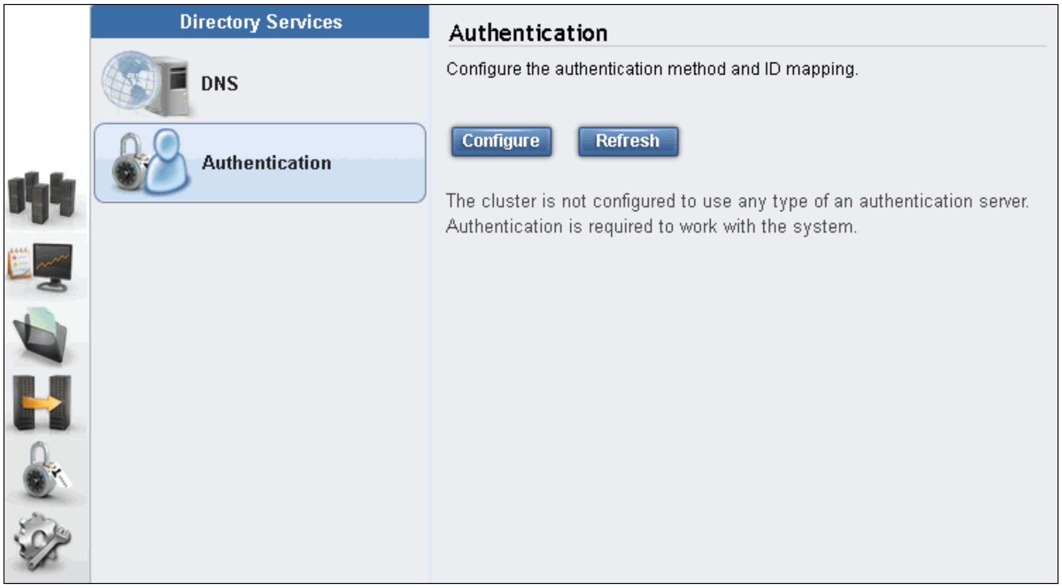


Figure 4-3 Select Directory Services to configure Authentication on SONAS

When you click **Directory Services**, a new page opens, as shown in Figure 4-4 on page 243.

2. Select **Authentication** from the left pane. No authentication method should be configured on the system. If authentication is already configured, it will be overwritten.



Authentication

Configure the authentication method and ID mapping.

The cluster is not configured to use any type of an authentication server. Authentication is required to work with the system.

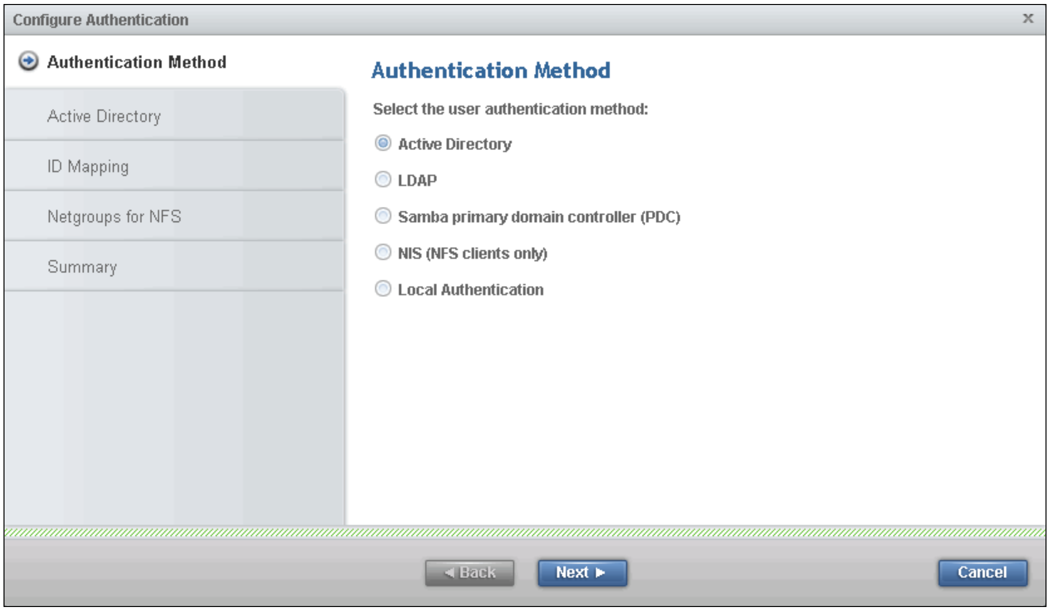
Figure 4-4 Choose Authentication to configure authentication for SONAS

Choosing Active Directory for configuring authentication

This section explains how to select AD authentication.

1. Complete the steps in “Starting the Directory Services configuration in the GUI” on page 242 before beginning this task.

   After you click **Authentication**, a window opens and asks for the type of authentication that you want to configure. For configuring AD, select **Active Directory** from the list. See Figure 4-5.



Authentication Method

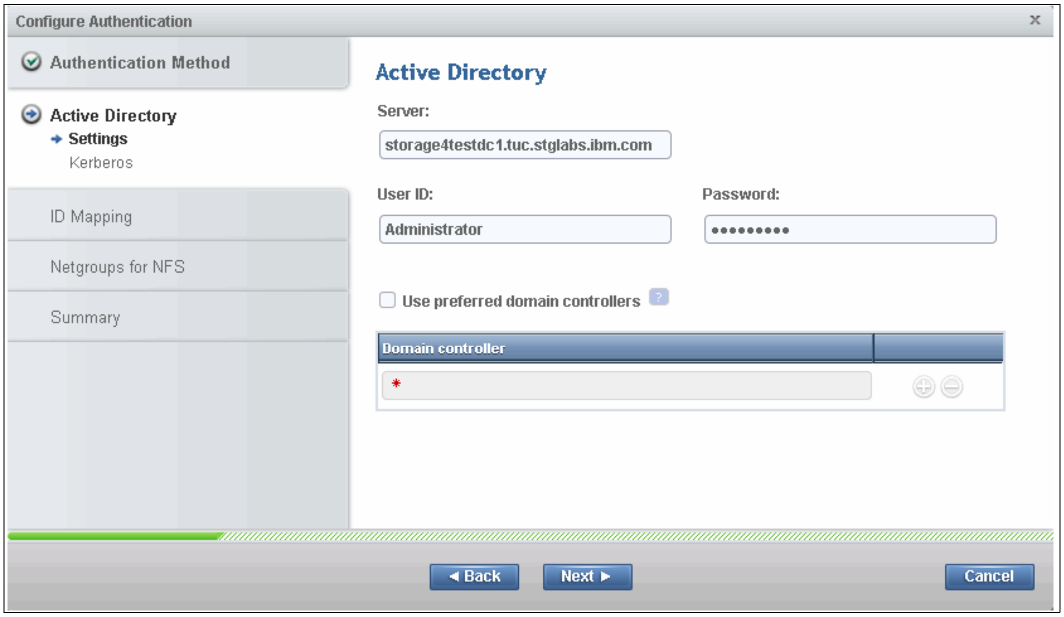
Select the user authentication method:

● Active Directory
○ LDAP
○ Samba primary domain controller (PDC)
○ NIS (NFS clients only)
○ Local Authentication

Figure 4-5 Choose Active Directory as the authentication method

2. Click **Next** to proceed. You see a new window. Complete the Server information and Administrative user information, including the User ID and Password.

   In this example, the `storage4testdc1.tuc.stglabs.ibm.com` server and administrator User ID are used. See Figure 4-6.



Active Directory

Server: storage4testdc1.tuc.stglabs.ibm.com

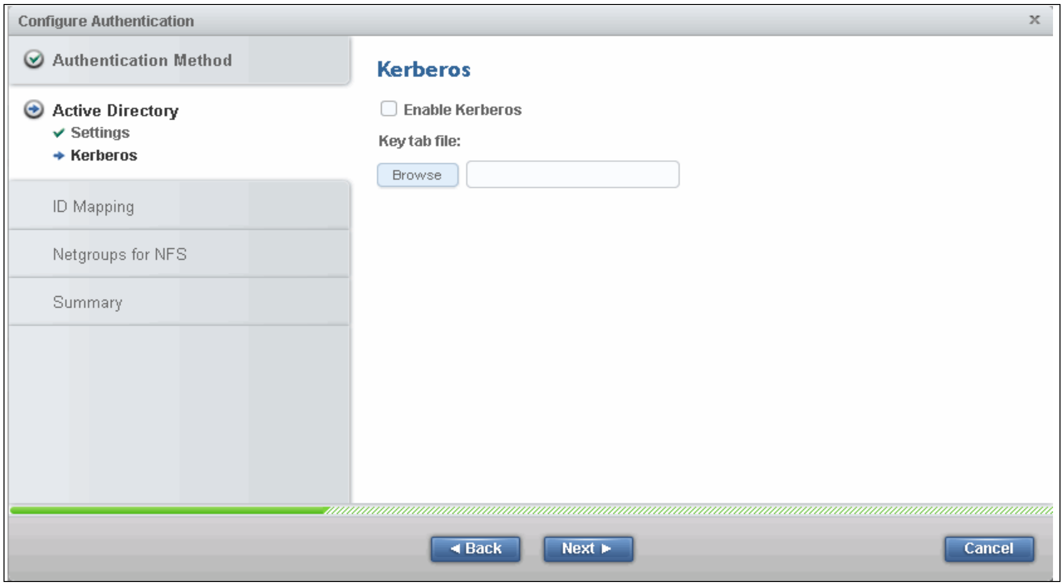
User ID: Administrator

Password: •••••••••

□ Use preferred domain controllers

Figure 4-6 Active Directory Configuration Details required

3. Click **Next** to proceed. You see a window like the window that is shown in Figure 4-7.
4. If you want to configure AD with Kerberos, provide the keytab file. See Figure 4-7.



Kerberos

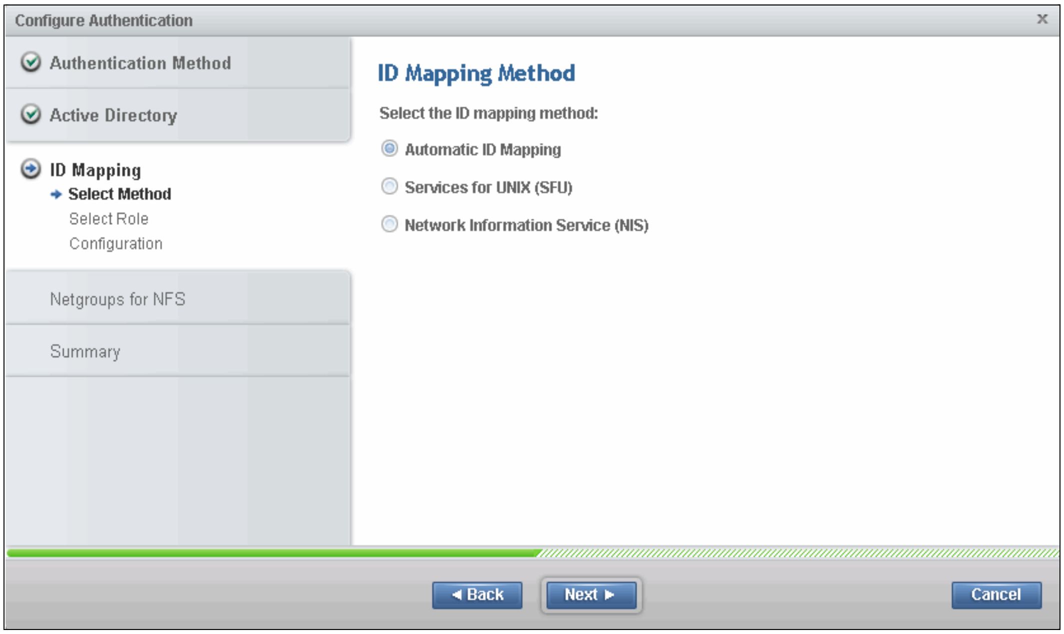
□ Enable Kerberos

Key tab file:

Figure 4-7 Select the keytab file for configuring AD with Kerberos

5. Click **Next** to proceed.

   A new window opens, where you select the ID mapping method to use. In Figure 4-8, Automatic ID mapping is selected. The other options available are SFU, which is described in “Configuring SONAS with Active Directory and NIS by using the CLI” on page 283, and NIS, which is described in 4.8.3, “Active Directory with NIS” on page 279.



ID Mapping Method

Select the ID mapping method:

● Automatic ID Mapping

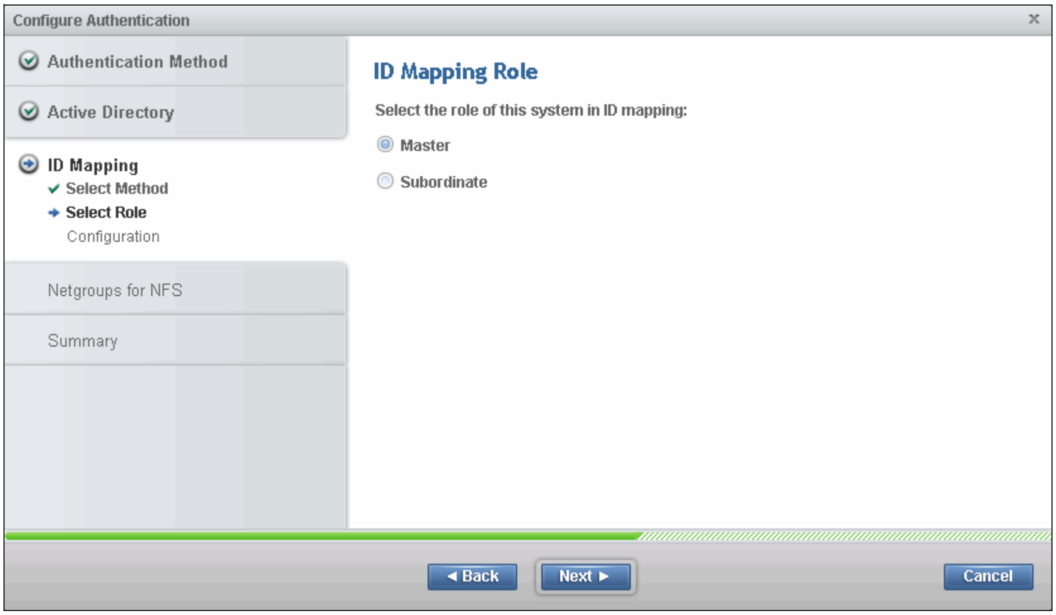
○ Services for UNIX (SFU)

○ Network Information Service (NIS)

Figure 4-8 Select the ID mapping method to use

6. Select a method and click **Next** to proceed.

7. In the window that opens, select the ID mapping role of the cluster, as shown in Figure 4-9. For more information about ID mapping roles, see 4.6.3, “Deterministic ID mapping support in SONAS V1.5.1” on page 261.



ID Mapping Role

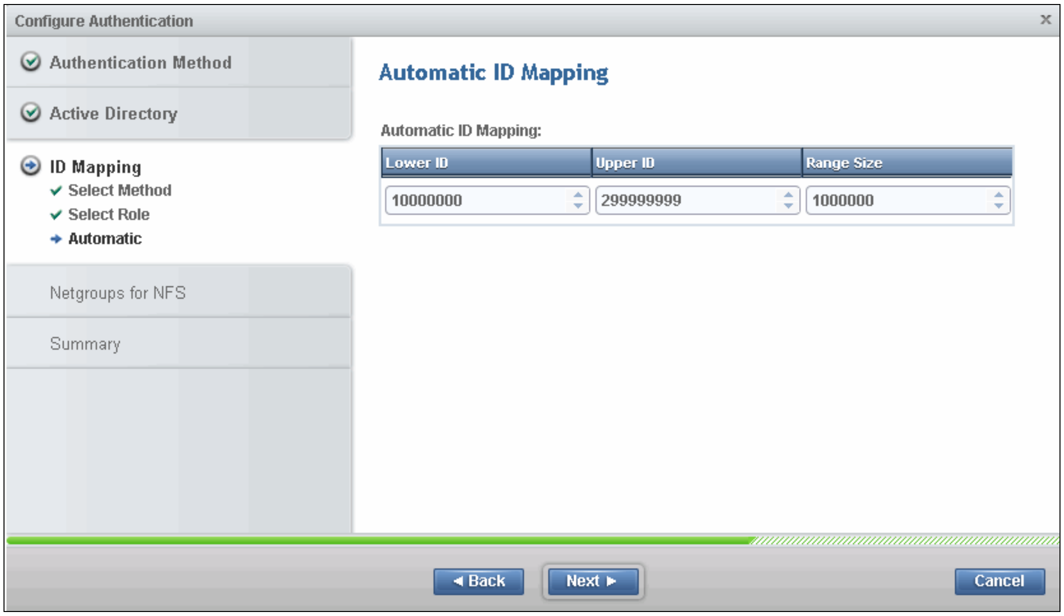
Select the role of this system in ID mapping:

● Master

○ Subordinate

Figure 4-9 Select the ID mapping role of the cluster

8. Select a role and click **Next** to proceed.
9. In the window that opens, select the ID mapping range size and the lower and upper limit for the ID range (see Figure 4-10) and click **Next** to proceed.



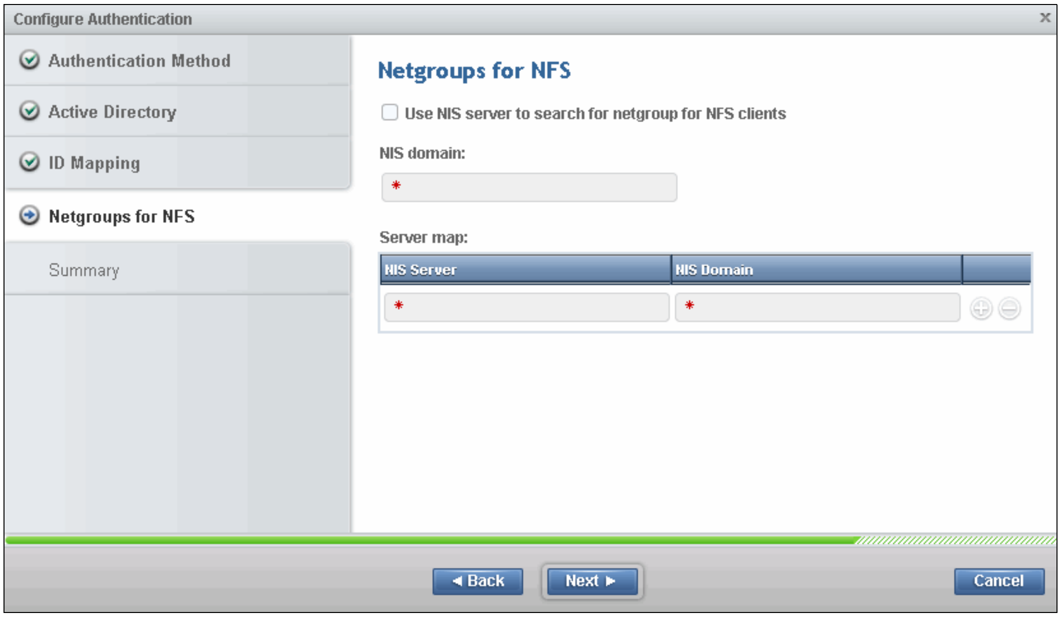
Automatic ID Mapping

Automatic ID Mapping:

Lower ID	Upper ID	Range Size
10000000	299999999	1000000

Figure 4-10 Select the ID map range and limits

10. In the window that opens (Figure 4-11), select the NIS server if you want to support netgroups for NFS clients and click **Next** to proceed.



Netgroups for NFS

□ Use NIS server to search for netgroup for NFS clients

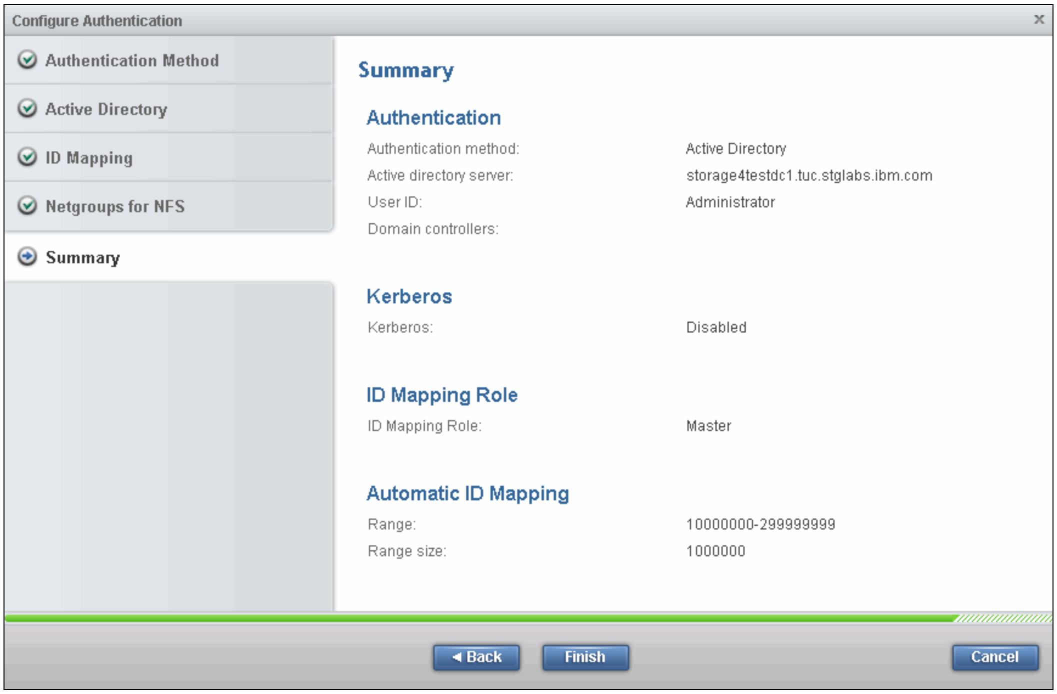
NIS domain: *

Server map:

NIS Server	NIS Domain
*	*

Figure 4-11 Select the NIS server for using netgroups

The final window opens. It displays a summary of the configuration information, as shown in Figure 4-12 on page 247.



Summary

Authentication

Authentication method:	Active Directory
Active directory server:	storage4testdc1.tuc.stglabs.ibm.com
User ID:	Administrator
Domain controllers:	

Kerberos

Kerberos:	Disabled

ID Mapping Role

ID Mapping Role:	Master

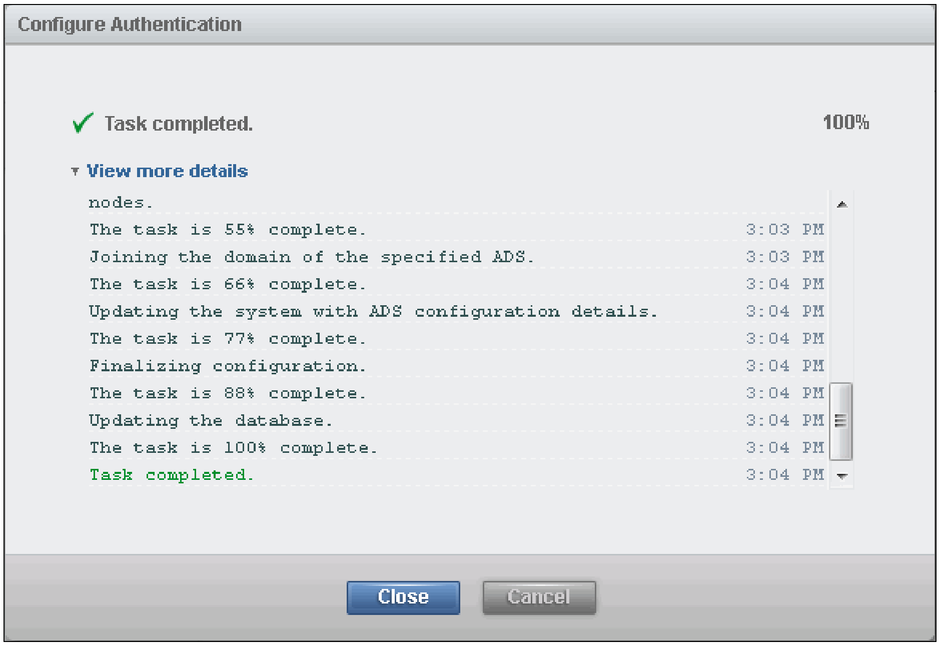
Automatic ID Mapping

Range:	10000000-299999999
Range size:	1000000

Figure 4-12 AD authentication configuration summary

11. When you are satisfied with the summary, click **Finish**.

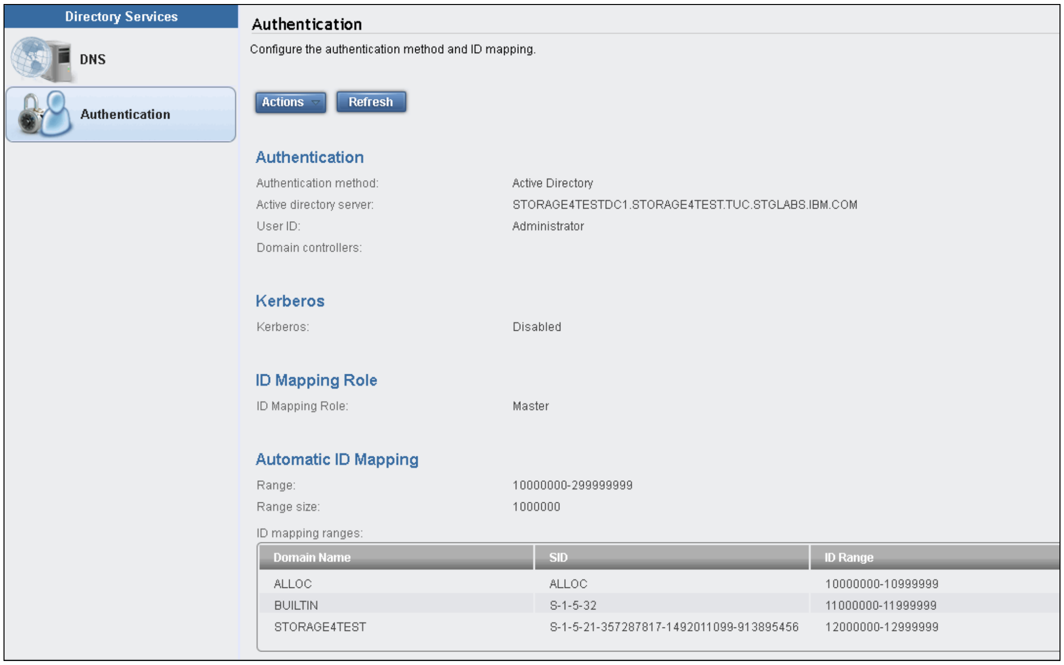
    The configuration starts. You see a window that shows the progress of the command that configures the authentication method that you selected. If the configuration is successful, you see the window that is shown in Figure 4-13. If it fails, you see an error. In this case, correct the error and try again.



```
nodes.
The task is 55% complete.                              3:03 PM
Joining the domain of the specified ADS.               3:03 PM
The task is 66% complete.                              3:04 PM
Updating the system with ADS configuration details.    3:04 PM
The task is 77% complete.                              3:04 PM
Finalizing configuration.                              3:04 PM
The task is 88% complete.                              3:04 PM
Updating the database.                                 3:04 PM
The task is 100% complete.                             3:04 PM
Task completed.                                        3:04 PM
```

Figure 4-13 Active Directory configuration successful

12. Click **Close** to return to the main window of the SONAS GUI, where you see the authentication that has been configured. See Figure 4-14.



Authentication

Configure the authentication method and ID mapping.

Authentication

Authentication method:	Active Directory
Active directory server:	STORAGE4TESTDC1.STORAGE4TEST.TUC.STGLABS.IBM.COM
User ID:	Administrator
Domain controllers:	

Kerberos

Kerberos:	Disabled

ID Mapping Role

ID Mapping Role:	Master

Automatic ID Mapping

Range:	10000000-299999999
Range size:	1000000

ID mapping ranges:

Domain Name	SID	ID Range
ALLOC	ALLOC	10000000-10999999
BUILTIN	S-1-5-32	11000000-11999999
STORAGE4TEST	S-1-5-21-357287817-1492011099-913895456	12000000-12999999

Figure 4-14   Active Directory is successfully configured

4.4.6 Configuring Active Directory with the CLI

The command for configuring AD on SONAS is **`cfgad`**. To configure SONAS with AD, complete the following steps:

1. Run **`lsauth`** to verify that no authentication method is configured. See Example 4-2.

Example 4-2   The lsauth CLI command shows that no authentication method is configured

```
# $ lsauth
EFSSG0571I Cluster furby.storage.tucson.ibm.com is not configured with any type
of authentication server.
```

2. Run **`cfgad`** to configure SONAS with AD. See Example 4-3. This example shows a cluster with name `furby.storage.tucson.ibm.com`, and AD server `storage4testdc1.tuc.stglabs.ibm.com`. See Example 4-5 on page 249 for help with the **`cfgad`** command.

Example 4-3   CLI command cfgad that is used to configure SONAS with Active Directory

```
$ cfgad --adServerName 9.11.139.2  --userName Administrator --password XXXX
--idMapRole master -c furby.storage.tucson.ibm.com
(1/9) Fetching the list of cluster Nodes.
(2/9) Check if cfgcluster has done the basic configuration successfully.
(3/9) Check whether Interface nodes are reachable from management node.
(4/9) Detection of AD server and fetching domain information from AD server.
(5/9) Check whether AD server is reachable from Interface nodes.
(6/9) Joining the domain of the specified ADS.
(7/9) Updating the system with ADS configuration details.
(8/9) Finalizing configuration.
```

```
(9/9) Updating the database.
EFSSG1000I The command completed successfully.
```

3. Run **lsauth** to verify that cluster is configured with AD. See Example 4-4.

Example 4-4   The lsauth CLI command shows that AD is configured successfully

```
# $ lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
idMappingMethod = auto
domain = STORAGE4TEST
clusterName = furby.storage.tucson.ibm.com
userName = Administrator
idMapRole = master
adHost = STORAGE4TESTDC1.STORAGE4TEST.TUC.STGLABS.IBM.COM
passwordServer = *
krbMode = off
realm = STORAGE4TEST.TUC.STGLABS.IBM.COM
EFSSG1000I The command completed successfully.
```

Example 4-5 shows the help for the **cfgad** command. The command accepts the server name or IP of the AD server and the user name and password for the authentication server. You can also specify a comma-separated list of domain controllers, which has priority over the other domain controller that is serving the same domain. This setting is optional. The command also accepts the range and range size of the UID/GID pool.

Example 4-5   Command help for the cfgad CLI command

```
$ cfgad --help
usage: cfgad  -s <adServerName> -u <userName> [-p <password>] --idMapRole
<idMapRole> [--idMapConfig <idMapConfig>] [--preferredDC <preferredDC>]
[--krbKeytabFile <krbKeytabFile>] [-c < clusterID | clusterName >]
Configures the Active Directory server-based authentication for the cluster.

Parameter           Description
-s, --adServerName  Specifies the name or IP address of the Active Directory
server against which the cluster will  be configured for authentication.
-u, --userName      Specifies the user name for joining the cluster to the
Active Directory domain.
-p, --password      Specifies the password of the user name
    --idMapRole     Specifies the ID map role of the cluster(master or
subordinate)
    --idMapConfig   Specifies the range and range size of the UID/GID pool.
    --preferredDC   Prioritizes the domain controllers.
    --krbKeytabFile Indicates the Kerberos keytab file to use.
-c, --cluster       The cluster scope for this command
```

Configuring AD with Extended NIS: After you configure AD by using the CLI as shown in this section, you can use the **cfgnis** command to configure extended NIS. For more information, see 4.8.3, “Active Directory with NIS” on page 279.

4.5 Configuring SONAS with Active Directory + SFU

SONAS with AD + SFU is the correct choice in the following circumstances:

- You use AD to store user information and user passwords.
- You plan to use NFS for NAS access for UNIX clients.
- You plan to use SONAS remote replication.

By default, SONAS configures AD without SFU/RFC2307. It can be configured to use SFU after the AD configuration that is done by the **`cfgad`** command completes.

The **`cfgsfu`** command is used to configure SFU with AD.

The configuration must be done for each trusted domain. For trusted domains that have no SFU configuration, the default mode (no SFU and no internal ID mapping) is used.

4.5.1 Prerequisites for configuring SONAS with AD + SFU

Here is a summary of the steps to configure SONAS with AD + SFU:

- SONAS is configured with AD authentication (for example, **`cfgad`** was run).
- Microsoft Windows Services for UNIX is installed and all User Name Mapping Maps are created.
- No files are stored on SONAS (see 4.5.2, "Limitations of SONAS with AD + SFU" on page 251).
- The **`cfgsfu`** command does not restrict the UID/GID range. However, use UIDs/GIDs greater than 1024 because using smaller UIDs/GIDS causes a UID/GID collision of the SONAS NAS user, SONAS *admin* user, and SONAS Linux components. This collision is a security concern (SONAS administrators might be able to touch customer data) and some SONAS commands (for example, **`lsquota`**) might report incorrect information. The ID range must not overlap with 10000000 - 299999999. This range is used by SONAS to generate automatically GIDs for groups that are not configured in SFU.
- The primary windows group that is assigned to an AD user must have a GID assigned. Otherwise, the user is denied access to the system.
- Each user in AD must have a valid UID and GID assigned so that they can mount and access SONAS shares. The UID and GID number that is assigned must be within the range that is used with the **`cfgsfu`** command. The users and groups in AD are mapped to a UID and GID that is specified in the UNIX Attribute tab for users and for groups. It is important to enter this value. The preferred practice is that this value is the UID and GID that the users have on NFS clients. Hence, a file on SONAS can now be accessed by user on AD through CIFS and also through NFS from a Linux client.

> **Note:** The primary UNIX group setting in AD is not respected by SONAS. SONAS always uses the primary Windows group as the primary group for the user. This configuration means that new files and directories that are created by a user through CIFS are owned by the user's primary Windows group and not by the primary UNIX group. For this reason, make the UNIX Attribute primary group the same as the Windows primary group that is defined for the user.

4.5.2 Limitations of SONAS with AD + SFU

Consider the following limitations for configuring SONAS authentication with AD and SFU:

- The range 10000000 - 299999999 is reserved for SONAS and must not be used in SFU.
- Do not add SFU after data is stored on SONAS. For more information, see “Understanding ID mapping and range for Active Directory” on page 240.

4.5.3 Understanding range and schema for AD + SFU

The **--configParameter** parameter that is required for configuring SFU is a combination of three parameters namely, **domainName**, **range**, and **schemaMode**:

- Domain Name: This value specifies the trusted domain of the AD server.
- Range: All the users or groups who want to access exports should have a UID and GID in the specified range. The range parameter must be in the format LowerID - UpperID, for example, 20000 - 30000. The allowed range typically is 1 - 4294967295, inclusive.

  The preferred practice is to have the lower range greater that 1024 to avoid conflict with the Management CLI users. If you run the command with a lower range less than 1024, a warning is generated that also asks for confirmation. You can use the **--force** option to override it.

  > **Important:** The specified range should not intersect with the range that is specified by using the **--idMapConfig** option of the AD server authentication configuration, where the default range is 10000000 - 299999999.

  Users and groups that have a UID or GID that do not fit in the range are denied access.
- Schema Mode: This parameter can be either `sfu` or `rfc2307`, depending on the operating system of the domain controllers. If the operating system of the domain to be joined is Microsoft Windows 2008 or Windows 2003 with R2 packages, use `rfc2307`. For Windows 2000 and Windows 2003 with SP1, use `sfu`.

To specify multiple tuples of domain, range, and schema mode, use ";" to separate the tuples. For example:

```
--configParameters "domain1,20000-30000,sfu;domain2,40000-50000,rfc2307"
```

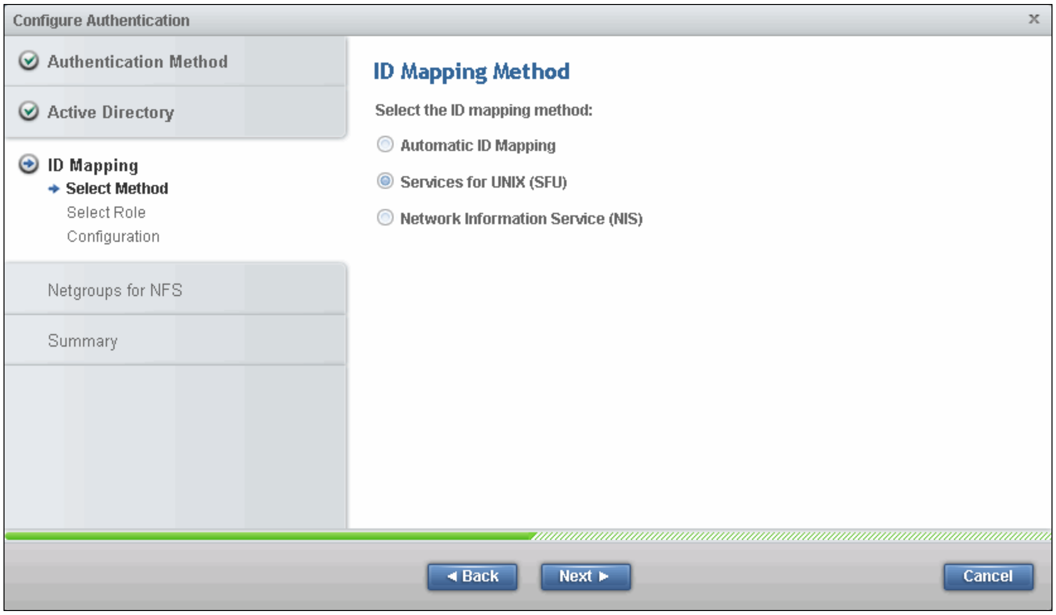
> **Important:** It is important to select the range carefully. Users and groups with a UID or GID that is out of range are denied access.

4.5.4 Configuring AD + SFU with the GUI

To configure AD + SFU with the GUI, complete the following steps:

1. Complete the steps in “Starting the Directory Services configuration in the GUI” on page 242 get to the Authentication window in the GUI.

2. In the window that opens after you click **Authentication**, select **Active Directory** from the list and follow the same steps until you reach Figure 4-8 on page 245. In this window, as shown in Figure 4-15, select **SFU** as the ID mapping method.



ID Mapping Method

Select the ID mapping method:

○ Automatic ID Mapping

● Services for UNIX (SFU)

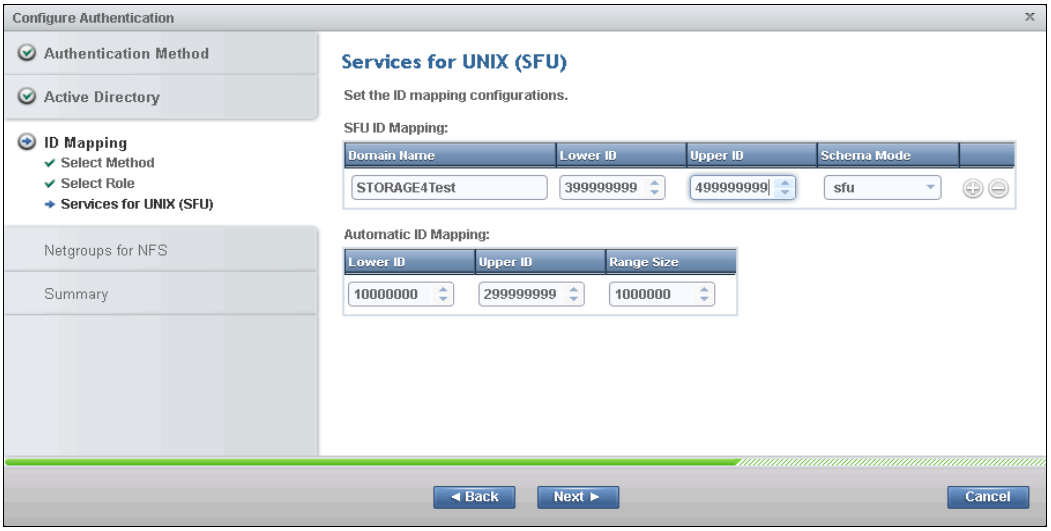
○ Network Information Service (NIS)

Figure 4-15 ID mapping selection

3. Click **Next** to proceed to the next step.

4. Select the ID mapping role of the cluster, as shown in Figure 4-9 on page 245, and click **Next** to proceed.

5. In the window that opens (Figure 4-16), enter the SFU domain name and ID map range. In this example, the STORAGE4Test domain and ID map range of 399999999 - 499999999 are used.



Services for UNIX (SFU)

Set the ID mapping configurations.

SFU ID Mapping:

Domain Name	Lower ID	Upper ID	Schema Mode
STORAGE4Test	399999999	499999999	sfu

Automatic ID Mapping:

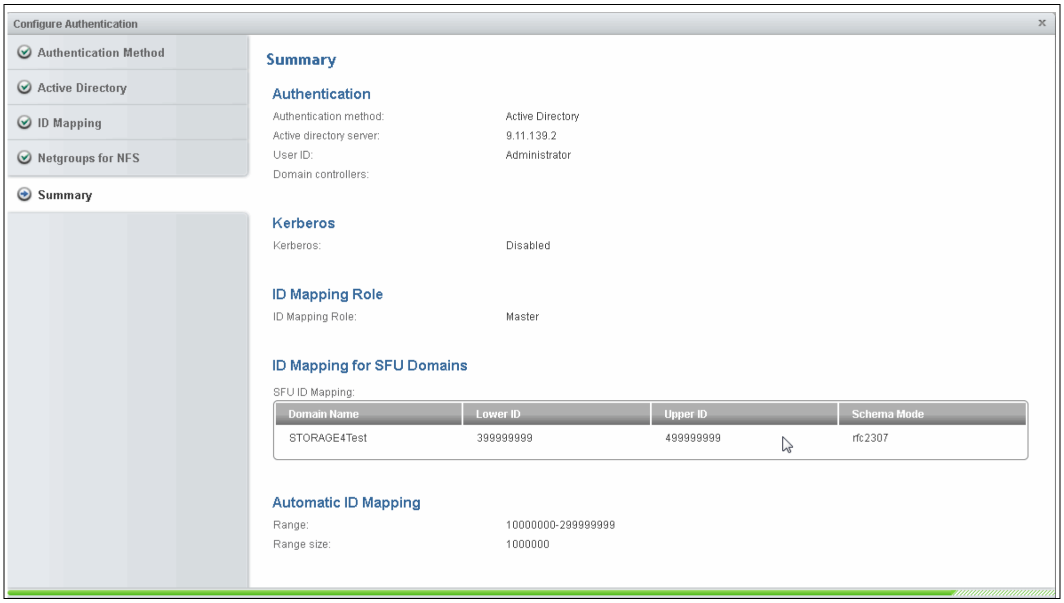
Lower ID	Upper ID	Range Size
10000000	299999999	1000000

Figure 4-16 AD with SFU details required

6. Click **Next**.

7. Enter the netgroup information in the window that opens, as shown in Figure 4-11 on page 246.
8. Click **Next**.

   The next window shows the configuration summary, as shown in Figure 4-17.



Summary

Authentication

Authentication method: Active Directory
Active directory server: 9.11.139.2
User ID: Administrator
Domain controllers:

Kerberos

Kerberos: Disabled

ID Mapping Role

ID Mapping Role: Master

ID Mapping for SFU Domains

SFU ID Mapping:

Domain Name	Lower ID	Upper ID	Schema Mode
STORAGE4Test	399999999	499999999	rfc2307

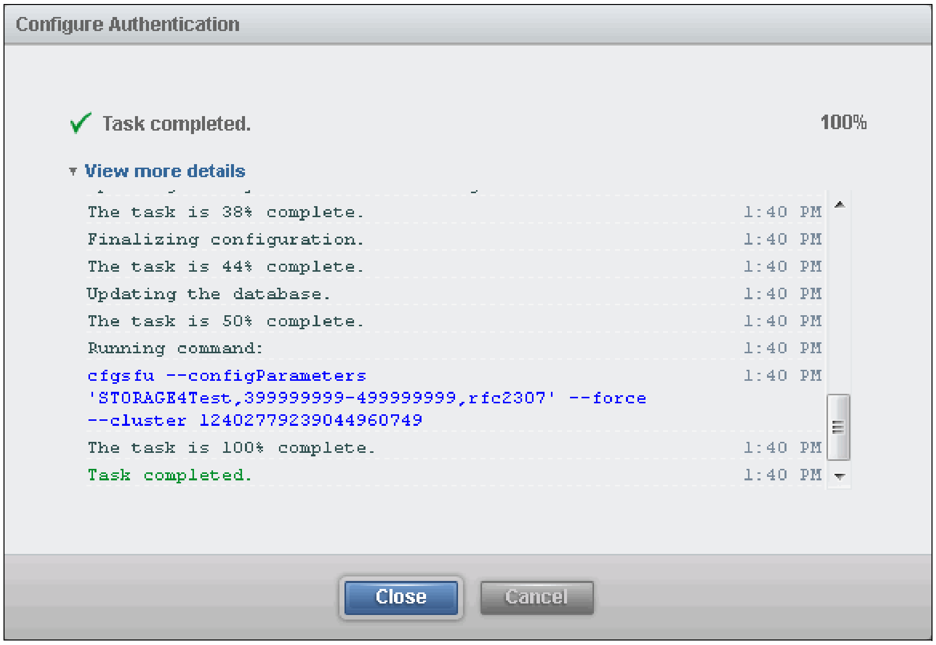
Automatic ID Mapping

Range: 10000000-299999999
Range size: 1000000

Figure 4-17 Configuration summary

9. Click **Finish** to complete the configuration.

   The configuration starts. You see a window that shows the progress of the command that runs to configure the authentication method that is chosen. If the configuration is successful, you see the window in Figure 4-18. If the configuration fails, you see an error. Correct the error and try again.

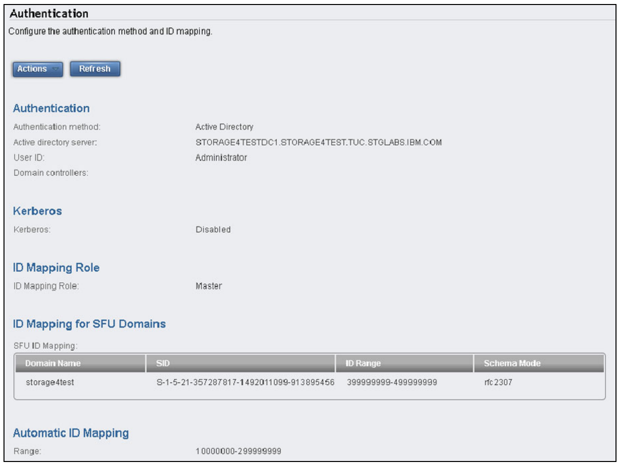


✓ Task completed. 100%

```
The task is 38% complete.                                          1:40 PM
Finalizing configuration.                                          1:40 PM
The task is 44% complete.                                          1:40 PM
Updating the database.                                             1:40 PM
The task is 50% complete.                                          1:40 PM
Running command:                                                   1:40 PM
cfgsfu --configParameters                                          1:40 PM
'STORAGE4Test,399999999-499999999,rfc2307' --force
--cluster 12402779239044960749
The task is 100% complete.                                         1:40 PM
Task completed.                                                    1:40 PM
```

Figure 4-18 Active Directory with SFU configuration successful

10. Click **Close** to return to the main authentication window, where you see the authentication that has been configured (Figure 4-19).



Authentication

Configure the authentication method and ID mapping.

Authentication

Authentication method: Active Directory

Active directory server: STORAGE4TESTDC1.STORAGE4TEST.TUC.STGLABS.IBM.COM

User ID: Administrator

Domain controllers:

Kerberos

Kerberos: Disabled

ID Mapping Role

ID Mapping Role: Master

ID Mapping for SFU Domains

SFU ID Mapping:

Domain Name	SID	ID Range	Schema Mode
storage4test	S-1-5-21-357287817-1492011099-913895456	399999999-499999999	rfc2307

Automatic ID Mapping

Range: 10000000-299999999

Figure 4-19 Active Directory with SFU configured

4.5.5 Configuring AD + SFU with the CLI

The **cfgsfu** command is used to configure SFU in SONAS. This command must be run after you run the **cfgad** command, which initially configures AD initially. To configure SONAS with SFU, complete the following steps:

1. Run **lsauth** to verify that the authentication method that is configured is AD. See Example 4-6. For more information about configuring AD, see "Configuring Active Directory with the CLI" on page 248.

Example 4-6 The lsauth CLI command shows that AD is already configured

```
$lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
idMappingMethod = auto
domain = STORAGE4TEST
clusterName = furby.storage.tucson.ibm.com
userName = Administrator
idMapRole = master
adHost = STORAGE4TESTDC1.STORAGE4TEST.TUC.STGLABS.IBM.COM
passwordServer = *
krbMode = off
realm = STORAGE4TEST.TUC.STGLABS.IBM.COM
EFSSG1000I The AD + SFUcommand completed successfully.
```

2. Run **cfgsfu** to configure SONAS with SFU, as shown in Example 4-7 on page 255. In this example, SFU is used with the VIRTUAL1 domain and a 399999999 - 499999999 ID map. Example 4-7 on page 255 shows an example of the **cfgsfu** command.

Example 4-7 The cfgsfu command to configure SONAS with SFU

```
$cfgsfu --configParameters 'STORAGE4Test,399999999-499999999,rfc2307'
EFSSG1000I The command completed successfully.
```

3. Run **lsauth** to verify that the cluster is configured for AD with SFU. See Example 4-8.

Example 4-8 The lsauth command shows that SFU is configured successfully

```
# lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
domain = STORAGE4TEST
idMappingMethod = sfu
clusterName = furby.storage.tucson.ibm.com
userName = Administrator
idMapRole = master
SFU_storage4test = ad,399999999-499999999,rfc2307
adHost = STORAGE4TESTDC1.STORAGE4TEST.TUC.STGLABS.IBM.COM
krbMode = off
passwordServer = *
realm = STORAGE4TEST.TUC.STGLABS.IBM.COM
EFSSG1000I The command completed successfully.
```

Example 4-9 shows the help for the **cfgsfu** command. The command accepts the configuration parameter of the SFU, which is a comma-separated list of domain name, IDmap range, and schema mode, which can be SFU or RFC2307. The command also accepts a cluster name or cluster ID.

Example 4-9 Help for the cfgsfu CLI command

```
$cfgsfu --help
usage: cfgsfu  --configParameters <configParameters> [--force] [-c < clusterID
| clusterName >]
Configures the cluster with Services For UNIX (SFU) user mapping services.

Parameter              Description
    --configParameters Specifies the parameters with which the Services for
UNIX (SFU) toolkit needs to be configured.
    --force           Forces the UID/GID even if the lower range value is less
than 1024.
-c, --cluster          The cluster scope for this command
```

4.6 Configuring Active Directory with NT4 and PDC

NT4/Samba PDC is a domain controller for Microsoft in Windows NT and Windows 2000. It is not supported by Microsoft. To support the controller, the open source and Samba community developed Samba PDC. It works similarly to AD.

NT4 domain users can act as SONAS users. Samba can serve as an NT4 server when configured as a primary domain controller (PDC).

Before SONAS can be configured with Samba PDC Server, you must verify the validity of PDC Server. For this purpose, the following checks must be done:

1. Validate that the Management node can create SSH connections to the Interface nodes.
2. Check whether each of the Interface nodes can ping the NT4 server.
3. Obtain the necessary parameters that are related to your authentication server. These parameters are required for the `cfgnt4` command.

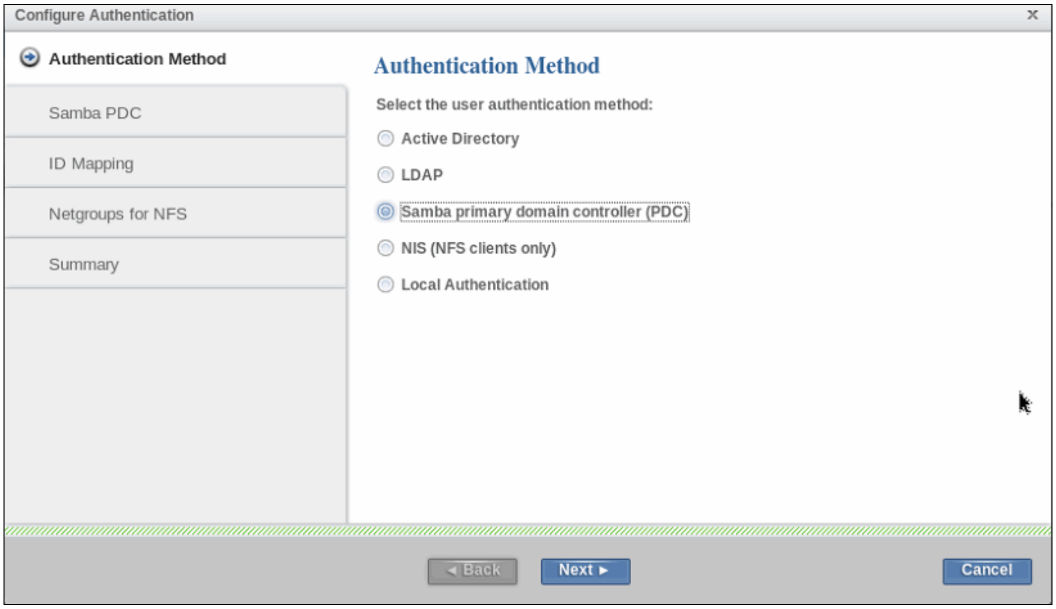
4.6.1 Configuring SONAS with Samba PDC/NT4 by using the GUI

To configure SONAS with Samba PDC/NT4 By using the GUI, complete the following steps:

1. Complete the steps in “Starting the Directory Services configuration in the GUI” on page 242 to get to the Authentication window.

   After you click **Authentication,** a window opens and asks for the type of authentication that you want to configure.

2. Select **Samba primary domain controller (PDC)** from the list. See Figure 4-20.



Authentication Method

Select the user authentication method:

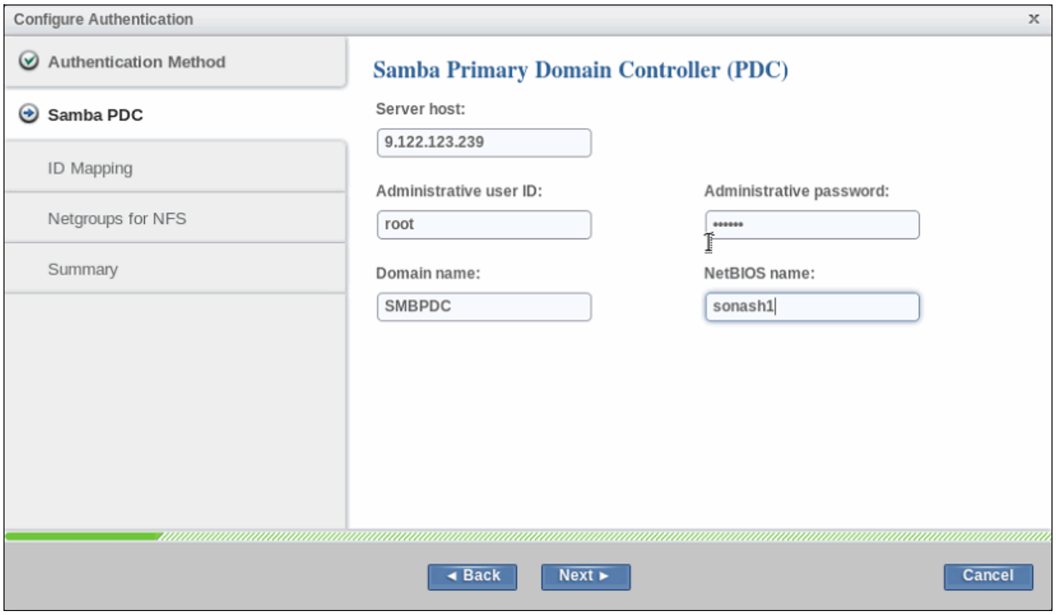
○ Active Directory
○ LDAP
● Samba primary domain controller (PDC)
○ NIS (NFS clients only)
○ Local Authentication

Figure 4-20 Choose Samba PDC/NT4 for the configuration

   You can also configure Extended NIS. This option is explained in 4.8.4, “Samba PDC/NT4 with NIS” on page 284.

3. Click **Next** to proceed.

   A window opens, as shown in Figure 4-21 on page 257.



Samba Primary Domain Controller (PDC)

Server host: 9.122.123.239

Administrative user ID: root

Administrative password: ••••••

Domain name: SMBPDC

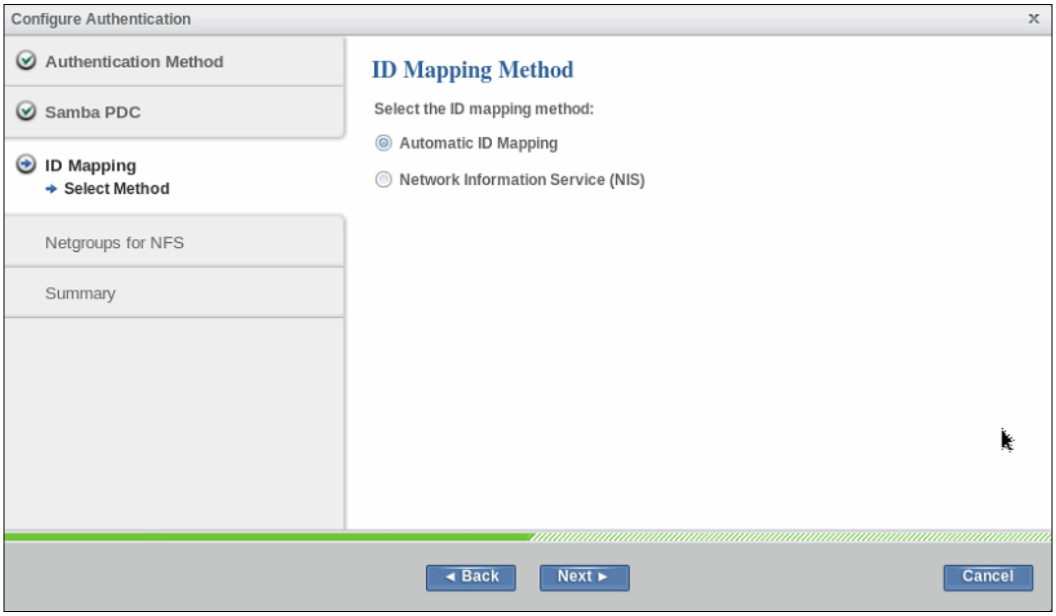
NetBIOS name: sonash1

Figure 4-21 Samba PDC/NT4 details that are required for configuration

4. Complete the Server name, Administrative user credentials, NT4 Domain name, and NetBios name and then click **Finish** to start the configuration.

   In this example, the NT4 server is `ldap1.virtual1.com`, the domain name is SMBPDC, the NetBios name is ldap1, and the root user credentials are provided.

   Click **Next** and select the ID mapping method to be used, as shown in Figure 4-22. In this example, automatic ID mapping is selected.



ID Mapping Method

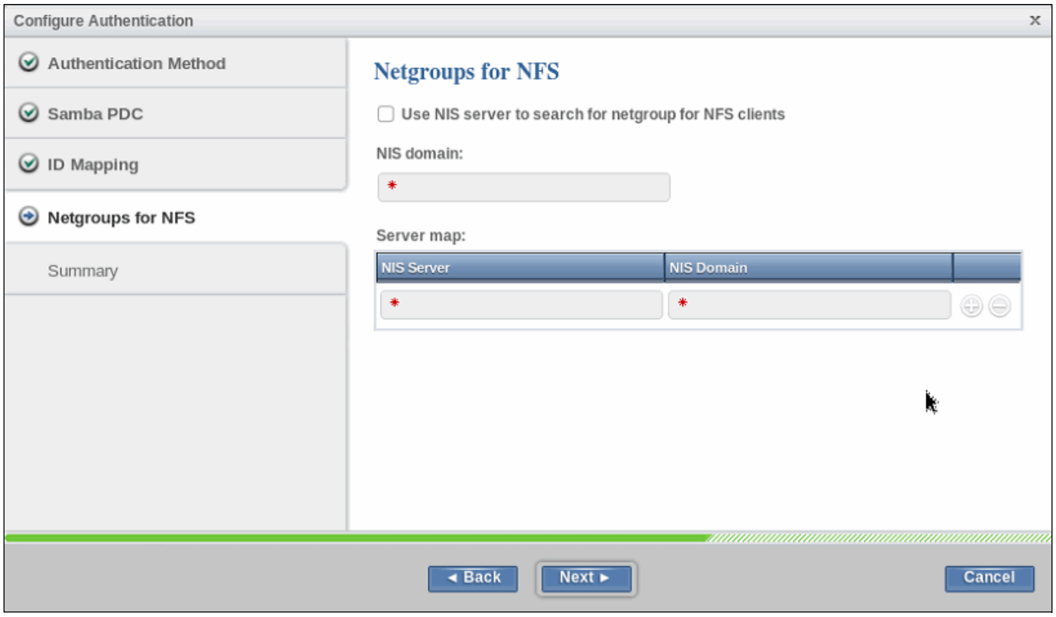
Select the ID mapping method:

● Automatic ID Mapping

○ Network Information Service (NIS)

Figure 4-22 Select ID mapping method

5. Click **Next** and enter the NIS server details if you want to use netgroups with NFS (see Figure 4-23).



☐ Use NIS server to search for netgroup for NFS clients

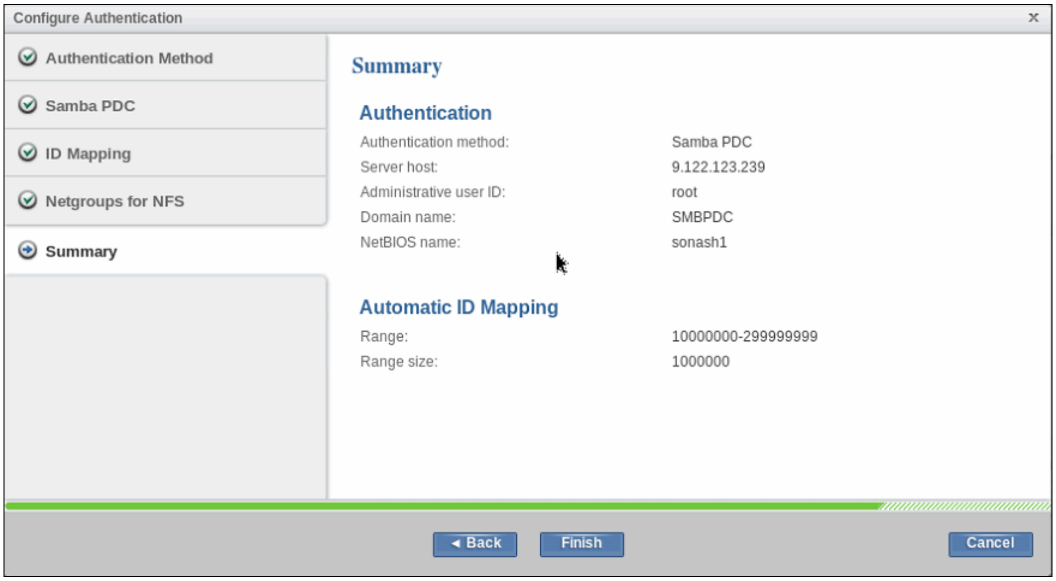
NIS domain: *

Server map:

NIS Server	NIS Domain
*	*

Figure 4-23 NIS server details for netgroups for NFS

6. Click **Next** and the next window shows the summary of the Samba PDC configuration option that is selected (see Figure 4-24).



Authentication

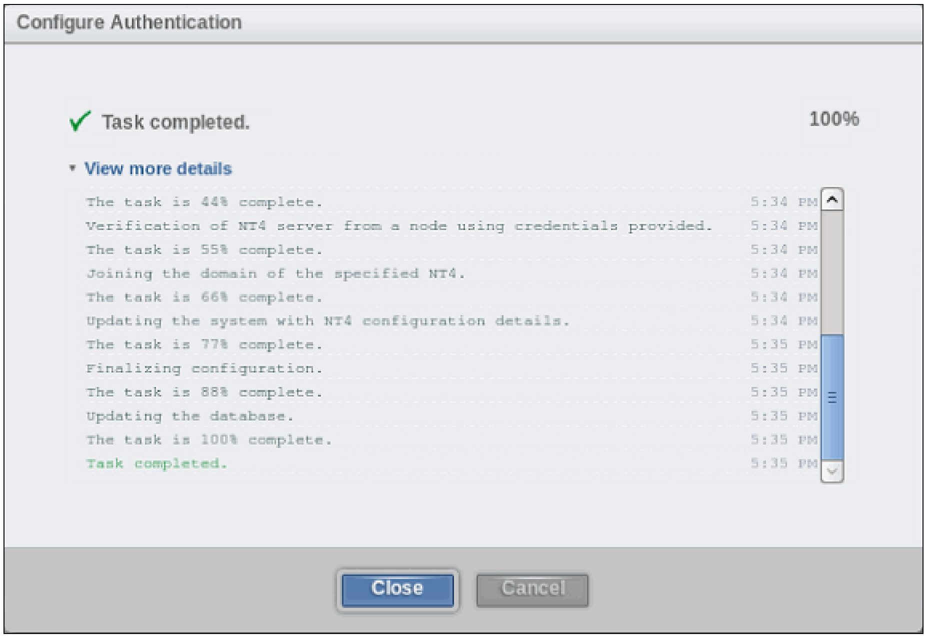
Authentication method: Samba PDC
Server host: 9.122.123.239
Administrative user ID: root
Domain name: SMBPDC
NetBIOS name: sonash1

Automatic ID Mapping

Range: 10000000-299999999
Range size: 1000000

Figure 4-24 Configuration summary

When you click **Finish,** the configuration for NT4 starts. You see a window that shows the progress of the command that runs to configure the authentication method that you chose. If the configuration is successful, you see the window that is shown in Figure 4-25 on page 259. If the configuration fails, you see an error. Correct the error and try again.



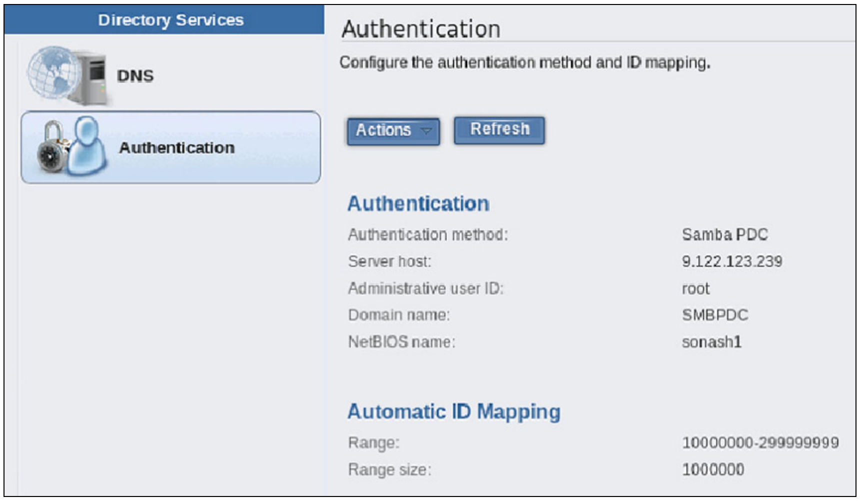
✓ Task completed. 100%

▾ View more details

```
The task is 44% complete.                                              5:34 PM
Verification of NT4 server from a node using credentials provided.     5:34 PM
The task is 55% complete.                                              5:34 PM
Joining the domain of the specified NT4.                               5:34 PM
The task is 66% complete.                                              5:34 PM
Updating the system with NT4 configuration details.                    5:34 PM
The task is 77% complete.                                              5:35 PM
Finalizing configuration.                                              5:35 PM
The task is 88% complete.                                              5:35 PM
Updating the database.                                                 5:35 PM
The task is 100% complete.                                             5:35 PM
Task completed.                                                        5:35 PM
```

Figure 4-25 NT4 configuration is successful

7. Click **Close** to return to the main authentication page, where you see the authentication that has been configured (see Figure 4-26).



Authentication

Configure the authentication method and ID mapping.

Authentication

Authentication method:	Samba PDC
Server host:	9.122.123.239
Administrative user ID:	root
Domain name:	SMBPDC
NetBIOS name:	sonash1

Automatic ID Mapping

Range:	10000000-299999999
Range size:	1000000

Figure 4-26 Samba PDC/NT4 configured

4.6.2 Configuring SONAS with Samba PDC/NT4 by using the CLI

The CLI command for configuring Samba PDC/NT4 on SONAS is **`cfgnt4`**. To configure SONAS with Samba PDC/NT4, complete the following steps:

1. Run **`lsauth`** to verify that no authentication method is configured. See Example 4-10.

Example 4-10 The lsauth CLI command shows that no authentication method is already configured on SONAS

```
# lsauth -c st001.virtual1.com
EFSSG0571I Cluster st001.virtual1.com is not configured with any type of authentication
server(ldap/ldap_krb/nt4/ad).
```

2. Run **`cfgnt4`** to configure SONAS with Samba PDC/NT4. See Example 4-11. In this example, the NT4 server `ldap1.virtual1.com`, NetBios name ldap1, and Domain name SMBPDC are used. For help with the **`cfgnt4`** command, see Example 4-13.

Example 4-11 The cfgnt4 command to configure SONAS with NT4

```
$cfgnt4 -s 9.122.123.239 --nt4NetbiosName sonash1 -d SMBPDC -u root -p test01
(1/9) Fetching the list of cluster Nodes.
(2/9) Check if cfgcluster has done the basic configuration successfully.
(3/9) Check whether Interface nodes are reachable from management node.
(4/9) Check whether NT4 server is reachable from cluster nodes.
(5/9) Verification of NT4 server from a node using credentials provided.
(6/9) Joining the domain of the specified NT4.
(7/9) Updating the system with NT4 configuration details.
(8/9) Finalizing configuration.
(9/9) Updating the database.
EFSSG1000I The command completed successfully.
```

3. Run **`lsauth`** to verify that cluster is configured with NT4. See Example 4-12.

Example 4-12 The lsauth command shows that NT4 is configured successfully

```
$lsauth
AUTH_TYPE = nt4
idMappingMethod = auto
domain = SMBPDC
nt4NetbiosName = sonash1
clusterName = st002.virtual1.com
nt4Host = 9.122.123.239
userName = root
EFSSG1000I The command completed successfully.
```

See Example 4-13 for help with the **`cfgnt4`** command. The command accepts the NT4 server name, NetBios name, Domain name, administrative user, and password along with the cluster name or cluster ID.

Example 4-13 Command help for cfgnt4

```
$cfgnt4 --help
usage: cfgnt4  -s <nt4Host> --nt4NetbiosName <nt4NetbiosName> -d <nt4Domain> -u <nt4AdminUser>
[-p <nt4AdminPw>] [-c < clusterID | clusterName >]
Configures the Microsoft Windows NT 4.0 (NT4) server on all the nodes present in the cluster
using the input values.

Parameter           Description
-s, --nt4Host       Specifies the NT4 server name.
```

```
    --nt4NetbiosName Specifies the NT4 server NetBIOS name.
-d, --nt4Domain      Specifies the NT4 server workgroup value.
-u, --nt4AdminUser   Specifies the NT4 server administrative user name.
-p, --nt4AdminPw     Specifies the NT4 server administrative user password.
-c, --cluster        The cluster scope for this command
```

4.6.3 Deterministic ID mapping support in SONAS V1.5.1

SONAS allows different approaches to SID to UID and GID mapping, which depend on the authentication method that is used. Common ID mapping methods that are used are NIS, SFU, and LDAP, which are hosted on an external server, and autorid and auto increment, which SONAS manages internally. Before SONAS V1.5.1, advanced functions such as asynchronous replication and IBM Active Cloud Engine® (ACE) are supported only when you have an external ID mapping mechanism, such as SFU and LDAP. Also, with internal ID mapping mechanisms such as autorid and auto increment, it is not possible to use advanced functions such as asynchronous replication because there is no way to maintain the same SID to UID mapping to better source and target clusters.

With the deterministic ID mapping support, you can set up two or more storage SONAS clusters with an AD replication relationship without requiring an external ID mapping source, such as SFU or NIS.

Deterministic ID mapping introduces roles that can be assigned to clusters while you configure authentication with AD. SONAS clusters in a customer environment can be designated as *master* and *subordinate*. The master idmap role, which can be set by using the **--idMapRole=master** parameter, is similar in behavior to the **cfgad** command before SONAS V1.5.1. You can then set up a secondary SONAS cluster with the subordinate role by using the **--idMapRole=subordinate** option with the **cfgad** command. In the subordinate role, the SONAS cluster is joined to the AD domain, but NAS services are not started and the autorid module is read only. If you try another command, such as **lsauth**, **chkauth**, or **chservices**, or try accessing the exports before your run **cfgidmap import**, there is no effect on ID maps.

The **cfgidmap** command, which is introduced in SONAS V1.5.1, should be used by the SONAS administrator to export the ID maps from the master SONAS cluster and then import them to the subordinate SONAS cluster. The exported ID maps from the master SONAS cluster are in XML format and have the information of all domains, including AD domains, well-known domains, built-in host domains, and their respective ID ranges. These exported ID maps can then be imported by using the **cfgidmap** command on the subordinate cluster. This command starts NAS services. After this step, both the SONAS clusters have identical ID maps.

Then, whenever more AD domains are added to the master SONAS cluster, the SONAS administrator must make sure that the new ID map configuration is exported from the master cluster and imported to the subordinate cluster. Now, importing the configuration does not restart NAS services.

Apart from supporting the export and import of ID maps, there are two other important changes that help maintain a consistent ID map across clusters that use the SONAS internal ID maps:

1. The new algorithm of the autorid module SIDs with RIDs larger than the **rangesize** value can be mapped because it supports multiple ranges for domains.
2. UIDs and GIDs for well-known domains, built-in domains, the SONAS system name (host) domain, and AD domains from ranges that are defined by the **cfgad** command are provided.

> **Note:** If there are SONAS systems earlier than Version 1.5.1. in async replication or remote cache relationship with a SONAS V1.5.1 system, they do not have the same ID for built-in domains.

Two new commands are introduced in SONAS V1.5.1 to support the use of deterministic internal ID mapping:

- **`cfgidmap`**
- **`lsidmap`**

The cfgidmap command

The **`cfgidmap`** command is used to export an ID mapping configuration from the master cluster to the subordinate cluster. It is also used to change the role of the SONAS cluster and compare the ID map configuration on the subordinate cluster. See the man page or the command help in the IBM Knowledge Center information for a detailed explanation:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/manpages/cfgidmap.html

Using the cfgidmap command to export the ID map

Example 4-14 shows the help for exporting the ID map by running the **`cfgidmap`** command.

Example 4-14 Help for exporting the ID map by running the cfgidmap command

```
$ cfgidmap --help
usage: cfgidmap operation  [--idMapRole <idMapRole>] [--fileName <fileName>]
[--force] [-c < clusterID | clusterName >]
Configure the cluster with ID mapping services

Parameter       Description
operation       The operation to be performed [import, export, compare,
changeRole]
    --idMapRole The ID map role of the cluster (operation: changeRole)
    --fileName  The file name for export or import or compare ID maps operation
(operation: export/import/compare)
    --force     Do not prompt for manual confirmation (operation:
changeRole/import)
-c, --cluster   The cluster scope for this command
```

Example 4-15 shows an example of the **`cfgidmap`** export operation.

Example 4-15 Run the cfgidmap command to export an ID map

```
$ cfgidmap export
(1/4) Fetching the list of cluster Nodes.
(2/4) Checking if basic configurations are present to execute the command
(3/4) Reading the internal ID Map database
(4/4) Writing ID Map data into XML file
Created ID map file /ftdc/idmap.xml on current cluster's active management node.
Copy this file on subordinate cluster's active management node at/ftdc/files/
EFSSG1000I The command completed successfully.
```

As shown in Figure 4-27, you can also use the GUI to export and import the ID map configuration. When you export, the exported XML file is stored in the local drive on which the SONAS GUI is running. When you import, the XML can be imported from the local drives.

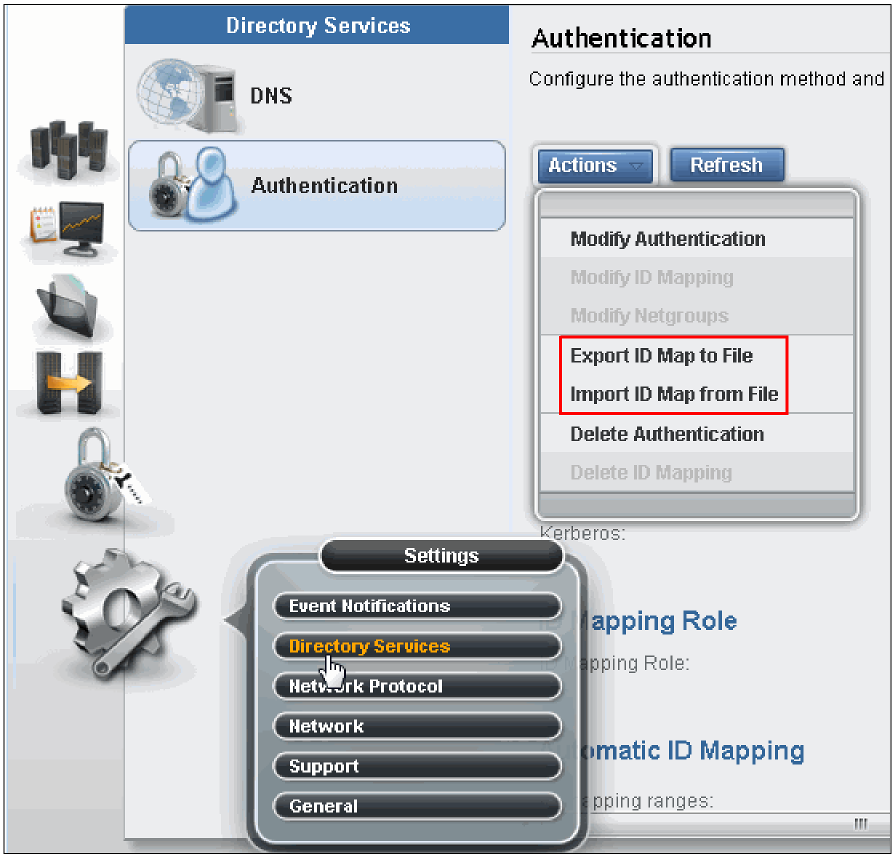


Figure 4-27 Use the GUI to import and export the ID map configuration

After the ID map export file is created, it must be copied to the SONAS cluster that is designated as the subordinate. This XML file has the information of all domains, including AD domains, well-known domains, built-in domains, host domains, and their respective IDs. If SFU ranges are configured, the XML file contains SFU domains and their ranges too.

Do not modify: The file that is created by the export operation is a system-generated file. Do not modify it manually. If you modify it, the import and comparison function might fail.

Download this file from the master cluster by running the **`scpuser`** command and then upload it to the subordinate cluster by running the **`scpuser`** command at the location `/ftdc/files`. After the file is uploaded to the subordinate cluster, you can either import the ID map configuration or run an ID map configuration comparison.

Note: Unless a domain user accesses the exports or a CLI user run **`chkauth -i u`** for a domain user, the domain ID map configuration for that domain is not created in the SONAS internal database.

Using the cfgidmap command to perform an import

You can now set a replication target cluster that runs the **cfgad** command similarly to how it runs on the source cluster, but this time with the **--idMapRole=subordinate** parameter. As described in “Using the cfgidmap command to export the ID map” on page 262, after the **cfgad** configuration is done, the cluster is joined to the AD domain, but NAS services are not started, and the shares are available for use only after the ID map configuration from the master cluster is imported. The NAS services (and winbind) are started and the share is available. Now, running the import command does not restart the NAS services.

Example 4-16 shows an example of running the **cfgidmap** command to import an ID map.

Example 4-16 cfgidmap import operation

```
$ cfgidmap import
(1/6) Fetching the list of cluster Nodes.
(2/6) Checking if basic configurations are present to execute the command
(3/6) Parsing the external XML file /ftdc/files/idmap.xml
(4/6) Reading the internal ID Map database
(5/6) Comparing external and internal ID Maps
(6/6) Updating internal ID Map database
EFSSG1000I The command completed successfully.
```

After the import operation, both the master and subordinate cluster have identical ID maps, so the UID and GID always are identical across sites.

If more domains are added to master cluster, the administrator must redo the **cfgidmap** export and import operation to synchronize the ID map configuration of the two clusters.

Note: There is no way to sync automatically the changes in the master cluster ID map configuration to the subordinate cluster.

Note: If SFU ranges are imported, the import operation restarts the winbind service every time. Also, only the Autorid and SFU ID mapping config is imported by the import operation. It does not import the NIS and LDAP ID mappings.

Using the cfgidmap command to compare ID maps

The **cfgidmap compare** operation that is shown in Example 4-17 is used to compare ID map configurations between the subordinate cluster and the ID map xml file, `/ftdc/files/idmap.xml`, which is exported and copied from the master cluster. You can export and import the ID map configuration file, as described in “Using the cfgidmap command to export the ID map” on page 262 and “Using the cfgidmap command to perform an import” on page 264 by using either the **cfgidmap** command or the SONAS GUI. The **cfgidmap compare** operation can be run only on the subordinate cluster.

Example 4-17 Example cfgidmap compare operation

```
$
cfgidmap compare
(1/5) Fetching the list of cluster Nodes.
(2/5) Checking if basic configurations are present to execute the command
(3/5) Parsing the external XML file /ftdc/files/idmap.xml
(4/5) Reading the internal ID Map database
(5/5) Comparing external and internal ID Maps
ID Maps on current and other system are identical
EFSSG1000I The command completed successfully.
```

The cfgidmap changeRole option

Run the **`cfgidmap changeRole`** command to change the ID map role of a cluster from master to subordinate and vice versa. Example 4-18 shows an example command where the cluster role is changed to master.

Example 4-18 Example cfgidmap change role operation

```
$ cfgidmap changeRole -idMapRole master
You have chosen to change ID map role from subordinate to master.Before
proceeding, it is required that you have imported all the id maps from the
previous master system. Have you ? [yes/no]: yes
(1/3) Fetching the list of cluster Nodes.
(2/3) Checking if basic configurations are present to execute the command
(3/3) Updating ID Map Role to master
You have changed ID mapping role of the system to master. Users and groups from
new domains will be allowed to access the system henceforth.
Execute lsidmap command to display the domains which are currently allowed to
access the system.
EFSSG1000I The command completed successfully.
```

A typical usage is when you might want to change the role of a subordinate cluster to master when the two clusters are set up for asynchronous replication and the master cluster for some reason becomes inaccessible.

> **Note:** It is the administrator’s responsibility to maintain the proper ID map roles on all the systems in the environment. If a system’s idmap role is *master* and it is changed to *subordinate*, then it is the administrator's responsibility to appoint a new master. Similarly, if a *subordinate* is changed to a *master*, it is administrator's responsibility to change the old master system to subordinate.

The lsidmap command

As shown in Example 4-19, the **`lsidmap`** command can be used to show the current ID map configuration for all the domains that are registered with the cluster. The ID map configuration also can be seen from the GUI, as shown in Figure 4-14 on page 248.

Example 4-19 Example lsidmap command to show the current ID map configuration

```
$lsidmap
ID Map Role (subordinate)
Domain       SID                                     Range             Source
ALLOC        ALLOC                                   10000000-10999999 auto
BUILTIN      S-1-5-32                                11000000-11999999 auto
STORAGE4TEST S-1-5-21-357287817-1492011099-913895456 12000000-12999999 auto
EFSSG1000I The command completed successfully.
```

Synchronizing the ID map between an existing SONAS cluster and a freshly installed SONAS cluster

If you have an existing SONAS cluster that has an ACL set on built-in groups, that is Administrators and Users, the GIDs of these groups are not copied to the new cluster. Therefore, these groups do not have the same IDs on both clusters.

Delete ACLs that use built-in groups before you set up advanced functions such as asynchronous replication or ACE and reapply them after the upgrade is done.

After this process, you can export the ID map configuration at the master cluster, configure the target replication cluster as “subordinate” by running the **`cfgad`** command, and import the ID map configuration.

Preferred practices for using deterministic ID mapping

Consider the following preferred practices for deterministic ID mapping:

- If you have a single SONAS cluster in the environment, configure it with the *master* ID map role.
- The authentication method should be the same for all SONAS clusters under consideration.
- For multiple SONAS clusters in the environment, designate the production cluster where the users typically access the storage exports as the master.
- In an asymmetric setup where one system has access to all domains, and another might have access to only a subset of domain, designate the system that has access to all domains as the master. In this case, before you can start using a domain on subordinate systems, a user from that domain must access the master system. If users do not have access to the master system, the domain can be registered by running the **`chkauth -i`** command. When this task is done, export the ID map configuration from the master and import it into the subordinate cluster.

4.7 Configuring SONAS with LDAP

The Lightweight Directory Access Protocol (LDAP) is an application protocol for reading and editing directories over an IP network. A directory in this sense is an organized set of records: for example, a telephone directory is an alphabetical list of persons and organizations with an address and phone number in each record.

LDAP is widely used to store USER and GROUP information for UNIX based environments. USER and GROUP information is stored in standard POSIX schema in LDAP. You must extend the schema to support CIFS and the Samba protocol, which requires adding more attributes to POSIX user objects (mainly to store the SID, Windows password hash, and domain information).

If SONAS is configured with LDAP authentication, all user and group information is fetched from LDAP by `/etc/nsswitch.conf` (NSS component configuration of the operating system) by using Pluggable Authentication Module (PAM) LDAP or Name Service Switch (NSS) LDAP.

SONAS with LDAP is good when mixed-mode clients such as Windows and UNIX client systems are used. LDAP with Kerberos is supported with CIFS and Samba protocol access. It is not supported for secured NFS, FTP, HTTP, and SCP protocol access.

This section describes configuring SONAS with plain LDAP and with LDAP along with Kerberos.

Before SONAS can be configured with an external LDAP server, you must verify the validity of the LDAP server. For this purpose, perform the following checks:

1. Validate that the Management node can create SSH connections to the Interface nodes.
2. Check whether at least one of the LDAP servers can be pinged. Verify that the LDAP server is up and responds by sending an LDAP query with filter “objectClass=*”.
3. Verify that the LDAP server is configured with at least a few users by sending an LDAP query with filter “ou=People”.

4. Before your configure SONAS with LDAP, the LDAP user information must be updated with unique Samba attributes in addition to the attributes that are stored for a normal LDAP user. Ensure that these required Samba attributes are present in the LDAP user entries. For more information, see “Updating LDAP user information with unique Samba attributes” on page 269.
5. The validation must be run on each of the Interface nodes. If SSL/TLS is being configured, the certificate is also checked at this stage. Also, if Kerberos is enabled, the keytab file is checked for entries of the type:
   - `cifs/<cluster-name>.<domain>`
   - `cifs/<hostname>`

Attention: The verification procedure does not change any configuration for an existing authentication.

For LDAP-KRB-CIFS configuration, plain LDAP is still accepted. If the requirement is to allow access only for Kerberos users, set some invalid passwords in the LDAP database for the users.

4.7.1 ID mapping methods that are available for SONAS with LDAP

With LDAP, SONAS does not need any external ID mapping method. The LDAP server itself can hold ID mappings. SONAS can fetch this ID mapping information from the LDAP server.

4.7.2 When to choose SONAS with LDAP

If SONAS must be used in the IT infrastructure where LDAP is used as a directory service and all user information is maintained in same LDAP server, then LDAP configuration for authentication is preferred.

If SONAS must be used with both CIFS and NFS data access protocols, LDAP can also be used because it stores ID mapping.

4.7.3 Limitations for Storwize V7000 Unified with LDAP

Storwize V7000 Unified with LDAP has the following limitations:

- Communication with the LDAP server is not secure.
- The user's SID is composed of the RID and the SID of the domain account that is created in the external LDAP server after the IBM SONAS system LDAP-based authentication configuration is done. For example, an IBM SONAS system that has the NetBios name st001 has an entry in the following format:

  ```
  dn: sambaDomainName=ST001,dc=sonasldap,dc=com
  sambaDomainName: ST001
  sambaSID: S-1-5-21-3315143710-1287377127-4281028267
  ```

  The user SID format is domain account SID-RID, so in this example all IBM SONAS system users must have an SID in the following format:

  ```
  S-1-5-21-3315143710-1287377127-4281028267-RID.
  ```

4.7.4 Configuring SONAS with multiple LDAP servers

For multiple LDAP servers, you must add the server certificates of all LDAP servers to a single certificate file. Certificates are created with a specific LDAP server name and this server name must match with the LDAP host names (otherwise, the configuration might fail). These multiple LDAP servers must be specified. Use a comma-separated list when you run the **cfgldap** command.

4.7.5 Prerequisites for configuring the IBM SONAS system with LDAP

To support the CIFS protocol, the schema must be extended with more attributes to the POSIX user object, mainly to store the SID, Windows password hash, and domain information.

This information is essential for LDAP to work correctly for CIFS access. The section “Setting up external LDAP server prerequisites” describes the prerequisites in detail.

Setting up external LDAP server prerequisites

Before you configure the IBM SONAS environment for external server LDAP integration, several external LDAP server prerequisites must be met:

- The external LDAP server must already be configured.
- Obtain in advance the administrative information for the external LDAP authentication server, such as the administration account, password, SSL certificate, and Kerberos keytab file.
- Ensure that the IBM SONAS nodes have proper connectivity to the external LDAP server, and vice versa.
- On each of the IBM SONAS nodes, you must synchronize the time with the external LDAP authentication server. Authentication does not work if the times on the IBM SONAS nodes and the external LDAP authentication server are not synchronized.
- Optionally, enable SSL or TLS encryption on the external LDAP server. Details about configuring SSL or TLS encryption on the server can be obtained from the *OpenLDAP Administrator's Guide* found at the following website:

  http://www.openldap.org/doc/admin24/
- Before you configure SONAS with LDAP, the LDAP user information must be updated with unique Samba attributes in addition to the attributes that are stored for a normal LDAP user. Ensure that these required Samba attributes are present in the LDAP user entries. For more information, see “Updating LDAP user information with unique Samba attributes” on page 269.
- A special administrative user for the IBM SONAS system must be created on the external LDAP server. This user might not have permission to create users; however, this user must at a minimum have permission to query users and groups that are created in the external LDAP server.
- The user information is used as Bind DN by Samba as it makes LDAP queries, as shown in the following example LDAP query:

  ```
  ldapsearch -x -D "cn=Manager,dc=example,dc=com" -w <password> "<query-filter>"
  ```

  The LDIF for this user is similar to the following example:

  ```
  dn: cn=Manager,dc=example,dc=com
  objectclass: organizationalRole
  cn: Manager
  ```

Updating LDAP user information with unique Samba attributes

To use Samba accounts, it is necessary to update LDAP user information with unique Samba attributes. The sample LDIF file in Example 4-20 shows the minimum required Samba attributes.

Example 4-20 Sample LDIF file

```
dn: cn=cifsuser,ou=People,dc=ibm,dc=com
changetype: modify
add : objectClass
objectClass: sambaSamAccount
-
add: sambaSID
sambaSID: (S-1-0-41200)
-
add:sambaPasswordHistory
sambaPasswordHistory: 00000000000000000000000000000000000000000000000000000000
-
add:sambaNTPassword
sambaNTPassword: (valid samba password hash )
-
add:sambaPwdLastSet
sambaPwdLastSet: 1263386096
-
add:SambaAcctFlags
sambaAcctFlags: [U          ]
```

Attributes must be separated with a dash as the first and only character on a separate line.

To create these attributes, complete the following steps:

1. Create the values for the **`sambaNTPassword`**, **`sambaPwdLastSet`**, and **`SambaAcctFlags`** attributes, which must be generated from a Perl module.

   An example module is provided at the following website:

   `http://search.cpan.org/~bjkuit/Crypt-SmbHash-0.12/SmbHash.pm`

2. Use a Perl script to generate the LM and NT password hashes. Example 4-21 shows an example script.

   Example 4-21 Perl script to generate LM and NT password hashes

   ```
   # cat /tmp/Crypt-SmbHash-0.12/gen_hash.pl
   #!/usr/local/bin/perl
   use Crypt::SmbHash;
   $username = $ARGV[0];
   $password = $ARGV[1];
   if ( !$password ) {
   print "Not enough arguments\n";
   print "Usage: $0 username password\n";
   exit 1;
   }
   $uid = (getpwnam($username))[2];
   my ($login,undef,$uid) = getpwnam($ARGV[0]);
   ntlmgen $password, $lm, $nt;
   printf "%s:%d:%s:%s:[%-11s]:LCT-%08X\n", $login, $uid, $lm, $nt, "U", time;
   ```

3. Generate the password hashes for any user. Example 4-22 shows an example script for the test01 user.

   Example 4-22 Example Perl script to generate password hashes

   ```
   # perl gen_hash.pl cifsuser test01
   :0:47F9DBCCD37D6B40AAD3B435B51404EE:82E6D500C194BA5B9716495691FB7DD6:[U
   ]:LCT-4C18B9FC
   ```

   The line in Example 4-22 shows the login name, UID, LM hash, NT hash, flags, and time. Each field is separated from the next by a colon. The login name and UID are omitted because the command was not run on the LDAP server.

4. After the password is generated, use the information to update the LDIF file in the format that is provided by Example 4-20 on page 269.
5. To generate the **sambaPwdLastSet** value, use the hexadecimal time value from step 3 after the dash character and convert it into decimal.

A valid Samba SID is required for a user to enable that user's access to an IBM SONAS share. To generate the samba SID, multiply the user's UID by 2 and add 1000. The user's SID should contain the Samba SID from the *sambaDomainName*, which is either generated or picked up from LDAP server, if it already exists.

The attributes that are shown in Example 4-23 for *sambaDomainName* LDIF entry are required.

Example 4-23 Attributes for sambaDomainName LDIF entry

```
dn: sambaDomainName=(IBM SONAS system name),dc=ibm,dc=com
sambaDomainName: (IBM SONAS system name)
sambaSID: S-1-5-21-1528920847-3529959213-2931869277
```

This entry can be created by the LDAP server administrator by writing and running a bash script similar to the one that is shown in Example 4-24.

Example 4-24 Sample bash script to generate an LDIF entry

```
sambaSID=
for num in 1 2 3 ;do
randNum=$(od -vAn -N4 -tu4 < /dev/urandom | sed -e 's/ //g')
if [ -z     "$sambaSID" ];then
sambaSID="S-1-5-21-$randNum"
else
sambaSID="${sambaSID}-$    {randNum}"
fi
done
echo $sambaSID
```

Next, use the Samba SID that is generated to create the LDIF file. The *sambaDomainName* must match the IBM SONAS system name.

Run the **cfgldap** command, which creates the *sambaDomainName* if it does not exist.

The *sambaSID* for every user should have the following format:

(samba SID for the domain)-(userID*2+1000).

For example:

```
S-1-5-21-1528920847-3529959213-2931869277-1102
```

To enable access to more than one IBM SONAS system, the domain SID prefix of all of the IBM SONAS systems must match. If you change the domain SID for an IBM SONAS system on the LDAP server, you must restart CTDB on that IBM SONAS system for the change to take effect.

To update the user's information, run the **`ldapmodify`** command, as shown in Example 4-25.

Example 4-25 Sample output for the ldapmodify command

```
# ldapmodify -h localhost -D cn=Manager,dc=ibm,dc=com -W -x -f
/tmp/samba_user.ldif
```

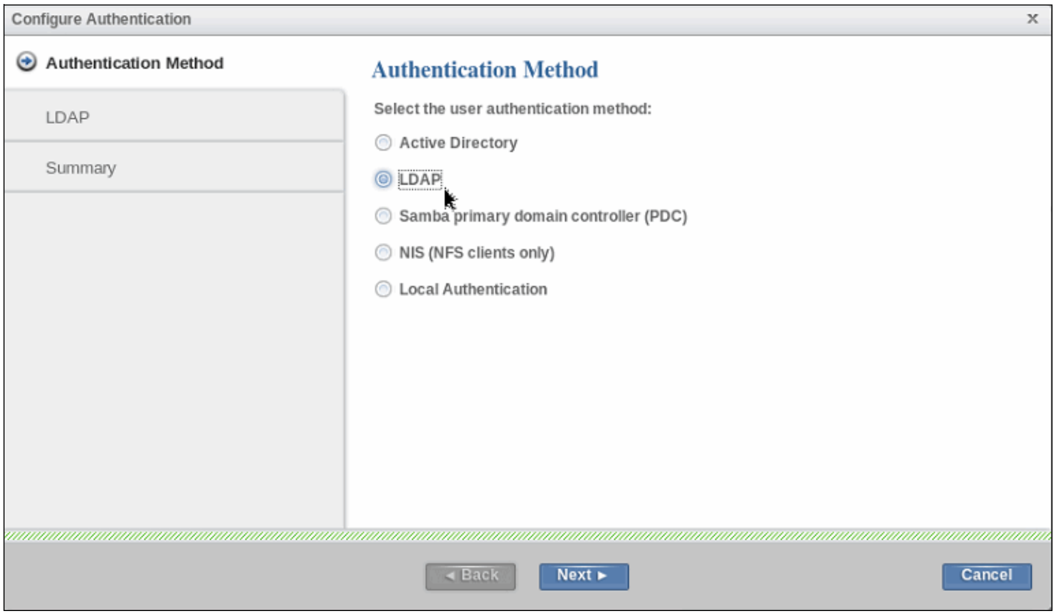
4.7.6 Configuring LDAP by using the GUI

To configure LDAP with the GUI, complete the following steps:

1. Complete the steps in “Starting the Directory Services configuration in the GUI” on page 242 to get to the Authentication page in the GUI.

   After you click **Authentication**, a window opens.

2. Select LDAP from the list, as shown in Figure 4-28.



Authentication Method

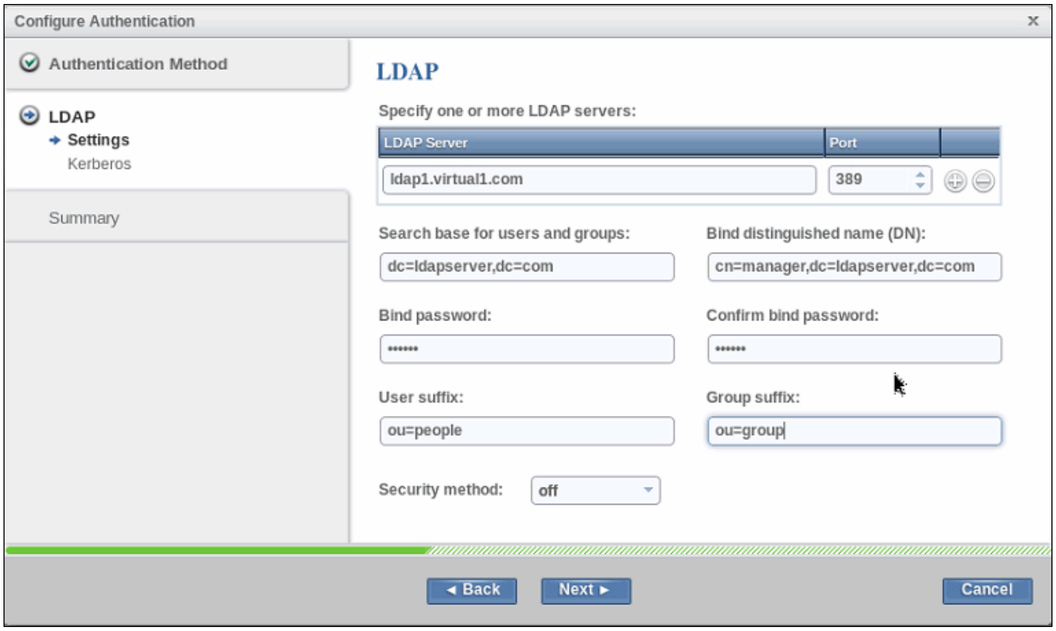
Select the user authentication method:

○ Active Directory
● LDAP
○ Samba primary domain controller (PDC)
○ NIS (NFS clients only)
○ Local Authentication

Figure 4-28 Choose LDAP for configuration

3. Click **Next** to proceed to the next step.

   A window (Figure 4-29) opens.



Configure Authentication

LDAP

Specify one or more LDAP servers:

LDAP Server	Port
ldap1.virtual1.com	389

Search base for users and groups: dc=ldapserver,dc=com

Bind distinguished name (DN): cn=manager,dc=ldapserver,dc=com

Bind password: ••••••

Confirm bind password: ••••••

User suffix: ou=people

Group suffix: ou=group

Security method: off

Figure 4-29 Required LDAP configuration details

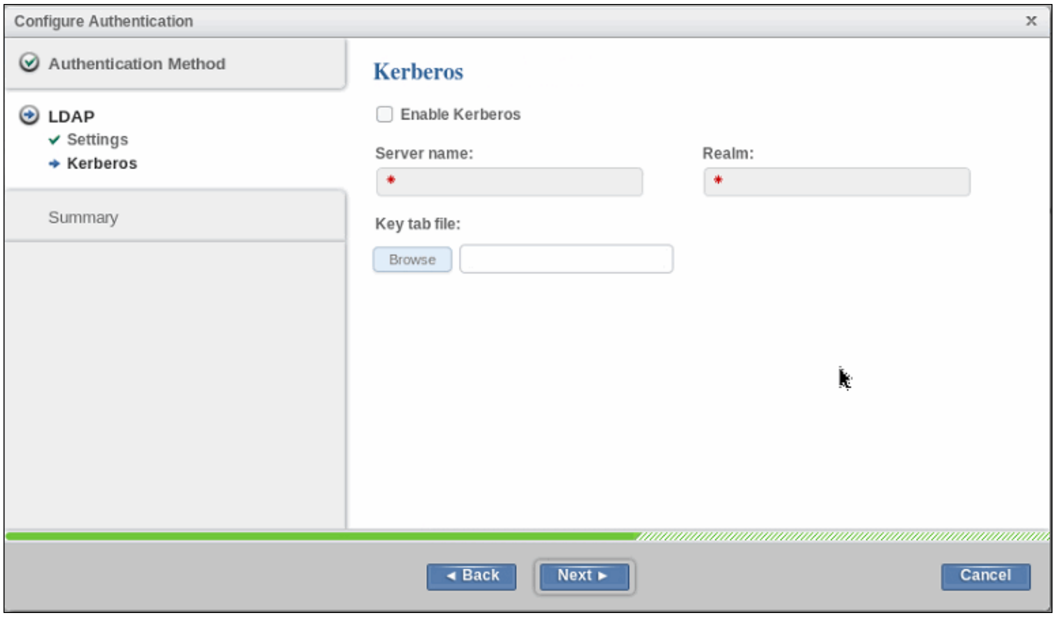
4. Complete the details, such as LDAP Server name, Search Base for users and groups, Bind Name, Bind Password, User and Group suffix, and the workgroup. Click **Finish** to start the configuration.

   In this example, the LDAP Server is `ldap1.virtual1.com`, the Search base is `dc=ldapserver,dc=com`, The Bind name is `cn=manager,dc=ldapserver,dc=com`. The User suffix is `ou=People`. The Group suffix is `ou=Group` and the workgroup is `virtual1`.

   If you want to enable Kerberos, check the **Enable Kerberos** box. Provide the Kerberos Server name, Realm, and the Key Tab file to proceed. Set the security method to off.

5. Click **Next**.

6. If you want to use Kerberos with LDAP, enter the Kerberos server details and the keytab file, as shown in Figure 4-30 on page 273.



Kerberos

☐ Enable Kerberos

Server name: *

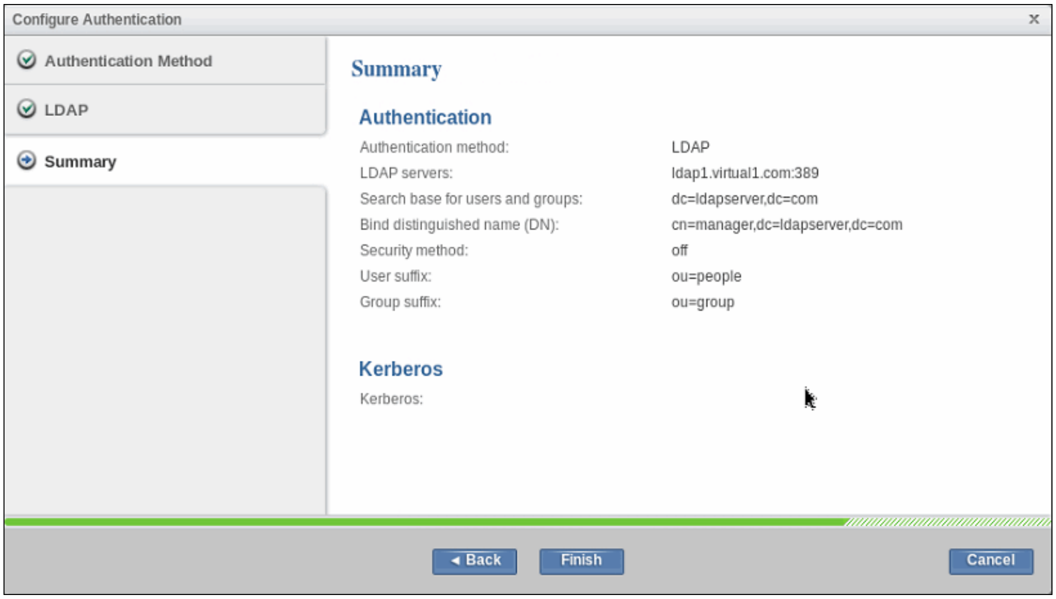
Realm: *

Key tab file:

Figure 4-30 Kerberos server details

7. Click **Next.**

   The next window summarizes the LDAP server configuration details, as shown in Figure 4-31.



Summary

Authentication

Authentication method:	LDAP
LDAP servers:	ldap1.virtual1.com:389
Search base for users and groups:	dc=ldapserver,dc=com
Bind distinguished name (DN):	cn=manager,dc=ldapserver,dc=com
Security method:	off
User suffix:	ou=people
Group suffix:	ou=group

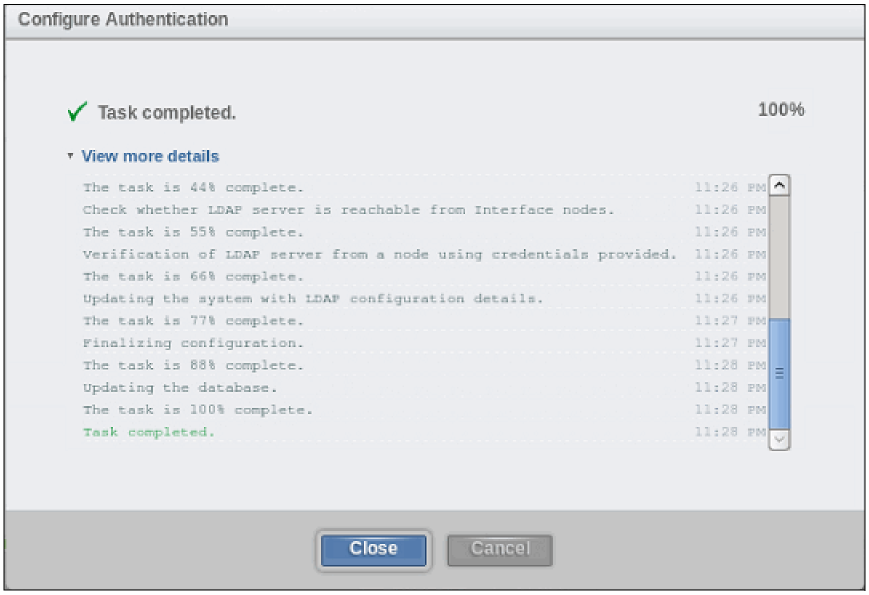
Kerberos

Kerberos:

Figure 4-31 LDAP server configuration summary

8. Click **Finish** to complete the configuration.

   When you click **Finish**, the configuration for LDAP starts. You see a window that shows the progress of the command that runs to configure the authentication method that you chose. If the configuration is successful, you see the window that is shown in Figure 4-32. If it fails, you see an error. In this case, correct the error and try again.



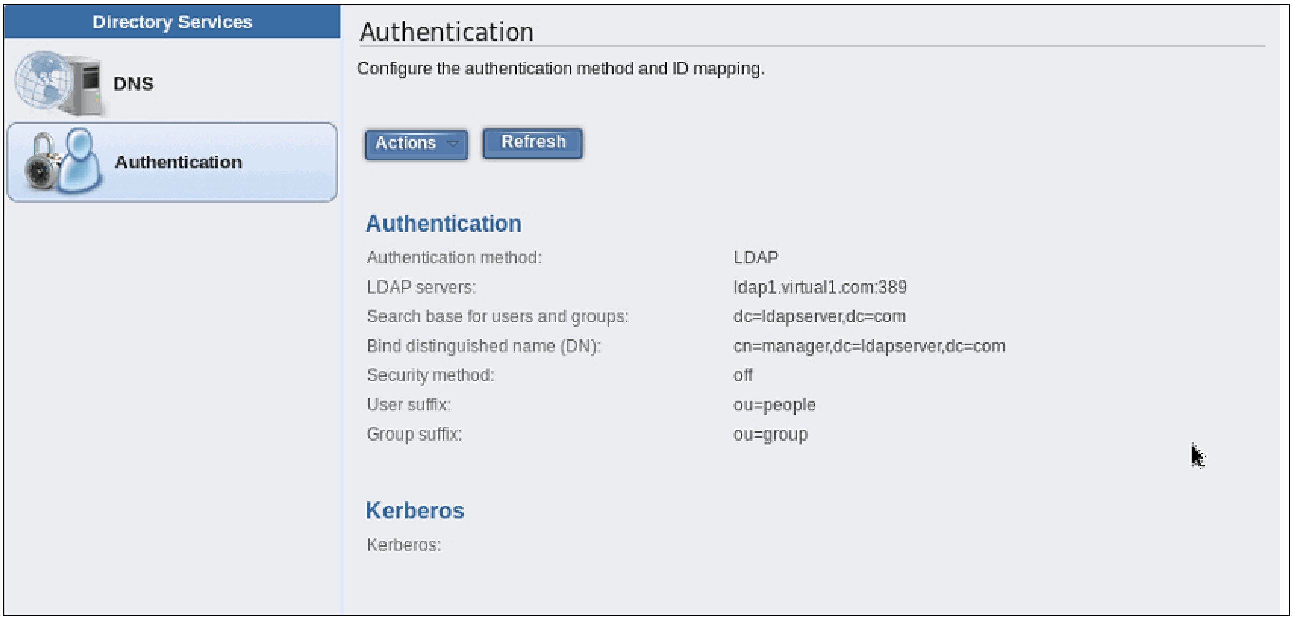
✓ Task completed. 100%

```
The task is 44% complete.                                            11:26 PM
Check whether LDAP server is reachable from Interface nodes.         11:26 PM
The task is 55% complete.                                            11:26 PM
Verification of LDAP server from a node using credentials provided. 11:26 PM
The task is 66% complete.                                            11:26 PM
Updating the system with LDAP configuration details.                 11:26 PM
The task is 77% complete.                                            11:27 PM
Finalizing configuration.                                            11:27 PM
The task is 88% complete.                                            11:28 PM
Updating the database.                                               11:28 PM
The task is 100% complete.                                           11:28 PM
Task completed.                                                      11:28 PM
```

Figure 4-32 LDAP configuration is successful

9. Click **Close** to return to the main authentication window.

   You see the authentication method that is configured (Figure 4-33).



Authentication

Configure the authentication method and ID mapping.

Authentication

Authentication method:	LDAP
LDAP servers:	ldap1.virtual1.com:389
Search base for users and groups:	dc=ldapserver,dc=com
Bind distinguished name (DN):	cn=manager,dc=ldapserver,dc=com
Security method:	off
User suffix:	ou=people
Group suffix:	ou=group

Kerberos

Kerberos:

Figure 4-33 LDAP is configured

4.7.7 Configuring SONAS with plain LDAP by using the CLI

The **cfgldap** command is used to configure LDAP with SONAS.

Before you run the command, you must obtain the required parameters for your authentication server. See Example 4-26 for the command help to understand the different parameter values that are required.

Example 4-26 Help for the cfgldap command

```
$cfgldap --help
usage: cfgldap  -s <ldapServers> -b <ldapBase> -D <ldapBindDn> [-p <ldapBindPw>] [-d <domain>]
[--krbKDC <krbKdc>] [--krbRealm <krbRealm>] [--krbKeytabFile <krbKeytabFile>] [--caCertFile
<caCertFile>] [--ldapUserSuffix <ldapUserSuffix>] [--ldapGroupSuffix <ldapGroupSuffix>]
[--sslMode <sslMode>] [-c < clusterID | clusterName >]
Configures an LDAP server or an LDAP with Kerberos server on all the nodes present in the
cluster with the input values.

Parameter           Description
-s, --ldapServers     Specifies the LDAP servers to be used.
-b, --ldapBase        Specifies the LDAP base to be used.
-D, --ldapBindDn      Specifies the LDAP base distinguished name (DN) to be used.
-p, --ldapBindPw      Specifies the LDAP bind password to be used.
-d, --domain          Specifies the domain name to be used.
    --krbKDC          Specifies the Kerberos server name to be used.
    --krbRealm        Indicates the realm for the Kerberos server to be used.
    --krbKeytabFile   Indicates the Kerberos keytab to be used.
    --caCertFile     Specifies the certificate file required when the security method is SSL or
TLS.
    --ldapUserSuffix  Specifies the LDAP user suffix to be used.
    --ldapGroupSuffix Specifies the LDAP group suffix to be used.
    --sslMode         Specifies whether the Secure Socket Layer (SSL) mode is to be used.
-c, --cluster         The cluster scope for this command
```

Configuring LDAP

To configure LDAP, complete the following steps:

1. Run the **lsauth** command to verify that no authentication method is configured on the cluster. See Example 4-27.

Example 4-27 The lsauth CLI command shows that no authentication method is configured

```
$lsauth -c st001.virtual1.com
EFSSG0571I Cluster st001.virtual1.com is not configured with any type of authentication server.
```

2. Run the **cfgldap** command to configure SONAS with LDAP (see Example 4-28). In this example, the LDAP server with name `ldap1.virtual1.com` is used. The Base values are `dc=ldapserver,dc=com` and the binding values are `cn=manager,dc=ldapserver,dc=com`.

Example 4-28 Example cfgldap command to configure SONAS with LDAP

```
$cfgldap -c st001.virtual1.com -d virtual1.com -b dc=ldapserver,dc=com -D
cn=manager,dc=ldapserver,dc=com -p secret -s ldap1.virtual1.com --sslMode off
(1/9) Fetching the list of cluster Nodes.
(2/9) Check if cfgcluster has done the basic configuration successfully.
(3/9) Check whether Interface nodes are reachable from management node.
(4/9) LDAP Configuration started.
```

```
(5/9) Check whether LDAP server is reachable from Interface nodes.
(6/9) Verification of LDAP server from a node using credentials provided.
(7/9) Updating the system with LDAP configuration details.
(8/9) Finalizing configuration.
(9/9) Updating the database.
EFSSG1000I The command completed successfully.
```

3. Run the **lsauth** command to verify that the cluster is configured with LDAP, as shown in Example 4-29.

Example 4-29 The lsauth command shows that LDAP is configured successfully

```
$lsauth -c st001.virtual1.com
AUTH_TYPE = ldap
ldapVersion = 3
clusterName = st001.virtual1.com
ldapUri = ldap://ldap1.virtual1.com
ldapBase = dc=ldapserver,dc=com
nscdMode = on
domainSID = S-1-5-21-2809281514-1709260598-260063778
ldapBindDn = cn=manager,dc=ldapserver,dc=com
krbMode = off
sslMode = off
EFSSG1000I The command completed successfully.
```

4.7.8 Configuring SONAS to use LDAP with Kerberos by using the CLI

SONAS can be configured to use LDAP with Kerberos (this process requires an existing Kerberos infrastructure). SONAS needs a keytab file, called `krb5.keytab`, to authenticate to the KDC.The keytab file must be uploaded first.

To configure LDAP with Kerberos, complete the following steps:

1. Run **lsauth** to verify that no authentication method is configured on the cluster, as shown in Example 4-30.

Example 4-30 The lsauth command shows that no authentication method is configured

```
$lsauth -c st001.virtual1.com
EFSSG0571I Cluster st001.virtual1.com is not configured with any type of authentication server.
```

2. Run **cfgldap** to configure SONAS with LDAP, as shown in Example 4-31.

   In this example, the LDAP server is `ldap1.virtual1.com`, the Search base is `dc=ldapserver,dc=com`, the Bind name is `cn=manager,dc=ldapserver,dc=com`, and the domain and workgroup is `virtual1`. The Kerberos-related values are Realm (`VIRTUAL1.COM`), Kerberos Server (`ldap1.virtual1.com`), and the Kerberos keytab file (`/etc/krb5.keytab`), which is copied from the LDAP server to the cluster at `/root/krb5.keytab`.

Example 4-31 Example cfgldap command to configure SONAS to use LDAP with Kerberos

```
$cfgldap -s ldap1.virtual1.com -b "dc=ldapserver,dc=com" -D
"cn=manager,dc=ldapserver,dc=com" -p "XXXX" -d virtual1 --krbKDC
ldap1.virtual1.com --krbRealm VIRTUAL1.COM --krbKeytabFile /root/krb5.keytab
(1/9) Fetching the list of cluster Nodes.
(2/9) Check if cfgcluster has done the basic configuration successfully.
```

```
(3/9) Check whether Interface nodes are reachable from management node.
(4/9) LDAP Kerberos Configuration started.
(5/9) Check whether LDAP server is reachable from Interface nodes.
(6/9) Verification of LDAP server from a node using credentials provided.
(7/9) Updating the system with LDAP Kerberos configuration details.
(8/9) Finalizing configuration.
(9/9) Updating the database.
EFSSG1000I The command completed successfully.
```

3. Run **lsauth** to verify that the cluster is configured for LDAP with Kerberos, as shown in Example 4-32.

Example 4-32 The lsauth command shows that the LDAP configured successfully

```
$lsauth -c st001.virtual1.com
ldapUserSuffix = ou=People
domain = virtual1
ldapUri = ldap://ldap1.virtual1.com
clusterName = st001
ldapGroupSuffix = ou=Group
nscdMode = on
krbKeytabFile = /etc/krb5.keytab
ldapBindDn = cn=manager,dc=ldapserver,dc=com
krbKdc = ldap1.virtual1.com
sslMode = off
AUTH_TYPE = ldap_krb
ldapVersion = 3
krbRealm = VIRTUAL1.COM
ldapBase = dc=ldapserver,dc=com
krbMode = on
EFSSG1000I The command completed successfully.
```

4.8 Configuring SONAS with Network Information Service

NIS is used in UNIX based environment for centralized user and other services management. NIS is used for keeping user, domain, and netgroup information. If you have a UNIX based environment, use NIS for user management and host name management so all systems have the same user information. NIS accesses NAS data stores by using the NFS protocol.

The netgroup is used to group the client system IP and host name, which can be specified when you create NFS exports. NIS is also used for user authentication for different services, such as SSH, FTP, and HTTP. However, SONAS does not support NIS authentication.

SONAS uses NIS for netgroup support and ID mapping. It uses the NIS default domain to resolve the netgroup even when multiple NIS domains are supported. The NIS client configuration needs the server and domain details of NIS server.

Multiple servers per domain and multiple domains are supported. This information is updated on all SONAS nodes by using the cnscm service. The default NIS domain name is required for setting up a SONAS node in NIS domain. Domain and server information is kept in the `/etc/yp.conf` file, and this file is updated on all SONAS nodes by using the cnscm service.

4.8.1 When to use SONAS with AD and NIS

SONAS with AD and NIS is the correct choice for the following conditions:

- You use Microsoft AD to store user information and user passwords.
- You plan to use NFS for NAS access.
- User and group IDs are stored and managed in NIS.

Three different modes of NIS configuration are supported:

- NIS for netgroup only and AD or Samba PDC/NT4 for authentication and AD increment ID mapping logic.
- NIS with ID mapping as an extension to AD or Samba PDC/NT4 and netgroup support.
- Plain NIS without any authentication, just for netgroup support. Only the NFS protocol is supported.

Prerequisites for configuring the SONAS system with AD and NIS

Here are the prerequisites for configuring SONAS with AD and NIS:

- No data files are stored on the SONAS system.
- AD authentication is configured on the SONAS system by running the **`cfgad`** command.
- NIS is used for all user ID mapping.
- All NIS user and group names must be in lowercase, with white space that is replaced by the underscore character (_). For example, an AD user name CAPITAL Name should have a corresponding name on NIS as capital_name.

Limitations for configuring the SONAS system with AD and NIS

Consider the following limitations:

- The low value of the idmapUserRange and the idmapGroupRange cannot be less than 1024.
- UNIX style names do not allow spaces in the name. For mapping AD users or group to NIS users, consider the following conversion on the NIS server:
  - Convert all uppercase alphabets to lowercase.
  - Replace all blank spaces with underscores.

    For example, an AD user or group name CAPITAL Name should have a corresponding name on NIS as capital_name.

Important: If this naming convention is not used for users, NIS mapping cannot happen, and ID mapping for such users follows the user map rules that are defined in the **`cfgnis --userMap`** option for that AD domain.

4.8.2 Understanding key parameters of the cfgnis command

This section describes several key parameters of the **`cfgnis`** command.

--extend { extend }

This parameter sets the extended configuration mode for the NIS over the current authentication type. Appropriate authentication types might be AD or NT4 (Samba PDC). In extended mode, NIS can also be used as an ID mapping mechanism. ID mapping functions can be activated by using the **`--useAsIdmap`** option.

--idmapUserRange { idmapUserRange }

This parameter sets the user ID range. The UIDs of the users, from the AD domain that has the user map rule set to AUTO, are assigned from this range. Samba also uses this range to map some of the well-known SIDs to UIDs. This option is mandatory when it is used with the **`--extend`** and **`--useAsIdmap`** options.

Use the format <lowerID of the range> - <higherID of the range>. LowerID cannot be less than 1024. See the following example:

```
--idmapUserRange 100000-200000
```

--idmapGroupRange { idmapGroupRange }

This parameter sets the group ID range. The GIDs of all the AD groups that have no NIS mapping are assigned from this range. Samba also uses this range to map some of the well-known SIDs to GIDs. This option is mandatory when it is used with the **`--extend`** and **`--useAsIdmap`** options.

Use the format <lowerID of the range> - <higherID of the range>. The LowerID cannot be less than 1024. See the following example:

```
--idmapGroupRange 100000-200000
```

Tips:

The absence of the **`--extend`** option indicates that the NIS will be configured in basic mode. The basic configuration supports NIS authentication for the NFS netgroup. It does not include any other protocol configuration and it does not support the ID mapping mechanism.

There can be only one mapping entry for a specified server. Multiple entries for the same NIS server trigger an error.

LowerID cannot be less than 1024.

4.8.3 Active Directory with NIS

Configuring SONAS system authentication with AD and NIS can be useful when you use AD to store user information and user passwords, user and group IDs are stored and managed in NIS, and you plan to use the NFS protocol for NAS access, or you plan to use the remote replication feature of the SONAS system.

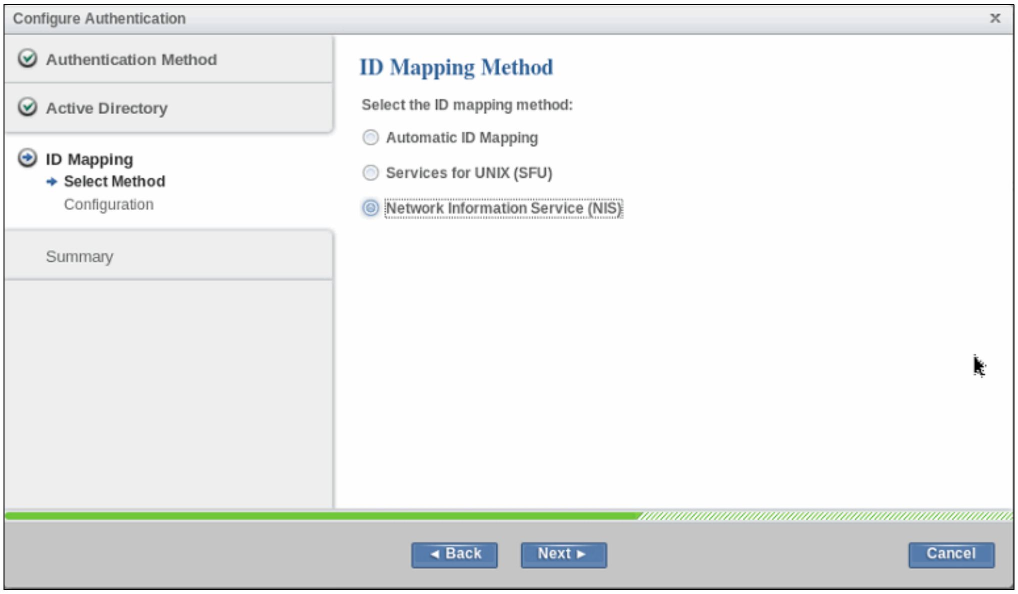
In this method, the cluster must already be configured with AD when you configure it with the CLI. When you use the GUI, it is done at the same time.

Configuring SONAS with AD and NIS by using the GUI

To use the GUI to configure SONAS with AD and NAS, complete the following steps:

1. To get to the Authentication page in the GUI, complete the steps in “Starting the Directory Services configuration in the GUI” on page 242. Follow the instructions in “Choosing Active Directory for configuring authentication” on page 243 until you reach Figure 4-8 on page 245.

2. To configure extended NIS with AD, select the NIS option, as shown in Figure 4-34.



ID Mapping Method

Select the ID mapping method:

○ Automatic ID Mapping

○ Services for UNIX (SFU)

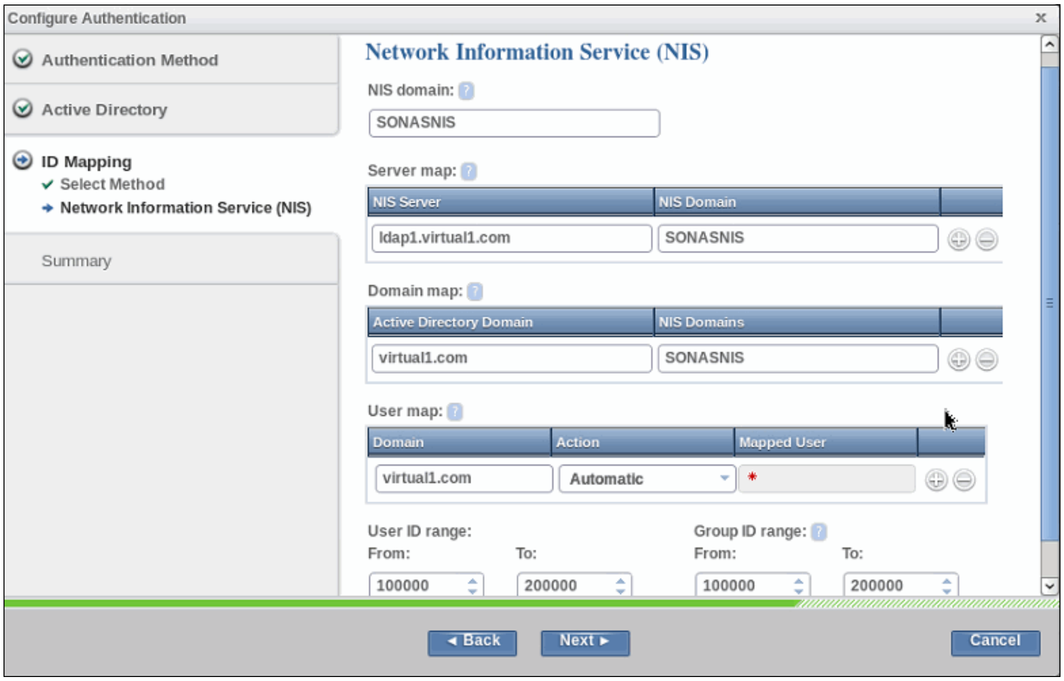
● Network Information Service (NIS)

Figure 4-34 Choose the AD with extended NIS configuration

3. Click **Next**.

4. Enter the NIS configuration details, including the Primary NIS domain name, NIS server, domain maps, user maps, type of user mapping (automatic in this example), and the user and group ID ranges.

   See Figure 4-35 for an example of the window. See the man page for the **`cfgnis`** command for a detailed explanation of these parameters.



Network Information Service (NIS)

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Domain map:

Active Directory Domain	NIS Domains
virtual1.com	SONASNIS

User map:

Domain	Action	Mapped User
virtual1.com	Automatic	*

User ID range: From: 100000 To: 200000

Group ID range: From: 100000 To: 200000

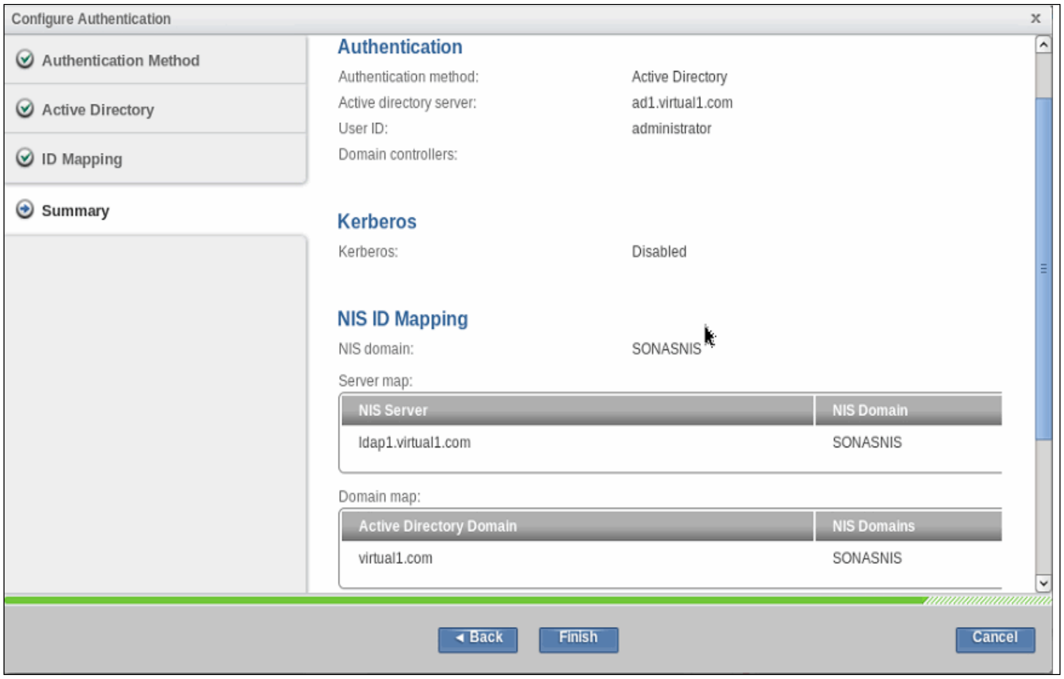
Figure 4-35 Required NIS configuration details

The UIDs of the users, from the AD domain that has the user map rule set to AUTO, are assigned from this range. Samba also uses this range to map some of the well-known SIDs to UIDs. The LowerID cannot be less than 1024. Similarly, you can set the group ID range. The GIDs of all the AD groups that have no NIS mapping are assigned from this range.

In this example, the NIS domain is `SONASNIS`, the NIS server is `ldap1.virtual1.com.`, the AD domain is `virtual1.com`, the user mapping for the domain virtual1.com is `auto`, and the range for UID and GID are specified.

5. Click **Next**.

   The next window shows the summary of the AD server and NIS configuration, as shown in Figure 4-36.



Authentication

Authentication method: Active Directory
Active directory server: ad1.virtual1.com
User ID: administrator
Domain controllers:

Kerberos

Kerberos: Disabled

NIS ID Mapping

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

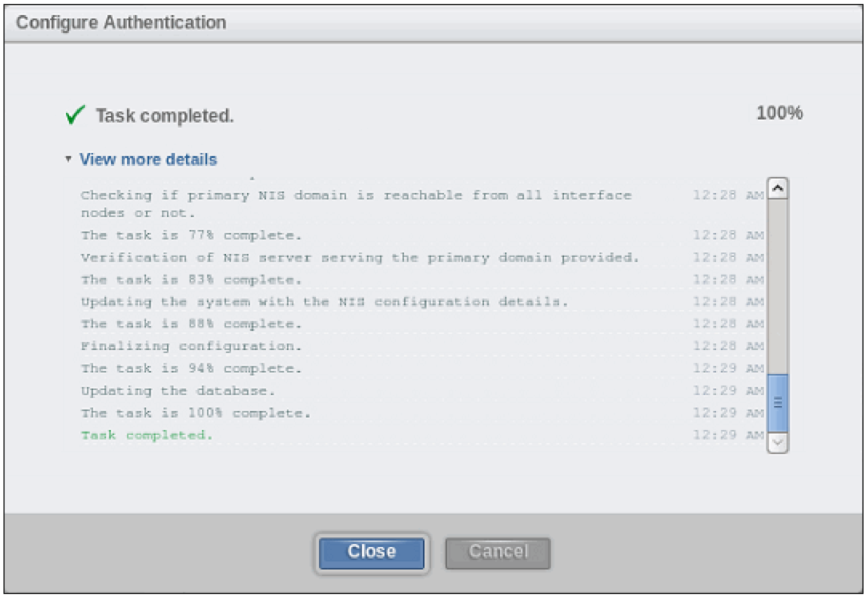
Domain map:

Active Directory Domain	NIS Domains
virtual1.com	SONASNIS

Figure 4-36 Extended NIS configuration summary

6. Click **Finish**.

   The configuration begins. If it is successful, you see a window like the window that is shown in Figure 4-37. In an error occurs, the configuration fails. Check the configuration values and retry the configuration.



✓ Task completed. 100%

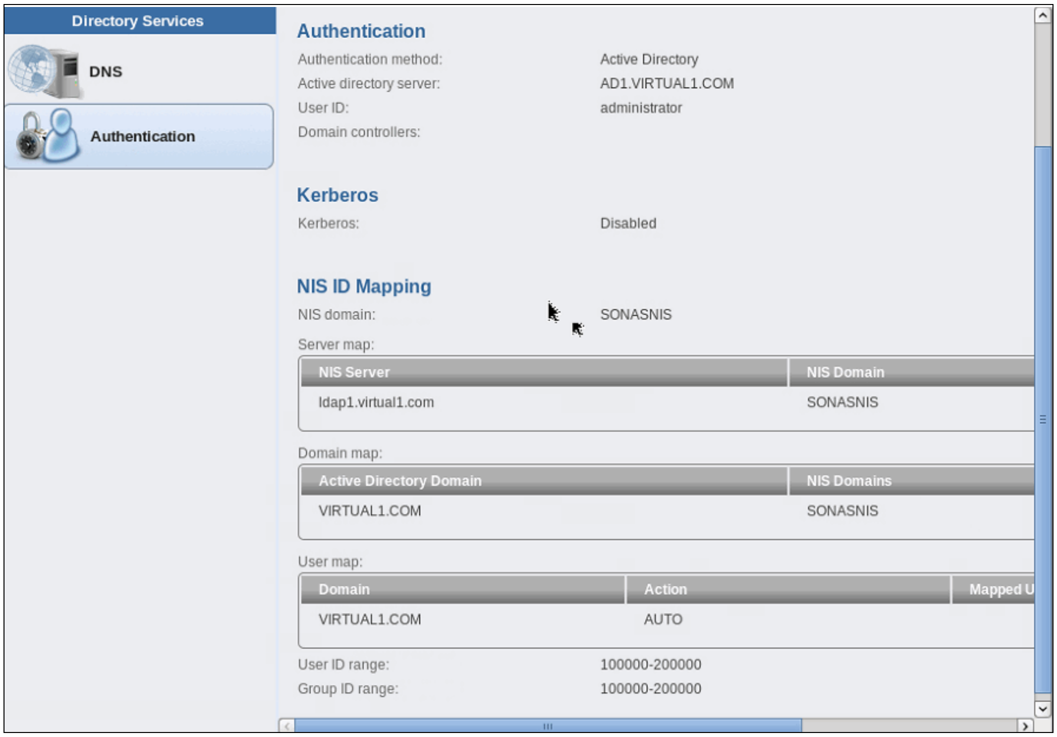
▾ View more details

```
Checking if primary NIS domain is reachable from all interface      12:28 AM
nodes or not.
The task is 77% complete.                                            12:28 AM
Verification of NIS server serving the primary domain provided.      12:28 AM
The task is 83% complete.                                            12:28 AM
Updating the system with the NIS configuration details.              12:28 AM
The task is 88% complete.                                            12:28 AM
Finalizing configuration.                                            12:28 AM
The task is 94% complete.                                            12:29 AM
Updating the database.                                               12:29 AM
The task is 100% complete.                                           12:29 AM
Task completed.                                                      12:29 AM
```

Figure 4-37 Configure Active Directory with extended NIS

7. Click **Close** to return to the main authentication window.

   The details of the configured Authentication method are displayed. In this case, it is AD with extended NIS. See Figure 4-38.



Authentication

Authentication method: Active Directory
Active directory server: AD1.VIRTUAL1.COM
User ID: administrator
Domain controllers:

Kerberos

Kerberos: Disabled

NIS ID Mapping

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Domain map:

Active Directory Domain	NIS Domains
VIRTUAL1.COM	SONASNIS

User map:

Domain	Action	Mapped U
VIRTUAL1.COM	AUTO	

User ID range: 100000-200000
Group ID range: 100000-200000

Figure 4-38 Active Directory with Extended NIS is configured

Configuring SONAS with Active Directory and NIS by using the CLI

To configure AD with Extended NIS, complete the following steps:

1. Run **lsauth** to verify that AD is already configured as the authentication method. See Example 4-33. For help about configuring AD, see 4.4.6, “Configuring Active Directory with the CLI” on page 248.

   Example 4-33 Active Directory is already configured on the cluster

   ```
   $lsauth -c st001.virtual1.com
   AUTH_TYPE = ad
   idMapConfig = 10000000-299999999,1000000
   idMappingMethod = auto
   domain = VIRTUAL1
   clusterName = st001.virtual1.com
   userName = administrator
   idMapRole = master
   adHost = AD1.VIRTUAL1.COM
   passwordServer = *
   krbMode = off
   realm = VIRTUAL1.COM
   EFSSG1000I The command completed successfully.
   ```

2. Run **cfgnis** to configure SONAS with Extended NIS. See Example 4-34. In this example, NIS server `ldap1.virtual1.com` and Domain name `SONASNIS` are used. For help with the **cfgnis** command, see Example 4-36 on page 284.

   Example 4-34 Example command to configure NIS for netgroup support

   ```
   $cfgnis --extend -d SONASNIS --serverMap ldap1.virtual1.com:SONASNIS
   (1/9) Check if NIS can be extended or not.Checking if authentication type is AD
   or NT4
   (2/9) Fetching the list of cluster Nodes.
   (3/9) Check whether Interface nodes are reachable from management node.
   (4/9) Check if primary NIS domain is served by any NIS server or not.
   (5/9) Checking if primary NIS domain is reachable from all interface nodes or
   not.
   (6/9) Verification of NIS server serving the primary domain provided.
   (7/9) Updating the system with the NIS configuration details.
   (8/9) Finalizing configuration.
   (9/9) Updating the database.
   EFSSG1000I The command completed successfully.
   ```

   d. Run **lsauth** to verify that the cluster is configured with AD with Extended NIS. See Example 4-35.

   Example 4-35 Active Directory with Extended NIS is configured

   ```
   $lsauth
   AUTH_TYPE = ad
   NIS_ServerMap = ldap1.virtual1.com:SONASNIS
   idMapConfig = 10000000-299999999,1000000
   domain = VIRTUAL1
   idMappingMethod = auto
   clusterName = st001.virtual1.com
   userName = administrator
   idMapRole = master
   ```

```
adHost = AD1.VIRTUAL1.COM
NIS_Domain = SONASNIS
passwordServer = *
realm = VIRTUAL1.COM
EFSSG1000I The command completed successfully.
```

See Example 4-36 for **cfgnis** command help.

Example 4-36 Help for the cfgnis command

```
$cfgnis --help
usage: cfgnis  [--extend] [-m] [--useAsIdmap] [-d <nisDomain>] [--serverMap <serverMap>]
[--domainMap <domainMap>] [--userMap <userMap>] [--idmapUserRange <idmapUserRange>]
[--idmapGroupRange <idmapGroupRange>] [--disableVerifyServer] [-c < clusterID | clusterName >]
Configures the Network Information Service (NIS).

Parameter                Description
    --extend             Sets the extended configuration mode for the NIS over the current
authentication type.
-m, --modify             Modifies the current configuration of the NIS.
    --useAsIdmap         Indicates that the NIS is to be used for ID mapping.
-d, --nisDomain          Sets the primary NIS domain into the registry.
    --serverMap          Specifies the mapping between the different NIS servers and NIS
domains.
    --domainMap          Specifies the mapping between the Active Directory and other NIS
domains.
    --userMap            Describes the action to be taken if the user does not have a mapping
entry in the NIS.
    --idmapUserRange     Sets the user ID range into the registry.
    --idmapGroupRange    Sets the group ID range into the registry.
    --disableVerifyServer Specifies whether the NIS server is to be verified before
configuration.
-c, --cluster            The cluster scope for this command
```

4.8.4 Samba PDC/NT4 with NIS

Configuring SONAS system authentication with NT4/Samba PDC and NIS can be useful when you use AD to store user information and user passwords, you plan to use the NFS protocol for NAS access, or you plan to use SONAS system remote replication, and the user IDs are stored and managed in NIS.

In this mode, SONAS supports both the CIFS and NFS protocols. Netgroups are also supported.

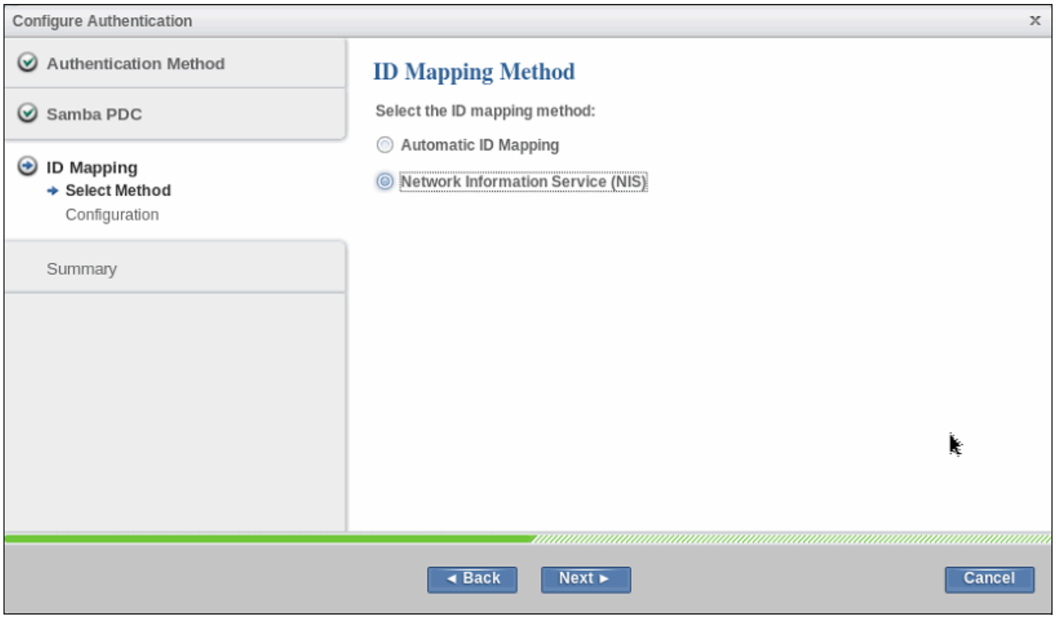
When you use the CLI, the cluster must already be configured with Samba PDC/NT4. When using the GUI, it is done at the same time.

Configuring SONAS with Samba PDC/NT4 and NIS by using the GUI

To configure SONAS with Samba PDF/NTU and NIS, complete the following steps:

1. Complete the steps in “Starting the Directory Services configuration in the GUI” on page 242 to get to the Authentication window in the GUI.
2. Follow the instructions for 4.6.1, “Configuring SONAS with Samba PDC/NT4 by using the GUI” on page 256 until you reach Figure 4-22 on page 257.

3. To configure extended NIS along with Samba PDC/NT4, select **NIS** as the ID Mapping Method, as shown in Figure 4-39.



ID Mapping Method

Select the ID mapping method:

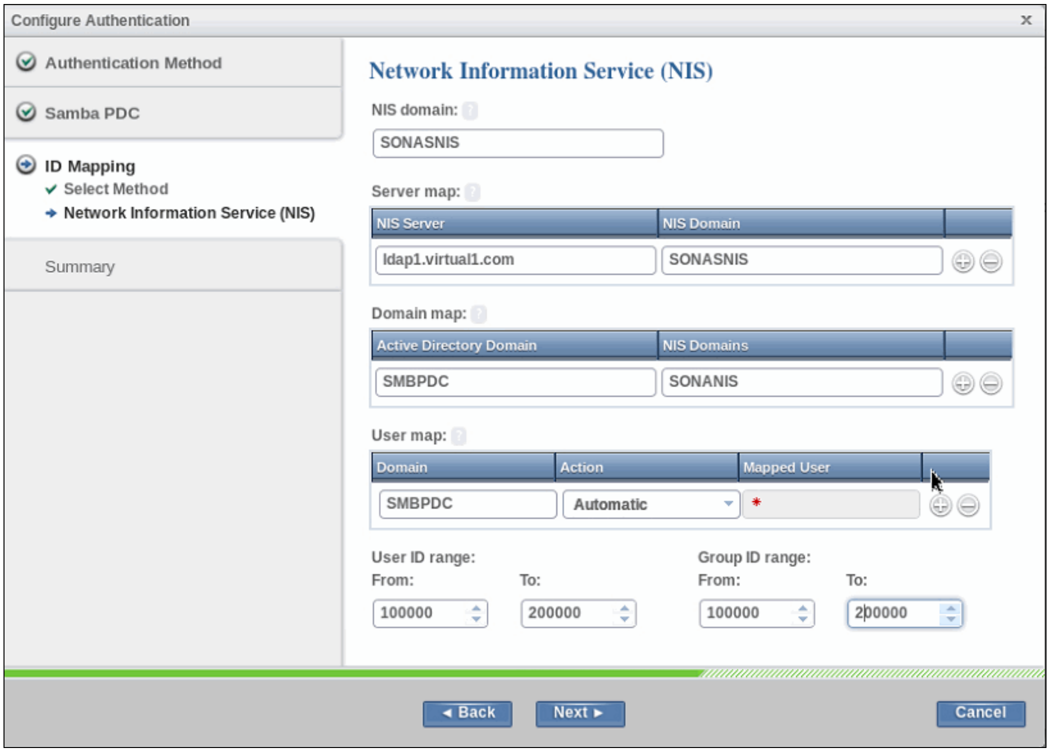
○ Automatic ID Mapping

● Network Information Service (NIS)

Figure 4-39 Choose Samba PDC/NT4 with Extended NIS

4. Click **Next**.

5. Enter the NIS configuration details. See Figure 4-40. Enter the Primary NIS domain name, NIS server, domain maps, user maps, type of user mapping (automatic in this example), and the user and group ID ranges. For a detailed explanation of these parameters, see the man page for the **`cfgnis`** command.



Network Information Service (NIS)

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Domain map:

Active Directory Domain	NIS Domains
SMBPDC	SONANIS

User map:

Domain	Action	Mapped User
SMBPDC	Automatic	*

User ID range: From: 100000 To: 200000

Group ID range: From: 100000 To: 200000

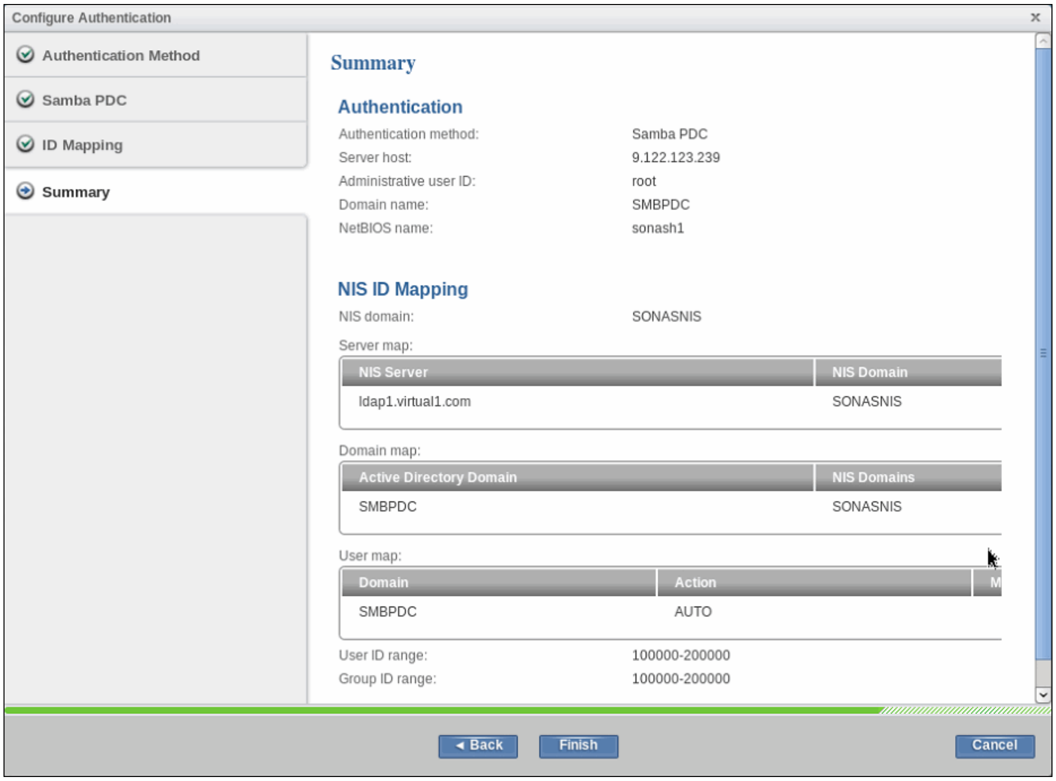
Figure 4-40 NIS details that are required for the Extended NIS configuration

The UIDs of the users from the AD domain that has the user map rule set to AUTO are assigned from this range. Samba also uses this range to map some of the well-known SIDs to UIDs. The LowerID cannot be less than 1024. Similarly, you can set the group ID range. The GIDs of all the AD groups that have no NIS mapping are assigned from this range.

In this example, the NIS domain is `SONASNIS`, the NIS server is `ldap1.virtual1.com.`, the Samba PDC domain is SMBPDC, the user mapping for the domain `virtual1.com` is auto, and the range for the UID and GID are specified.

6. Click **Next**.

The next window shows the summary of the configuration that is entered so far. See Figure 4-41 on page 287.



Summary

Authentication

Authentication method: Samba PDC
Server host: 9.122.123.239
Administrative user ID: root
Domain name: SMBPDC
NetBIOS name: sonash1

NIS ID Mapping

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Domain map:

Active Directory Domain	NIS Domains
SMBPDC	SONASNIS

User map:

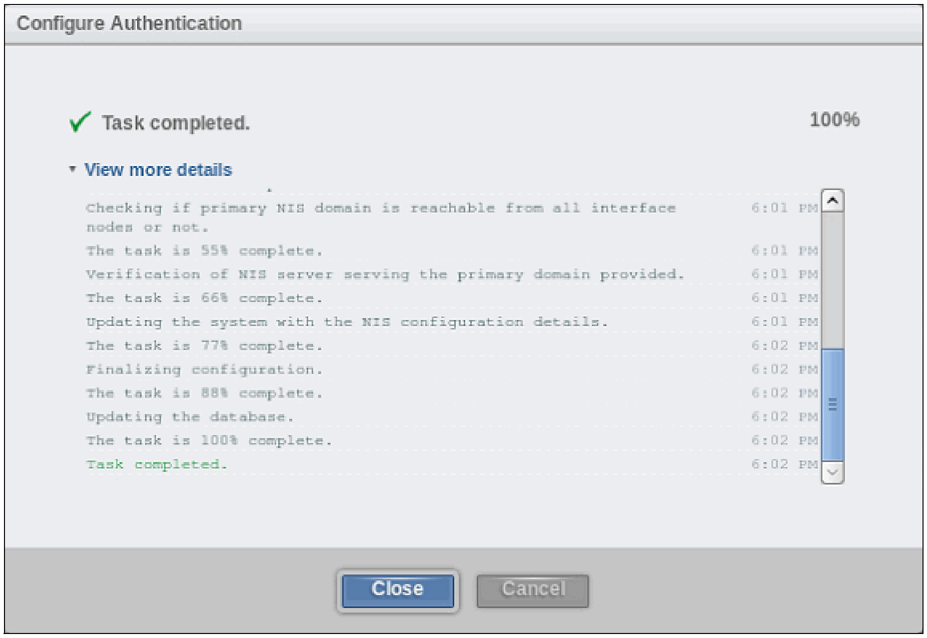
Domain	Action	M
SMBPDC	AUTO	

User ID range: 100000-200000
Group ID range: 100000-200000

Figure 4-41 Samba PDC/NT4 configuration summary

7. Click **Finish**.

   The configuration begins and, if it is successful, a window similar to Figure 4-42 opens. If there is an error, the configuration fails. Check the configuration values and try the configuration again.



✓ Task completed. 100%

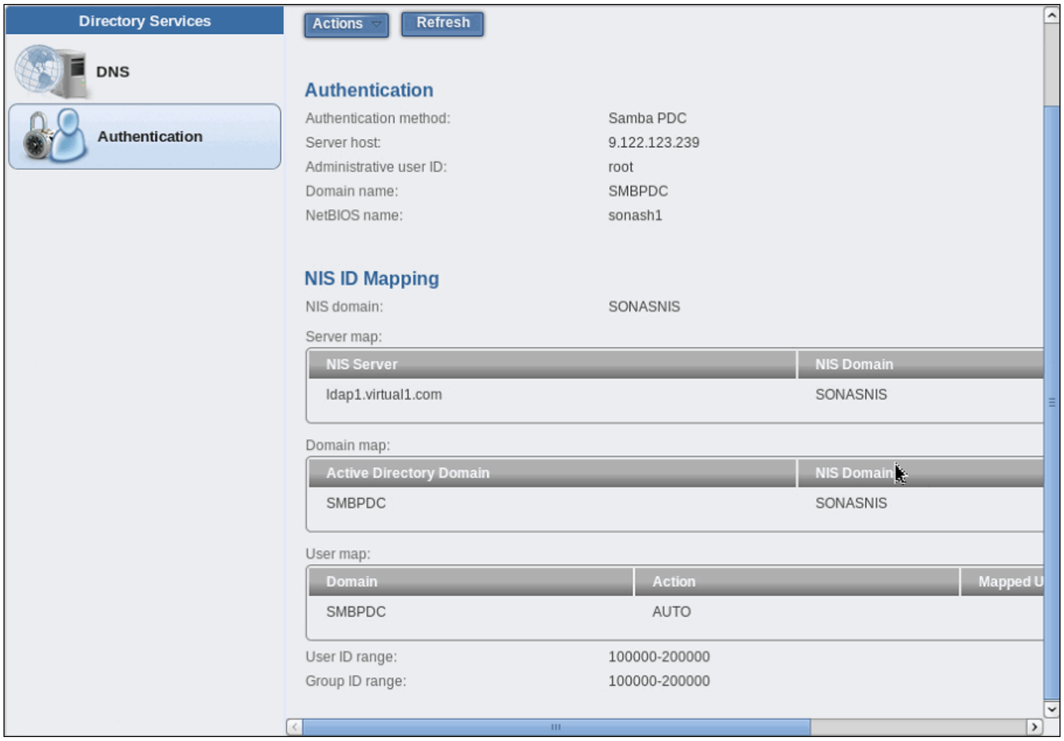
▾ View more details

```
Checking if primary NIS domain is reachable from all interface      6:01 PM
nodes or not.
The task is 55% complete.                                            6:01 PM
Verification of NIS server serving the primary domain provided.      6:01 PM
The task is 66% complete.                                            6:01 PM
Updating the system with the NIS configuration details.              6:01 PM
The task is 77% complete.                                            6:02 PM
Finalizing configuration.                                            6:02 PM
The task is 88% complete.                                            6:02 PM
Updating the database.                                               6:02 PM
The task is 100% complete.                                           6:02 PM
Task completed.                                                      6:02 PM
```

Figure 4-42 Samba PDC/NT4 configuration is successful

8. Click **Close** to return to the main authentication window.

   This window shows the details of the configured authentication method. In this case, it is Samba PDC/NT4 with extended NIS. See Figure 4-43.



Authentication

Authentication method:	Samba PDC
Server host:	9.122.123.239
Administrative user ID:	root
Domain name:	SMBPDC
NetBIOS name:	sonash1

NIS ID Mapping

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Domain map:

Active Directory Domain	NIS Domain
SMBPDC	SONASNIS

User map:

Domain	Action	Mapped U
SMBPDC	AUTO	

User ID range:	100000-200000
Group ID range:	100000-200000

Figure 4-43 Samba PDC/NT4 is configured

Configuring SONAS with Samba PDC/NT4 and NIS by using the CLI

To configure Samba PDC/NT4 with Extended NIS, complete the following steps:

1. Run `lsauth` to verify that Samba PDC/NT4 is configured as the authentication method. See Example 4-37. For help configuring Samba PDC/NT4, see 4.6.2, “Configuring SONAS with Samba PDC/NT4 by using the CLI” on page 260.

Example 4-37 Active Directory configured on the cluster

```
# lsauth -c st001.virtual1.com
AUTH_TYPE = nt4
idMappingMethod = auto
domain = SMBPDC
nt4NetbiosName = ldap1
clusterName = st001
nt4Host = ldap1.virtual1.com
userName = root
EFSSG1000I The command completed successfully.
```

2. Run **cfgnis** to configure SONAS with Extended NIS. See Example 4-38. In this example, NIS server `ldap1.virtual1.com` and domain name `SONASNIS` are used. For help with the **cfgnis** command, see Example 4-36 on page 284.

Example 4-38 Command to configure NIS for netgroup support

```
# cfgnis --extend -d SONASNIS --serverMap ldap1.virtual1.com:SONASNIS
(1/9) Check if NIS can be extended or not.Checking if authentication type is AD
or NT4
(2/9) Fetching the list of cluster Nodes.
(3/9) Check whether Interface nodes are reachable from management node.
(4/9) Check if primary NIS domain is served by any NIS server or not.
(5/9) Checking if primary NIS domain is reachable from all interface nodes or
not.
(6/9) Verification of NIS server serving the primary domain provided.
(7/9) Updating the system with the NIS configuration details.
(8/9) Finalizing configuration.
(9/9) Updating the database.
EFSSG1000I The command completed successfully.
```

3. Run **lsauth** to verify that the cluster is configured with AD with Extended NIS, as shown in Example 4-39.

Example 4-39 Active Directory with Extended NIS configured

```
# lsauth -c st001.virtual1.com
AUTH_TYPE = nt4
NIS_ServerMap = ldap1.virtual1.com:SONASNIS
domain = SMBPDC
idMappingMethod = auto
clusterName = st001
nt4NetbiosName = ldap1
userName = root
nt4Host = ldap1.virtual1.com
NIS_Domain = SONASNIS
EFSSG1000I The command completed successfully.
```

4.8.5 Plain NIS without authentication (only for netgroup support)

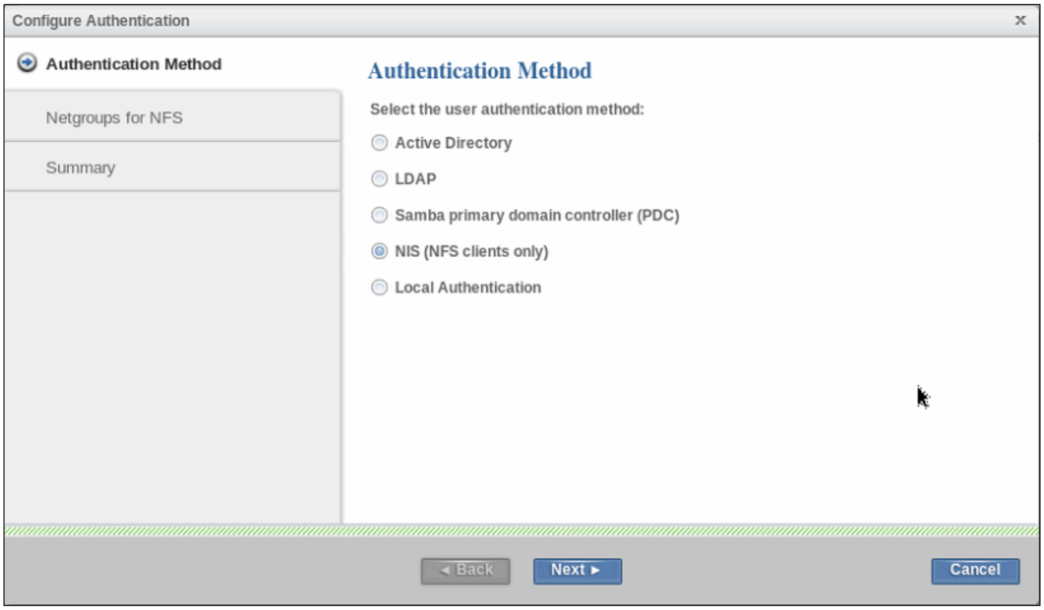
This mode is used only to provide netgroup support for NFS clients. In this mode, only the NFS protocol is supported over UNIX; all other protocols are disabled.

Configuring SONAS with NIS by using the GUI

To configure SONAS with NIS, complete the following steps:

1. Complete the steps in “Starting the Directory Services configuration in the GUI” on page 242 to get to the Authentication window in the GUI.
2. Click **Authentication**.

3. In the window that opens, select **NIS (NFS Clients only)**. See Figure 4-44.



Configure Authentication

Authentication Method

Select the user authentication method:

○ Active Directory

○ LDAP

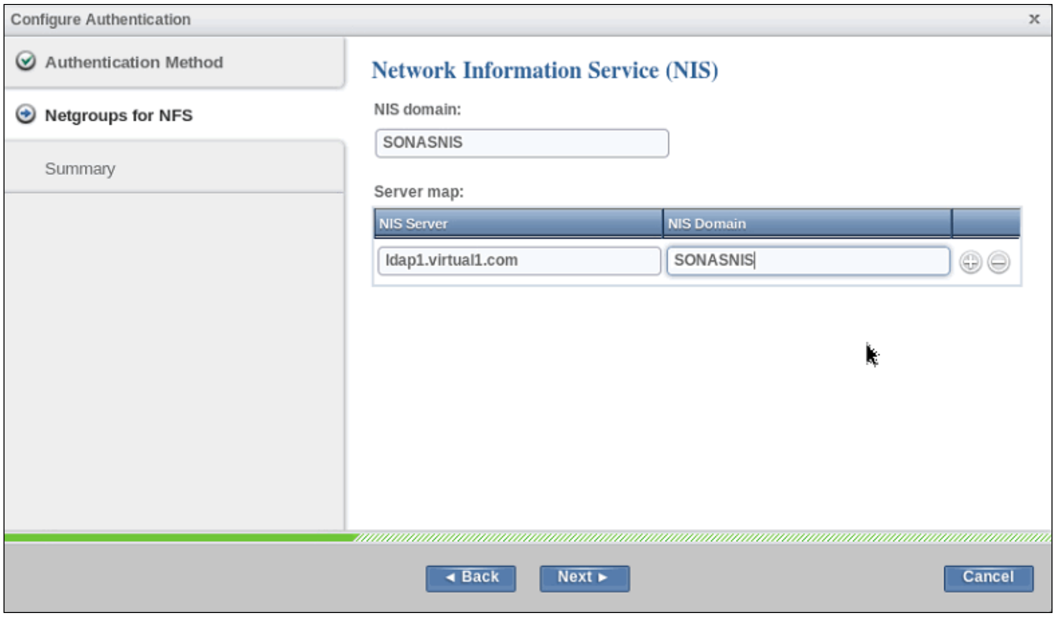
○ Samba primary domain controller (PDC)

● NIS (NFS clients only)

○ Local Authentication

Figure 4-44 Choose Plain NIS as SONAS authentication

4. Click **Next**.
5. Enter the NIS details, including the NIS Server and the NIS Domain name. See Figure 4-45. In this example, `ldap1.virtual1.com` is the NIS server and `SONASNIS` is the primary domain.



Configure Authentication

Network Information Service (NIS)

NIS domain: SONASNIS

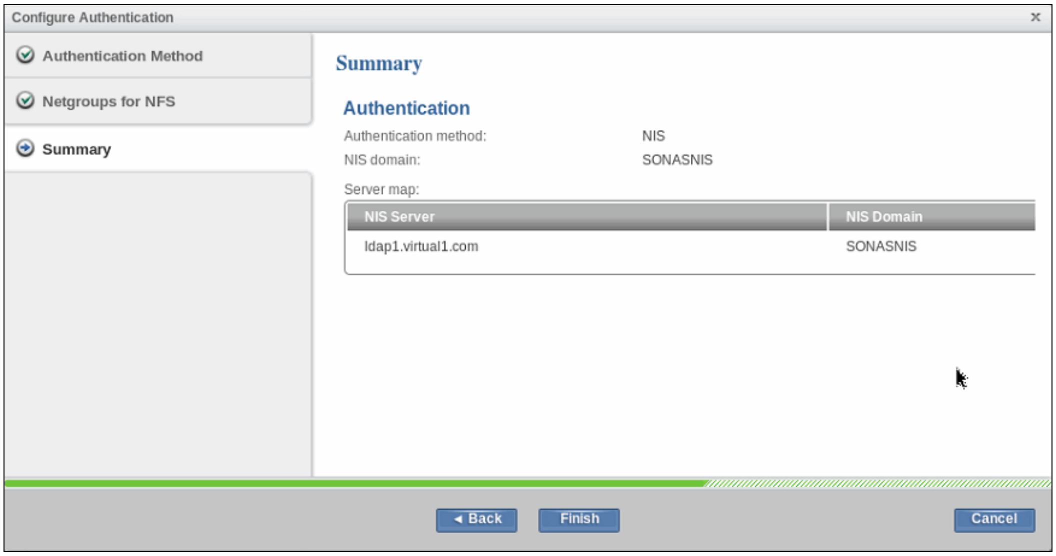
Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Figure 4-45 NIS configuration details that are required

6. Click **Next**.

The next window shows the NIS configuration summary (Figure 4-46 on page 291).



Summary

Authentication

Authentication method: NIS

NIS domain: SONASNIS

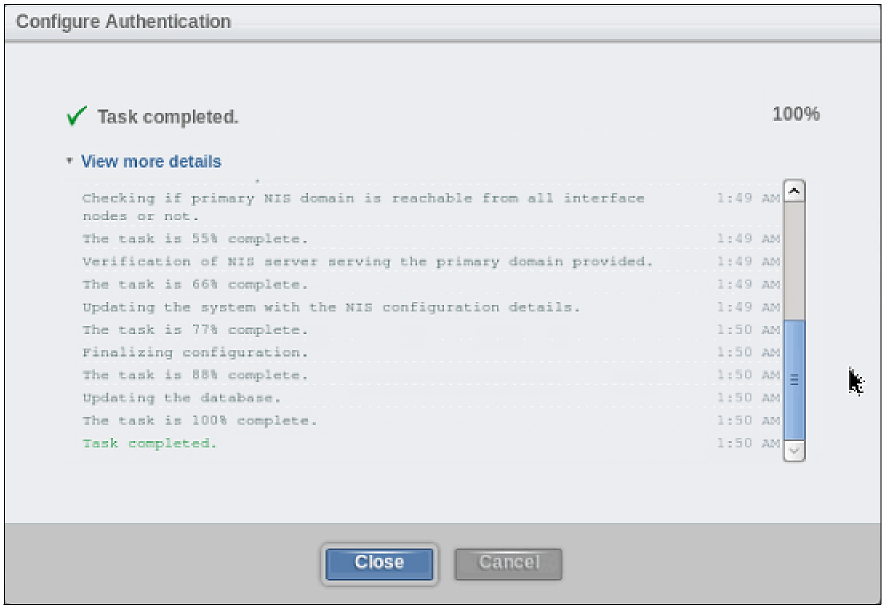
Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Figure 4-46 NIS configuration summary

7. Click **Finish** to proceed.

   The configuration begins. If it is successful, you see a window similar to the one that is shown in Figure 4-47. In there is an error, the configuration fails. Check the configuration values and try the configuration.



✓ Task completed. 100%

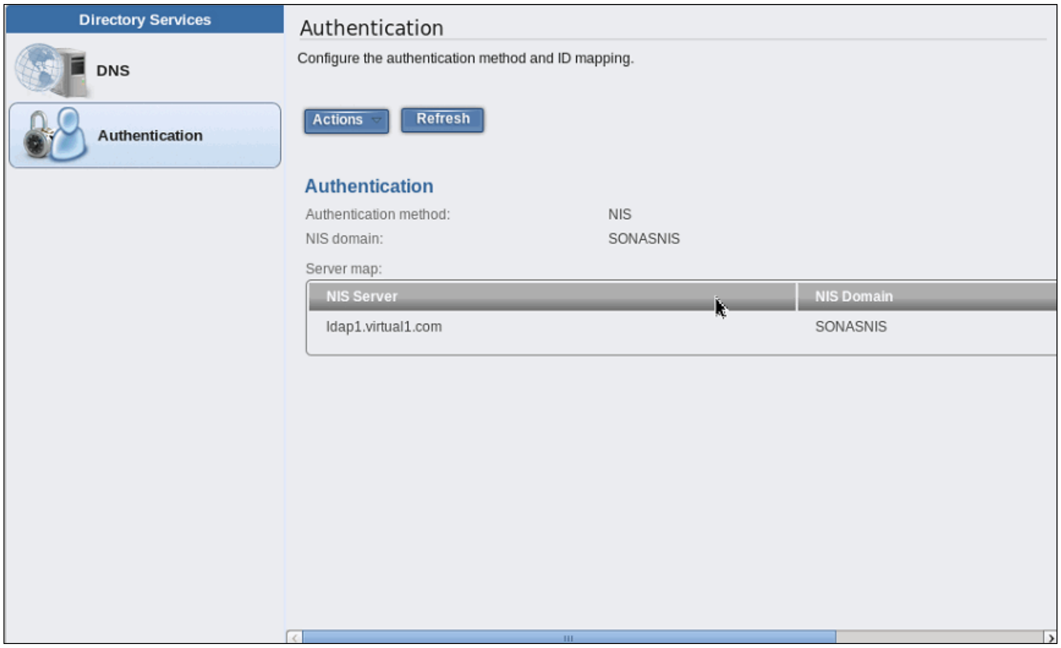
▾ View more details

```
Checking if primary NIS domain is reachable from all interface      1:49 AM
nodes or not.
The task is 55% complete.                                            1:49 AM
Verification of NIS server serving the primary domain provided.      1:49 AM
The task is 66% complete.                                            1:49 AM
Updating the system with the NIS configuration details.             1:49 AM
The task is 77% complete.                                            1:50 AM
Finalizing configuration.                                            1:50 AM
The task is 88% complete.                                            1:50 AM
Updating the database.                                               1:50 AM
The task is 100% complete.                                           1:50 AM
Task completed.                                                      1:50 AM
```

Figure 4-47 NIS configured successfully

8. Click **Close** to return to the main authentication window.

   This window shows the details of the configured authentication method. See Figure 4-48.



Authentication

Configure the authentication method and ID mapping.

Authentication

Authentication method: NIS

NIS domain: SONASNIS

Server map:

NIS Server	NIS Domain
ldap1.virtual1.com	SONASNIS

Figure 4-48 NIS is configured on the cluster

Configuring SONAS with NIS by using the CLI

To configure SONAS with NIS, complete the following steps:

1. Run **lsauth** to verify that no authentication method is configured on the cluster. See Example 4-40.

Example 4-40 The lsauth command shows that no authentication server is configured

```
$lsauth -c st001.virtual1.com
EFSSG0571I Cluster st001.virtual1.com is not configured with any type of
authentication server.
```

2. Configure NIS by running **cfgnis**, as shown in Example 4-41. Do not use the **--extend** parameter because NIS is being configured alone on the system and not with an existing AD or NT4.

Example 4-41 Example cfgnis command to configure SONAS with NIS

```
$cfgnis  -d SONASNIS --serverMap ldap1.virtual1.com:SONASNIS
(1/9) Check if basic NIS can be configured on the system.
(2/9) Fetching the list of cluster Nodes.
(3/9) Check whether Interface nodes are reachable from management node.
(4/9) Check if primary NIS domain is served by any NIS server or not.
(5/9) Checking if primary NIS domain is reachable from all interface nodes or
not.
(6/9) Verification of NIS server serving the primary domain provided.
(7/9) Updating the system with the NIS configuration details.
(8/9) Finalizing configuration.
(9/9) Updating the database.
EFSSG1000I The command completed successfully.
```

3. Run **lsauth** to verify that the cluster is configured with plain NIS, as shown in Example 4-42.

Example 4-42 Active Directory with extended NIS configured

```
$lsauth
AUTH_TYPE = nis
NIS_ServerMap = ldap1.virtual1.com:SONASNIS
idMappingMethod = none
clusterName = st001.virtual1.com
NIS_Domain = SONASNIS
EFSSG1000I The command completed successfully.
```

4.9 Configuring SONAS with local authentication

Starting with Version 1.5.1, SONAS supports configuring authentication by using a local authentication server that is hosted internally by SONAS. Local authentication supports both Windows and UNIX clients with some limitations. Hence, for mixed environments with both Windows and UNIX clients, this mechanism can be considered.

Local authentication eliminates the requirement for an external authentication server and provides the capability to do user authentication and ID mapping from within the SONAS system. This method reserves an ID range and allocates UID and GID on a first-come, first-served incremental basis as the default.

A CLI or GUI user with only the SecurityAdmin role can create, modify, and delete users and groups that are stored on the internal authentication server that is used for authentication and ID mapping for the NFS, CIFS, HTTPS, SCP, and FTP NAS protocols. Netgroups, and Kerberos for NFS or CIFS, are not supported.

New **mknasgroup**, **rmnasgroup**, **lsnasgroup**, **mknasuser**, **chnasuser**, **lsnasuser**, and **rmnasuser** commands can be used for local data access user and group management. The **cfglocalauth** command is used to configure the SONAS cluster for local authentication.

The SONAS internal authentication server cannot be used as an external authentication server for other systems, including any other Storwize V7000 Unified system. The local data access user and local data access user group names that are stored on the internal authentication server are case-insensitive.

You can create a maximum of 100 local data access user groups and a maximum of 1000 local data access users. You can assign a local data access user to a maximum of 16 local data access user groups.

Note: User and user group names of system, CLI, and data access and local authentication users and user groups should not be the same. Using duplicate names might cause unexpected behavior.

4.9.1 Choosing SONAS with local authentication

Plan to use SONAS local authentication server when you do not have an external server for authentication and ID mapping and you require a limited number of data access users and groups. You can easily manage the replication of user and group data between the production and target cluster if asynchronous support is wanted.

4.9.2 Limitations of SONAS with local authentication

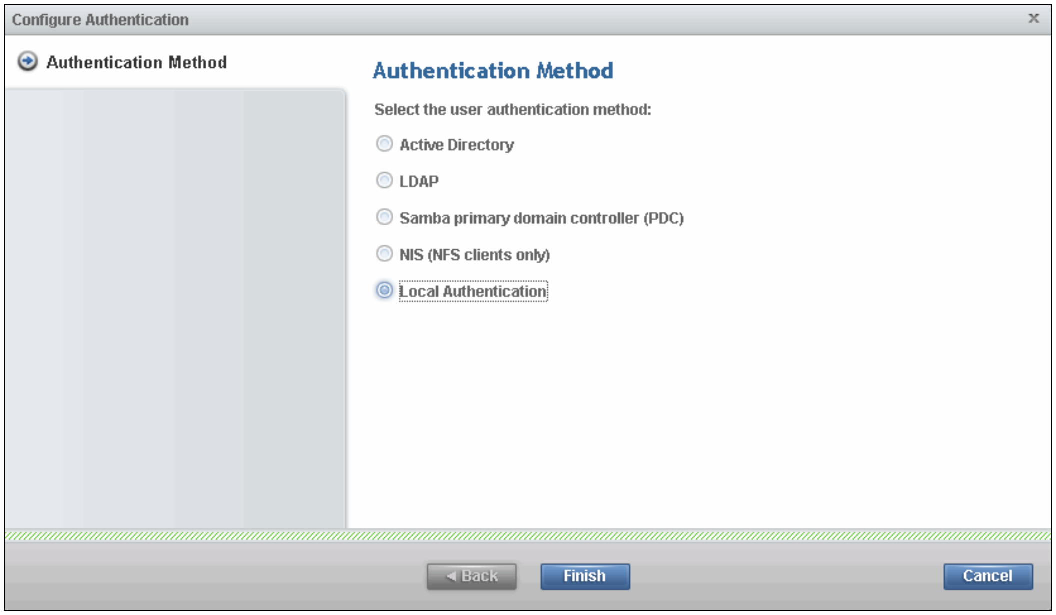
SONAS with local authentication has the following limitations:

- There is no support for the migration of existing authentication server user and group data to the local authentication server.
- There is no support for migration of user and group data from the local authentication server to an external authentication server.
- NAS user and group names are case-sensitive.
- NAS user and group names cannot collide with the CLI and system users.
- There is no support for secure NFS and CIFS. SONAS local authentication does not support Kerberized access.
- The local authentication server that is hosted inside SONAS cannot be used as an external directory server.
- The local authentication server does not support NFS netgroups.
- There are no local data access user password policies (except for a minimum length password).
- Stand-alone Windows clients (not part of any domain) lack ACL update capabilities.
- User names and IDs that are used with local authentication must be the same as those that are used on NFSv4 clients.
- There is support for a maximum of 1000 users and 100 groups. A user can belong to only 16 groups. A group can consist of 1000 users.
- Async replication is not supported.

4.9.3 Configuring local authentication by using the GUI

To configure local authentication with the GUI, complete the following steps:

1. Click **Settings** → **Directory services** → **Authentication**.
2. Click **Configure** and select **Local Authentication**, as shown in Figure 4-49 on page 295.



Authentication Method

Select the user authentication method:

○ Active Directory

○ LDAP

○ Samba primary domain controller (PDC)

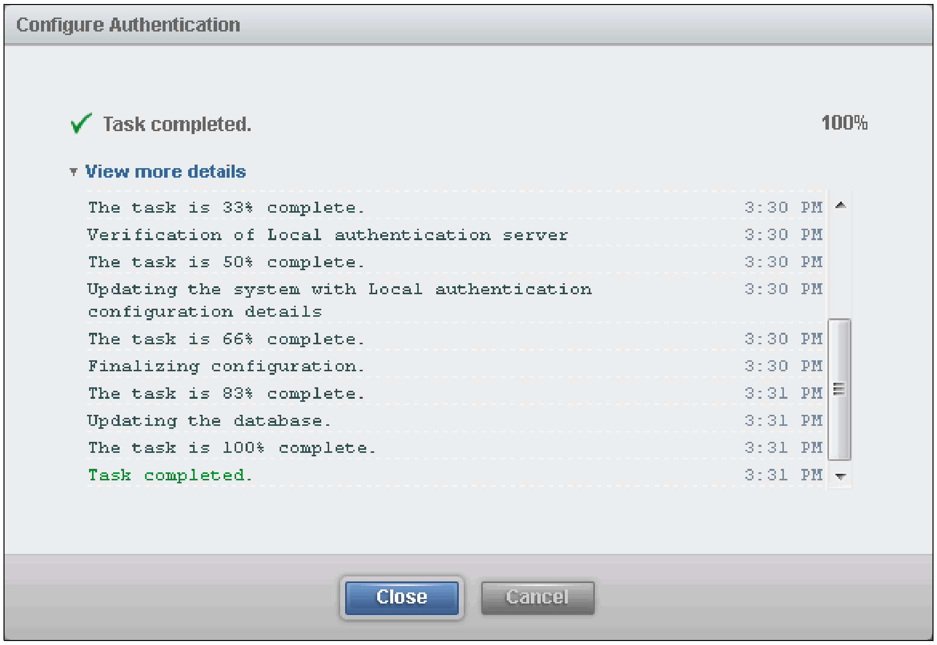
○ NIS (NFS clients only)

● Local Authentication

Figure 4-49 Select Local Authentication

3. Click **Finish** to proceed.

   The configuration begins. If it is successful, you see the window that is shown in Figure 4-50.



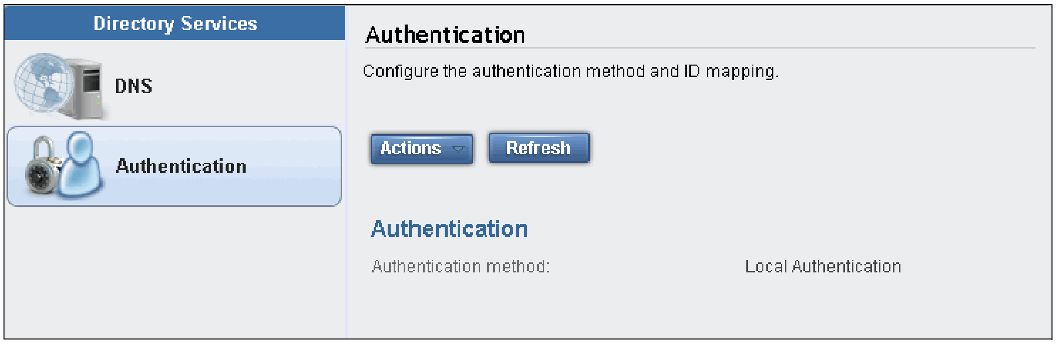
✓ Task completed. 100%

View more details

```
The task is 33% complete.                                   3:30 PM
Verification of Local authentication server                 3:30 PM
The task is 50% complete.                                   3:30 PM
Updating the system with Local authentication               3:30 PM
configuration details
The task is 66% complete.                                   3:30 PM
Finalizing configuration.                                   3:30 PM
The task is 83% complete.                                   3:31 PM
Updating the database.                                      3:31 PM
The task is 100% complete.                                  3:31 PM
Task completed.                                             3:31 PM
```

Figure 4-50 Local authentication is configured successfully

4. Click **Close** to return to the main authentication window. This window shows the details of the configured authentication method. See Figure 4-51.



Authentication method: Local Authentication

Figure 4-51 Local authentication is configured on the cluster

4.9.4 Configuring local authentication by using the CLI

To configure local authentication with the CLI, complete the following steps:

1. Run **lsauth** to verify that no authentication method is configured on the cluster. See Example 4-43.

Example 4-43 Sample lsauth output with no authentication configured

```
$lsauth -c furby.storage.tucson.ibm.com
EFSSG0571I Cluster furby.storage.tucson.ibm.com is not configured with any type
of authentication server.
```

2. Configure local authentication by running **cfglocalauth**, as shown in Example 4-44.

Example 4-44 Configure local authentication

```
$cfglocalauth
(1/6) Local authentication server is already configured and hence skipping
server setup.
(2/6) Local authentication configuration started
(3/6) Verification of Local authentication server
(4/6) Updating the system with Local authentication configuration details
(5/6) Finalizing configuration.
(6/6) Updating the database.
EFSSG1000I The command completed successfully.
```

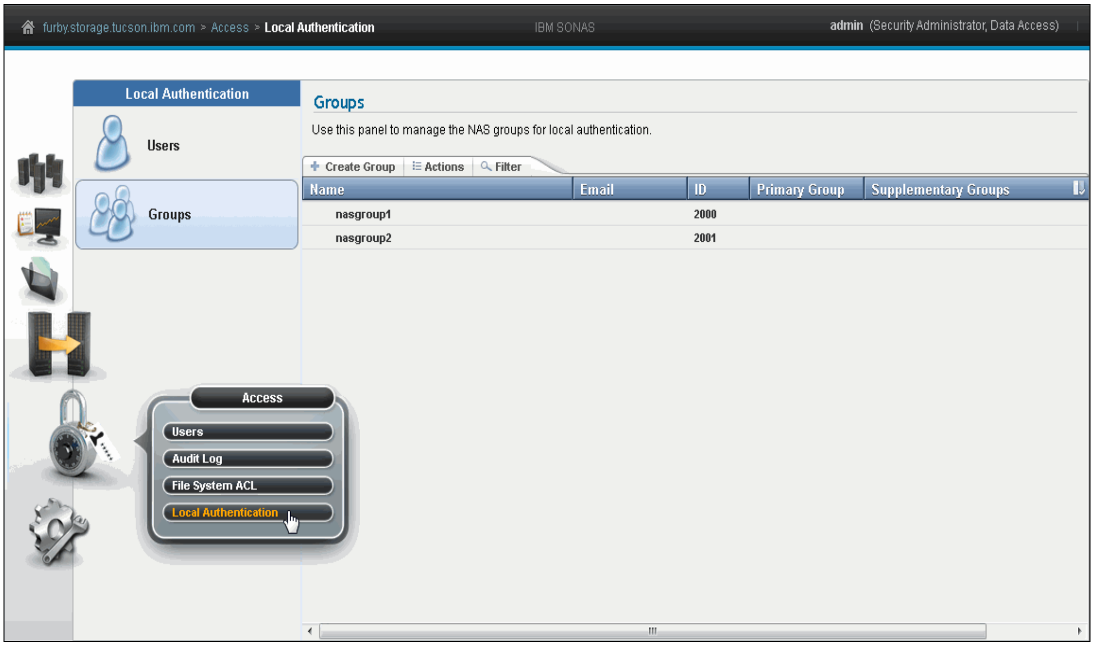
3. Run **lsauth** to verify that the cluster is configured with local authentication, as shown in Example 4-45.

Example 4-45 Local authentication is configured

```
$lsauth
AUTH_TYPE = local
LocalAuthServerName = furby
idMappingMethod = none
clusterName = furby.storage.tucson.ibm.com
LocalAuthServerSID = S-1-5-21-3527331916-285169719-1775312056
EFSSG1000I The command completed successfully.
```

4.9.5 Managing local NAS users and groups by using the GUI

SONAS V1.5.1 provides support for managing local data access users and groups who can access the NAS shares from the management GUI. You can select **Local Authentication** from the **Access** menu, as shown in Figure 4-52.



Groups

Use this panel to manage the NAS groups for local authentication.

Name	Email	ID	Primary Group	Supplementary Groups
nasgroup1		2000		
nasgroup2		2001		

Figure 4-52 Local authentication user/group management

Local data access groups and users can be created by an authorized CLI or GUI administrative user from the window that is shown in Figure 4-52.

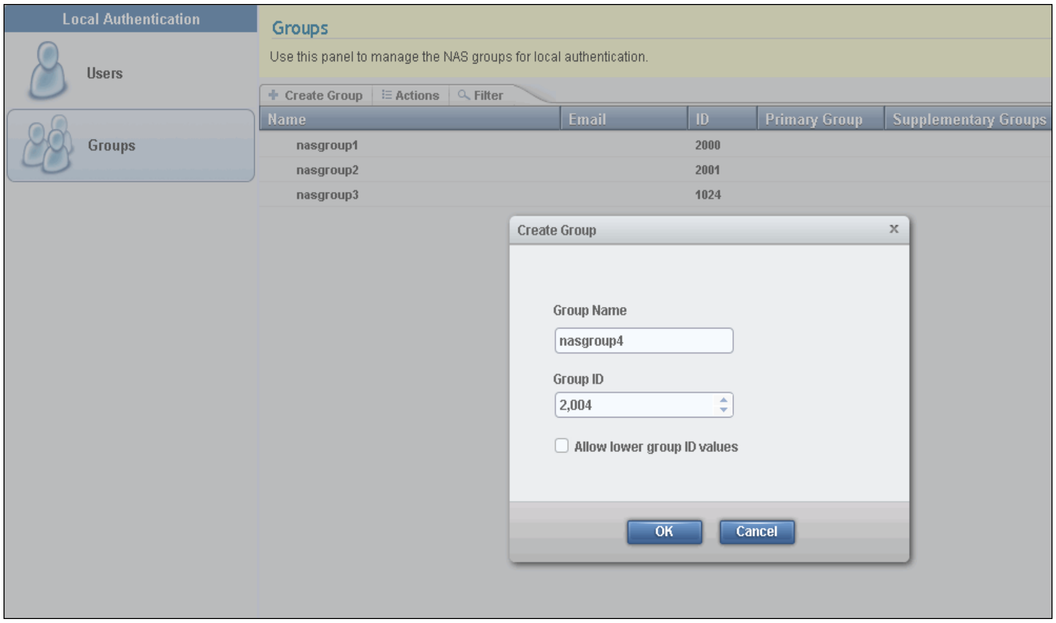
Before you create users, you must create the group to which users belong.

To create a group, complete the following steps:

1. Click **Create Group**.

2. Add the group details in the window that opens, as shown in Figure 4-53.

   A group name can have a minimum of 3, and a maximum of 32, ASCII characters. You can specify a GID or let the SONAS system automatically assign a GID. Automatically assigned GIDs start with 1024. It is a preferred practice to make the GID equal or greater than 1024 to avoid conflicts with SONAS CLI and system groups. You can create groups with a GID less that 1024 by selecting **Allow lower group ID values**. The maximum GID value is 2,147,482,648.



Use this panel to manage the NAS groups for local authentication.

Name	Email	ID	Primary Group	Supplementary Groups
nasgroup1		2000		
nasgroup2		2001		
nasgroup3		1024		

Group Name: nasgroup4

Group ID: 2,004

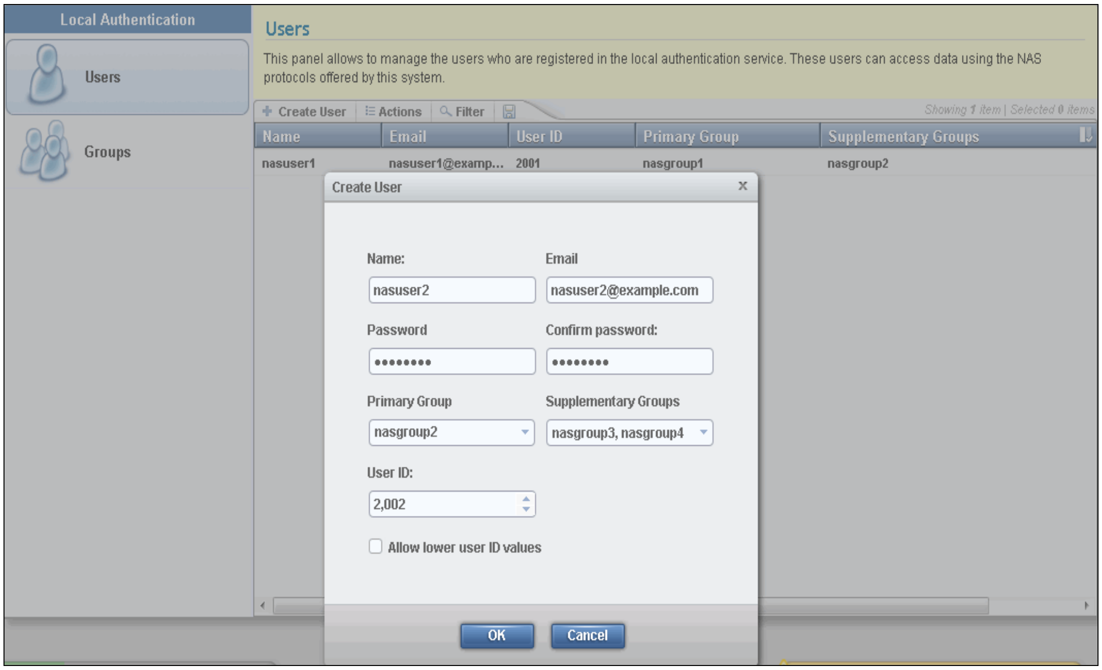
□ Allow lower group ID values

Figure 4-53 Enter new group details

As an authorized CLI/GUI administrative user, you can delete groups by clicking the **Actions** menu, as shown in Figure 4-52 on page 297, and selecting **Delete Group**. The local data access user group that is planned to be removed cannot be the primary group for any local data access user.

An authorized CLI and GUI administrative user can view existing users and create users from the window that is shown in Figure 4-52 on page 297. Click **Users** on the left and enter the user details in the window that opens, as shown in Figure 4-54 on page 299.

A user name must have a minimum of 3, and a maximum of 32, ASCII characters. You must also enter the password (which must be a minimum of eight characters in length) and select the primary group from the defined local data access user group. You can also enter the email address, and select up to 15 supplementary local data access groups for the specified user. Each specified supplementary group name in the list must already be defined as a local data access group. You can specify a UID or let the SONAS system automatically assign a UID. Automatically assigned UIDs start with 1024. It is a preferred practice to have UID equal or greater than 1024 to avoid conflicts with SONAS CLI and system users. You can create users with a UID lesser that 1024 by selecting **Allow lower user ID values**. The maximum UID value is 2,147,482,648.



Users

This panel allows to manage the users who are registered in the local authentication service. These users can access data using the NAS protocols offered by this system.

Name	Email	User ID	Primary Group	Supplementary Groups
nasuser1	nasuser1@examp...	2001	nasgroup1	nasgroup2

Create User

Name: nasuser2

Email nasuser2@example.com

Password ••••••••

Confirm password: ••••••••

Primary Group nasgroup2

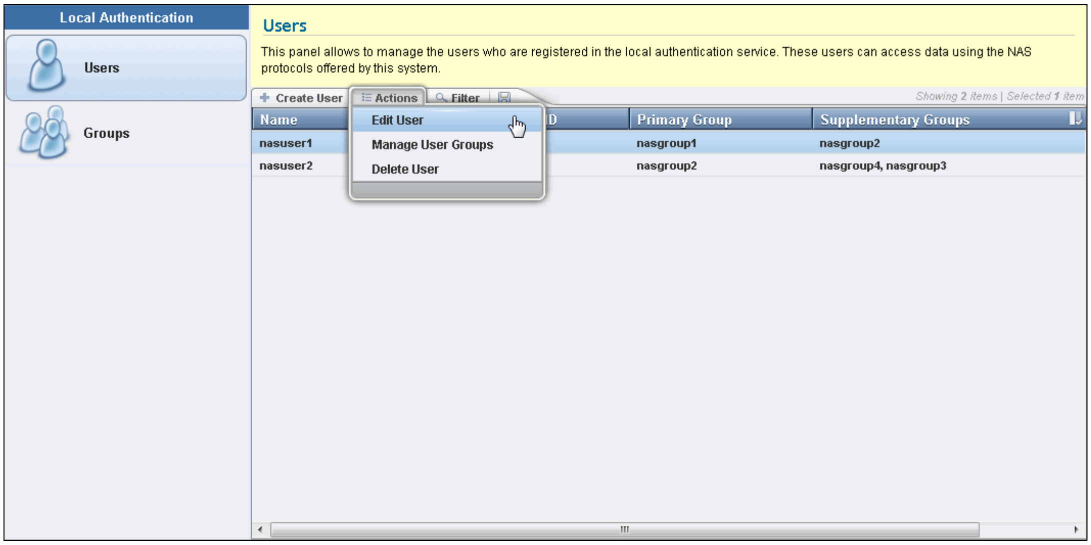
Supplementary Groups nasgroup3, nasgroup4

User ID: 2,002

☐ Allow lower user ID values

Figure 4-54 Add user information

To change the attributes of a local data access user on the local authentication server, select the user, click **Actions** and then **Edit user**, as shown in Figure 4-55.



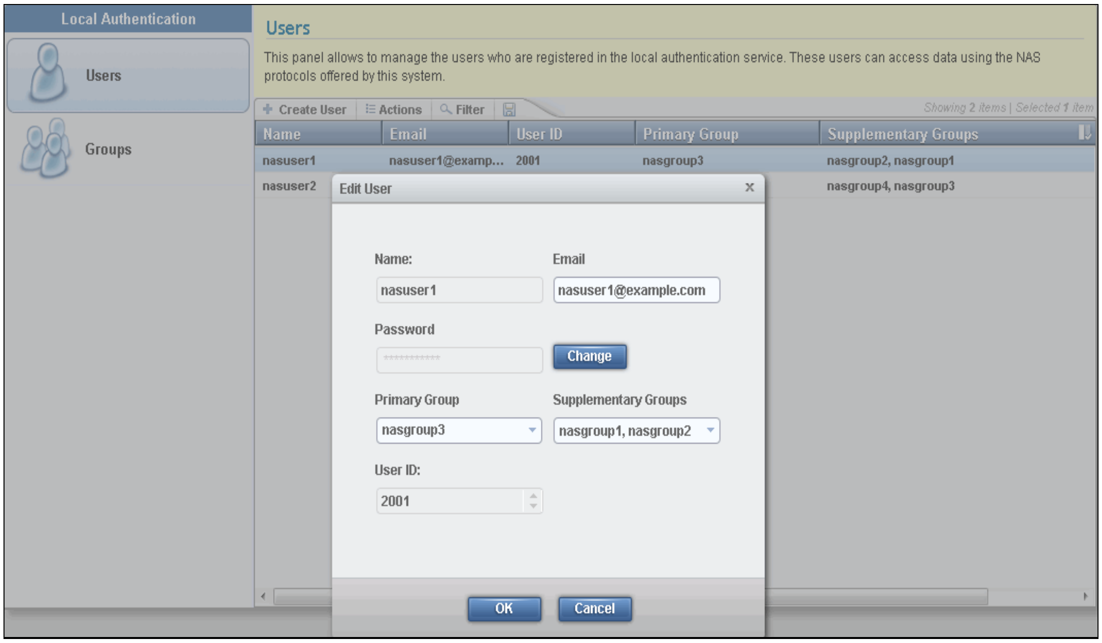
Users

This panel allows to manage the users who are registered in the local authentication service. These users can access data using the NAS protocols offered by this system.

Name	D	Primary Group	Supplementary Groups
nasuser1		nasgroup1	nasgroup2
nasuser2		nasgroup2	nasgroup4, nasgroup3

Figure 4-55 Modify user attributes

Make the changes in the window that opens, as shown in Figure 4-56. An authorized CLI/GUI administrative user can optionally change the password, which must be a minimum of eight characters in length. You can modify a primary group assignment. When you modify the current primary group, it automatically becomes part of the supplementary groups for the specified user. You can also modify the email address for the selected user. You can also add or remove supplementary groups by selecting or clearing the groups from the **Supplementary Groups** menu. The UID and user name cannot be changed.



Users

This panel allows to manage the users who are registered in the local authentication service. These users can access data using the NAS protocols offered by this system.

Showing 2 items | Selected 1 item

Name	Email	User ID	Primary Group	Supplementary Groups
nasuser1	nasuser1@examp...	2001	nasgroup3	nasgroup2, nasgroup1
nasuser2				nasgroup4, nasgroup3

Name: nasuser1

Email: nasuser1@example.com

Password: ************

Primary Group: nasgroup3

Supplementary Groups: nasgroup1, nasgroup2

User ID: 2001

Figure 4-56 Modify user attributes

Authorized CLI and GUI administrative users can also delete users by clicking the **Actions**, as shown in Figure 4-55 on page 299, and then selecting **Delete User**.

4.9.6 Managing local NAS users and groups with the CLI

SONAS V1.5.1 provides support for managing local NAS users and groups who can access the NAS shares by using the following commands:

`mknasgroup`	Create a local data access user group.
`lsnasgroup`	List all the local data access user groups and their attributes.
`rmnasgroup`	Delete a local data access user group.
`mknasuser`	Create a local data access user.
`chnasuser`	Modify the attributes of a local data access user.
`lsnasuser`	List all the local data access users and its attributes.
`rmnasuser`	Delete a local data access user.

For more information, see the corresponding man pages and the “Authentication using the local authentication server” topic in the IBM Knowledge Center found at the following address:

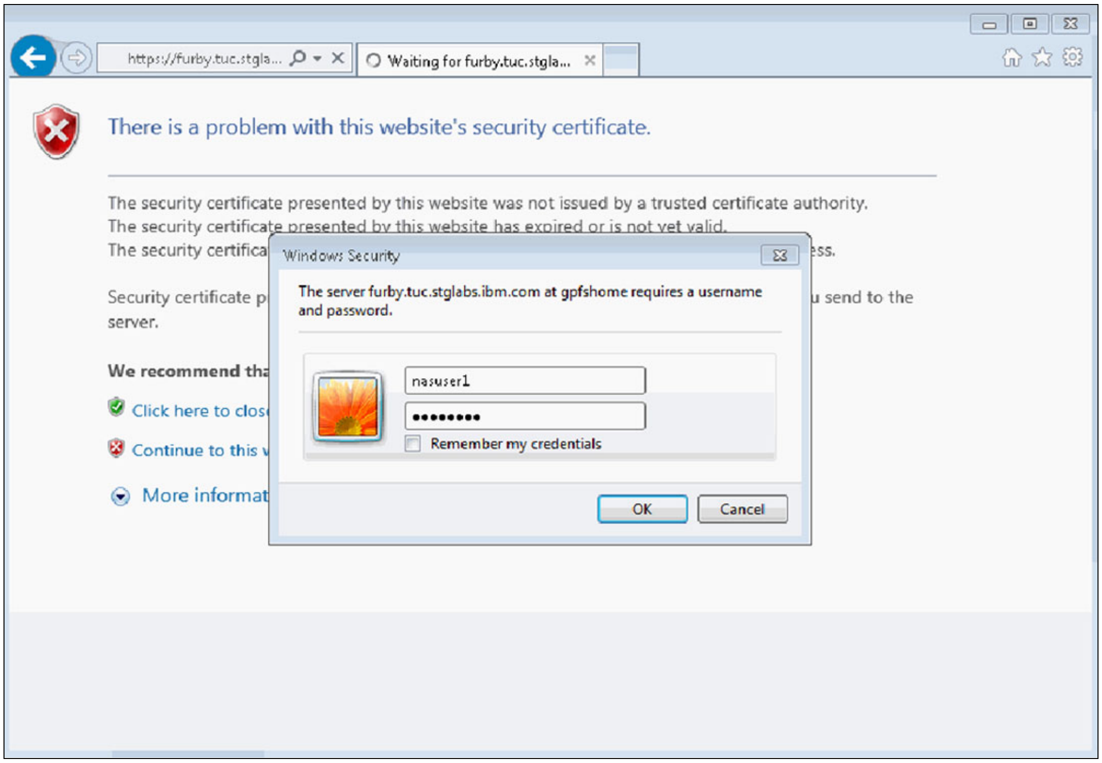
http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_authentication_OpenLDAPinternal.html

Changing the local data access user password

The CLI commands are only available to an authorized CLI and GUI administrative user. Other users can change passwords through a portal that is hosted by the system at the following address:

```
http://<Public_IP or System_Name of system/changepassword.html
```

Figure 4-57 shows changing the password for user nasuser1. You must enter the local data access user name and password.



There is a problem with this website's security certificate.

The security certificate presented by this website was not issued by a trusted certificate authority.
The security certificate presented by this website has expired or is not yet valid.
The security certifica ess.

Security certificate p u send to the server.

We recommend tha

Click here to clos

Continue to this w

More informat

Windows Security

The server furby.tuc.stglabs.ibm.com at gpfshome requires a username and password.

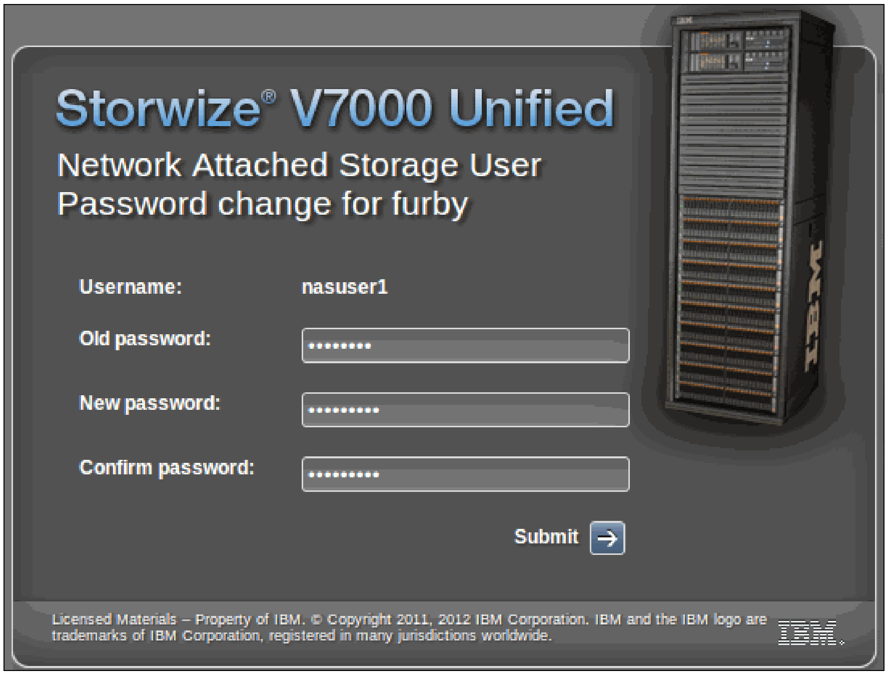
nasuser1

••••••••

□ Remember my credentials

Figure 4-57 Access the local data access user password change portal by using Internet Explorer

In the next window, which is shown in Figure 4-58, you can change the password by entering the current password and the new password and clicking **Submit**.



Storwize® V7000 Unified

Network Attached Storage User
Password change for furby

Username: nasuser1

Old password: ••••••••

New password: •••••••••

Confirm password: •••••••••

Licensed Materials – Property of IBM. © Copyright 2011, 2012 IBM Corporation. IBM and the IBM logo are trademarks of IBM Corporation, registered in many jurisdictions worldwide.

Figure 4-58   Change the local data access user password

Preferred practices for using local authentication

Here are preferred practices for using local authentication:

- Specify the UID and GID when you create local data access users and groups. For increased control over the UID and GID, use the CLI command options to specify the ID values instead of allowing the SONAS system to generate ID values.
- Define local data access users and groups for NFS access from UNIX and Linux hosts. To prevent UID or GID conflicts, ensure that user and group identities on the host system are consistent with IDs on Storwize V7000 Unified systems. Host IDs can be displayed by using the UNIX **id** command, as shown in Example 4-46.

Example 4-46   Display host IDs

```
id username_on_AIX_or_LINUX_host
```

4.10 Listing the authentication method that is configured on SONAS

You can use the **lsauth** command to list the authentication settings of the cluster.

The **lsauth** command retrieves the authentication settings of the cluster from the database and returns a map in either a human-readable format or in a format that can be parsed. For command help, see Example 4-47 on page 303.

Example 4-47 Help for the lsauth command

```
$lsauth     --help
usage: lsauth  [-r] [-Y] [-c < clusterID | clusterName >]
Lists authentication settings of a cluster.

Parameter     Description
-r, --refresh Refreshes configuration.
-Y            Shows parseable output.
-c, --cluster The cluster scope for this command
```

Example 4-48 shows an example of **lsauth** output for a cluster. In this example, the cluster is configured with AD.

Example 4-48 The lsauth command showing the authentication configuration of an existing cluster

```
$lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
domain = SONASDM
idMappingMethod = auto
clusterName = fstsonas01
userName = Administrator
adHost = SONASPB11.sonasdm.storage.tucson.ibm.com
passwordServer = *
realm = sonasdm.storage.tucson.ibm.com
EFSSG1000I The command completed successfully.
```

4.11 Checking the authentication setting for the cluster

The **chkauth** command checks the authentication settings of a cluster.

The command checks the authentication settings on the cluster by using following features:

- Node configuration
- Node synchronization
- Availability of the server
- Extra read operations on the servers
- User authentication
- User information

Example 4-49 shows the **chkauth** command help.

Example 4-49 Help for the chkauth command

```
$chkauth --help
usage: chkauth  [--nodesInSync <nodesInSync>] [--verifyAllNodeConf
<verifyAllNodeConf>] [--ping <ping>] [-t <checkSecret>] [-i <userInfo>] [-a
<authenticateUser>] [-u <userName>] [-p <userPwd>] [--validateEncryption
<validateEncryption>] [--ldapPassword <ldapBindPassword>] [--verifyNISServer
<validateServer>] [--nisDomain <nisDomain>] [--serverList <serverList>] [-Y] [-c <
clusterID | clusterName >]
Checks authentication settings of a cluster.

Parameter                Description
```

```
    --nodesInSync        Designates the nodes in sync.
    --verifyAllNodeConf  Verifies configuration on all interface nodes
    --ping               Pings the authentication server.
-t, --checkSecret        Performs actions with directory service.
-i, --userInfo           Fetches user information.
-a, --authenticateUser   Authenticates the user.
-u, --userName           Specifies the user name to be used to log in.
-p, --userPassword       Specifies the password to be used to log in.
    --validateEncryption Validates SSL certificates and Kerberos keytab file
    --ldapPassword       Designates the LDAP bind password.
    --verifyNISServer    Verifies NIS Server
    --nisDomain          Provides the NIS domain server.
    --serverList         Provides the NIS server list.
-Y                       Shows the parseable output.
-c, --cluster            The cluster scope for this command
```

If none of the options are specified, node synchronization is called by default. See Example 4-50. If more than one option is specified, an exception error is displayed.

Example 4-50 Example chkauth command with the default option

```
# chkauth
ALL NODES IN CLUSTER ARE IN SYNC WITH EACH OTHER
EFSSG1000I The command completed successfully.
```

4.12 Cleaning up authentication

You can clean up the authentication by running the **cleanupauth** command. The command cleans up the authentication and ID map configuration. You must provide confirmation before the command runs. By default, only the authentication configuration is deleted. After the command runs successfully, the system is left with no authentication configuration, and no services (FTP, SCP, HTTP, CIFS, or NFS) are available. See Example 4-51.

Example 4-51 CLI command to clean up authentication that is configured

```
$cleanupauth --help
usage: cleanupauth  [--idMapDelete] [--force] [-c < clusterID | clusterName >]
Clears authentication configuration settings from the system.

Parameter         Description
    --idMapDelete Deletes the current ID mappings from the database used for
storing IDs.
    --force       Do not prompt for manual confirmation.
-c, --cluster     The cluster scope for this command
```

If the **--idMapDelete** option is used, only the ID Mapping information is deleted.

To delete the configuration information from the system, run **cleanupauth** without this option first. Then, run **cleanupauth** again with the **--idMapDelete** option to delete the ID mappings.

This method ensures that both the configuration information of previously configured authentication and all ID mappings are deleted.

Tip: Only security administrators can run this command.

Important: Running the `cleanupauth` command cleans up the configuration. Use caution when you run this command. If there are users from other domains accessing SONAS, they might lose access.

4.13 Working with the ID map cache

All the UIDs and GIDs, whether generated automatically or in SFU or NIS, are cached. For every positive login of a user, the UID and GID are stored for seven days. For negative login, where login to the cluster fails, the UID and GID are cached for two minutes.

It is not advised to change the UID or GID after it is generated. If the administrator must do it, run the `rmidmapcacheentry` command.

The `rmidmapcacheentry` command removes the ID map cache entry of a specified user or group. The command affects the data access for only those NAS users or groups whose name or SID are specified in the `--name` or `--sid` options. The command does not change file ownerships, update quota reports (the quota is still accounted against the original UID or GID), or ACLs.

The user must manually traverse the file system and correct ACLs and file ownerships after such an operation.

The command checks whether the currently connected user is affected by this change (has the UID/GID in question) and terminates the CIFS connection if so, which disconnects the user. Termination happens immediately regardless of whether files are still open for reading or writing.

This command must be run when the UID or GID of a user or group is changed in the external server (SFU or NIS). This command can be run only by the security administrator.

Important: It is a preferred practice not to change the UID or GID set for users and groups in an AD + SFU or AD + NIS environment. These values are cached on SONAS and are referred to until it is flushed out. Lookup might return old UID or GID values, so users and groups might face access denied issues. The cache flushes every seven days for a positive UID and GID.

The administrator must run `rmidmapcacheentry` to flush a cache if a UID or GID must be modified. After this command is run, access is denied to data that has ACLs with the old UID and GID. Ownerships and ACLs for this UID or GID must be reapplied by the administrator.

Therefore, changing a UID or GID must be planned carefully and is preferably done before data is put on the SONAS.

4.14 Working with SONAS authorization

SONAS authorization checks to see whether the authenticated user has the access permissions to access SONAS and its data.

There can be two types of users who need to access SONAS. The first type is users who need to access SONAS to manage the system. The second type is the users and groups who need to access the data that is stored on SONAS.

4.14.1 SONAS administrative users

SONAS administrative users are local users on the SONAS appliance. They typically access the SONAS to do administrative tasks such as, but not limited to, creating file systems and file sets, setting quotas, creating exports (shares), managing notifications, and monitoring the system.

The SONAS administrator has no access to the data that is stored on SONAS. The administrator works on a restricted shell, which allows only SONAS commands to be run.

> **Note:** With SONAS V1.5.1, support for adding remote users (data access users from AD or LDAP) as administrators was added.

SONAS administrators can have different profiles or roles. For more information, see the “Predefined user role definitions” topic in the IBM Knowledge Center found at the following website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_user_roles.html?lang=en

4.14.2 SONAS client users

Client users access the data that is stored on the system. They access SONAS by using CIFS or NFS. They are authenticated with an external authentication server such as AD or LDAP.

Clients can access data on SONAS through an export. They do not have direct access to the data that is stored on SONAS. Also, these users can access the data if they have the correct permissions to the data. Exports must first be created with the correct ACLs and inheritance that is configured to manage their security.

ACLs can be modified from the Windows client for a CIFS export. For NFS only, clients must be added to the export at creation time or later to provide access to the users. The management GUI also allows adding ACLs to the shares that are created.

5

IBM Active Cloud Engine

IBM Active Cloud Engine (ACE), the core of IBM SONAS and IBM Storwize V7000 Unified storage, provides users with the ability to manage files efficiently, locate the data of interest rapidly, and get that data to where it is needed seamlessly.

This chapter provides an overview of ACE and a detailed description of the Policy, Integrated Lifecycle Management, and Remote Caching components.

This chapter describes the following topics:

- Active Cloud Engine: Overview
- Setting up remote caching
- Remote caching administration tasks
- SONAS policy management
- Creating and managing policies

5.1 Active Cloud Engine: Overview

The ACE is an advanced form of multiple site replication that is designed by IBM Research. ACE is designed to address an IT world where cloud storage is the goal, yet where you must face both islands of storage and islands of clouds. Different types of cloud implementations cannot dynamically exchange data, and moving data between cloud vendors and providers can be equally difficult.

The ACE capability is designed to address these issues by providing a means to exchange data in a cloud storage environment dynamically, immediately, and in an on-demand manner, between multiple geographically dispersed locations and even between different cloud storage providers. ACE does it by extending the SONAS capability for a centrally auto-managed, single highly scalable high-performance namespace, to a truly distributed worldwide, geographically dispersed global namespace.

In Figure 5-1, the SONAS users see the appearance of a single global NAS storage namespace, even though the namespace is physically distributed among multiple geographically dispersed locations. This common global namespace is achieved by the ACE enabled systems that are working together.

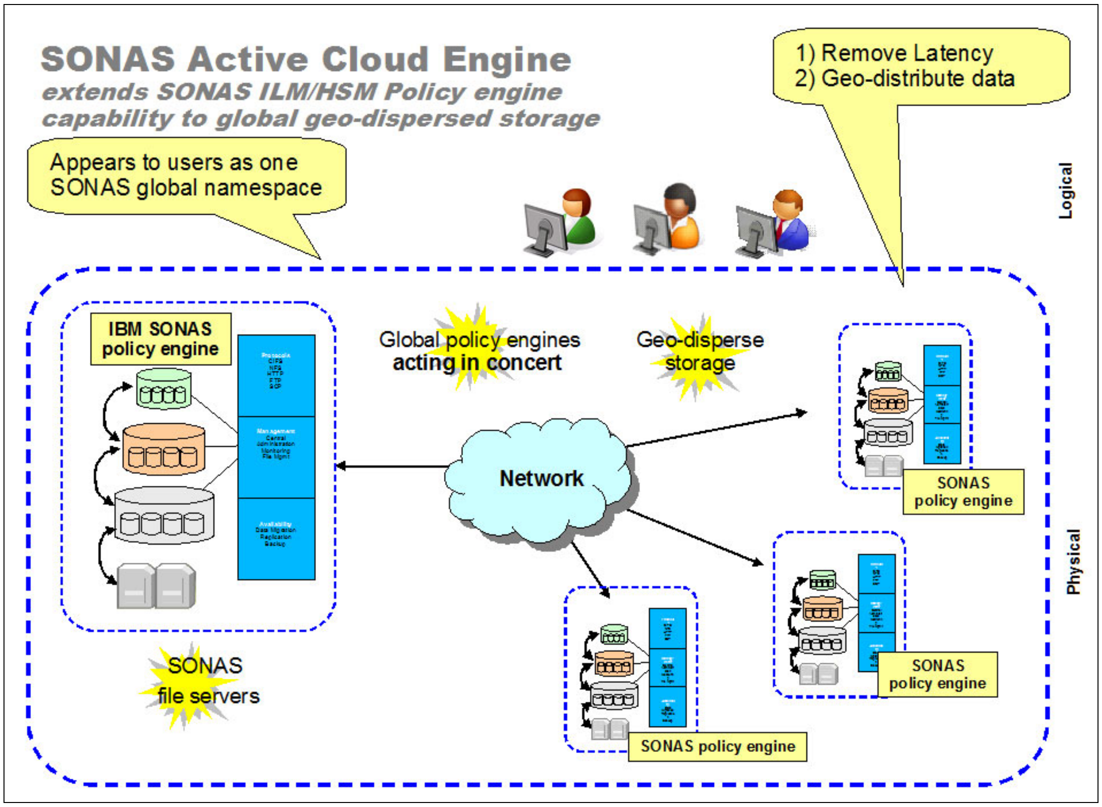
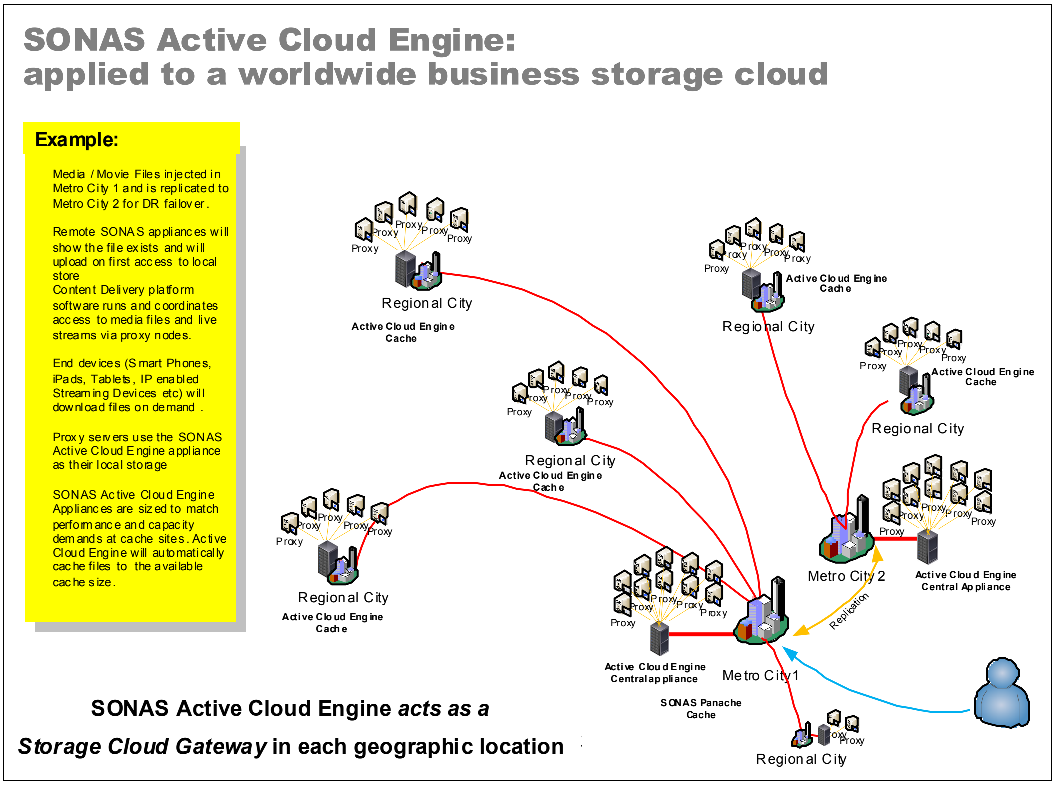


Figure 5-1 Manage SONAS storage centrally and automatically on a global basis

The number of Active Cloud Engine sites is not fixed or limited; you can assemble as many ACE sites together as you want. ACE achieves a common global geo-graphically dispersed namespace and enables efficient multi-site global file management, as follows:

- Automatically caching files at remote sites, providing local latency reduction and higher availability while reducing bandwidth costs
- Providing automatic pre-fetching of complete or partial files, that is, *on-demand pull*
- Providing *capability to push* files and updates to either the remote sites, or back to the home sites, according to policy
- Virtualizing all of the Active Cloud Engine sites to a single view, and all of the physically separated SONAS, GPFS, or Storwize V7000 Unified system, thus providing a single global namespace

The power of the ACE can be understood by examining its usage when applied to a worldwide business storage cloud, as shown in Figure 5-2.



**SONAS Active Cloud Engine:
applied to a worldwide business storage cloud**

Example:

Media / Movie Files injected in Metro City 1 and is replicated to Metro City 2 for DR failover.

Remote SONAS appliances will show the file exists and will upload on first access to local store
Content Delivery platform software runs and coordinates access to media files and live streams via proxy nodes.

End devices (Smart Phones, iPads, Tablets, IP enabled Streaming Devices etc) will download files on demand .

Proxy servers use the SONAS Active Cloud Engine appliance as their local storage

SONAS Active Cloud Engine Appliances are sized to match performance and capacity demands at cache sites. Active Cloud Engine will automatically cache files to the available cache size.

SONAS Active Cloud Engine *acts as a*

***Storage Cloud Gateway* in each geographic location**

Figure 5-2 Active Cloud Engine deployed in a worldwide business storage cloud

The Active Cloud Engine appears to be a standard NFS file server at each site; any NFS clients, proxies, or users in that site can access the Active Cloud Engine storage through NFS.

More importantly, each ACE also acts as a *gateway* to *all other* Active Cloud Engine sites, presenting a common view and common global namespace of all other sites in the cloud storage. With proper authority, all users at an individual site can see the entire single global cloud storage namespace across all sites.

Among this cloud storage Active Cloud Engine collection, files are moved automatically or on demand, and cached at the Active Cloud Engine in the local site. The central locations can also request data that is at any of the remote sites, or data can be pushed from any site to any other sites.

Interoperability is assured because each site sees the ACE gateway as a standard NFS server. Therefore, other NAS storage, proxies, or NAS clients at any site can be of *any* vendor or cloud storage deployment that supports standard NFS.

By fusing standard NFS protocols with IBM SONAS central-policy-managed storage capabilities, ACE bridges not only multiple geo-dispersed SONAS sites, but also provides for dynamic, automatic, and policy-managed cloud storage data movement between multiple different cloud storage implementations or vendors at each site.

ACE provides a unique, high-performance capability to auto-align geographically dispersed or local cloud storage, and this storage is automatically managed and synchronized across all sites by using ACE capabilities. ACE provides the ability to make a multi-site and multi-company cloud storage an *active* reality.

The next section describes how ACE implements these capabilities.

5.1.1 Active Cloud Engine: Global in more detail

ACE provides “home clusters” and “cache clusters”. Cache clusters act as front-end, wide area network (WAN) cache access points, which can transparently access the *entire* collection of ACE systems.

Active Cloud Engine consists of a client/server architecture:

- The home cluster *(server)* provides the primary storage for data.
- The cache cluster *(clients)* can read or write cache data that is exported to them.

These concepts are shown in Figure 5-3 on page 311.

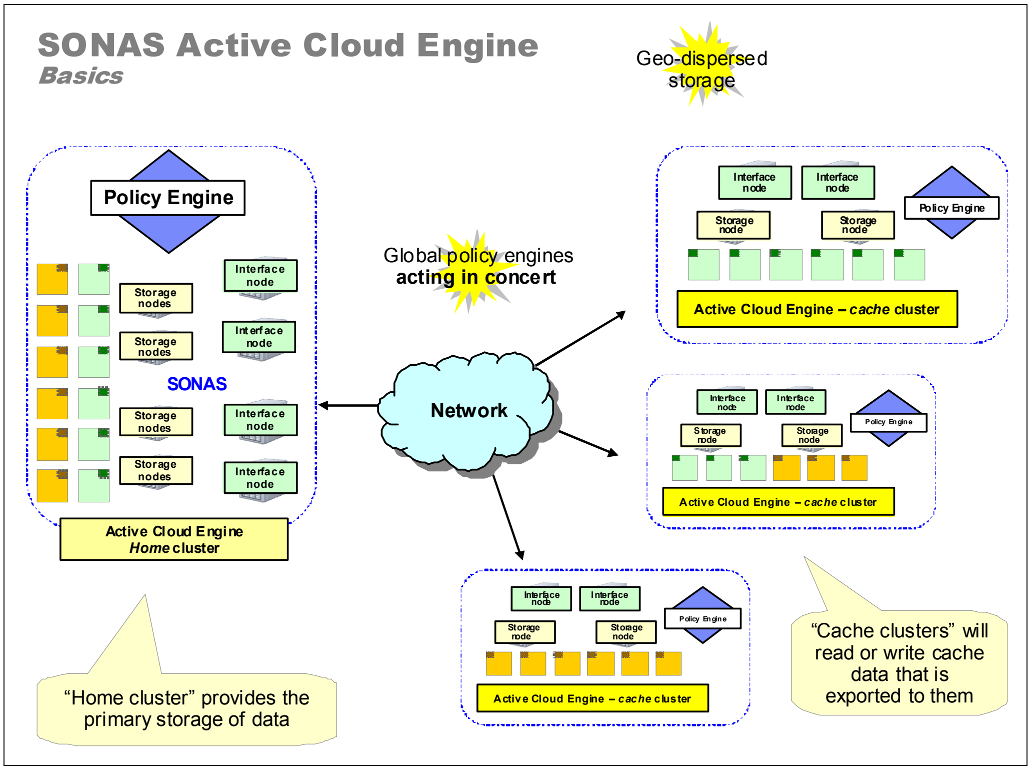

SONAS Active Cloud Engine
Basics
Geo-dispersed storage
Policy Engine
Storage nodes
Interface node
SONAS
Active Cloud Engine Home cluster
Global policy engines acting in concert
Network
Active Cloud Engine – cache cluster
Storage node
Interface node
Policy Engine
"Home cluster" provides the primary storage of data
"Cache clusters" will read or write cache data that is exported to them

Figure 5-3 Active Cloud Engine basics

The following terminology applies to the Active Cloud Engine:

- *Home*: The home cluster exports a file set by using the standard NFS protocol over one or more Interface nodes as defined by policies:
  - Users are verified as having permission to access the file from the home cluster.
  - A home is any NFS mount point in the file system. Home designates which site is the *owner* of the data in a cache relationship. The *writing* of the data might be at a different location in the ACE configuration. However, the home cluster is the owner of the relationships.
  - The NFS mount point can be any SONAS file system or file set in a file system. At any point in time, there is one home location. You can have virtually an unlimited number of cache relationships, and home locations can be changed or moved in coordination with the cache relationships.
- *Cache*: A cache cluster is a SONAS, GPFS, or Storwize V7000 Unified file set. The cache cluster does most of the work in an ACE environment, including the following tasks:
  - Interface nodes in the cache cluster communicate with the home clusters.
  - Cache cluster Interface nodes communicate with the home clusters by using standard NFS.
  - The cache cluster presents the virtualized image of *all* authorized Active Cloud Engine namespaces to the local user as a single virtualized global namespace view.

These concepts are illustrated in Figure 5-4.

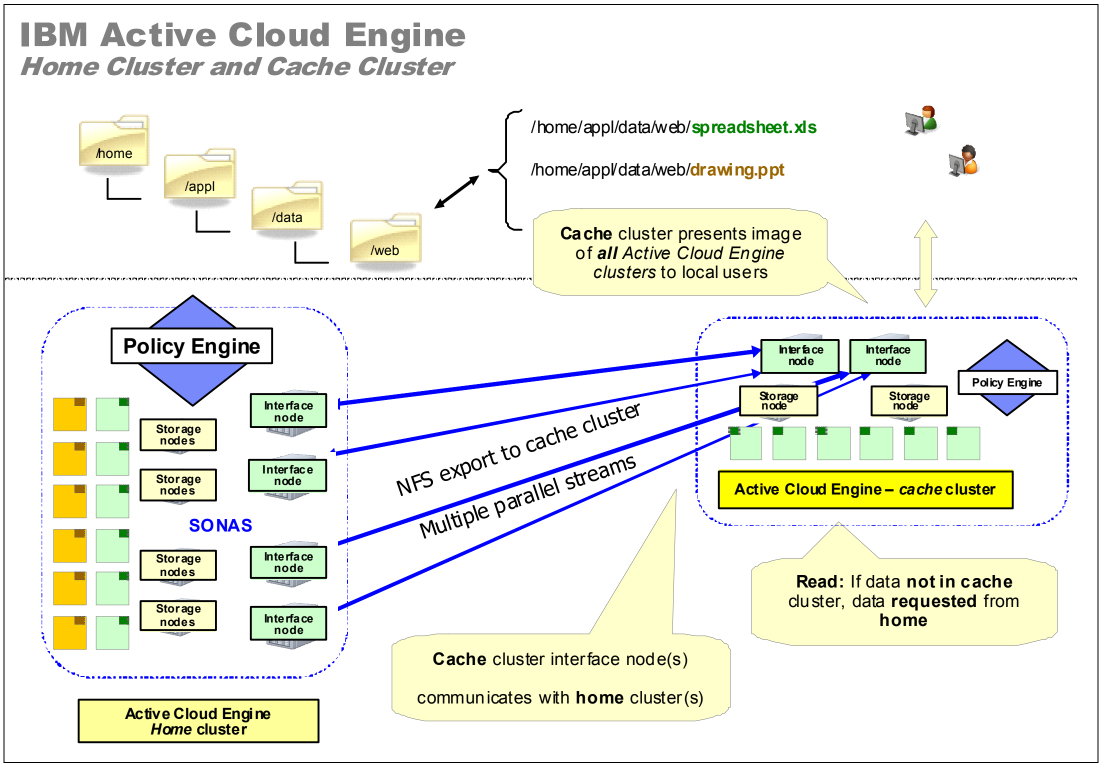

IBM Active Cloud Engine
Home Cluster and Cache Cluster
/home
/appl
/data
/web
/home/appl/data/web/spreadsheet.xls
/home/appl/data/web/drawing.ppt
Cache cluster presents image of all Active Cloud Engine clusters to local users
Policy Engine
Storage nodes
Storage nodes
Storage nodes
Storage nodes
Interface node
Interface node
Interface node
Interface node
SONAS
Active Cloud Engine Home cluster
NFS export to cache cluster
Multiple parallel streams
Interface node
Interface node
Storage node
Storage node
Policy Engine
Active Cloud Engine – cache cluster
Read: If data not in cache cluster, data requested from home
Cache cluster interface node(s) communicates with home cluster(s)

Figure 5-4 Active Cloud Engine - home cluster and cache cluster

When there is a read request at the local cache, existing file data at home is pulled into the cache on demand. Multiple Interface nodes and multiple NFS connections are used to provide high performance.

If the cache is in write mode (for any particular file set, a single writer across the global namespace is supported), new data is written to the local cache. This data is asynchronously pushed back to home, while still maintaining a local copy in cache.

As of ACE V1.4, four modes are allowed for an individual Active Cloud Engine cache cluster file set. Different file sets can have different modes:

- *Read Only*: Cache cluster can read data, but no write data changes are allowed. This mode is useful for data broadcast, for latency and bandwidth reduction, and for localizing data to the local users.
- *Single Writer*: The cache cluster is defined as the only writing location in the cloud storage global namespace. The home cluster receives the writes, does not change the data, and the home cluster acts as central repository so that multiple other cache clusters can read the data. It is useful for remote importation of data, transmission to home, and then reflecting that data to all other authorized points in the cloud storage.

- *Local Update*: Data is read or pushed from home and cached at the local cache cluster. Changes are allowed locally, but changes are not pushed back to home. After data is locally changed, the home cluster and cache cluster relationship is marked to indicate that cache and home are no longer in sync for this file.
- *Independent Writer*: Multiple cache clusters can write to a single home cluster, enabling multiple sites to be writers for the same file set if the cache writes to different files. The cache cluster sites perform revalidation periodically and pull the new data from home, which includes the changes that are done by other peer sites that were pushed to home. If multiple caches write to same file, the file's data is updated on home as each cache pushes them back. The sequence of updates is indeterministic. Independent Writer mode does not ensure file locking or write ordering from cache to home.

Read Only and Single Writer caches can be changed (at the level of the file set) to any other mode, with appropriate coordination within the Active Cloud Engine environment.

Single Writer and Independent Writer modes cannot coexist, so a remote caching source cannot have both Single Writer and Independent Writer clients. Independent Writer mode does not support peer snapshot or manual resynchronization.

The fundamental ACE implementation requirement is that a unique high-level directory naming schema is in place, thus allowing all Active Cloud Engine sites to identify each ACE site by a unique high-level directory tree name. An example of this naming schema is shown in Figure 5-5.

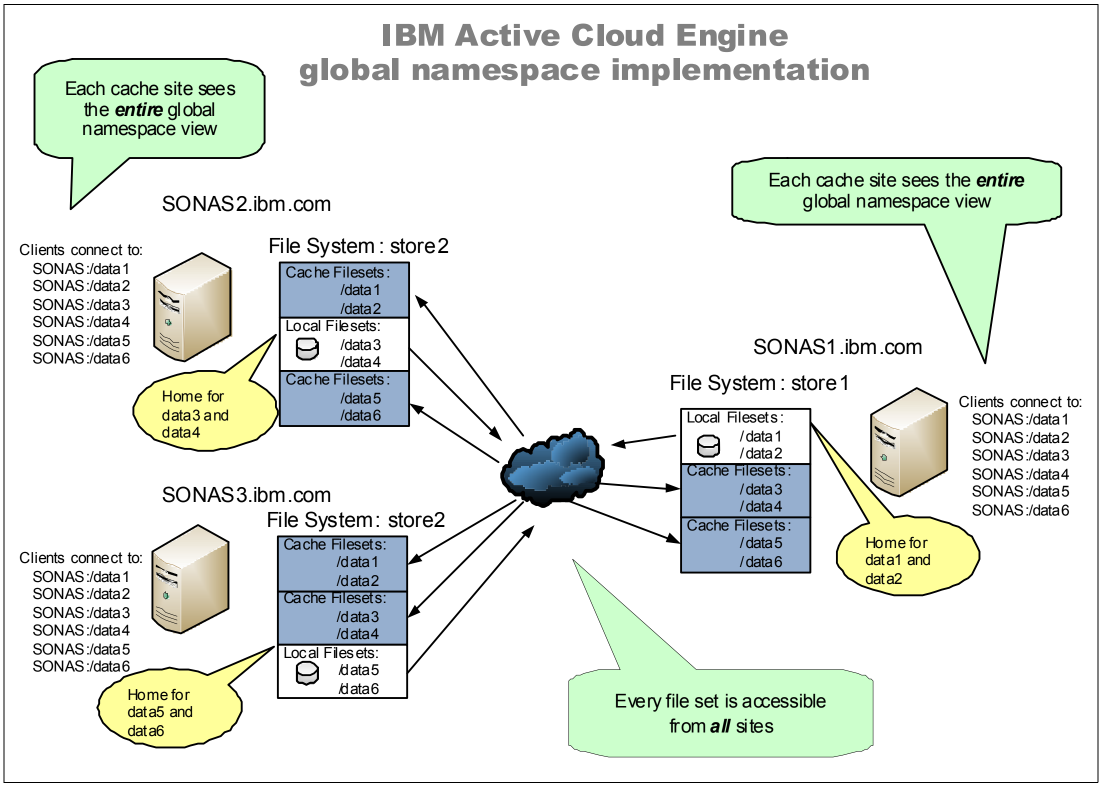


Figure 5-5 Active Cloud Engine - sample directory structure implementation

After the unique directory name is identified, it becomes the mount point for all of the other ACE sites. As each remote cluster is mounted, the cache cluster mode (Read Only, Single Writer, or Local Update) is specified and validated. After the mode is validated and mounted, all of the ACE functions work.

ACE is designed to use standard directory tree structures as the unique identifiers for each ACE site to minimize integration effort and maximize time to implementation and time to value.

5.1.2 Global Active Cloud Engine: Summary

SONAS architectural capabilities for scale-out parallelism, high performance, and local ILM/HSM main storage policy management are extended to multiple geographic locations and to cloud storage by the global Active Cloud Engine capabilities.

The global Active Cloud Engine is flexible in capability and configuration, highly performant in nature, and provides a new technology capability for centrally managing and deploying geographically dispersed cloud storage. Cascading is allowed and encouraged.

Figure 5-6 shows an example of an advanced ACE configuration, which illustrates the type of flexibility that is possible.

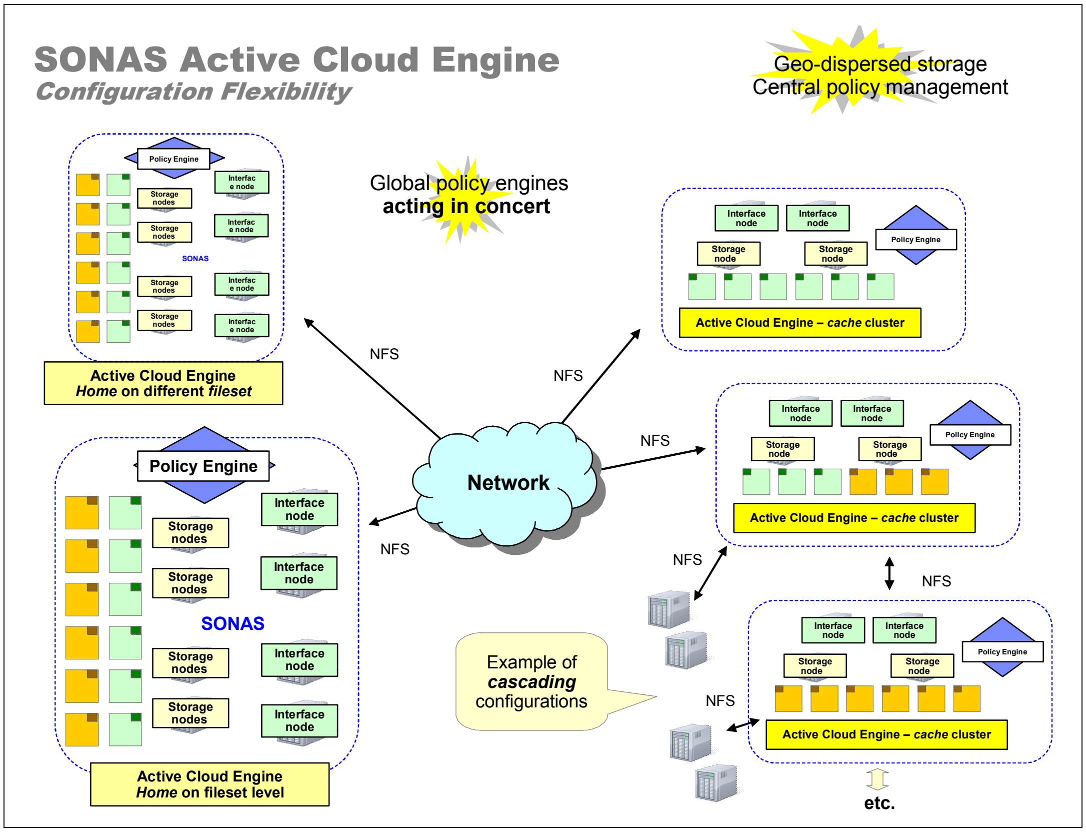


Figure 5-6 Active Cloud Engine configuration flexibility

The use of NFS as the standard interface and mount point provides interoperability, extending the Active Cloud Engine capability beyond SONAS, GPFS, and Storwize V7000 Unified to any NFS-capable storage, clients, and cloud storage implementation.

The creative applications for ACE in today's geographically dispersed, multiple-organization cloud storage environments is limited only by your business imagination. ACE brings a new paradigm of possibilities to cloud storage.

5.2 Setting up remote caching

This section provides information about configuring a remote caching relationship in SONAS systems.

Remote caching can be configured by using the CLI and the GUI. SSH logon is needed across the home and the cache systems. Figure 5-7 shows a remote caching environment.

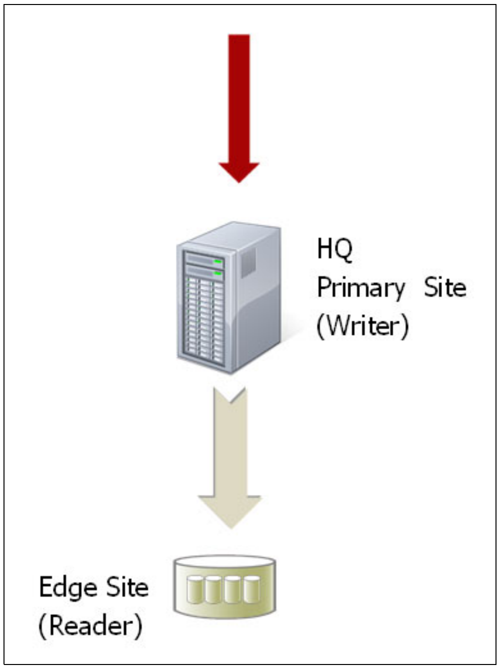


Figure 5-7 Remote caching example

5.2.1 Remote caching setup requirements

This section describes the requirements for setting up remote caching.

Ports must be open in both directions for the public, management, and service IP addresses of the home and the cache systems. The following ports are used for remote caching communication between the home and the cache.

- 22 for SSH
- 111 for PortMapper
- 1081 for HTTP
- 800, 2049, 32765, 32767, 32768, and 32769 for NFS

The home system's Management node and Interface nodes must communicate with the cache systems Management node and Interface nodes over the user network.

The time on home and cache systems should be synchronized. The preferred practice is to synchronize them to the same NTP server.

Consistent authentication mappings between the home and cache systems are required for proper authentication control and management of the files that are transferred through remote caching between the home and the cache systems. All systems that are participating must be managed by a common Active Directory with Services for UNIX (SFU), or by LDAP. An Active Directory server without the Services for UNIX extension is not supported.

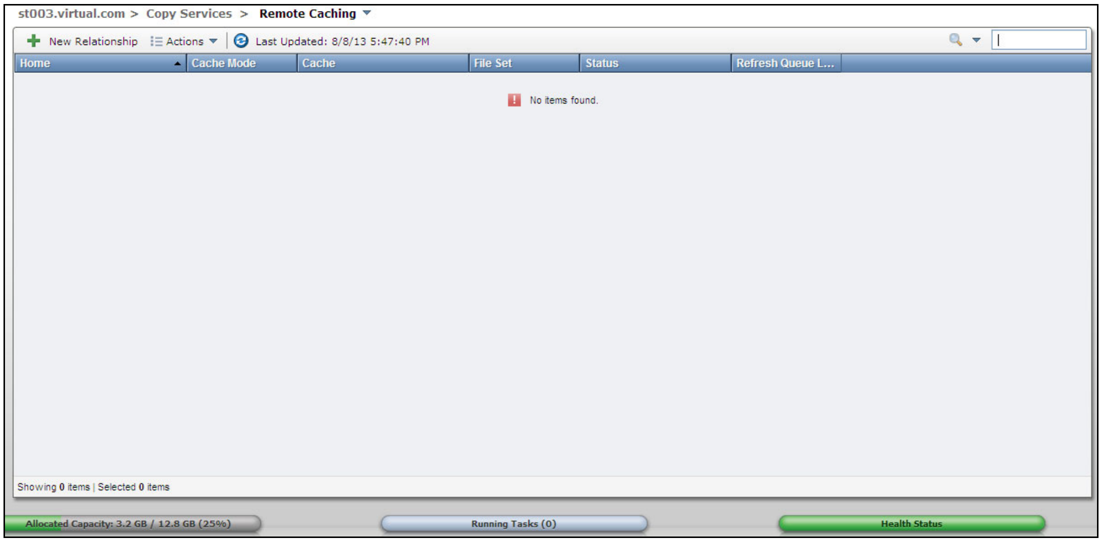
5.2.2 Setting up remote caching by using the GUI

This section describes a scenario to establish a remote caching relationship between two SONAS systems. Configuration is done only on the home system. On the home system, a file set named `homefileset` is created. On the cache system, the file set is named `cachedhomefileset`. Here are the IP addresses of the systems:

- Home system IP: `10.0.0.30`
- Cache system IP: `10.0.0.10`

To set up remote caching, complete the following steps:

1. Log in to the home system and click **Copy Services** → **Remote Caching**, as shown in Figure 5-8.



st003.virtual.com > Copy Services > Remote Caching

Home	Cache Mode	Cache	File Set	Status	Refresh Queue L...

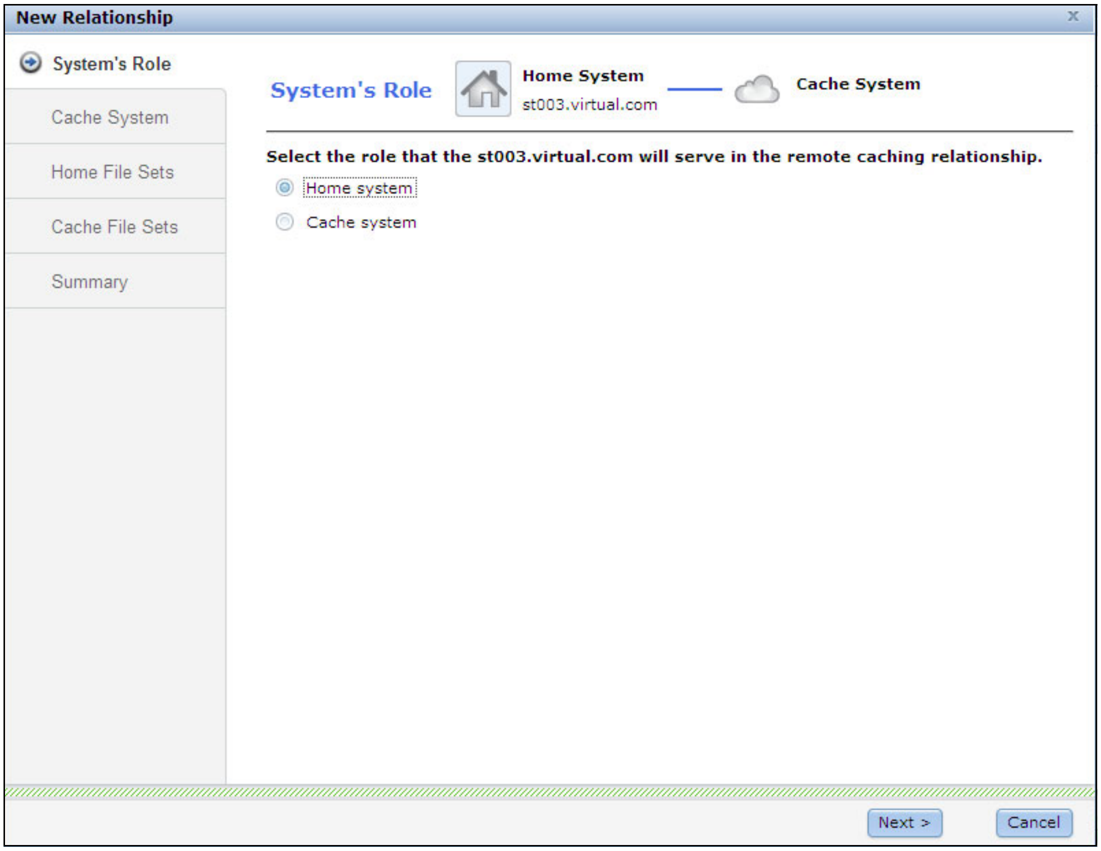
No items found.

Showing 0 items | Selected 0 items

Allocated Capacity: 3.2 GB / 12.8 GB (25%)

Figure 5-8 Remote Caching view

2. Click **New Relationship**. A window opens, as shown in Figure 5-9 on page 317. Select the home system from the list and click **Next** to continue.



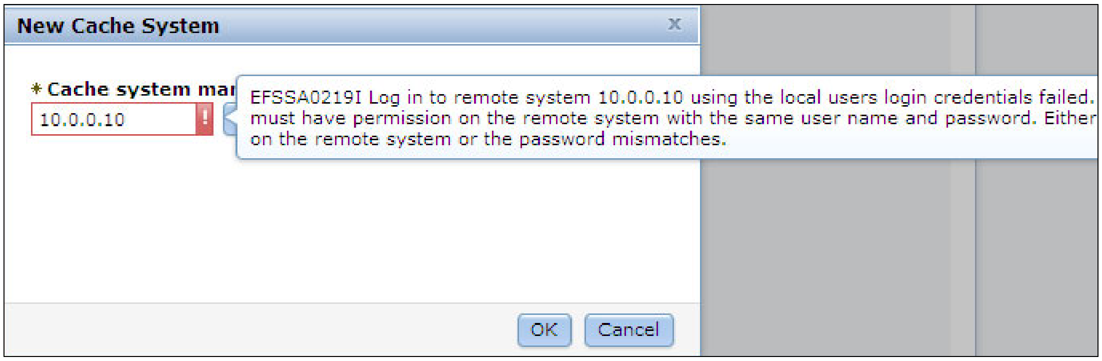
Select the role that the st003.virtual.com will serve in the remote caching relationship.

● Home system

○ Cache system

Figure 5-9   New relationship setup window

3. On the next window, click **+New Cache System**. Enter the cache systems's IP address and click **Find**, as shown in Figure 5-10.



* Cache system mar

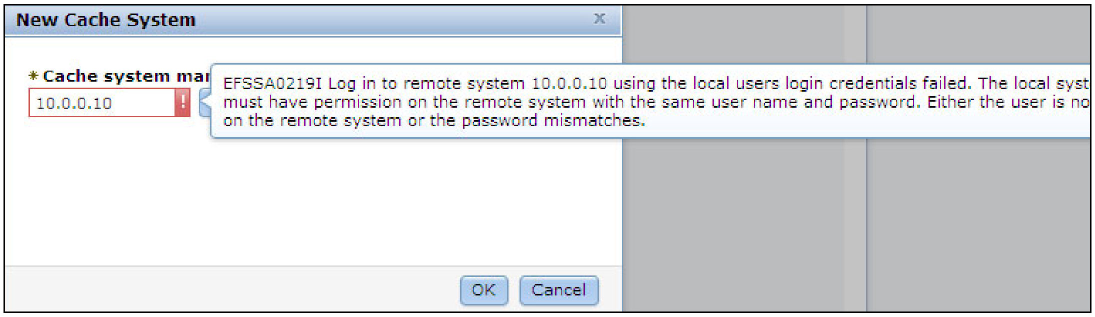
10.0.0.10

EFSSA0219I Log in to remote system 10.0.0.10 using the local users login credentials failed.
must have permission on the remote system with the same user name and password. Either
on the remote system or the password mismatches.

Figure 5-10   Add New Cache System window

The user name and password that are used to log in to the management GUI to create the remote caching relationship *must* exist in the related remote systems. For example, if you log in as the Administrator user with the user name *admin* and password *admin* to run the setup wizard, an Administrator user with the user name *admin* and password *admin* must exist in the system that supports the related cache file set. If you must create the home system and the cache file set with different user settings, use the CLI commands to create the remote caching relationship.

Figure 5-11 shows the error message that is received when the user name and password that are used to log in to the management GUI to create the remote caching relationship do not exist in the related remote systems.

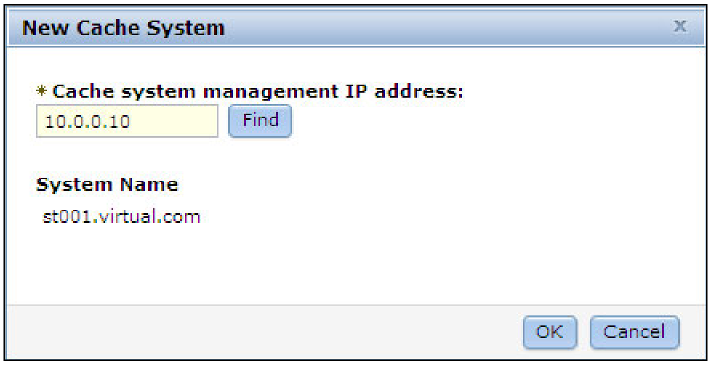


10.0.0.10

EFSSA0219I Log in to remote system 10.0.0.10 using the local users login credentials failed. The local syst must have permission on the remote system with the same user name and password. Either the user is no on the remote system or the password mismatches.

Figure 5-11 Error message that indicates that local and remote user login credentials do not match

4. After a successful search, the system name of the cache system is displayed, as shown in Figure 5-12. Click **OK** to continue**.**

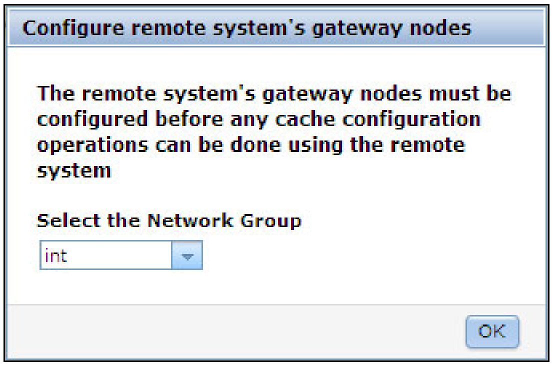


* Cache system management IP address: 10.0.0.10

System Name
st001.virtual.com

Figure 5-12 New Cache System window with System Name displayed

5. To enable remote caching on a SONAS system, configure the specified Interface nodes in the system to function as the caching gateway nodes that exchange data with other systems (see Figure 5-13). In the window that opens, select the network group that includes the caching nodes on the remote system and press **OK** to continue.



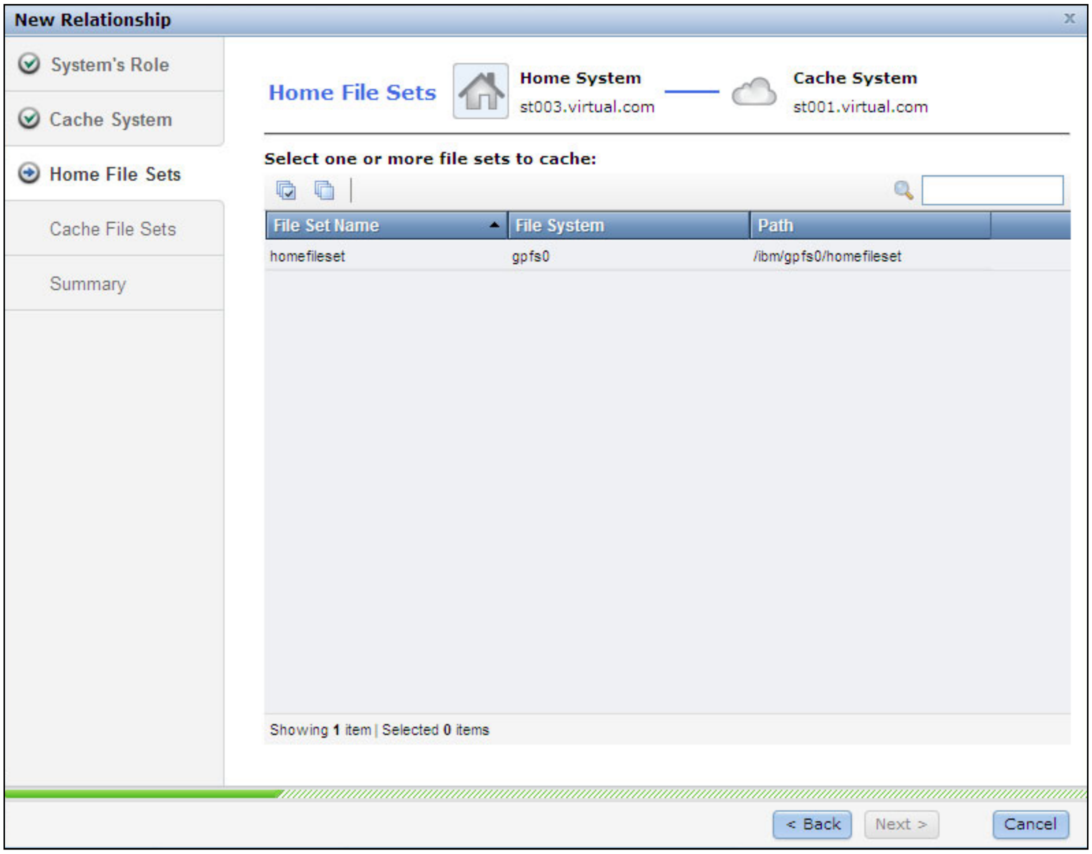
The remote system's gateway nodes must be configured before any cache configuration operations can be done using the remote system

Select the Network Group: int

Figure 5-13 Configure remote system's gateway nodes window

Tip: If a network group is used to configure the caching gateway nodes, the nodes that are members of that network group at the time that the command is submitted are configured as caching gateway nodes. If that network group membership changes after the nodes were configured as caching gateway nodes, the network group membership change does not affect which nodes are configured as caching gateway nodes.

6. The New Relationship window (Figure 5-14) opens. Select one or more file sets for caching and click **Next**.



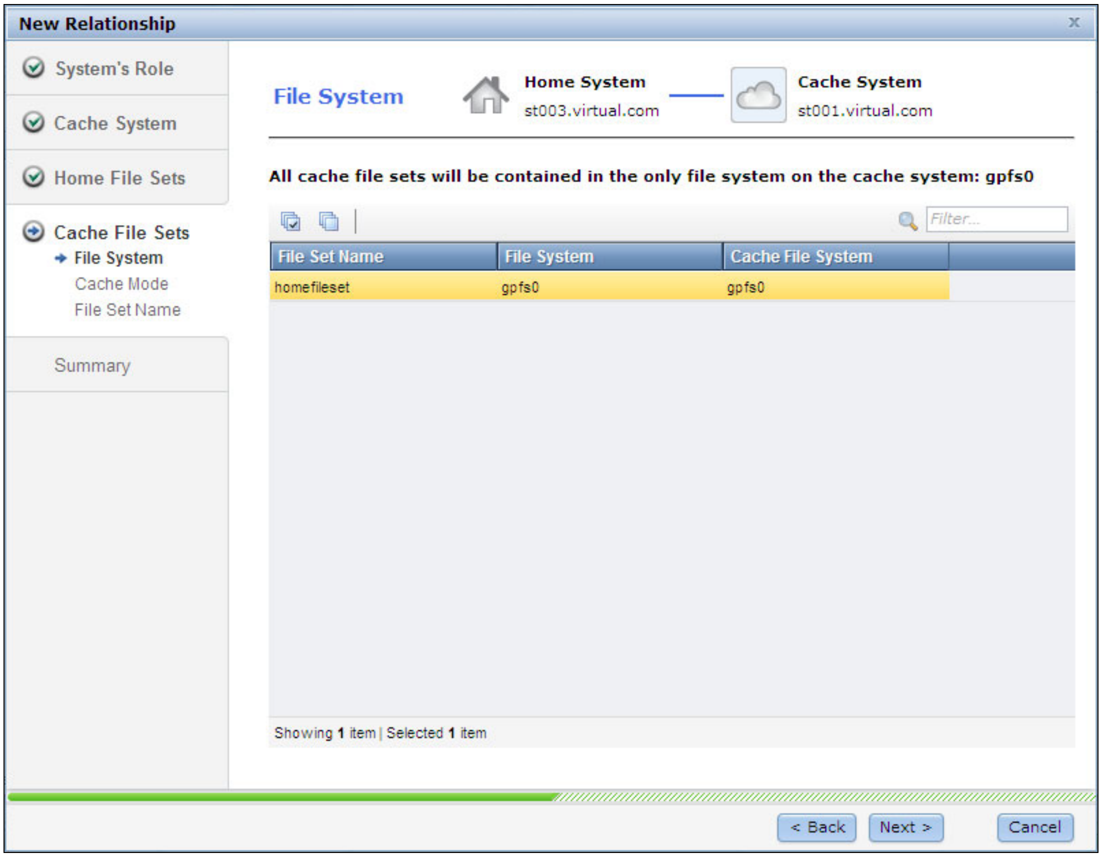
Select one or more file sets to cache:

File Set Name	File System	Path
homefileset	gpfs0	/ibm/gpfs0/homefileset

Showing 1 item | Selected 0 items

Figure 5-14   New Relationship window - select file sets to cache

7. The next window (Figure 5-15) summarizes the selected file sets and file systems. In this case, you cache only `homefileset`, which is placed in the `gpfs0` file system. Click **Next** to continue.



All cache file sets will be contained in the only file system on the cache system: gpfs0

File Set Name	File System	Cache File System
homefileset	gpfs0	gpfs0

Showing 1 item | Selected 1 item

Figure 5-15 Summary of the selected file sets and file systems

8. The next step is to select the cache mode. The following cache modes are available:

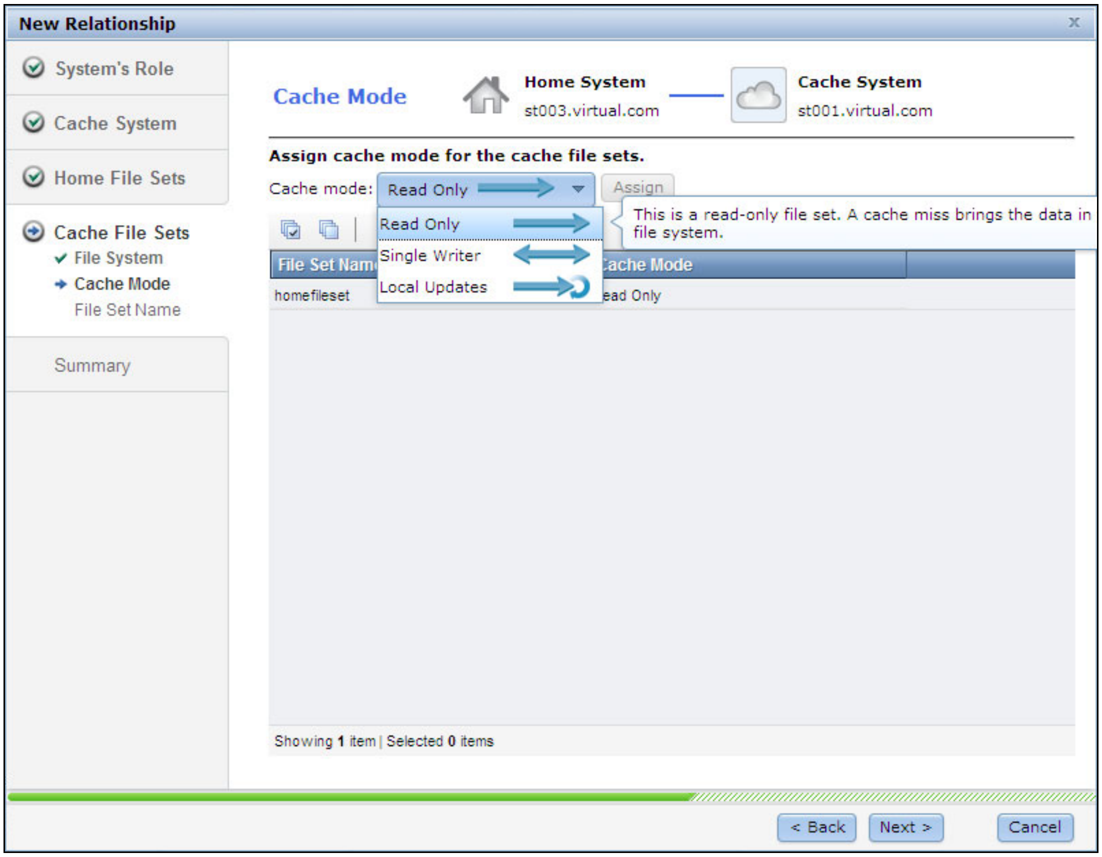
   – Single Writer mode

   – Read Only mode

     In Read Only mode, if a file is removed and re-created at the home system after the file is cached in the remote Read Only cache, the file is moved into the `.ptrash` subdirectory to avoid application errors. These files are removed automatically when the file set soft quota limit is reached, or when the **`ctlwcache`** command is run with the **`--evict`** option.

   – Local Update mode

Select the **Read Only** mode, as shown in Figure 5-16. Click **Next** to continue.



Assign cache mode for the cache file sets.

Cache mode: Read Only

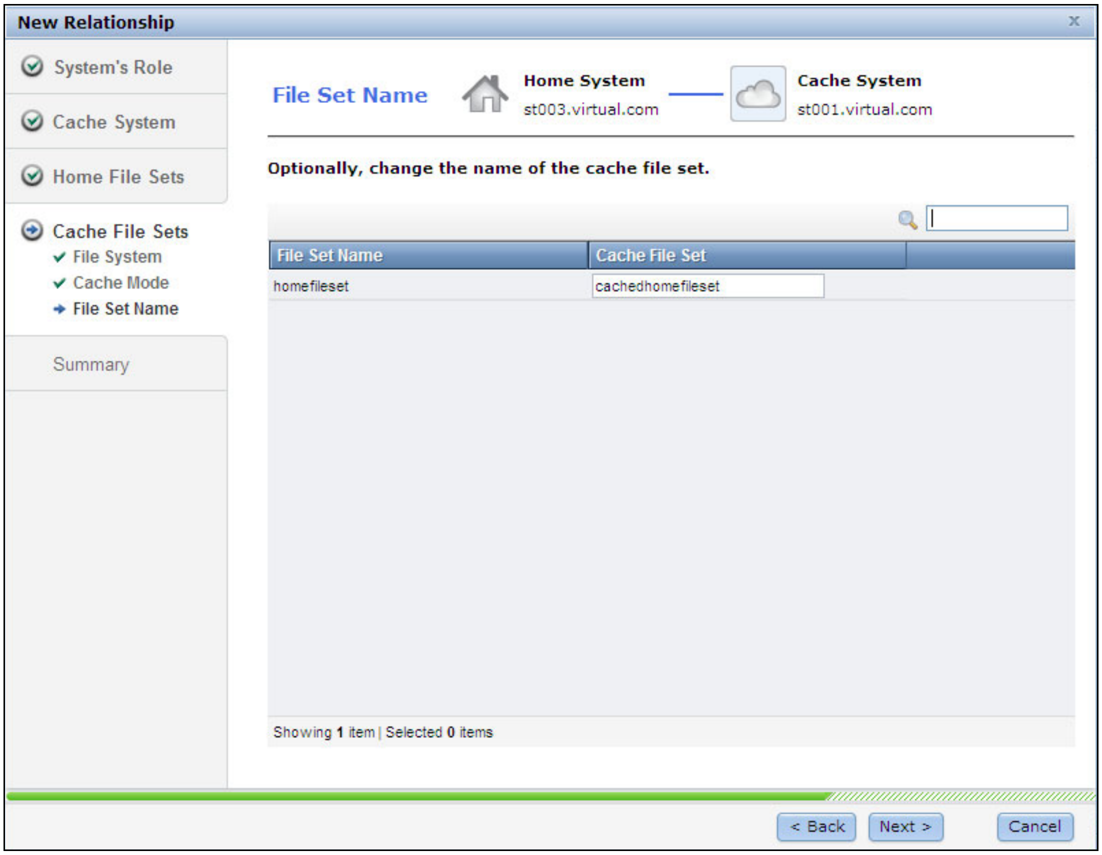
This is a read-only file set. A cache miss brings the data in file system.

File Set Nam	ache Mode
homefileset	ead Only

Showing 1 item | Selected 0 items

Figure 5-16 Cache Mode - selecting Read Only mode

9. In the next window, you can optionally change the file set name on the cache system. As shown in Figure 5-17, change the file set name to `cachedhomefileset`. Click **Next** to continue.



New Relationship

File Set Name

Home System
st003.virtual.com

Cache System
st001.virtual.com

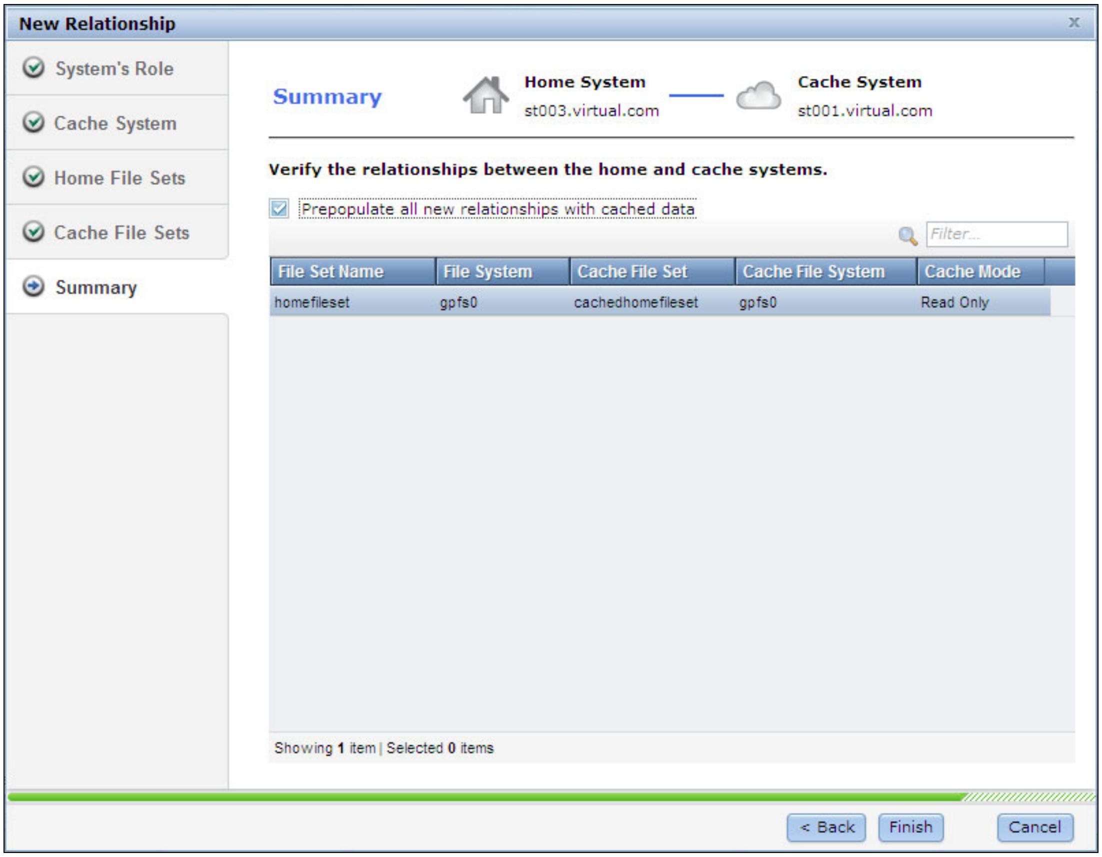
Optionally, change the name of the cache file set.

File Set Name	Cache File Set
homefileset	cachedhomefileset

Showing 1 item | Selected 0 items

Figure 5-17 Cache File Set name change

10. The last step is to verify the settings. To cache all of the data in a file set from the home file set to its cache file set, select the **Prepopulate all new relationships with cached data** check box and click **Finish** to save and activate the settings, as shown in Figure 5-18 on page 323.



Verify the relationships between the home and cache systems.

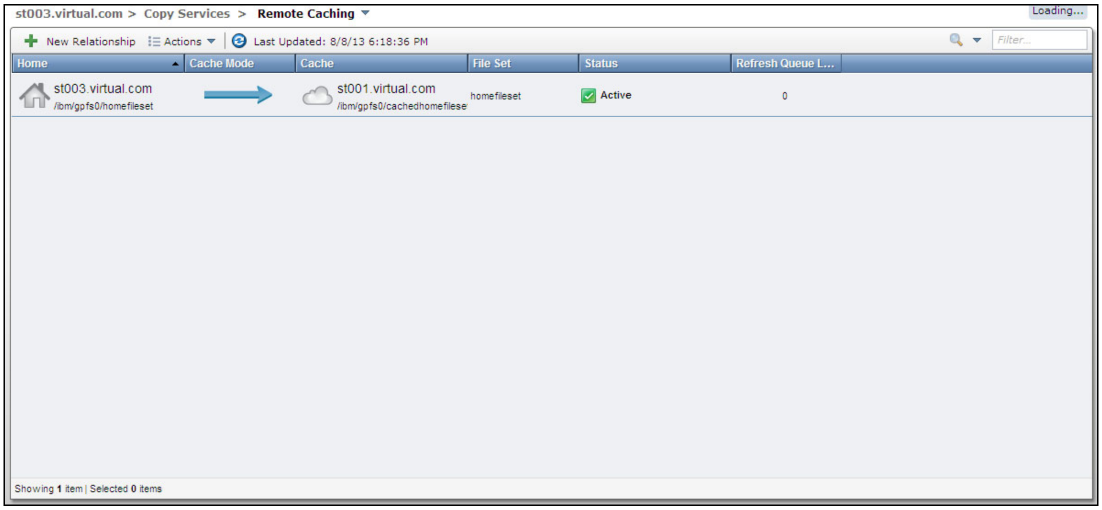
☑ Prepopulate all new relationships with cached data

File Set Name	File System	Cache File Set	Cache File System	Cache Mode
homefileset	gpfs0	cachedhomefileset	gpfs0	Read Only

Showing **1** item | Selected **0** items

Figure 5-18 Summary of the relationship between home and cache systems

11. After a successful configuration, the relationship is displayed, as shown in Figure 5-19.



Last Updated: 8/8/13 6:18:36 PM

Home	Cache Mode	Cache	File Set	Status	Refresh Queue L...
st003.virtual.com /ibm/gpfs0/homefileset	→	st001.virtual.com /ibm/gpfs0/cachedhomefilese	homefileset	✓ Active	0

Showing **1** item | Selected **0** items

Figure 5-19 Display of the new home and cache system relationship

5.2.3 Setting up remote caching by using the CLI

This section describes a scenario to establish a remote caching relationship between two SONAS systems by using the CLI. Configuration is done only on the home system.

On the home system, assume that there is a file set that is named `homefileset`. On the cache system, you call the file set `cachedhomefileset`. There is also a share with the name `cacheNFSshare`.

Here are the IP addresses that are used in this scenario:

- Home system IP: `10.0.0.30`
- Home system public IP: `10.0.0.131` and `10.0.0.132`
- Cache system IP: `10.0.0.10`

Setting up the home system

To set up the home system, complete the following steps:

1. Log in to the home system by using the CLI. Create a remote cache home export by running **`mkwcachesource`**, as shown in Example 5-1. When you create a home export, you must provide the cache system management IP for the **`--client`** option. Use the **`lsnwmgmt`** command on the cache system to determine the management IP. Use the **`lsexport`** command for the share details.

Example 5-1 Sample mkwcachesource command output

```
[st003.virtual.com]$ lsexport
Name          Path                   Protocol Active Timestamp
cacheNFSshare /ibm/gpfs0/homefileset NFS      true   8/8/13 8:43 PM
EFSSG1000I The command completed successfully.

[st003.virtual.com]$ mkwcachesource cacheNFSshare /ibm/gpfs0/homefileset
--client '10.0.0.10(ro)'
EFSSG1000I The command completed successfully.
```

Note: For read-only and local-updates modes, configure the access-mode as `ro`.

2. Verify that the remote cache home export is created by running **`lswcachesource`**, as shown in Example 5-2.

Example 5-2 Sample lswcachesource command output

```
[st003.virtual.com]$ lswcachesource
WCache-Source Name WCache-Source Path     ClientClusterId
ClientClusterName WCache-Source Access Mode Is Cached Remote system name
cacheNFSshare      /ibm/gpfs0/homefileset 12402779239030444814
st001.virtual.com ro                       no        st001.virtual.com
EFSSG1000I The command completed successfully.
```

Setting up the cache system

To set up the cache system, complete the following steps:

1. Log in to the cache system by using the CLI. A cache system must have one or more gateway nodes to communicate with the home system. Configure one or more gateway nodes on the cache system by running **`mkwcachenode`**. Run **`lsnode`** or **`lsnwgroup`** to determine the nodes that are available, as shown in Example 5-3 on page 325.

Example 5-3 Sample lsnwgroup command output

```
[st001.virtual.com]$ lsnwgroup
Network Group Nodes                    Interfaces
DEFAULT       mgmt001st001
int           int001st001,mgmt002st001 ethX0
EFSSG1000I The command completed successfully.

[st001.virtual.com]$ mkwcachenode --nodelist int001st001,mgmt002st001
EFSSG1000I The command completed successfully.
```

2. Verify the gateway nodes by running **`lsnode -v`**. In the command output, if the `Is Cache` field has the value `yes`, the node is a gateway node. If the `Is Cache` field has the value `no`, the node is a non-gateway node. For more details, see Example 5-4.

Example 5-4 Sample lsnode command output

```
[st001.virtual.com]$ lsnode -v -r
Hostname     IP           Description             Role                 Product
version Connection status GPFS status CTDB status Username Is manager Is quorum
Daemon ip address Daemon version Is Cache Recovery master Monitoring enabled
Ctdb ip address OS name         OS family Serial number Last updated
int001st001  172.31.132.1                         interface
1.4.1.0-40      OK                active      active      root     yes
yes      172.31.132.1     1224          yes     yes           yes
172.31.132.1    RHEL 6.1 x86_64 Linux     FD3B853       8/8/13 9:07 PM
mgmt001st001 172.31.136.2 active Management node  management,interface
1.4.1.0-40      OK                active      active      root     yes
no       172.31.136.2     1224          no      no            yes
172.31.136.2    RHEL 6.1 x86_64 Linux     417C29B       8/8/13 9:07 PM
mgmt002st001 172.31.136.3 passive Management node management,interface
1.4.1.0-40      OK                active      active      root     yes
no       172.31.136.3     1224          yes     no            yes
172.31.136.3    RHEL 6.1 x86_64 Linux     E94EE2F       8/8/13 9:07 PM
strg001st001 172.31.134.1                         storage
1.4.1.0-40      OK                active                  root     no
yes      172.31.134.1     1224          no      no            no
127.0.0.1       RHEL 6.1 x86_64 Linux     635E44F       8/8/13 9:07 PM
strg002st001 172.31.134.2                         storage
1.4.1.0-40      OK                active                  root     no
yes      172.31.134.2     1224          no      no            no
127.0.0.1       RHEL 6.1 x86_64 Linux     B6EEBEF       8/8/13 9:07 PM
EFSSG1000I The command completed successfully.
```

3. Create a read-only cache file set by running **`mkwcache`**, as shown in Example 5-5.

Example 5-5 Sample mkwcache command output

```
[st001.virtual.com]$ mkwcache gpfs0 cachedhomefileset
/ibm/gpfs0/cachedhomefileset --cachemode read-only --homeip 10.0.0.30
--remotepath 10.0.0.131:/ibm/gpfs0/homefileset
EFSSG1000I The command completed successfully.
```

In this example, the file system name is `gpfs0`.

The value for the **`--cachemode`** option can be `read-only`, `single-writer`, or `local-updates`. Read-only is specified because it is the type of cache file set that you want to create.

The value for the **--homeip** option is the home system's management IP. In the example, the value is 10.0.0.30.

4. Verify that the cache file set is created by running **lswcache**, as shown in Example 5-6.

Example 5-6 Sample lswcache command output

```
[st001.virtual.com]$ lswcache gpfs0
ID Name              Status Path                        CreationTime   Comment
RemoteFilesetPath                 CacheState CacheMode Remote system name
1  cachedhomefileset Linked /ibm/gpfs0/cachedhomefileset 8/8/13 8:51 PM
10.0.0.131:/ibm/gpfs0/homefileset enabled    read-only st003.virtual.com
EFSSG1000I The command completed successfully.
```

5.2.4 Setting up Independent Writers with the CLI

This section describes a scenario to establish a remote caching relationship between two IBM SONAS systems by using the CLI and configuring Independent Writer (IW) support. Configuration is done only on the home system and cache system. On the home system, assume that there is a file set that is named homefset1. On the cache system, call it iwcachfset1.

Here are the IP addresses that are used in this scenario:

- Home system IP: 9.118.47.55
- Cache system IP: 9.118.37.192

Setting up the home and cache systems

To set up the home system and cache systems, complete the following steps:

1. Log in to the home system with the CLI. Create a home file set by running **mkfset**. Then, create a remote cache home export by running **mkwcachesource**, as shown in Example 5-7. When you create a home export, you must provide the cache system management IP for the **--client** option. Run **lsnwmgmt** on the cache system to determine the management IP. Run **lswcachesource** for the remote cache home export details.

Example 5-7 Sample mkwcachesource command output

```
$ mkfset gpfs0 homefset1 --link
(1/3) Creating file set
EFSSG0070I File set homefset1 created successfully.
(2/3) Linking file set
EFSSG0078I File set homefset1 successfully linked.
(3/3) Setting owner
EFSSG1000I The command completed successfully.

$ mkwcachesource homeshare /ibm/gpfs0/homefset1 --client '9.118.37.192(iw)'
EFSSG1000I The command completed successfully.

$ lswcachesource
WCache-Source Name WCache-Source Path   ClientClusterId      ClientClusterName
WCache-Source Access Mode Is Cached Remote system name
homeshare          /ibm/gpfs0/homefset1 10632285293179704851 06tdph0.ibm
iw                        no       06tdph0.ibm
EFSSG1000I The command completed successfully.
```

2. Log in to the cache system with the CLI. Create a cache file set named `iwcachefset1` by running **`mkwcache`**, as shown in Example 5-8. The **`--cachemode independent-writer`** option is specified to enable Independent Writer support.

Example 5-8 Sample mkwcache command output

```
$ mkwcache fs0 iwcachefset1 /ibm/fs0/iwcachefset1 --cachemode
independent-writer --homeip 9.118.47.55 --remotepath /ibm/gpfs0/homefset1
EFSSG1000I The command completed successfully.

$ lswcache fs0
ID Name         Status Path                  CreationTime    Comment
RemoteFilesetPath               CacheState CacheMode          Remote system
name
15 iwcachefset1 Linked /ibm/fs0/iwcachefset1 4/18/14 2:05 PM
9.118.47.59:/ibm/gpfs0/homefset1 enabled    independent-writer pluto.in.ibm.com
EFSSG1000I The command completed successfully.
```

3. List the newly created cache file set by running **`lswcache`**.

5.3 Remote caching administration tasks

This section describes the administration tasks for remote caching.

5.3.1 Managing caching gateway nodes

The management of caching gateway nodes involves listing the caching gateway nodes and removing the caching gateway function from nodes.

Listing caching gateway nodes

To determine which nodes were configured as caching gateway nodes, run **`lsnode`** with the **`-v`** option, as shown in Example 5-9.

> **Tip:** The **`-Y`** option specifies to display the command output as colon-delimited fields. You might want to use the **`-r`** option to force a refresh of the node data before it is displayed. Otherwise, the displayed information might be stale.

Example 5-9 List caching gateway nodes

```
[st001.virtual.com]$ lsnode -v
Hostname     IP          Description            Role                 Product
version Connection status GPFS status CTDB status Username Is manager Is quorum
Daemon ip address Daemon version Is Cache Recovery master Monitoring enabled Ctdb
ip address OS name         OS family Serial number Last updated
int001st001  172.31.132.1                      interface          1.4.1.0-40
OK                active      active      root     yes        yes
172.31.132.1      1224           yes      yes            yes
172.31.132.1    RHEL 6.1 x86_64 Linux     FD3B853       8/8/13 9:07 PM
mgmt001st001 172.31.136.2 active Management node management,interface 1.4.1.0-40
OK                active      active      root     yes        no
172.31.136.2      1224           no       no             yes
172.31.136.2    RHEL 6.1 x86_64 Linux     417C29B       8/8/13 9:07 PM
```

```
mgmt002st001 172.31.136.3 passive Management node management,interface 1.4.1.0-40
OK                 active      active      root     yes        no
172.31.136.3       1224          yes       no            yes
172.31.136.3     RHEL 6.1 x86_64 Linux     E94EE2F      8/8/13 9:07 PM
strg001st001 172.31.134.1                       storage           1.4.1.0-40
OK                 active                  root     no         yes
172.31.134.1       1224          no        no            no
127.0.0.1        RHEL 6.1 x86_64 Linux     635E44F      8/8/13 9:07 PM
strg002st001 172.31.134.2                       storage           1.4.1.0-40
OK                 active                  root     no         yes
172.31.134.2       1224          no        no            no
127.0.0.1        RHEL 6.1 x86_64 Linux     B6EEBEF      8/8/13 9:07 PM
EFSSG1000I The command completed successfully.
```

The nodes with `yes` in the cache column are caching gateway nodes.

Removing the caching gateway function from nodes

To remove the caching gateway function from nodes, specify either the node names or node IP addresses in a comma-separated list or specify a network group by running **rmwcachenode**, as shown in Example 5-10.

Example 5-10 Remove the caching gateway function from nodes

```
$ rmwcachenode --nodelist mgmt002st004
This removes the wan-caching gateway node
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
```

You can use the **-f** or **--force** option to prevent the normal manual removal confirmation prompt from being displayed.

5.3.2 Managing caching source shares and exports on the home file set

You must first create a caching-enabled share or export on the home file set by running **mkwcachesource** before you can cache that share or export in a client system. Run **lswcachesource** on the home file set to list the caching-enabled shares and exports on that system. Run **rmwcachesource** to remove a cache-enabled share or export. Run **chwcachesource** to change a cache-enabled share or export configuration on the home file set, for example, to add or remove a caching client system.

To specify a system when you are using any of these four commands, use the **-c** or **--cluster** option of the command and specify either the system ID or the system name. If the **-c** and **--cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

Creating a caching source share or export

To enable remote caching for a file set, specify a share or export name and an existing file system path by running **mkwcachesource**, as shown in Example 5-11.

Example 5-11 Create a caching source share or export

```
$ mkwcachesource trial_ro /ibm/gpfs0/trial_ro --client ’10.0.100.10(ro)’
EFSSG1000I The command completed successfully.
```

Specify a comma-separated list of management IP addresses with the **--client** option. Use one client for each of the systems that you want to enable to cache the file set. Follow each IP with either **(rw)** or **(ro)** to specify whether the system that is managed by that IP address can access the file set in read/write or read only mode. The entire parameter value must be enclosed by the single quotation mark character at each end. Only one system can be defined to access the file set in read/write mode at any particular time. In Example 5-11 on page 328, the share name is `trial_ro` and the file system path is `/ibm/gpfs0/trial_ro`.

Listing caching source shares and exports

To show information about caching sources that are configured on a home system, run **lswcachesource**, as shown in Example 5-12.

Example 5-12 List caching source shares and exports

```
$ lswcachesource
WCache-Source Name WCache-Source Path     ClientClusterId    ClientClusterName
WCache-Source Access Mode Is Cached
wcache_ro         /ibm/gpfs0/wcache/wcache 12402779243267445246
st001.virtualad.ibm.com ro                      no
EFSSG1000I The command completed successfully.
```

To show information about all of the home shares and exports that are configured for a particular client system, use the **--clientclusterid** option and specify the system ID of the client system.

You can use the **--details** option to display whether the system is actively using the cache share or export, as shown in Example 5-13.

Example 5-13 List detailed caching source shares and exports

```
$ lswcachesource -details
WCache-SourceNameWCache-SourcePath WCache-SourceClients
Last Update
wcache_ro         /ibm/gpfs0/wcache/wcache
10.0.0.18(no_root_squash,no_wdelay,insecure,nohide,fsid=408829840);10.0.0.19(no_ro
ot_squash,no_wdelay,insecure,nohide,fsid=408829840);10.0.0.111(no_root_squash,no_w
delay,insecure,nohide,fsid=408829840);10.0.0.112(no_root_squash,no_wdelay,insecure
,nohide,fsid=408829840) 10/19/11 1:20 AM
EFSSG1000I The command completed successfully.
```

Only the listed IP addresses are permitted to access the corresponding share or export.

Removing a caching source share or export

To remove a cache source file set configuration on a home file set and delete the share or export, run **rmwcachesource** and specify the share or export name to be removed. You must respond `yes` to the confirmation prompt to remove a cache source file set configuration, as shown in Example 5-14.

Example 5-14 Remove a caching source share or export

```
$ rmwcachesource myCachedData
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
```

You can use the **-f** or **--force** option to prevent the normal manual removal confirmation prompt from being displayed.

Changing a caching source share or export

To add a caching client system to a configured caching share or export on the home file set, specify the share or export name by running **chwcachesource**. Use the **--addclient** option to specify the management IP address of the caching system that you are adding, suffixed by **(rw)** or **(ro)** for its access mode. The entire parameter value is enclosed by the single quotation mark character at each end, as shown in Example 5-15.

Example 5-15 Change a caching source share or export

```
$ chwcachesource testShare --addclient '10.0.0.100(rw)'
EFSSG1000I The command completed successfully.
```

This command option must also be used after you change a cache system ID to enable configuration or reconfiguration of client caches by using the new ID.

To remove a caching client system from a configured caching share or export on the home file set, specify the share or export name by running **chwcachesource** and use the **--removeclient** option to specify the client system ID to be removed. The client system ID is displayed by the **lswcachesource** command. If there is only one client system that is configured for a share or export, it cannot be removed with the **chwcachesource** command.

You can use the **-u** or **--updateKeys** option to update the remote cache internal user keys by using the client remote cache's user key definitions. One scenario where it is required is when the system ID of a client system is changed.

Use the **-f** or **--force** option to force the command submission and prevent normal output display.

5.3.3 Managing cache file sets

Before you can create a cache file set on a client system by running **mkwcache**, you must first create the caching-enabled share or export on the home file set by running **mkwcachesource**. You must also enable wide area network (WAN) caching on the client SONAS system by running **mkwcachenod**, and to configure specified Interface nodes in that system to function as the caching gateway nodes that exchange data with other systems.

Run **lswcache** to list the cache file sets on a system. You can use the **rmwcache** command to remove a cache file set. You can run **chwcache** to change the configured parameters of a cache file set. Run **chcfg** to set certain cache attributes at the client cache system level, and run **ctlwcache** to perform caching maintenance operations on the cache file set. You can run **lswcachestate** to display the status of gateway nodes and client shares and exports, and to list pending and completed cache operations.

> **Tip:** If the home file system or file set is destroyed and re-created or restored at the home site, the cache relationship is lost. It must be re-established by running **chwcache** and specifying the path by using the **--remotepath** and **--homeip** options.

To specify a system when you are using the **mkwcache**, **lswcache**, **rmwcache,** and **chwcache** commands, use the **-c** or **--cluster** option of the command and specify either the system ID or the system name. If the **-c** and **--cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

Creating a cache file set

To create a cache version of a file set on a client system, run **mkwcache** and specify a device, a file set name, and file system junction path.

For device, specify the device name of the file system that you want to contain the new cache file set on the client system. The device name of the file system does not need to be fully qualified; for example, `fs0` is as acceptable as `/dev/fs0`. The device name must be unique within a GPFS cluster, and the device cannot be an existing entry in `/dev`.

The file system junction path specifies the name of the junction, must be a file system path, and cannot refer to an existing file system object. If the file system junction path is not specified, the default is the file system mount point/file system name.

Tip: You cannot link any file set, whether dependent or independent, into a cache file set.

You must also specify the home system IP address and path of the file set cache source by using the **--remotepath** required option of the **mkwcache** command in the format `home system IP address:exported path` (see Example 5-16). It is the share or export that you want to cache to the remote file set; this share or export must have previously been created on the home file set by running the **mkwcachesource** command. The home system IP address is a public IP address of the home system. If the home system IP address is not specified (which requires that the colon separator is also omitted), an internal public IP of the home system is selected based on a round-robin algorithm. The exported path can be either the full path specification or just the name of the exported file set. In the first instance in Example 5-16, the home system IP address is specified; in the second instance, it is omitted.

Example 5-16 The --remotepath option of the mkcache command

```
--remotepath 10.0.0.41:/ibm/gpfs0/wcacheSource
--remotepath /ibm/gpfs0/wcacheSource
```

You must also use the **--homeip** required option to specify the IP address of the active Management node of the home system.

You must specify either the **-h** option to specify the hard limit of disk usage for the file set, or the **-s** option to specify the soft limit. These values are used for automated cache eviction.

You can use the **--cachemode** option to specify the remote caching mode of the cache version of the file set that you are creating on the remote system. Valid values are `read-only`, `local-updates`, and `single-writer`. If the option is not used or its value is not specified, the default value is `read-only`.

Note: If the file set is being created with Single Writer caching mode, the system verifies whether the home file set share or export allows the specified remote cache file set to be in Single Writer mode. If write permission is not configured for the cache file set, an error message is displayed.

After the cache mode is set to `local-updates`, you cannot change the cache mode to any other mode value.

To change the cache mode of a file set, the file system that contains the cache file set to be changed must be unmounted.

In Example 5-17, the device is gpfs0, the file set name is test_2, and the file system junction path is /ibm/gpfs0/test_2.

Example 5-17 -mkcache with the --homeip option

```
$ mkwcache gpfs0 test_2 /ibm/gpfs0/test_2 --cachemode read-only --remotepath
'10.0.100.133:/ibm/gpfs0/--homeip '10.0.100.10' -h 1024
EFSSG1000I The command completed successfully.
```

You can use the **-i** option to specify the maximum number of inodes that can be allocated and the number of inodes that the system immediately preallocates to the cache version of the file set, which is separated by a colon. You can use the multiplicative suffixes M for millions and either K or k for thousands, for example, **-i 1M:500K**.

You can use the **--FileOpenRefreshInterval** option to specify the number of seconds since the last time that the file was opened, after which revalidation with the file version on the home file set must be performed. Similarly, you can use the **--FileLookupRefreshInterval** option to specify the number of seconds since the last time that a file lookup occurred, after which revalidation with the file version on the home file set must be performed. A file lookup is when the file metadata is retrieved without opening the file for a read or write, which usually occurs when you display or report the file metadata. The **--remotepath 10.0.0.41:/ibm/gpfs0/wcacheSource --remotepath /ibm/gpfs0/wcacheSource --DirOpenRefreshInterval** and **--DirLookupRefreshInterval** options specify the corresponding time latencies for directories instead of files.

Tip: Setting a lower value for a refresh interval provides greater consistency between the home and remote system data. If the home file set data changes frequently, you might want to set refresh intervals as low as 0 for critical data consistency.

You can specify a time interval in seconds after which queued updates are flushed when a gateway node becomes disconnected by using the **--DisconnectTimeout** option. A value of disable for this option forces synchronous operations. You can specify a time interval after which files and directories in the cache file set expire when the cache file set is in disconnected mode by using the **--ExpirationTimeout** option.

This option is disabled by default. You must enable this option by specifying a value if you want to use the **ctlwcache** command **--expire** option to mark all of the cache file set contents as expired. You can specify the number of seconds before a queued write must be sent to the home file set by using the **--AsyncDelay** option. This delay might be useful for write-intensive applications that frequently write to the same file set. Delaying writes to the home file set provides the opportunity to send a single write containing the most current version of data that results from multiple writes, thus reducing WAN traffic. However, setting a higher value decreases the currency of data on remote system.

When you create a cache file set, if any of the following options or their corresponding values are not specified, the default value that is listed in Table 5-1 is used.

Table 5-1 Cache file set options

Option	Default value
CacheMode	1 (Read-only)
fileOpenRefreshInterval	30 seconds
fileLookupRefreshInterval	30 seconds
dirOpenRefreshInterval	60 seconds

Option	Default value
dirLookupRefreshInterval	60 seconds
asyncDelay	15 seconds
DisconnectTimeout	60 seconds
expirationTimeout	disabled
i (max inodes and # inodes to preallocate	100,000

You can specify a comment that is displayed in the output of the **lswcache** command by using the **-t** option. The comment must be fewer than 256 characters in length. If the comment text contains a space character, the comment text must be enclosed in quotation marks.

Listing cache file sets

To show information about cache file sets on a remote system, run **lswcache** and specify a device name, which does not need to be fully qualified, as shown in Example 5-18.

Example 5-18 Sample lswcache command

```
$ lswcache gpfs0
ID Name Status Path CreationTime Comment RemoteFilesetPath CacheState CacheMode
3 test_2 Linked /ibm/gpfs0/test_2 5/11/11 1:03 PM
10.0.100.133:/ibm/gpfs0/testing3_home_default enabled read-only
1 trialing_lu Linked /ibm/gpfs0/trialing 5/10/11 5:43 PM
10.0.100.133:/ibm/gpfs0/trialing enabled local-updates
3 test_2 Linked /ibm/gpfs0/test_2 5/11/11 1:03 PM
10.0.100.133:/ibm/gpfs0/testing3_home_default enabled read-only
1 trialing_lu Linked /ibm/gpfs0/trialing 5/10/11 5:43 PM
10.0.100.133:/ibm/gpfs0/trialing enabled local-updates
```

You might want to use the **-r** option to force a refresh of the cache file sets data before it is displayed; otherwise, the displayed information might be stale.

You can use the **-v** or **--verbose** option to also display this information for a cache file set:

- RootInode (the number of the root inodes)
- ParentID (The parent file set identifier; -- displays if none exists.)
- Async delay
- File open refresh interval
- File lookup refresh interval
- Directory open refresh interval
- Directory lookup refresh interval
- Expiration timeout
- Disconnect timeout

You can use the **-Y** option to specify to display the command output as colon-delimited fields.

Removing a cache file set

To remove a cache file set from a remote system and delete its contents after unlinking any child file sets, run **rmwcache** and specify the file system name and the file set to be removed. You must respond yes to the confirmation prompt to remove a cache file set, as shown in Example 5-19.

Example 5-19 Sample rmwcache command output

```
$ rmwcache /dev/fs0 myCachedData
All data in the fileset will be deleted.
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
```

You can use the **-f** or **--force** option to prevent the normal manual removal confirmation prompt from being displayed.

This command cannot be used to remove a file set that is not a cached file set.

Changing a cache file set

To change a cache file set configuration on a remote system, specify the device and file set name by running **chwcache**. As with the **mkwcache** command, the device name of the file system does not need to be fully qualified; for example, `fs0` is as acceptable as `/dev/fs0`. The device name must be unique within a GPFS cluster.

You can change the following values of the **mkwcache** command options by running **chwcache**:

- **-t** (to change the comment)
- **--cachemode**
- **--FileOpenRefreshInterval**
- **--FileLookupRefreshInterval**
- **--DirOpenRefreshInterval**
- **--DirLookupRefreshInterval**
- **--ExpirationTimeout**
- **--DisconnectTimeout**
- **--AsyncDelay**
- **--remotepath**
- **--homeip** (This option must be specified when the --remotepath option is specified.)
- **-i** (maximum number of inodes and maximum number of inodes to preallocate)

> **Tip:** If the file set is being created with single-writer caching mode, the system verifies whether the home file set share or export allows the specified remote cache file set to be in single-writer mode. If write permission is not configured for the cache file set, an error message is displayed.

After the cache mode is set to `local-updates`, you cannot change the cache mode to any other mode value.

To change the cache mode of a file set, the file system that contains the cache file set to be changed must be unmounted.

Example 5-20 on page 335 changes the home IP of a Single Writer mode file set named finalTest to a different system.

Example 5-20 Sample chwcache command syntax

```
$ chwcache gpfs0 finalTest --remotepath 10.0.100.10:/ibm/gpfs0/final --homeip
10.0.100.10
```

You must respond with `yes` to submit the change request. For more information, see "Moving the Single Writer mode of a file set to a different client system" on page 344.

You can use the **-f** or **--force** option to force the command submission and prevent normal output display.

You can display the cache file set configuration by running **lswcache**.

Setting select cache attributes at the system level

You can use the **chcfg** command to set the following cache attributes at the remote cache system level:

- **--cachemode**
- **--FileOpenRefreshInterval**
- **--FileLookupRefreshInterval**
- **--DirOpenRefreshInterval**
- **--DirLookupRefreshInterval**
- **--ExpirationTimeout**
- **--DisconnectTimeout**
- **--AsyncDelay**

Tip: If an attribute is not specified at the file set level and the same attribute is defined at the system level by running **chcfg**, the system-level attribute value is used. When both file-set-level and system-level definitions are present for the same attribute, the file-set-level attribute value has precedence over the system-level attribute value.

Performing caching maintenance operations on the cache system

You can use the **ctlwcache** command and specify the device and cache file set name to resynchronize client cache data to the home file set, to flush the cache I/O to the home file set, to evict cached data, and to expire or unexpire all of the data in a file set.

You can use the **--resync** option when there are inconsistencies between the home and cache versions of data to ensure that the home version is consistent with the cache version.

You can use the **--flushqueue** option to flush a client file set's I/O queue to the home file set. The command completes when the queue is empty. You can monitor the queue by using the **--connectionstatus** option of the **lswcachestate** command.

You can use the **--expire** option to mark all of the cache file set contents as expired.

You can use the **--unexpire** option to remove the expired status from all of the cache file set contents that are marked as expired.

Tip: The **--expire** and **--unexpire** options fail if the **--ExpirationTimeout** option of the cache file set is disabled, which is the default value.

You can use the **--evict** option to remove cache data if the soft limit is set for the disk usage by the file set, either when the cache file set is created by running **mkwcache**, or later by running **setquota** with the **-j** option to specify the file set. You can use the **lsquota** command with the **-j** option to specify the file set to display the current soft limit. You cannot use the **--evict** option if the soft limit is not defined, or is zero. If you use the **--evict** option, you must specify the **--evictsafelimit** suboption. The **--evict** option takes the following suboptions:

- If you use the **--evict** option, you must use the **--evictsafelimit** suboption and specify the target quota limit for eviction that is used as the low watermark. The value that is specified with the **--evictsafelimit** suboption must be less than the value for the soft limit, which can be displayed by running **lsquota** with the **-j** option to specify the file set. The value that is specified by the **--evictsafelimit** suboption overrides the soft limit value during the manual eviction operation.
- You can use the **--log** suboption and specify the location of the eviction log file. By default, the logged information is appended to `mmfs.log`.
- You can use the **--evictorder** suboption and specify either LRU or SIZE to indicate the eviction queue ordering sequence.
- You can use the **--evictfilenamepattern** suboption and specify a value that is a file name pattern regular expression.
- You can use the **--evictminfilesize** suboption and specify the minimum file size in KB.
- You can use the **--evictmaxfilesize** suboption and specify the maximum file size in KB.

Displaying the remote caching status

You can use the **lswcachestate** command and specify a file system to display the connection status of client cache shares and exports, and to list the status of pending and completed cache operations. You can optionally specify a file set name to limit the display to the specified file set.

You can use the **--connectionstatus** option to display the status for all of the cache gateway nodes and client shares and exports. The displayed connection status includes the remote file set status, the cache gateway that is assigned to the specified file set, the status of the gateway, and the status of the queue.

The remote file set status can have one of the following values:

- `Active`: The cache is active and the NFS mount from home was successful. The value `Active` is valid for all caching modes.
- `Inactive`: The cache has not connected to home. This state changes when new operations are initiated that require cache to contact home.
- `Unmounted`: The cache cannot mount NFS from home.
- `Disconnected`: The cache cannot connect to home.
- `Expired`: The cache is disconnected from home after the expiration timeout that is set for the cache occurs.
- `NeedsResync`: The cache has detected inconsistencies between home and cache, and these inconsistencies are being fixed automatically. This state is transitional.
- `Dirty`: The cache has data that is not sent to home.

The cache gateway status can have one of the following values:

- `Active`: The queue is active.
- `FlushOnly`: The queue waits for application requests.

- `QueueOnly`: Operations are queued but not flushed. It is a temporary state.
- `Dropped`: A transient state that occurs because of external changes.
- `Recovery`: A WAN cache recovery is in progress.

Example 5-21 shows the format of the **`--connectionstatus`** option.

Example 5-21 List the cache gateway status

```
$ lswcachestate gpfs0 --connectionstatus
Fileset Path Remote Fileset Status Cache-Gateway Assigned Cache-Gateway Status
QueueLength QueueNumExec
sw_fset1 gpfs0 10.0.100.132:/ibm/gpfs0/fset1 Active int003st001 Active 0 7
ro_fset1 gpfs0 10.0.100.131:/ibm/gpfs0/fset1 Active mgmt001st001 Active 0 6
```

You can use the **`--actionstatus`** option to display the cache operations currently pending and the cache operations that have completed since the most recent start. Possible status values include `RUNNING`, `COMPLETED`, and `FAILED`. A status of `RUNNING` indicates that the task is still progressing in the background. A status of `COMPLETED` indicates that the action was performed successfully. A status of `FAILED` indicates that the task did not complete, and a possible error reason is reflected in the output.

Example 5-22 shows the format of the **`--actionstatus`** option.

Example 5-22 Sample lswcachestate command with the --actionstatus option

```
$ lswcachestate gpfs0 --actionstatus
FilesetName FilesystemName WCacheOperationId Status Message Timestamp
sw_fset1 gpfs0 evict_20111014073421 FAILED wan-caching operation evict failed for
fileset sw_fset1 on node mgmt001st001 Please see logs for details. 10/14/11 7:35
AM
ro_fset1 gpfs0 evict_20111014073351 FAILED wan-caching operation evict failed for
fileset ro_fset1 on node mgmt001st001 Please see logs for details. 10/14/11 7:34
AM
ro_fset1 gpfs0 unexpire_20111014073257 FINISHED wan-caching operation unexpire
completed successfully for fileset ro_fset1 10/14/11 7:33 AM
sw_fset1 gpfs0 resync_20111014073229 FINISHED wan-caching operation resync
completed successfully for fileset sw_fset1 10/14/11 7:33 AM
sw_fset1 gpfs0 flushPending_20111014073218 FINISHED wan-caching operation
flushPending completed successfully for fileset sw_fset1 10/14/11 7:33 AM
```

You can use the **`-Y`** option to specify that the output is displayed in colon-delimited format.

5.3.4 Managing cache prepopulation

Remote caching fetches complete files on demand from the home file set to the remote system cache in real time during normal operation. The prepopulation feature moves files into the cache in batch mode so that they are already present in the cache when they are accessed by an application. Prepopulation of cache before an application is started can reduce the network delay when the application starts. Prepopulation can also be used to proactively manage WAN traffic patterns by moving files over the WAN during a period of lower WAN usage in anticipation that it might otherwise be accessed during a period of higher WAN usage. You can use the r**`unprepop`** command to specify a file system, file set, and policy to prepopulate cache. You can use the **`lsprepop`** command to specify a file system and file set to display the status of a prepopulation operation.

To specify a system when you are using the **runprepop** and **lsprepop** commands, use the **-c** or **--cluster** option and specify either the system ID or the system name. If the **-c** and **--cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

Running cache prepopulation

To cache all of the data in a file set from the home file set to its cache file set on the remote system, run **runprepop** and specify the file system, the file set name, and the policy to use for prepopulation. The specified policy can use only the LIST rule and cannot contain any REPLICATE, MIGRATE, or DELETE rules. For more information, see 5.5, “Creating and managing policies” on page 355. Prepopulation does revalidation of locally stored cached file attributes at the remote system against the corresponding home file set attributes and fetches a file only if the locally stored cached file attributes differ from the attributes of the corresponding file at the home file set.

Displaying the cache prepopulation status

To display the status of a prepopulation operation, run **lsprepop**. You can optionally limit the display to a file system by using either the **-d** or **--filesystem** option and you can optionally limit the display to a file set by using either the **-f** or **--filesetName** option. You can also use the **--latestPrePops** option to limit the display to the specified number of the most recent prepopulation status entries, as shown in Example 5-23.

Example 5-23 Display the cache prepopulation status

```
$ lsprepop -l 3
Cluster ID filesystem FilesetName Status Message Last update Timestamp
12402779238972964251 gpfs0 xcache118 FINISHED FINISHED 10/12/11 12:53 PM
12402779238972964251 gpfs0 xcache120 FINISHED FINISHED 10/12/11 12:53 PM
12402779238972964251 gpfs0 xcache117 FINISHED FINISHED 10/12/11 12:53 PM
```

You can use the **-Y** option to specify display in colon-delimited format, as shown in Example 5-24.

Example 5-24 Display the cache prepopulation status with the -Y option

```
$ lsprepop -Y -l 2
lsprepop:Prepop:HEADER:version:reserved:reserved:Cluster
ID:filesystem:FilesetName:Status:Message:Last
lsprepop:Prepop:0:1:::12402779238972964251:gpfs0:xcache120:FINISHED:FINISHED:2011-
10-12 12.53.38:
lsprepop:Prepop:0:1:::12402779238972964251:gpfs0:xcache117:FINISHED:FINISHED:2011-
10-12 12.53.30:
```

5.3.5 Managing partial file caching

The default behavior of Active Cloud Engine WAN caching is that, when a file is requested, it fetches the complete file in cache. For large files, this process can take a long time, which depends on the available network bandwidth. To alleviate this behavior, Active Cloud Engine provides some file-set-level control parameters, which you can use to configure the number of blocks to be cached when a file is requested. This configuration helps remove unnecessary penalties on cache disk space, reduces network delays because of clogged bandwidth, and allows for faster determination of whether the correct file is accessed.

In SONAS versions before Version 1.5, the entire file is cached, so reading any part of a file (except the first block) causes all the contents of the file to be fetched from home and stored locally in the cache. SONAS V1.5 introduces a configurable Partial file caching option that fetches only the blocks that are accessed and uses the network and local disk space more efficiently.

The amount of data to cache per file can be controlled with the **`--readprefetchthreshold`** option on the cache file set by running **`mkwcache`** and **`chwcache`**. This option optimizes data transfer and data storage in cache according to the application requirements, as shown in Example 5-25.

Example 5-25 Change the read prefetch threshold for a file set

```
chwcache <device> <cachefileset> --readprefetchthreashold <file percentage>
```

The value range for the **<file percentage** → option is 0 - 100. The thresholds have the following meanings:

0 Specify this value to prefetch the entire file when the application that is accessing the file starts reading the file. This value is the default value.

1-99 Percentage of file size that must be cached before the entire file is prefetched. For example, if you specify 60 as the threshold value, SONAS triggers full file prefetch when the application that is accessing the file completes reading 60% of the file.

100 Specify this value to disable full file prefetching.

5.3.6 Managing cache file set peer snapshots

The cache file set peer snapshot function provides a generic infrastructure for obtaining an application-consistent point-in-time copy of data in the cache file set that can be integrated with other functions to meet specific customer requirements, such as backup and recovery. The peer snapshot pushes all of the snapshot data in cache to the home file set so that the home data is consistent with cache, and then takes a snapshot of the corresponding home data. This process results in a pair of peer snapshots, one each at the cache and home file sets, which refers to the same consistent copy. Peer snapshots can be created manually on demand or on a defined schedule.

> **Note:** If cache is disconnected from the home file set when the cache snapshot is created, the cache notes that the peer snapshot on the home file set has not been created. When the home file set becomes reconnected to cache, the cache attempts to resume the creation of the missing peer snapshot at the home file set if conditions permit. Certain events in the intervening time interval might prevent resuming the creation of the peer snapshot at the home file set.

To specify a system when you are using the peer snapshot commands, use the **`-c`** or **`--cluster`** option of the command and specify either the system ID or the system name. If the **`-c`** and **`--cluster`** options are omitted, the default system, as defined by the **`setcluster`** command, is used.

Creating a peer snapshot

To create a pair of peer snapshots, one each on the remote and home file set, run **mkpsnap** and specify a file system, a file set name, and peer snapshot name. For file system, specify the device name of the file system that you want to contain the peer snapshot on the cache file set. The device name of the file system does not need to be fully qualified; for example, fs0 is as acceptable as /dev/fs0. The device name must be unique within a GPFS cluster, and the device cannot be an existing entry in /dev. The file system must be unique within a device. The peer snapshot file name is created by prefixing the specified peer snapshot name with system-generated values by using the following format:

peer snapshot name-psnap-system uid-file system uid-file set uid-timestamp

The maximum number of snapshots per file set is 100. Example 5-26 creates a peer snapshot that is named test on the file system that is named gpfs0 and in the file set named sw_fset.

Example 5-26 Create a peer snapshot

```
$ mkpsnap gpfs0 sw_fset test
If the maximum limit of psnaps is reached, this will automatically delete the
oldest snapshot.
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
```

Displaying peer snapshot information

To list peer snapshot information, run **lspsnap**. You can optionally specify a file system. If no file system is specified, information for all peer snapshots on the system is displayed. If you specify a file system, you can specify that, if the maximum limit of **spsnap** is reached, the file system automatically deletes the oldest snapshot. Also, you can optionally specify a file set. If you specify a file system but no file set, information for all peer snapshots on the file system is displayed.

You can use the **-r** or **--refresh** option to force a refresh of the peer snapshot information before it is displayed; otherwise, the displayed information might be stale. When you use this option, a message is displayed before the peer snapshot information is displayed, as shown in Example 5-27.

Example 5-27 Display peer snapshot information

```
EFSSG0015I Refreshing data.
Filesystem name Fileset name Snapshot ID Status Creation ID Timestamp
gpfs0 sw_fset test valid 5/19/11 12:46 PM 3 5/19/11 12:46 PM
gpfs0 sw_fset abc valid 5/19/11 9:39 AM 1 5/19/11 12:44 PM
EFSSG1000I The command completed successfully.
```

A status of valid indicates that there is a corresponding peer snapshot on the home file set. A status of invalid indicates that a corresponding peer snapshot does not exist on the home file set.

You can use the **-u** or **--usage** option to include usage data in the display of the peer snapshot information. This usage data is similar to the usage data that is displayed by the **lssnapshot** command. A message is displayed before the data is displayed to indicate that the peer snapshot information and usage data is refreshed, as shown in Example 5-28 on page 341.

Example 5-28 List the peer snapshot status

```
$ lspsnap -c 12402814423332523984 -u
EFSSG0015I Refreshing data.
Filesystem name Fileset name Snapshot ID Status Creation ID Used (metadata) Used
(data) Timestamp
gpfs0 sw_fset test valid 5/19/11 12:46 PM 3 0 0 5/19/gpfs0 sw_fset abc valid
5/19/11 9:39 AM 1 0 0 5/19/EFSSG1000I The command completed successfully.
```

Tip: Because the **-u** or **--usage** option does the same peer snapshot information refresh as the **-r** or **--refresh** option, and additionally refreshes and displays usage information, you cannot specify both options in the same command submission instance.

You can use the **-v** option to include both the system ID and usage data in the display of the peer snapshot information. This usage data is similar to the usage data that is displayed by the **lssnapshot** command, as shown in Example 5-29.

Example 5-29 List the peer snapshot status

```
lspsnap -c 12402814423332523984 -v
Filesystem name Fileset name Snapshot ID Status Creation ID Used (metadata) Used
(data) Cluster gpfs0 sw_fset test valid 5/19/11 12:46 PM 3 0 0
12402814423332523984 gpfs0 sw_fset abc valid 5/19/11 9:39 AM 1 0 0
12402814423332523984 EFSSG1000I The command completed successfully.
```

Because the **-r** or **--refresh** options and the **-u** or **--usage** options are *not* specified in Example 5-29, the displayed information might be stale.

You can use the **-Y** option to specify the display of the command output as colon-delimited fields.

Removing a peer snapshot pair

To remove a peer snapshot pair, run **rmpsnap** and specify a file system, a file set name, and peer snapshot name. For file system, specify the device name of the file system that contains the peer snapshot that you want to remove on the cache file set. The device name of the file system does not need to be fully qualified; for example, fs0 is as acceptable as /dev/fs0. The device name must be unique within a GPFS cluster. The file system must be unique within a device. You must respond yes to the confirmation prompt to remove the peer snapshot pair, as shown in Example 5-30.

Example 5-30 Remove a peer snapshot pair

```
$ rmpsnap gpfs0 sw_fset test -c 12402814423332523984
Do you really want to perform the operation (yes/no - default no):yes
EFSSG0021I The psnap test has been successfully removed.
EFSSG1000I The command completed successfully.
```

You can use the **-f** or **--force** option to prevent the normal manual removal confirmation prompt from being displayed.

Scheduling automated cache file set peer snapshot tasks

Peer snapshot tasks can be created by running **mkpsnaptask** to automatically submit peer snapshot creation requests on a defined schedule. These tasks can be changed running **chpsnaptask** or removed by running **rmpsnaptask**, and information about these tasks can be displayed by running **lspsnaptask**.

To specify a system when you are using the peer snapshot commands, use the `-c` or `--cluster` option of the command and specify either the system ID or the system name. If the `-c` and `--cluster` options are omitted, the default system, as defined by the `setcluster` command, is used.

Creating a peer snapshot task

To create a peer snapshot, run `mkpsnaptask` and specify a file system, a file set name, and task name. For file system, specify the device name of the file system that you want to contain the peer snapshot on the cache file set. The device name of the file system does not need to be fully qualified; for example, `fs0` is as acceptable as `/dev/fs0`. The device name must be unique within a GPFS cluster, and the file system must also be unique within a device.

Use either the `--hourInterval` or the `--minuteInterval` option, or both, to specify the time latency since the last submission of the task for the next subsequent task submission. Use the `--hourInterval` option to specify the number of hours, which can be any non-negative integer. If the value is zero or the option is not specified, the `--minuteInterval` option must be specified with a nonzero value. Use the `--minuteInterval` option to add the specified number of minutes to the value of the `--hourInterval` option. Valid values are 0, 15, 30, and 45; however, if the `--hourInterval` option is not specified or is specified with a value of zero, the `--minuteInterval` option must be specified as one of the following nonzero integers: 15, 30, or 45.

Note: The initiation of a snapshot creation instance by a create peer snapshot task occurs only at 00, 15, 30, and 45 minutes after the hour. A snapshot for a scheduled interval might not be created by the create peer snapshot task if the specified time latency has not passed since the completion of the most recent snapshot creation by that task.

You can use the `--psnapLimit` option to specify the number of peer snapshots to be retained for a file set; the maximum is 100 and the default when the option or its value is not specified is 10. When the limit is exceeded, the oldest snapshot of the file set is removed when a new snapshot is created.

A peer snapshot task that is named `task` is automatically submitted every four hours and 45 minutes is created with the command that is shown in Example 5-31 for the file set named `sw_fset` on file system named `gpfs0`.

Example 5-31 Create a peer snapshot task

```
$ mkpsnaptask gpfs0 sw_fset task -c 12402814423332523984 --hourInterval 4
--minuteInterval 45 --psnapLimit EFSSG1000I The command completed successfully.
```

Changing a peer snapshot task

To change a create peer snapshot task, run `chpsnaptask` and specify the task name. You can change the `--hourInterval`, `--minuteInterval`, and `--psnapLimit` values of the task, subject to the limitations that are described in "Creating a peer snapshot task" regarding the usage of the `mkpsnaptask` command. Example 5-32 changes the peer snapshot task named *task* to be submitted every three hours, and specifies that the system retain up to a maximum of 25 versions of these peer snapshots.

Example 5-32 Change a peer snapshot task

```
$ chpsnaptask task --hourInterval 3 --minuteInterval 0 --psnapLimit 25 -c
12402814423332523984
EFSSG1000I The command completed successfully.
```

You can use the **--suspend** option to prevent automatic submission of scheduled peer snapshot tasks, as shown in Example 5-33.

Example 5-33 Change a peer snapshot task by using the --suspend option

```
$ chpsnaptask task --suspend -c 12402814423332523984
EFSSG1000I The command completed successfully.
```

The task status is displayed as `Suspended` when you run **lspsnaptask** while the task is in the suspended state.

You can use the **--resume** option to change the status of a suspended scheduled peer snapshot task to `Active` so that automatic submission of its scheduled peer snapshot tasks resumes, as shown in Example 5-34.

Example 5-34 Resume a peer snapshot task

```
$ chpsnaptask task --resume -c 12402814423332523984
EFSSG1000I The command completed successfully.
```

Displaying peer snapshot task information

To display information about peer snapshot tasks, run **lspsnaptask**. You can optionally specify a file system; if no file system is specified, information for all peer snapshot tasks on the system is displayed. If you specify a file system, you can also optionally specify a file set; if you specify a file system but no file set, information for all peer snapshot tasks on the file system is displayed. Example 5-35 displays all of the peer snapshot task for the specified system.

Example 5-35 Display the peer snapshot task information

```
$ lspsnaptask -c 12402814423332523984
TaskID Filesystem Fileset HourInterval MinuteInterval PsnapLimit Status
task gpfs0 sw_fset 4 45 50 Active
EFSSG1000I The command completed successfully.
```

You can use the **-v** or **--verbose** option to include the system ID and last date and time that the task was submitted in the information that is displayed, as shown in Example 5-36.

Example 5-36 Display the verbose peer snapshot task information

```
$ lspsnaptask -c 12402814423332523984 -v
TaskID Filesystem Fileset HourInterval MinuteInterval PsnapLimit Status ClusterID
LastExecuted
task gpfs0 sw_fset 4 45 50 Active 12402814423332523984 1/1/70 1:EFSSG1000I The
command completed successfully.
```

You can use the **-Y** option to specify the display of the command output as colon-delimited fields, as shown in Example 5-37.

Example 5-37 Display the peer snapshot task information in semicolon-delimited output

```
$ lspsnaptask -c 12402814423332523984 -Y
lspsnaptask:PsnapTask:HEADER:version:reserved:reserved:TaskID:Filesystem:Fileset:H
ourInterval:MinuteInterval:lspsnaptask:PsnapTask:0:1:::task:gpfs0:sw_fset:4:45:50:
Active:
```

Removing a peer snapshot task

To remove a peer snapshot task, run **rmpsnaptask** and specify the peer snapshot task name. You must respond yes to the confirmation prompt to remove the specified peer snapshot task, as shown in Example 5-38, which removes the peer snapshot task named task.

Example 5-38 Remove a peer snapshot task

```
$ rmpsnaptask task -c 12402814423332523984
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
```

You can use the **-f** or **--force** option to prevent the normal manual removal confirmation prompt from being displayed.

Moving the Single Writer mode of a file set to a different client system

The following example scenario describes how to change dynamically the designated Single Writer role of a file set to a different remote system.

> **Tip:** You cannot change the mode of a client cache file set from Local Update mode to any other mode. You can change the mode of a client cache file set from Read Only to Single Writer and vice versa, and from both Read Only and Single Writer to Local Update.

In this example, cache system 1 is initially configured as Single Writer and cache system 2 is initially configured as Read Only. To reverse the roles so that system 2 becomes Single Writer and system 1 becomes Read Only, complete the following steps:

On client system 1:

1. Run **lswcachestate** to verify that Remote Fileset Status and Cache-Gateway Status are active, and that the queue length is zero, as shown in Example 5-39.

Example 5-39 The lswcachestate command

```
# lswcachestate gpfs0 tij_ro --connectionstatus -r
FilesetName FilesystemName Remote Fileset Path Remote Fileset Status Cache-Gateway
Assigned Cache-Gateway Status QueueLength QueueNumExec
tij_ro gpfs0 9.118.46.43:/ibm/gpfs0/tij Active int001st001 Active 0 35405
```

2. Unmount the file system that contains the cache file set to be changed.
3. Run **chwcache** to change the mode of the remote file set to Read Only, as shown in Example 5-40.

Example 5-40 The chwcache command

```
$ chwcache fileSystem filesetName --cachemode read-only
```

> **Note:** The **chwcache** command unlinks the file set, so any application attempt to access the file set results in an error. A preferred practice is to remove application access to the file set before running the command.
>
> The **chwcache** command fails if there are pending updates that are not synchronized. In this case, create a file in cache. This triggers the system to perform a recovery and resynchronization. Then, resubmit the **chwcache** command.

On the home system:

1. Run **chwcachesource** to change system 1 to Read Only, as shown in Example 5-41.

Example 5-41 The chwcachesource command

```
$ chwcachesource sharename --removeclient clientsystem1 system id
$ chwcachesource sharename --addclient 'clientsystem1 IP address(ro)'
```

2. Run **chwcachesource** to change system 2 to Single Writer (rw) mode.

On client system 2:

1. Run **runprepop** to ensure that all of the files in the file system are cached and current on client system 2, as shown in Example 5-42.

Example 5-42 The runpreop command

```
$ runprepop fileSystem filesetName policyName
```

> **Tip:** `policyName` is the name of a policy that was created with the **mkpolicy** command by using the **-R** option to create a set of rules that fetch all of the files in the file set when started by the **runprepop** command.

2. Unmount the file system that contains the cache file set to be changed.
3. Run **chwcache** to change the mode of the remote file set to Single Writer (sw), as shown in Example 5-43.

Example 5-43 The chwcache command

```
$ chwcache fileSystem filesetName --cachemode single-writer
```

> **Note:**The **chwcache** command unlinks the file set, so any application attempt to access the file set results in an error. A preferred practice is to remove application access to the file set before running the command.

5.3.7 Using Active Cloud Engine for migration

The ACE function can be used for migrating data from other heritage NAS devices to SONAS. This task is accomplished by creating an Active Cloud Engine cache relationship to the heritage NAS device that must be migrated to SONAS. At the completion of the migration procedure, the Active Cloud Engine cache file set must be converted to a normal file set.

When the migration is complete and the heritage system is removed from the infrastructure, the Active Cloud Engine function remains in, and puts the file set in, the disconnected state. In this state, any new requests try to probe the home system based on revalidation intervals to validate data. This revalidation impacts performance because of home not being available or NFS mounts failing.

Another usability aspect is that Active Cloud Engine file sets can be monitored and changed only with Active Cloud Engine commands. For example, to increase inodes, an administrator must run **chwcache** for Active Cloud Engine file sets and not **chfset**, which is used or normal regular file sets.

The Active Cloud Engine file set can be converted to a regular file set by using the **`--disablewcache`** option with the **`chwcache`** command. This option does the following things:

- It validates that the cache is RO or LU.
- For other modes (SW or IW), it fails with the message `Disable wcache is only allowed for RO/LU wcaches`.
- It verifies that there are no exports for the file set path.
- It unlinks the file set.
- It runs **`mmchfileset`** and disables the target home.
- If the home is SONAS, it deactivates the cache file set count on the home cluster. This step is not needed where the home is not SONAS.
- It updates the CMDB database tables with null values for remote variables and cache state, mode, and connection status.

The command must to be run only after all data is migrated from the heritage storage to SONAS. The function allows only an Active Cloud Engine file set to be converted to a normal file set. A normal file set cannot be converted to Active Cloud Engine file set with this function.

5.4 SONAS policy management

This section provides information about how you can create and use SONAS policies. It describes the following topics:

- Policies
- Policy rules
- Policy command-line syntax
- Creating and managing policies
- Policy creation walkthrough

5.4.1 Policies

SONAS provides a means to automate the management of files by using policies and rules. Properly managing your files allows you to use and balance efficiently your premium and less expensive storage resources.

GPFS supports these policies:

- File placement policies are used to place automatically newly created files in a specific file system.
- File management policies are used to manage files during their lifecycle by moving them to another file system pool, copying them to archival storage, changing their replication status, or deleting them.

A policy is a set of rules that describes the lifecycle of user data based on the file's attributes. Each rule defines an operation or definition, such as migrate to a pool and replicate the file. There are three uses for rules:

- Initial file placement
- File management
- Restoring file data

Placement policy

When a file is created or restored, the placement policy determines the location of the file's data and assigns the file to a file system pool. All data that is written to that file is placed in the assigned file system pool.

The placement policy that defines the initial placement of newly created files and the rules for placement of restored data must be installed into SONAS. If a SONAS system does not have a placement policy that is installed, all the data is stored in the system pool. Only one placement policy can be installed at a time. If you switch from one placement policy to another, or change a placement policy, that action has no effect on existing files. However, newly created files are always placed according to the currently installed placement policy.

Management policy

The management policy determines file management operations, such as migration and deletion. You can define the file management rules and install them in the file system together with the placement rules. In either case, policy rules for placement or migration can be intermixed. Over the life of the file, data can be migrated to a different pool any number of times, and files can be deleted or restored. File management rules can also be used to control the space usage of online file system pools. When the usage for an online pool exceeds the specified high threshold value, SONAS can be configured to trigger an event that can automatically start a policy and reduce the utilization of the pool.

Error checking for file-placement policies

SONAS does error checking for file-placement policies in the following phases:

- When you install a new policy, GPFS checks the basic syntax of all the rules in the policy.
- GPFS also checks all references to file system pools. If a rule in the policy refers to a file system pool that does not exist, the policy is not installed and an error is returned.
- When a new file is created, the rules in the active policy are evaluated in order. If an error is detected, all the subsequent rules are skipped, and SONAS returns an EINVAL error code to the application.
- Otherwise, the first applicable rule is used to store the file data.

Default file-placement policy: When a file system is first created, the default file-placement policy is to assign all files to the system storage pool.

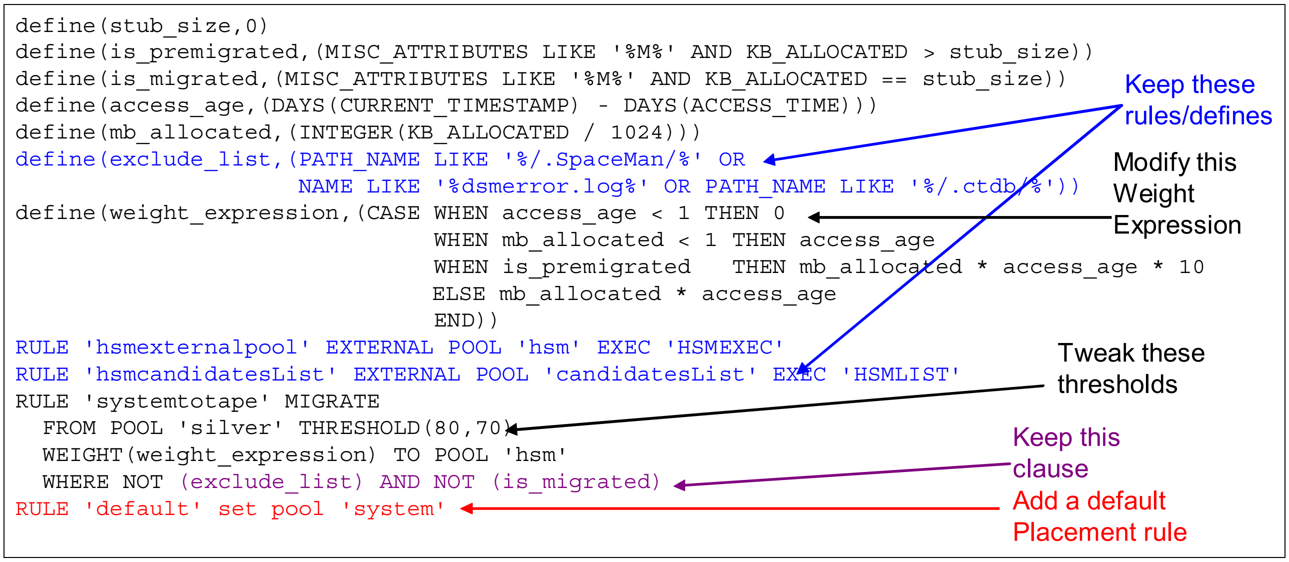
5.4.2 Policy rules

A policy rule is an SQL-like statement that tells SONAS system what to do with the data for a file in a specific pool if the file meets specific criteria. A rule can apply to any file that is being created or only to files that are being created within a specific file set or group of file sets.

Rules specify conditions that, when true, cause the rule to be applied. Here are some of these conditions:

- Date and time when the rule is evaluated, that is, the current date and time
- Date and time when the file was last accessed
- Date and time when the file was last modified
- File set name
- File name or extension
- File size
- User ID and group ID

GPFS evaluates policy rules in order, from first to last, as they appear in the policy. The first rule that matches determines what is to be done with that file. For example, when a client creates a file, GPFS scans the list of rules in the active file-placement policy to determine which rule applies to the file. When a rule applies to the file, GPFS stops processing the rules and assigns the file to the appropriate pool. If no rule applies, an EINVAL error code is returned. For details, see Figure 5-20.



```
define(stub_size,0)
define(is_premigrated,(MISC_ATTRIBUTES LIKE '%M%' AND KB_ALLOCATED > stub_size))
define(is_migrated,(MISC_ATTRIBUTES LIKE '%M%' AND KB_ALLOCATED == stub_size))
define(access_age,(DAYS(CURRENT_TIMESTAMP) - DAYS(ACCESS_TIME)))
define(mb_allocated,(INTEGER(KB_ALLOCATED / 1024)))
define(exclude_list,(PATH_NAME LIKE '%/.SpaceMan/%' OR
                     NAME LIKE '%dsmerror.log%' OR PATH_NAME LIKE '%/.ctdb/%'))
define(weight_expression,(CASE WHEN access_age < 1 THEN 0
                               WHEN mb_allocated < 1 THEN access_age
                               WHEN is_premigrated   THEN mb_allocated * access_age * 10
                               ELSE mb_allocated * access_age
                               END))
RULE 'hsmexternalpool' EXTERNAL POOL 'hsm' EXEC 'HSMEXEC'
RULE 'hsmcandidatesList' EXTERNAL POOL 'candidatesList' EXEC 'HSMLIST'
RULE 'systemtotape' MIGRATE
  FROM POOL 'silver' THRESHOLD(80,70)
  WEIGHT(weight_expression) TO POOL 'hsm'
  WHERE NOT (exclude_list) AND NOT (is_migrated)
RULE 'default' set pool 'system'
```

Figure 5-20 Sample policy syntax constructs

Policy rule types

There are eight types of policy rules that allow you to define specific actions that GPFS can implement on the file data. Each rule has clauses that control candidate selection, namely when the rule is allowed to match a file, what files it matches, the order to operate on the matching files, and more attributes to show for each candidate file. Different clauses are permitted on different rules that are based on the semantics of the rule.

Here is an explanation of these rules and their respective syntax diagrams:

1. File-placement rules.

   File-placement rules specify which file system and allocation are used upon file creation. Changing the rules that apply to a file's placement does not cause the file to be moved. See Example 5-44.

   Example 5-44 Placement rule

   ```
   RULE ['RuleName']
      SET POOL 'PoolName'
         [LIMIT (OccupancyPercentage)]
         [REPLICATE (DataReplication)]
         [FOR FILESET (FilesetName[,FilesetName]...)]
      [WHERE SqlExpression]
   ```

   Example 5-45 creates a rule that is named `datfiles`. The newly created `.dat` files are saved to storage pool named `poolfordatfiles`.

   Example 5-45 File-placement rule example

   ```
   RULE 'datfiles' SET POOL 'poolfordatfiles' WHERE UPPER(name) like '%.DAT'
   ```

2. File migration rules. File migration rules allowed the policy manager to coordinate file migrations from one file system to another file system pool or to external pools. See Example 5-46.

Example 5-46 File migration rule

```
RULE ['RuleName'] [WHEN TimeBooleanExpression]
   MIGRATE
      FROM POOL 'FromPoolName'
      [THRESHOLD (HighPercentage[,LowPercentage[,PremigratePercentage]])]]
      [WEIGHT (WeightExpression)]
   TO POOL 'ToPoolName'
      [LIMIT (OccupancyPercentage)]
      [REPLICATE (DataReplication)]
      [FOR FILESET (FilesetName[,FilesetName]...)]
      [SHOW (['String'] SqlExpression)]
      [SIZE (numeric-sql-expression)]
      [WHERE SqlExpression]
```

In Example 5-47, there is no `FROM POOL` clause, so regardless of their current storage pool placement, all files from the named file sets are subject to migration to storage pool pool2.

Example 5-47 Sample migration rule syntax

```
RULE 'migraterule' MIGRATE TO POOL 'pool2' FOR FILESET('root','fset1')
```

3. File deletion rules. A file that matches this rule becomes a candidate for deletion. See Example 5-48.

Example 5-48 File deletion rule

```
RULE ['RuleName'] [WHEN TimeBooleanExpression]
   DELETE
      [FROM POOL 'FromPoolName'
      [THRESHOLD (HighPercentage[,LowPercentage])]]
      [WEIGHT (WeightExpression)]
      [FOR FILESET (FilesetName[,FilesetName]...)]
      [SHOW (['String'] SqlExpression)]
      [SIZE (numeric-sql-expression)]
      [WHERE SqlExpression]
```

The rule in Example 5-49 creates a rule that is named `mpg`. All files have the `.mpg` extension, and file sizes that are larger than 20123456 are deleted from the system.

Example 5-49 Sample DELETE rule syntax

```
RULE 'mpg' DELETE WHERE lower(NAME) LIKE '%.mpg' AND FILE_SIZE>20123456
```

4. File exclusion rules. A file that matches this rule is excluded from further rule evaluation. When specified in a LIST rule, EXCLUDE indicates that any matching files be excluded from the list. See Example 5-50.

Example 5-50 File exclusion rule

```
RULE ['RuleName'] [WHEN TimeBooleanExpression]
   EXCLUDE
      [FROM POOL 'FromPoolName']
      [FOR FILESET (FilesetName[,FilesetName]...)]
   [WHERE SqlExpression]
```

The example rule in Example 5-51, called Xsuper, excludes all .mpg files that belong to USERID 200 from deletion.

Example 5-51 Sample EXCLUDE rule syntax

```
RULE 'Xsuper' EXCLUDE WHERE USER_ID=200
RULE 'mpg' DELETE WHERE lower(NAME) LIKE '%.mpg' AND FILE_SIZE>20123456
```

Tip: Specify the EXCLUDE rule before rules that might match the file that is being excluded. You cannot define a list and what to exclude from the list in a single rule. You must define two LIST statements, one specifying which files are in the list, and one specifying what to exclude from the list.

5. File list rules. File list rules identify a file list generation rule. A particular file might match more than one list rule, but are included in a particular list only once. ListName provides the binding to an EXTERNAL LIST rule that specifies the executable program to use when you process the generated list. See Example 5-52.

Example 5-52 File list rule

```
RULE ['RuleName'] [WHEN TimeBooleanExpression]
   LIST 'ListName'
      EXCLUDE
      [DIRECTORIES_PLUS]
      [FROM POOL 'FromPoolName'
      [THRESHOLD (HighPercentage[,LowPercentage])]]
      [WEIGHT (WeightExpression)]
      [FOR FILESET (FilesetName[,FilesetName]...)]
      [SHOW (['String'] SqlExpression)]
      [SIZE (numeric-sql-expression)]
   [WHERE SqlExpression]
```

Example 5-53 shows how to exclude files that contain the word test from the LIST rule named allfiles.

Example 5-53 Sample EXTERNAL RULE syntax

```
RULE EXTERNAL LIST 'allfiles' EXEC '/u/brownap/policy/CHE/exec.list'
RULE 'exclude_allfiles' LIST 'allfiles' EXCLUDE where name like '%test%'
```

6. File restore rules. When a file is restored, the placement policy determines the location of the file's data and assigns the file to a file system pool. See Example 5-54 on page 351.

Example 5-54 Fire restore rule

```
RULE ['RuleName']
  RESTORE TO POOL 'PoolName'
    [LIMIT (OccupancyPercentage)]
    [REPLICATE (DataReplication)]
    [FOR FILESET (FilesetName[,FilesetName]...)]
 [WHERE SqlExpression]
```

Example 5-55 shows a rule to restore files and assign them to the system pool.

Example 5-55 Sample RESTORE syntax

```
RULE ’RestFromExt’ RESTORE TO POOL ’system’ WHERE ...
```

7. External storage pool definition rules. This type of rule defines an external file system pool. This rule does not match files, but instead defines the binding between the policy language and the external storage manager that implements the external storage. See Example 5-56.

Example 5-56 External storage pool definition rule

```
RULE [’RuleName’]
   EXTERNAL POOL ’PoolName’
      EXEC ’InterfaceScript’
      [OPTS ’OptionsString ...’]
      [ESCAPE ’%SpecialCharacters’]
   [SIZE sum-number]
```

Example 5-57 shows a rule to define the `hsm` external file system pool.

Example 5-57 Sample EXTERNAL POOL syntax

```
RULE 'hsmexternalpool' EXTERNAL POOL 'hsm' EXEC 'HSMEXEC' OPTS '-l LOGID'
```

Rules must adhere to a specific syntax, as documented in the SONAS IBM Knowledge Center. This syntax is similar to the SQL language because it contains statements such as `WHEN (TimeBooleanExpression)` and `WHERE SqlExpression`. Rules also contain SQL expression clauses that allow you to reference various file attributes as SQL variables and combine them with SQL functions and operators. Depending on the clause, an SQL expression must evaluate to either true or false, a numeric value, or a character string. Not all file attributes are available to all rules.

Macro defines

Policies can be coded by using defines, which also are called macro defines. These defines are named variables that are used to make rules easier to read. For example, the statement in Example 5-58 creates a define that is named mb_allocated and sets it to the size of the file in MB.

Example 5-58 Macro define statement

```
define(mb_allocated,(INTEGER(KB_ALLOCATED / 1024)))
```

Defines offer a convenient way to encapsulate *weight expressions* to provide common definitions across the policy. These common exclusions are typical:

- The “special file” migration exclusion definition: Always use it when you are migrating.
- The “migrated file” migration exclusion definition: Always use it when you are migrating.

Peered policies

Peered policies contain placement rules only. Defines are not required for peered ILM policies. Placement rules select files by user-defined criterion or policy (see Example 5-59).

Example 5-59 Sample placement rule

```
RULE 'P1' set pool 'system' where upper(name) like '%SO%'
RULE 'P1' set pool 'system' where upper(name) like '%TOTALLY%'
```

Peered pools must contain a default placement rule that by default puts files in the lower performance pool, and then select groups of files by using rules for placement into the higher performance pool. A sample placement rule is shown in Example 5-60.

Example 5-60 Sample default placement rule

```
RULE 'default' set pool 'slowpool'
```

Tiered policies

Tiered policies contain both migration rules and optional placement rules. This type of policy requires the defines that are contained in the sample TEMPLATE-ILM policy. You can also encapsulate weight expression as a define. Optional placement rules select files by policy. Here are some preferred practices for migration rules:

- Make sure that at least one threshold exists as a safety net, even if you are using other rules.
- Include exclusion clauses for migrated and special files in migration rules even if you are not using HSM, so they can be added later.
- Non-threshold migration needs an associated **cron** job to trigger it, as described later for migration filters.

The policy is terminated by the default placement rule that is shown in Example 5-61.

Example 5-61 Default placement rule

```
RULE 'default' set pool 'system'
```

In the example, the default setting of a higher performance pool is used because subsequent tiering cascades data from high-performance to low-performance pools.

HSM policies

Use the defines from the TEMPLATE-HSM rules. You can again encapsulate weight expression as a define and optionally have placement rules to select files by policy.

Follow these preferred practices for the migration rules:

- External pool rules: Use rules from template.
- Threshold: Make sure at least one exists as a safety net even if you are using other rules.
- Always include exclusion clauses (migrated and special files) in migration rules.
- Non-threshold migration: This setting needs an associated **cron** job to trigger it. You might want to have a “time” clause to prevent running on threshold trigger.
- Levels: Define at least one rule for each migration “level” (system → pool2, pool2 → hsm).
- External pool rules: Use rules from a template.

Remember to terminate the policy with a default placement rule.

Remote caching policies

Here are considerations for remote caching policies:

- Periodically run parallel inode scans at home.
- Select files and directories based on policy criteria.
- Include user-defined metadata in xattrs or other file attributes.
- Use a SQL like construct to select the caching policy, as shown in Example 5-62.

Example 5-62 Sample remote caching policy

```
RULE LIST ‘prefetchlist' WHERE FILESIZE > 1GB  AND MODIFICATION_TIME  >
CURRENT_TIME- 3600 AND USER_ATTR1 = “sat-photo” OR USER_ATTR2 = “classified”
```

- The cache pre-fetches selected objects
- Run asynchronously in the background
- Parallel multi-node prefetch
- Can callout when complete

Policy triggers

Policies can be applied to a file system or be only in the SONAS database.

The *File system policy* trigger can be “active”. One policy can be used per file system. It is loaded from the database (**setpolicy**).

The database policies can have one of the following states:

- “inactive”: They are not running
- “default”: Quick path to recalling a policy; it is a database state only.

Triggers control when policies are activated. Policies do something only if they are triggered. SONAS provides the following types of triggers:

- Manual trigger: The **runpolicy** command allows a database policy to be run.
- Automated triggers, also referred to as callbacks, are triggered by a threshold:
  - The SONAS GPFS file system manager detects that disk space is running below the low threshold that is specified in the current policy rule, and raises a *lowDiskSpace* event.
  - The lowDiskSpace event initiates a SONAS GPFS migration callback procedure.
  - The SONAS GPFS migration callback runs the SONAS script that is defined for that callback.
  - The SONAS script runs the active file system policy.
- Cron: In SONAS, **cron** activates the default file system policy.

When SONAS identifies that a threshold is reached, it triggers a new lowspace event every two minutes while the fill level of the file system is above the threshold. SONAS knows that a migration was already triggered, so it ignores the new trigger and it does not do any additional processing; the migration that started earlier continues execution.

Weight expressions

Weight expressions are used with threshold migration rules. The threshold limits the amount of data that is moved and the weight expression determines the order of files that are being migrated so that files with the highest weight are moved first and until the threshold is satisfied.

Code the weight expression as a define because it makes a rule easier to read, as shown in the rule in Example 5-63.

Example 5-63 Threshold migration rule

```
RULE 'systemtosilver' MIGRATE FROM POOL 'system' THRESHOLD(15,10)
WEIGHT(weight_expression) TO POOL 'silver' WHERE NOT (exclude_list) AND NOT
(is_migrated)
```

The weight expression is shown in Example 5-64.

Example 5-64 Define a weight expression

```
define(weight_expression,(CASE WHEN access_age < 1 THEN 0  WHEN mb_allocated < 1
THEN access_age WHEN is_premigrated   THEN mb_allocated * access_age * 10 ELSE
mb_allocated * access_age  END))
```

The previous two statements are simpler to read than the combined statements in Example 5-65.

Example 5-65 Combined threshold migration rule and weight expression

```
RULE 'systemtosilver' MIGRATE FROM POOL 'system' THRESHOLD(15,10) WEIGHT(CASE WHEN
access_age < 1 THEN 0  WHEN mb_allocated < 1 THEN access_age WHEN is_premigrated
THEN mb_allocated * access_age * 10 ELSE mb_allocated * access_age  END) TO POOL
'silver' WHERE NOT (exclude_list) AND NOT (is_migrated)
```

Migration filters

Migration filters are used to control what gets migrated and when. Exclusion rules, or filters, must include the following files:

- Migrated and special files: These files must be used from the templates.
- Optionally, small files: Leave small files behind for efficiency if they can fit on disk (a threshold plus a weight rule might do this anyway, so this might not be a useful rule).
- The fine print: Small files are not migrated to auxiliary storage, and cannot be recovered from the auxiliary storage. Although HSM can be used to recover files, it is *not* preferable and is *not* supported as a customer action. Customers must be using backup/restore; in that case, if they run coupled with backup, the small files are backed up, just not migrated.

Time filters can be useful when coupled with **cron** jobs, for example, running a **cron** every Sunday at 4:05 AM; perhaps you are flushing many files that are not accessed for a week.

5.5 Creating and managing policies

This section describes what policies and rules consist of, including examples of policies and rules, and it describes the SONAS commands that manage policies and rules. It illustrates how to create a file system pool and extend a file system to use the file system pool. It then shows how to create and apply data allocation policies. For more information, see the SONAS information in the IBM Knowledge Center, found at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_policies_topic_welcome.html?lang=en

5.5.1 Setting up policies with the GUI

To set up policies on a file system, complete the following steps:

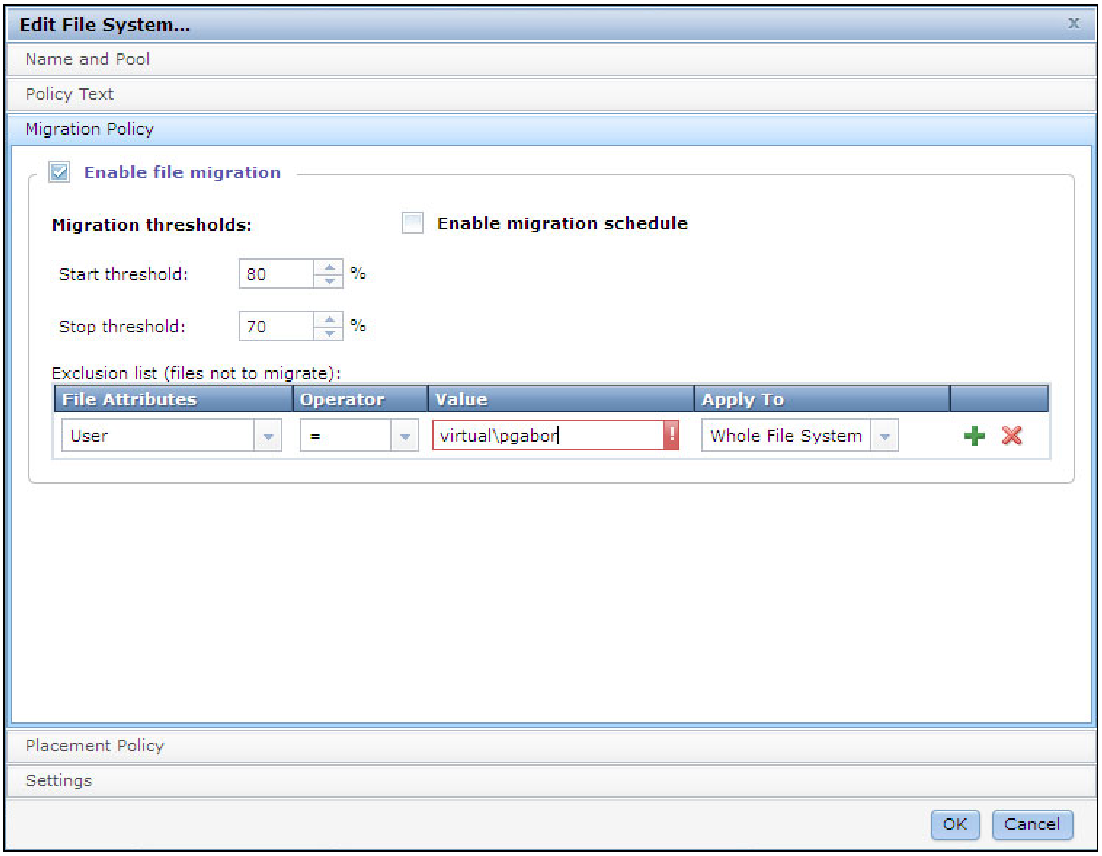
1. To set up policies on a file system, log in to the SONAS GUI and click **Files** → **File Systems**.
2. To modify the existing file system, select the file system and then click **Actions** → **Edit File System**. Click the Migration Policy bar to expand it.

3. In the expanded Migration Policy window that is shown in Figure 5-21, complete the following tasks:

   – Select the **Enable file migration** check box.

   – Set up the migration threshold values. In this example, the start threshold value is 80% and the stop threshold value is 70%. If the file system becomes 80% full, the migration starts until it drops down to 70%.You can set a up migration schedule and also an exclude list.

     For migration and exclude list preferred practices, see *IBM SONAS Best Practices*, SG24-8051.



☑ Enable file migration

Migration thresholds:

Start threshold: 80 %

Stop threshold: 70 %

☐ Enable migration schedule

Exclusion list (files not to migrate):

File Attributes	Operator	Value	Apply To
User	=	virtual\pgabor	Whole File System

Figure 5-21 Edit file system window - Migration Policy setup view

4. To set up the file placement rule, click the **Placement Policy** header. In the expanded Placement Policy window that is shown in Figure 5-22 on page 357, complete the following tasks:

   – Select the **Enable file placement** check box.

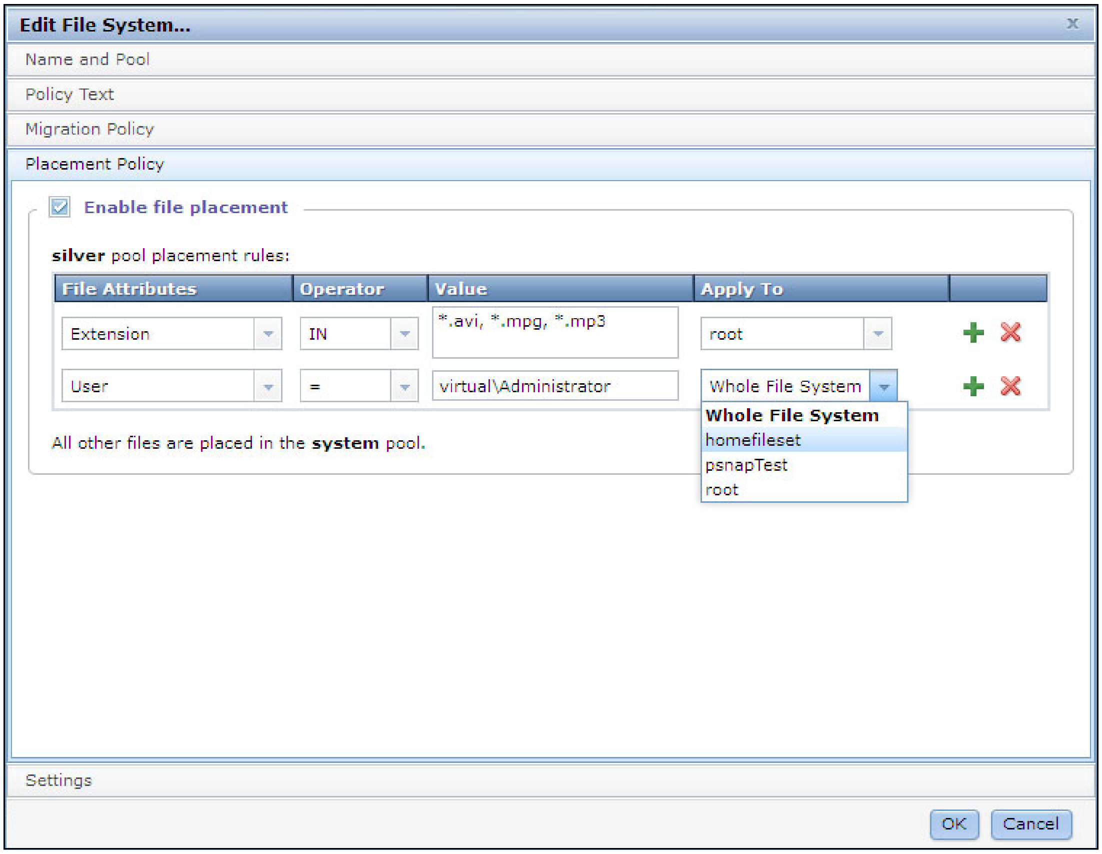
     In the GUI, there are only a few predefined file attributes:

     - Extension
     - File Set Name
     - Group
     - User

     In this example, you are creating two placement policies based on the Extension and User file attributes. All files that have the avi, mpg, or mp3 extension and all files that are owned by virtual\Administrator are placed into the silver pool.

More detailed placement policies can be created from the CLI. For examples, see *IBM SONAS Best Practices*, SG24-8051.

Figure 5-22 shows the Edit File System window.



Edit File System...

Name and Pool

Policy Text

Migration Policy

Placement Policy

☑ Enable file placement

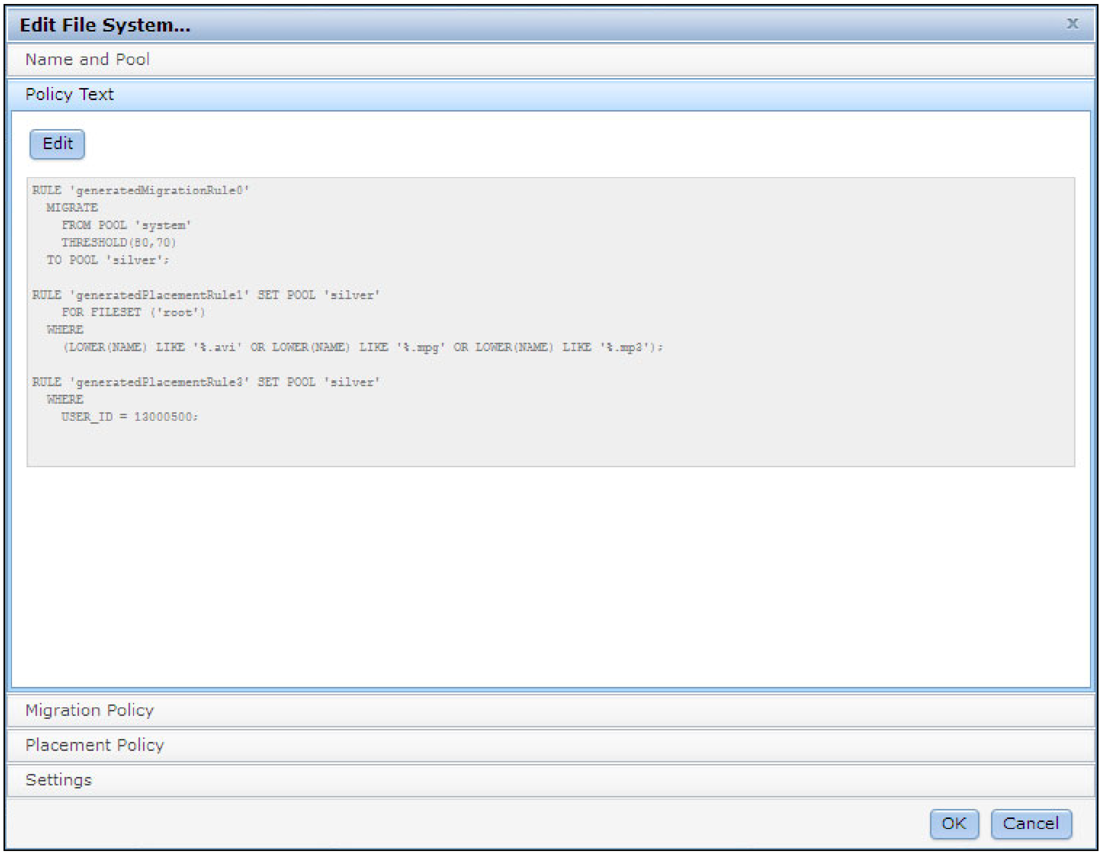
silver pool placement rules:

File Attributes	Operator	Value	Apply To
Extension	IN	*.avi, *.mpg, *.mp3	root
User	=	virtual\Administrator	Whole File System

All other files are placed in the **system** pool.

Figure 5-22 Edit file system window - placement policy setup view

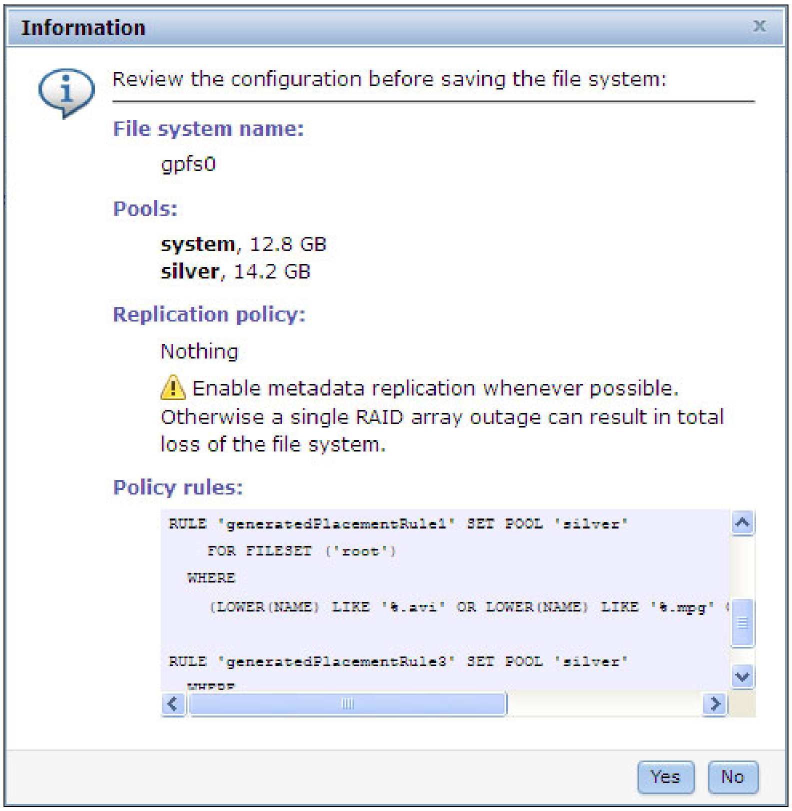
5. You can review the policy in the Policy text by clicking the **Policy Text** header, as shown in Figure 5-23. Click **OK** to save and activate the policies.



```
RULE 'generatedMigrationRule0'
  MIGRATE
    FROM POOL 'system'
    THRESHOLD(80,70)
  TO POOL 'silver';

RULE 'generatedPlacementRule1' SET POOL 'silver'
    FOR FILESET ('root')
  WHERE
    (LOWER(NAME) LIKE '%.avi' OR LOWER(NAME) LIKE '%.mpg' OR LOWER(NAME) LIKE '%.mp3');

RULE 'generatedPlacementRule3' SET POOL 'silver'
  WHERE
    USER_ID = 13000500;
```

Figure 5-23 Edit file system window - Policy Text view

6. An information window opens, as shown in Figure 5-24 on page 359. Review the configuration before you save the file system. Click **Yes** to finalize the settings.



Review the configuration before saving the file system:

File system name:

gpfs0

Pools:

system, 12.8 GB
silver, 14.2 GB

Replication policy:

Nothing

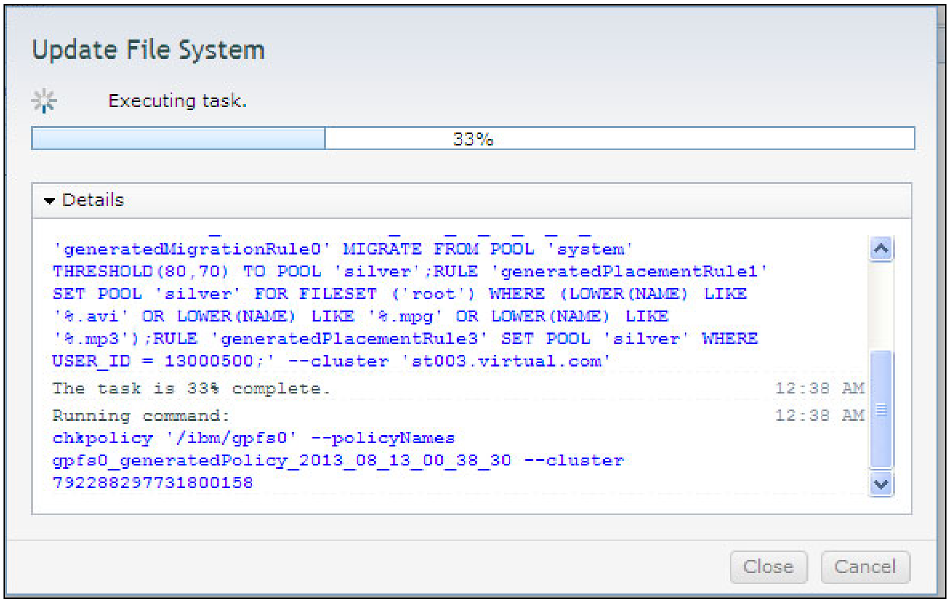
⚠ Enable metadata replication whenever possible. Otherwise a single RAID array outage can result in total loss of the file system.

Policy rules:

```
RULE 'generatedPlacementRule1' SET POOL 'silver'
    FOR FILESET ('root')
  WHERE
    (LOWER(NAME) LIKE '%.avi' OR LOWER(NAME) LIKE '%.mpg' (

RULE 'generatedPlacementRule3' SET POOL 'silver'
  WHERE
```

Figure 5-24 Configuration review window

During the setup, a progress window opens, as shown in Figure 5-25. Click **Close** to close the window.



Executing task.

33%

```
'generatedMigrationRule0' MIGRATE FROM POOL 'system'
THRESHOLD(80,70) TO POOL 'silver';RULE 'generatedPlacementRule1'
SET POOL 'silver' FOR FILESET ('root') WHERE (LOWER(NAME) LIKE
'%.avi' OR LOWER(NAME) LIKE '%.mpg' OR LOWER(NAME) LIKE
'%.mp3');RULE 'generatedPlacementRule3' SET POOL 'silver' WHERE
USER_ID = 13000500;' --cluster 'st003.virtual.com'
The task is 33% complete.                                   12:38 AM
Running command:                                            12:38 AM
chkpolicy '/ibm/gpfs0' --policyNames
gpfs0_generatedPolicy_2013_08_13_00_38_30 --cluster
792288297731800158
```

Figure 5-25 Policy setup progress window

5.5.2 Setting up policies by using the CLI

The CLI commands and output in the following examples show a quick walkthrough of how to set up a second storage pool by using CLI commands:

1. Example 5-66 shows how to create a new pool that is named `silver` by running **`lsdisk`** to list all the available disks. The **`chdisk`** and **`chfs`** commands are used to create the pool.

Example 5-66 Second storage pool creation example with the CLI

```
[st003.mgmt001st003 ~]# lsdisk
Name                           File system Failure group Type          Pool
Status Availability Timestamp
array0_sas_60001ff07975001890901  gpfs0     1         dataAndMetadata system
ready  up          8/13/13 10:55 PM
array0_sas_60001ff07975001890902  gpfs0     1         dataAndMetadata system
ready  up          8/13/13 10:55 PM
array0_sas_60001ff07975001890903  gpfs0     1         dataAndMetadata system
ready  up          8/13/13 10:55 PM
array0_sas_60001ff07975001890907  gpfs0    1                          system
                   8/13/13 10:55 PM
array0_sas_60001ff07975001890908            1                        system
ready              8/13/13 10:55 PM
..........
EFSSG1000I The command completed successfully.

[st003.mgmt001st003 ~]# chdisk array0_sas_60001ff07975001890907 --pool silver
--usagetype dataOnly

[st003.mgmt001st003 ~]# chfs gpfs0 --add array0_sas_60001ff07975001890907 --force
--noverify
[st003.mgmt001st003 ~]# lspool
Filesystem Name   Size     Usage Available fragments Available blocks Disk list
gpfs0      silver 14.24 GB 0%    256 kB 14.24 GB array0_sas_60001ff07975001890907
gpfs0      system 12.78 GB 3.6%  2.10 MB 12.32 GB
array0_sas_60001ff07975001890901;array0_sas_60001ff07975001890902;array0_sas_60001
ff07975001890903
EFSSG1000I The command completed successfully.
```

2. Run **`mkpolicy`** and specify a policy name to create a policy as shown in Example 5-67. You must specify rules by using the **`-R`** or **`--rules`** option. If you specify more than one rule, multiple rules must be separated by the semicolon character. The entire rule, or set of rules, must be enclosed in double quotation characters.

   In this scenario, the policy name is `policysample`. One migration rule based on threshold and two placement rules based on extension and user ID are created.

Example 5-67 CLI policy creation example by running mkpolicy

```
[st003.mgmt001st003 ~]# mkpolicy policysample -R "\
RULE 'migrationrule' MIGRATE FROM POOL 'system' THRESHOLD(80,70) TO POOL 'silver'; \
RULE 'placementrule' SET POOL 'silver' WHERE (LOWER(NAME) LIKE '%.avi' OR LOWER(NAME) LIKE
'%.mpg' OR LOWER(NAME) LIKE '%.mp3'); \
RULE 'placementbyID' SET POOL 'silver' WHERE USER_ID = 13000500"
EFSSG1000I The command completed successfully.
```

3. The next step is to verify the policy. Running **lspolicy** with the **-P policyname** parameter lists your policy rules, as shown in Example 5-68. In the example, all of the rules only for the policy that is named `policysample` are listed.

Example 5-68 Policy rule listing example

```
[st003.mgmt001st003 ~]# lspolicy -P policysample
Policy Name  Declaration Name Default Declarations
policysample migrationrule    N       RULE 'migrationrule' MIGRATE FROM POOL 'system'
THRESHOLD(80,70) TO POOL 'silver'
policysample placementrule    N       RULE 'placementrule' SET POOL 'silver' WHERE
(LOWER(NAME) LIKE '%.avi' OR LOWER(NAME) LIKE '%.mpg' OR LOWER(NAME) LIKE '%.mp3')
policysample placementbyID    N       RULE 'placementbyID' SET POOL 'silver' WHERE USER_ID
= 13000500
EFSSG1000I The command completed successfully.
```

4. It is a preferred practice to first validate a policy by running **chkpolicy** before setting the policy. This action ensures that the policy does what is expected. For example, to test the policy that is named `policysample` against the `gpfs0` file system, run the command that is shown in Example 5-69.

Example 5-69 Policy validating example

```
[st003.mgmt001st003 ~]# chkpolicy gpfs0 -P policysample
EFSSG1000I The command completed successfully.
```

5. To set the active policy for the file system, run **setpolicy**. For example, to set a policy named `policysample` as the policy for the file system that is named `gpfs0`, run the command that is shown in Example 5-70.

Example 5-70 Policy activation example

```
[st003.mgmt001st003 ~]# setpolicy gpfs0 -P policysample
EFSSG1000I The command completed successfully.
```

6. Run **lspolicy -A** to list the policies that are applied on the system. The policy information that is listed includes the policy name, the time when the policy was applied, the user who applied the policy, and the device on which the policy is applied, as shown in Example 5-71.

Example 5-71 List the applied policy example

```
[st003.mgmt001st003 ~]# lspolicy -A
Cluster          Device Policies     Applied Time     Who applied it?
st003.virtual.com gpfs0  policysample 8/13/13 10:58 PM cliuser1
EFSSG1000I The command completed successfully.
```

7. Run **lspolicy -D** to list all the policies on a device and to make sure that the default rule exists, as shown in Example 5-72.

Example 5-72 List the policy for a device example

```
[st003.mgmt001st003 ~]# lspolicy -D gpfs0
RULE 'migrationrule' MIGRATE FROM POOL 'system' THRESHOLD(80,70) TO POOL 'silver'
RULE 'placementrule' SET POOL 'silver' WHERE (LOWER(NAME) LIKE '%.avi' OR
LOWER(NAME) LIKE '%.mpg' OR LOWER(NAME) LIKE '%.mp3')
RULE 'placementbyID' SET POOL 'silver' WHERE USER_ID = 13000500
RULE 'default' SET POOL 'system'
EFSSG1000I The command completed successfully.
```

5.5.3 Managing policies

This section describes commands to manage policies. It provides sample syntax.

The mkpolicy command

The **mkpolicy** command creates a policy template with a name and a list of one or more rules. The policy and rules are stored in the SONAS management database. A validation of the rules is not done at creation. The command syntax is shown in Example 5-73.

Example 5-73 Sample mkpolicy command

```
mkpolicy policyName [-CP <policyName> | -R <rules>] [-D]
```

The policy has a name and a set of rules that are specified with the **-R** option. The **-D** option sets the default policy for a file system. Optionally, a policy can be created by copying an existing policy or a predefined policy template by running **mkpolicy -CP oldpolicy**. The policy is later applied to a SONAS file system.

The rules for a policy must be entered as a single string and separated by semicolons, and there must be no leading or trailing blanks that surround the semicolons. This task can be accomplished one of two ways:

1. The first method is to enter the rule as a single long string.
2. The second method uses the Linux line continuation character (backslash) to enter rules, as shown in Example 5-74.

Example 5-74 Rule that is entered by using the continuation character

```
mkpolicy ilmtest -R "\
RULE 'gtktosilver' SET POOL 'silver' WHERE NAME LIKE '%gtk%';\
RULE 'ftktosystem' SET POOL 'system' WHERE NAME LIKE '%ftk%';\
RULE 'default' SET POOL 'system'"
```

The chpolicy command

The **chpolicy** command modifies an existing policy by adding, appending, or deleting rules. The **rmpolicy** command can remove a policy from the SONAS database, but it does not remove a policy from a file system.

To change a policy, run **chpolicy**, specify the policy name, and use either the **--add** or **--remove** option. Run the commands that are shown in Example 5-75 to change a policy that is named `hsmpolicy`, to remove the current `systemtotape` rule, and add a rule that is named `systemtotape`.

Example 5-75 Sample chpolicy command

```
# chpolicy hsmpolicy --remove systemtotape
# chpolicy hsmpolicy --add "RULE 'systemtotape' MIGRATE FROM POOL 'system'
THRESHOLD(80,70) WEIGHT(weight_expression) TO POOL 'hsm' WHERE NOT (exclude_list)
AND NOT (is_migrated)"
```

You can also add a rule and insert it before another rule by using the **--before** option. For example, to include rule that is named `PoolXtotape` before rule named `systemtotape`, submit the command that is shown in Example 5-76 on page 363.

Example 5-76 Sample chpolicy command that uses the --before parameter

```
# chpolicy hsmpolicy --add "RULE 'PoolXtotape' MIGRATE FROM POOL 'poolX' TO POOL
'hsm' WHERE NOT (is_migrated)" --before 'systemtotape'"
```

Example 5-77 is an example of the **chpolicy** command with the **--add** option.

Example 5-77 Sample chpolicy command that uses the --add parameter

```
[st002.virtual.com]$ chpolicy ilmtest --add "RULE 'default' SET POOL 'system'"
EFSSG1000I The command completed successfully.
[st002.virtual.com]$ lspolicy -P ilmtest
Policy Name Declaration Name Default Declarations
ilmtest     gtktosilver      N       RULE 'gtktosilver' SET POOL 'silver' WHERE
NAME LIKE '%gtk'
ilmtest     default          N       RULE 'default' SET POOL 'system'
EFSSG1000I The command completed successfully.
```

The chkpolicy command

The **chkpolicy** command allows you to check policy syntax and to test the policy, as shown in Example 5-78.

Example 5-78 Sample chkpolicy command testing the policy syntax

```
chkpolicy deviceOrPath  -P <policyNames> [-T] [-N <nodeNames>] [-c < clusterID |
clusterName >]
```

Here, **<deviceOrPath>** specifies the file system and **<policyNames>** specifies the policy that is contained in the database to be tested. Without the **-T** option, the policy is checked only for correctness against the file system. Using the **-T** option tests the policy and outputs the result of applying the policy to the file system and showing which files are migrated. The **<nodeNames>** parameter indicates whether the **chkpolicy** command should be run on the specified list of nodes of the cluster where this option is used. If no cluster nodes are specified, the **chkpolicy** command is run either on an arbitrary node from the list of HSM enabled nodes or on an arbitrary node. Run **showlog** to view the output from the **chkpolicy** command.

Example 5-79 shows a **chkpolicy** command with the **showlog** command specified.

Example 5-79 Check policies for correctness example

```
[st002.virtualad.ibm.com]$ chkpolicy gpfs0 -P ilmtest -T
EFSSA0184I The policy is started on gpfs0 with JobID 2.
[st002.virtual.com]$ showlog 2
Primary node: mgmt002st002
Job ID : 2
[I] GPFS Current Data Pool Utilization in KB and %
system  697344  2342912 29.763986%
[I] 4156 of 275968 inodes used: 1.505972%.
[I] Loaded policy rules from /var/opt/IBM/sofs/PolicyFiles/policy1319581067426.
Evaluating MIGRATE/DELETE/EXCLUDE rules with CURRENT_TIMESTAMP =
2011-10-25@22:17:48 UTC
parsed 1 Placement Rules, 0 Restore Rules, 0 Migrate/Delete/Exclude Rules,
        0 List Rules, 0 External Pool/List Rules
RULE 'gtktosilver' SET POOL 'silver' WHERE NAME LIKE '%gtk'
```

```
[I] Directories scan: 53 files, 36 directories, 0 other objects, 0 'skipped' files
and/or errors.
[I] Inodes scan: 49 files, 34 directories, 6 other objects, 12 'skipped' files
and/or errors.
[I] Summary of Rule Applicability and File Choices:
 Rule#  Hit_Cnt KB_Hit  Chosen  KB_Chosen       KB_Ill  Rule
[I] Filesystem objects with no applicable rules: 83.
[I] GPFS Policy Decisions and File Choice Totals:
 Chose to migrate 0KB: 0 of 0 candidates;
 Chose to premigrate 0KB: 0 candidates;
 Already co-managed 0KB: 0 candidates;
 Chose to delete 0KB: 0 of 0 candidates;
 Chose to list 0KB: 0 of 0 candidates;
 0KB of chosen data is illplaced or illreplicated;
Predicted Data Pool Utilization in KB and %:
system  697344  2342912 29.763986%
----------------------------------------------------------
End of log - chkpolicy completed
----------------------------------------------------------
EFSSG1000I The command completed successfully.
```

The lspolicy command

Multiple named policies can be stored in the SONAS database. Policies can be listed by running **`lspolicy`**. Running **`lspolicy`** without arguments returns the name of all the policies that are stored in the SONAS database. Specifying **`-P policyname`** lists all the rules in a policy, and specifying **`lspolicy -A`** lists file systems with applied policies. Example 5-80 shows the **`list`** command output.

Example 5-80 List policies

```
[st002.virtualad.ibm.com]$ lspolicy
Policy Name  Declarations (define/RULE)
default      default
ilmtest      gtktosilver
TEMPLATE-HSM
stub_size,is_empty,is_premigrated,is_migrated,access_age,mb_allocated,weight_expre
ssion,hsmexternalpool,systemtotape
TEMPLATE-ILM
stub_size,is_empty,is_premigrated,is_migrated,access_age,mb_allocated,exclude_list
,weight_expression,systemtosilver
EFSSG1000I The command completed successfully.
[st002.virtualad.ibm.com]$ lspolicy -P ilmtest
Policy Name Declaration Name Default Declarations
ilmtest     gtktosilver      N       RULE 'gtktosilver' SET POOL 'silver' WHERE
NAME LIKE '%gtk'
EFSSG1000I The command completed successfully.
[st002.virtualad.ibm.com]$ lspolicy -A
Cluster          Device Policy Set Name Policies Applied Time Who applied it?
st002.virtualad.ibm.com gpfs0  SYS             null     N/A
EFSSG1000I The command completed successfully.
```

The setpolicy command

A named policy that is stored in the SONAS database can be applied to a file system by running **setpolicy**, as shown in Example 5-81. Policies that are set with the **setpolicy** command become the active policy for a file system. The active policy controls the allocation and placement of new files in the file system. The **setpolicy -D** command can also be used to remove an active policy for a file system.

Example 5-81 Sample setpolicy command output

```
[st002.virtualad.ibm.com]$ setpolicy gpfs0 -P ilmtest
EFSSG1000I The command completed successfully.
[st002.virtual.com]$ lspolicy -A
Cluster           Device Policy Set Name Policies Applied Time     Who applied it?
st002.virtualad.ibm.com gpfs0  ilmtest         ilmtest  10/26/11 1:21 AM admin
EFSSG1000I The command completed successfully.
```

Changing or replacing an active policy

To replace an active policy's placement rules with the default GPFS placement rule, run **setpolicy -D**.

To change an existing active policy, you must run **setpolicy -D** as the first of three steps. This procedure is not recommended because it opens a window where the wanted policy configuration is not in place, and auto-migration or some other unintended consequence might occur. The second step is to run **chpolicy** to modify the policy, and then run **setpolicy** to set the modified policy as the active policy.

Running and stopping policies by using the runpolicy command

The **runpolicy** command runs a policy on a file system. Either the default policy, or the policy set on the file system that uses the **setpolicy** command, can be run by specifying the **-D** option. Another policy that is stored in the SONAS database can be run by specifying the **-P** option (see Example 5-82). The **runpolicy** command runs migration and deletion rules.

Example 5-82 Sample runpolicy command output

```
[st002.virtualad.ibm.com]$ runpolicy gpfs0 -P ilmtest
EFSSA0184I The policy is started on gpfs0 with JobID 4.
[st002.virtual.com]$ showlog 4
Primary node: mgmt002st002
Job ID : 4
[I] GPFS Current Data Pool Utilization in KB and %
silver  4608    1990656 0.231481%
system  693760  1757184 39.481352%
[I] 4156 of 275968 inodes used: 1.505972%.
[I] Loaded policy rules from /var/opt/IBM/sofs/PolicyFiles/policy1319586479577.
Evaluating MIGRATE/DELETE/EXCLUDE rules with CURRENT_TIMESTAMP =
2011-10-25@23:48:01 UTC
parsed 2 Placement Rules, 0 Restore Rules, 0 Migrate/Delete/Exclude Rules,
        0 List Rules, 0 External Pool/List Rules
RULE 'gtktosilver' SET POOL 'silver' WHERE NAME LIKE '%gtk'
RULE 'default' SET POOL 'system'
[I] Directories scan: 68 files, 36 directories, 0 other objects, 0 'skipped' files
and/or errors.
[I] Summary of Rule Applicability and File Choices:
 Rule#  Hit_Cnt KB_Hit  Chosen  KB_Chosen        KB_Ill  Rule
[I] Filesystem objects with no applicable rules: 85.
[I] GPFS Policy Decisions and File Choice Totals:
 Chose to migrate OKB: 0 of 0 candidates;
 Chose to premigrate OKB: 0 candidates;
 Already co-managed OKB: 0 candidates;
 Chose to delete OKB: 0 of 0 candidates;
 Chose to list OKB: 0 of 0 candidates;
 OKB of chosen data is illplaced or illreplicated;
Predicted Data Pool Utilization in KB and %:
silver  4608    1990656 0.231481%
system  695552  1757184 39.583333%
[I] A total of 0 files have been migrated, deleted or processed by an EXTERNAL
EXEC/script;
       0 'skipped' files and/or errors.
----------------------------------------------------------
End of log - runpolicy completed
----------------------------------------------------------
EFSSG1000I The command completed successfully.
```

The stoppolicy command

The **stoppolicy** command stops running policy jobs, depending on the parameters that are specified (see Example 5-83). Auto-migrations can be stopped only by providing the time parameter. Be careful! Stopped auto-migrations are restarted within two minutes if the condition that triggered this policy run (the auto-migration policy that is applied against the file system; check the status with **lspolicy -A** and then **lspolicy -P <policy_name>**) is still valid. Also, stopping auto-migrations can lead to out-of-space conditions that must always be avoided.

Example 5-83   Sample stoppolicy command output

```
[st002.virtual.com]$ stoppolicy gpfs0
EFSSG1000I The command completed successfully.
```

The mkpolicytask command

The **mkpolicytask** command creates a SONAS **cron** job, a scheduled operation, which applies the currently applied policy on a file system at a specified time. The **mkpolicytask** command takes the file system as an argument (see Example 5-84).

Example 5-84   Sample mkpolicytask command

```
[st002.virtualad.ibm.com]$ mkpolicytask gpfs0 --hour 22 -P ilmtest
EFSSG0019I The task StartRunPolicy has been successfully created.
EFSSG1000I The command completed successfully.
```

To remove scheduled policy tasks from a file system, run **rmpolicytask** with the file system as the argument.

It is important that the task schedule allows sufficient time for the policy to complete before it runs again, so that policy tasks do not overlap. For example, if the policy has a rule that specifies that at 80% usage (that is, migrate to 75%), you are migrating 5% of the file system. Based on the transfer rate between file system pools and the system load, you must schedule the policy task with a sufficiently large time interval between start times to complete each migration before the scheduled task submits another iteration of the migration policy.

6

Backup and recovery, availability, and resiliency functions

This chapter describes Scale Out Network Attached Storage (SONAS) components and external products that can be used to provide data availability and resiliency. It also provides information about the Tivoli Storage Manager integration.

This chapter describes the following topics:

- Backup and recovery of files in a SONAS cluster
- Configuring SONAS to use HSM
- Replication of SONAS data
- SONAS Snapshots
- Disaster recovery

6.1 High availability and data protection in base SONAS configurations

A SONAS cluster offers many high availability and data protection features that are part of the base configuration and do not need to be ordered separately. SONAS is a grid-like storage solution. By design, all the components in a SONAS cluster are redundant, so there is no single point of failure. For example, you can have multiple Interface nodes for client access, and data can be replicated cross multiple storage pods. The software components that are included in the SONAS cluster also offer high availability functions. For example, the SONAS General Parallel File System (GPFS) is accessed concurrently from multiple Interface nodes and offers data protection through synchronous replication and snapshots. For more information, see Chapter 3, "Installation and configuration for SONAS Gateway solutions" on page 97.

SONAS also includes Tivoli Storage Manager Client software for data protection and backup to an external Tivoli Storage Manager Server, and asynchronous replication functions to send data to a remote SONAS or file server.

Data is accessed through Interface nodes, and Interface nodes are deployed in netgroups of two or more to ensure data accessibility if an Interface node is no longer accessible. The SONAS software stack manages services availability and access failover between multiple Interface nodes. This configuration allows clients to continue accessing data if an Interface node is unavailable. The SONAS Cluster Manager is composed of four fundamental components for data access failover:

- The Cluster Trivial Database (CTDB) monitors services and restarts them on an available node, offering concurrent access from multiple nodes with locking for data integrity.
- DNS performs IP address resolution and round-robin IP load balancing.
- NTP keeps timing in sync between the clustered devices.
- The file sharing protocol includes error retry mechanisms.

These four components, together with a retry mechanism in the file sharing protocols, make SONAS a high availability file sharing solution.

This chapter introduces the SONAS high availability and data protection functions and describes how these features can be applied in your environment to protect your data.

6.1.1 Cluster Trivial Database

The CTDB is used for two major functions, as described here.

Overview

CTDB is used for two major functions. First, it provides a clustered manager that can scale well to large numbers of nodes. The second function that it offers is the control of the cluster. CTDB controls the public IP addresses that are used to publish the NAS services and moves them between nodes. Using monitoring scripts, CTDB determines the health state of a node. If a node has problems, such as broken services or network links, the node becomes unhealthy. In this case, CTDB migrates all public IP addresses to healthy nodes and sends CTDB *"tickle-acks"* to the clients so that they reestablish the connection. CTDB also provides the API to manage cluster IP addresses, add and remove nodes, and ban and disable nodes.

CTDB must be healthy on each node of the cluster for SONAS to work correctly. When services are down for any reason, the state of CTDB might go down. CTDB services can be restarted on a node by using either the SONAS GUI or the command-line interface (CLI). It is also possible to change CTDB configuration parameters, such as public addresses, log file information, and debug level.

Suspending and resuming nodes

You can use the SONAS administrator CLI to do multiple operations on a node.

The **suspendnode** and **resumenode** commands provide control of the status of an Interface node in the cluster. The **suspendnode** command suspends a specified Interface node. It does this by banning the node at the CTDB level. A banned node does not participate in the cluster and does not host any records for the CTDB. The IP addresses for a suspended node are taken over by another node and no services are hosted on the suspended node.

6.1.2 DNS performs IP address resolution and load balancing

DNS is easily configured in the GUI, as shown in Figure 6-1.

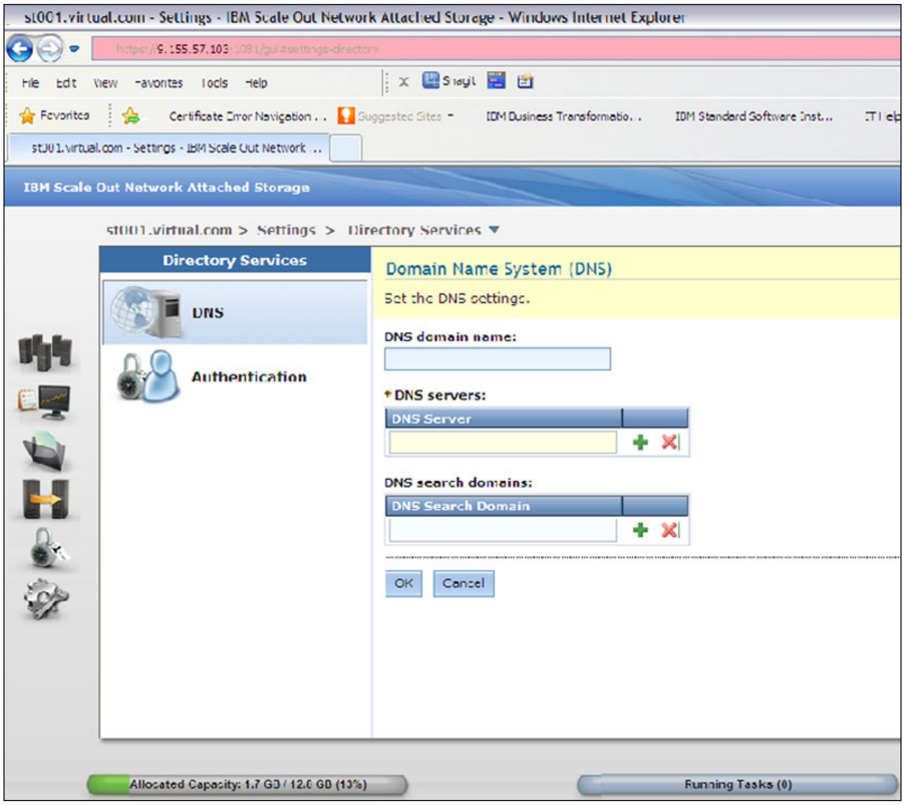

st001.virtual.com - Settings - IBM Scale Out Network Attached Storage - Windows Internet Explorer
https://9.155.57.103:1081/gui#settings-directory
File Edit View Favorites Tools Help
Favorites
Certificate Error Navigation...
Suggested Sites
IBM Business Transformatio...
IBM Standard Software Inst...
st001.virtual.com - Settings - IBM Scale Out Network ...
IBM Scale Out Network Attached Storage
st001.virtual.com > Settings > Directory Services
Directory Services
DNS
Authentication
Domain Name System (DNS)
Set the DNS settings.
DNS domain name:
* DNS servers:
DNS Server
DNS search domains:
DNS Search Domain
OK
Cancel
Allocated Capacity: 1.7 GB / 12.0 GB (13%)
Running Tasks (0)

Figure 6-1 GUI DNS configuration wizard

When a problem occurs on a SONAS Interface node or in the network that connects the client to the SONAS Interface node, the result depends on multiple factors, such as the file-sharing protocol that is in use and specific SONAS configuration parameters. The following paragraphs illustrate various failover considerations.

All requests from a client to a SONAS cluster for data access are serviced through the SONAS public IP addresses. These public IP addresses are similar to virtual addresses because, in general, the client can access the same service, at various moments in time, over various public IP addresses. SONAS Interface nodes can have multiple public IP addresses for load balancing and IP failover. For example, the **lsnwinterface -x** command displays all public addresses in the Interface nodes, as shown in Figure 6-2. This figure shows two Interface nodes, *int001st002* and *int002st002,* each with two public IP addresses that are assigned on interfaces *eth1* and *eth2*. The Management node is also shown but it does not host any public IP addresses.

As shown in Figure 6-2, in normal operating conditions, each Interface node has two public IP addresses.

```
[[SONAS]$ lsnwinterface -x
Node                      Interface MAC               Master/Slave Up/Down IP-Addresses
int001st002.virtual.com  eth0      02:1c:5b:00:01:01              UP
int001st002.virtual.com  eth1      02:1c:5b:00:01:02              UP      10.0.1.121
int001st002.virtual.com  eth2      02:1c:5b:00:01:03              UP      10.0.2.122
int002st002.virtual.com  eth0      02:1c:5b:00:02:01              UP
int002st002.virtual.com  eth1      02:1c:5b:00:02:02              UP      10.0.1.122
int002st002.virtual.com  eth2      02:1c:5b:00:02:03              UP      10.0.2.121
mgmt001st002.virtual.com eth0      02:1c:5b:00:00:01              UP
mgmt001st002.virtual.com eth1      02:1c:5b:00:00:02              UP
mgmt001st002.virtual.com eth2      02:1c:5b:00:00:03              UP
```

Figure 6-2 Public IP addresses before IP address failover

Figure 6-3 shows that, after a node failover, all public IP addresses were moved to Interface node int002st002, and node int001st002 is hosting no IP addresses.

```
[[SONAS]$ Node                     Interface MAC              Master/Slave Up/Down IP-Addresses
int001st002.virtual.com  eth0      02:1c:5b:00:01:01              UP
int001st002.virtual.com  eth1      02:1c:5b:00:01:02              UP
int001st002.virtual.com  eth2      02:1c:5b:00:01:03              UP
int002st002.virtual.com  eth0      02:1c:5b:00:02:01              UP
int002st002.virtual.com  eth1      02:1c:5b:00:02:02              UP      10.0.1.121,10.0.1.122
int002st002.virtual.com  eth2      02:1c:5b:00:02:03              UP      10.0.2.121,10.0.2.122
mgmt001st002.virtual.com eth0      02:1c:5b:00:00:01              UP
mgmt001st002.virtual.com eth1      02:1c:5b:00:00:02              UP
mgmt001st002.virtual.com eth2      02:1c:5b:00:00:03              UP
```

Figure 6-3 Public IP addresses after IP address failover

6.1.3 Network Time Protocol setup

There are two main tasks to set up the Network Time Protocol (NTP):

1. Configuring an NTP server on the active Management node

   It is important for log and application consistency that all nodes in the cluster maintain synchronized timing. For this reason, having a valid NTP server and alternative that is defined is important to system and service availability.

2. Configuring an NTP server on the active Management node

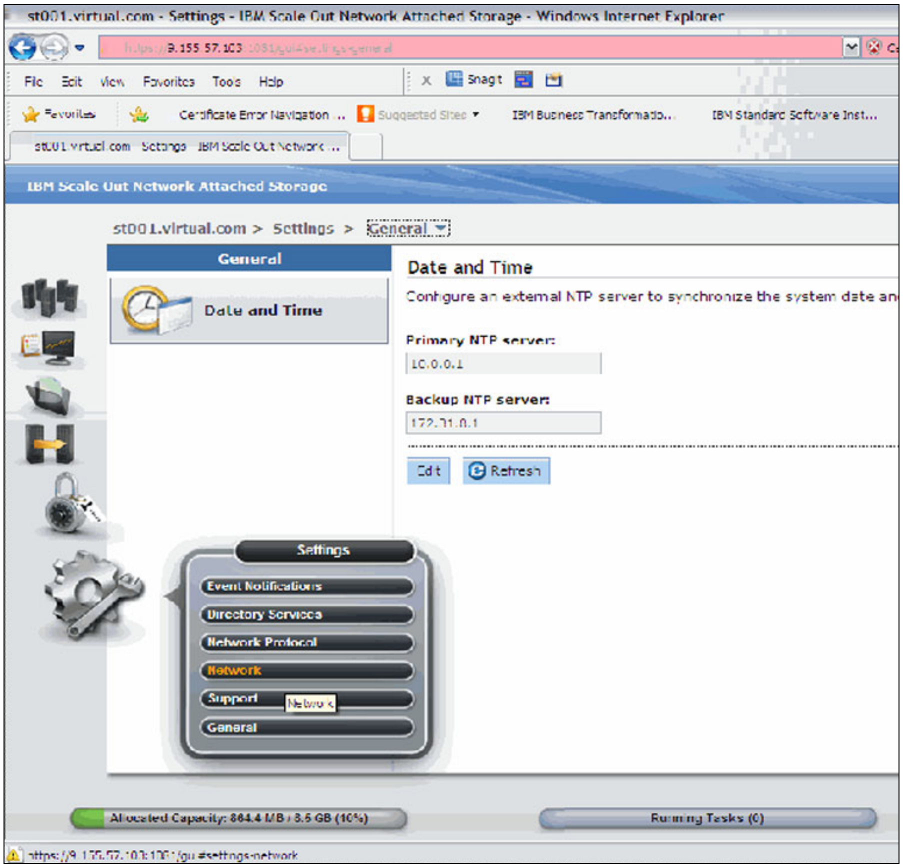
   Configure one or more external NTP servers on the active Management node for time synchronization.

To synchronize the system date and time on all of the nodes in the system, the active Management node must be configured to synchronize its time with an external NTP server. The active Management node is used by the other system members as their time source so that all of the nodes of the system are time-synchronized. To minimize the occurrence of Kerberos token errors that can result from client systems not being synchronized with the SONAS system, you can configure the same NTP server to which the client systems refer.

> **Note:** When you configure a Gateway solution, the back-end storage device should also point to the same NTP service for clock alignment in troubleshooting.

GUI navigation

To work with this function in the management GUI, log on to the GUI and click **Settings** → **Networks**, as shown in Figure 6-4.



Date and Time

Configure an external NTP server to synchronize the system date an

Primary NTP server: 10.0.0.1

Backup NTP server: 172.31.0.1

st001.virtual.com - Settings - IBM Scale Out Network Attached Storage - Windows Internet Explorer
File Edit View Favorites Tools Help
Favorites
Certificate Error Navigation ...
Suggested Sites
IBM Business Transformatio...
IBM Standard Software Inst...
IBM Scale Out Network Attached Storage
st001.virtual.com > Settings > General
General
Date and Time
Edit
Refresh
Settings
Event Notifications
Directory Services
Network Protocol
Network
Support
Network
General
Allocated Capacity: 864.4 MB / 3.6 GB (10%)
Running Tasks (0)

Figure 6-4 GUI NTP configuration wizard

CLI usage

To configure the active Management node to use one or multiple external NTP servers, run **`setnwntp`**.

setnwntp command example

In the following example, two NTP servers that use the IP addresses `10.0.0.10` and `10.0.0.11` are specified:

```
# setnwntp 10.0.0.10,10.0.0.11
```

To set one or more external NTP servers on the Management node, run **setnwntp**.

setnwntp command syntax

The following example shows the **setnwntp** command syntax:

```
setnwntp ip[,...,ip] [-c { clusterID | clusterName }]
```

In some cases, you might decide to move the Management service IP addressees to specified ports in your SONAS Management nodes. To accommodate that change, run **chnwmgt**.

The **chnwmgt** command changes Management node-related IP configurations. If two Management nodes are available (primary and secondary), both of them are updated.

Run this command with caution to prevent losing the connection.

For example, to move the Management IP traffic from the 1 GbE interfaces to the 10 GbE interfaces, you can use the following command structure:

```
chnwmgt --interface ethX1
```

This command moves the IP that is associated with the Management IP and the Service IP addresses from the default 1 GbE port (ethX0) to the first set of 10 GbE ports eth*X*1.

6.2 Backing up and restoring file data

The first thing to consider for the protection of your SONAS cluster is the nature of the configuration data protection, or protecting your ability to recover from a Management node failure.

This form of data protection is managed internally in the cluster, and not by an external replication or data protection service.

The active Management node can be backed up by running **backupmanagementnode**. The active Management node backup enables the replacement of the Management node if there is a failure of that node.

To back up the active Management node manually, run **backupmanagementnode**. This command does not create a scheduled backup task. For more information, including task parameters and default values, see the **backupmanagementnode** information in the SONAS IBM Knowledge Center (http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/manpages/backupmanagementnode.html?lang=en). The command creates an archive that includes all the components. To specify components to back up, you must use the optional **--component** parameter.

Only the active Management node is backed up by the **backupmanagementnode** command.

Tip: A subsequent restore operation restores the active Management node from the same system on which the Management node archive was created. If a Storage node or an Interface node is reinstalled after an archive is created, the archive is no longer valid for use by a restore operation.

To back up the active Management node to the `/persist/mgmtbackup` path on the strg001st001 Storage node, run the following command:

```
# backupmanagementnode --targethost strg001st001 --targetpath
/persist/mgmtbackup
```

To back up to a USB flash drive, run the following command:

```
# backupmanagementnode --mount /media/usb
```

Backing up to a USB flash drive requires an 8 GB USB flash drive.

Scheduling a backup of the active Management node

You can schedule backups of the active Management node by using either the CLI or the management GUI.

To schedule a backup of the active Management node from the CLI, run **`mktask`**.

The command in Example 6-1 makes the `BackupMgmtNode` task run at 17:00 every day, where the backup is on the strg001st001 Storage node in the `/backupdir/` directory.

Example 6-1 CLI command creating a daily task to back up the Management node configurations

```
[mgmt001st001 ~]#mktask BackupMgmtNode --hour 17 --minute 00  -p "strg001st001
/backupdir/"
```

There are several options available for data protection in the SONAS solution:

- Tivoli Storage Manager is one option that has extensive support and integration development. It is the only solution that allows for the mix of data protection and tape-based storage tiering (that uses the IBM HSM components of the Tivoli Storage Manager product). Tivoli Storage Manager provides for complete backup of the file system and full file system restore or individual file or directory restore.
- The SONAS solution differs slightly from most enterprise Tivoli Storage Manager solutions in that the SONAS cluster itself drives the backup and restore activity (not the Tivoli Storage Manager Server).
- NDMP backups are also supported through several of the primary NDMP backup solution vendors. NDMP backup is described in 6.11, “NDMP” on page 478. NDMP also offers a few benefits that are not supported by Tivoli Storage Manager data protection. One example is that NDMP backup offers support for file-set-level data protection (with full or incremental backup), and the Tivoli Storage Manager solution manages all backups at a filesystem level.

Tivoli Storage Manager data protection solution

This section describes the Tivoli Storage Manager data protection solution, its configuration, use, and common tasks.

Tivoli Storage Manager backup and restore processing backs up files to, and restores files from, an external Tivoli Storage Manager Server by using the embedded Tivoli Storage Manager backup and archive client. Tivoli Storage Manager backup and restore processing is controlled by the **`cfgbackupfs`**, **`cfgtsmnode`**, **`chbackupfs`**, **`lsbackupfs`**, **`lstsmnode`**, **`rmbackupfs`**, **`startbackup`**, **`showlog`**, **`showerrors`**, **`startrestore`**, **`stopbackup`**, and **`stoprestore`** commands.

IBM Tivoli Storage Manager, working together with IBM SONAS, provides an end-to-end comprehensive solution for backup and restore, archival, and HSM.

6.2.1 How IBM SONAS works with Tivoli Storage Manager

This section reviews and compares the Tivoli Storage Manager terminology and processes to explain how it works with IBM SONAS for the following tasks:

- Backing up and restoring files
- Archiving and retrieving them
- Migrating and recalling them (HSM)

Tivoli Storage Manager terminology

If you use Tivoli Storage Manager to *back up* files (which calls the Tivoli Storage Manager backup and archive client code on the Interface nodes), copies of the files are created on the Tivoli Storage Manager Server external storage, and the original files remain in your local file system. To obtain a backed up file from Tivoli Storage Manager storage, for example, if the file is deleted accidentally from the local file system, you *restore* the file.

If you use Tivoli Storage Manager to *archive* files to Tivoli Storage Manager storage, those files are removed from your local file system, and if needed later, you *retrieve* them from Tivoli Storage Manager storage.

If you use Tivoli Storage Manager to *migrate* SONAS files to external storage (which starts the Tivoli Storage Manager HSM client code on the Interface nodes), you move the files to external storage that is attached to the Tivoli Storage Manager Server, and Tivoli Storage Manager replaces the file with a *stub file* in the SONAS file system. You can accept the default stub file size, or if you want, specify the size of your Tivoli Storage Manager HSM stub files to accommodate needs or applications that want to read headers or read initial portions of the file. To users, the files appear to be online in the file system. If the migrated file is accessed, Tivoli Storage Manager HSM automatically initiates a *recall* of the full files from their migration location in external Tivoli Storage Manager attached storage. The effect on the user is simply an elongated response time while the file is being recalled and reloaded into internal SONAS storage. You can also initiate recalls proactively if you want.

Support for Tivoli Storage Manager

The SONAS system contains a Tivoli Storage Manager Client that can work with your Tivoli Storage Manager Server system to do high-speed data backup and recovery operations.

SONAS has special use for high-speed backup by the Tivoli Storage Manager Server product. The SONAS scan engine quickly identifies incremental changes in the file system, and then passes the list of changed, new, or deleted files directly to the Tivoli Storage Manager Server. This special use of Tivoli Storage Manager has the following effects:

- Avoids the need for Tivoli Storage Manager Server to walk the directory trees to identify changed files.
- Reduces the backup window to the time that is needed to copy changes to the external Tivoli Storage Manager managed storage.

Even though SONAS uses the normal Tivoli Storage Manager Client packages, the backup procedure is implemented in a different way than a plain Tivoli Storage Manager Client. SONAS is designed for scalability, so a normal file system traversal is much too slow. Instead, it uses the file system scan engine to improve the process of finding candidates for backup and dealing with file lists. It also includes more than one node in the backup process to reduce the time window that is needed to process the backup. Based on this implementation, the normal Tivoli Storage Manager Client GUI cannot be used and the Tivoli Storage Manager Server cannot initiate the scheduled backup jobs.

You can take backups for a file system, but not for a specific file, or a path. For example, there cannot be a selectable backup. However, there can be a restore on a file or a path level. The backup process in SONAS is not a disaster recovery solution because it takes much time to restore all files when you are dealing with a large (for example, petabyte) file system. You can use asynchronous replication to protect your environment against a disaster. Similar to backup, the SONAS HSM implementation is different from the standard HSM client that is used with GPFS.

SONAS supports LAN backup that uses the Tivoli Storage Manager Server. You have the choice of IBM premium Linear Tape-Open (LTO) Tape Libraries or any Tivoli Storage Manager Server supported tape or tape data deduplication device. Also, the Tivoli Storage Manager Server product provides full support for IBM tape-encryption technology products.

6.2.2 Methods to back up a SONAS cluster

SONAS is a storage device that stores your file data, so it is important to develop an appropriate file data protection and backup plan to be able to recover data in case of disaster, accidental deletion, or data corruption.

Overview

This section describes how to back up data that is contained in the SONAS cluster by using either Tivoli Storage Manager or another ISV backup product solutions (such as NDMP-based backups and data replication). It does not describe the backup of SONAS configuration information.

> **Tip:** Tivoli Storage Manager and NDMP backup are not supported concurrently in any SONAS cluster.

SONAS cluster configuration information is stored on the Management node in multiple repositories. SONAS offers the **`backupmanagementnode`** command to back up SONAS cluster configuration information. The **`backupmanagmentnode`** command can also be used to back up SONAS configuration data to other nodes in the cluster (to protect that data in the instance that the Management node is lost). The use of this command is described in 6.10.1, “Backing up SONAS configuration information” on page 473.

SONAS clusters are preinstalled with Tivoli Storage Manager to act as a Tivoli Storage Manager Client to back up file systems. The SONAS Tivoli Storage Manager Client requires an external, customer-supplied Tivoli Storage Manager Server and license.

Tivoli Storage Manager licenses

Licenses are based on the Interface nodes that pass data to the Tivoli Storage Manager Server. The minimum licenses that are required are for two Tivoli Storage Manager clients, based on the Interface processor count. The interface processor count minimum is a single processor with six cores in two Interface nodes.

6.2.3 Tivoli Storage Manager Client and server concepts and considerations

The Tivoli Storage Manager Client that is integrated into SONAS is at Version 7.1, and this client version is compatible with Tivoli Storage Manager servers at Versions 7.1, 6.3, and 6.2. For more Tivoli Storage Manager Client to server compatibility information, see the following website:

http://www.ibm.com/support/docview.wss?uid=swg21053218

The Tivoli Storage Manager Client runs on the SONAS Interface nodes and each Interface node can open up to eight sessions to the Tivoli Storage Manager Server. Multiple Interface nodes can initiate proportionally more sessions to the Tivoli Storage Manager Server.

For example, 10 Interface nodes can initiate up to 80 Tivoli Storage Manager sessions. In a typical configuration, you can set the Tivoli Storage Manager Server **maxsess** parameter to a value of 100 for SONAS. If the Tivoli Storage Manager Server cannot handle so many sessions, it might be necessary to reduce the number of Interface nodes that are involved in a backup because server sessions that hang or are disconnected might cause incomplete or failed backups.

Mount requests: As each node can start up to eight parallel sessions, the Tivoli Storage Manager Client **maxnummp** parameter must be set to eight. This setting means that a Tivoli Storage Manager Client node can initiate up to eight mount requests for Tivoli Storage Manager sequential media on the server.

If the backup destination is physical tape drives, the number of tape drives that are available for backup must be equal or greater than the maximum number of sessions.

SONAS LAN-based backup with Tivoli Storage Manager

SONAS supports LAN backup through the preinstalled Tivoli Storage Manager backup and archive client that runs on Interface nodes. Only LAN backup is supported. *LAN-free backup is not supported or implemented.* Tivoli Storage Manager uses the backup component. The archiving component is not used. All backup and restore operations are run by using the SONAS GUI or CLI commands. Native server-based Tivoli Storage Manager commands are not supported. The Tivoli Storage Manager Client is configured to try backup of open files and continue without backing up the file after a set number of retries.

The Tivoli Storage Manager backup path length is limited to 1024 characters, including both file and directory path length. File names must *not* use the following characters: " or ' or linefeed (0x0A). *Databases must be shut down or frozen before a backup occurs to put them into a consistent state*. Backup jobs are run serially: that is, only one backup job for one file system can run at one point in time.

Tivoli Storage Manager database sizing

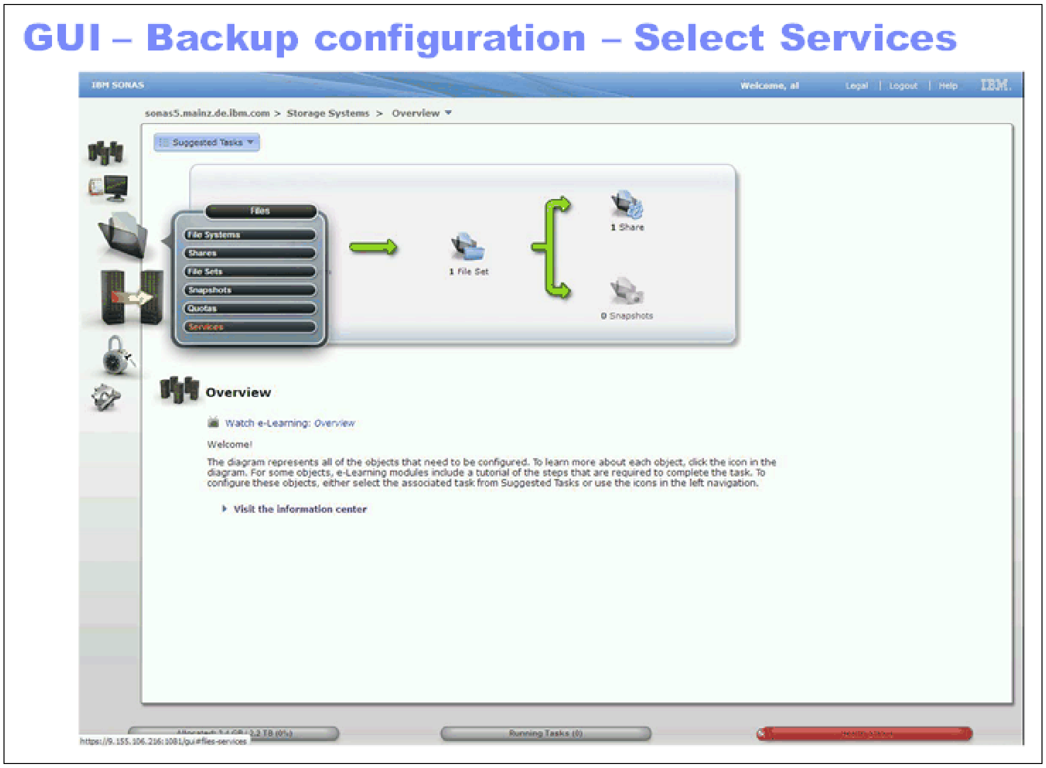
The Tivoli Storage Manager Server and the Tivoli Storage Manager Server database must be sized based on the number of files to be backed up. Each file that is backed up is an entry in the Tivoli Storage Manager database and each file entry in the Tivoli Storage Manager database uses 400 - 600 bytes or around 0.5 KB, so you can give a rough estimate of the size of the database by multiplying the number of files by the average file entry size. For example, a total of 200 million files use around 100 GB of Tivoli Storage Manager database space.

As of Tivoli Storage Manager V6.2, the maximum preferred size for one Tivoli Storage Manager database is 1000 GB. When large numbers of files must be backed up, you might need to deploy multiple Tivoli Storage Manager servers. The smallest SONAS that can be handled by a Tivoli Storage Manager Server is a file system, so this means that only one particular Tivoli Storage Manager Server can back up and restore files for a particular file system. When you have *n* file systems, you can have 1 - *n* Tivoli Storage Manager servers.

6.2.4 Configuring Interface nodes and file systems for Tivoli Storage Manager

This section describes the process for configuring Tivoli Storage Manager backup services from the GUI. Complete the following steps:

1. Select the **Services** icon from the GUI, as shown in Figure 6-5.



GUI – Backup configuration – Select Services

Overview

Watch e-Learning: Overview

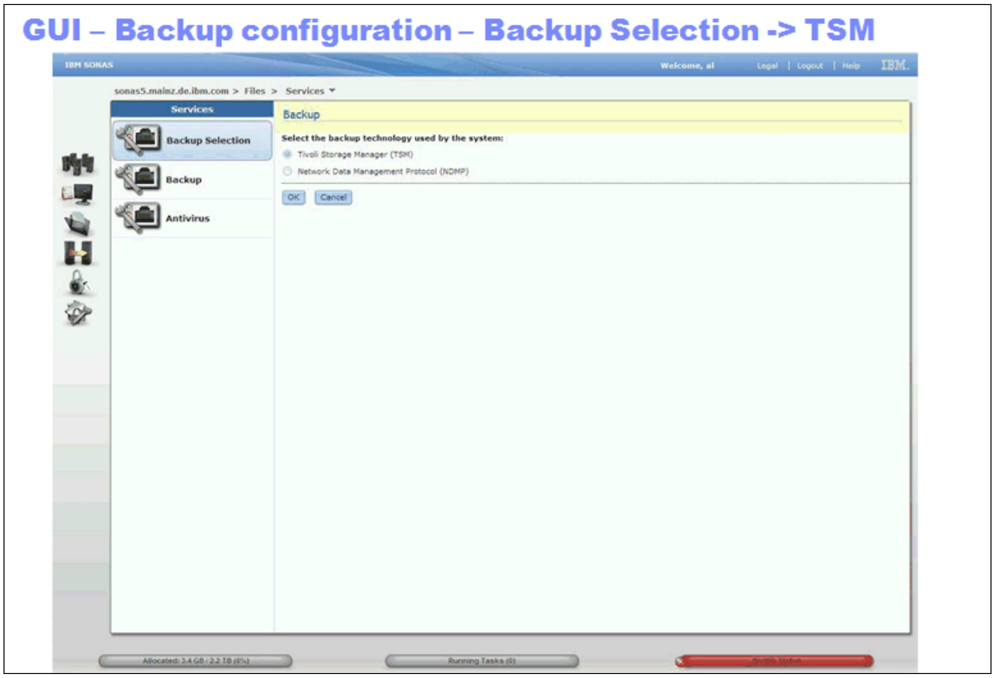
Welcome!

The diagram represents all of the objects that need to be configured. To learn more about each object, click the icon in the diagram. For some objects, e-Learning modules include a tutorial of the steps that are required to complete the task. To configure these objects, either select the associated task from Suggested Tasks or use the icons in the left navigation.

Visit the information center

Figure 6-5 Services window in SONAS V1.3 GUI

2. Select the **Tivoli Storage Manager** configuration, as shown in Figure 6-6.



GUI – Backup configuration – Backup Selection -> TSM

Backup

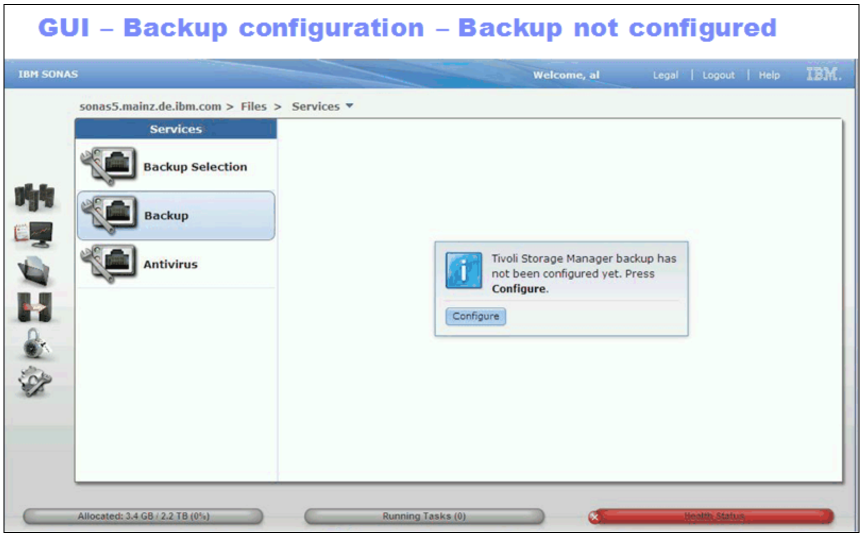
Select the backup technology used by the system:

● Tivoli Storage Manager (TSM)

○ Network Data Management Protocol (NDMP)

Figure 6-6 Tivoli Storage Manager - selection from GUI backup configuration window

3. Select **Backup** and click **Configure**, as shown in Figure 6-7.

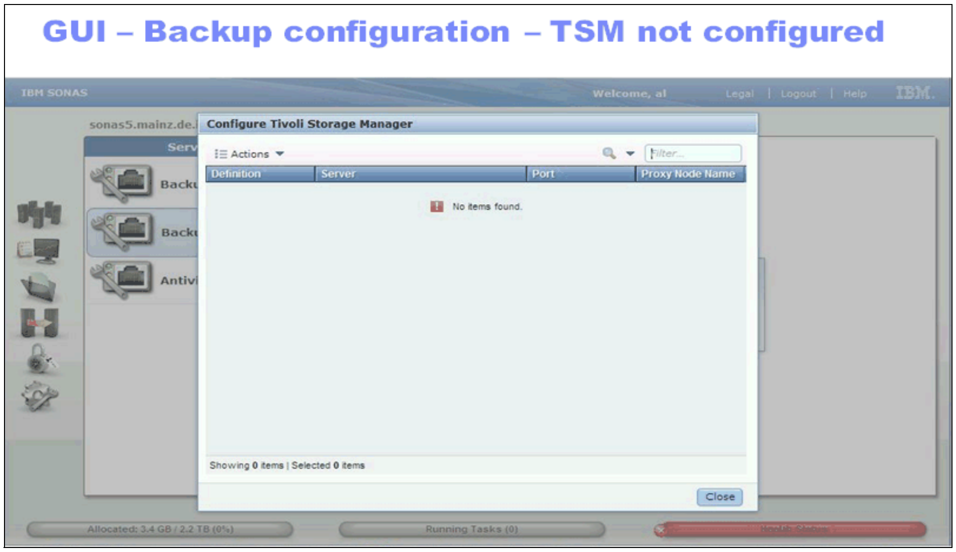


GUI – Backup configuration – Backup not configured
IBM SONAS
Welcome, al
Legal | Logout | Help
sonas5.mainz.de.ibm.com > Files > Services
Services
Backup Selection
Backup
Antivirus
Configure
Allocated: 3.4 GB / 2.2 TB (0%)
Running Tasks (0)
Health Status

Tivoli Storage Manager backup has not been configured yet. Press **Configure**.

Figure 6-7 First-time backup configuration option from the GUI

4. Select **Actions** from the Backup configuration window, as shown in Figure 6-8.

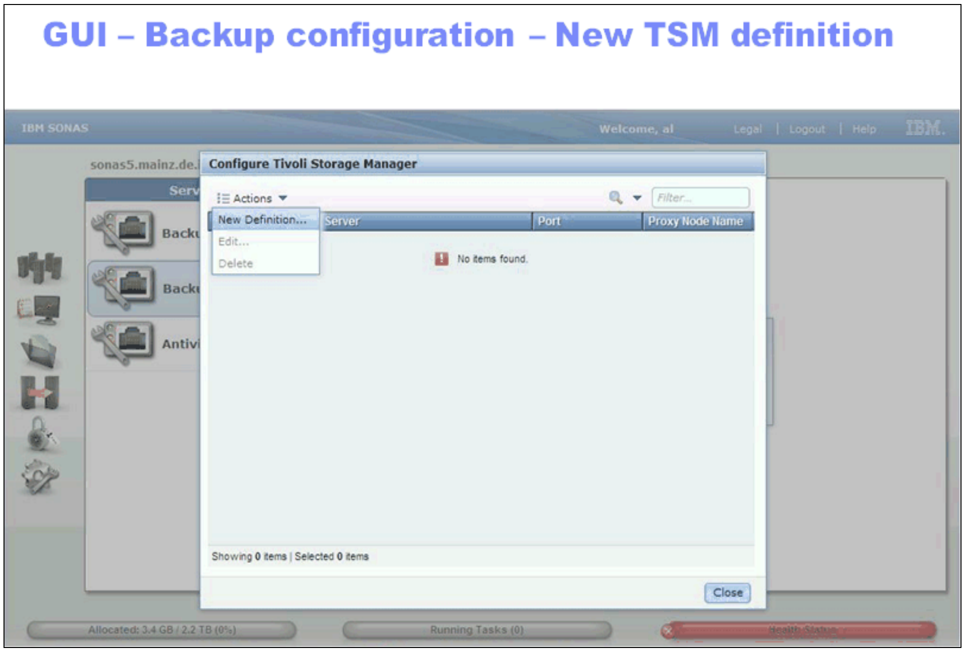


No items found.

Showing 0 items | Selected 0 items

Figure 6-8 First configuration shows an empty configuration list

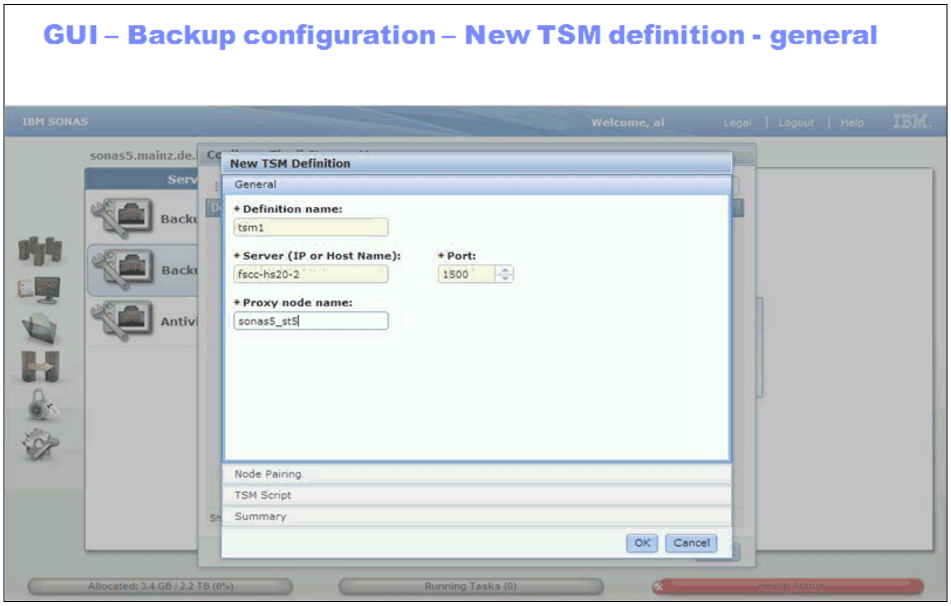
5. Select **New Definition** from the Actions window, as shown in Figure 6-9.



GUI – Backup configuration – New TSM definition

Figure 6-9 Create a Tivoli Storage Manager configuration definition

6. Enter the required information in the New Tivoli Storage Manager Definition window, as shown in Figure 6-10.



GUI – Backup configuration – New TSM definition - general

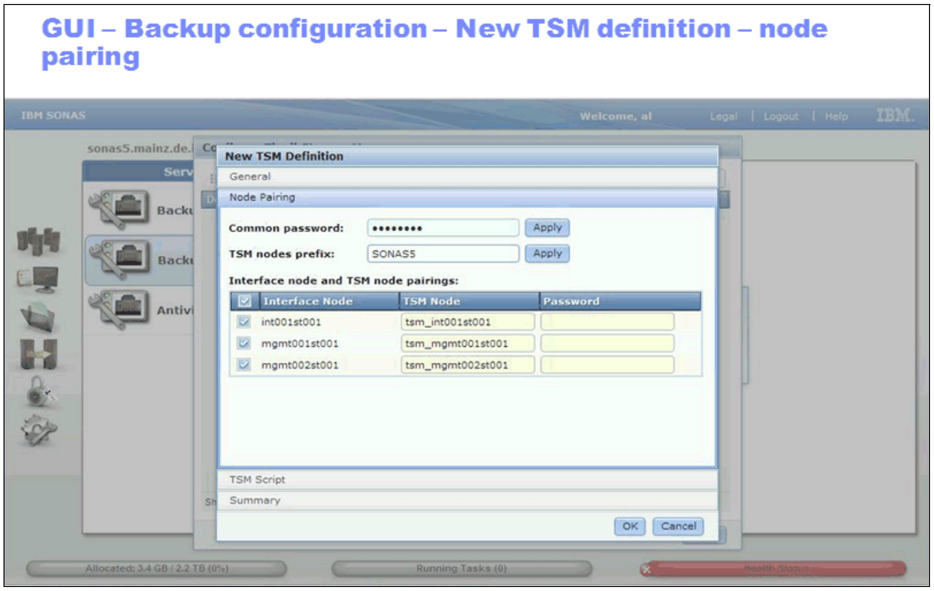
New TSM Definition

General

* Definition name: tsm1
* Server (IP or Host Name): fscc-hs20-2
* Port: 1500
* Proxy node name: sonas5_st5

Figure 6-10 Tivoli Storage Manager - definition boxes from the SONAS backup configuration GUI

7. Input the required information in to the Tivoli Storage Manager node pairing options fields, as shown in Figure 6-11.



GUI – Backup configuration – New TSM definition – node pairing

New TSM Definition

Common password: ••••••••• Apply

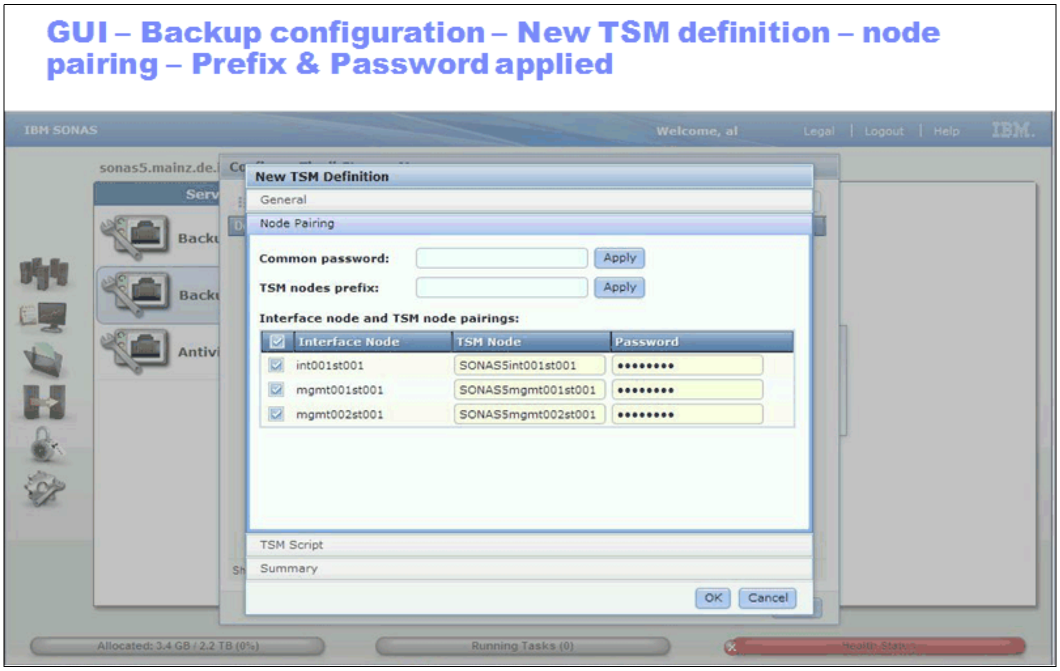
TSM nodes prefix: SONAS5 Apply

Interface node and TSM node pairings:

☑	Interface Node	TSM Node	Password
☑	int001st001	tsm_int001st001	
☑	mgmt001st001	tsm_mgmt001st001	
☑	mgmt002st001	tsm_mgmt002st001	

Figure 6-11 Tivoli Storage Manager - complete the definition node pairing option fields

8. Set the node pairing passwords, as shown in Figure 6-12.



GUI – Backup configuration – New TSM definition – node pairing – Prefix & Password applied

New TSM Definition

Common password: Apply

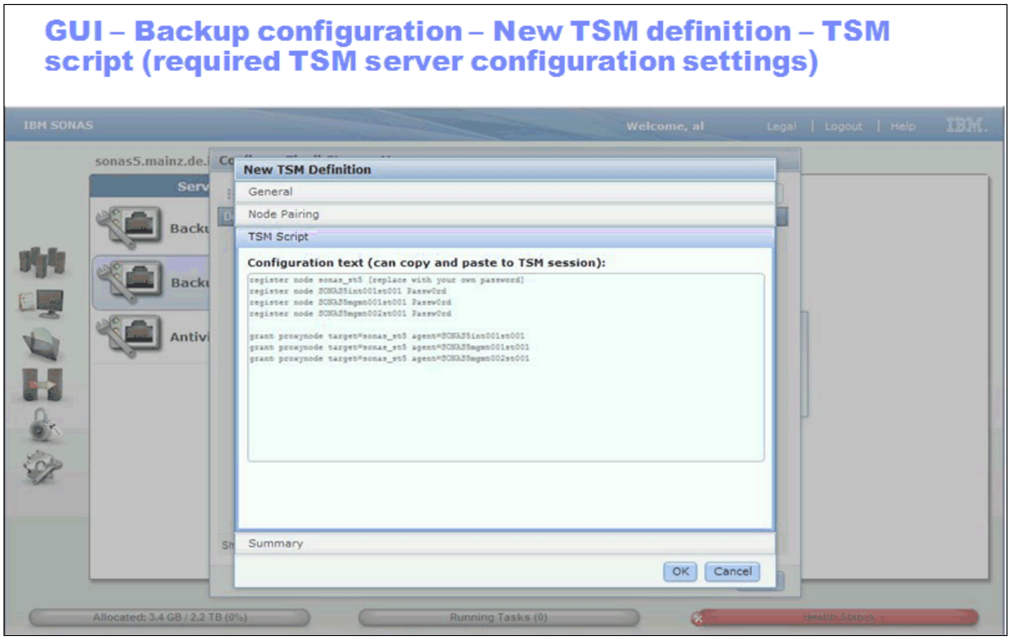
TSM nodes prefix: Apply

Interface node and TSM node pairings:

☑	Interface Node	TSM Node	Password
☑	int001st001	SONAS5int001st001	•••••••••
☑	mgmt001st001	SONAS5mgmt001st001	•••••••••
☑	mgmt002st001	SONAS5mgmt002st001	•••••••••

Figure 6-12 Set the node pair passwords

9. Enter the Tivoli Storage Manager Server scripts to run on the Tivoli Storage Manager Server, as shown in Figure 6-13.

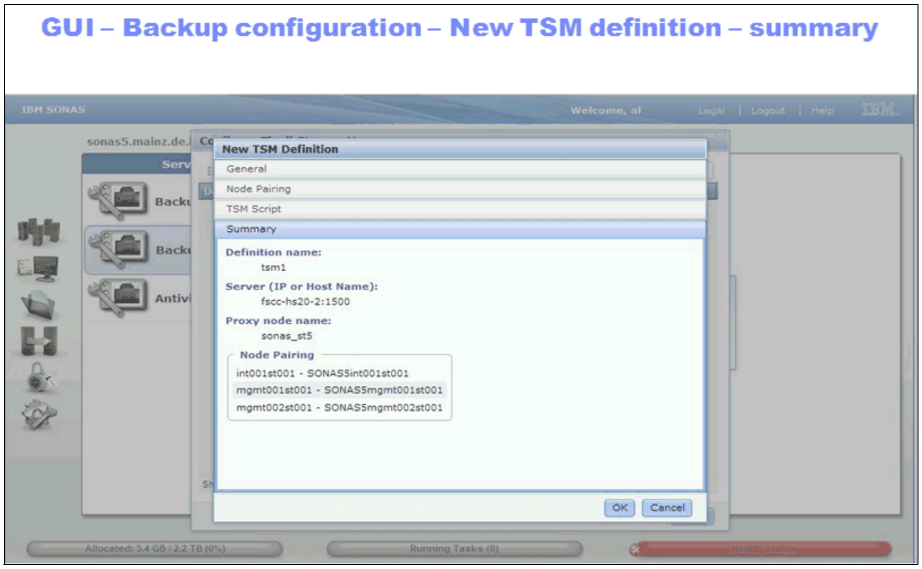


Configuration text (can copy and paste to TSM session):

```
register node sonas_st5 [replace with your own password]
register node SONAS5int001st001 Passw0rd
register node SONAS5mgmt001st001 Passw0rd
register node SONAS5mgmt002st001 Passw0rd

grant proxynode target=sonas_st5 agent=SONAS5int001st001
grant proxynode target=sonas_st5 agent=SONAS5mgmt001st001
grant proxynode target=sonas_st5 agent=SONAS5mgmt002st001
```

Figure 6-13 The Tivoli Storage Manager - server script window - display exact commands to run on the server

10. Review the summary of the Tivoli Storage Manager definition configuration for accuracy before running the commands on the Tivoli Storage Manager Server, as shown in Figure 6-14.



Definition name:
tsm1

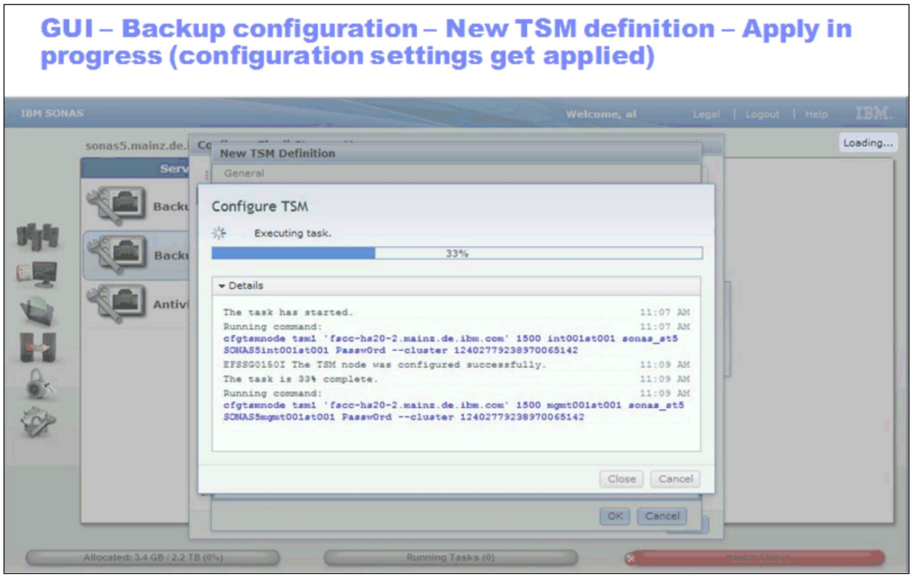
Server (IP or Host Name):
fscc-hs20-2:1500

Proxy node name:
sonas_st5

Node Pairing
int001st001 - SONAS5int001st001
mgmt001st001 - SONAS5mgmt001st001
mgmt002st001 - SONAS5mgmt002st001

Figure 6-14 The GUI Tivoli Storage Manager server definition summary

11. Apply the configuration, as shown in Figure 6-15. The Details window shows the commands that are being run in the background and the progress of the Tivoli Storage Manager configuration.



GUI – Backup configuration – New TSM definition – Apply in progress (configuration settings get applied)

Configure TSM

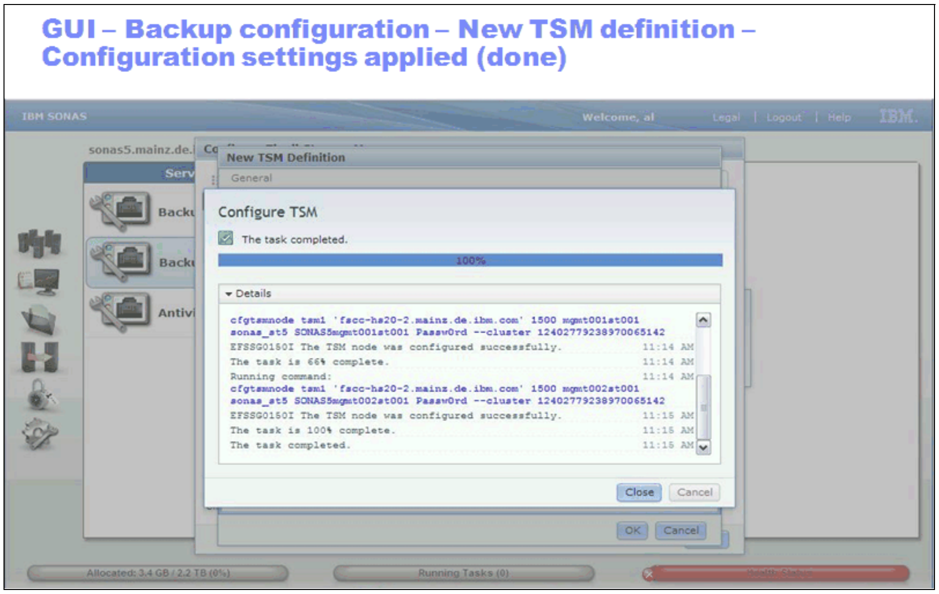
Executing task.

33%

```
The task has started.                                              11:07 AM
Running command:                                                   11:07 AM
cfgtsmnode tsm1 'fscc-hs20-2.mainz.de.ibm.com' 1500 int001st001 sonas_st5
SONAS5int001st001 Passw0rd --cluster 12402779238970065142
EFSSG0150I The TSM node was configured successfully.               11:09 AM
The task is 33% complete.                                          11:09 AM
Running command:                                                   11:09 AM
cfgtsmnode tsm1 'fscc-hs20-2.mainz.de.ibm.com' 1500 mgmt001st001 sonas_st5
SONAS5mgmt001st001 Passw0rd --cluster 12402779238970065142
```

Figure 6-15 Display of progress of the configuration (with CLI details)

12. When the task is complete, click **Close**, as shown in Figure 6-16.



GUI – Backup configuration – New TSM definition – Configuration settings applied (done)

Configure TSM

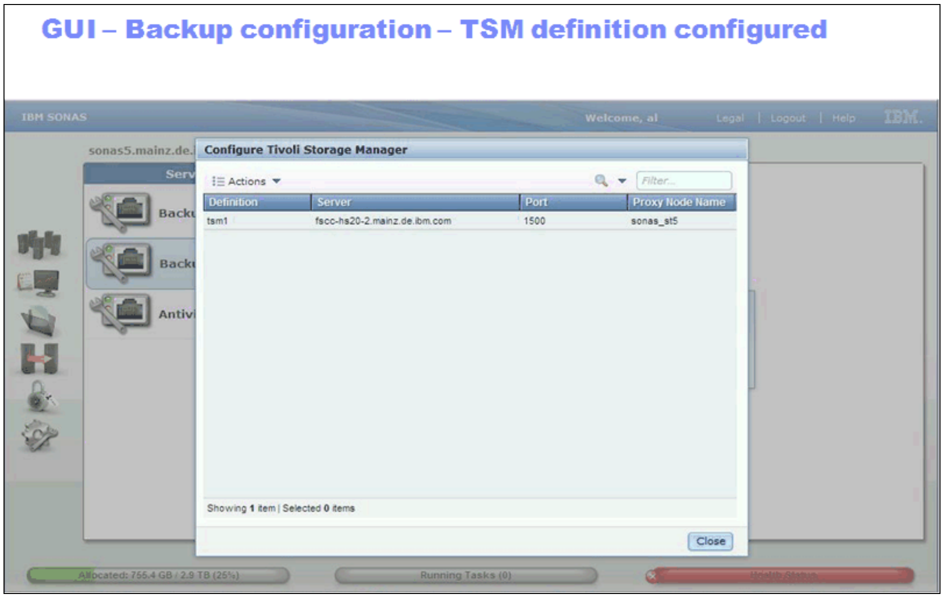
The task completed.

100%

```
cfgtsmnode tsm1 'fscc-hs20-2.mainz.de.ibm.com' 1500 mgmt001st001
sonas_st5 SONAS5mgmt001st001 Passw0rd --cluster 12402779238970065142
EFSSG0150I The TSM node was configured successfully.          11:14 AM
The task is 66% complete.                                     11:14 AM
Running command:                                              11:14 AM
cfgtsmnode tsm1 'fscc-hs20-2.mainz.de.ibm.com' 1500 mgmt002st001
sonas_st5 SONAS5mgmt002st001 Passw0rd --cluster 12402779238970065142
EFSSG0150I The TSM node was configured successfully.          11:15 AM
The task is 100% complete.                                    11:15 AM
The task completed.                                           11:15 AM
```

Figure 6-16 GUI Tivoli Storage Manager configuration successful

You can now see the newly defined configuration in the definition list, as shown in Figure 6-17.



GUI – Backup configuration – TSM definition configured

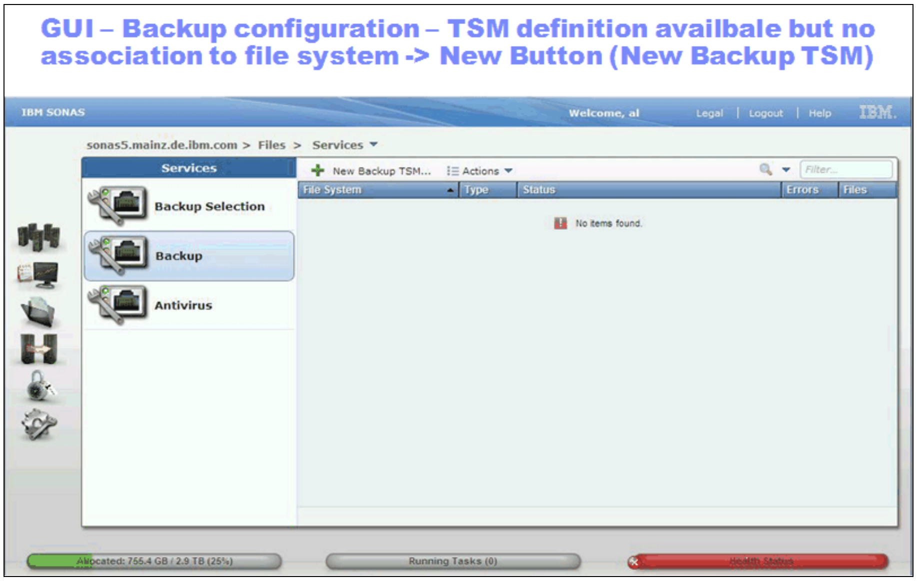
Configure Tivoli Storage Manager

Definition	Server	Port	Proxy Node Name
tsm1	fscc-hs20-2.mainz.de.ibm.com	1500	sonas_st5

Showing 1 item | Selected 0 items

Figure 6-17 The new configuration appears in the list of definitions

13. Begin associating a file system with the Tivoli Storage Manager service backup configuration, as shown in Figure 6-18.



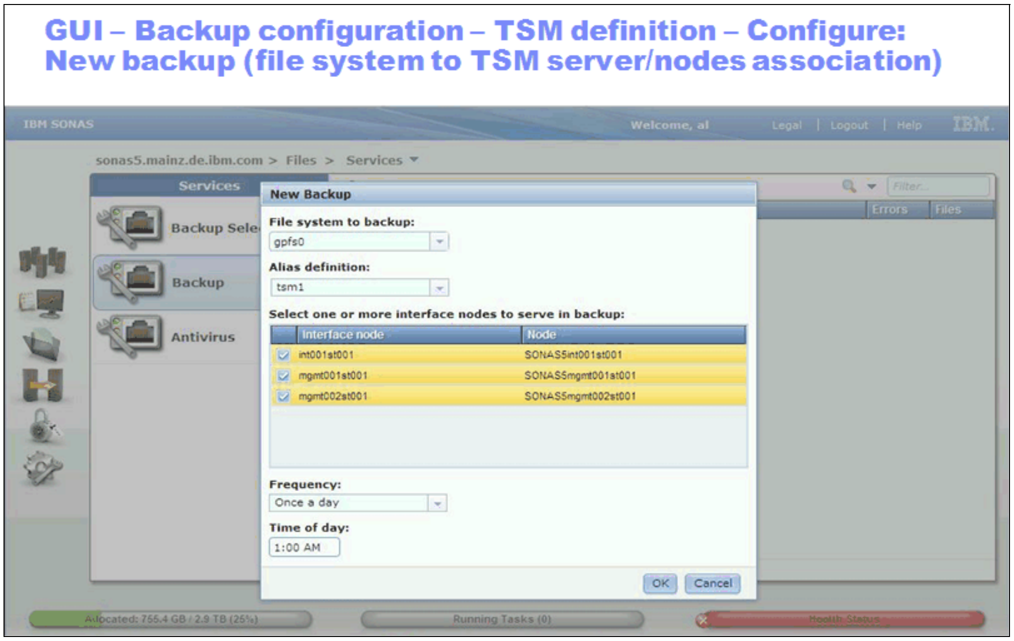
GUI – Backup configuration – TSM definition availbale but no association to file system -> New Button (New Backup TSM)

File System	Type	Status	Errors	Files

No items found.

Figure 6-18 Associate a file system with the Tivoli Storage Manager service backup

14. From the Actions menu, click **New Backup** and define the file system target, as shown in Figure 6-19.



GUI – Backup configuration – TSM definition – Configure:
New backup (file system to TSM server/nodes association)

New Backup

File system to backup: gpfs0

Alias definition: tsm1

Select one or more interface nodes to serve in backup:

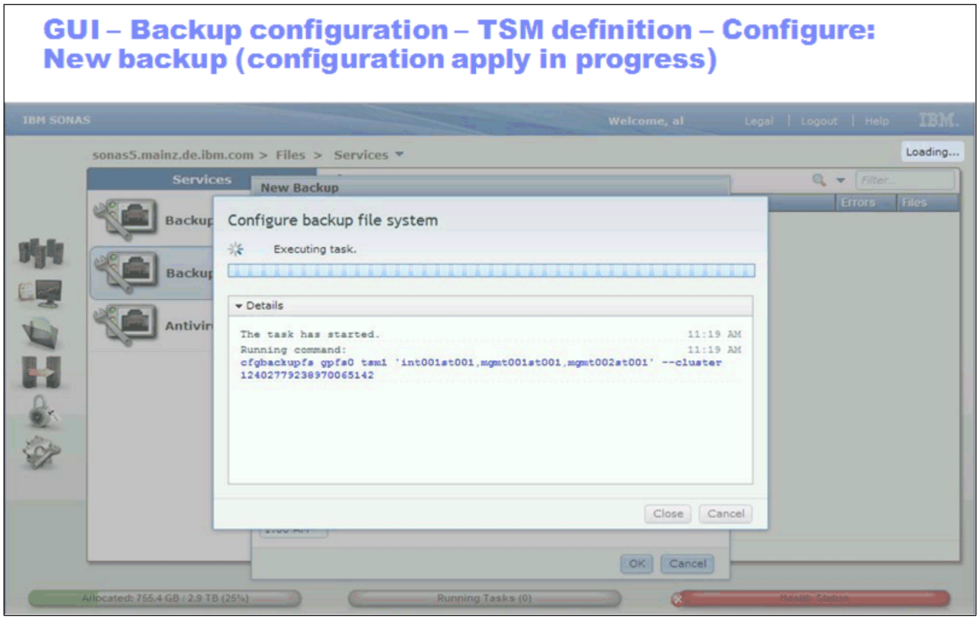
	Interface node	Node
☑	int001st001	SONAS5int001st001
☑	mgmt001st001	SONAS5mgmt001st001
☑	mgmt002st001	SONAS5mgmt002st001

Frequency: Once a day

Time of day: 1:00 AM

Figure 6-19 Tivoli Storage Manager - definition backup target configuration information

15. Click **OK** to start the configuration. The Details window shows the associated CLI commands that are being run and the status, as shown in Figure 6-20.



GUI – Backup configuration – TSM definition – Configure:
New backup (configuration apply in progress)

Configure backup file system

Executing task.

```
The task has started.                                                  11:19 AM
Running command:                                                       11:19 AM
cfgbackupfs gpfs0 tsm1 'int001st001,mgmt001st001,mgmt002st001' --cluster
12402779238970065142
```

Figure 6-20 Status with CLI details

16. After the configuration successfully completes, confirm that the file system is listed in the GUI backup window, as shown in Figure 6-21.

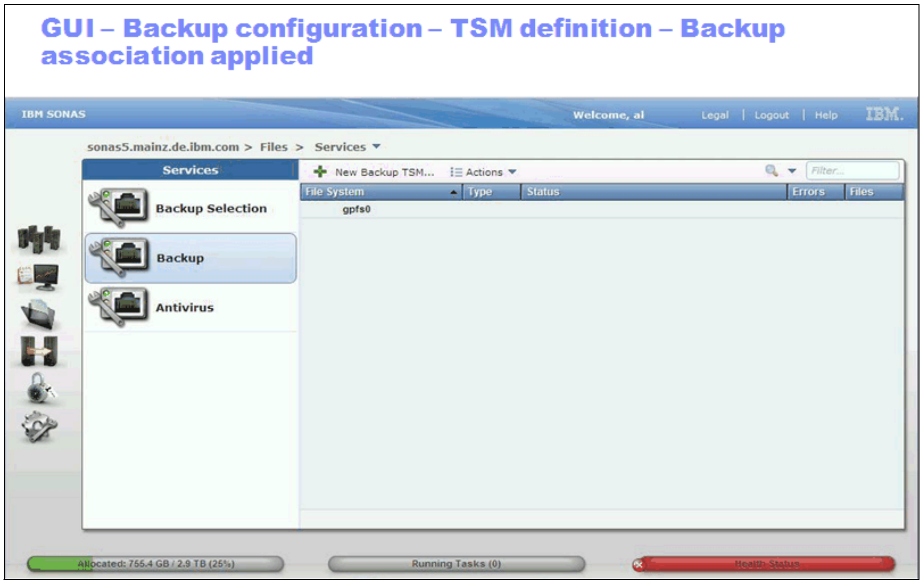


Figure 6-21 GUI backup window - file system list

6.2.5 Performing GUI-based Tivoli Storage Manager backups and restores

To do the file system backup, complete the following steps:

1. Start the file system backup by clicking **File System** → **Start backup**, as shown in Figure 6-22. The first backup is a full backup.

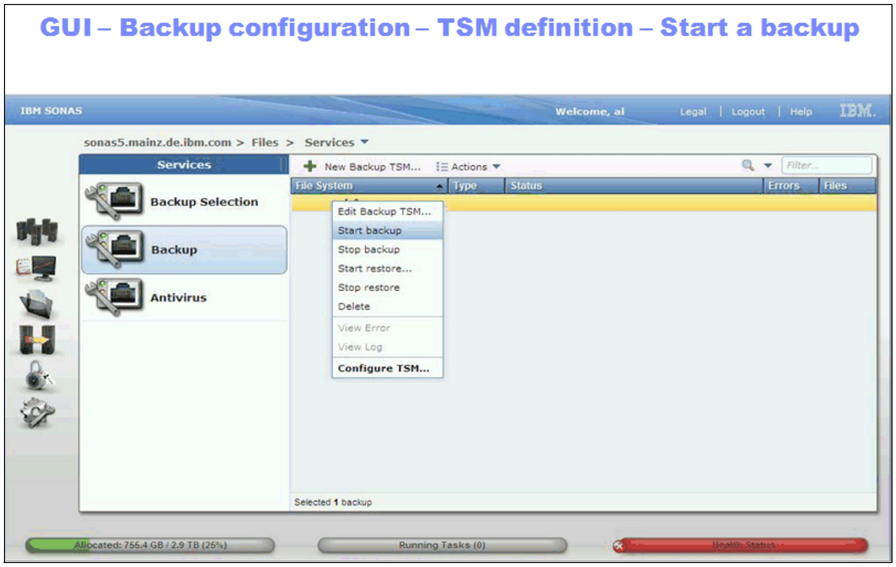
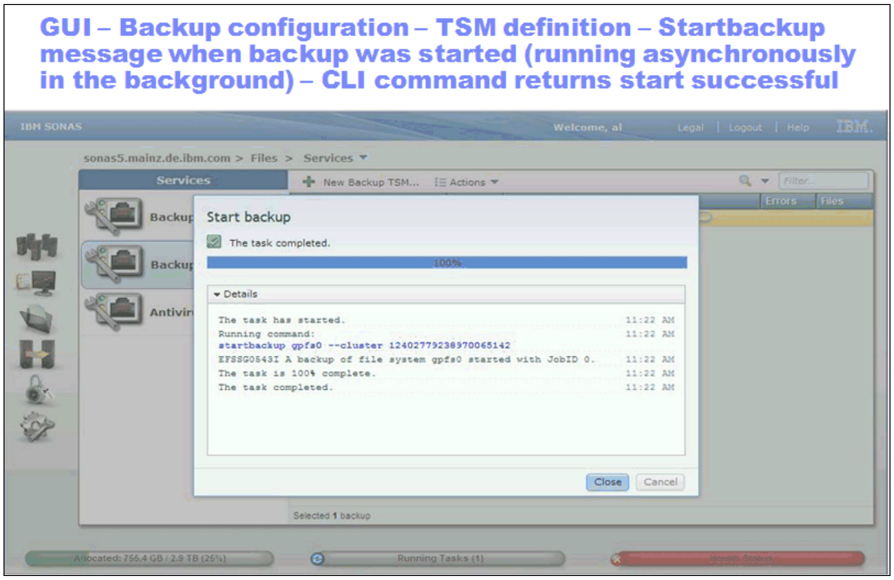


Figure 6-22 Start a file system backup

2. Confirm a successful start of the backup, as shown in Figure 6-23. Capture the output if you want. Click **Close**.



GUI – Backup configuration – TSM definition – Startbackup message when backup was started (running asynchronously in the background) – CLI command returns start successful

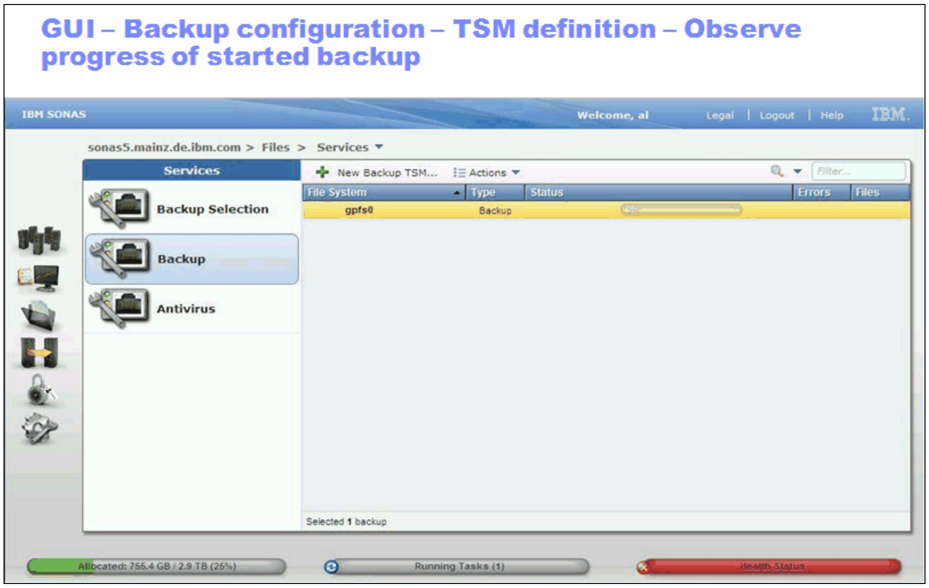
Start backup

☑ The task completed.

```
The task has started.                                    11:22 AM
Running command:                                         11:22 AM
startbackup gpfs0 --cluster 12402779238970065142
EFSSG0543I A backup of file system gpfs0 started with JobID 0.   11:22 AM
The task is 100% complete.                               11:22 AM
The task completed.                                      11:22 AM
```

Figure 6-23 Backup start confirmation status

You can watch the progress status of the backup as indicated by the progress bar under the Status field, as shown in Figure 6-24.

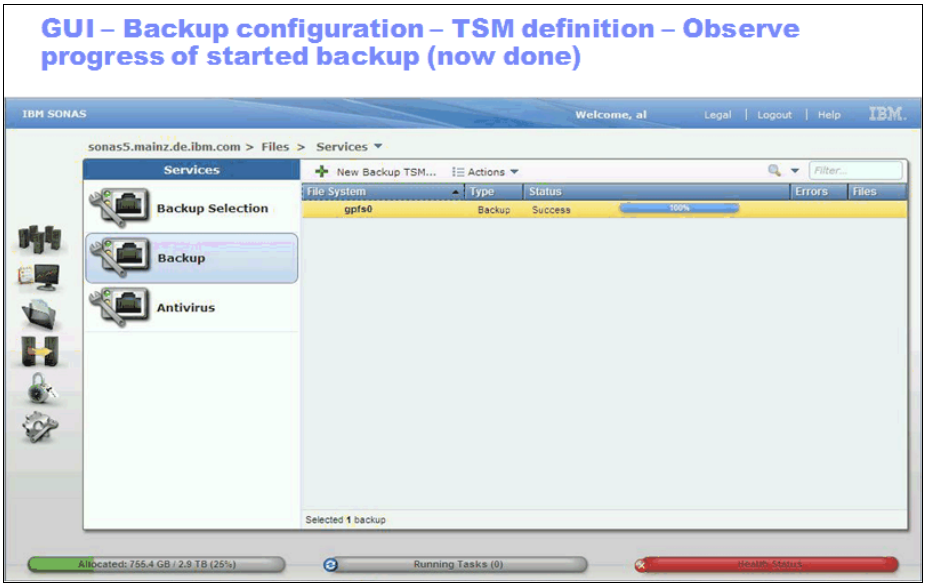


GUI – Backup configuration – TSM definition – Observe progress of started backup

File System	Type	Status	Errors	Files
gpfs0	Backup			

Figure 6-24 Backup progress status view at start

Figure 6-25 shows the progress at 100% and the status as Success.

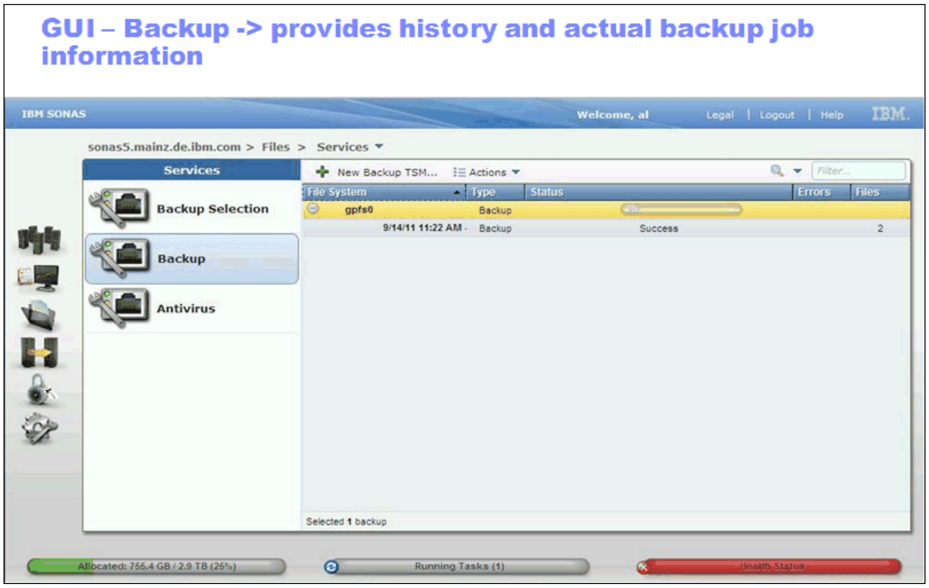

GUI – Backup configuration – TSM definition – Observe progress of started backup (now done)

File System	Type	Status		Errors	Files
gpfs0	Backup	Success	100%		

Selected 1 backup

Figure 6-25 Backup progress status view at 100%

3. Run a backup again to see the history of the backups in the Backup list, as shown in Figure 6-26.

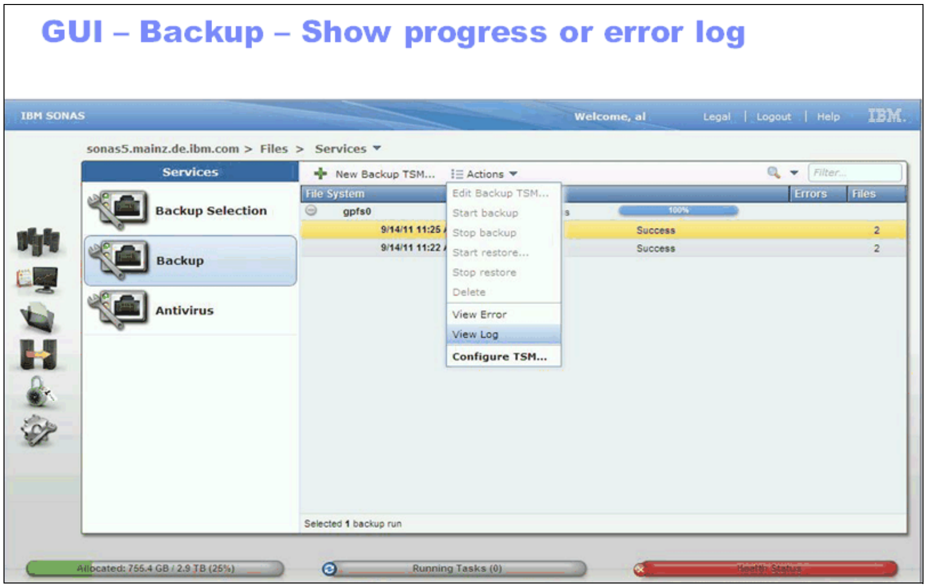

GUI – Backup -> provides history and actual backup job information

File System	Type	Status		Errors	Files
gpfs0	Backup		0%		
9/14/11 11:22 AM -	Backup		Success		2

Selected 1 backup

Figure 6-26 A new instance of backup still shows the history of previous backups

4. Click **Services** → **Actions** and select **View Log** (see Figure 6-27), which allows you to view backup logs for errors.



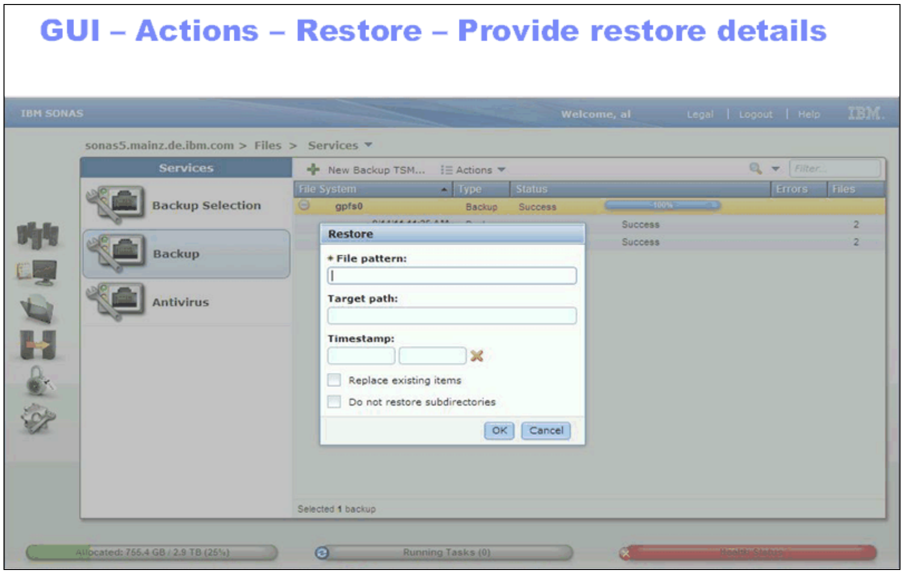
GUI – Backup – Show progress or error log

File System		Errors	Files
gpfs0			
9/14/11 11:25	Success		2
9/14/11 11:22	Success		2

Selected 1 backup run

Figure 6-27 The Actions menu provides options to view logs

Next, test a file restore from the Backup Actions menu (Figure 6-28).



GUI – Actions – Restore – Provide restore details

File System	Type	Status	Errors	Files
gpfs0	Backup	Success		
		Success		2
		Success		2

Restore

* File pattern:

Target path:

Timestamp:

□ Replace existing items

□ Do not restore subdirectories

OK Cancel

Selected 1 backup

Figure 6-28 Restore window from the Backup Actions menu

6.2.6 Tivoli Storage Manager CLI options for SONAS V1.3

Tivoli Storage Manager backup and restore processing is controlled by the **`cfgbackupfs`**, **`cfgtsmnode`**, **`chbackupfs`**, **`lsbackupfs`**, **`lstsmnode`**, **`rmbackupfs`**, **`startbackup`**, **`showlog`**, **`showerrors`**, **`startrestore`**, **`stopbackup`**, and **`stoprestore`** commands.

This section shows the commands, their description, and syntax.

The cfgbackupfs command

The **cfgbackupfs** command defines the Tivoli Storage Manager server on which a file system should be backed up, and the nodes that will back up the file system.

Here is the command syntax:

```
cfgbackupfs fileSystem tsmServerAlias nodes [-c { clusterID | clusterName }]
```

- **fileSystem**: Specifies the GPFS file system that is backed up.
- **tsmServerAlias**: Specifies the Tivoli Storage Manager server stanza on which the file system is backed up.
- **nodes**: Specifies the backup nodes that back up the file system.

Here are some examples of this command:

- **cfgbackupfs gpfs0 tsm001st001 mgmt001st001**
- **cfgbackupfs gpfs0 tsm001st001 mgmt001st001,mgmt002st001**

The cfgtsmnode command

The **cfgtsmnode** command configures the Tivoli Storage Manager node by defining the node name and node password, and by adding a Tivoli Storage Manager server stanza. This configuration must be done for each Tivoli Storage Manager server and node combination.

Here is the command syntax:

```
cfgtsmnode tsmServerAlias tsmServerAddress tsmServerPort nodeName virtualNodeName
clientNode clientNodePassword [--adminport port] [-c { clusterID | clusterName }]
```

- **tsmServerAlias**: Specifies the stanza of the Tivoli Storage Manager server that is registered with the Management node.
- **tsmServerAddress**: Specifies the address or the IP address of the Tivoli Storage Manager server that is registered with the Management node.
- **tsmServerPort**: Specifies the port of the Tivoli Storage Manager server that is registered with the Management node.
- **nodeName**: Specifies the GPFS node or host name.
- **virtualNodeName**: Specifies the virtual node name that is used as a common node name by all Management nodes, which is used to store the data. The **virtualNodeName** must be registered on the Tivoli Storage Manager server.
- **clientNode**: Specifies the Tivoli Storage Manager node (the customer can use any name for this node) where this command is to be run. The **clientNode** must be registered on the Tivoli Storage Manager server.
- **clientNodePassword**: Specifies the password to be used when the client node is registered on the Tivoli Storage Manager server. The password is Tivoli Storage Manager client-node specific; that is, it is not the virtual node name password.

Using unlisted arguments can lead to an error.

Here are some examples of this command:

- **cfgtsmnode tsmserver mytsmserver.com 1500 mgmt001st001 virtnode managementnode1 Password1**
- **cfgtsmnode tsmserver mytsmserver.com 1500 mgmt002st001 virtnode managementnode2 Password2**

The chbackupfs command

The **chbackupfs** command can change the Tivoli Storage Manager backup node list to a configured file system with a Tivoli Storage Manager association.

Here is the command syntax:

```
chbackupfs fileSystem { --add nodeList | --remove nodeList } [-c { clusterID |
clusterName }]
```

The **fileSystem** argument specifies the GPFS file system that is configured for backup.

Using unlisted arguments can lead to an error.

Here are some examples of this command:

- **chbackupfs gpfs0 --add int001st001**
- **chbackupfs gpfs0 --add int001st001 --remove int002st001,int003st001**

The lsbackupfs command

The **lsbackupfs** command lists the file system to Tivoli Storage Manager server and backup node associations.

Here is the command syntax:

```
lsbackupfs [-c { clusterID | clusterName }] [-Y]
```

Using unlisted arguments can lead to an error.

Here is an example of this command:

```
lsbackupfs
```

The lstsmnode command

The **lstsmnode** command lists all the defined and reachable Tivoli Storage Manager nodes in the cluster. Unreachable, but configured nodes are not displayed.

Here is the command syntax:

```
lstsmnode [nodeName] [-c { clusterID | clusterName }] [-Y] [--validate]
```

The **nodeName** argument specifies the node where the Tivoli Storage Manager server stanza information displays. If this argument is omitted, all the Tivoli Storage Manager server stanza information, for reachable client nodes within the current cluster, is displayed.

Using unlisted arguments can lead to an error.

Here is an example of this command with some sample outputs:

```
lstsmnode
```

Here is an example output for a SONAS system:

```
Node name   Virtual node name TSM server name  TSM server address TSM node name
int001st001 sonas_st1         TSMserver        9.155.106.19       int001st001
int002st001 sonas_st1         TSMserver        9.155.106.19       int002st001
int003st001 sonas_st1         TSMserver        9.155.106.19       int003st001
```

Here is an example output for an IBM Storwize V7000 Unified system:

```
Node name    Virtual node name TSM server name  TSM server address TSM node name
mgmt001st001 ifs_st1           TSMserver        9.155.106.19       mgmt001st001
mgmt002st001 ifs_st1           TSMserver        9.155.106.19       mgmt002st001
```

The rmbackupfs command

The **rmbackupfs** command removes a file system to Tivoli Storage Manager server association.

Here is the command syntax:

```
rmbackupfs fileSystem
```

The fileSystem argument specifies the GPFS file system for which the association is deleted. File system names do not need to be fully qualified. For example, `fs0` is as acceptable as `/dev/fs0`. However, file system names must be unique within a GPFS cluster. Do not specify an existing entry in `/dev`.

Using unlisted arguments can lead to an error.

Here is an example of this command:

```
rmbackupfs gpfs0
```

The startbackup command

The startbackup command starts the backup process.

Here is the command syntax:

```
startbackup [fileSystems] [--sync | --synconly ]
```

The **fileSystems** argument identifies the name of the file system to be backed up. If a file system is not provided, all registered file systems are backed up. The device name of the file system must be provided to this command without the `/dev` prefix. For example, `gpfs0` is accepted, but `/dev/gpfs0` is not. The available file system device names are retrievable by the **lsbackupfs** command.

Using unlisted arguments can lead to an error.

Here is an example of this command:

```
startbackup gpfs0
```

The showlog command

The **showlog** command shows the latest log file (if any) for a specific job.

Here is the command syntax:

```
showlog {jobID | JobCategory:fileSystem} [-c { clusterID | clusterName }] [--count
numberOfLines] [-t time]
```

- **jobID**: Specifies the unique ID of the job to display the log. Run **lsjobstatus** to discover the corresponding jobID value.
- **JobCategory**: Specifies the job. Possible values are `(abbreviations allowed): bac(kup)`, `res(tore)`, `rec(oncile)`, `chk(policy)`, `run(policy)`, and `aut(opolicy)`.
- **fileSystem**: Specifies the device name of the file system.

Using unlisted arguments can lead to an error.

Here are some examples of this command:

- **showlog 15**: Shows the log for a job with jobID 15.
- **showlog backup:gpfs0**: Shows the backup log for the latest backup job that is done for file system gpfs0.
- **showlog 15 -count 20**: Shows only the last 20 lines of the log for a job with jobID 15.
- **showlog backup:gpfs0 -t 03.05.2011 14:18:21.184**: Shows the backup log taken of file system gpfs0 at the date and time that is specified.

The showerrors command

The **showerrors** command shows the recent restore errors of a file system. It can also show all restore errors (**--all**) or the restore errors of a specified time (**-t**).

Here is the command syntax:

```
showerrors {jobID | JobCategory:fileSystem} [-c { clusterID | clusterName }]
[--count numberOfLines] [-t time]
```

- **jobID**: Specifies the unique ID of the job to display the log. Run **lsjobstatus** to discover the corresponding jobID value.
- **JobCategory**: Specifies the job. Possible values are (abbreviations allowed): bac(kup), res(tore), rec(oncile), chk(policy), run(policy), and aut(opolicy).
- **fileSystem**: Specifies the device name of the file system.

Using unlisted arguments can lead to an error.

Here are some examples of this command:

- **showerrors 15**: Shows the error log for a job with jobID 15.
- **showerrors backup:gpfs0**: Shows the backup error log for the latest backup job that is done for file system gpfs0.
- **showerrors 15 -count 20**: Shows only the last 20 lines of the error log for a job with jobID 15.
- **showerrors backup:gpfs0 -t 03.05.2011 14:18:21.184**: Shows the backup error log that is taken of file system **gpfs0** at the date and time that is specified.

The startrestore command

The **startrestore** command restores the file system on the specified current file pattern.

Here is the command syntax:

```
startrestore filePattern [-v] [-R | -T targetPath] [-t timestamp] [--nosubdirs]
[--nqr]
```

The **filePattern** argument specifies the file pattern where the file system is mounted and will be restored. You cannot restore multiple file systems at the same time. So, for example, if you have two file systems, one on /ibm/gpfs0, and one on /ibm/gpfs1, when you enter the following statement, there is no result:

```
startrestore "/ibm/*"
```

Here are some **filePattern** examples:

- To restore the whole file system, run the following command:

  ```
  startrestore /ibm/gpfs0/
  ```

- To restore the abc file, run the following command:

  ```
  startrestore /ibm/gpfs0/abc
  ```

- To restore all the files that are in dir1 of the **/ibm/gpfs0** file system, run the following command:

  ```
  startrestore /ibm/gpfs0/dir1/
  ```

- To restore a directory without restoring any of its files, run the following command:

  ```
  startrestore /ibm/gpfs0/dir1
  ```

Using unlisted arguments can lead to an error.

Here is an example of this command:

```
startrestore /gpfs0 -R
```

The stopbackup command

The **stopbackup** command stops the running Tivoli Storage Manager backup session on the cluster.

Here is the command syntax:

```
stopbackup { --all | -d filesystem | -j <jobId>} [-f] [-c { clusterID |
clusterName }]
```

Here is an example of this command:

```
stopbackup -c yolanda.bud.hu.ibm.com
```

The example stops the running of the Tivoli Storage Manager backup session on the yolanda cluster.

The stoprestore command

The **stoprestore** command stops running Tivoli Storage Manager restore sessions on the cluster.

Here is the command syntax:

```
stoprestore {-d filesystem | --all | -j jobID} [-f] [-c { clusterID | clusterName
}]
```

The **-f** option forces the restore status to “failed” if stopping the restore failed because the primary node is down and cannot be restarted.

Here is an example of this command:

```
stoprestore --all -c yolanda.bud.hu.ibm.com
```

The example stops all the running Tivoli Storage Manager restore sessions on the yolanda cluster.

The querybackup command

The **querybackup** command queries the backup summary for the specified file pattern.

Here is the syntax of the command:

```
querybackup filePattern [-i] [-d] [-q] [--nosubdirs] [--filesonly | --dirsonly]
[--fromdate [--fromtime]] [--todate [--totime]]
```

The **`filePattern`** argument specifies the file pattern where the file system is mounted.

Here are some **`filePattern`** examples:

- `/ibm/gpfs0/`: Queries the whole file system.
- `/ibm/gpfs0/abc`: Queries the `abc` file.
- `/ibm/gpfs0/abc?e*`: Queries all the files starting with "abc", plus one character, plus "e", plus any following characters.
- `/ibm/gpfs0/dir1/`: Queries all the files in dir1 of the `/ibm/gpfs0` file system.
- `/ibm/gpfs0/dir1/*`: Queries all the files that are in `dir1` of the `/ibm/gpfs0` file system.
- `/ibm/gpfs0/dir1`: Queries the directory without querying any of its files.

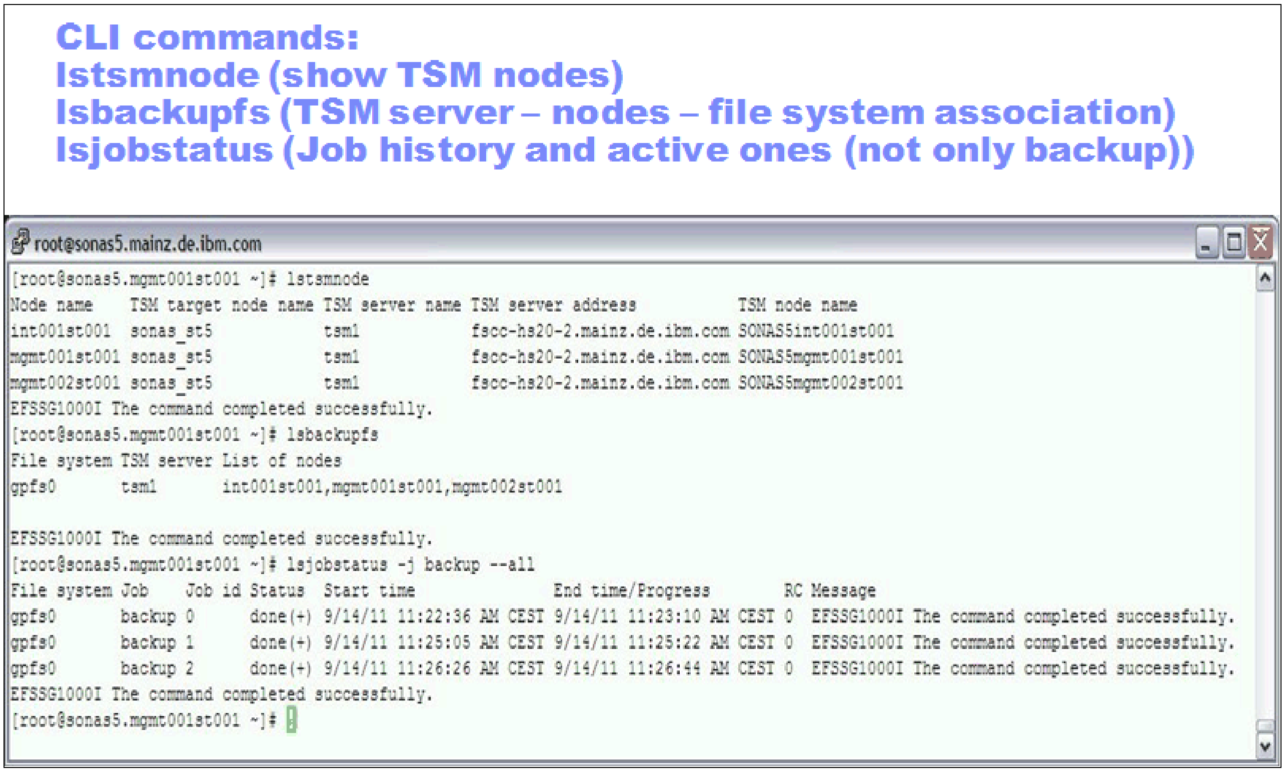
Using unlisted arguments can lead to an error.

Here are some examples of this command:

- **`querybackup /ibm/gpfs0/`**
- **`querybackup /ibm/gpfs0/ -i -d -q`**
- **`querybackup /ibm/gpfs0/ --fromdate 2011-01-01 --todate 2011-03-31 --filesonly`**

Additional backup commands

Figure 6-29 shows information about the **`lstsmnode`**, **`lsbackupfs`**, and **`lsjobstatus`** backup tool commands.

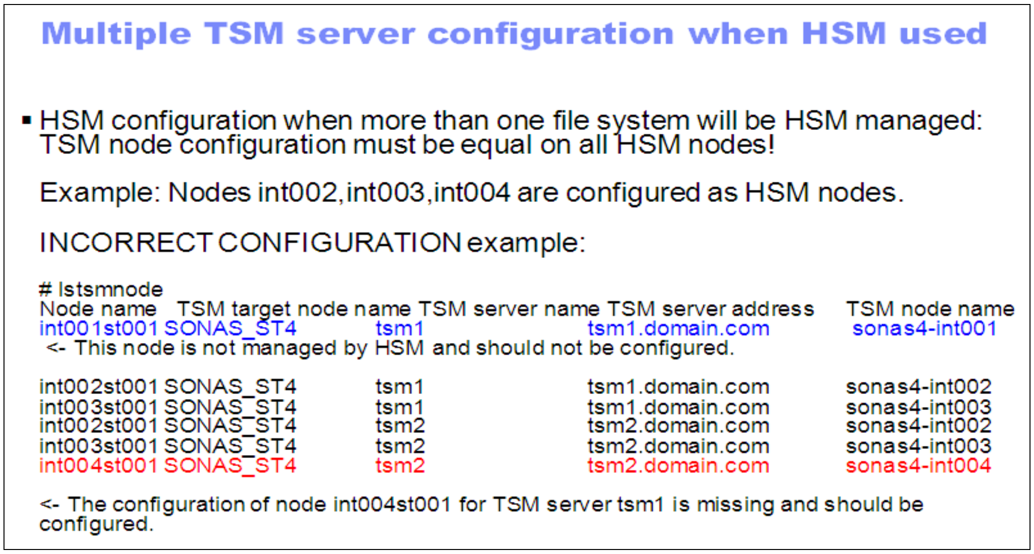


CLI commands:
lstsmnode (show TSM nodes)
lsbackupfs (TSM server – nodes – file system association)
lsjobstatus (Job history and active ones (not only backup))

```
[root@sonas5.mgmt001st001 ~]# lstsmnode
Node name    TSM target node name TSM server name TSM server address          TSM node name
int001st001  sonas_st5            tsm1            fscc-hs20-2.mainz.de.ibm.com SONAS5int001st001
mgmt001st001 sonas_st5            tsm1            fscc-hs20-2.mainz.de.ibm.com SONAS5mgmt001st001
mgmt002st001 sonas_st5            tsm1            fscc-hs20-2.mainz.de.ibm.com SONAS5mgmt002st001
EFSSG1000I The command completed successfully.
[root@sonas5.mgmt001st001 ~]# lsbackupfs
File system TSM server List of nodes
gpfs0       tsm1       int001st001,mgmt001st001,mgmt002st001

EFSSG1000I The command completed successfully.
[root@sonas5.mgmt001st001 ~]# lsjobstatus -j backup --all
File system Job    Job id Status  Start time                End time/Progress         RC Message
gpfs0       backup 0      done(+) 9/14/11 11:22:36 AM CEST 9/14/11 11:23:10 AM CEST 0  EFSSG1000I The command completed successfully.
gpfs0       backup 1      done(+) 9/14/11 11:25:05 AM CEST 9/14/11 11:25:22 AM CEST 0  EFSSG1000I The command completed successfully.
gpfs0       backup 2      done(+) 9/14/11 11:26:26 AM CEST 9/14/11 11:26:44 AM CEST 0  EFSSG1000I The command completed successfully.
EFSSG1000I The command completed successfully.
[root@sonas5.mgmt001st001 ~]# 
```

Figure 6-29   Example of backup tool commands

Figure 6-30 shows the Tivoli Storage Manager Server configuration confirmation by running **lstsmnode**.



Multiple TSM server configuration when HSM used

- HSM configuration when more than one file system will be HSM managed: TSM node configuration must be equal on all HSM nodes!

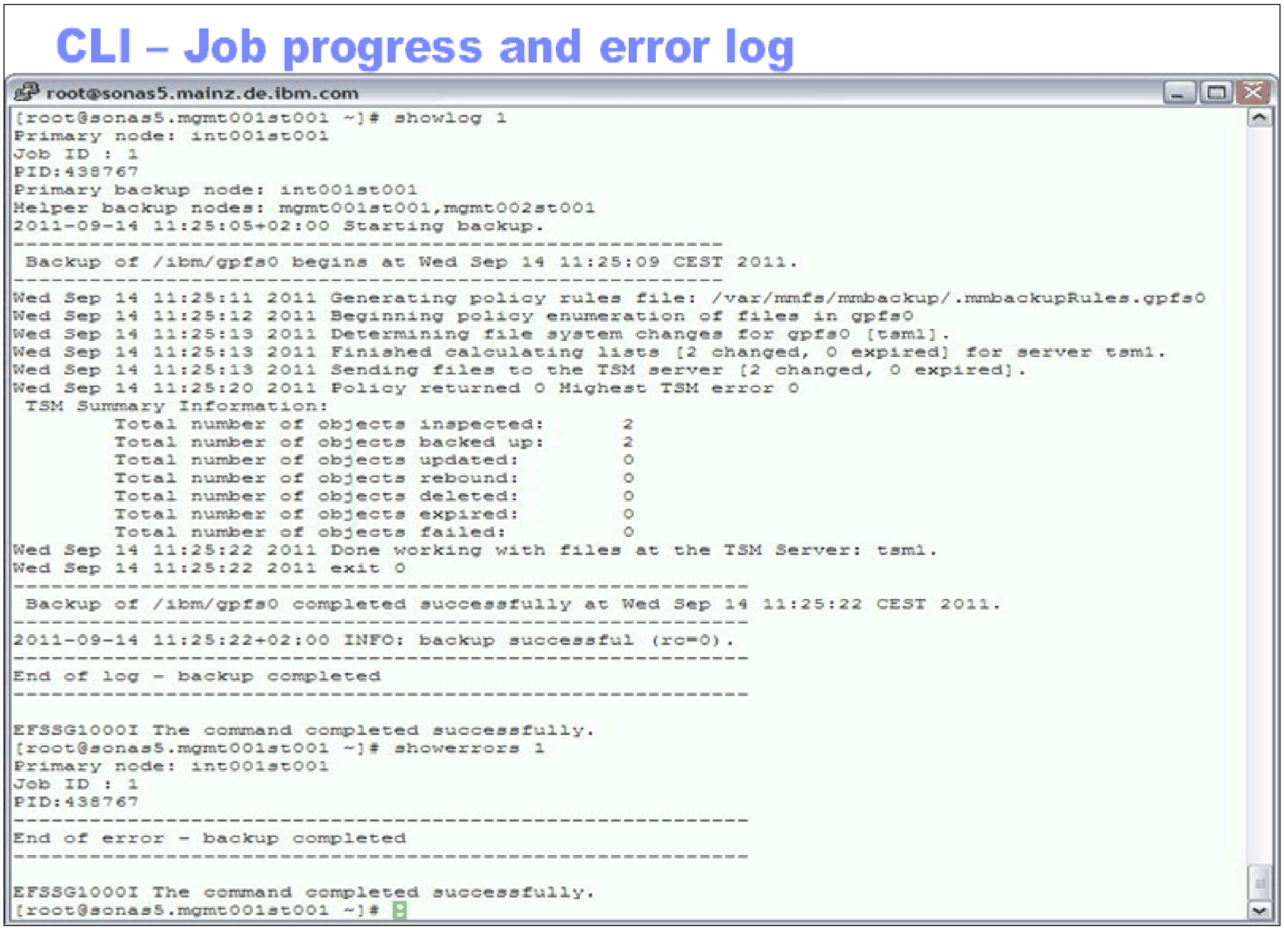
Example: Nodes int002,int003,int004 are configured as HSM nodes.

INCORRECT CONFIGURATION example:

```
# lstsmnode
Node name   TSM target node name TSM server name TSM server address    TSM node name
int001st001 SONAS_ST4            tsm1            tsm1.domain.com       sonas4-int001
 <- This node is not managed by HSM and should not be configured.

int002st001 SONAS_ST4            tsm1            tsm1.domain.com       sonas4-int002
int003st001 SONAS_ST4            tsm1            tsm1.domain.com       sonas4-int003
int002st001 SONAS_ST4            tsm2            tsm2.domain.com       sonas4-int002
int003st001 SONAS_ST4            tsm2            tsm2.domain.com       sonas4-int003
int004st001 SONAS_ST4            tsm2            tsm2.domain.com       sonas4-int004

<- The configuration of node int004st001 for TSM server tsm1 is missing and should be
configured.
```

Figure 6-30 SONAS backup configuration check

Figure 6-31 shows the List Job progress by running **showlog** and the appropriate job ID.

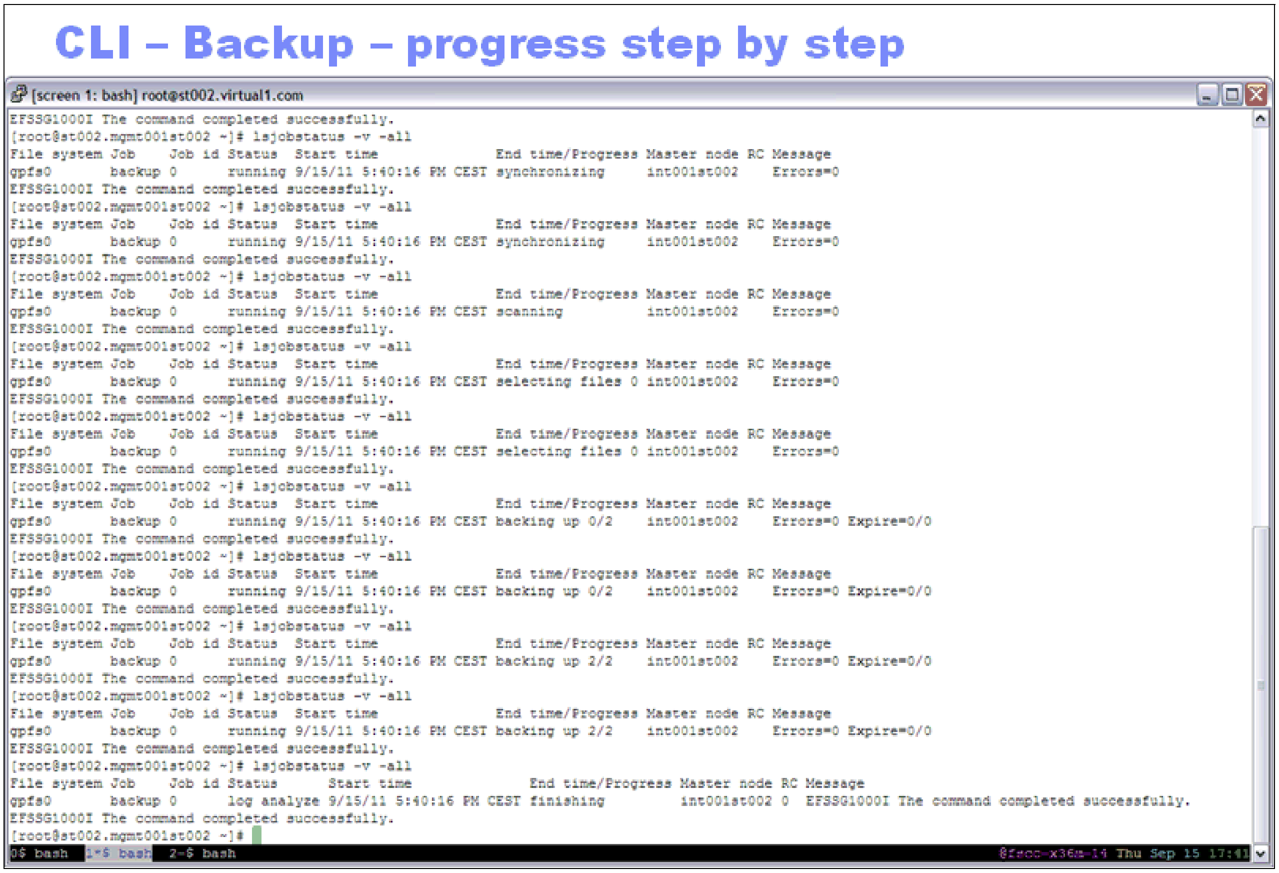


CLI – Job progress and error log

```
[root@sonas5.mgmt001st001 ~]# showlog 1
Primary node: int001st001
Job ID : 1
PID:438767
Primary backup node: int001st001
Helper backup nodes: mgmt001st001,mgmt002st001
2011-09-14 11:25:05+02:00 Starting backup.
----------------------------------------------------------
 Backup of /ibm/gpfs0 begins at Wed Sep 14 11:25:09 CEST 2011.
----------------------------------------------------------
Wed Sep 14 11:25:11 2011 Generating policy rules file: /var/mmfs/mmbackup/.mmbackupRules.gpfs0
Wed Sep 14 11:25:12 2011 Beginning policy enumeration of files in gpfs0
Wed Sep 14 11:25:13 2011 Determining file system changes for gpfs0 [tsm1].
Wed Sep 14 11:25:13 2011 Finished calculating lists [2 changed, 0 expired] for server tsm1.
Wed Sep 14 11:25:13 2011 Sending files to the TSM server [2 changed, 0 expired].
Wed Sep 14 11:25:20 2011 Policy returned 0 Highest TSM error 0
 TSM Summary Information:
        Total number of objects inspected:      2
        Total number of objects backed up:      2
        Total number of objects updated:        0
        Total number of objects rebound:        0
        Total number of objects deleted:        0
        Total number of objects expired:        0
        Total number of objects failed:         0
Wed Sep 14 11:25:22 2011 Done working with files at the TSM Server: tsm1.
Wed Sep 14 11:25:22 2011 exit 0
------------------------------------------------------------
 Backup of /ibm/gpfs0 completed successfully at Wed Sep 14 11:25:22 CEST 2011.
------------------------------------------------------------
2011-09-14 11:25:22+02:00 INFO: backup successful (rc=0).
------------------------------------------------------------
End of log - backup completed
------------------------------------------------------------

EFSSG1000I The command completed successfully.
[root@sonas5.mgmt001st001 ~]# showerrors 1
Primary node: int001st001
Job ID : 1
PID:438767
------------------------------------------------------------
End of error - backup completed
------------------------------------------------------------

EFSSG1000I The command completed successfully.
[root@sonas5.mgmt001st001 ~]#
```

Figure 6-31 Showlog progress list

Figure 6-32 shows some tasks that use the **lsjobstatus -v -all** command.



CLI – Backup – progress step by step

```
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST synchronizing     int001st002    Errors=0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST synchronizing     int001st002    Errors=0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST scanning          int001st002    Errors=0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST selecting files 0 int001st002    Errors=0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST selecting files 0 int001st002    Errors=0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST backing up 0/2    int001st002    Errors=0 Expire=0/0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST backing up 0/2    int001st002    Errors=0 Expire=0/0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST backing up 2/2    int001st002    Errors=0 Expire=0/0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status  Start time              End time/Progress Master node RC Message
gpfs0       backup 0      running 9/15/11 5:40:16 PM CEST backing up 2/2    int001st002    Errors=0 Expire=0/0
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]# lsjobstatus -v -all
File system Job    Job id Status      Start time              End time/Progress Master node RC Message
gpfs0       backup 0      log analyze 9/15/11 5:40:16 PM CEST finishing         int001st002 0  EFSSG1000I The command completed successfully.
EFSSG1000I The command completed successfully.
[root@st002.mgmt001st002 ~]#
0$ bash  1*$ bash  2-$ bash                                                                          @fscc-x36m-14 Thu Sep 15 17:41
```

Figure 6-32 Step-by-step capture by running lsjobstatus -v -all

6.2.7 Common routines for managing Tivoli Storage Manager backup and restore

This section describes common SONAS tasks for managing Tivoli Storage Manager backup integration. Information about these tasks is available in the Help menus and IBM Knowledge Center for the SONAS V1.3 Release.

Configuring the backup

To configure the backup, run **cfgbackupfs**. For example, to back up the gpfs0 file system to the Tivoli Storage Manager server name tsmserver1 on the Interface node name int001st001, run the following command:

```
# cfgbackupfs gpfs0 tsmserver1 int001st001
```

The Interface node is specified when you configure the Tivoli Storage Manager Server stanza. More than one Interface node can be specified in a comma-separated list.

Adding or removing an Interface node or node list

To add or remove an Interface node or node list from a configured file system backup, run **chbackupfs** and specify the node or node list with the **--add** or **--remove** option.

Listing configured backups

To list the configured backups, run **`lsbackupfs`**. Example 6-2 shows a sample command and the output.

Example 6-2 Sample lsbackupfs command

```
# lsbackupfs
File system TSM server List of nodes Status Start time End time Message Last update
gpfs0       tsmserver1  int001st001   NOT_STARTED N/A       N/A        1/15/10 4:39 PM
```

A daily scheduled backup of the specified file system is created with the default run time of 2 a.m. This time can be altered in the GUI by clicking **Files** → **Services** → **Backup**, or with the CLI.

Manual backup

To run a manual backup, run **`startbackup`**. If you specify a comma-separated list of file systems when you submit the command, backups are started for those file systems. If no file system is specified, all file systems that have backups configured begin their backups. For example, to start backing up the file system `gpfs0`, run the following command:

```
# startbackup gpfs0
```

Listing status messages, completion dates, and times

To list status messages and completion dates and times, run **`lsbackupfs`**, as shown in Example 6-3.

Example 6-3 Sample lsbackupfs command and output

```
# lsbackupfs gpfs0
Filesystem Date Message
gpfs0      20.01.2010 02:00:00.000  G0300IEFSSG0300I The filesystem gpfs0 backup
started.
gpfs0      19.01.2010 12:30:52.087  G0702IEFSSG0702I The filesystem gpfs0 backup
was done successfully.
gpfs0      18.01.2010 02:00:00.000  G0300IEFSSG0300I The filesystem gpfs0 backup
started.
# lsbackupfs
File system TSM server List of nodes Status Start time End time Message
Last update
gpfs0       SONAS_SRV_2 int001st001   RUNNING 2/18/10 12:30 PM N/A
log:/var/log/cnlog/cnbackup/cnbackup_gpfs0_20100218123051.log, on host:
int001st001 2/18/10 12:30 PM
```

Monitoring backup progress

You can monitor the progress of the backup process by running **`query session`** in the Tivoli Storage Manager administrative CLI client. Run this command twice and compare the values in the Bytes Recvd column of the output. Incremental values indicate that the process is in progress, and identical values indicate that the backup process stopped.

> **Tip:** If a CTDB failover, CTDB stop, or connection loss occurs on an Interface node, any backup that is running then on that Interface node cannot send its data, and the overall backup fails.

Changing an existing backup configuration

To add or remove Interface nodes from a backup configuration, any running backups must be stopped and removed, Tivoli Storage Manager nodes must be added or removed, and the backup configuration must be re-created.

Scheduling the Tivoli Storage Manager file system backups

File system backups that use Tivoli Storage Manager can be scheduled by using the GUI or CLI.

Restoring a file system

Files that are backed up through the Tivoli Storage Manager integration can be restored by running **startrestore**.

To work with this function in the management GUI, log on to the GUI and click **Files** → **File Services**.

Before you restore a file system, determine whether a backup is running and when backups were completed by running **lsbackupfs** and specifying the file system. For example, the command to display the gpfs0 file system backup listing shows the output in the format that is shown in Example 6-4.

Example 6-4 Sample lsbackupfs command and output

```
# lsbackupfs gpfs0
Filesystem Date Message
gpfs0      20.01.2010 02:00:00.000    G0300IEFSSG0300I The filesystem gpfs0 backup started.
gpfs0      19.01.2010 06:10:00.123    G0702IEFSSG0702I The filesystem gpfs0 backup was done
successfully.
gpfs0      15.01.2010 02:00:00.000    G0300IEFSSG0300I The filesystem gpfs0 backup started.
```

The startrestore command

Restore the backup by running **startrestore** and specify a file system name pattern. You cannot restore two files systems at the same time, so the file pattern cannot match more than one file system name. Use the **-t** option to specify a date and time in the format dd.MM.yyyy HH:mm:ss.SSS to restore files as they existed then. If a time is not specified, the most recently backed up versions are restored. For example, to restore the /ibm/gpfs0/temp/* file pattern to its backed-up state as of January 19, 2010 at 12:45 PM, run the following command:

```
# startrestore "/ibm/gpfs0/temp/*" -t "19.01.2010 12:45:00.000"
```

> **Tip:** The **-R** option overwrites files, and has the potential to overwrite newer files with older data.

Determining whether restore is running

Run **lsbackupfs** to determine whether a restore operation is running. The Message field displays RESTORE_RUNNING if a restore operation is running on a file system.

Monitoring the progress of the restore process

You can monitor the progress of the restore process by running **QUERY SESSION** in the Tivoli Storage Manager administrative CLI client. Run this command twice and compare the values in the Bytes Sent column of the output. Incremental values indicate that the process is in progress, and identical values indicate that the restore process has stopped.

Error message: The following error message can occur while you are restoring millions of files:

```
ANS1030E The operating system refused a TSM request for memory allocation.
2010-07-09 15:51:54-05:00 dsmc return code: 12
```

If the file system is managed by Tivoli Storage Manager for Space Management, you can break down the restore into smaller file patterns, or subdirectories that contain fewer files.

If the file system is not managed by Tivoli Storage Manager for Space Management, try to force a no-query-restore (NQR) by altering the path that is specified for the restore to include all files by putting a wildcard ("*") after the file system path. For example:

```
# startrestore "ibm/gfps0/*"
```

This command attempts a no query restore, which minimizes memory issues with the Tivoli Storage Manager Client because the Tivoli Storage Manager Server does the optimization of the file list. If you are still unable to restore many files at the same time, break down the restore into smaller file patterns, or subdirectories that contain fewer files.

Stopping a running restore session

You can use the CLI to stop a running Tivoli Storage Manager restore session. To stop a running restore session, run **`stoprestore`**.

Listing backup configurations

Run **`lsbackupfs`** to list backup configurations for a file system. You can also use the GUI to work with this function.

To work with this function in the management GUI, log on to the GUI and click **Files** → **File Services**.

Displaying backup configurations

This section describes how to display the backup configurations.

CLI command

To display the backup configurations, run **`lsbackupfs`**.

For each file system, the display includes the file system, the Tivoli Storage Manager Server, the Interface nodes, the status of the backup, the start time of the backup, the end time of the most previous completed backup, the status message from the last backup, and the last update. Example 6-5 shows a backup that started on 1/20/2010 at 2 a.m.

Example 6-5 Sample lsbackupfs command and output

```
[root@examplemgmt.mgmt001st001 ~]# lsbackupfs
File system TSM server  List of nodes
gpfs0       SONAS_SRV_2 int001st001,int002st001
Status  Start time      End time         Message                          Last update
RUNNING 1/20/10 2:00 AM 1/19/10 11:15 AM INFO: backup successful (rc=0). 1/20/10 2:00 AM
```

GUI navigation

To work with this function in the management GUI, log on to the GUI and click **Files** → **Services** → **Backup**.

Listing file system backups

Completed and in-progress file system backups can be listed by running **lsbackupfs**.

CLI command

To list the running and completed backups, run **lsbackupfs**. Specify a file system to show completed and running backups for only that file system. For example, to list the backups for file system gpfs0, submit the command that is shown in Example 6-6.

Example 6-6 Sample lsbackupfs command

```
lsbackupfs gpfs0
The output displays in the following format:
[root@examplemgmt.mgmt001st001 ~]# lsbackupfs gpfs0
Filesystem Date Message
gpfs0      20.01.2010 02:00:00.000  G0300IEFSSG0300I The filesystem gpfs0 backup
started.
gpfs0      19.01.2010 16:08:12.123  G0702IEFSSG0702I The filesystem gpfs0 backup
was done successfully.
gpfs0      15.01.2010 02:00:00.000  G0300IEFSSG0300I The filesystem gpfs0 backup
started.
```

> **Tip:** The output of the **lsbackupfs** command shows only backup session results for the past 7 days.

GUI navigation

To work with this function in the management GUI, log on to the GUI and click **Files** → **Services** → **Backup**.

Viewing backup and restore results

Using CLI commands, you can view the results of a Tivoli Storage Manager backup or restore.

To view the logs of a previous backup or restore operation, run **showlog** or, in the GUI, click **Files** → **Services** → **Backup**.

The specific operation can be identified by using the JobID, Job, FileSystem, ClusterID, Count, and time options. For example:

- **showlog 15** shows the log for a job with jobID 15.
- **showlog backup:gpfs0** shows the backup log for the latest backup job that is done for file system gpfs0.
- **showlog 15 -count 20** shows only the last 20 lines of the log for a job with jobID 15.
- **showlog backup:gpfs0 -t “03.05.2011 14:18:21.184”** shows the backup log taken of file system gpfs0 at the date and time specified.

Viewing errors that are related to backup or restore operations

To view the errors from a previous backup or restore operation, run **showerrors**.

Overview

The specific operation can be identified by using the JobID, Job, FileSystem, ClusterID, Count, and time options. For example:

- **showlog 15** shows the log for a job with jobID 15.
- **showlog backup:gpfs0** shows the backup log for the latest backup job that is done for file system gpfs0.
- **showlog 15 -count 20** shows only the last 20 lines of the log for a job with jobID 15.
- **showlog backup:gpfs0 -t "03.05.2011 14:18:21.184"** shows the backup log taken of file system gpfs0 at the date and time specified.

GUI navigation

To work with this function in the management GUI, log on to the GUI and click **Files** → **Services** → **Backup**.

6.3 Tivoli Storage Manager server-side data deduplication

Data deduplication is a method for eliminating redundant data. Only one instance of the data is retained on storage media, such as disk or tape. Other instances of the same data are replaced with a pointer to the retained instance.

Although Tivoli Storage Manager V 6.2 and later supports both client-side (source-side) and server-side (target-side) deduplication, the SONAS system does not support Tivoli Storage Manager client-side deduplication as a Tivoli Storage Manager Client. Tivoli Storage Manager V6.1 and later supports server-side deduplication, which can be used with the SONAS system configured as a Tivoli Storage Manager Client. For more information about Tivoli Storage Manager, see Tivoli Storage Manager publications.

6.4 Tivoli Storage Manager server-side common operations guide

This section provides information about common server-side operations for Tivoli Storage Manager.

> **Tip:** The following information concerns Tivoli Storage Manager services when Tivoli Storage Manager is installed and specifically dedicated to the IBM SONAS solution. The tips and examples that are provided in this document apply specifically to Tivoli Storage Manager Server and might not be pertinent to your specific configuration. Detailed information is beyond the scope of this publication. Consult your Tivoli Storage Manager expert before you use this information in your production environments.

6.4.1 Data mobility and validation

A common concern is that data that remains on tape for an extended time can be invalid or damaged because of the nature of magnetic tape. Tivoli Storage Manager works in the background to keep data alive. As data in the backup status expires, the population thresholds of media decrease, and by default Tivoli Storage Manager meets those thresholds and migrate the remaining contents of media to a fresh media device. This process also validates the readability of the data against the metadata in the migration process. This process is managed by the server and the library and not by SONAS directly. The Tivoli Storage Manager Server tracks the data locations that are related to each device and client, and data is not deleted from the source tape until it is validated on the target tape.

Logging in to the Tivoli Storage Manager Server

Before you do any Tivoli Storage Manager administrative tasks, it is necessary to log in to the Tivoli Storage Manager Server. To do so, complete the following steps:

1. Telnet to the Tivoli Storage Manager Server TSMSRV01 (TSM-ServerName) by running the following command:

   ```
   #telnet tsmsrv01
   ```

2. Log in to the Tivoli Storage Manager Server Administrative CLI by running this command:

   ```
   #dsmadmc –id=admin –pass=adminpassword
   ```

Starting the Tivoli Storage Manager Server instance

To start the Tivoli Storage Manager Server, complete the following steps:

1. Telnet to the Tivoli Storage Manager Server TSMSRV01.
2. Log in as tsm1client.
3. Run the following commands:

   ```
   # cd /home/tsm1client/tsm1client
   # nohup /opt/tivoli/tsm/server/bin/dsmserv –q &
   ```

Daily queries

Tivoli Storage Manager typically runs without any user intervention. However, you must check it occasionally. Run the following Tivoli Storage Manager commands daily to check the status of the server.

Database and Log usage

Check the Tivoli Storage Manager database and log files and make sure that their percent utilization is not too high by running the following commands:

```
tsm: TSM1CLIENT> Query DBase
tsm: TSM1CLIENT> Query DBase Format=Detail
tsm: TSM1CLIENT> Query LOG
tsm: TSM1CLIENT> Query LOG Format=Detail
```

If utilization is too high, you can add more space.

Client events

For Tivoli Storage Manager Server initiated schedules only, check that the backup/archive schedules did not fail by running the following command:

```
tsm: TSM1CLIENT> Query  EVent  *  *  Type=Admin BEGINDate=-1 BEGINTime=17:00
(Format=Detail)
```

The first “*” is for the domain name.

The second “*” is for the schedule name.

Administrative events

Check that the administrative command schedules did not fail by running the following commands:

```
tsm: TSM1CLIENT> Query  Event  *  Type=Administrative BEGINDate=TODAY
(Format=Detail)
```

The “*” is for the schedule name.

6.4.2 Tape operations

The following topics describe various tape operations.

Scratch tapes

Check for the number of available scratch tapes by running the following command:

```
tsm: TSM1CLIENT> RUN Q_SCRATCH
```

`Query_Scratch` is a user-defined server command script.

Read-only tapes

Check for any tapes with an access setting of “read-only” by running the following command:

```
tsm: TSM1CLIENT> Query VOLume ACCess=READOnly
```

If any tapes are in RO mode, check the Tivoli Storage Manager Server activity log for related errors.

Unavailable tape

Check for any tapes with an access setting of “unavailable” by running the following command”:

```
tsm: TSM1CLIENT> Query VOLume ACCess=UNAVailable
```

If there are any tapes in this mode, check the Tivoli Storage Manager Server activity log for related errors and take appropriate actions.

Checking in new tapes

New tapes must be labeled by Tivoli Storage Manager before use. To insert new tapes into the 3584 library for use, insert them into the library I/O station; then, from the Tivoli Storage Manager Administrative CLI, run the following commands to check in these tapes as SCRATCH:

```
tsm: TSM1CLIENT> checkin libv 3584lIB search=BULK status=scratch checklabel=bar
tsm: TSM1CLIENT>query request (make note of REPLY number)
tsm: TSM1CLIENT>reply <REPLY NUMBER>
```

Checking in “existing” tapes as scratch tapes

Tapes that already have labels on them and *no* longer contain valid data can be checked into the 3584 Library as scratch tapes. Insert those tapes into the 3584 I/O station.

From the Tivoli Storage Manager Administrative CLI, run the following commands to check in those tapes:

```
tsm: TSM1CLIENT> checkin libv 3584lIB search=BULK status=scratch checklabel=bar
tsm: TSM1CLIENT>query request (make note of REPLY number)
tsm: TSM1CLIENT>reply <REPLY NUMBER>
```

Checking in “existing” tapes as private tapes

Tapes that already are labeled by Tivoli Storage Manager and contain data can be checked into the 3584 Library as private tapes. Insert those tapes into the 3584 I/O station.

From the Tivoli Storage Manager Administrative CLI, run the following commands to check in those tapes as private:

- **`tsm: TSM1CLIENT> CHECKIn  LIBVolume  3584LIB   SEARCH=BULK  STATUS=PRIvate CHECKLabel=BAR`**
- **`tsm: TSM1CLIENT>query request`** (Note the REPLY number.)
- **`tsm: TSM1CLIENT>reply <REPLY NUMBER>`**

Offsite processing (DRM)

To move offsite tapes to the vault and retrieve empty tapes for reuse, complete these procedures daily:

1. Create a Tivoli Storage Manager DB Backup to send offsite with the tape media to be sent offsite and wait for the process to complete before you do the next step by running the following command:

   ```
   tsm: TSM1CLIENT> backup db type=dbsnapshot devc=lto5class
   ```

2. You can pre-determine which tapes (Tivoli Storage Manager database and Tivoli Storage Manager data volumes) are to be taken out of the library by running the following command:

   ```
   tsm: TSM1CLIENT> Query  DRMedia  WHERESTate=MOuntable Source=DBSnapshot
   ```

3. The following command runs the tasks for the creation of the offsite media, including checking out, from the tape library, the tapes that are available to be sent offsite:

   ```
   tsm: TSM1CLIENT> move DRMedia WHERESTate=MOuntable Source=DBSnapshot
   tostate=VAULT
   ```

   Remove the tapes from the 3584 Library I/O station and send them to the offsite location.

4. The following command lists tapes that can be recalled from the offsite location for reuse:

   ```
   tsm: TSM1CLIENT> Query  DRMedia  * WHERESTate=VAULTRETrieve Source=DBSnapshot
   ```

5. If there are tapes available that were retrieved from offsite and can be checked in to the library, complete the following tasks:

   – Insert the tapes into the 3584 I/O slots and close the I/O door.

   – On the front panel of the 3584, select **Assign tapes to library_a** and run the following commands:

   ```
   tsm: TSM1CLIENT> checkin libv 3584lIB search=BULK status=scratch
   checklabel=bar
   tsm: TSM1CLIENT>query request (make note of REPLY number)
   tsm: TSM1CLIENT>reply <REPLY NUMBER>
   ```

Moving data off bad tapes

Occasionally, tapes get media errors and must be removed from the library. You can identify those tapes by running the following commands:

```
tsm: TSM1CLIENT> Query VOLume ACCESS=READOnly Format=Detail
tsm: TSM1CLIENT> Query VOLume ACCESS=UNAVailable Format=Detail
```

Usually (but not always), tapes that changed from the“read/write” status to a “read-only” status have write errors. Tapes that change from a “read/write” status to an “unavailable” status have read errors.

The first thing to do with these tapes is to track which ones are changing. For the ones that change to “read-only”, you might want to change them back to “read/write” and see whether the problem recurs by running the following command:

```
tsm: TSM1CLIENT> UPDate VOLume volume_name ACCESS=READWrite
```

volume_name is the name of the volume with which you are having problems.

If these tapes change back to the “read-only” status a second time, move the data off those tapes and remove those tapes from the library by running the following commands. Also, move the tapes that changed to the “unavailable” status the first time.

- **`tsm: TSM1CLIENT> MOVe Data volume_name STGpool=storagepool_name`**

  volume_name is the name of the volume with which you are having problems and *storagepool_name* is the storage pool where you want to move that data.

- **`tsm: TSM1CLIENT> CHECKOut LIBVolume library_name volume_name REMOVE=Yes`**

  library_name is the name of the library the Tivoli Storage Manager Server uses and volume_name is the name of the volume whose data was moved.

If you cannot move the data off the tapes that are suspected to be bad, then you must either delete those tapes volumes from within Tivoli Storage Manager and check them out of the library, or restore the data from a copy storage pool tape volume.

If the tape was part of the offsitepool storage pool, then you can delete the data from the tape by running following command, and the next time the **`backup stg`** command runs, that data is copied to another **`OFFSITEPOOL`** storage pool volume:

```
tsm: TSM1CLIENT> DELete VOLume volume_name DISCARDDATA=Yes
```

volume_name is the name of the volume with which you are having problems.

After you run this command, all data that was backed up to that tape volume is gone. To remove the tape from the library, if needed, run **`checkout libvolume`**.

If the tape was part of a TAPEPOOL storage pool, and you need the data that was on the tape, you must restore that data from the OFFSITEPOOL storage pool. To determine which OFFSITEPOOL storage pool volumes you need, run the following command:

```
tsm: TSM1CLIENT> RESTORE Volume volume_name Preview=Yes
```

volume_name is the name of the volume with which you are having problems.

Retrieve these tapes from the vault, check them into the library, and set their access to `READONLY`. Then, run the following command to restore the damaged volume:

```
tsm: TSM1CLIENT> RESTORE Volume volume_name
```

volume_name is the name of the volume with which you are having problems.

This command marks the volume access as DESTROYED and attempts to restore all the data that was on it to another volume in the same storage pool.

After that process completes, change the access of the tapes that are used from the OFFSITEPOOL storage pool to OFFSITE and check them back out of the library.

Also, check the tape with which you are having problems out of the library.

6.5 Using the Tivoli Storage Manager HSM client

SONAS offers an HSM integration option that allows you to save cluster disk capacity by keeping low use data on a tape-based storage platform (useful to many of the existing clients). It is accomplished by sending (migrating) data to external storage devices that are managed by Tivoli Storage Manager.

Similar to the data protection preparation, this solution requires that an external Tivoli Storage Manager Server is available to manage the data and that a specific storage pool is defined on that Tivoli Storage Manager Server to store the data that is targeted on tape devices. This solution uses tapes as virtual file system capacity that aligns (virtually) within the same file system as the original data as a virtual storage tier. Policies manage placement of that data or movement of that data from disk repositories to the tape target file system.

The SONAS client-side software is already installed, which simplifies implementation. All that is required from a licensing perspective is client licenses that are based on the number of Interface nodes in the cluster and processors. The license is then installed on the Tivoli Storage Manager Server (not the client or SONAS solution itself).

The data on tape obviously enhances the importance of reliability of the access to that data from the Interface nodes. If the Tivoli Storage Manager Server environment is for any reason down for maintenance, then the file system does not offer clients access to that data. So, it is important that if you decide to place file system data in HSM policy management, you either harden your Tivoli Storage Manager Server environment for high availability or live with the expectation that, if the Tivoli Storage Manager service is down, that data is not available. Although it is out of scope to define the HA requirements of the Tivoli Storage Manager Server environment in the context of this document, be aware of the effect of this consideration.

The Tivoli Storage Manager HSM clients that are run in the SONAS Interface nodes and use the Ethernet connections within the Interface nodes to connect to the external, client-provided, Tivoli Storage Manager Server. The primary goal of the HSM support is to provide a high performance HSM link between a SONAS subsystem and an external tape subsystem. SONAS HSM support has the following requirements:

- One or more external Tivoli Storage Manager servers must be provided and the servers must be accessible through the external Ethernet connections on the Interface nodes.
- The **cfgtsmnode** command must be run to configure the Tivoli Storage Manager environment.
- SONAS GPFS policies drive migration, so Tivoli Storage Manager HSM automigration must be disabled.

Every Interface node has a Tivoli Storage Manager HSM client that is installed alongside with the standard Tivoli Storage Manager backup/archive client. An external Tivoli Storage Manager Server is attached to the Interface node through the Interface node Ethernet connections. The Tivoli Storage Manager HSM client supports the SONAS GPFS file system by using the Data Management API (DMAPI).

Before you configure HSM to a file system, you must complete the Tivoli Storage Manager initial setup that uses the **cfgtsmnode** command, as illustrated in 6.2.4, “Configuring Interface nodes and file systems for Tivoli Storage Manager” on page 377. SONAS HSM uses the same Tivoli Storage Manager Server that was configured for the SONAS Tivoli Storage Manager backup client, and using the same server allows Tivoli Storage Manager to clone data between the Tivoli Storage Manager Server backup storage pools and HSM storage pools.

With the SONAS Tivoli Storage Manager Client, one Tivoli Storage Manager Server stanza is provided for each GPFS file system. Therefore, one GPFS file system can be connected to one single Tivoli Storage Manager Server. Multiple GPFS file systems can use either the same or various Tivoli Storage Manager servers. Multiple Tivoli Storage Manager servers might be needed when you have many files in a file system.

Attention: You cannot remove SONAS HSM without help from IBM.

Unlike the backup and restore options for SONAS with Tivoli Storage Manager, the SONAS HSM client must be configured to run on all the Interface nodes in the SONAS cluster. Migrated files can be accessed from any node, so the Tivoli Storage Manager HSM client must be active on all the nodes. As it becomes an extension of your file system, any client that is connected to a SONAS Interface node to access data must be able to access the HSM data through that same connection. All SONAS HSM configuration commands are run by using the SONAS CLI and not the GUI.

The files in HSM policies appear to the clients as regular files. A stub file exists on disk and, when accessed, it is called and served from tape. The size of the stub file is programmable, so you can make all stubs 16 K or even 1 MB, and that affects how much of the file is immediately accessible while the rest of the file is called up.

Also, other behaviors in HSM file access and migration policies are programmable. For example, by default, all data that is destined for HSM tape, before migration, requires a copy in backups. It can be modified, but it is a preferred practice to persist with this default setting. You can also choose to migrate data back to disk when it is recalled from archive. You can choose to evacuate the disk-based copy when pushed to tape or leave a copy on disk. There might also be other options.

HSM offers two kinds of recall: transparent recall and policy-based recalls when large numbers of files are recalled based on policies. Transparent recall acts only on an individual file. A request is sent to the Tivoli Storage Manager Server that mounts the media and recalls the file. When large numbers of files that were migrated to multiple tape cartridges must be recalled, Tivoli Storage Manager must mount each of them to do the recall. Tivoli Storage Manager offers a tape-optimized recall to speed up the recall process of batches of HSM migrated files from sequential storage. It is designed to speed up the recall with the following features:

- Avoids frequent tape mount and unmount operations.
- Avoids excessive tape seek operations.
- Enables tape drives to go to streaming mode, if possible, for optimal performance. Streaming mode is possible only if the recalled files are located contiguously on the tape.

- Orders and recalls files in separate steps so that you can restrict the recall operations on certain tapes.
- Recalls files from several tape drives in parallel to increase recall throughput.

The Tape Optimized Recall function is automatically available in SONAS, and so there are no specific commands or parameters to run it.

> **Tip:** HSM can be configured to support complex or clever conditions that might not be identified in this document. As with other externally managed ISV type solutions, it is of great value to include a Tivoli Storage Manager and HSM expert in the planning phase of your HSM solution. Use the information here to provide a quick start on what you need to consider and a basic understanding on how to deploy an HSM solution with SONAS, along with a few preferred practice considerations for using HSM.

6.5.1 SONAS HSM concepts

When you use SONAS HSM, new and most frequently used files remain on your local file systems, and the files that you use less often are automatically migrated to storage media that is managed by an external Tivoli Storage Manager Server. Migrated files still appear local and are transparently migrated to and retrieved from the Tivoli Storage Manager Server. Files can also be prioritized for migration according to their size and the number of days since they were last accessed, which allows users to maximize local disk space. Enabling space management for a file system can provide the following benefits:

- Extends local disk space by using storage on the Tivoli Storage Manager Server
- Takes advantage of lower-cost storage resources that are available in your network environment
- Allows for automatic migration of old and large files to the Tivoli Storage Manager Server
- Helps to avoid out-of-disk space conditions on client file systems

To migrate a file, HSM sends a copy of the file to a Tivoli Storage Manager Server and replaces the original file with a stub file on the local file system. A stub file is a small file that contains the information that is required to locate and recall a migrated file from the Tivoli Storage Manager Server. It also makes it appear as though the file is still on your local file system. Similar to backups and archives, migrating a file does not change the access time (atime) or permissions for that file.

SONAS storage management policies control and automate the migration of files between storage pools and external storage.

A feature of automatic migration is the premigration of eligible files. The HSM client detects this condition and migrates automatically eligible files to the Tivoli Storage Manager Server. This migration process continues to migrate files until the file system utilization falls below the defined low threshold value. At that point, the HSM client premigrates files. To premigrate a file, HSM copies the file to Tivoli Storage Manager storage and leaves the original file intact on the local file system (that is, no stub file is created).

An identical copy of the file is both on the local file system and in Tivoli Storage Manager storage. The next time migration starts for this file system, HSM can quickly change premigrated files to migrated files without having to spend time copying the files to Tivoli Storage Manager storage. HSM verifies that the files have not changed since they were premigrated and replaces the copies of the files on the local file system with stub files. When automatic migration is performed, premigrated files are processed before resident files because this allows space to be freed in the file system more quickly.

A file that is managed by HSM can be in multiple states:

Resident A resident file is on the local file system. For example, a newly created file is a resident file.

Migrated A migrated file is a file that was copied from the local file system to Tivoli Storage Manager storage and replaced with a stub file.

Premigrated A premigrated file is a file that was copied from the local file system to Tivoli Storage Manager storage but has not been replaced with a stub file. An identical copy of the file is both on the local file system and in Tivoli Storage Manager storage. A file can be in the premigrated state after premigration. If a file is recalled but not modified, it is also in the premigrated state.

To return a migrated file to your workstation, access the file in the same way as you might access a file that is on your local file system. The HSM recall daemon automatically recalls the migrated file from Tivoli Storage Manager storage. This process is referred to as transparent recall.

6.5.2 Configuring SONAS HSM

This section describes the commands that are used to configure SONAS HSM.

The **cfghsmnodes** command configures a specified list of nodes to be used by HSM, and unconfigures Management nodes, which are in the list, from use for HSM. At least two nodes must be provided in the node list because Management nodes must be contacted during this configuration. All Management nodes must be accessible. All nodes that are provided in the list of nodes to be enabled must be configured for the Tivoli Storage Manager Server that is provided by the argument. The **cfgtsmnode** command must be used to enable the nodes before you use this **cfghsmnodes** command. This command activates or deactivates the basic HSM functions in the nodes, as declared by the provided Tivoli Storage Manager Server alias argument. You can use the **lstsmnode** command to display the existing Tivoli Storage Manager server alias definition.

The cfghsmnodes command

To configure SONAS HSM, run **cfghsmnodes** to validate the connection to Tivoli Storage Manager and set up HSM parameters. It validates the connection to the provided Tivoli Storage Manager Server and it registers the migration callback.

This script is run as follows:

```
cfghsmnodes <TSMserver_alias> <intNode1,intNode2,...,intNodeN> [ -c <clusterId |
clusterName> ]
```

- **<TSMserver_alias>** is the name of the Tivoli Storage Manager Server set that is up by the backup/archive client,
- **<intNode1,intNode2,...>** is the list of Interface nodes that run HSM to the attached Tivoli Storage Manager Server.
- **<clusterId>** or **<clusterName>** is the cluster identifier.

Figure 6-33 shows the **cfghsmnodes** syntax.

Name: **cfghsmnodes**

Function: Configures nodes to be enabled for HSM.

Syntax: **cfghsmnodes tsmServerAlias nodeName1,nodeName2,...,nodeNameN [-c { clusterID | clusterName }]**

Arguments:

- **tsmServerAlias**

  Specifies the name of the Tivoli Storage Manager server that is registered with the Management node. The name can contain only ASCII alphanumeric, '_', '-', '+', '.' and '&' characters. The maximum length is 64 characters.

- **nodeName1,nodeName2,...,nodeNameN**

  Lists the host names of the Management nodes that should participate in the HSM migration and recall processes, in a comma-separated list. A valid node list argument might be, for example, **mgmt001st001,mgmt002st001**. For a sample output, see the `nodeName` field of the **lstsmnode** command.

Using unlisted arguments can lead to an error.

Figure 6-33 cfghsmnodes command reference

The cfghsmfs command

Configure the file system to be managed by HSM by running **cfghsmfs**.

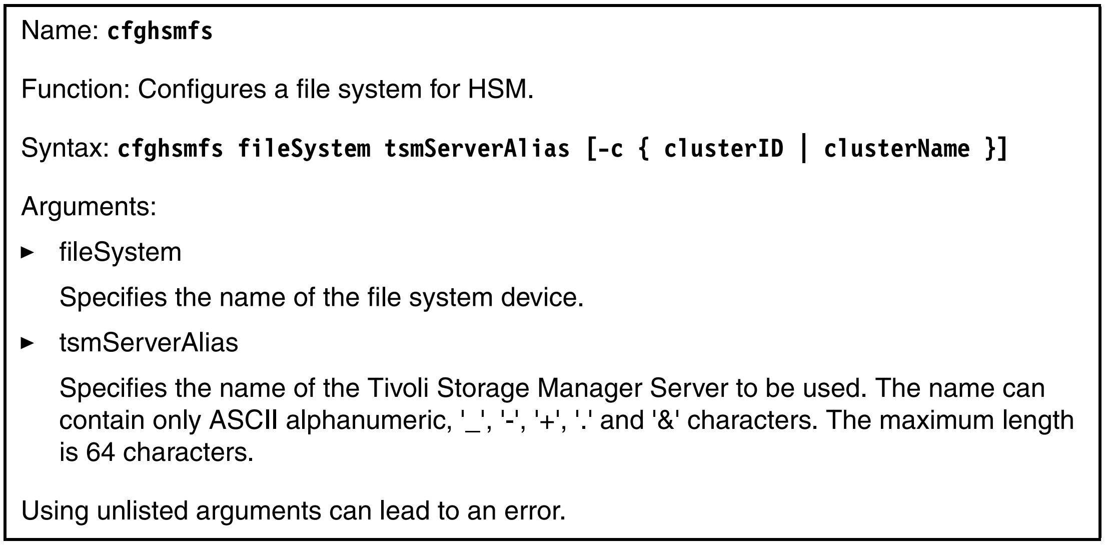
The **cfghsmfs** command configures the specified file system to be HSM-managed by using the specified Tivoli Storage Manager. This command sets the HSM-relevant parameters for CIFS, adds the file system that is HSM-managed to the network, and configures the file system to support HSM operations.

Then, run the **cfghsmfs** command as follows:

```
cfghsmfs <TSMserv> <filesystem> [-P pool] [-T(TIER/PEER)] [-N <ifnodelist>] [-S
stubsize]
```

- **<TSMserv>** is the name of the Tivoli Storage Manager Server that is set up with the **cfgtsmnode** command.
- **<filesystem>** is the name of the SONAS file system to be managed by HSM.
- **<pool>** is the name of the user pool.
- **TIER/PEER** specifies whether the system pool and the specified user pool are set up as TIERed or PEERed.
- **<ifnodelist>** is the list of Interface nodes that interface with the Tivoli Storage Manager Server for this file system and **<stubsize>** is the HSM stub file size in bytes.

Figure 6-34 shows the **cfghsmfs** syntax.



Name: **cfghsmfs**

Function: Configures a file system for HSM.

Syntax: **cfghsmfs fileSystem tsmServerAlias [-c { clusterID | clusterName }]**

Arguments:

- fileSystem

  Specifies the name of the file system device.

- tsmServerAlias

  Specifies the name of the Tivoli Storage Manager Server to be used. The name can contain only ASCII alphanumeric, '_', '-', '+', '.' and '&' characters. The maximum length is 64 characters.

Using unlisted arguments can lead to an error.

Figure 6-34 cfghsmfs command reference

The lstsmnode command

Validate the Tivoli Storage Manager node configuration by running **lstsmnode**. The **lstsmnode** command lists all the defined and reachable Tivoli Storage Manager nodes in the cluster. Unreachable, but configured, nodes are not displayed.

Figure 6-35 shows the **lstsmnode** syntax.

Name: **lstsmnode**

Function: Lists all the defined and reachable Tivoli Storage Manager nodes in the cluster. Unreachable, but configured nodes are not displayed.

Syntax: **lstsmnode [nodeName] [-c { clusterID | clusterName }] [-Y] [--validate]**

Argument:

nodeName specifies the node where the Tivoli Storage Manager server stanza information displays. If this argument is omitted, all the Tivoli Storage Manager server stanza information, for reachable client nodes within the current cluster, is displayed.

Using unlisted arguments can lead to an error.

Examples:

```
lstsmnode
```

Example output on a SONAS system:

```
Node name Virtual node name TSM server name  TSM server address TSM node name
int001st001 sonas_st1        TSMserver       9.155.106.19      int001st001
int002st001 sonas_st1        TSMserver       9.155.106.19      int002st001
int003st001 sonas_st1        TSMserver       9.155.106.19      int003st001
```

Example output on a Storwize V7000 Unified system:

```
Node name Virtual node name TSM server name  TSM server address TSM node name
mgmt001st001 ifs_st1          TSMserver       9.155.106.19      mgmt001st001
mgmt002st001 ifs_st1          TSMserver       9.155.106.19      mgmt002st001
```

Figure 6-35 lstsmnode command reference

6.5.3 HSM diagnostic tests

For debugging purposes, there are three commands that can be used:

- **lshsm** runs HSM diagnostic tests on all client-facing nodes.
- **lshsmlog** shows the HSM error log output (`/var/log/dsmerror.log`).
- **lshsmstatus** shows the HSM status.

The lshsm command

The **lshsm** command runs HSM diagnostic tests on all client-facing nodes or on just one node. If HSM-enabled file systems are discovered, the following values are displayed:

- HSM file system name
- File system state
- Migrated size
- Premigrated size
- Migrated files
- Premigrated files
- Unused i-nodes
- Free size

Figure 6-36 shows the **lshsm** syntax.

Name: **lshsm**

Function: Lists all the HSM-enabled file systems in the cluster.

Syntax: **lshsm [nodeName] [-c { clusterID | clusterName }] [-Y]**

Argument:

nodeName specifies the node where the HSM enabled file systems are checked. If this argument is omitted, all client-facing nodes are checked. If the system is configured and operating correctly, all client-facing nodes should display an identical HSM configuration.

Using unlisted arguments can lead to an error.

Figure 6-36 lshsm command reference

The lshsmlog command

The **lshsmlog** command displays the HSM log entries of the recent HSM errors from the nodes in a human-readable format or as a parsable output. The log files do not contain the success messages, so the command cannot show them.

Figure 6-37 shows the **lshsmlog** syntax.

Name: **lshsmlog**

Function: Lists the HSM log messages.

Syntax: **lshsmlog [-c { clusterID | clusterName }] [--count numOfLines] [-Y]**

Example:

```
lshsmlog --count 2
```

The example displays the last two log entries.

Example result:

```
Date                Node            MSG-ID     Message
08-04-2010 17:28:41  garfield        ANS9020E   Could not establish a session
with a TSM server or client agent.
08-04-2010 17:28:41  garfield        ANS1017E   Session rejected: TCP/IP
connection failure
```

Figure 6-37 lshsmlog command reference

The lshsmstatus command

The **lshsmstatus** command retrieves status information about the HSM-enabled nodes of the managed clusters and returns a list in either a human-readable format or in a format that can be parsed. By specifying either the ID or the name of the cluster, the list includes the HSM status of the nodes that belong to that cluster.

Figure 6-38 shows the **lshsmstatus** syntax.

Name: **lshsmstatus**

Function: Lists the status of the HSM-enabled nodes in the cluster.

Synopsis: **lshsmstatus [-c { clusterID | clusterName }] [-Y] [-v]**

Use the **lshsmstatus** command to verify the Tivoli Storage Manager for Space Management configuration, as in the following example:

```
# lshsmstatus
Output is displayed in the following format:
Managed file system: gpfs0   Mountpoint: (/ibm/gpfs0)
-------------------------------------------------------------------------------
Status OK: All HSM nodes have fs mounted. gpfs0 owned by node: int001st001
***** Show HSM daemon status of HSM configured nodes **********************
nodename        watchd          recalld         failover        fs owned
                                                status
-------------------------------------------------------------------------------
int001st001:   OK (1)         OK (3)            active     /ibm/gpfs0,
int003st001:   OK (1)         OK (3)            active
```

Figure 6-38 lshsmstatus command reference

Deeper knowledge of HSM possibilities and diagnostic tests can be learned through HSM-specific documentation. Before you consider snapshot enhancements in SONAS V1.3 (as described in 6.5.5, “File cloning in SONAS V1.3” on page 419), review the HSM policy structures in 6.5.4, “HSM sample policies” on page 415.

6.5.4 HSM sample policies

The SONAS system provides policy templates for data migration policies to facilitate policy creation and implementation.

Available templates

No template contains a default rule. You must create a default policy and set that policy at the same time that you set your migration policy. In most cases, implement a default rule. By using this method, you are not required to copy a default rule to the policy that implements the migration of your file system pool. For example, to set a default policy that is named *default* at the same time as a Tivoli Storage Manager for Space Management migration policy named *hsmpolicy,* run **setpolicy** with a comma that separates the two policy names in the value for the **-P** option. It is shown in the following example:

```
# setpolicy gpfs0 -P default,hsmpolicy
```

The TEMPLATE-HSM policy template specifies migration of data between a secondary file system pool that is named *silver* and an external file system pool that is named *hsm*, as shown in the following example:

```
# lspolicy -P TEMPLATE-HSM
```

Copy the Sample policy to preserve the Integrity of the initial template. Use the copy of the template and modify the definitions. For example, you can adjust the `stub_size`, `access_age`, `exclude_list`, and `systemtotape` pool (or its thresholds). After the template is modified, it must be enabled before it is activated.

Figure 6-39 shows several sample HSM policy template definition changes.

```
Policy Name  Declaration Name  Default Declarations
TEMPLATE-HSM stub_size         N       define(stub_size,0)
TEMPLATE-HSM is_premigrated    N       define(is_premigrated,(MISC_ATTRIBUTES LIKE '39'
AND KB_ALLOCATED > stub_size))
TEMPLATE-HSM is_migrated       N       define(is_migrated,(MISC_ATTRIBUTES LIKE '39' AND
KB_ALLOCATED == stub_size))
TEMPLATE-HSM access_age        N       define(access_age,(DAYS(CURRENT_TIMESTAMP) -
DAYS(ACCESS_TIME)))
TEMPLATE-HSM mb_allocated      N       define(mb_allocated,(INTEGER(KB_ALLOCATED /
1024)))
TEMPLATE-HSM exclude_list      N       define(exclude_list,(PATH_NAME LIKE
'%/.SpaceMan/%' OR NAME LIKE '%dsmerror.log%'
OR PATH_NAME LIKE '%/.ctdb/%'))
TEMPLATE-HSM weight_expression N       define(weight_expression,(CASE WHEN access_age <
1 THEN 0
WHEN mb_allocated < 1 THEN access_age WHEN is_premigrated   THEN mb_allocated *
access_age * 10 ELSE mb_allocated * access_age  END))
TEMPLATE-HSM hsmexternalpool   N       RULE 'hsmexternalpool' EXTERNAL POOL 'hsm' EXEC
'HSMEXEC'
TEMPLATE-HSM hsmcandidatesList N       RULE 'hsmcandidatesList' EXTERNAL POOL
'candidatesList' EXEC 'HSMLIST'
TEMPLATE-HSM systemtotape      N       RULE 'systemtotape' MIGRATE FROM POOL 'silver'
THRESHOLD(80,70) WEIGHT(weight_expression)
TO POOL 'hsm' WHERE NOT (exclude_list) AND NOT (is_migrated)
```

Figure 6-39 Sample HSM policy

Configuring Tivoli Storage Manager for Space Management

The SONAS system can be configured to use Tivoli Storage Manager for Space Management for migrating data to an external file system pool.

To configure Tivoli Storage Manager for Space Management, the Interface nodes must be configured with the Tivoli Storage Manager servers, as in the following example:

```
upd mgmt standard standard standard spacemgtech=auto migrequiresb=n
assign defmgmt standard standard standard
validate pol standard standard
activate pol standard standard
```

For more information, see 6.2.4, “Configuring Interface nodes and file systems for Tivoli Storage Manager” on page 377. To configure Tivoli Storage Manager for Space Management, complete the following steps:

1. Enable the Tivoli Storage Manager for Space Management client with the server by running **cfghsmnodes**, specifying the Tivoli Storage Manager Server name and the Interface nodes that are to be used for doing Tivoli Storage Manager for Space Management operations, as in the following example:

   ```
   # cfghsmnodes tsmserver int001st001,int002st001
   ```

   > **Important:** At least two Interface nodes must be configured for Tivoli Storage Manager for Space Management.

2. Enable Tivoli Storage Manager for Space Management on the file system by running **cfghsmfs**, specifying the file system name and the Tivoli Storage Manager Server name, as shown in the following example:

   ```
   # cfghsmfs gpfs0 tsmserver
   ```

   > **Tip:** When you configure Tivoli Storage Manager for Space Management for multiple file systems with multiple Tivoli Storage Manager servers, there is one thing you must do. Ensure that all of the Tivoli Storage Manager for Space Management nodes are configured for all of the Tivoli Storage Manager servers for the file systems that are managed by Tivoli Storage Manager for Space Management.

   In the following examples, there are a total of five Interface nodes. Only three of them (int002st001 int003st001, and int004st001) are configured for Tivoli Storage Manager for Space Management on two file systems with two Tivoli Storage Manager servers (tsm1 and tsm2). Both a correct configuration and an incorrect configuration are shown.

   Example 6-7 shows a correct configuration.

   Example 6-7 Correct Tivoli Storage Manager for Space Management configuration

   ```
   # lstsmnode
        Node name  TSM target node name TSM server name TSM server address
   TSM node name
         int002st001 SONAS_ST4            tsm1       tsm1.domain.com
   sonas4-int002st001
         int003st001 SONAS_ST4            tsm1       tsm1.domain.com
   sonas4-int003st001
         int004st001 SONAS_ST4            tsm1       tsm1.domain.com
   sonas4-int004st001
         int002st001 SONAS_ST4            tsm2       tsm2.domain.com
   sonas4-int002st001
         int003st001 SONAS_ST4            tsm2       tsm2.domain.com
   sonas4-int003st001
         int004st001 SONAS_ST4            tsm2       tsm2.domain.com
   sonas4-int004st001
   ```

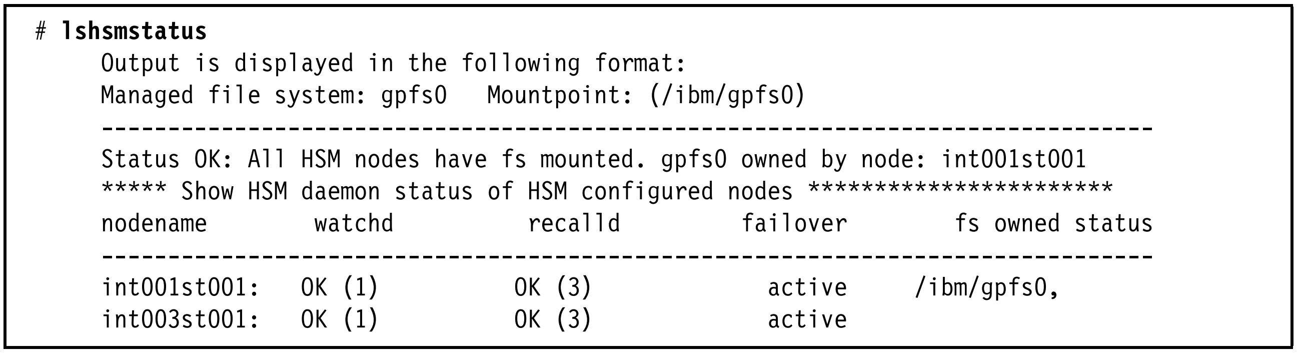
   Example 6-8 shows an incorrect configuration.

   Example 6-8 Incorrect Tivoli Storage Manager for Space Management configuration

   ```
   # lstsmnode
        Node name  TSM target node name TSM server name TSM server address
   TSM node name
         int001st001 SONAS_ST4            tsm1       tsm1.domain.com
   sonas4-int001st001<- This node is not managed by HSM and therefore it is not
   required to be configured, but this optional configuration is acceptable for
   using the node as a Tivoli Storage Manager backup node.
         int002st001 SONAS_ST4            tsm1       tsm1.domain.com
   sonas4-int002st001
         int003st001 SONAS_ST4            tsm1       tsm1.domain.com
   sonas4-int003st001
         int002st001 SONAS_ST4            tsm2       tsm2.domain.com
   sonas4-int002st001
         int003st001 SONAS_ST4            tsm2       tsm2.domain.com
   sonas4-int003st001
         int004st001 SONAS_ST4            tsm2       tsm2.domain.com
   sonas4-int004st001
   ```

```
# <- The configuration of node int004st001 for TSM server tsm1 is missing and
#should be configured.
```

3. Run **lshsmstatus** to verify the Tivoli Storage Manager for Space Management configuration, as shown in Figure 6-40.



```
# lshsmstatus
     Output is displayed in the following format:
     Managed file system: gpfs0   Mountpoint: (/ibm/gpfs0)
     -------------------------------------------------------------------------------
     Status OK: All HSM nodes have fs mounted. gpfs0 owned by node: int001st001
     ***** Show HSM daemon status of HSM configured nodes ***********************
     nodename        watchd          recalld         failover        fs owned status
     -------------------------------------------------------------------------------
     int001st001:   OK (1)          OK (3)             active     /ibm/gpfs0,
     int003st001:   OK (1)          OK (3)             active
```

Figure 6-40 Running lshsmstatus to verify the Tivoli Storage Manager for Space Management configuration

4. Run **mkpolicy** to create a default policy with a rule that states that the default pool is the system pool, as shown in the following example:

   ```
   # mkpolicy default -R "RULE 'default' set pool 'system'" -D
   ```

   > **Tip:** This default policy is used when you are applying policies to implement Tivoli Storage Manager for Space Management configurations in this scenario.

   Creating a policy with a default rule avoids the requirement to add a default rule to the policy template that is copied in step 5. If you skip step 5, you must add a default rule to your copied policy.

5. Copy the Tivoli Storage Manager for Space Management template policy to a new policy by running **mkpolicy**, specifying a name for the new policy and copying the TEMPLATE-HSM policy. For example, to create a policy that is named hsmpolicy, run the following command:

   ```
   # mkpolicy hsmpolicy -CP TEMPLATE-HSM
   ```

   Another option is to provide the **-R** option and define the rules instead of copying a template. For more information, see the **chpolicy** command.

6. Adapt the policy to reflect your environment by removing non-relevant rules and adding rules that can accomplish the migration. The following example configures a system to migrate from the system pool to the external hsm pool:

   ```
      # chpolicy hsmpolicy --remove systemtotape
         # chpolicy hsmpolicy --add "RULE 'systemtotape' MIGRATE FROM POOL
   'system' THRESHOLD(80,70)
         WEIGHT(weight_expression) TO POOL 'hsm' WHERE NOT (exclude_list) AND NOT
   (is_migrated)"
   ```

   For more information about policies and rules, see *GPFS Advanced Administration Guide Version 3 Release 3*, SC23-5182.

7. To avoid migration of SONAS system files during ILM migration, change the HSM/ILM policy templates that are contained in your DB for the exclude list to add items, as in the following example:

   ```
   define(exclude_list,(PATH_NAME LIKE '%/.SpaceMan/%' OR NAME LIKE
   '%dsmerror.log%' OR PATH_NAME LIKE '%/.ctdb/%' OR PATH_NAME LIKE '%/.sonas/%'
   OR PATH_NAME LIKE '%/.mmbackupCfg/%' OR NAME LIKE '%.quota')).
   ```

   Make similar changes to policies that are applied to file systems.

8. The preferred practice is to validate the policy by running **chkpolicy** before you run or apply the policy to ensure that the policy performs as intended. For example, to validate the policies that are named *default* and *hsmpolicy* against the gpfs0 file system, run the following command:

   ```
   # chkpolicy gpfs0 -P default,hsmpolicy -T
   ```

9. Set the active policy for the file system by running **setpolicy**. For example, to set the default policy and *hsmpolicy* as the policy for the file system gpfs0, run the following command:

   ```
   # setpolicy gpfs0 -P default,hsmpolicy
   ```

   The default and modified template policies are set for the file system, which creates a system-to-external pool Tivoli Storage Manager for Space Management configuration.

Reconciling Tivoli Storage Manager and Tivoli Storage Manager for Space Management files

Because the Tivoli Storage Manager implementation of Tivoli Storage Manager for Space Management is a client and server design, updates at the client are not immediately synchronized with the server. A reconcile operation is required to do this synchronization. SONAS uses an accelerated process that includes a two-way orphan check that enhances reconcile performance.

6.5.5 File cloning in SONAS V1.3

File cloning is typically used to clone a file to use it in application or development instances while it preserves the integrity of the original file. In this regard, it can be considered another form of data protection.

Description

The **mkclone** command creates a clone from a source file. The source file can become and serve as the immutable parent (*not* susceptible to change), or a parent file can be specified as the immutable parent. If the parent file has no child files, it can be deleted by running **rm nameFile**.

> **Tip:** If a clone file is moved out of the file system of its parent, the parent inode information is lost.

Examples

Consider the following examples:

- **mkclone -s someFile -t someFileClone -p someFileParent**

  someFileParent is created and made immutable. The file data of someFile and someFileClone are stored in someFileParent. someFile and someFileClone can be modified. A list of file clones against someFile lists someFileClone.

- **mkclone -s someFile -t someFileClone**

  `someFile` is made the file parent and is made immutable. The file data of `someFile` and `someFileClone` are stored in `someFile`. `SomeFile` cannot be modified, but `someFileClone` can be modified.

  Figure 6-41 shows the syntax of the **mkclone** command.

> Name: **mkclone**
>
> Function: Creates a clone.
>
> Syntax: **mkclone -s, --source filePath -t, --target filePath [-p, --parent filePath][-c, --cluster { clusterID | clusterName }]**
>
> You must specify the absolute path of each file.
>
> Arguments:
>
> - **-c, --cluster { clusterID | clusterName }**: Selects the cluster for the operation. Use either the clusterID or the clusterName to identify the cluster. Optional. If this option is omitted, the default cluster, as defined with the **setcluster** command, is used.
> - **-s, --source filePath**: Identifies the source file. If the source and parent files are the same, the source file becomes the immutable parent.
> - **-t, --target filePath**: Identifies the target file, which is the newly created clone file.
> - **-p, --parent filePath**: Identifies the parent file. If the parent file is omitted, the source file becomes the immutable parent file. Optional.
>
> Using unlisted options can lead to an error.

Figure 6-41 mkclone - for creating file clones

Description of the lsclone command

The **lsclone** command retrieves information that is found on clone files or parent files, which includes depth and the parent inode. Because a clone file can also be a parent file, multiple levels of clone files can exist. You can view the number of levels in the Depth column. If a file does not have a parent file, the depth value is 0; the clones for this file have a depth value of 1, a clone of a clone is then 2, and so on. The Parent field displays whether a file is a parent file, and the Parent inode column displays the inode of the parent file. As a result, a clone file that has a depth value less than the value in the Parent inode field does not have “Yes” displayed in the Parent column.

> **Tip:** If a clone file is moved out of the file system of its parent file, the parent inode information is lost.

Clone file example

This example displays information about a clone file. The following column headers display: Parent, Depth, Parent inode, and File name. The Parent inode contains the inode number of the parent.

```
lsclone /ibm/gpfs0/testFile
Parent Depth Parent inode File name
yes    0                 /ibm/gpfs0/testFile
EFSSG1000I The command completed successfully.
```

```
lsclone /ibm/gpfs0/someFile
Parent Depth Parent inode File name
no     1     9479         /ibm/gpfs0/someFile
EFSSG1000I The command completed successfully.
```

Figure 6-42 shows the **lsclone** command syntax.

Name: **lsclone**

Function: Lists information about a clone file.

Syntax: **lsclone cloneFile [-Y] [-c { clusterID | clusterName }]**

Arguments:

- **cloneFile**: Specifies the clone file. Use the absolute path to specify the `cloneFile`.

  Using unlisted arguments can lead to an error.
- **-c, --cluster { clusterID | clusterName }**: Selects the cluster for the operation. Use either the clusterID or the clusterName to identify the cluster. Optional. If this option is omitted, the default cluster, as defined with the **setcluster** command, is used.
- **-Y**: Creates parsable output. Optional.

Using unlisted options can lead to an error.

Figure 6-42 lsclone syntax example

6.6 Snapshots

Snapshots are a simple and powerful tool that can be used as a way to protect files and file data from accidental deletion or client-driven corruption.

A snapshot of an entire file system or of an independent file set can be created to preserve the contents of the file system or the independent file set at a single point in time. The storage that is needed for maintaining a snapshot is because of the required retention of a copy of all of the data blocks that were changed or deleted after the time of the snapshot, and is charged against the file set quota.

Snapshots are read-only; changes can be made only to the normal, active files and directories, not to the snapshot.

The snapshot function allows a backup or mirror program to run concurrently with user updates and still obtain a consistent copy of the file system or file set as of the time that the snapshot was created. Snapshots also provide an online backup capability that allows easy recovery from common problems, such as accidental deletion of a file, and comparison with older versions of a file.

Snapshots are managed by an automated background process that self-initiates once a minute. The snapshot management service creates and deletes snapshots based on the system time at process initiation and the attributes of snapshots rules that are created and then associated with file systems, or file sets, or both, by the system administrator. There are two steps to configure the snapshot management for a file system or file set. First, create the rule or rules and then associate the rule or rules with the file system or file set. A user must have the Snapshot Administrator role to perform snapshot management functions.

6.6.1 Snapshot rules

A snapshot rule indicates the frequency and timing of the creation of snapshots, and also indicates the retention of the snapshots that are created by the rule. The retention attributes indicate how many snapshots are retained for the current day and for the previous days, weeks, and months. One snapshot can be retained for each previous day, week, or month that is identified, and is the last snapshot that is taken in that day, week, or month.

The **mksnaprule** command creates a snapshot rule and the **lssnaprule** command displays snapshot rules. The **rmsnaprule** command removes a snapshot rule, and the **chsnaprule** command is used to change the attributes of an existing snapshot rule. Running the **chsnaprule** command preserves and manages existing snapshots that are associated with the rule that is changed; this is not the case if the rule is unassociated from the file system or file set, the rule is deleted, and a rule is created and associated with the file system or file set.

The file system or file set in a snapshot rule association defines the scope of files that are managed by the rule in that association. One or more snapshot rules can be associated with a file system or file set. This association might be necessary if the snapshot creation timing and frequency that you want to configure is not consistent over time. For example, multiple rules are required if you want to create a snapshot hourly Monday through Friday, but only twice daily on Saturday and Sunday. A single rule can be associated with multiple file systems or file sets, but each snapshot rule association has only one rule and only one file system or file set. Snapshot rule associations can be set, changed, or removed at any time after the snapshot rule and the file system or file set are created.

Every snapshot instance that is created by a snapshot rule association has an entry in a database to allow management by the snapshot management service. The Snapshot Administrator can also create a snapshot that specifies a file system or an independent file set and an associated rule that has a frequency of onDemand. Because the onDemand snapshot instance is created by a rule association, it has an entry in the database and is managed by the snapshot management service. The service does not create a scheduled snapshot in this case, but does delete the snapshot instances based on the associated rule's retention option settings. The snapshot management service evaluates onDemand snapshot instances every 15 minutes for deletion eligibility.

The Snapshot Administrator can also create a snapshot manually, in which case the snapshot is not associated with or managed by any snapshot rule, and must be managed manually. A manually created snapshot is retained until it is manually deleted. If the manual snapshot is associated with a rule, it is treated as the most recent snapshot in the set of snapshots that are maintained by the rule, and most likely, results in the deletion of an older snapshot.

The **lssnapops** command displays all queued and running snapshot operations in chronological sequence, including invocations of the background process and creation and deletion of snapshot instances. The system administrator must manage the snapshot rules and their associations with file systems and file sets to ensure optimal performance. If operations for a particular rule are still in progress when the current instance is initiated, a warning is logged. The automated background process is queued if the previously initiated process is not completed. It is serialized with other GPFS create and delete operations and also with other instances of running **setsnapnotify** command to generate a warning if the number of operations threshold is exceeded.

Tip: This snapshot management does not replace the ability to create snapshots with scheduled tasks that are created with the **mktask** command by using the MkSnapshotCron template; these snapshots must be managed manually because they are not created by the snapshot management service.

The **mksnapassoc** command creates an association between a snapshot rule and a file system or a file set, and the **rmsnapassoc** command removes an association between a snapshot rule and a file system or a file set. When an association is removed, the previously associated snapshots become unmanaged and can be deleted or must be managed manually.

The **mksnapshot**, **lssnapshot**, and **rmsnapshot** commands create, list, and remove snapshots. You can use the **-j** option with each of these commands to specify a particular file set.

Considerations:

- A snapshot of a file creates a file that captures the user data and user attributes from the original file. The snapshot file is independent from the original file and is space-efficient; only modified blocks are written to the snapshot, and reads to unmodified data are directed to the original file. Because snapshots are not copies of the entire file system or file set, they should not be used as protection against media failures.
- File systems or file sets that are managed by external file system-pool-support software, such as Tivoli Storage Manager, must have DMAPI enabled. For files in file systems or file sets where DMAPI is enabled, the snapshot of a file is not automatically managed by DMAPI, regardless of the state of the original file. The DMAPI attributes from the original file are not inherited by the snapshot.

6.6.2 Listing the snapshot notification option configuration

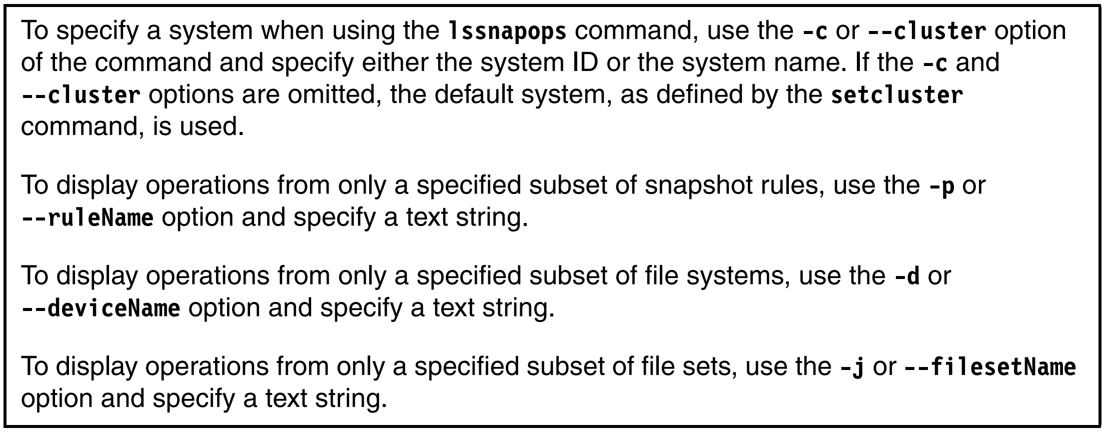
This section explains how to list the snapshot notification option configuration.

Overview

You can use the CLI or the GUI to view a list of all the snapshot events that can be configured to send notifications.

The lssnapops command

Figure 6-43 provides an overview and the syntax of the **lssnapops** command.



To specify a system when using the **lssnapops** command, use the **-c** or **--cluster** option of the command and specify either the system ID or the system name. If the **-c** and **--cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

To display operations from only a specified subset of snapshot rules, use the **-p** or **--ruleName** option and specify a text string.

To display operations from only a specified subset of file systems, use the **-d** or **--deviceName** option and specify a text string.

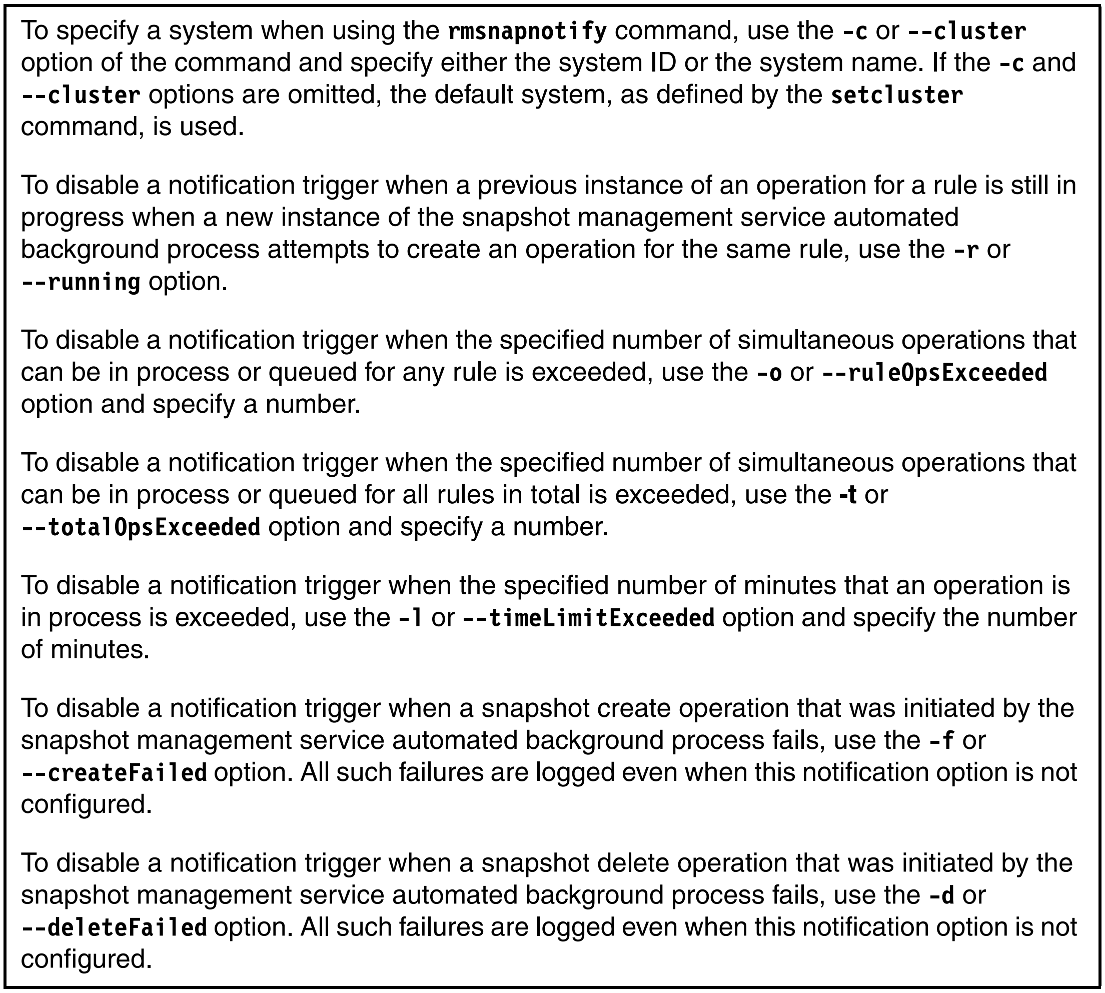
To display operations from only a specified subset of file sets, use the **-j** or **--filesetName** option and specify a text string.

Figure 6-43 lssnapops command reference

6.6.3 Disabling snapshot notification options

The **`rmsnapnotify`** command disables notification options for the snapshot management service. Notifications can be used to monitor conditions that might indicate automated snapshot management performance symptoms. If no notifications are configured, the default is that no notification is sent for any condition, event, or threshold.

Figure 6-44 shows the syntax of the **`rmsnapnotify`** command.



To specify a system when using the **`rmsnapnotify`** command, use the **`-c`** or **`--cluster`** option of the command and specify either the system ID or the system name. If the **`-c`** and **`--cluster`** options are omitted, the default system, as defined by the **`setcluster`** command, is used.

To disable a notification trigger when a previous instance of an operation for a rule is still in progress when a new instance of the snapshot management service automated background process attempts to create an operation for the same rule, use the **`-r`** or **`--running`** option.

To disable a notification trigger when the specified number of simultaneous operations that can be in process or queued for any rule is exceeded, use the **`-o`** or **`--ruleOpsExceeded`** option and specify a number.

To disable a notification trigger when the specified number of simultaneous operations that can be in process or queued for all rules in total is exceeded, use the **`-t`** or **`--totalOpsExceeded`** option and specify a number.

To disable a notification trigger when the specified number of minutes that an operation is in process is exceeded, use the **`-l`** or **`--timeLimitExceeded`** option and specify the number of minutes.

To disable a notification trigger when a snapshot create operation that was initiated by the snapshot management service automated background process fails, use the **`-f`** or **`--createFailed`** option. All such failures are logged even when this notification option is not configured.

To disable a notification trigger when a snapshot delete operation that was initiated by the snapshot management service automated background process fails, use the **`-d`** or **`--deleteFailed`** option. All such failures are logged even when this notification option is not configured.

Figure 6-44 rmsnapnotify command reference

6.6.4 Creating snapshot rules

This section explains how to create snapshot rules.

Overview

A snapshot rule defines the snapshot creation frequency and timing, and snapshot retention, and can be associated with file systems and files sets. The maximum number of snapshots that can exist at any one time for a file system or a file set is 224.

mksnaprule command

The **mksnaprule** command creates a snapshot rule with a specified alphanumeric unique name that has a maximum of 256 characters. The command fails if an attempt is made to create a rule with a name that exists. A warning is displayed if the attributes of the new rule are identical to an existing rule with a different name, and the option to proceed is prompted. A snapshot rule defines the snapshot creation frequency and timing, and snapshot retention, and can be associated with file systems and files sets by running **mksnapassoc**. If no **mksnaprule** options are specified, the default rule options create a single snapshot daily at 00:00:00, where only the snapshot from the previous day is retained. Figure 6-45 provides details and syntax of the **mksnaprule** command.

The **mksnaprule** command fails if the rule retention options that are specified would result in retaining more than the 224 maximum number of snapshots per file system or file set. Likewise, the **mksnapassoc** command fails if the newly associated rule, when added to all of the rules that are currently associated with the file set, would increase the total number of retained snapshots past the 224 limit.

You must specify the snapshot rule name, which must be unique within a system. You can use the **--snap-prefix** option to set the name prefix for snapshots that are created by the newly created rule. The default is no prefix, in which case the snapshot file name is the GMT date and time at the time that the snapshot is created. If the default is not used, the snapshots are not visible from Windows clients.

The **mksnaprule** command has many options to allow as much flexibility as possible, but this flexibility also makes it possible to specify option value combinations that create a snapshot rule that is self-contradictory or that makes no sense. The command fails such attempts with information about the conflicting options that were specified.

The following options can be used to define the retention policy, which defines how many snapshots are kept for a period:

- **--maxHoursOrMinutes**
- **--maxDays**
- **--maxWeeks**
- **--maxMonths**

None, one, or several of these options can be used. If none are used, the default is one per period, as specified by using the **-q** or **--frequency** options. Which retention options are considered valid depends on the frequency and time options that are chosen. For example, a maximum for a day does not make sense if the frequency is monthly; the command failure output displays details.

Figure 6-45 mksnaprule command reference

6.6.5 Changing snapshot rules

You can change existing snapshot rules without changing any associations of the rule with a file system or file set. Using the **`chsnaprule`** command preserves and manages existing snapshots that are associated with the rule that is changed; this is not the case if the rule is unassociated from the file system or file set, the rule is deleted, and a rule is created and associated with the file system or file set. Figure 6-46 shows the syntax of the **`chsnaprule`** command.

Name: **`chsnaprule`**

Function: Changes a snapshot rule.

Syntax: **`chsnaprule ruleName {-k | -r} [--snap-prefix snapshotNamePrefix] [-q frequency] [-x everyXMinutes] [-a minutesAfterHour,...minutesAfterHour] [-h hour,...,hour] [-d dayOfWeek,...,dayOfWeek] [-w week,...,week] [-n dayOfMonth,...dayOfMonth] [-m month,...,month] [-e HH{:MM:SS},HH{:MM:SS}] [--maxHoursOrMinutes maxHoursOrMinutes] [--maxDays maxDays] [--maxWeeks maxWeeks] [--maxMonths maxMonths] [-f] [-c {clusterID | clusterName}]`**

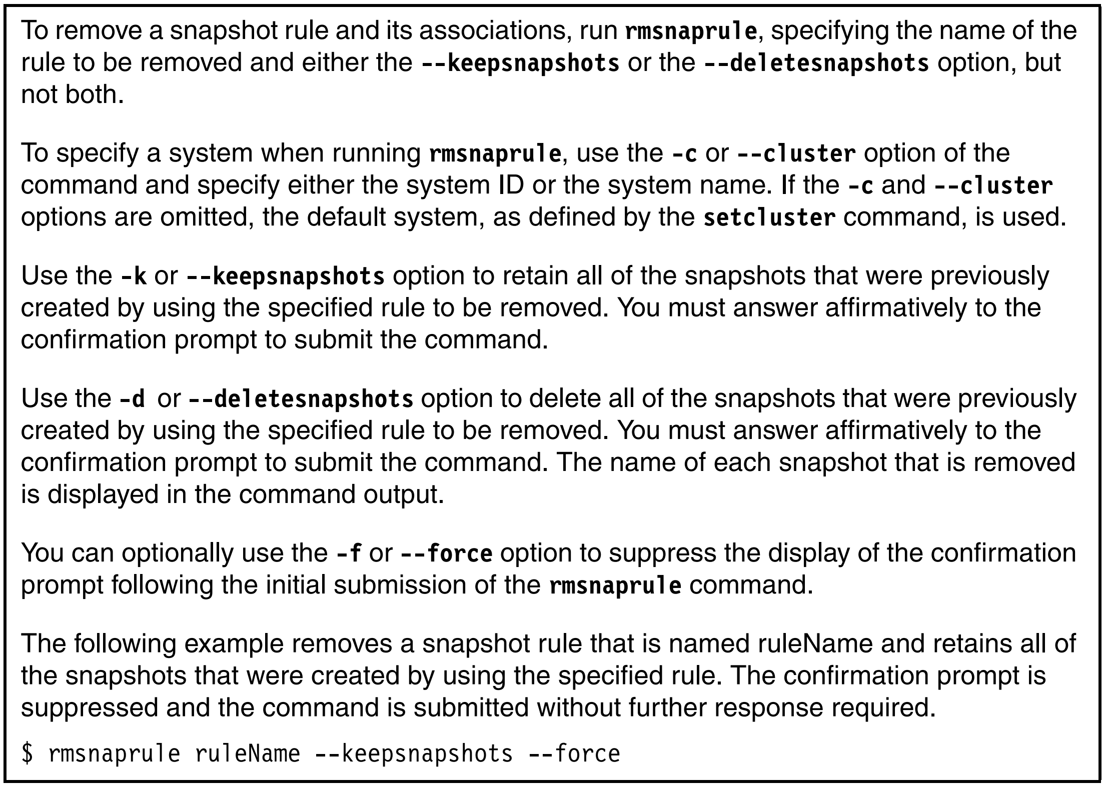
Argument: The **`ruleName`** option identifies the snapshot rule. Snapshot rule names must be unique within a cluster.

Using unlisted arguments can lead to an error.

Figure 6-46 chsnaprule command reference

6.6.6 Removing snapshot rules

The **rmsnaprule** command deletes an existing snapshot rule and all of its associations with file systems and file sets. You must specify either the **--keepsnapshots** or the **--deletesnapshots** option, but not both. These options retain or remove existing snapshots that were created by using the removed rule. If these snapshots are retained, the administrator is responsible for manually managing the retained snapshots. Figure 6-47 shows the **rmsnaprule** command and syntax.



To remove a snapshot rule and its associations, run **rmsnaprule**, specifying the name of the rule to be removed and either the **--keepsnapshots** or the **--deletesnapshots** option, but not both.

To specify a system when running **rmsnaprule**, use the **-c** or **--cluster** option of the command and specify either the system ID or the system name. If the **-c** and **--cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

Use the **-k** or **--keepsnapshots** option to retain all of the snapshots that were previously created by using the specified rule to be removed. You must answer affirmatively to the confirmation prompt to submit the command.

Use the **-d** or **--deletesnapshots** option to delete all of the snapshots that were previously created by using the specified rule to be removed. You must answer affirmatively to the confirmation prompt to submit the command. The name of each snapshot that is removed is displayed in the command output.

You can optionally use the **-f** or **--force** option to suppress the display of the confirmation prompt following the initial submission of the **rmsnaprule** command.

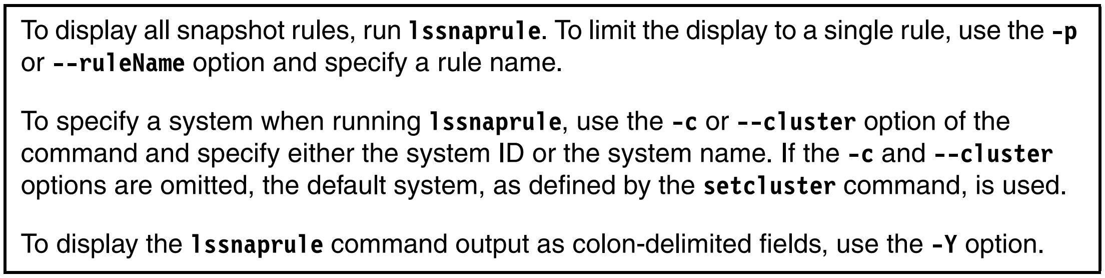
The following example removes a snapshot rule that is named ruleName and retains all of the snapshots that were created by using the specified rule. The confirmation prompt is suppressed and the command is submitted without further response required.

```
$ rmsnaprule ruleName --keepsnapshots --force
```

Figure 6-47 rmsnaprule command reference

6.6.7 Displaying snapshot rules

The **lssnaprule** command displays snapshot rules. If no operations match the parameter value, an error is displayed. Figure 6-48 shows the syntax of the **lssnaprule** command.



To display all snapshot rules, run **lssnaprule**. To limit the display to a single rule, use the **-p** or **--ruleName** option and specify a rule name.

To specify a system when running **lssnaprule**, use the **-c** or **--cluster** option of the command and specify either the system ID or the system name. If the **-c** and **--cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

To display the **lssnaprule** command output as colon-delimited fields, use the **-Y** option.

Figure 6-48 lssnaprule command reference

6.6.8 Snapshot considerations

Because snapshots are not copies of the entire file system, they must not be used as protection against physical media failure; snapshots are appropriate and effective for protection against logical failures and data corruption that is caused by the application. SONAS uses a redirect-on-write, GPFS based snapshot technology. It does not use the hardware snapshot integration tools from the underlying storage. It secures the metadata pointers of the data in its current state and as changes are written for the writes to new locations.

Because the technology uses redirect-on-write, snapshot invocation can be almost instantaneous, and it uses no additional capacity until the data in a snapshot set changes. It is important to keep a grasp of data change rates as you increase the number of active snapshots on any environment to manage capacity growth effectively.

A snapshot file is independent from the original file because it contains only the user data and user attributes of the original file. For DMAPI-managed file systems, the snapshot is not DMAPI-managed regardless of the DMAPI attributes of the original file because the DMAPI attributes are not inherited by the snapshot.

For example, consider a base file that is a stub file because the file contents were migrated by Tivoli Storage Manager HSM to offline media; the snapshot copy of the file is not managed by DMAPI because it has not inherited any DMAPI attributes. Therefore, referencing a snapshot copy of a Tivoli Storage Manager HSM managed file does not cause Tivoli Storage Manager to initiate a file recall.

6.6.9 VSS snapshot integration

This section explains how to do VSS snapshot integration.

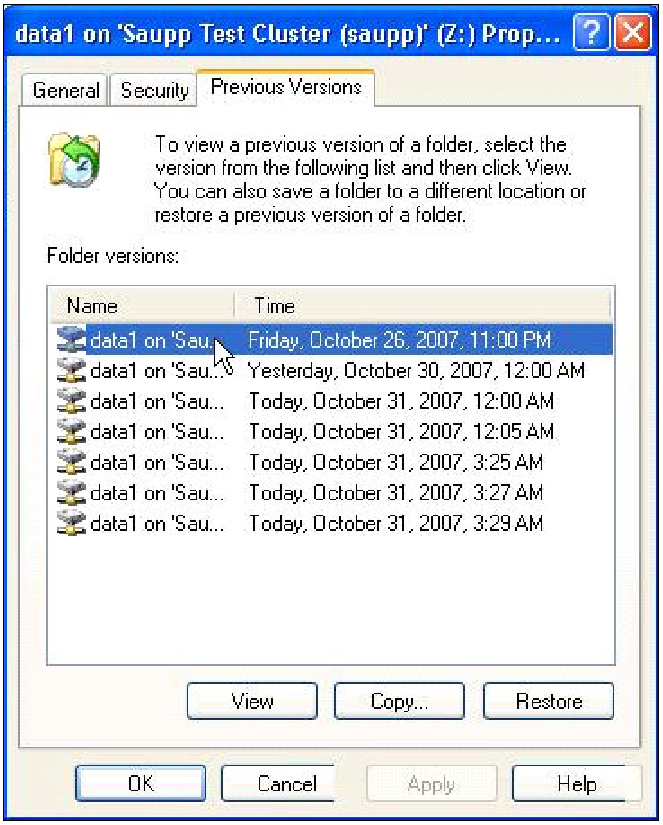
Overview

You must follow a naming convention if you want to integrate snapshots into a Microsoft Windows environment.

Microsoft Windows offers a feature that is called Volume Shadow Copy Service (VSS). SONAS integrates into VSS seamlessly, but only snapshots that use a name in the format `@GMT-yyyy.MM.dd-HH.mm.ss` are visible in the “Previous version” window of Windows Explorer. Snapshots that are created by using the CLI automatically adhere to this naming convention.

Snapshot name format

The example in Figure 6-49 shows the correct name format for a snapshot (`@GMT-2008.08.05-23.30.00`) that can be viewed on Microsoft Windows under “Previous version”.



To view a previous version of a folder, select the version from the following list and then click View. You can also save a folder to a different location or restore a previous version of a folder.

Folder versions:

Name	Time
data1 on 'Sau...	Friday, October 26, 2007, 11:00 PM
data1 on 'Sau...	Yesterday, October 30, 2007, 12:00 AM
data1 on 'Sau...	Today, October 31, 2007, 12:00 AM
data1 on 'Sau...	Today, October 31, 2007, 12:05 AM
data1 on 'Sau...	Today, October 31, 2007, 3:25 AM
data1 on 'Sau...	Today, October 31, 2007, 3:27 AM
data1 on 'Sau...	Today, October 31, 2007, 3:29 AM

Figure 6-49 Example Windows Explorer folder previous versions tab

6.6.10 Snapshot creation and management

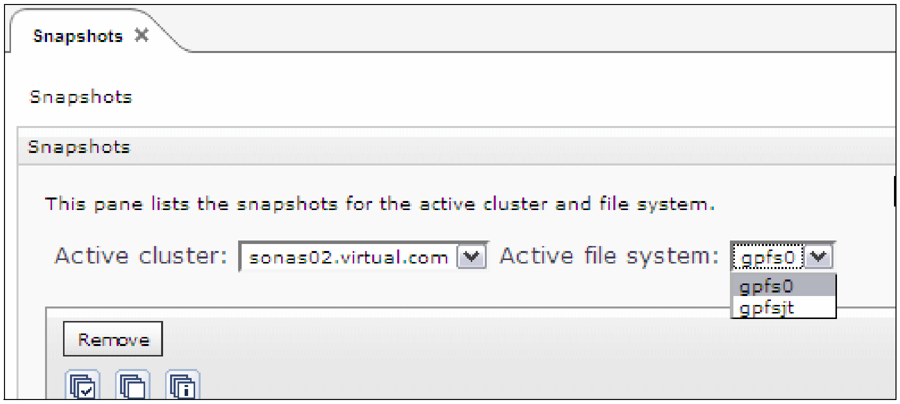
This section shows how to create and manage SONAS snapshots by using both the CLI and the GUI. SONAS snapshot commands create a snapshot of the entire file system at a specific point in time. Snapshots appear in a hidden subdirectory of the root directory called `.snapshots`.

It also shows you how to create and manage snapshot rules and retention.

Creating snapshots from the GUI

To create a snapshot of a sample file system called `gpfsjt` through the SONAS GUI, complete the following steps:

1. Log in to the SONAS management GUI.
2. Click **Files** → **Snapshots**.
3. Select the active cluster and the file system of which you want to take a snapshot, as shown in Figure 6-50.



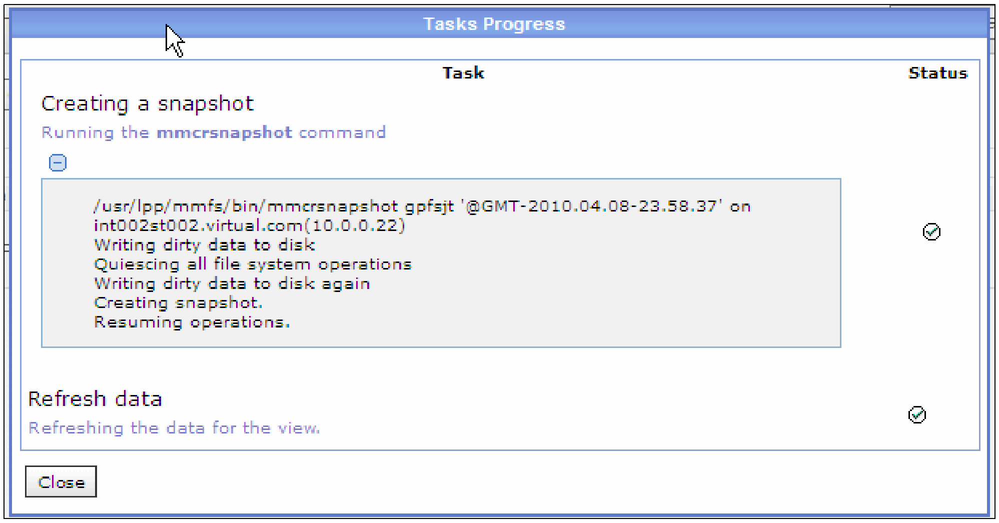
This pane lists the snapshots for the active cluster and file system.

Active cluster: sonas02.virtual.com

Active file system: gpfs0

Figure 6-50 Select cluster and file system for snapshot

4. Click **Create new snapshot**.
5. You are prompted for a name for the new snapshot; accept the default name if you want the snapshot to be integrated with Windows VSS previous versions and click **OK** to proceed.
6. You see a task progress indicator window, as shown in Figure 6-51. You can monitor task progression by using this window.



Creating a snapshot

Running the **mmcrsnapshot** command

```
/usr/lpp/mmfs/bin/mmcrsnapshot gpfsjt '@GMT-2010.04.08-23.58.37' on
int002st002.virtual.com(10.0.0.22)
Writing dirty data to disk
Quiescing all file system operations
Writing dirty data to disk again
Creating snapshot.
Resuming operations.
```

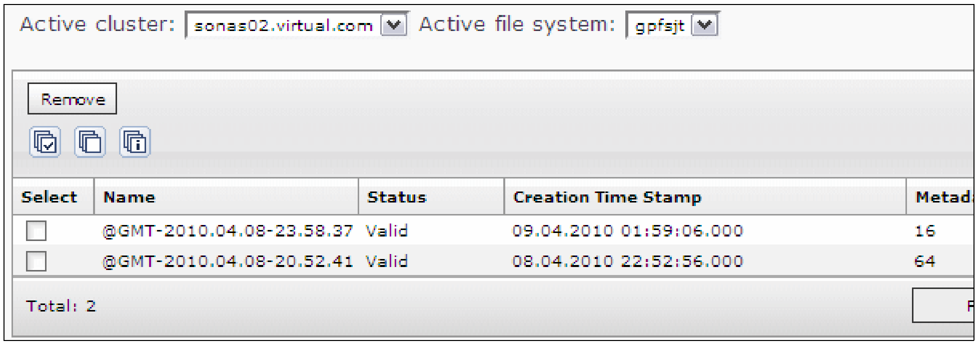
Refresh data

Refreshing the data for the view.

Figure 6-51 Snapshot task progress indicator

7. You can close the task progress window by clicking **Close**.

8. You are now presented with the list of available snapshots, as shown in Figure 6-52.



Active cluster: sonas02.virtual.com Active file system: gpfsjt

Select	Name	Status	Creation Time Stamp	Metad
□	@GMT-2010.04.08-23.58.37	Valid	09.04.2010 01:59:06.000	16
□	@GMT-2010.04.08-20.52.41	Valid	08.04.2010 22:52:56.000	64

Total: 2

Figure 6-52 List of completed snapshots

Creating and listing snapshots from the CLI

You can create snapshots from the SONAS CLI by running **mksnapshot**, as shown in Figure 6-53.

```
[SONAS]$ mksnapshot gpfsjt
EFSSG0019I The snapshot  @GMT-2010.04.09-00.32.43 has been successfully created.
```

Figure 6-53 Create a snapshot

To list all snapshots from all file systems, you can run **lssnapshot**, as shown in Figure 6-54. The command retrieves data about the snapshots of a managed cluster from the database and returns a list of snapshots.

```
[SONAS]$ lssnapshot
Cluster ID   Device name    Path    Status    Creation    Used (metadata)   Used (data)   ID   Timestamp
72..77 gpfsjt @GMT-2010.04.09-00.32.43 Valid  09.04.2010 02:32:43.000    16     0      5     20100409023246
72..77 gpfsjt @GMT-2010.04.08-23.58.37 Valid  09.04.2010 01:59:06.000    16     0      4     20100409023246
72..77 gpfsjt @GMT-2010.04.08-20.52.41 Valid  08.04.2010 22:52:56.000    64     1      1     20100409023246
```

Figure 6-54 List all snapshots for all file systems

The ID Timestamp field is the same for all snapshots, and it indicates the time stamp of the last SONAS database refresh. The **lssnapshots** command with the **-r** option forces a refresh of the snapshots data in the SONAS database by scanning all cluster snapshots before it retrieves the data for the list from the database.

Removing snapshots

Snapshots can be removed by running **rmsnapshot** or from the GUI. For example, to remove a snapshot for file system `gpfsjt` by using the CLI, proceed as shown in Figure 6-55 by completing the following steps:

1. Run **lssnapshot** for file system `gpfsjt`.
2. Choose a snapshot to remove by choosing that snapshot's name, for example, `@GMT-2010.04.08-23.58.37`.
3. Run **rmsnapshot** with the name of the file system and the name of the snapshot.
4. To verify whether the snapshot was removed, run **lssnapshot** again and check that the removed snapshot is no longer present.

```
[SONAS]$ lssnapshot -d gpfsjt
ClusID Devname Path                     Status Creation           Used (metadata) Used (data) ...
72..77 gpfsjt  @GMT-2010.04.09-00.32.43 Valid  09.04.2010 02:32:43.000 16         0           ...
72..77 gpfsjt  @GMT-2010.04.08-23.58.37 Valid  09.04.2010 01:59:06.000 16         0           ...
72..77 gpfsjt  @GMT-2010.04.08-20.52.41 Valid  08.04.2010 22:52:56.000 64         1           ...

[SONAS]$ rmsnapshot gpfsjt      @GMT-2010.04.08-23.58.37

[SONAS]$ lssnapshot -d gpfsjt
ClusID DevName Path                     Status Creation           Used (metadata) Used (data) ...
72..77 gpfsjt  @GMT-2010.04.09-00.32.43 Valid  09.04.2010 02:32:43.000 16         0           ...
72..77 gpfsjt  @GMT-2010.04.08-20.52.41 Valid  08.04.2010 22:52:56.000 64         1           ...
```

Figure 6-55 Remove snapshots

Scheduling snapshots at regular intervals

To automate the task of creating snapshots at regular intervals, you can create a repeating SONAS task based on the snapshot task template called **MkSnapshotCron**. For example, to schedule a snapshot every five minutes on file system `gpfsjt`, run the command that is shown in Figure 6-56.

```
[SONAS]$ mktask MkSnapshotCron --parameter "sonas02.virtual.com gpfsjt" --minute */5
EFSSG0019I The task MkSnapshotCron has been successfully created.
```

Figure 6-56 Create a task to schedule snapshots

To create scheduled **cron** tasks, you must run **mktask**; it is not possible to create **cron** tasks from the GUI. To list the snapshot task that you created, run **lstask**, as shown in Figure 6-57.

```
[[SONAS]$ lstask -t cron
Name           Description                                Status Last run Runs on   Schedule
MkSnapshotCron   This is a cronjob for scheduled snapshots. NONE N/A Mgmt node Runs at every 5th minute.
```

Figure 6-57 List scheduled tasks

To verify that snapshots are being correctly performed, run **lssnapshot**, as shown in Figure 6-58 on page 433.

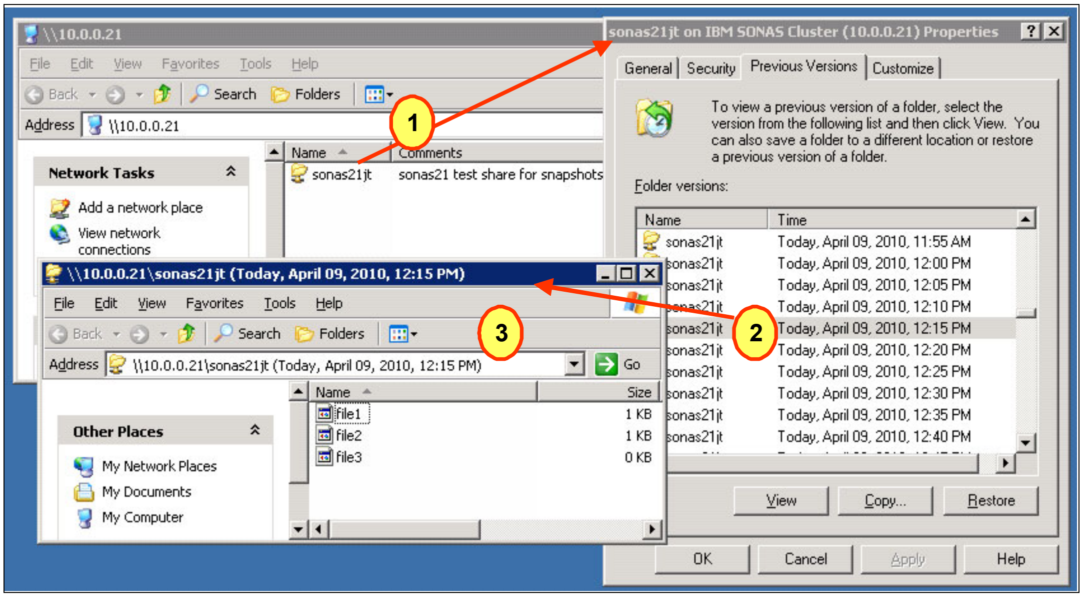
```
[SONAS]$ lssnapshot
Cluster ID     Device name Path                 Status Creation          Used (metadata) Used (data) ID
72..77 gpfsjt  @GMT-2010.04.09-03.15.06 Valid 09.04.2010 05:15:08.000  16              0           9
72..77 gpfsjt  @GMT-2010.04.09-03.10.08 Valid 09.04.2010 05:10:11.000  16              0           8
72..77 gpfsjt  @GMT-2010.04.09-03.05.03 Valid 09.04.2010 05:05:07.000  16              0           7
72..77 gpfsjt  @GMT-2010.04.09-03.00.06 Valid 09.04.2010 05:00:07.000  16              0           6
72..77 gpfsjt  @GMT-2010.04.09-00.32.43 Valid 09.04.2010 02:32:43.000  16              0           5
72..77 gpfsjt  @GMT-2010.04.08-20.52.41 Valid 08.04.2010 22:52:56.000  64              1           1
```

Figure 6-58 List snapshots

Viewing previous versions in the Microsoft Windows operating system

Snapshots that are created with a naming convention such as `@GMT-yyyy.MM.dd-HH.mm.ssname` are visible in the “Previous version” window of the Windows Explorer, as shown in Figure 6-59. The snapshots are only visible at the export level. To see the previous versions for an export, complete the following steps:

1. Open a Windows Explorer window to see the share for which you want previous versions to be displayed. In this example, `\\10.0.0.21` is the server and `sonas21jt` is the share.
2. Right-click the `sonas21jt` share name to open the `sonas21jt` share properties window as shown in step 1 in the diagram.
3. To select a time stamp for which you want to see the previous versions, double-click Today, April 09, 2010, 12:15 PM, as shown in step 2 in the diagram.
4. You are now presented with a pane (step 3) showing the previous versions of files and directories that are contained in the `sonas21jt` folder.


\\10.0.0.21
File Edit View Favorites Tools Help
Back Search Folders
Address \\10.0.0.21
Network Tasks
Add a network place
View network connections
1
2
3
sonas21jt on IBM SONAS Cluster (10.0.0.21) Properties
General Security Previous Versions Customize
\\10.0.0.21\sonas21jt (Today, April 09, 2010, 12:15 PM)
Address \\10.0.0.21\sonas21jt (Today, April 09, 2010, 12:15 PM)
Go
Other Places
My Network Places
My Documents
My Computer
View Copy... Restore
OK Cancel Apply Help

Name	Comments
sonas21jt	sonas21 test share for snapshots

To view a previous version of a folder, select the version from the following list and then click View. You can also save a folder to a different location or restore a previous version of a folder.

Folder versions:

Name	Time
sonas21jt	Today, April 09, 2010, 11:55 AM
sonas21jt	Today, April 09, 2010, 12:00 PM
sonas21jt	Today, April 09, 2010, 12:05 PM
sonas21jt	Today, April 09, 2010, 12:10 PM
sonas21jt	Today, April 09, 2010, 12:15 PM
sonas21jt	Today, April 09, 2010, 12:20 PM
sonas21jt	Today, April 09, 2010, 12:25 PM
sonas21jt	Today, April 09, 2010, 12:30 PM
sonas21jt	Today, April 09, 2010, 12:35 PM
sonas21jt	Today, April 09, 2010, 12:40 PM

Name	Size
file1	1 KB
file2	1 KB
file3	0 KB

Figure 6-59 View previous versions in the Microsoft Windows operating system

6.7 Local and remote replication

Data replication functions create a second copy of the file data and are used to offer a certain level of protection against data unavailability. Replication generally offers protection against component unavailability, such as a missing storage device or storage pod, but does not offer protection against logical file data corruption. When you replicate data, you usually want to send it to a reasonable distance as a protection against hardware failure or a site disaster event that makes the primary copy of data unavailable. For disaster protection, data is normally sent to a remote site at a reasonable distance from the primary site.

6.7.1 Synchronous versus asynchronous replication

Data replication can occur in two ways, depending when the acknowledgment to the writing application is returned: it can be *synchronous* or *asynchronous*. With synchronous replication, both copies of the data are written to their storage repositories before an acknowledgment is returned to the writing application. With asynchronous replication, one copy of the data is written to the primary storage repository, an acknowledgment is returned to the writing application, and then the data is written to the auxiliary storage repository. Asynchronous replication can be further broken down into *continuous* or *periodic* replication, depending on the frequency that batches of updates are sent to the auxiliary storage. The replication taxonomy is illustrated in Figure 6-60.

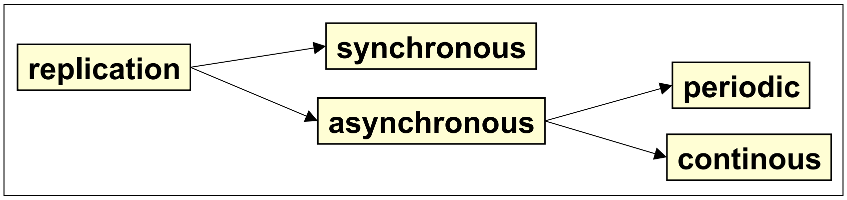


Figure 6-60 Replication types

Asynchronous replication is normally used when the additional latency because of the distance becomes problematic because it causes an unacceptable elongation to response times to the primary application.

6.7.2 Block-level versus file-level replication

Replication can occur at various levels of granularity. It can be *block level* when you replicate a disk or LUN and it can be *file level* when you replicate files or a portion of a file system, such as a directory or a file set.

File-level replication can either be either *stateless* or *stateful*. Stateless file replication occurs when you replicate a file to a remote site and then lose track of it. Stateful replication tracks and coordinates updates that are made to the local and remote file to maintain the two copies of the file in sync.

6.7.3 SONAS cluster replication

Replication can occur inside one single SONAS cluster or between a local SONAS cluster and a remote SONAS cluster. The term *intracluster replication* refers to replication between storage pods in the same SONAS cluster, and *intercluster replication* occurs between one SONAS cluster and a remote destination that can be a separate SONAS cluster or a file server. With intracluster replication, the application does not need to be aware of the location of the file, and failover is transparent to the application itself. For intercluster replication, the application must be aware of the file's location and must connect to the new location to access the file.

Figure 6-61 shows two SONAS clusters with *file1* replicated with intracluster replication and *file2* replicated with intercluster replication.

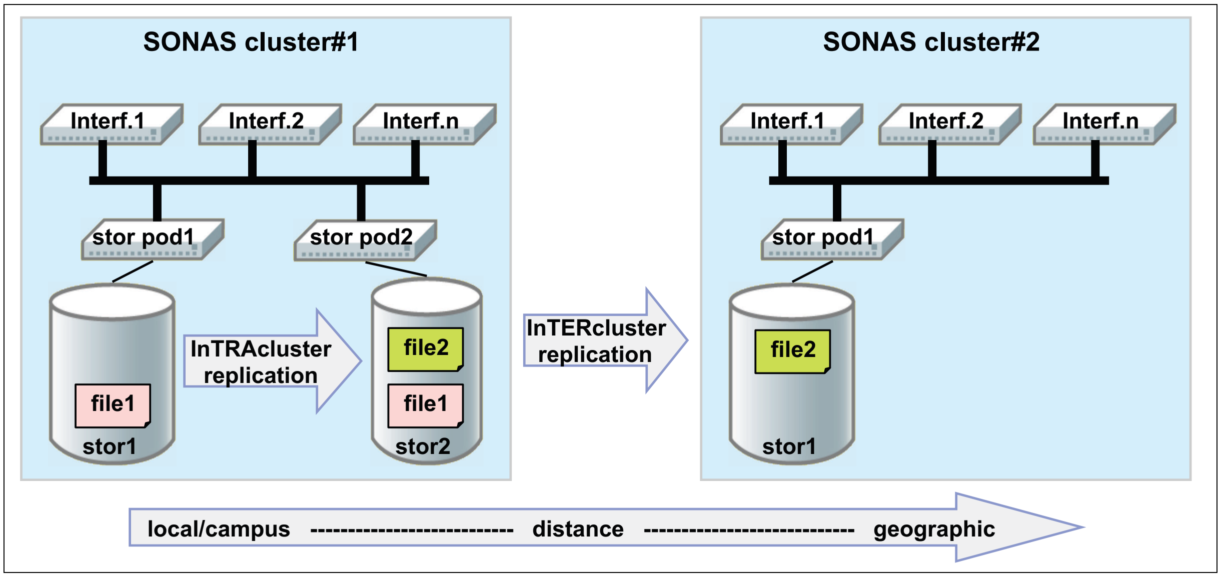


Figure 6-61 Replication options

Table 6-1 shows the possible SONAS replication scenarios.

Table 6-1 SONAS replication solutions

Type	Intracluster or intercluster	Stateful or stateless	Local or Remote distance
Synchronous	Intracluster	Stateful	Local
Asynchronous	Intercluster	Stateless	Remote

6.7.4 Local synchronous replication

Local synchronous replication is implemented within a single SONAS cluster, so it is defined as intracluster replication. Synchronous replication is protection against total loss of a whole storage building block or storage pod and it is implemented by writing all data blocks to two storage building blocks that are part of two separate failure groups. Synchronous replication is implemented by using separate GPFS failure groups. Currently, synchronous replication applies to an entire file system and not to the individual file set.

However, SONAS V1.3 and later versions include a file cloning capability that is described in 6.5.5, "File cloning in SONAS V1.3" on page 419.

With synchronous replication, because the writes are acknowledged to the application only when both writes are complete, write performance is dictated by the slower storage building block. High latencies can degrade performance, so it is a short distance replication mechanism. Synchronous replication requires an InfiniBand connection between both sites and an increase in distances can decrease the performance.

Tip: Synchronous replication between sites is not supported in SONAS.

Another use case is protection against total loss of a complete site. In this scenario, a complete SONAS cluster (including Interface and Storage nodes) is split across two sites. The data is replicated between both sites, so that every block is written to a building block on both sites. For correct operation, the administrator must define correct failure groups. For the two-site scenario, you need one failure group for each site. For SONAS V1.3 and later versions, this use case is not applicable because all InfiniBand switches are in the same rack and unavailability of this rack stops SONAS cluster communications.

Synchronous replication does not distinguish between the two storage copies. SONAS does not have a preferred failure group concept where it sends all reads; the reads are sent from disks in both failure groups.

Synchronous replication in the SONAS file system offers the following replication choices:

- No replication at all
- Replication of metadata only
- Replication of data and metadata

From a reliability perspective, it is preferable that metadata replication is always used for file systems within the SONAS cluster. Synchronous replication can be established at file system creation time or later when the file system already contains data. Depending on when replication is applied, various procedures must be followed to enable synchronous replication. Synchronous replication requires that the disks belong to two distinct failure groups to ensure that the data and metadata is not replicated to the same physical disks. It is preferable that the various failure groups are defined on various storage enclosures and storage controllers to make sure that failover is possible if a physical disk component becomes unavailable.

Synchronous replication has the following prerequisites:

- Two separate failure groups must be present.
- The two failure groups must have the same number of disks.
- The same number of disks from each failure group and the same disk usage type must be assigned to the file system.

Establishing synchronous replication at file system creation

Synchronous replication across failure groups can be established as an option at file system creation time by using either the GUI or the **mkfs** command and specifying the **-R** option. This option sets the level of replication that is used in this file system and can be one of the following values:

- **none**, which means no replication at all
- **meta**, which indicates that the file system metadata is synchronously mirrored
- **all**, which indicates that the file system data and metadata is synchronously mirrored

Establishing synchronous replication after file system creation

Establishing synchronous replication after file system creation cannot be done by using the GUI; it requires the CLI interface. To enable synchronous replication, complete the following steps:

1. Enable synchronous replication with the change file system (**chfs**) command and specify the **-R** option.
2. Redistribute the file system data and metadata with the **restripefs** command.

This section shows how to enable synchronous replication on an existing file system that is called `gpfsjt`.

You can use the **lsdisk** command to see the available disks and the **lsfs** command to see the file systems, as shown in Figure 6-62.

```
[SONAS]$ lsdisk
Name     File system Failure group Type            Pool     Status Availability Timestamp
Name     File system Failure group Type            Pool     Status Availability Timestamp
gpfs1nsd gpfs0       1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs2nsd gpfs0       1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs3nsd gpfsjt      1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs4nsd             1             dataAndMetadata userpool ready               4/13/10 1:55 AM
gpfs5nsd             1             dataAndMetadata system   ready               4/13/10 1:55 AM
gpfs6nsd             2             dataAndMetadata userpool ready               4/13/10 1:55 AM

[SONAS]$ lsfs
Cluster Devicen Mountpoint  .. Data replicas Metadata replicas Replication policy Dmapi
sonas02 gpfs0   /ibm/gpfs0  .. 1             1                 whenpossible       F
sonas02 gpfsjt  /ibm/gpfsjt .. 1             1                 whenpossible       T
```

Figure 6-62 Disks and file system before replication

By using the example in Figure 6-62, you can verify the number of disks that are assigned to the `gpfsjt` file system in the **lsdisk** output and see that only one disk that is called `gpfs3nsd` is used. To create the synchronous replica, you need the same number of disks as the number of disks that are assigned to the file system. From the **lsdisk** output, you can also verify that there are enough free disks that are not assigned to any file system. You can use the disk called `gpfs5nsd` to create the data replica.

The disk that is called `gpfs5nsd` is in failure group 1 as the primary disk, and you must assign the disk to a separate failure group 2 by running **chdisk**, as shown in Figure 6-63. Then, verify the disk status with the **lsdisk** command. Also, verify that the new disk, `gpfs5nsd`, is in the same pool as the current disk `gpfs3nsd`.

```
[SONAS]$ chdisk gpfs5nsd --failuregroup 2
EFSSG0122I The disk(s) are changed successfully!

[SONAS]$ lsdisk
Name     File system Failure group Type            Pool     Status Availability Timestamp
gpfs1nsd gpfs0       1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs2nsd gpfs0       1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs3nsd gpfsjt      1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs4nsd             1             dataAndMetadata userpool ready               4/13/10 2:15 AM
gpfs5nsd             2             dataAndMetadata system   ready               4/13/10 2:15 AM
gpfs6nsd             2             dataAndMetadata userpool ready               4/13/10 2:15 AM
```

Figure 6-63 Assign a new failure group to a disk

Add the new disk to the gpfsjt file system by running **chfs -add**, as shown in Figure 6-64, and verify the outcome by running **lsdisk**.

```
[SONAS]$ chfs gpfsjt -add gpfs5nsd
The following disks of gpfsjt are formatted on node mgmt001st002.virtual.com:
    gpfs5nsd: size 1048576 KB
Extending Allocation Map
Checking Allocation Map for storage pool 'system'
  52 % complete on Tue Apr 13 02:22:03 2010
 100 % complete on Tue Apr 13 02:22:05 2010
Completed adding disks to file system gpfsjt.
mmadddisk: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
EFSSG0020I The filesystem gpfsjt has been successfully changed.

[SONAS]$ lsdisk
Name     File system Failure group Type            Pool     Status Availability Timestamp
gpfs1nsd gpfs0       1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs2nsd gpfs0       1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs3nsd gpfsjt      1             dataAndMetadata system   ready  up           4/12/10 3:03 AM
gpfs5nsd gpfsjt      2             dataAndMetadata system   ready  up           4/13/10 2:26 AM
gpfs4nsd             1             dataAndMetadata userpool ready               4/13/10 2:26 AM
gpfs6nsd             2             dataAndMetadata userpool ready               4/13/10 2:26 AM

[SONAS]$ lsfs
Cluster Devicen Mountpoint  .. Data replicas Metadata replicas Replication policy Dmapi
sonas02 gpfs0   /ibm/gpfs0  .. 1             1                 whenpossible       F
sonas02 gpfsjt  /ibm/gpfsjt .. 1             1                 whenpossible       T
```

Figure 6-64 Add a disk to a file system

From the **lsdisk** output, you can see that gpfs5nsd is assigned to the gpfsjt file system. In the **lsfs** output, only one copy of data and metadata are displayed, as shown in the Data replicas and Metadata replicas columns. To activate data and metadata replication, you must run **chfs -R**, as shown in Figure 6-65.

```
[SONAS]$ chfs gpfsjt -R all
EFSSG0020I The filesystem gpfsjt has been successfully changed.

[SONAS]$ lsfs
Cluster DevicenMountpoint      Data replicas Metadata replicas Replication policy Dmapi
sonas02 gpfs0   /ibm/gpfs0  .. 1             1                 whenpossible       F
sonas02 gpfsjt  /ibm/gpfsjt .. 2             2                 whenpossible       T
```

Figure 6-65 Activate data replication

The **lsfs** command now shows that there are two copies of the data in the gpfsjt file system.

Run **restripefs** with the replication switch to redistribute data and metadata, as shown in Figure 6-66.

```
[SONAS]$ restripefs gpfsjt --replication
Scanning file system metadata, phase 1 ...
Scan completed successfully.
Scanning file system metadata, phase 2 ...
  64 % complete on Thu Apr 15 23:11:00 2010
  85 % complete on Thu Apr 15 23:11:06 2010
 100 % complete on Thu Apr 15 23:11:09 2010
Scan completed successfully.
Scanning file system metadata, phase 3 ...
Scan completed successfully.
Scanning file system metadata, phase 4 ...
Scan completed successfully.
Scanning user file metadata ...
EFSSG0043I Restriping of filesystem gpfsjt completed successfully.
[root@sonas02.mgmt001st002 dirjt]#
```

Figure 6-66 Restripefs to activate replication

SONAS does not offer any command to verify that the file data is being replicated. To verify the replication status, connect to SONAS as a root user and run **mmlsattr** with the **-L** switch, as illustrated in Figure 6-67. The report shows the metadata and data replication status; you can see that you have two copies for both metadata and data.

```
[root@sonas02.mgmt001st002 userpool]# mmlsattr -L *
file name:            f1.txt
metadata replication: 2 max 2
data replication:     2 max 2
immutable:            no
flags:
storage pool name:    system
fileset name:         root
snapshot name:

file name:            f21.txt
metadata replication: 2 max 2
data replication:     2 max 2
immutable:            no
flags:
storage pool name:    userpool
fileset name:         root
snapshot name:
```

Figure 6-67 Verify that file data is replicated

File system synchronous replication can also be disabled by running **chfs**, as shown in the following example:

```
chfs gpfsjt -R all
```

After you change the file system attributes, the **restripefs** command must be run to remove replicas of the data, as shown in the following example:

```
restripefs gpfsjt --replication
```

6.7.5 Remote asynchronous replication

This section provides information about how you can create and use SONAS replication.

Introduction

Asynchronous replication allows replication of file systems across long distances or to low-performance, high-capacity storage systems.

The ability to continue operations in the face of a regional disaster is handled through asynchronous replication that is provided by the SONAS system. Asynchronous replication allows for one or more file systems within a SONAS file name space to be defined for replication to another SONAS system over the customer network infrastructure. Files that were created, modified, or deleted at the primary location are carried forward to the remote system at each invocation of the asynchronous replication.

The asynchronous replication process looks in a specified file system of the source SONAS system for files that changed since the last replication cycle was started for that file system, and uses the **rsync** tool to move efficiently only the changed portions of a file to the target system. In addition to the file contents, all extended attribute information about the changed file is also replicated to the remote system. File set information is not replicated.

The file-based movement allows the source and destination file trees to be of differing sizes and configurations, if the destination file tree is large enough to hold the contents of the files from the source. Differing configurations allow for options such as local synchronous copies of the file tree to be used at the source location, but not used at the destination. This allows for great flexibility in tailoring the solution for many different needs.

Asynchronous replication is configured in a single direction one-to-one relationship, such that one site is considered the source of the data, and the other is the target. The replica of the file system at the target remote location is intended to be used in read-only mode until a disaster or other source file system downtime occurs. During a file system failure recovery operation, failback is accomplished by defining the replication relationship from the original target back to the original source.

Example

Figure 6-68 illustrates the high-level picture of the replication relationship.

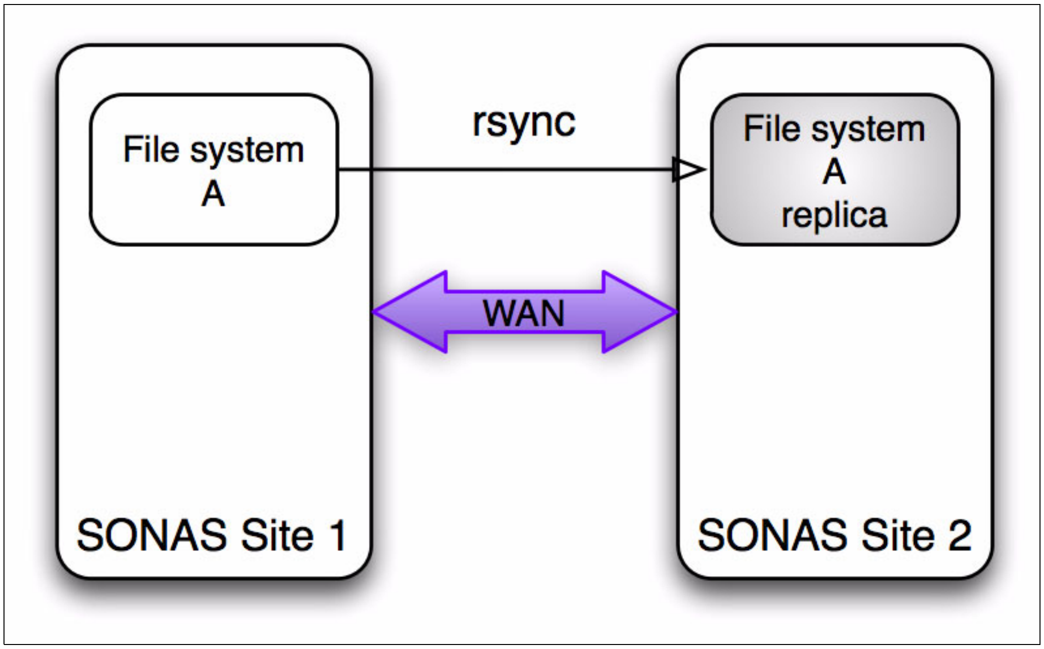


Figure 6-68 Replication relationship

6.7.6 Async replication topologies

For business continuance in a disaster, SONAS currently supports a 1:1 relationship between SONAS systems. Each SONAS is an independent system from one another. The connectivity between the systems is through the customer network between the customer-facing network adapters in the Interface nodes.

The systems must be capable of routing network traffic between one another by using the customer supplied IP addresses or fully qualified domain name (FQDN) of the Interface nodes.

Async replication in single direction

One of the topologies is a relationship where there is a distinct primary and secondary SONAS system. The SONAS at site 2 is a backup of the system at site 1, and maintains no other file systems other than replicas for site 1. The second system can be used for testing purposes, continuing production in a disaster, or for restoring the primary site after a disaster.

Figure 6-69 illustrates the relationship between the primary and secondary sites for this scenario.

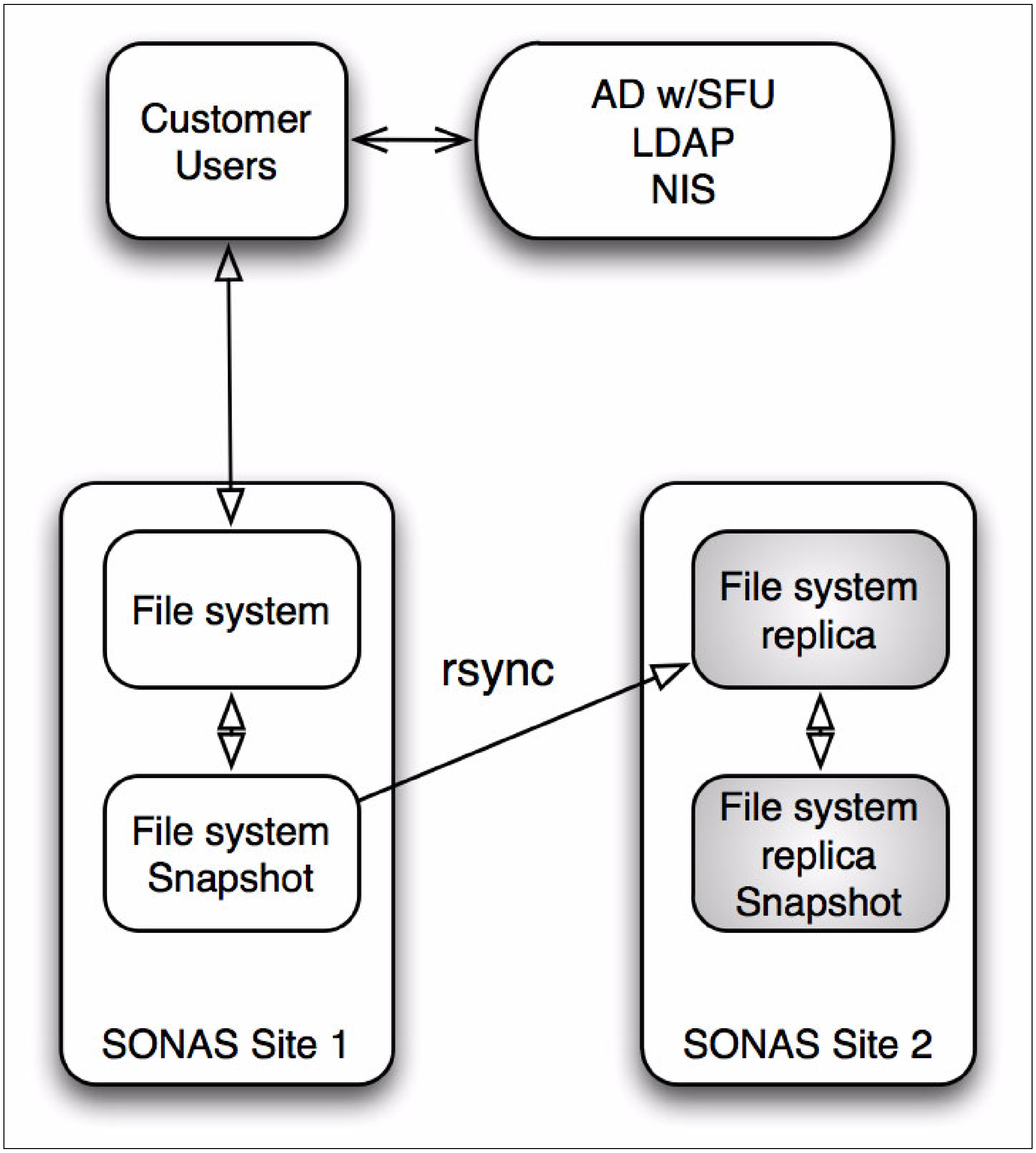


Figure 6-69 Single direction async replication

Async replication in two directions

The second topology is when the second site exports shares of a file system in addition to holding mirrors of a file tree from the primary site (see Figure 6-70). This scenario is when the SONAS at both sites is used for production I/O, in addition to being the target mirror for the other SONAS system's file structure. This mirroring can be in both directions, such that both SONAS systems have their own file trees, in addition to the having the file tree of the other. Or it can be that both have their own file tree, and only one has the mirror of the other.

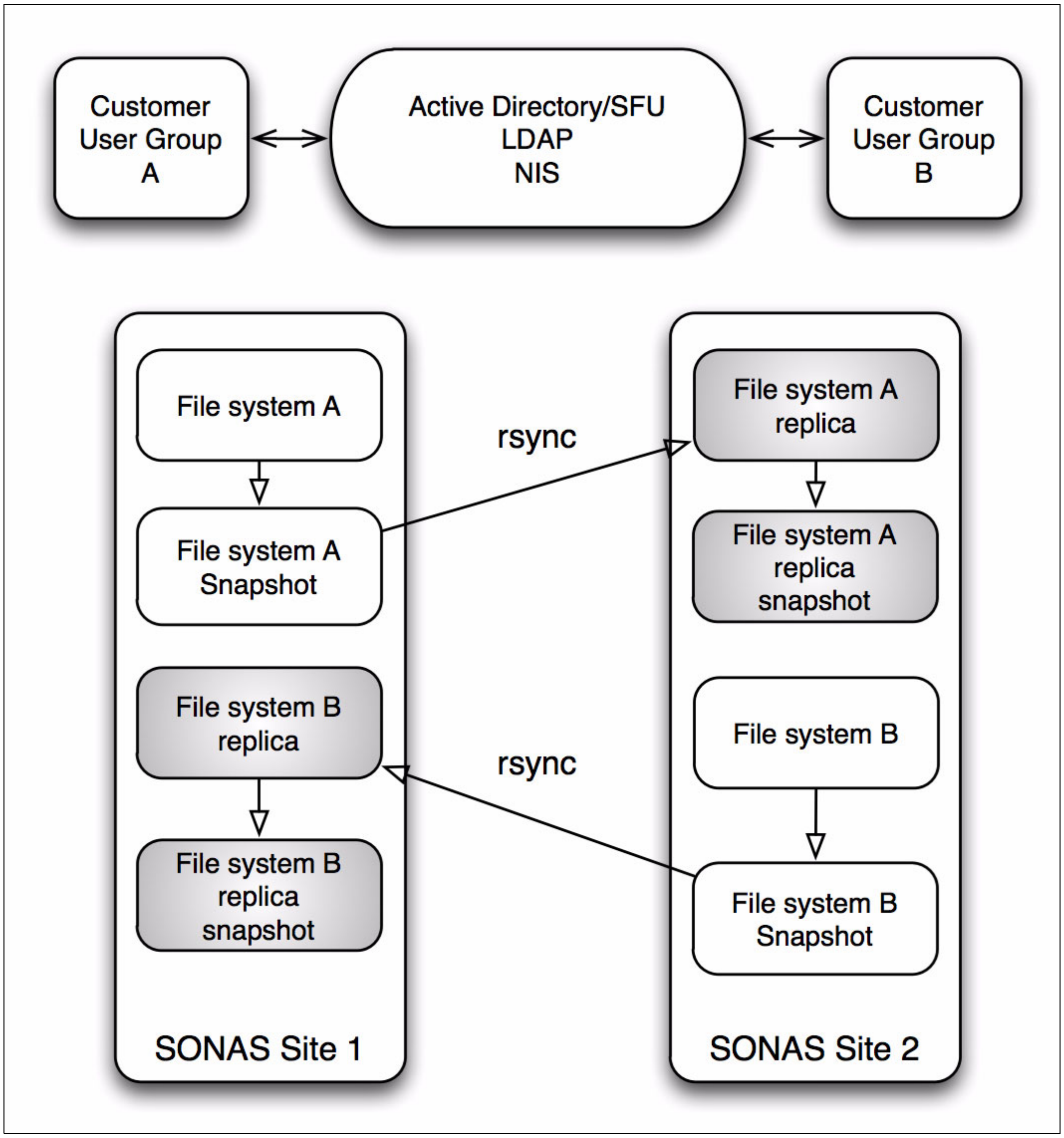


Figure 6-70 Async replication in two directions

Tip: In most cases, the cluster must be sized for two-way replication. Assuming that you are replicating two file systems in both directions, a cluster must have twice the spindle count to serve each file system at the capacity and performance level that is evenly divided.

6.8 Managing asynchronous replication

To configure and manage asynchronous replication, you can use the GUI or the CLI.

6.8.1 Introduction

The async replication function is intended to be a function that is run on a periodic basis to create a replica of a file system's contents on a source SONAS system to a file system on a destination SONAS. When it is started, the following major steps are done during the replication process:

1. A snapshot of the source file system is created.
2. The source file system snapshot is scanned to identify files and directories that were created, modified, or deleted since the last asynchronous replication completed.
3. The changed contents (the specific changed blocks) are identified for files or directories that have changed.
4. Changed contents (file blocks) are replicated to the target system.
5. A snapshot of the target file system is created.
6. The source file system snapshot is removed.

Overview

The source and target snapshots can be configured to be omitted from the replication process, but it is not recommended. The source-side snapshot creates a point-in-time image of the source file system when the async replication process is started. Async then uses this snapshot to look for changes and to use this source as the basis for the replication to the destination.

Use the target system only in read-only mode except when the target data is being used as the primary data source, for example, during disaster recovery. The target system can later be configured to replicate asynchronously its contents back to the primary system to reestablish the file system contents back to the previous version of the file system as it existed when the copy was created on the source system. Asynchronous replication can be configured bidirectionally so that one system is the primary site for some file systems and a second system is the primary site for other file systems, but no single file system’s asynchronous replication is concurrently bidirectional. However, if asynchronous replication is not bidirectional for a single file system, you cannot replicate a file system.

The destination snapshot creates an image of the destination file system at the time async replication completes. This process creates an image of the file system, which can be used for issues or errors between replications or during the next async update.

The Management node of the source SONAS system is the node that async is initiated on and controls the async operation. Async is designed to spread the scan and replication work across a defined number of source and destination Interface nodes to have parallel efforts to complete quickly the replication task. The source Management node coordinates and distributes the work elements to the configured Interface nodes.

All changes are tracked by the source-side SONAS system, which carries these changes forward to the destination through async replication. The destination system should be used only in R/O mode until such time that it is required to be made R/W to provide business continuance operations. This configuration prevents changes from being made to the destination system that are independent of the source SONAS system’s visibility.

If required, the secondary SONAS system can be configured to replicate asynchronously its contents back to the primary SONAS to re-establish the file system contents back to the original SONAS.

Consistent authentication mappings between the source and the target are required. Authentication management must be provided by an Active Directory with Services for UNIX (SFU) extension, by Network Information Service (NIS), or by an LDAP server. Active Directory server without the SFU extension is not supported.

Asynchronous replication is compatible with Tivoli Storage Manager Hierarchical Storage Management for Windows management of files in both source and target. Policies for the source can differ from policies that are implemented for the target. New or changed files should not be moved by Tivoli Storage Manager Hierarchical Storage Management for Windows to auxiliary storage before they are moved to the asynchronous replication target system because asynchronous replication causes the recall of a file to primary storage on the source system so that it can be moved to the target system. For simplicity, give the asynchronous replication source and target system the same Tivoli Storage Manager Hierarchical Storage Management for Windows configuration, capabilities, and management policies.

Requirements

Observe the following requirements:

- The active Management node and the Interface nodes of the source system must communicate over the network with the active Management node and Interface nodes of the target system.
- The target system file system must be large enough, with enough free space to allow for replication of the source file system along with enough free space to accommodate snapshots.
- Sufficient network bandwidth is required to replicate all of the file system delta changes with a latency that is sufficient to meet Recovery Point Objective (RPO) needs during peak utilization.
- The active Management node and Interface nodes of the source system must be able to communicate with the active Management node and Interface nodes of the target system over the customer network.
- TCP port 1081 is required on the source and target systems for the configuration process to establish secure communications from the target active Management node to the source active Management node by using SSH.
- TCP port 22 is required on the source and target systems for `rsync` to use SSH to transfer encrypted file changes from the source active Management node and Interface nodes to the target active Management node and Interface nodes.
- For replication in both directions or for potential failback after a recovery, ports 1081 and 22 should be open in both directions.
- The customer has either an LDAP NIS or AD with SFU environment that is resolvable across their sites, or is mirrored/consistent across their sites such that the SONAS at each site is able to authenticate from each location.
- The authentication mechanism is the same across both locations.
- The time synchronization across both sites is sufficient to allow for successful authentication with SONAS systems.

6.8.2 Configuring asynchronous replication

Before asynchronous replication can occur between two sites, communication between the participating systems and the replication configuration must be established.

Prerequisites

You must configure the asynchronous replication relationship between the two systems before you configure file system replication. Asynchronous replication is normally used between source and target systems where distance might affect response time because of bandwidth shortages. Only changed blocks of a file are transferred to the target system, rather than the entire file, which can simplify and quicken restore operations.

The replica of the file system at the destination system is intended to be used in read-only mode until a disaster or other source file system downtime occurs. During a file system failure-recovery operation, failback is accomplished by defining the replication relationship from the original target system back to the original source system and replicating the data back to the original source.

Information needed

Before you set up asynchronous and file system replication, you need the following information to complete replication configuration:

- The public IP address of the Management node of the source system is needed when you configure asynchronous replication on the target system.
- The public IP address of the Management node of the target system and the public IP addresses of the target Interface nodes are needed when you configure asynchronous replication on the source system.

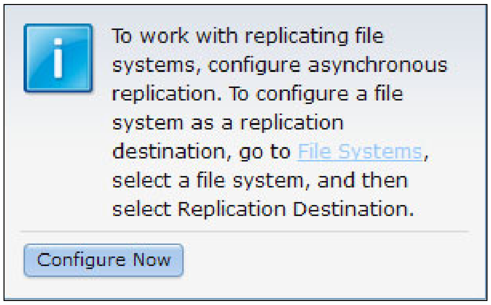
6.8.3 GUI replication configuration

This section describes the steps to configure replication with the GUI.

Configuring asynchronous replication on the target system

Complete the following steps:

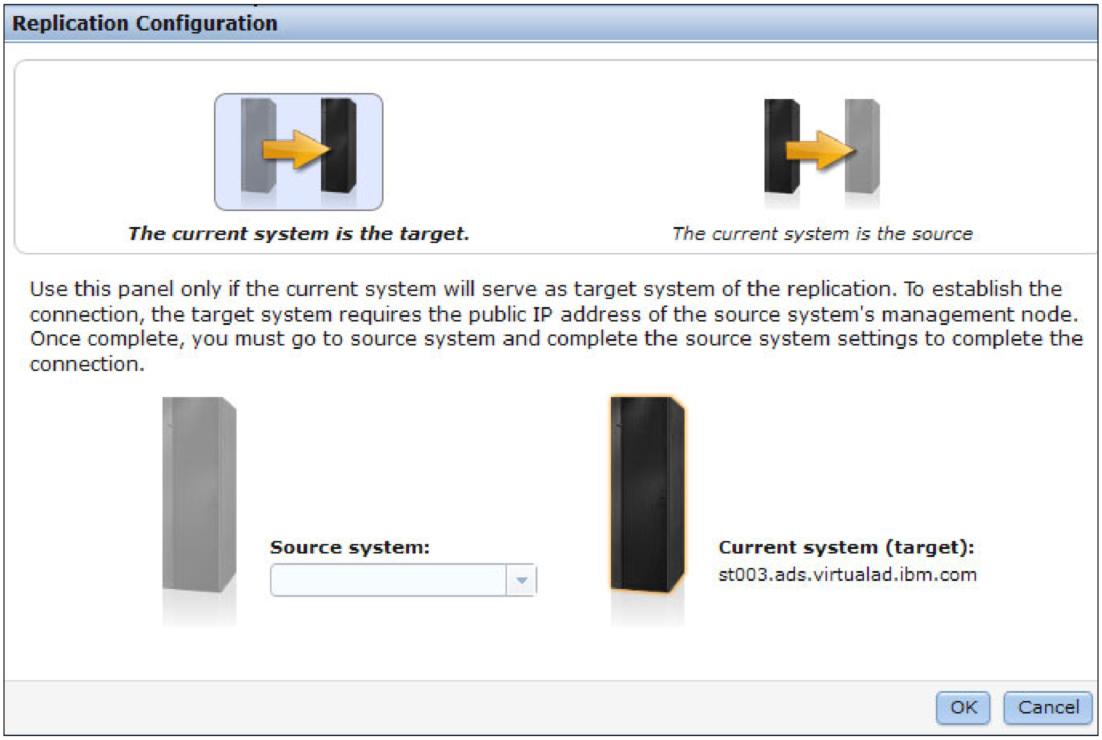
1. On the target system, click **Copy Services** → **Replication**.
2. Click **Actions** → **Configure**. You can see the message that is shown in Figure 6-71.



To work with replicating file systems, configure asynchronous replication. To configure a file system as a replication destination, go to File Systems, select a file system, and then select Replication Destination.

Figure 6-71 Replication configuration with the GUI

3. Click **The current system is the target**, as shown in Figure 6-72.



Replication Configuration

The current system is the target.

The current system is the source

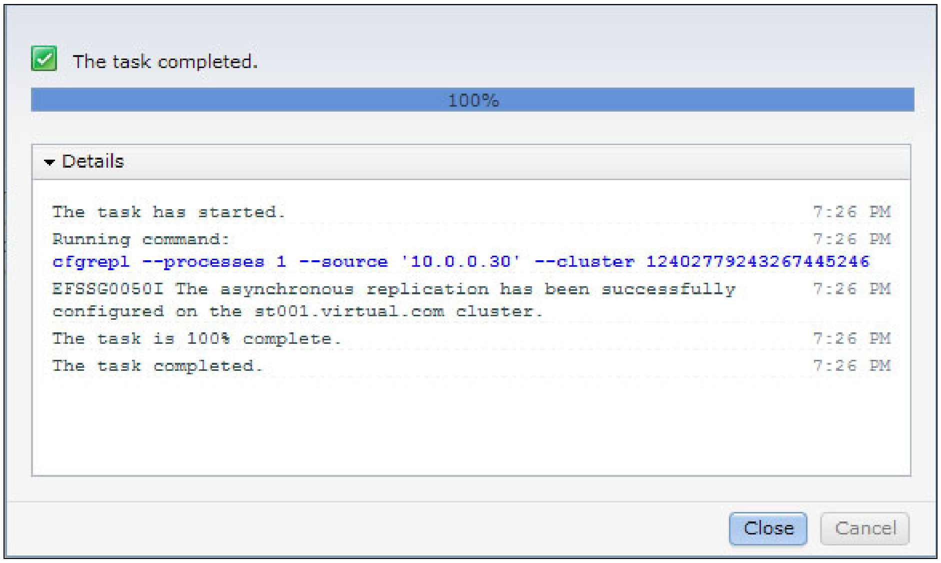
Use this panel only if the current system will serve as target system of the replication. To establish the connection, the target system requires the public IP address of the source system's management node. Once complete, you must go to source system and complete the source system settings to complete the connection.

Source system:

Current system (target): st003.ads.virtualad.ibm.com

Figure 6-72 Replication configuration - specify the role of the current system

4. Enter the public IP address for the management service of the source system and click **OK**.
5. When this task is finished, click **Close** to continue (see Figure 6-73).



The task completed.

```
The task has started.                                                  7:26 PM
Running command:                                                       7:26 PM
cfgrepl --processes 1 --source '10.0.0.30' --cluster 12402779243267445246
EFSSG0050I The asynchronous replication has been successfully          7:26 PM
configured on the st001.virtual.com cluster.
The task is 100% complete.                                             7:26 PM
The task completed.                                                    7:26 PM
```

Figure 6-73 Replication configuration task status and CLI commands that are run

6. Click **Files** → **File Systems**.

7. Select a file system and from the **Action** menu (or right-click the file system), select the **Replication Destination** check box, as shown in Figure 6-74.

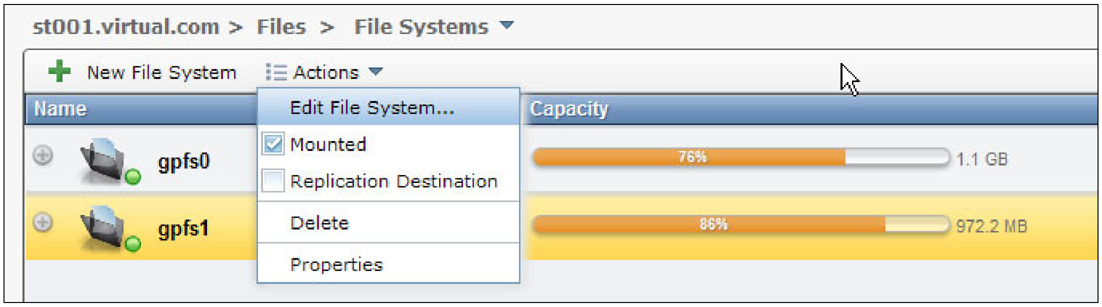
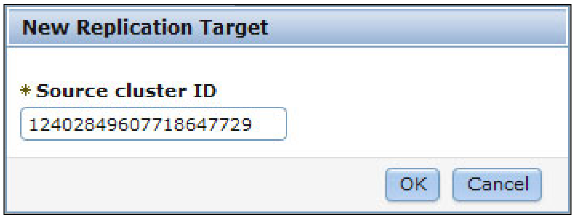


Figure 6-74 Replication Destination selection

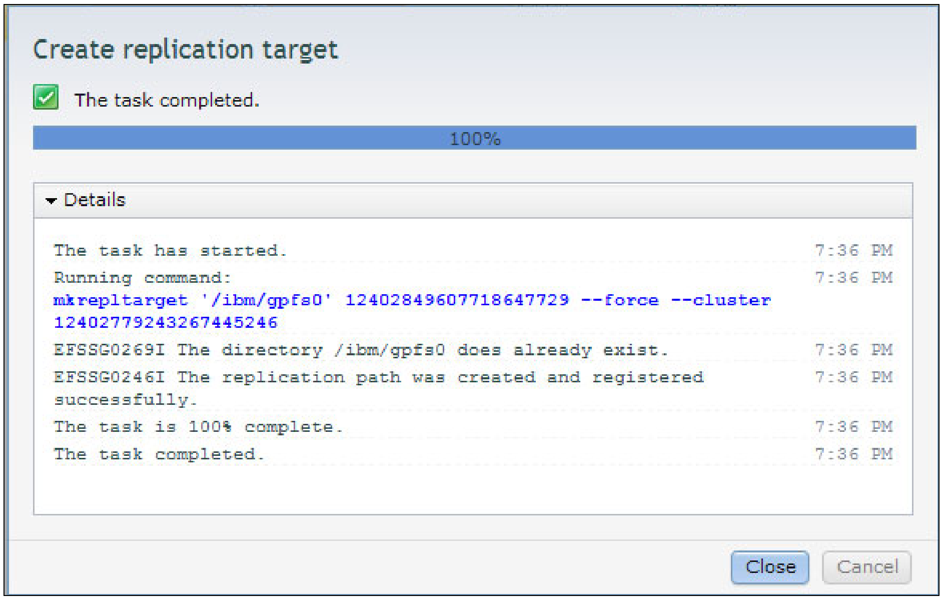
8. Enter the Source systems Cluster ID. On the source system, run `lscluster` from the CLI or go to Monitoring system details to get the Cluster ID (Figure 6-75).



* Source cluster ID: 12402849607718647729

Figure 6-75 New Replication Target - Source cluster ID specification

9. Click **OK** to finish the configuration on the target system. Figure 6-76 shows the progress of the replication target process.



☑ The task completed.

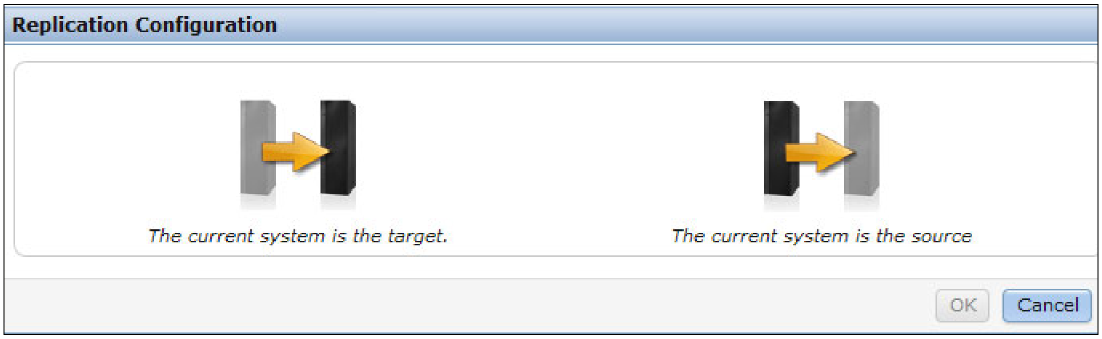
```
The task has started.                                              7:36 PM
Running command:                                                   7:36 PM
mkrepltarget '/ibm/gpfs0' 12402849607718647729 --force --cluster
12402779243267445246
EFSSG0269I The directory /ibm/gpfs0 does already exist.            7:36 PM
EFSSG0246I The replication path was created and registered         7:36 PM
successfully.
The task is 100% complete.                                         7:36 PM
The task completed.                                                7:36 PM
```

Figure 6-76 Create replication target progress

Configuring asynchronous replication on the source system

Complete the following steps:

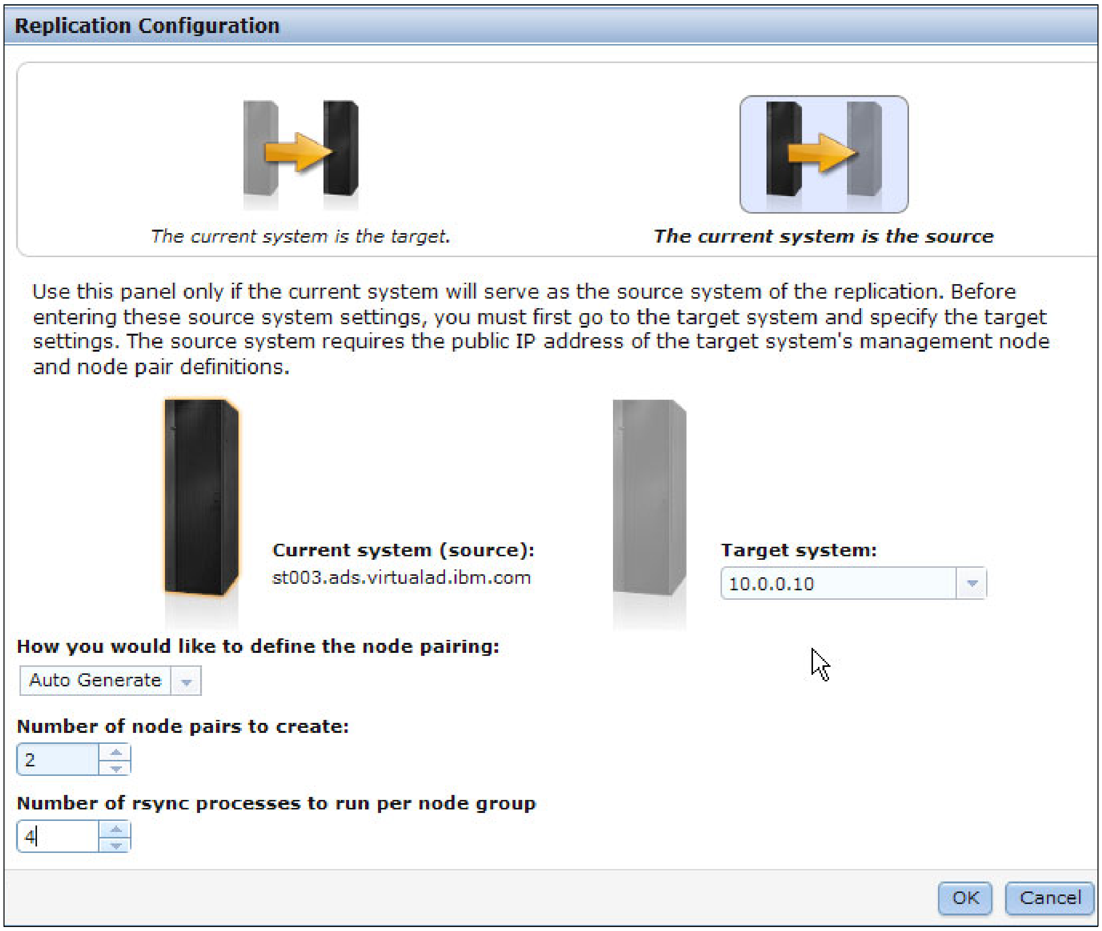
1. On the source system, click **Copy Services** → **Replication**.
2. Click **Actions** → **Configure**.
3. Click **The current system is the source**, as shown in Figure 6-77.



Replication Configuration
The current system is the target.
The current system is the source
OK
Cancel

Figure 6-77 Current system as replication source

4. Enter the public IP address for the management service of the target system and select the method of creating node pairs between the source and target (see Figure 6-78). You can choose to automatically generate node pairing or manually define node pairs.



Replication Configuration
The current system is the target.
The current system is the source

Use this panel only if the current system will serve as the source system of the replication. Before entering these source system settings, you must first go to the target system and specify the target settings. The source system requires the public IP address of the target system's management node and node pair definitions.

Current system (source):
st003.ads.virtualad.ibm.com

Target system:
10.0.0.10

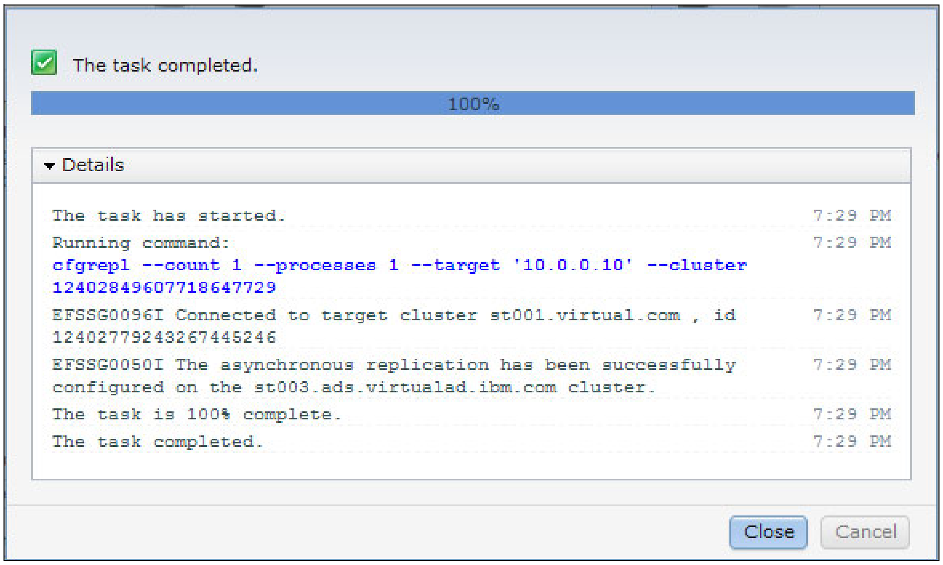
How you would like to define the node pairing:
Auto Generate

Number of node pairs to create:
2

Number of rsync processes to run per node group
4

Figure 6-78 Specify the IP address and method for node pairing

5. Click **OK** to save the settings. Monitor the task and view the corresponding CLI commands (see Figure 6-79).

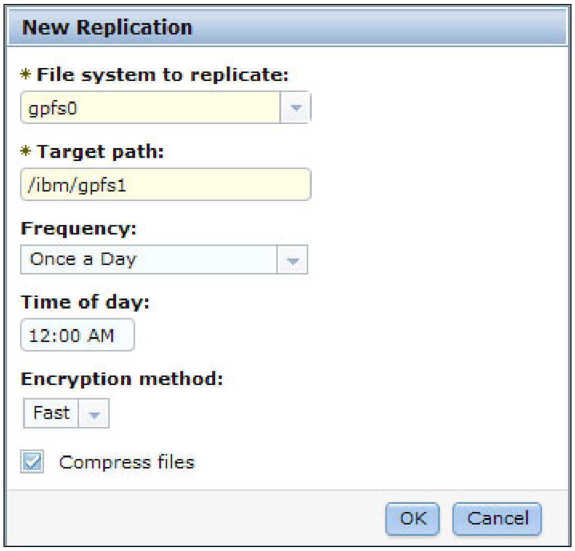


```
The task has started.                                        7:29 PM
Running command:                                             7:29 PM
cfgrepl --count 1 --processes 1 --target '10.0.0.10' --cluster
12402849607718647729
EFSSG0096I Connected to target cluster st001.virtual.com , id   7:29 PM
12402779243267445246
EFSSG0050I The asynchronous replication has been successfully   7:29 PM
configured on the st003.ads.virtualad.ibm.com cluster.
The task is 100% complete.                                   7:29 PM
The task completed.                                          7:29 PM
```

Figure 6-79 Task progress and CLI commands that are generated and run in the background

6. Click **OK** to save the settings.
7. Click **New Replication** from the menu.
8. Specify the file system and the path on the target system where the data from the specified file system is replicated. When you set the target path, you have two options:
   - *Target path* is the root directory of the target file system. The source file system contents are copied to the root of a target file system, such that the target directory tree matches that of the source from the file system mount point. Use this method to ensure equivalent failover and failback between the source and target systems.
   - Target path is a directory within the target file system. The source file system contents are copied to a specified directory within the target file system. By specifying a directory, you can replicate multiple source file systems to a single target file system. However, in failback scenarios where data is replicated from the target file system to the original source file system, the directory structure is changed and requires changes to applications and users of these files.
9. Select the frequency of resynchronization operations for node pairs. Resynchronization compares the source and target files for differences and then copies only the changed information about the target to ensure that the contents of the target match the source.
10. Select the frequency and the time when the changed file system data is replicated to the target system.
11. Select the appropriate encryption method to protect data as it is replicated between the systems. Strong encryption provides a more complex algorithm that provides more protection to data as it is transmitted; however, it can slow transmission for data. Fast encryption transmits data more quickly, but does not provide as much protection of the data.

12. Optionally, you can select to compress data as it is written to the destination system (see Figure 6-80). If the data supports compression, compression reduces the amount of data that is transferred over the network and can make replication faster.



New Replication

* File system to replicate: gpfs0

* Target path: /ibm/gpfs1

Frequency: Once a Day

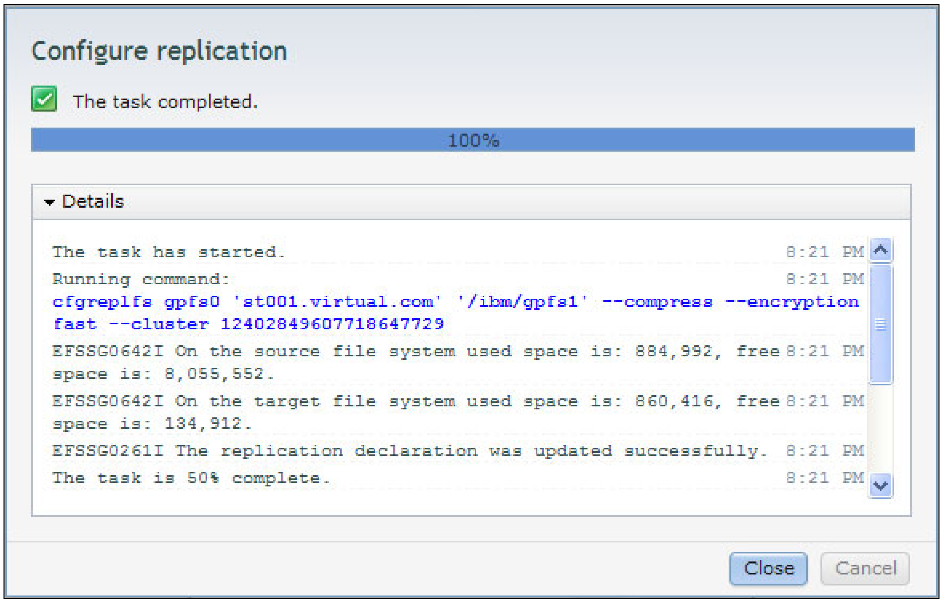
Time of day: 12:00 AM

Encryption method: Fast

☑ Compress files

Figure 6-80 New replication options

13. Click **OK** to continue. The task progress window opens, as shown in Figure 6-81.



Configure replication

☑ The task completed.

100%

Details

```
The task has started.                                              8:21 PM
Running command:                                                   8:21 PM
cfgreplfs gpfs0 'st001.virtual.com' '/ibm/gpfs1' --compress --encryption
fast --cluster 12402849607718647729
EFSSG0642I On the source file system used space is: 884,992, free 8:21 PM
space is: 8,055,552.
EFSSG0642I On the target file system used space is: 860,416, free 8:21 PM
space is: 134,912.
EFSSG0261I The replication declaration was updated successfully.  8:21 PM
The task is 50% complete.                                          8:21 PM
```

Figure 6-81 Configure replication task progress

6.8.4 CLI usage

Run **cfgrepl** to configure the source and target nodes for asynchronous replication and to identify which nodes participate.

Prepare the systems for communication by running **cfgrepl** on the target system by using the **--source** option and providing the source system Management node IP address, as shown in Example 6-9.

Example 6-9 The cfgrepl command

```
[admin@st001.mgmt002st001 ~]# cfgrepl  --source 10.0.0.30
EFSSG0050I The asynchronous replication has been successfully configured on the
st001.virtual.com cluster.
EFSSG1000I The command completed successfully.
```

This command defines the system with which the target SONAS system can be paired.

> **Tip:** This command unlocks the target SONAS for the specific source SONAS that is defined in the command. This step prevents accidental misconfiguration. When the source now goes to pair with this target, it compares the system against the one defined here to validate it is the correct relationship. Only the **--source** option and its value are required. The other parameters have meaning only when you are configuring the relationship on the source system.

On the source Management node, run **cfgrepl** to specify the target Management node IP address and the source-to-target Interface node pairings. When you are specifying node pairs, use the Interface node name for the source Interface nodes. Run **lsnode** to obtain this value. The target node must be an external IP address reachable over the WAN (see Example 6-10).

Example 6-10 Example cfgrepl command with target Management node

```
[admin@st003.mgmt001st003 ~]# cfgrepl --target 10.0.0.10 --pairs
int001st003:10.0.0.112
EFSSG0096I Connected to target cluster st001.virtual.com , id 12402779243267445246
EFSSG0050I The asynchronous replication has been successfully configured on the
st003.ads.virtualad.ibm.com cluster.
EFSSG1000I The command completed successfully.
```

You can use the **-n** option to specify the number of Interface nodes to use in the replication configuration rather than specifying node pairs. If the **-n** option is used, the system automatically selects the specified number of node pairings from available Interface nodes.

You can use the **--processes** option to specify the number of parallel processes per node. The default is 10. For systems where network bandwidth and sharing CPU with other workloads is not a concern, increasing the number of processes can provide significant performance improvements to the overall replication process.

If you use the **--target** option, you can also use the **--forcekeyupdate** option to exchange the SSH keys between the source and target systems. This option is required if the target system has been reinstalled, which generates a new set of SSH keys. This option can also be used if errors indicate that there is a problem with the SSH communication because the keys are not matching the keys that were stored during the initial configuration of the system.

On the target system, create a target path by running **mkrepltarget**, providing a path for the source system's data and the source system's ID. To determine the system ID, run **lscluster**. The **--force** parameter is required when you are specifying a target directory that exists; this includes the base directory of a file system. For example, to make a target directory on the target /ibm/gpfs1 for the data of system 12402849607718647729, run the command that is shown in Example 6-11.

Example 6-11 Example mkrepltarget command

```
[admin@st001.mgmt002st001 ~]# mkrepltarget /ibm/gpfs1 12402849607718647729 --force
EFSSG0269I The directory /ibm/gpfs1 does already exist.
EFSSG0246I The replication path was created and registered successfully.
EFSSG1000I The command completed successfully.
```

Tip: When you create a replication target path, the preferred practice is to choose the base directory of the target file system. If a different directory is used for replication, a failback writes data back with the different directory structure, and thus the recovered file system is not the same as before the failure.

On the source node, run **cfgreplfs** to set up the replication relationship for the source file system. For example, to set up the file system gfps0 to replicate to /ibm/gpfs1 on target system 12402779243267445246, run the command that is shown in Example 6-12.

Example 6-12 A cfgreplfs command example

```
[admin@st003.mgmt001st003 ~]# cfgreplfs gpfs0 12402779243267445246 /ibm/gpfs1
EFSSG0642I On the source file system used space is: 887,040, free space is:
8,053,504.
EFSSG0642I On the target file system used space is: 861,440, free space is:
133,888.
EFSSG0261I The replication declaration was updated successfully.
EFSSG1000I The command completed successfully.
```

Considerations:

- The default configuration specifies that the SONAS system creates snapshots at the source and target node. Although you can disable this default by specifying the **--nosourcesnap** or **--notargetsnap** options, it is not preferable to use them.
- Use the **--nosourcesnap** option only when all write activity is quiesced on the source file system to create a data consistent point. The write activity must remain quiesced until the asynchronous replication process completes to maintain a recoverable copy of the data on the target.
- Consider the **--nosourcesnap** option only when there are no inter-file relationships on the source file system that must have a consistent point-in-time copy of the group of files on the target system. Use the **--notargetsnap** option only if it is the last asynchronous replication to the target file system.

You can use the **--compress** option to specify that this relationship uses software compression to reduce the network bandwidth that is required for the transmission of changes. More node resources are required to do this compression. Consider these resources in addition to the network bandwidth savings. You can use the **--encryption** option and specify either strong or fast to designate which encryption cipher is used. The strong encryption cipher maps to the AES encryption standard; the fast cipher uses the arcfour encryption standard, which is not as strong as the AES standard but results in a much higher transfer rate.

6.8.5 Starting and stopping asynchronous replication

Run **startrepl** to start asynchronous replication. Run **stoprepl** to stop asynchronous replication. One scenario where these commands are useful is for satisfying the requirement to stop asynchronous replication before you do a SONAS code upgrade, and to start asynchronous replication after a SONAS code upgrade completes. You can also use the GUI to work with this function.

GUI usage

To work with this function in the management GUI, log on to the GUI and click **Copy Services** → **Replication**. Click the **Actions** tab, as shown in Figure 6-82.

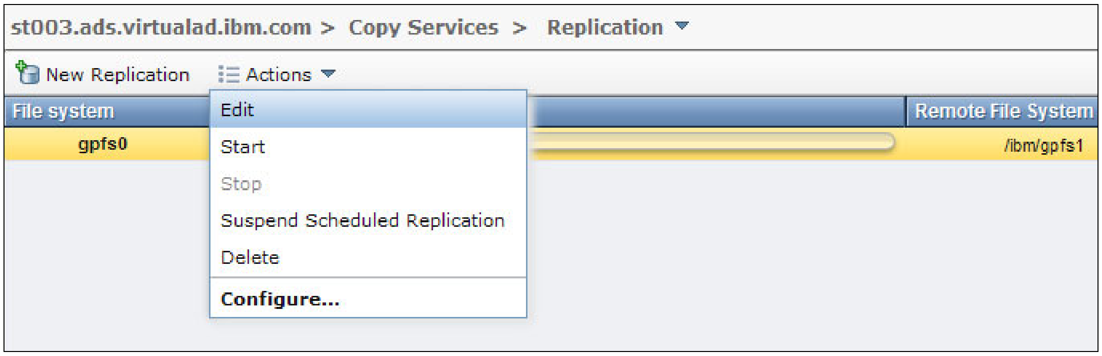


Figure 6-82 GUI to start and stop asynchronous replication

CLI usage

To start asynchronous replication, run **startrepl** by specifying the file system to be replicated. For example, to start a replication of the source system for the file system gpfs0, submit the command that is shown in Example 6-13.

Example 6-13 A startrepl command example

```
[admin@st003.mgmt001st003 ~]# startrepl gpfs0
EFSSG0062I The asynchronous replication has been successfully started with logID:
20111028223950.
EFSSG1000I The command completed successfully.
```

Considerations:

- Use the **--fullsync** option to request that all of the files in the source file system are checked against the target system to identify any changes. If the option is not specified, only new or changed files that are flagged as such on the source system are replicated.
- The **--fullsync** option extends the time that is required to do the replication. Normal recurring replications do not require the **--fullsync** option, which is used for the initial failback replication after a disaster that required the use of the target SONAS system as the primary, or if the target SONAS system has changed, for example, either a different SONAS system or a new target within the same source SONAS system.

To stop replication for a file system, from the source system, run **stoprepl** and specify the file system (see Example 6-14).

Example 6-14 A stoprepl command example

```
[admin@st003.mgmt001st003 ~]# stoprepl  gpfs0
EFSSG0288C The stop request for the given file system has already been accepted.
```

This command stops asynchronous replication gracefully on the source system for the specified file system. The **stoprepl** command is a graceful stop request. It waits for all of the currently running **rsync** processes to complete copying their current file list, and then stops the entire asynchronous replication. If the entire list of all of the files to replicate has already been sent to the **rsync** processes before the stop request is sent, the command waits for all replication to complete and then exits. The **--kill** option can be added if the graceful stop request is not being recognized because of error states, and stops the replication immediately.

6.8.6 Listing asynchronous replication

The asynchronous replication configuration can be listed by using the GUI or the CLI.

GUI usage

To work with this function in the management GUI, log on to the GUI and click **Copy Services** → **Replication**.

CLI usage

From the source system, run **lsreplcfg** to display source-to-target system replication configurations, as shown in Example 6-15.

Example 6-15 lsreplcfg command example

```
[root@st003.mgmt001st003 ~]# lsreplcfg
Source Cluster Name         Target Cluster Name Target Mgmt IP
st003.ads.virtualad.ibm.com st001.virtual.com   10.0.0.10
EFSSG1000I The command completed successfully.
[root@st003.mgmt001st003 ~]# lsreplcfg  -v
Source Cluster Name        Target Cluster Name Target Mgmt IP Node Pairs
Processes
st003.ads.virtualad.ibm.com st001.virtual.com   10.0.0.10
int001st003:10.0.0.112 1
EFSSG1000I The command completed successfully.
```

Specify the **-v** or**--verbose** option to display node pairs in an additional data column.

Listing asynchronous replication file system relationships

From the source system, run **lsreplfs** to list the file system relationships to the target paths and to list which snapshots are configured to be created, as shown in Example 6-16.

Example 6-16 List async replication file system relationships

```
[root@st003.mgmt001st003 ~]# lsreplfs
filesystem target path snapshots     rules compress encryption
gpfs0      /ibm/gpfs1  source&target       no
EFSSG1000I The command completed successfully.
```

Listing replication targets

From the target system, run **lsrepltarget** to display the configured replication target paths and associated source system IDs, as shown in Example 6-17.

Example 6-17 List replication targets

```
[admin@st001.mgmt002st001 ~]# lsrepltarget
SourceClusterId      TargetPath
12402849607718647729 /ibm/gpfs1
EFSSG1000I The command completed successfully.
```

Listing currently running and previous replications

From the source system, run **lsrepl** to display currently running and previous replication operation information for that source system, as shown in Example 6-18.

Example 6-18 List current and previous replications

```
[admin@st003.mgmt001st003 ~]# lsrepl
filesystem log Id         status  description time
gpfs0      20111028223950 STARTED initiated   10/28/11 10:39 PM
EFSSG1000I The command completed successfully.
```

Listing asynchronous replication results

From the source system, run **showreplresults** to display the results of a replication. Either the **--errors** parameter must be specified to display errors only, or the **--logs** parameter must be specified to display the log. The **--loglevel** parameter can optionally be specified with the **--logs** parameter to limit the log display entries. If not specified, the default log level is 1. The log ID and file system name must always be specified. The log ID can be determined by running **lsrepl**. Example 6-19 lists the replication errors for the specified parameters.

Example 6-19 List asynchronous replication results

```
[root@st003.mgmt001st003 ~]# showreplresults gpfs0 -e 20111028223950
File: cnreplicate.log.20111028223950_gpfs0
Replication ID: gpfs0
Node: src mgmt001st003
------------------------------------
------------------------------------
File: async_repl.log
Replication ID: gpfs0
Node: src mgmt001st003
------------------------------------
------------------------------------
File: scan.log
Replication ID: gpfs0
Node: src mgmt001st003
------------------------------------
------------------------------------
File: async_repl.log.1
Replication ID: gpfs0
Node: src int001st003
------------------------------------
------------------------------------
File: async_repl_remote.log
Replication ID: gpfs0
```

```
Node: dest mgmt002st001
------------------------------------
------------------------------------
====================================
 Log summary
====================================
2011-10-28 22:40:26+02:00 S mgmt001st003 cnreplicate [L1]
================================================================================
2011-10-28 22:40:26+02:00 S mgmt001st003 cnreplicate [L1] Performance summary of
Rsyncs
2011-10-28 22:40:26+02:00 S mgmt001st003 cnreplicate [L1]
--------------------------------------------------------------------------------
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1] TOTAL:
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]   transfer : 22 B (22
bytes)
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]   file     : 0 B (0
bytes)
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]   async_repl elapsed
time  : 0 day(s) 0 hours 0 mins 8 secs (8 sec)
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]
throughput  : 0 MB/sec
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]   cnreplicate elapsed
time : 0 day(s) 0 hours 0 mins 35 secs (35 sec)
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]
throughput : 0 MB/sec
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L1]
================================================================================
2011-10-28 22:40:27+02:00 S mgmt001st003 cnreplicate [L0] Exiting overall
replication process with 0 (success).
------------------------------------
EFSSG1000I The command completed successfully.
```

6.8.7 Removing and changing the asynchronous replication configuration

To change the file system relationship or the target directory, the relationship or target directory must be removed with the appropriate CLI command and then readded. When you remove a target directory, ensure that the file system relationship was removed previously. You can also use the GUI to work with this function.

GUI navigation

To work with this function in the management GUI, log on to the GUI and click **Copy Services** → **Replication**.

CLI usage

To remove an asynchronous replication source to target file system relationship with the **rmreplfs** command, from the source system, run **rmreplfs**, specifying the source file system. For example, to remove the replication relationship for the file system `gpfs0`, run the command that is shown in Example 6-20.

Example 6-20 rmreplfs command example

```
[admin@st003.mgmt001st003 ~]# rmreplfs gpfs0
EFSSG0581I You are about to delete the replication relationship and its change
tracking information. It is recommended to delete the change tracking information
unless a new target for this relationship expects only future delta changes.
Do you really want to perform the operation (yes/no - default no):yes
EFSSG0242I Deleted 1 file system entries for replication.
EFSSG1000I The command completed successfully.
```

Run **rmrepltarget** to remove a target directory from asynchronous replication. The file system relationship must first be removed by running **rmreplfs** before you can remove the replication target. To remove the replication target directory, from the target system, run **rmrepltarget**, specifying the target directory, which is not available for asynchronous replication after the command completes. For example, to make the `/ibm/gpfs0` directory unavailable for asynchronous replication, run the command that is shown in Example 6-21.

Example 6-21 Remove a target directory from asynchronous replication

```
[admin@st001.mgmt002st001 ~]# rmrepltarget /ibm/gpfs1
Do you really want to perform the operation (yes/no - default no):yes
EFSSG0267I Replication target entry removed.
EFSSG1000I The command completed successfully.
```

Tip: When removing a target path with the **rmrepltarget** command, you are prompted whether the contents of the specified target directory are to be deleted. You can use the **--clean** option to also delete the directory and its contents.

6.8.8 Asynchronous replication disaster recovery

Recovering a file system by using asynchronous replication requires that a replication relationship from the target site to the source is configured and started.

About this task

After the source site fails, you must set the target site as the new source site, replicating back to the source.

Procedure

Where the previous replication relationship was Site A replicating to Site B, configure the asynchronous replication and reverse the source and target site information so that Site B now replicates to Site A. For more details, see 6.8.2, “Configuring asynchronous replication” on page 446 and transpose the source and target information.

Start the replication that is configured in step 1 by running **startrepl**, and specify the **--fullsync** parameter. For more information, see 6.8.5, “Starting and stopping asynchronous replication” on page 454.

If the amount of data to be replicated back to the Site A is large, multiple replications from Site B to Site A might be required until modifications to Site B can be suspended to do a final replication to catch Site A backup. Do *not* use the **--fullsync** option for these incremental replications.

When data is verified as having been replicated accurately on Site A, then Site A can be reconfigured as the primary site. Remove any replication tasks that are going from Site B to Site A by running **rmtask**.

6.8.9 Cleaning up asynchronous replication results

Run **cleanrepl** to clean up previous asynchronous replication results.

Procedure

Run **lsrepl** to determine the log ID of the asynchronous replication result to remove from the system (see Example 6-22).

Example 6-22 Sample lsrepl command output

```
filesystem log Id         status     description time
gpfs0      20110826183559 RUNNING    The replication task is 90% complete (910 out
of 1007 files)                                          8/26/11 6:36 PM
gpfs1      20110826180415 SUCCESSFUL A source or destination node was unreachable
during the replication and a failover of the node occurred. 8/26/11 6:04 PM
gpfs1      20110826183554 RUNNING    7/8 Replication task for asynchronous
replication process done                                        8/26/11 6:36 PM
gpfs2      20110826180420 SUCCESSFUL A source or destination node was unreachable
during the replication and a failover of the node occurred. 8/26/11 6:04 PM
gpfs2      20110826183556 RUNNING    7/8 Replication task for asynchronous
replication process done                                        8/26/11 6:36 PM
    EFSSG1000I The command completed successfully.
```

Removing log files

To remove log files, complete the following steps:

- To remove a single log file, run **cleanuprepl** with the **--logfiles** option, specifying the file system name and the log ID, separated by the colon character.

  For example, to remove the result from 5/26/2010 in Example 6-22, run the following command:

  ```
  # cleanuprepl --logfiles gpfs1:20110826180415
  ```

- To remove all of the log files for a file system, run **cleanuprepl** with the **--logfiles** option, specifying the file system followed by a space character and the parameter value `all`.
- To remove all but the most recent n number of log files for a file system, run **cleanuprepl** with the **--logfiles** option, specifying the file system and the number of most recent log files to retain, separated by the colon character.
- To remove an asynchronous replication lock, run **cleanuprepl** with the **--clearlock** option, specifying the file system from which to remove the lock.

> **Tip:** A lock is placed on a replication relationship when there is a condition that requires intervention to correct. Continued asynchronous replication attempts before the condition is corrected might hinder the corrective action.

6.8.10 Scheduling an established asynchronous replication task

This section describes how to schedule an asynchronous replication task so that it is submitted on a regular schedule.

Scheduling an asynchronous replication task that depends on a previously defined relationship that was established on the source side by running **cfgreplfs**. That definition established the source-to-target relationship between the file systems and the optional parameters for how the replication is done.

> **Tip:** If a replication for a specified file system is still in progress when the scheduler triggers a new replication task for that file system, the new replication request fails.

GUI navigation

To work with this function in the management GUI, log on to the GUI and click **Copy Services** → **Replication**.

CLI usage

Run **mkrepltask** to schedule an asynchronous replication task for a previously defined relationship that was established on the source side by running **mkrepltask**. Specify the source file system name for which the replication relationship was defined, and specify the schedule on which the task is to be submitted by using parameters for minute, hour, dayOfWeek, dayOfMonth, and month, as described in the man page for the **cfgreplfs** command. The following example creates a **cron** job task for submitting an asynchronous replication task for the file system gpfs0 every hour on the hour.

> **Tip:** Time is designated in the 24-hour format. The options **--dayOfWeek** and **--dayOfMonth** are mutually exclusive.

```
# mkrepltask gpfs0 --minute 0
```

Run **rmrepltask**, specifying the file system to remove a task that submits an asynchronous replication task on its defined schedule. For example:

```
# rmrepltask gpfs0
```

To specify a system for either command, use the **-c** or **—-cluster** option and specify either the system ID or the system name. If the **-c** and **—-cluster** options are omitted, the default system, as defined by the **setcluster** command, is used.

6.9 Asynchronous replication limitations

Asynchronous replication allows replication of file systems across long distances or to low-performance, high-capacity storage systems.

6.9.1 Limitations for disaster recovery

Keep these limitations in mind when you are using the asynchronous replication function:

- The asynchronous replication relationship is configured as a one-to-one relationship between the source and target.
- The entire file system is replicated in asynchronous replication. Although you can specify paths on the target system, you cannot specify paths on the source system.
- The source and target cannot be in the same system.
- Asynchronous replication processing on a file system can be impacted by the number of migrated files within the file system. Asynchronous replication on a source file system causes migrated files to be recalled and brought back into the source file system during the asynchronous replication processing.
- File set information that is on the source system is not copied to the target system. The file tree on the source is replicated to the target, but the fact that it is a file set is not carried forward to the target system's file tree. File sets must be created and linked on the target system before initial replication because a file set cannot be linked to an existing folder.
- Quota information is also not carried forward to the target system's file tree. Quotas can be set after initial replication as required, by using quota settings from the source system.
- Active Directory (AD) Only, and AD with NIS that uses SONAS internal UID and GID mapping are not supported by asynchronous replication because the mapping tables in the SONAS system clustered trivial database (CTDB) are not transferred by asynchronous replication. If asynchronous replication is used, the user ID mapping must be external to the SONAS system.

Networking

For the first occurrence of running asynchronous replication, you might want to consider transporting the data to the remote site physically at first and have replication take care of changes to the data. Asynchronous replication is no faster than a simple copy operation.

Ensure that adequate bandwidth is available to finish replication on time.

Disk I/O

There is no mechanism for throttling asynchronous replication. GPFS balances the load between asynchronous replication and other processes.

Path names

Source and target root paths that are passed as parameters must not contain a space, single or double quotation mark, “`”, “:”, “\”, “\n”, or any white-space characters.

6.9.2 Considerations for disaster recovery

The ability to continue operations in the face of a regional disaster is primarily handled through the async replication mechanism of the SONAS appliance. Again, async replication allows for one or more file systems within an SONAS file name space to be defined for replication to another SONAS system over the customer network infrastructure. As the name async implies, files that are created, modified, or deleted at the primary location are propagated to the remote system some time after the change of the file in the primary system.

The async replication process looks for changed files in a defined file system of the source SONAS since the last replication cycle was started against it, and using IBM hardened/enhanced versions of the **rsync** tools to move efficiently only the changed portions of a file from one location to the next. In addition to the file contents, all extended attribute information about the file is also replicated to the remote system.

Async replication is defined in a single direction, such that one site is considered the source of the data, and the other is the target, as illustrated in Figure 6-83. The replica of the file system at the remote location must be used in a Read-Only mode until it is needed to become usable if there is a disaster.

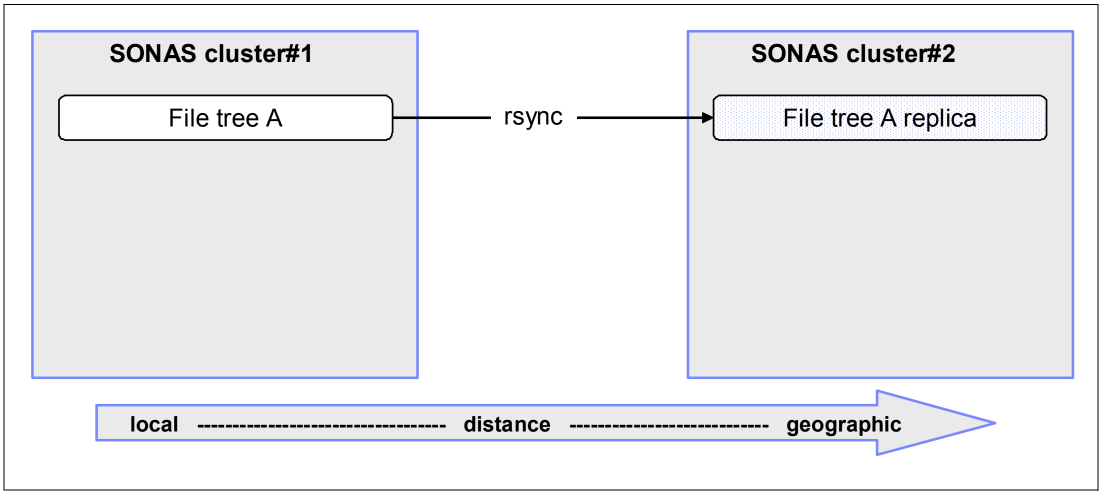

SONAS cluster#1
File tree A
rsync
SONAS cluster#2
File tree A replica
local ---------------------------------- distance --------------------------- geographic

Figure 6-83 Async replication source and target

The SONAS Interface nodes are defined as the elements for doing the replication functions. When you are using async replication, the SONAS system detects the modified files from the source system, and moves only the changed contents from each file to the remote destination to create an exact replica. By moving only the changed portions of each modified file, the network bandwidth is used efficiently.

The file-based movement allows the source and destination file trees to be of differing sizes and configurations if the destination file system is large enough to hold the contents of the files from the source.

Async replication allows all or portions of the data of a SONAS system to be replicated asynchronously to another SONAS system, and if there is an extended outage or loss of the primary system, the data that is kept by the backup system is accessible in R/W mode by the customer applications. Async replication also offers a mechanism to replicate the data back to the primary site after the outage or new system is restored.

The backup system also offers concurrent R/O access to the copy of the primary data testing/validation of the disaster recovery mirror. The data at the backup system can be accessed by all of the protocols in use on the primary system. You can take an R/W snapshot of the replica, which can be used to allow for full function disaster recovery testing against your applications. Typically, the R/W snapshot is deleted after the disaster recovery test concludes.

File shares that are defined at the production site are not automatically carried forward to the secondary site, and must be manually redefined by the customer for the secondary location, and these shares must be defined as R/O until such time that they need to do production work against the remote system in full R/W, for example, for business continuance in the face of a disaster. Redefinition to R/W shares can be done by using the CLI or GUI.

The relationship between the primary and secondary site is a 1:1 basis: one primary and one secondary site. The scope of an async replication relationship is on a file system basis. Preferred practices must be followed to ensure that the HSM systems are configured and managed to avoid costly performance impacts during the async replication cycles that can be because the file was migrated to auxiliary storage before it is replicated and must to be recalled from auxiliary storage for replication to occur.

Because these conditions can become complex for many customers, it is a preferred practice to request an IBM Services consultation for preparing a complete disaster recovery solution for your cluster and remote site replication scenarios.

User authentication and mapping requirements

Async replication requires coordination of the customer's Windows SID domain information to the UID/GID mapping that is internal to the SONAS cluster because the ID mapping from the Windows domain to the UNIX UID/GID mapping is not exchanged between the SONAS systems. As the mappings are held external to the SONAS system in one of LDAP, NIS, or with AD with Microsoft SFU, the external customer servers hold mapping information and must have coordinated resolution between their primary and secondary sites.

Async replication is only usable for installations that use LDAP, NIS, or AD with the SFU extensions. Standard AD, without SFU, is not sufficient. The reason is that async replication can move only the files and their attributes from one site to the next. Therefore, the UID and GID information which GPFS maintains is carried forward to the destination. However, Active Directory supplies only a SID (Windows authentication ID), and the CIFS server inside of the SONAS maintains a mapping table of this SID to the UID/GID that is kept by GPFS. This CIFS server mapping table is not carried forward to the destination SONAS.

Therefore, when users attempt to talk to the SONAS at the remote site, they do not have a mapping from their Active Directory SID to the UID/GID of the destination SONAS, and their authentication does not work properly, for example, users might map to the wrong user's files.

LDAP, NIS, and AD with SFU maintain the SID to UID/GID mapping external to the SONAS, and therefore, if their authentication mechanism is visible to the SONAS at the source and the destination site, they do not have a conflict with the users and groups.

The following assumptions are made for the environment that supports async replication:

- One of the following authentication mechanisms: either an LDAP or AD with SFU environment that is resolvable across their sites, or is mirrored and consistent across their sites such that the SONAS at each site is able to authenticate from each location.
- The authentication mechanism is the same across both locations.
- The time synchronization across both sites is sufficient to allow for successful authentication.

Async replication operation

The primary function of the async replication is to make a copy of the customer data, including file system metadata, from one SONAS system to another over a standard IP network. The design also attempts to minimize network bandwidth usage by moving only the portions of the file that were modified to the destination system.

Here are the primary elements of the async replication operation:

- The SONAS code does key replication tasks, such as scanning for changed files, removing files that are deleted at the source on the destination, and recovery and retry of failures.
- The UNIX `rsync` replication tool compares the source/destination files for differences, and moves and writes only the delta information about the destination to ensure that the destination matches the source.

Async replication considerations

This section highlights key considerations of async replication design and operation that must be understood:

- Replication is done on a file system basis, and file sets on the source SONAS cluster do not retain the file set information that is on the destination SONAS cluster. The file tree on the source is replicated to the destination, but the fact that it is a file set, or any quota information, is not carried forward to the destination cluster's file tree.
- The path to source and target root paths that are passed as parameters must not contain a space, single or double quotation mark characters, “`”, “:”, “\”, “\n”, or any white-space characters. The underlying paths within the directory tree being replicated are allowed to have them.
- The network bandwidth that is required to move large amounts of data, such as the first async replication of a large existing file system or the failback to an empty SONAS after a disaster, takes much time and network bandwidth to move the data. Other means of restoring the data, such as physical restore from a backup, is a preferred means of populating the destination cluster to reduce greatly the restore time and reduce the burden on the network.
- For disk I/O, the I/O performance is driven by GPFS and its ability to load balance across the nodes that are participating in the file system. Async replication performance is driven by metadata access for the scan part, and customer data access for the `rsync` movement of data. The number and classes of disks for metadata and customer data are an important part of the overall performance.
- File set information that is on the source system is not copied to the target system. The file tree on the source is replicated to the target, but the fact that it is a file set is not carried forward to the target system's file tree. File sets must be created and linked on the target system before initial replication because a file set cannot be linked to an existing folder.
- Quota information is also not carried forward to the target system's file tree. Quotas can be set after initial replication as required, by using quota settings from the source system.
- Active Directory (AD) Only, and AD with NIS that uses SONAS internal UID/GID mapping, are not supported by asynchronous replication because the mapping tables in the SONAS system clustered trivial database (CTDB) are not transferred by asynchronous replication. If asynchronous replication is used, the user ID mapping must be external to the SONAS system.
- Tivoli Storage Manager HSM stub files are replicated as regular files, and an HSM recall is done for each file, so they can be omitted by using the CLI.

HSM in an async replication environment

Async replication can coexist with SONAS file systems that are being managed by the Tivoli Storage Manager HSM software, which moves files that are held within a SONAS file system to and from an auxiliary storage media such as tape.

The key concept is that the Tivoli Storage Manager HSM client hooks into the GPFS file system within the SONAS to replace a file that is stored within the SONAS with a *stub file,* which makes it appear to the user that the file still exists in the SONAS GPFS file system on disk after it is moved to the auxiliary storage device. When the file is accessed, the Tivoli Storage Manager HSM client suspends the GPFS request for data within the file until it retrieves the file from the auxiliary storage device and replaces it in the SONAS primary storage. Now, the file can be accessed directly again by the users through the SONAS.

The primary function of this move is to allow for the capacity of the primary storage to be less than the amount of data it is holding by using the secondary (cheaper and slower) storage to retain the overflow of data. The following subsections describe key implications for using the HSM functions with file systems that are being backed up for disaster recovery purposes with async replication.

Source and destination primary storage capacities

The primary storage on the source and destination SONAS systems must be reasonably balanced in terms of capacity. Because HSM allows for the retention of more data than primary storage capacity and async replication is a file-based replication, planning must be done to ensure that the destination SONAS system has enough storage to hold the entire contents of the source data (both primary and auxiliary storage) contents.

HSM management at destination

If the destination system uses HSM management of the SONAS storage, enough primary storage at the destination must be considered to ensure that the change delta to be replicated over into its primary storage is part of the DR process. If the movement of the data from the destination location's primary to auxiliary storage is not fast enough, the replication process can outpace this movement, which causes a performance bottleneck in completing the disaster recovery cycle.

Therefore, the capacity of the destination system to move data to the auxiliary storage must be sufficiently configured to ensure that enough data was pre-migrated to the auxiliary storage to account for the next async replication cycle. The amount of data to be replicated can be achieved without waiting for movement to auxiliary storage. For example, enough Tivoli Storage Manager managed tape drives must be allocated and operational, and there must be enough media to ensure that enough data can be moved from the primary storage to tape to ensure that enough space is available for the next wave of replicated data.

Replication intervals with HSM at the source location

Planning must be done to ensure that the frequency of the async replication is such that the changed data at the source location is still in primary storage when the async process is initiated. This requires a balance with the source primary storage capacity, the change rate in the data, and the frequency of the async replication scan intervals.

If changed data is moved from primary to auxiliary storage before the async process can replicate it to the destination, the next replication cycle must recall it from the auxiliary storage back to the primary to copy it to the destination. The number of files that must be recalled back into primary storage and the duration to move them back into primary storage directly impacts the time that the async process needs to finish replicating.

SONAS async replication configurations

For business continuance in a disaster, SONAS supports asynchronous replication between two SONAS systems in a 1:1 relationship. The SONAS systems are distinct from one another, such that they are independent clusters with a non-shared InfiniBand infrastructure, separate interface, storage, and Management nodes, and so on. The connectivity between the systems is through the customer network between the customer-facing network adapters in the Interface nodes. The local and remote SONAS systems do not require the same hardware configuration in terms of nodes or disks; only the space at the secondary site must be enough to contain the data that is replicated from the primary site.

The systems must be able to route network traffic between each other by using the customer-supplied IP addresses or fully qualified domain names on the Interface nodes.

Async replication in a single direction

There are two primary disaster recovery topologies for a SONAS system. The first is where the second site is a standby disaster recovery site, such that it maintains a copy of file systems from the primary location only. It can be used for testing purposes, for continuing production in a disaster, or for restoring the primary site after a disaster.

Figure 6-84 illustrates the relationship between the primary and secondary sites for this scenario.

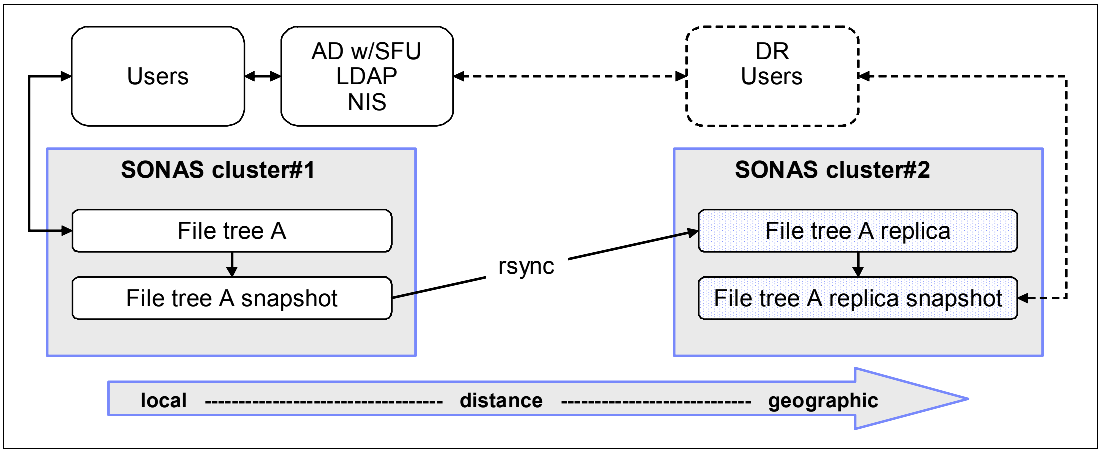

Users
AD w/SFU
LDAP
NIS
DR
Users
SONAS cluster#1
File tree A
File tree A snapshot
rsync
SONAS cluster#2
File tree A replica
File tree A replica snapshot
local ---------------------------------- distance -------------------------- geographic

Figure 6-84 Async replication with a single active direction

Async replication in two active directions

The second scenario, which is shown in Figure 6-85 on page 467, is when the second site exports shares of a file system in addition to holding mirrors of a file tree from the primary site. This scenario is when the SONAS at both sites is used for production I/O, in addition to being the target mirror for the other SONAS system's file structure. This mirroring can be in both directions, such that both SONAS systems have their own file trees, in addition to the having the file tree of the other, or both have their own file tree, and only one has the mirror of the other.

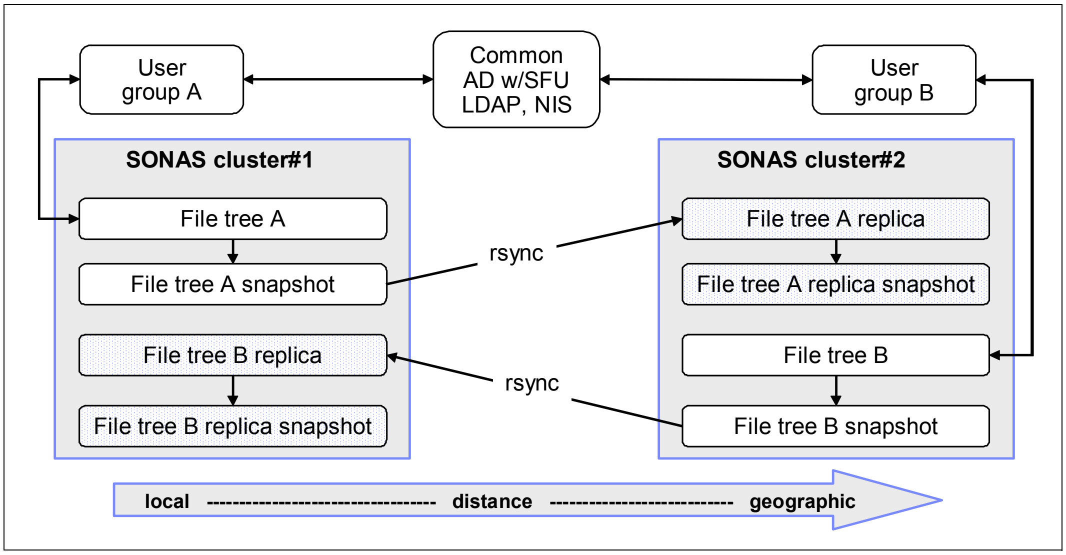

User group A
Common AD w/SFU LDAP, NIS
User group B
SONAS cluster#1
File tree A
File tree A snapshot
File tree B replica
File tree B replica snapshot
rsync
rsync
SONAS cluster#2
File tree A replica
File tree A replica snapshot
File tree B
File tree B snapshot
local ---------------------------------- distance --------------------------- geographic

Figure 6-85 Bidirectional async replication and snapshots

6.9.3 Async replication process

The async replication process has the following main steps:

1. Create a local snapshot of the source file system.
2. Scan and collect a full file path list with the stat information.
3. Build a new, changed, and deleted file and directory list, including hard links.
4. Distribute **rsync** tasks among defined nodes that are configured to participate in async replication.
5. Remove deleted files and create hard links on the remote site.
6. Create a remote snapshot of a replica file system if indicated in the async command.
7. Remove a local snapshot if created from a specified async command.

Async replication tools, by default, create a local snapshot of the file tree that is being replicated, and use the snapshot as the source of the replication to the destination system. It is the preferred method, as it creates a well-defined point-in-time of the data that is being protected against a disaster. The scan and resulting **rsync** commands must be run against a stable, non-changing file tree, which provides a known state of the files to be coordinated with the destination. Async replication does have a parameter that tells the system to skip the creation of the snapshot of the source, but the scan and following rsync are done on changing files. This configuration has the following implications:

- Inconsistent point-in-time value of the destination system, as changes to the tree during the async process might cause files that are scanned and replicated first to be potentially from an earlier state than the files later in the scan.
- Files that are changed after the scan cycle takes place are omitted from the replication.
- A file can be in flux during the **rsync** movement.

The name of the snapshot is based on the path to the async replication directory on the destination system, with the extension `_cnreplicate_tmp` appended to it. For example, if the destination file tree for async is `/ibm/gpfsjt/async`, then the resulting snapshot directory is created in the source file system:

```
/ibm/gpfs0/.snapshots/ibm_gpfsjt_async_cnreplicate_tmp
```

These snapshots are alongside any other snapshots that are created by the system as a part of user request. The async replication tool ensures that it operates only on snapshots that it created with its own naming convention. These snapshots count towards the 256 snapshot limit per file system, and can therefore be accounted for with the other snapshots that are used by the system. After the successful completion of async replication, the snapshot that is created in the source file system is removed.

After the completion of the async replication, a snapshot of the file system that contains the replica target is done. The name of the snapshot is based on the destination path to the async replication directory with the `extension _cnreplicate_tmp` appended to it.

As with source snapshots, these snapshots are alongside any other snapshots that are created by the system as a part of the user request. The async replication tool ensures that it operates only on snapshots that it created with this naming convention. These snapshots count towards the 256 snapshot limit per file system, and can therefore be accounted for with the other snapshots that are used by the system.

Replication frequency and recovery point objective considerations

To ensure that data in the remote SONAS sites is as current as possible and has a small recovery point objective (RPO), it seems natural to run the async replication as frequently as possible. The frequency of the replication must account for a number of factors:

- The change rate of the source data
- The number of files that are contained within the source file tree
- The network between SONAS systems, including bandwidth, latency, and sharing aspects
- The number of nodes that are participating in the async replication

A replication cycle must complete before a new cycle can be started. The key metric in determining the time that it takes for a replication cycle to complete is the time that it takes to move the changed contents of the source to the destination based on the change rate of the data and the network capabilities.

For example, a 10 TB file tree with a 5% daily change rate must move 500 GB of data over the course of a day (5.78 MBps on average over the day). Daily change rates are probably not consistent over the 24-hour period, and must be based on the maximum change rate per hour over the day. The required network bandwidth to achieve it is based on the RPO. With an RPO of 1 hour, enough network bandwidth is needed to ensure that the maximum change rate over the day can be replicated to the destination in under an hour.

Part of the async replication algorithm is the determination of the changed files, which can be a CPU- and disk-intensive process that must be accounted for as part of the impact. Continually running replications below the required RPO can cause undue impact to other workloads that use the system.

Async replication scenarios

Before doing async replication, verify that the following conditions are met:

- Ensure that you have consistent Active Directory with SFU or LDAP authentication across the sites that are participating in the disaster recovery environment.
- The mapping of users across both sites must be consistent from the Windows SID Domain to UNIX UID and GID.
- Ensure that there is sufficient storage at the destination for holding a replica of the source file tree and its associated snapshots.
- The network between the source and the destination must support SSH connections and **`rsync`** operations.

- The network between the source and destination Interface nodes need sufficient bandwidth in to account for the change rate of data that is being modified at the source between replicas, and the required RTO/RPO objectives to meet disaster recovery criteria.
- Define an async relationship between the Interface nodes of the source and destination, define the target file system, and create the source/destination file system relationship with the **`cfgreplf`**, **`mkrepltarget`**, and **`cfgreplfs`** commands.

Doing async replications

Here are the considerations and actions to protect the data against an extended outage or disaster at the primary location. The protection is accomplished by carrying out async replications between the source and destination systems.

- Perform async replication between source and destination SONAS systems. Replication can be carried out manually or by scheduled operation.
  - Manually run **`startrepl`** to initiate an async replication cycle against the directory tree structure that is specified in the command for the source and destination locations.
  - Define an automated schedule for the async replication to be carried out by the system on defined directory tree structures.
- Monitor the stats of the current and previous async replication processes to ensure a successful completion. Async replication raises a CIM indication to the Health Center, which can be configured to generate SMTP and SNMP alerts.

Disaster recovery testing

Define shares as read-only (R/O) to the destination file tree for accessing file resources at destination.

- Modification of the destination file tree as part of the validation of data or testing DR procedures must not be done. Changes to the destination file tree are not tracked, and cause the destination to differ from the source.
- FTP, HTTP, and SCP shares cannot be created R/O, and are a risk factor in being able to modify the target directory tree. Modifications to the target directory tree are not tracked by the DR recovery process, and can lead to discrepancies between the source and target file tree structures.

You must access the disaster recovery location file structure as read-only. You must create the shares at the destination site, which are used to access the data from the disaster recovery location.

Business continuance

The steps for enabling the recovery site involve the following major components:

1. Perform a baseline file scan of the file tree replica that is used as the target for the async replication.
2. Define shares/exports to the file tree replica.
3. Continue production operation against the remote system.

The baseline scan establishes the state of the remote system files that was last received by the production site, which tracks the changes that are made from this point forward. For the configuration where the secondary site was strictly only a backup for the production site, establishing the defined shares for the replica to enable it for production is the primary consideration. Figure 6-86 illustrates this scenario.

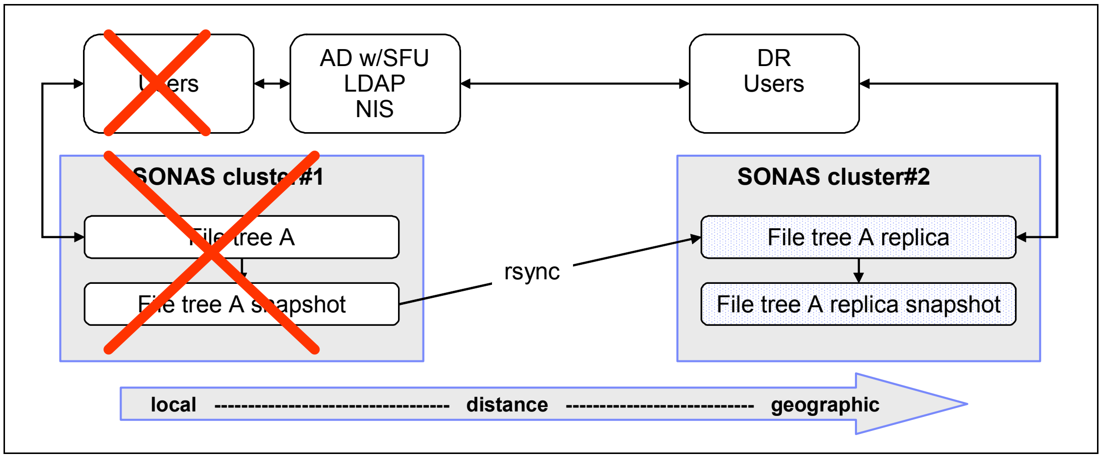

Users
AD w/SFU
LDAP
NIS
DR
Users
SONAS cluster#1
File tree A
File tree A snapshot
rsync
SONAS cluster#2
File tree A replica
File tree A replica snapshot
local ---------------------------------- distance --------------------------- geographic

Figure 6-86 Business continuance - active-passive production site failure

If the second site contains its own production file tree in addition to replicas, then the failure also impacts the replication of its production file systems back to the first site, as illustrated in Figure 6-87.

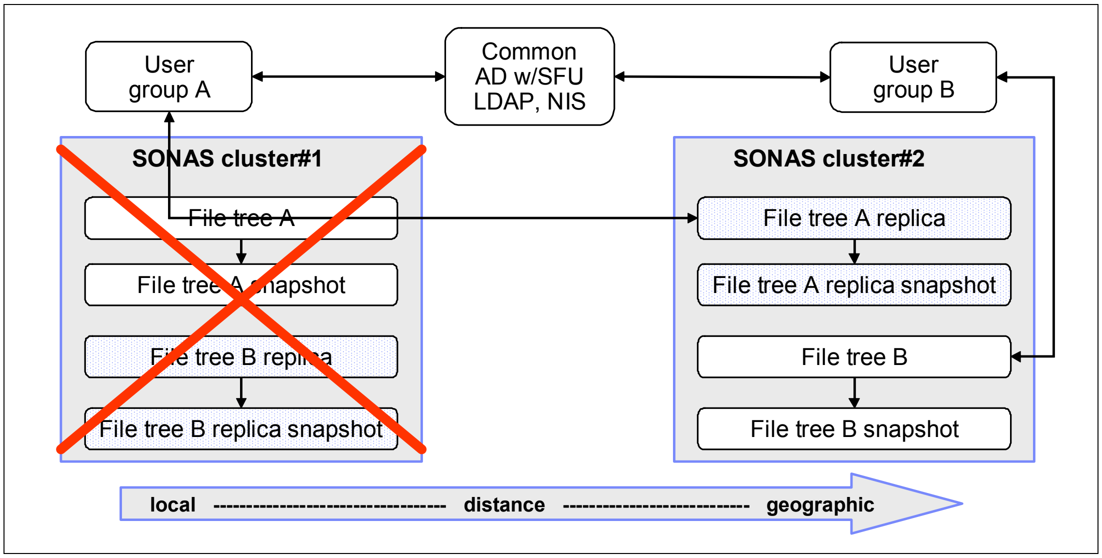


Figure 6-87 Business continuance - active-active production site failure

Here are the steps to recover at the disaster recovery site:

1. Run **`startrepl`** with the **`-S`** parameter to run a scan only on the destination system to establish a point in time of the current file tree structure. This command allows the system to track changes to the destination file tree to help with delta file update back to the original production system.
2. Define shares to destination file systems as R/W by running **`mkexport`**, or change existing R/O shares that are used for validation and testing to R/W by running **`chexport`**.
3. Proceed with R/W access to data at the disaster recovery location against the file tree.

Recovery from a site disaster

The recovery of a SONAS system at a site after an extended outage depends on the scope of the failure. The following primary scenarios are from the resulting outage:

- The failing site was lost, and no data was retained.
- The failing site had an extended outage, but data was retained.
- The failing site had an extended outage, and an unknown amount of data was lost.

Recovery from a data corruption disaster

The recovery of a SONAS system from a data corruption disaster is most likely to be an extended outage if recent snapshot recovery testing does not yield an uncorrupted state.

If the underlying storage becomes corrupted for any reason, then it is safe to assume that all data is corrupted (local or remotely replicated).

Assumptions are as follows:

- The failing site was lost, and no data that was retained is usable.
- The only way to recover that data is from backup.
- For a large cluster, the recovery time objectives stretch to the technology restoration capabilities of the underlying backup and restore technology.
- In most cases, this event affects the file system data and metadata, but not necessarily the cluster configuration.

If only a file or directory space is corrupted, restore from tape is much faster and easier to manage from a time requirement perspective.

For these reasons, it is important to consider not only a replication strategy but a snapshot and backup solution with your SONAS installation.

Recovery to an empty SONAS system

If the failing site was lost, the recovery must take place against an empty system, either a new site location with a new SONAS system or the previous SONAS system that was restored but contains none of the previously stored data. For the purposes of this publication, assume that the SONAS system is installed, configured with IP addresses, and the connections to authentication servers are complete so that you can bring the system to an online state.

The recovery steps for an active-passive configuration are as follows:

1. Configure the async replication policies such that the source to destination relationship moves from the secondary site to the new primary site. For the new primary site, you must enable it to be the destination of an async relationship and create a target file tree for async replication. For the secondary site, you configure it as an async source and define the async relationship with its file tree as the source and the one configured on the new primary site as the target.
2. Perform async replication back to the new primary site. It can take a long time to transfer the entire contents electronically. The time is based on the amount of data and the network capabilities.
3. Halt production activity to the secondary site and do another async replication to ensure that primary and secondary sites are identical.
4. Perform a baseline scan of the primary site file tree.
5. Define exports or shares to the primary site.
6. Begin production activity to the primary site.

7. Configure async replication of the source and destination nodes to direct replication back from the new primary site to the secondary site.
8. Resume the original async replication of the primary to the secondary site as previously defined before the disaster.

Figure 6-88 illustrates disaster failback to an empty SONAS.

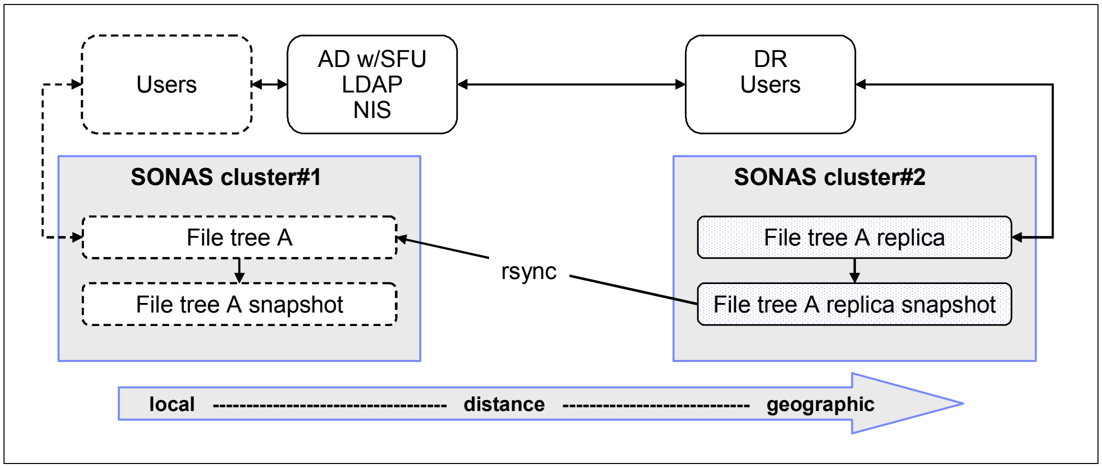


Figure 6-88 Disaster failback to an empty SONAS

In the scenario where the second site was used for both active production usage and as a replication target, the recovery is as illustrated in Figure 6-89.

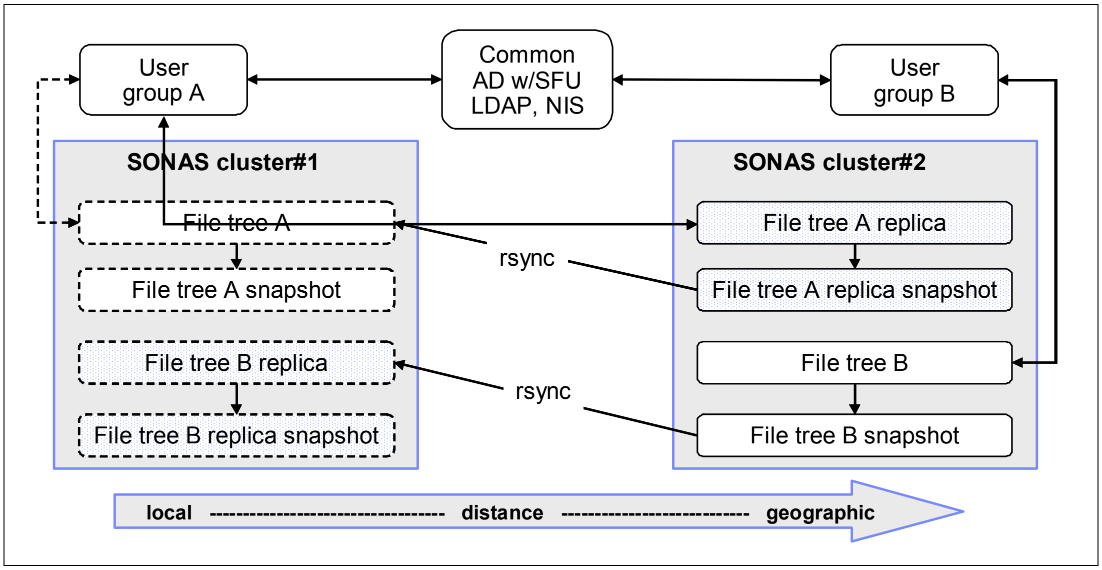


Figure 6-89 Failback to an empty SONAS in an active-active environment

The loss of the first site also lost the replica of the second's site file systems, which needs to be replicated back to the first site. The recovery steps for an active-active configuration are outlined as follows:

1. Configure the async replication policies such that the source to destination moves from the secondary site to the new primary site for file tree A.
2. Perform the async replication with the “full” replication parameter back from file tree A to the new primary site; the time to transfer the entire contents electronically can be a long time, based on the amount of data and the network capabilities.
3. Halt production activity to the secondary site and do another async replication to ensure that the primary and secondary sites are identical.
4. Perform a baseline scan of file tree A at site 1.
5. Define exports and shares to file tree A at site 1.
6. Begin production activity to file tree A at site 1.
7. Configure the async replication of the source and destination nodes to direct replication back from the new primary site to the secondary site for file tree A.
8. Resume the original async replication of file tree A from the new primary site to the secondary site.
9. For the first async replication of file tree B from the secondary site to the new primary site, ensure that the *full* replication parameter is used. It ensures that all contents from file tree B are sent from the secondary site to the new primary site.

6.10 Disaster recovery methods

To rebuild a SONAS cluster in the case of a disaster that caused the whole SONAS cluster to become unavailable, two types of data are required:

- The data that is contained on the SONAS cluster
- The SONAS cluster configuration files

The data that is contained in the SONAS cluster can be backed up to a backup server, such as Tivoli Storage Manager or another supported NDMP backup product. Another option if the NSD and underlying storage remain intact is to recover file systems from snapshots, and finally to recover the data from a remote intracluster replica of the data to a remote cluster or file server.

The cluster configuration data can be backed up with the **backupmanagmentnode** command.

6.10.1 Backing up SONAS configuration information

SONAS configuration information can be backed up with the **backupmanagementnode** command. This command makes a backup from the local Management node, where the command is running, and stores it on another remote host or server.

Use this command to back up one or more of the following SONAS configuration components:

- auth
- callhome
- cimcron
- ctdb
- derby

- misc
- role
- sonas
- ssh
- user
- yum

Use this command to specify how many previously preserved backup versions must be kept and how many older backups are deleted. The default value is three versions. You can also specify the target host name where the backup is stored, by default the first found Storage node of the cluster, and the target directory path within the target host where the backup is stored, by default `/var/sonas/managementnodebackup`. The example in Figure 6-90 shows the **backupmanagementnode** command that is used to back up Management node configuration information for the components, auth, ssh, ctdb, and derby.

```
[root@sonas02 bin]# backupmanagementnode --component auth,ssh,ctdb,derby
EFSSG0200I The Management node mgmt001st002.virtual.com(10.0.0.20) has been successfully backedup.

[root@sonas02 bin]# ssh strg001st002.virtual.com ls /var/sonas/managementnodebackup
mgmtbak_20100413041835_e2d9a09ea1365d02ac8e2b27402bcc31.tar.bz2
mgmtbak_20100413041847_33c85e299643bebf70522dd3ff2fb888.tar.bz2
mgmtbak_20100413041931_547f94b096436838a9828b0ab49afc89.tar.bz2
mgmtbak_20100413043236_259c7d6876a438a03981d1be63816bf9.tar.bz2
```

Figure 6-90 Activate data replication

Attention: Although administrator backup of Management node configuration information is allowed and documented in the manuals, the procedure to restore the configuration information is not documented and must be done under the guidance of IBM support personnel.

The restoration of configuration data is done by running **cnmgmtconfbak**, which is used by the GUI when building a new Management node. The **cnmgmtconfbak** command can also be used to list the available archives. The command requires that you to specify **--targethost <host>** and **--targetpath <path>** to any backup, restore, or list. Figure 6-91 on page 475 shows the command syntax and how to get a list of available backups.

```
[root@sonas02]# cnmgmtconfbak
Usage: /opt/IBM/sofs/scripts/cnmgmtconfbak <command> <mandatory_parameters> [<options>]
commands:
    backup  - Backup configuration files to the bak server
    restore - Restore configuration files from the bak server
    list    - List all available backup data sets on the selected server
mandatory parameters:
    --targethost - Name or IP address of the backup server
    --targetpath - Backup storage path on the server
options: [-x] [-v] [-u N *] [-k N **]
    -x          - Debug
    -v          - Verbose
    --component - Select data sets for backup or restore (if archive contains
                   data set. (Default:all - without yum!)
                   Legal component names are:
            auth, callhome, cim, cron, ctdb, derby, role, sonas, ssh, user, yum, misc
            (Pls. list them separated with commas without any white space)
only for backup
    -k|--keep   - Keep N old bak data set (default: keep all)
only for restore
    -p|--fail_on_partial  - Fail if archive does not contain all required components
    -u|--use              - Use Nth bak data set (default: 1=latest)

[root@sonas02]# cnmgmtconfbak list --targethost strg001st002.virtual.com --targetpath       (..cont..)
               /var/sonas/managementnodebackup
1 # mgmtbak_20100413043236_259c7d6876a438a03981d1be63816bf9.tar.bz2
2 # mgmtbak_20100413041931_547f94b096436838a9828b0ab49afc89.tar.bz2
3 # mgmtbak_20100413041847_33c85e299643bebf70522dd3ff2fb888.tar.bz2
4 # mgmtbak_20100413041835_e2d9a09ea1365d02ac8e2b27402bcc31.tar.bz2
```

Figure 6-91 Configuration backup restore command

> **Remote server:** You can back up the configuration data to a remote server that is external to the SONAS cluster by specifying the **`--targethost`** parameter. The final copy of the archive file is done by the **`scp`** command, so the target remote server can be any server to which you have a passwordless access established. Establishing passwordless access to a remote server requires root access to the SONAS cluster.

6.10.2 Restoring data from a traditional backup

The data that is contained in the SONAS cluster can be backed up to a backup server, such as Tivoli Storage Manager or another supported backup product. Using that backup, it is possible to recover all the data that was contained in the SONAS cluster. For more information about backup and restore procedures, see 6.2, “Backing up and restoring file data” on page 372.

6.10.3 Restoring data from a remote replica

SONAS data can also be recovered from SONAS data replicas that are stored on a remote SONAS cluster or on a file server that is the target for SONAS asynchronous replication. To recover data that is stored on a remote system, you can use utilities such as **`xcopy`** and **`rsync`** to copy the data back to the original location.

The copy can be done from one of two places:

- From a SONAS Interface node on the remote system by using asynchronous replication to realign the data
- From an external SONAS client that mounts the shares for both the remote system, which contains a copy of the data to be restored, and for the local system that needs to be repopulated with data

The first method requires that the remote system is a SONAS cluster, and the second method works regardless of the type of remote system.

For more information about how to recover from an asynchronous replica, see “Recovery from a site disaster” on page 471.

6.10.4 Restoring cluster configuration data from a Management node backup

The **`backupmanagementnode`** command makes a backup from the local Management node, where the command is running, and stores it on another host or server.

> **Tip:** Use this method only for configurations where there is a single dedicated Management node.

For *Storwize V7000 Unified* and for *dual Management nodes*, the backup is automatic. Never run this command from the CLI in these particular cases.

Syntax of backupmanagementnode

Here is the syntax for the **`backupmanagementnode`** command:

```
backupmanagementnode [--component components] [--keep number] [--targethost host]
[--targetpath path] [--mount mountName] [-v]
```

The command has the following options:

- **`--component components`**

  Lists the components that must be backed up. If this option is not present, the components without yum are backed up. Valid component names are auth, callhome, cim, cron, ctdb, misc, role, sonas, ssh, user, and yum. Selected components should be listed in a comma-separated list with no white space. Optional.

- **`--keep number`**

  Specifies how many backups must be kept. Old backups are deleted. With this option, you can optimally use space on the device. The default value is `3`. Optional.

- **`--targethost host`**

  Specifies a target host name where the backup is stored. The default value is the first found Storage node of the cluster. If this option is omitted with IBM Storwize V7000 Unified, then the default value is the active Management node. Optional.

- **`--targetpath host`**

  Specifies a target path within the target host where the backup is stored. The default value is `/var/sonas/managementnodebackup`. Optional.

- **`--mount mountName`**

  Specifies a mount name where the backup is stored. The mount point is detected automatically.

- **`-v,--verbose`**

  Prints additional data columns. Optional.

Using unlisted options can cause an error.

Management node role failover procedures

The following procedures either restart the management service or initiate a management service failover from the node hosting the active Management node role to the node that is hosting the passive Management node role.

After completing, the node that previously hosted the active Management node role now hosts the passive Management node role. The node that previously hosted the passive Management node role now hosts the active Management node role.

> **Tip:** All of these tasks require a user that is configured as a CLI admin. Other users cannot do these tasks.

Determining the service IP for the Management node roles

Use this procedure to identify the service IP addresses for the nodes that host the Management node roles.

You need the service IP address of a node that hosts a Management node role to do a management failover from the node that hosts the active Management node role to the node that hosts the passive Management node role, when the active Management node fails and the current management IP does not respond.

For more information about this process, see the IBM Knowledge Center at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/trbl_dtrm_srvc_ip_mgmt_nd.html?lang=en

Management node role failover

If you want to initiate a Management node failover or Management node role failover on a good system, see the IBM Knowledge Center at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/trbl_mgmt_flovr_prcdrs.html?lang=en

6.11 NDMP

The SONAS system supports NDMP, which is an open standard protocol for NAS backup and restore functions.

The SONAS system supports NDMP V4, which is provided by compatible Data Management Applications (DMAs) such as Symantec Veritas NetBackup. Full and incremental backup and restore of file system data is provided by capturing all data and all metadata by using file system snapshots. An NDMP backup session provides backup of a specific directory, a set of directories in a file system, or all of the files and subdirectories that are contained within a file system. The name length of files and directories that are backed up or restored by using NDMP is limited to a maximum of 255 characters. Multiple directories within the same file system, and multiple file systems, can be backed up or restored concurrently. All extended attributes, including access control list (ACL) information, are also stored for every file and directory in a backup. File set information is not backed up or restored.

An NDMP restore session restores all of the files and directories in the backed up structure along with their extended attributes, including ACL information. A snapshot is used to provide a point-in-time copy for a backup. It is the snapshot of the directory structure that is backed up. The use of a snapshot accounts for files that might be open or in use during the backup.

6.11.1 SONAS NDMP supported physical configuration

There are two primary methods in which NDMP can be used as an interface with SONAS:

- Two-way (or remote) NDMP
- Three-way NDMP

The following sections describe the two-way (or remote) and three-way NDMP configurations in depth.

Two-way (or remote) SONAS NDMP configuration

The two-way NDMP consists of an external data management application, such as Symantec NetBackup, running on a server external to the SONAS system. The data management application has some form of storage hierarchy, such as a tape library that it manages for the storage of backup data.

In addition to the external data management application (Symantec NetBackup), an Ethernet network connects the data management application to the SONAS Interface nodes on which the NDMP server is running. NDMP control and data traffic flows across this network between the external data management application and the SONAS Interface nodes on which the NDMP server is running.

Typically, this network is a high-speed 10 Gb Ethernet network that handles the volume of data that is being backed up or restored. However, nothing prevents this network from being a 1 Gb Ethernet network.

Figure 6-92 on page 479 shows an example of a two-way SONAS NDMP configuration. It includes an external data management application (Symantec NetBackup) that is running on a server external to the SONAS system. The data management application has an IBM System Storage® TS7650G ProtecTIER Deduplication Gateway attached to it through an 8 Gbps Fibre Channel (FC) storage area network (SAN).

The System Storage TS7650G ProtecTIER Deduplication Gateway has an IBM System Storage DS5000™ storage controller that is attached to it through 8 Gbps FC links. It in turn has IBM System Storage DS5020 disk storage expansion units that are attached to it through 8 Gbps FC links. The data management application is connected to a 10 Gbps Ethernet network, to which the SONAS Interface nodes are attached. In Figure 6-92, the lines are intended to show the type and speed on the connections between the various physical components and do not necessarily represent the number of physical links.

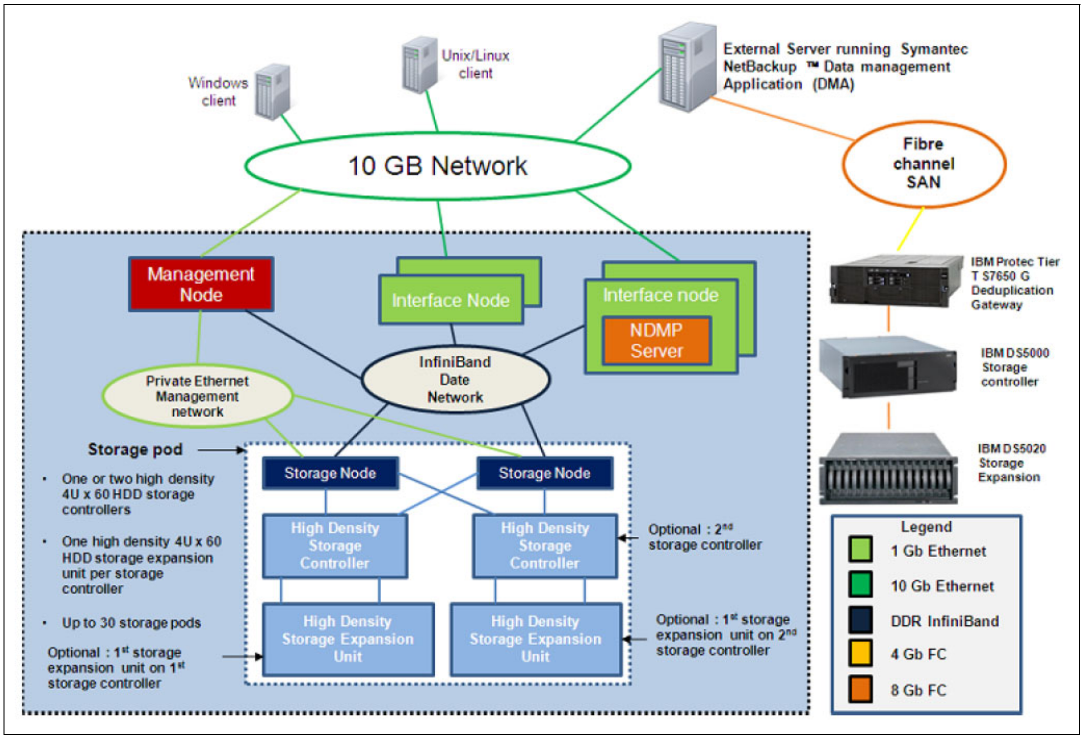

Windows client
Unix/Linux client
External Server running Symantec NetBackup ™ Data management Application (DMA)
10 GB Network
Fibre channel SAN
Management Node
Interface Node
Interface node
NDMP Server
InfiniBand Date Network
Private Ethernet Management network
Storage pod
• One or two high density 4U x 60 HDD storage controllers
• One high density 4U x 60 HDD storage expansion unit per storage controller
• Up to 30 storage pods
Optional : 1st storage expansion unit on 1st storage controller
Storage Node
Storage Node
High Density Storage Controller
High Density Storage Controller
High Density Storage Expansion Unit
High Density Storage Expansion Unit
Optional : 2nd storage controller
Optional : 1st storage expansion unit on 2nd storage controller
IBM Protec Tier T S7650 G Deduplication Gateway
IBM D S5000 Storage controller
IBM D S5020 Storage Expansion
Legend
1 Gb Ethernet
10 Gb Ethernet
DDR InfiniBand
4 Gb FC
8 Gb FC

Figure 6-92 Two-way SONAS NDMP configuration

Three-way SONAS NDMP configuration

In a three-way SONAS NDMP implementation, an NDMP tape server is installed on a server external to the SONAS system and separate from the server on which the Symantec NetBackup data management application is running. Some form of storage device, such as a virtual tape library or real tape library and tape drives, are attached to the server that is running the NDMP tape server.

The NDMP control traffic flows between the Symantec NetBackup data management application and the SONAS Interface nodes. NDMP data traffic flows between the SONAS Interface nodes and the NDMP tape server. In this scenario, because only NDMP control traffic is flowing between the Symantec NetBackup data management application and the SONAS Interface nodes, the Symantec NetBackup data management application does not need to be on a high-speed network. The Symantec NetBackup (data management application) can use a lower-speed, 1 Gbps Ethernet network to connect the SONAS Interface nodes and the external server on which the NDMP tape server is running.

However, NDMP data traffic (the data that is being backed up or restored) is flowing between the SONAS Interface nodes and the external server on which the NDMP tape server is running. Therefore, make the Ethernet network between the NDMP tape server and the SONAS Interface nodes running the NDMP server a high-speed, 10 Gbps Ethernet network.

Figure 6-93 shows an example of an NDMP three-way configuration. In this example, the SONAS Interface nodes (used for file serving) and the data management applications are on a 1 Gbps Ethernet network. Other SONAS Interface nodes, which are used for NDMP backup and restore, are on a 10 Gbps Ethernet network along with the NDMP tape server. The server on which the NDMP tape server is running is connected to an 8 Gbps FC SAN along with an IBM System Storage TS3500 tape library with some FC attached tape drives, such as IBM Linear Tape-Open data cartridges, Generation 3, Generation 4, or Generation 5 tape drives.

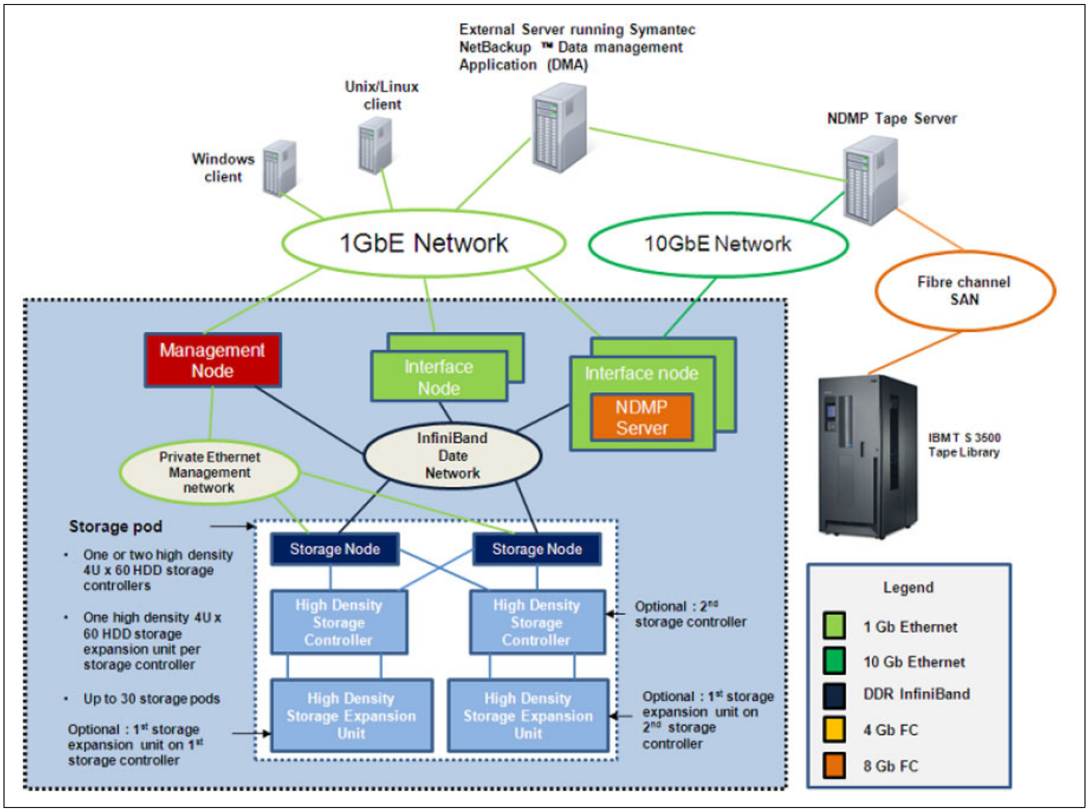

External Server running Symantec NetBackup ™ Data management Application (DMA)
Unix/Linux client
Windows client
NDMP Tape Server
1GbE Network
10GbE Network
Fibre channel SAN
Management Node
Interface Node
Interface node
NDMP Server
InfiniBand Date Network
Private Ethernet Management network
IBM T S 3500 Tape Library
Storage pod
• One or two high density 4U x 60 HDD storage controllers
• One high density 4U x 60 HDD storage expansion unit per storage controller
• Up to 30 storage pods
Optional : 1st storage expansion unit on 1st storage controller
Storage Node
Storage Node
High Density Storage Controller
High Density Storage Controller
High Density Storage Expansion Unit
High Density Storage Expansion Unit
Optional : 2nd storage controller
Optional : 1st storage expansion unit on 2nd storage controller
Legend
1 Gb Ethernet
10 Gb Ethernet
DDR InfiniBand
4 Gb FC
8 Gb FC

Figure 6-93 Three-way SONAS NDMP configuration

6.11.2 Fundamentals of the SONAS NDMP feature

The following points explain the fundamentals of the SONAS NDMP feature:

- An NDMP V4 compliant data server is available on every Interface node of the SONAS system. There is a provision to create a set of Interface nodes that are part of an NDMP_NODE_GROUP. This set of Interface nodes is paired with network group IP addresses that can be assigned to a specific network port. This network port is associated with the NDMP service.
- The NDMP server that runs on an Interface node provides for both data and control connections to servers that are external to the SONAS system on which a data management appliance is running.
- The ability to configure the NDMP parameters for this set of Interface nodes (the NDMP_NODE_GROUP) is provided through CLI commands that are run from the Management node.

- In addition to the CLI commands that are available to store and retrieve NDMP configuration parameters, a set of CLI commands provide interfaces to view NDMP session information and NDMP log information, and to stop currently running NDMP sessions.
- An NDMP backup session provides a backup of a specific directory in a GPFS file system and all files and subdirectories that are contained within it. In addition to the basic data of the files and directories, all extended GPFS attributes are saved for every file and directory. To provide the provision to back up a directory structure at a particular point in time, a snapshot is used, and it is the snapshot of the directory structure that is backed up. This snapshot also accounts for files that might be open or in use during the backup as a point-in-time representation of the file that is backed up by NDMP.
- An NDMP restore session restores all of the files and directories in the proper structure of subdirectories, and so on. In addition to the file contents, the GPFS extended attributes are also restored.

6.11.3 Configuring NDMP for the SONAS system

The main components of the NDMP configuration for the SONAS system are the data server, the NDMP tape server, and the data management application (DMA).

Overview

The two primary NDMP interfaces are the data server and the NDMP tape server. The data server reads data in an NDMP data stream from a disk device and writes NDMP data to disk. The NDMP tape server reads NDMP stream data from, or writes NDMP data to, a direct-attached storage device. (The NDMP tape server in the IBM Knowledge Center refers to the NDMP function that interfaces with any supported direct-attached storage device without regard to the actual storage device type to which it connects). The DMA controls NDMP data movement, including backup and restore operations.

The data server for SONAS system support of NDMP is software that runs on each of several SONAS Interface nodes that are configured collectively as an NDMP node group. The NDMP data server is integrated into the overall SONAS code stack and is installed on Interface nodes in the same way as any other software components that are running on the Interface nodes. Each Interface node in an NDMP node group contains an identical copy of its NDMP node group configuration. A data server can be started on each Interface node, so all of the Interface nodes are eligible to be configured as part of an NDMP node group to interact with a DMA. Configuring multiple NDMP node groups is not supported. An Interface node can be a member of a maximum of one NDMP node group. If an NDMP session begins on one Interface node and fails because the Interface node fails, the session can be restarted on another Interface node in its NDMP node group.

The DMA server is external to the SONAS system on a Linux, AIX, Microsoft Windows, or other platform, and connects to the data servers on each of the Interface nodes in an NDMP node group by using Ethernet.

The NDMP tape server function can be provided from the DMA server, which is called a remote configuration, or it can be provided by a separate external server as part of what is referred to as a three-way configuration.

Using the CLI

To configure an NDMP node group to be used for the backup, create a network group, create a network, attach the network group to the network, associate the network group with the NDMP node group, and configure and activate the NDMP node group, complete the following steps:

1. Create an NDMP node group by running **`cfgndmp --create`**. For example, create an NDMP node group that is named ndmpg1, as shown in Example 6-23.

   Example 6-23 Create an NDMP node group with cfgndmp --create

   ```
   [admin@st001.mgmt001st001 ~]# cfgndmp ndmpg1 --create
   NDMP group successfully created. Set your required NDMP parameters before
   activating NDMP.
   EFSSG1000I The command completed successfully.
   ```

   > **Tip:** An NDMP configuration can be defined only for an NDMP node group. The NDMP configuration parameters of a single Interface node cannot be changed individually. An NDMP node group must be created by running **`cfgndmp`** before any other NDMP node group configuration parameters can be set, and an NDMP node group is created and configured before an NDMP backup or restore can be configured on the DMA server. These initial configuration steps must be performed once for each NDMP node group. To change a configuration option when the NDMP node group has already been activated, the NDMP node group must be deactivated, the option changed, and then the NDMP node group must be activated.

2. Create a network group that includes all of the Interface nodes that service NDMP requests for the NDMP node group by running **`mknwgroup`**. For this example, assume that you have two Interface nodes, int001st001 and mgmt002st001, and use them to create a network group that is named ndmp_group, as shown in Example 6-24.

   Example 6-24 Create an NDMP network group

   ```
   [admin@st001.mgmt001st001 ~]# mknwgroup ndmp_group int001st001,mgmt002st001
   Reconfiguring NAT gateway 10.0.0.11/24
   EFSSG0087I NAT gateway successfully removed.
   EFSSG0086I NAT gateway successfully configured.
   EFSSG1000I The command completed successfully.
   ```

3. Create a network by running **`mknw`**, as shown in Example 6-25.

   Example 6-25 Create a network

   ```
   [admin@st001.mgmt001st001 ~]# mknw 17.0.0.0/24 0.0.0.0/0:17.0.0.1 --add
   17.0.0.100,17.0.0.101
   EFSSG1000I The command completed successfully.
   ```

4. Attach the network to the network group by running **`attachnw`**. For this example, assume that there is a 10 Gb card in each Interface node and that ethX0 can be used to access the bonded ports, as shown in Example 6-26.

   Example 6-26 Attach a network to a network group

   ```
   [admin@st001.mgmt001st001 ~]# attachnw 17.0.0.0/24 ethX0 -g ndmp_group
   EFSSG0015I Refreshing data.
   EFSSG1000I The command completed successfully.
   ```

5. Associate the network group with the NDMP node group by running **cfgndmp --networkGroup**, as shown in Example 6-27.

Example 6-27 Associate a network group to an NDMP node group

```
[admin@st001.mgmt001st001 ~]# cfgndmp ndmpg1 --networkGroup ndmp_group
This will clean NDMP configuration from previous Network group attached to
NDMPG1 node group if there were any.
Do you really want to perform the operation (yes/no - default no):yes
Network group configured for this NDMP group.
EFSSG1000I The command completed successfully.
```

NDMP node group: Only one network group can be associated with an NDMP node group. Because of the tight coupling between an NDMP node group and the associated network group, a particular Interface node can exist only in one NDMP node group, just as it can exist only in a single network group. For each NDMP node group that is configured, there can be only one unique associated network group. Each network group can be associated with only one NDMP node group. Any valid network group can be associated with an NDMP node group, including the network group that was created at system creation.

6. List the Interface nodes that are contained in an NDMP node group by running **lsnwgroup**, as shown in Example 6-28.

Example 6-28 List Interface nodes in an NDMP node group

```
[root@st001.mgmt001st001 ~]# lsnwgroup
Network Group Nodes                   Interfaces
DEFAULT       mgmt001st001
int                                   ethX0
ndmp_group    int001st001,mgmt002st001 ethX0
EFSSG1000I The command completed successfully.
```

7. Set the data port range of the NDMP node group by running **cfgndmp --dataTransferPortRange**. When the fields NDMP_PORT and DATA_TRANSFER_PORT_RANGE are blank, that means that there are no restrictions (see Example 6-29).

Example 6-29 Data port range set by running cfgdmp

```
[admin@st001.mgmt001st001 ~]# cfgndmp  NDMPG1 --dataTransferPortRange 2048-2098
Data port range configured for this NDMP group.
EFSSG1000I The command completed successfully.
```

8. Add file system mount point paths to the NDMP node group configuration by running **cfgndmp --addPaths**, as shown in Example 6-30.

Example 6-30 File system mount point paths by running cfgndmp

```
[admin@st001.mgmt001st001 ~]# cfgndmp  NDMPG1 -addPaths /ibm/gpfs0,/ibm/gpfs1
Backup/recovery configured for this NDMP group.
EFSSG1000I The command completed successfully.
```

9. Activate the NDMP node group by running **`cfgndmp --activate`**, as shown in Example 6-31.

Example 6-31 Activate the NDMP node group

```
[root@st001.mgmt001st001 ~]# cfgndmp NDMPG1 --activate
NDMP group activated.
EFSSG1000I The command completed successfully.
```

NDMP backup prefetch: A default NDMP backup prefetch configuration is assigned to a newly created NDMP node group, with the function deactivated. Optionally, run **`cfgndmpprefetch`** to change the NDMP backup prefetch configuration and activate the function for improved NDMP backup performance. For more information, see 6.11.7, "Configuring NDMP backup prefetch" on page 491.

10. Verify that NDMP has started on all of the nodes in the NDMP node group by running **`lsndmp --ndmpServiceStatus`**, as shown in Example 6-32.

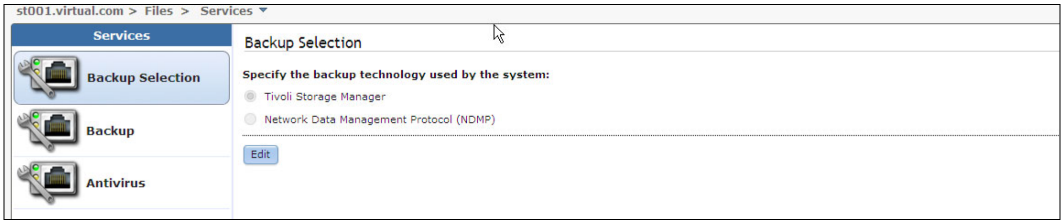
Example 6-32 Verify that NDMP has started by running lsndmp

```
[root@st001.mgmt001st001 ~]# lsndmp --ndmpServiceStatus
Nodes group name Node                     Ndmp service status
NDMPG1           mgmt002st001(172.31.136.3) RUNNING
NDMPG1           int001st001(172.31.132.1)  RUNNING
EFSSG1000I The command completed successfully.
```

Using the GUI to configure NDMP

To configure NDMP by using the GUI, complete the following steps:

1. Click **Files** → **Services and Backup Selection**, as shown in Figure 6-94.



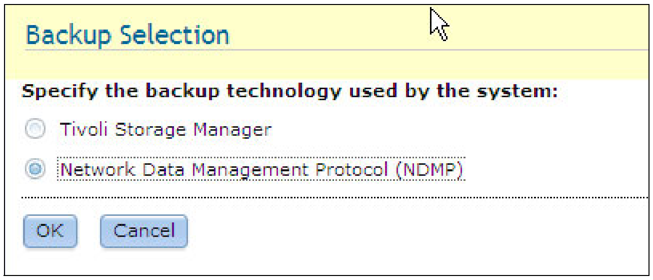
Specify the backup technology used by the system:

○ Tivoli Storage Manager

○ Network Data Management Protocol (NDMP)

Figure 6-94 Backup Selection window to configure NDMP

2. In the Backup Selection window, click **Network Data Management Protocol**, as shown in Figure 6-95. Click **OK** to continue.



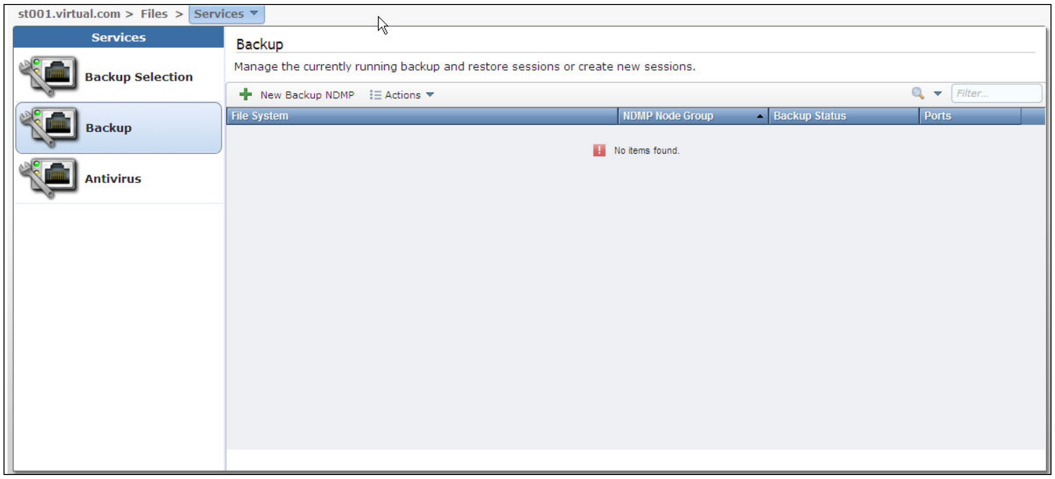
Specify the backup technology used by the system:

○ Tivoli Storage Manager

● Network Data Management Protocol (NDMP)

Figure 6-95 Click Network Data Management Protocol (NDMP)

3. From the main window, click **File** → **Services** and click **Backup**, as shown in Figure 6-96.



Backup

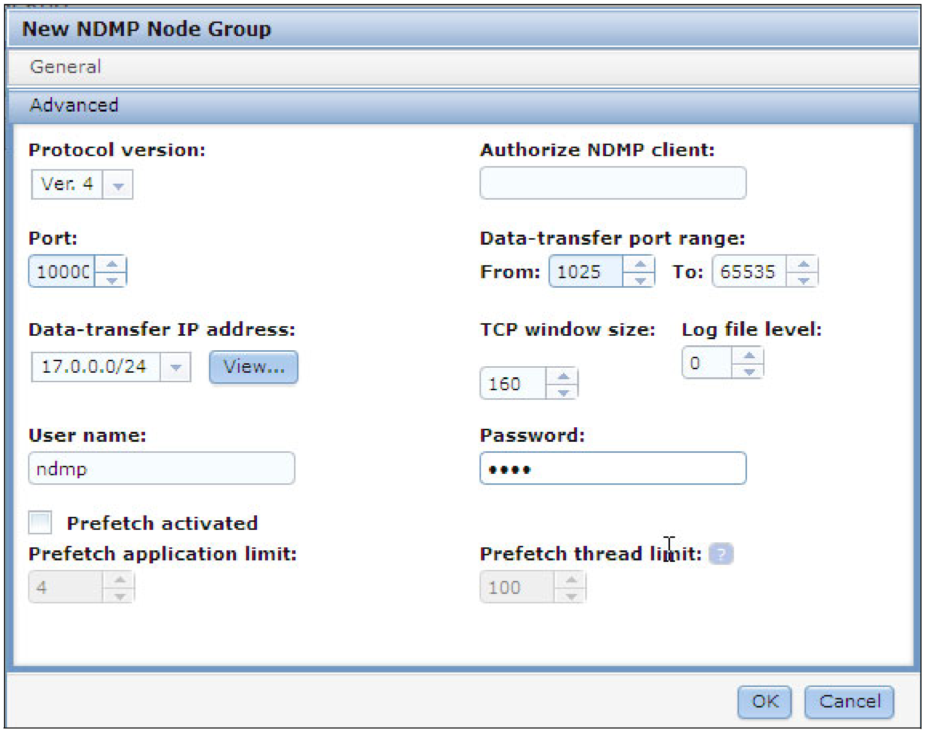
Manage the currently running backup and restore sessions or create new sessions.

File System	NDMP Node Group	Backup Status	Ports

No items found.

Figure 6-96 Backup window to manage sessions

4. From the New NDMP Node Group window, enter the required information, as shown in Figure 6-97.



Protocol version: Ver. 4

Authorize NDMP client:

Port: 1000C

Data-transfer port range: From: 1025 To: 65535

Data-transfer IP address: 17.0.0.0/24

TCP window size: 160

Log file level: 0

User name: ndmp

Password: ••••

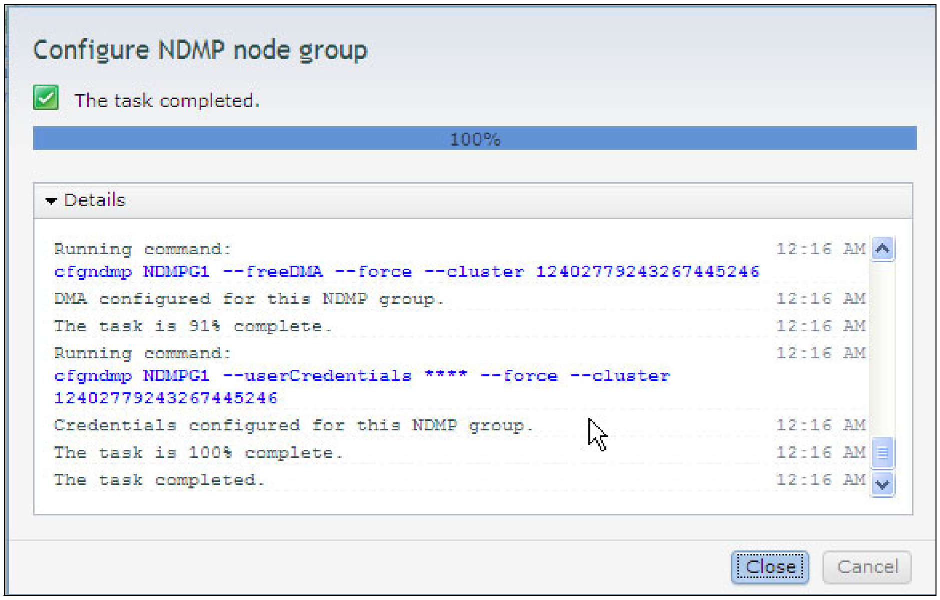
☐ Prefetch activated

Prefetch application limit: 4

Prefetch thread limit: 100

Figure 6-97 New NDMP Node Group information

5. After you click **OK**, the Configure NDMP node group window that shows the progress of the task opens, as shown as in Figure 6-98.



Configure NDMP node group

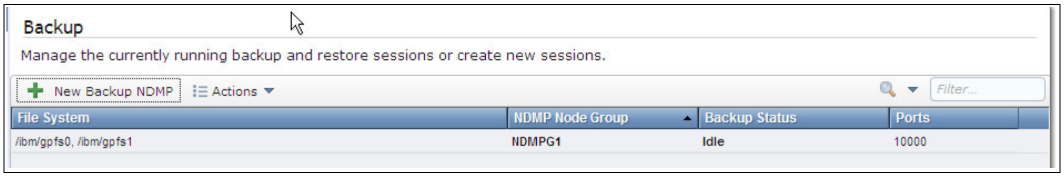
The task completed.

100%

```
Running command:                                                  12:16 AM
cfgndmp NDMPG1 --freeDMA --force --cluster 12402779243267445246
DMA configured for this NDMP group.                               12:16 AM
The task is 91% complete.                                         12:16 AM
Running command:                                                  12:16 AM
cfgndmp NDMPG1 --userCredentials **** --force --cluster
12402779243267445246
Credentials configured for this NDMP group.                       12:16 AM
The task is 100% complete.                                        12:16 AM
The task completed.                                               12:16 AM
```

Figure 6-98 Configure NDMP node group progress window

In the Backup window, the newly created NDMP Node Group is displayed, as shown in Figure 6-99.



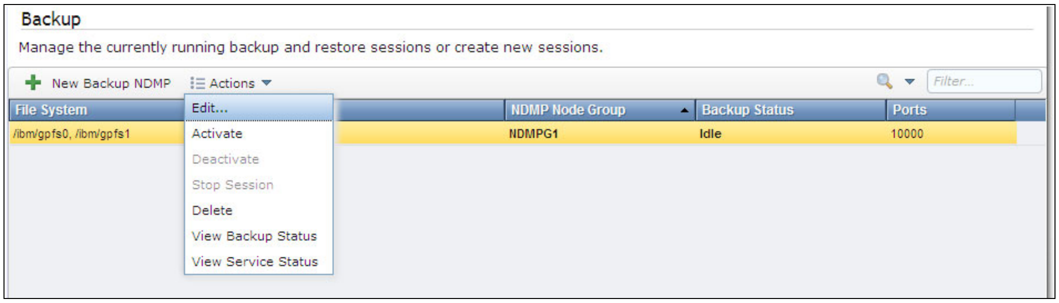
Backup

Manage the currently running backup and restore sessions or create new sessions.

File System	NDMP Node Group	Backup Status	Ports
/ibm/gpfs0, /ibm/gpfs1	NDMPG1	Idle	10000

Figure 6-99 Newly created NDMP node group displayed

6. Under the File System list, select the newly created file system and click **Actions** → **Activate**, as shown in Figure 6-100.



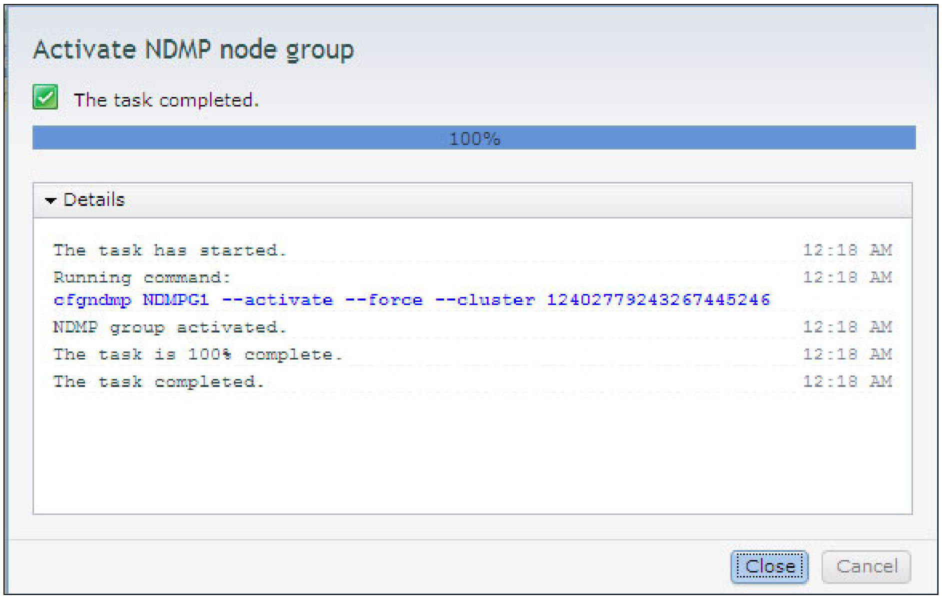
Backup

Manage the currently running backup and restore sessions or create new sessions.

File System	NDMP Node Group	Backup Status	Ports
/ibm/gpfs0, /ibm/gpfs1	NDMPG1	Idle	10000

Figure 6-100 Activate backup for the NDMP file system

The progress of the Activate NDMP node group window is displayed. The CLI commands that are run in the background are shown in Figure 6-101 on page 487.



Activate NDMP node group

☑ The task completed.

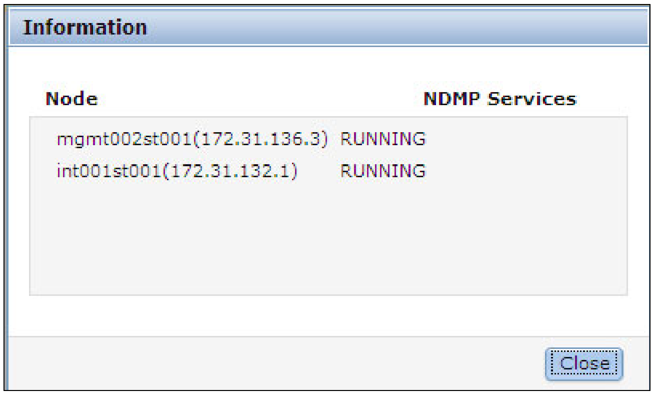
100%

Details

```
The task has started.                                        12:18 AM
Running command:                                             12:18 AM
cfgndmp NDMPG1 --activate --force --cluster 12402779243267445246
NDMP group activated.                                        12:18 AM
The task is 100% complete.                                   12:18 AM
The task completed.                                          12:18 AM
```

Figure 6-101   Status window that shows the NDMP group configuration

Figure 6-102 shows the NDMP services running on the nodes.



Information

Node	NDMP Services
mgmt002st001(172.31.136.3)	RUNNING
int001st001(172.31.132.1)	RUNNING

Figure 6-102   Information of NDMP Services on nodes

6.11.4 Viewing an NDMP session

This section describes how to view an NDMP session.

Overview

The operations that are described in this section are done only from the CLI.

Using the CLI

Complete the following steps:

1. Run **`lsndmpsession`** with the **`-n`** or **`--nodes`** option to view NDMP sessions that are running on specified Interface nodes, or with the **`-g`** or **`--nodeGroup`** option to view NDMP sessions that are running on specified NDMP node groups. Only Interface nodes can be specified when you are using the **`-n`** or **`--nodes`** option. Multiple nodes and multiple node groups in a list must be separated with commas.

If no nodes or node groups are specified, the output displays information for all of the Interface nodes in the system. To determine which Interface nodes are running NDMP sessions for an NDMP node group, run the command that is shown in Example 6-33, which shows which Interface nodes have NDMP sessions that are running and which sessions are running on the nodes.

Example 6-33 Determine Interface nodes running NDMP sessions

```
[admin@st001.mgmt001st001 ~]# lsndmpsession -g DEMONDMP
HOST SESSION_ID SESSION_TYPE BYTES_TRANSFERRED NDMP_VERSION AGE MB/SECONDS
LOCATION
int002st001 678964 DATA_RECOVER 3482892664832 25-28-51 42.52
EFSSG1000I The command completed successfully.
```

In Example 6-34, the node number 2 int002st001 is specified.

Example 6-34 The lsndmpsession command with specific node specified

```
[admin@st001.mgmt001st001 ~]# lsndmpsession -n int002st001 -v
HOST SESSION_ID SESSION_TYPE BYTES_TRANSFERRED NDMP_VERSION START_TIME DMA_IP
DATA_IP DATA_STATE TARGET_PATH PREP_DIR_PATH CURRENT_PATH DIR_PROCESSED
FILE_PROCESSED MOVER_IP MOVER_STATE AVERAGE_THRUPUTHOST CURRENT_THRUPUT DEVICE
int002st001 678964 DATA_RECOVER 3481312821248 1320186540 10.1.60.83 10.1.60.83
ACTIVE 0 0 10.1.60.115 IDLE 37972434.790000 43869798.000000
EFSSG1000I The command completed successfully.
```

2. Run **`lsndmpsession`** with the **`-i`** or **`--sessionID`** option to view the verbose information of the NDMP session that is identified by the specified session ID (SID) running on the specified node. In Example 6-35, the SID 17067 is specified:

   Example 6-35 The lsndmpsession command with parameters to view verbose information

   ```
   [admin@st001.mgmt001st001 ~]# lsndmpsession -n int002st001 -i 17067
   ```

   NDMP session information similar to Example 6-36 is displayed.

   Example 6-36 NDMP session information displayed

   ```
   SESSION_ID = 17067
   SESSION_TYPE = DATA_BACKUP
   START_TIME = 1256074914
   DMA_IP = 10.1.5.9
   DATA_IP = 10.1.5.12
   DATA_STATE = ACTIVE
   TARGET_PATH = /home/user1
   PREP_DIR_PATH = /home/.SnapShotDir/user1
   CURRENT_PATH = /home/.SnapShotDir/user1/source/fs/lnk.h
   DIR_PROCESSED = 87
   FILE_PROCESSED = 36002
   MOVER_IP = 10.168.105.14
   MOVER_STATE = IDLE
   BYTES_TXFERRED = 2298511448
   AVERAGE_THRUPUT = 33846454.51
   CURRENT_THRUPUT = 49825019.00
   ```

3. To view verbose NDMP session information for all NDMP sessions that are running on nodes that are specified by the **-n**, **--nodes**, **-g**, **--nodeGroup**, **-c** or **--cluster** options, use the **-v** or **--verbose** option of the **lsndmpsession** command without specifying the **-i** or **--sessionID** options.

6.11.5 Stopping an NDMP session

This section describes how to stop an NDMP session.

Overview

The operations that are described in this section can be done only from the CLI.

Using the CLI

Complete the following steps:

1. Run **stopndmpsession** with the **-g** or **--nodeGroup** option to stop all of the NDMP sessions that are running on the specified node group. In Example 6-37, the node group ndmpg1 is specified.

Example 6-37 Stop all NDMP commands with the stopndmpsession command

```
[admin@st001.mgmt001st001 ~]# stopndmpsession -g ndmpg1
This will stop specified NDMP sessions.
Do you really want to perform the operation (yes/no - default no):yes
Sessions killed on host are : {int001st001=SESSIONS KILLED,
mgmt002st001=SESSIONS KILLED}
EFSSG1000I The command completed successfully.
```

Considerations:

- When you run **stopndmpsession**, the stopped backup processes must still do subprocesses, such as cleaning up snapshots, and therefore might not stop immediately.
- After you run **stopndmpsession**, run **lsndmpsession** to ensure that all related SONAS NDMP backup sessions on all of the involved Interface nodes have stopped.
- Although only the NDMP sessions that are running on the specified Interface node or node group are stopped, the overall backup process fails because not all NDMP sessions completed.
- Also, although this command stops the specified NDMP sessions that are running on the SONAS system, those sessions might be restarted, depending on the settings of the backup software.
- If the **stopndmpsession** command is used for stopping NDMP backup sessions, disable the automated resubmit option on the backup software DMA; otherwise, unintended backup jobs might restart.

2. Run **stopndmpsession** with the **-n** or **--nodes** option to stop all of the NDMP sessions that are running on the specified Interface node. In Example 6-38, the node int001st001 is specified.

Example 6-38 Stop NDMP sessions with the stopndmpsession command

```
[admin@st001.mgmt001st001 ~]# stopndmpsession -n int001st001
This will stop specified NDMP sessions
Do you really want to perform the operation (yes/no - default no):yes
(1/2) Kill session started on hosts : int001st001
Kill session on host : int001st001
Kill session on host : int001st001 done
(2/2) All possible session killing done on hosts : int001st001. Killing done is
: {int001st001=SESSIONS KILLED}
Sessions killed on host are : {int001st001=SESSIONS KILLED}
```

6.11.6 Viewing NDMP log information

This section describes how to view the NDMP log information.

Overview

The operations that are described in this section are done only from the CLI.

Using the CLI

Run **lsndmplog** with the **-g** or **--nodeGroup** option to view all of the NDMP log information that is related to the specified node group. In Example 6-39, the node group NDMPG1 is specified.

Example 6-39 lsndmplog command to view NDMP log information

```
[root@st001.mgmt001st001 ~]# lsndmplog -g NDMPG1

int001st001

/VAR/LOG/CNLOG/NDMP.LOG LOGS
IO:CN 11/03 00:02:02.962627 2161738:9162b7c0
comm.c:ndmpRun:906 0001: Starting ndmpd listener on port 10000 of all IP address
IO:CN 11/03 00:02:02.975743 2162529:9162b7c0
comm.c:ndmpRun:982 0005: ndmpd started
IO:CN 11/03 00:18:33.827665 2312841:dd5f77c0
comm.c:ndmpRun:906 0001: Starting ndmpd listener on port 10000 of all IP address
IO:CN 11/03 00:18:33.837252 2313619:dd5f77c0
comm.c:ndmpRun:982 0005: ndmpd started
.....
```

The most recent 10 lines of the NDMP log files for the Interface nodes in the specified node group are displayed by the CLI interface.

Run **lsndmplog** with the **-o** or **--outputLogFilePath** option to save the log file as a temporary file, as shown in Example 6-40.

Example 6-40 Save NDMP log files for Interface nodes

```
[root@st001.mgmt001st001 ~]# lsndmplog -n int001st001 -o
/ibm/gpfs0/sharename/ndmpout.log
int001st001
```

```
EFSSA0183C Log file /var/log/cnlog/ndmp.log.old on Node int001st001 has not been
created.
EFSSG1000I The command completed successfully.
```

6.11.7 Configuring NDMP backup prefetch

This section describes how to configure an NDMP backup prefetch.

Overview

The operations that are described in this section can be done only from the CLI.

The NDMP backup prefetch function navigates the directory that is being backed up. It reads files in advance of the files that are coming due to be backed up. The prefetch function opens files in read-only mode and places the files in the cache of the Interface node for improved backup performance. Run **`cfgndmpprefetch`** to configure, activate, and deactivate the NDMP backup prefetch function.

> **Consideration:** NDMP backup prefetch is designed to work on files that are less than or equal to 1 MB. NDMP backup prefetch does not work for a file system that has a block size that is greater than 1 MB.

If NDMP is active for the specified NDMP node group when prefetch is activated or deactivated, or an NDMP prefetch configuration is changed, NDMP on the specified NDMP node group must be deactivated and then activated for the prefetch activation, deactivation, or configuration change to be implemented. It occurs if the user responds with “yes,” “y,” or “Y” to the prompt `Command requires NDMP to be deactivated before configuration can change. Deactivating NDMP will stop all NDMP sessions currently in progress and not allow new NDMP sessions to start for this NDMP node group. Do you really want to perform the operation (yes/no - default no):`.

Using the CLI

To activate the NDMP backup prefetch feature, run **`cfgndmpprefetch`** with the **`--activate`** option and specify the NDMP node group, as shown in Example 6-41.

Example 6-41 NDMP backup prefetch feature activation

```
[root@st001.mgmt001st001 ~]# cfgndmpprefetch NDMPG1 --activate
EFSSG0448W Command requires NDMP to be deactivated before configuration can
change. Deactivating NDMP will stop all NDMP sessions currently in progress and
not allow new NDMP sessions to start for this NDMP node group. Depending on how
many NDMP sessions will be killed (if any), there may be some delay.
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
```

If NDMP is not active on the specified NDMP node group, a message is displayed indicating that prefetch for NDMP is activated when NDMP is activated.

6.12 Immutability support

Immutability support, which is also called non-eraseable, non-rewriteable (NENR) file support, is available in SONAS as a request Price Quotation (RPQ). An RPQ feature is not a standard feature of a product, but is offered after the customer explicitly requests it. When enabled, this RPQ enables the setting of the immutability flag on files and folders in SONAS.

The immutability flag can be turned on and off by running **`mmchattr -i {yes | no} filename`** against a given file or directory that must be set to immutable. The RPQ enables the **`mmchattr`** command to be part of the list of allowed **`sudo`** commands, and enables RPQ-approved customers to use GPFS's built-in and supported basic file immutability. This function is not for compliance and timed retention, and is not supported for backup, replication, or HSM. Attributes for a file can be listed by running **`mmlsattr -L`** ***`filename`***.

The **`mmchattr`** command can be used continuously or periodically in *clumpy* batches, which are event-driven batches. The command can be automated in a script. Ideally, the script unexports the location of the files to be made immutable, then re-exports it when done, to ensure that no writes occur while the attribute is being set. Setting or resetting the immutable attribute during any read operation should have no effect on the read operation. Setting the attribute while a write is in flight might permit the write to succeed or it might cause it to fail the write, depending on the exact timing. However, it should not corrupt the file on any write that is atomically consistent. If the application is performing multiple discrete inconsistent writes, that is, non-atomic writes, then file inconsistency might result. If an application performs multiple write operations before it achieves file consistency, and the attribute is set during that application's writes, file inconsistency might result. This is why ideally the data is unexported while a mass attribute setting is done.

Immutability is not preserved across backup, HSM, or replication. However, these operations can be running when the attribute is set or reset, except that the attribute cannot be set on the *target* of any restore or asynch replication or WAN caching operation while the operation is running. This restriction is to prevent the *corrupted file* scenario of multiple non-atomic writes in flight.

Snapshots remain achievable. SONAS does allow for normal snapshot behavior of files and directories that are set as immutable. Restores of snapshots are not allowed to over-write the original files. Restores to folders that are listed as immutable are also not allowed. However, you can restore immutable files to folders that are not set as immutable.

7

SONAS administration

This chapter provides information about how you can use the GUI and CLI to administer your SONAS. Daily administrator tasks are described and examples are provided.

This chapter describes the following topics:

- Using the management interface
- SONAS administrator tasks list
- Cluster management
- File system management
- Creating and managing shares
- Disk management
- User management
- Services management
- Scheduling tasks in SONAS
- Health Center
- Call Home
- Assist On-site
- Logs: Uploading and downloading
- Network settings
- Event notifications

7.1 Using the management interface

The SONAS can be accessed with a GUI and a command-line interface (CLI). The GUI has different administrative window to do administrative tasks. You can use the CLI to administer the system with commands.

Both the CLI and GUI provide details and help for each task and command. The CLI also contains the man pages that you can use to get more information about a specific command. The GUI tasks are made to be self-explanatory and also have tooltips for every text box or command, which gives more information about what will be done. There is also the Help icon, which redirects you to SONAS IBM Knowledge Center.

This chapter explains most of the important and commonly used commands for both the GUI and the CLI. The screen captures that are used in this chapter are taken from different SONAS systems.

For more information about the steps that are required to access the GUI, see “Add Cluster to CLI” in *IBM SONAS Installation Guide*, GA32-0715.

7.1.1 GUI tasks

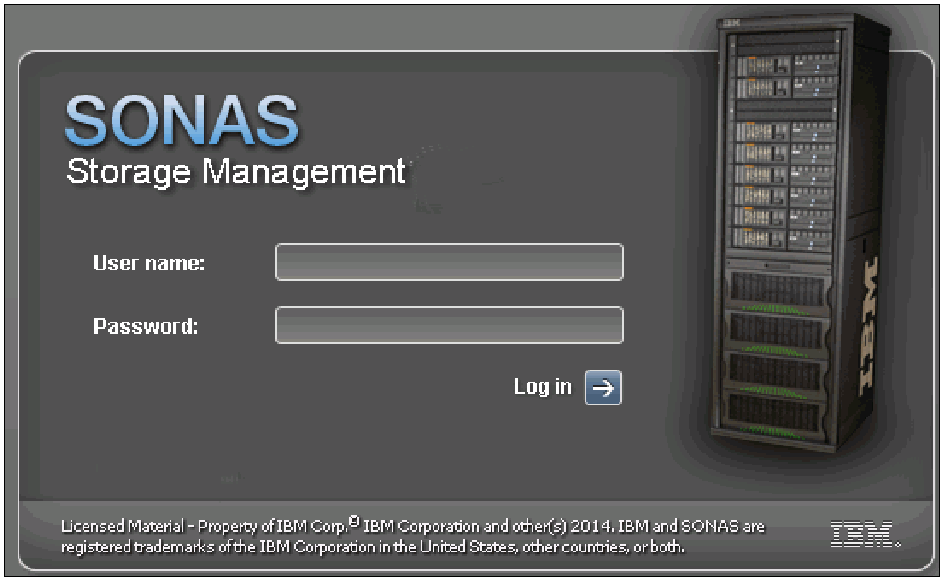
GUI tasks are tasks that you can do with the graphical interface of the SONAS. You log in to the Management node with an Internet browser, such as Internet Explorer or Firefox. On the URL bar, enter this link:

```
https://management_node_name_or_ip_address:1081
```

In this example, you have a Management node with IP: 9.11.137.220. You can access the GUI with this link:

```
https://9.11.137.220:1081
```

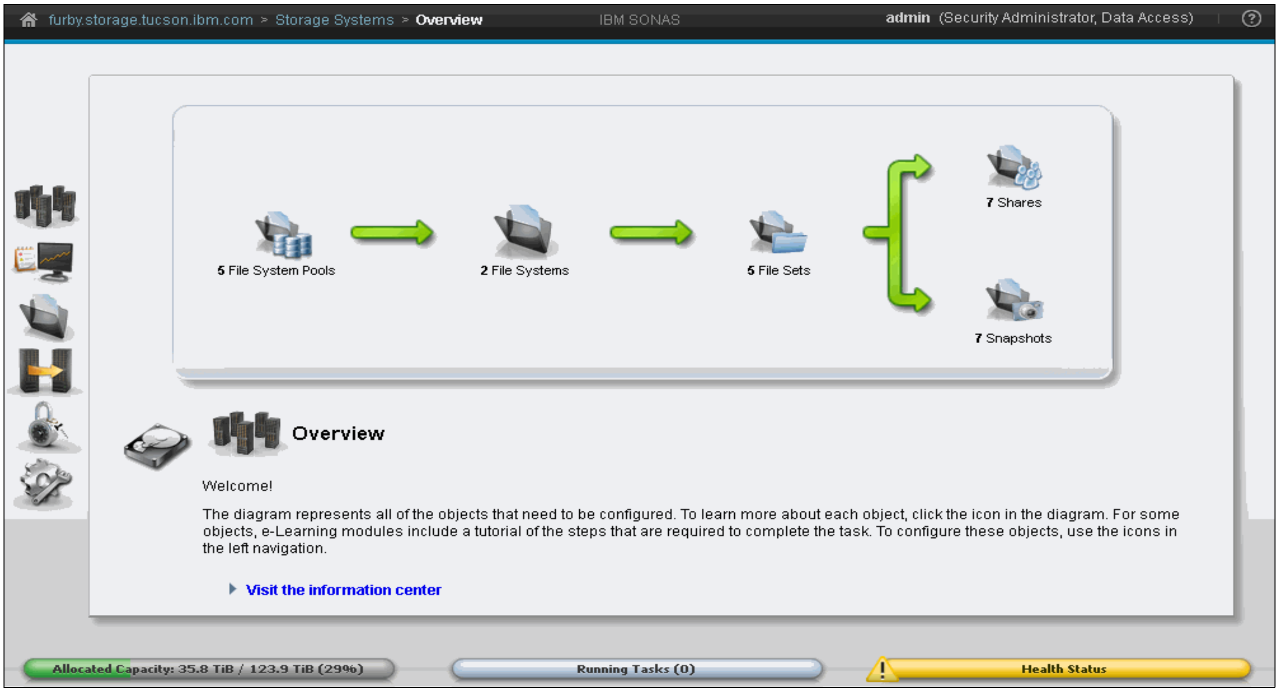
Figure 7-1 shows the login window. You need to enter the login name and password and click **Log in** to log in to a Management Interface.



Licensed Material - Property of IBM Corp.© IBM Corporation and other(s) 2014. IBM and SONAS are registered trademarks of the IBM Corporation in the United States, other countries, or both.

Figure 7-1 SONAS Management GUI asking for login details

If your browser does not support Java based web pages, there is an option for Low graphics mode for the window where you enter login credentials. When you are logged in, you see the window that is shown in Figure 7-2.


furby.storage.tucson.ibm.com > Storage Systems > Overview
IBM SONAS
admin (Security Administrator, Data Access)
5 File System Pools
2 File Systems
5 File Sets
7 Shares
7 Snapshots
Overview
Visit the information center
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Welcome!

The diagram represents all of the objects that need to be configured. To learn more about each object, click the icon in the diagram. For some objects, e-Learning modules include a tutorial of the steps that are required to complete the task. To configure these objects, use the icons in the left navigation.

Figure 7-2 SONAS GUI when logged in as the admin user

Figure 7-2 illustrates the various areas on the GUI navigation window. To the left, you have the main navigational pane, which allows you to select the area that you want to view or the task that you want to do. On the top, you see the currently logged-in administrative user name. Icons in the navigation pane represent six areas: Storage Systems, Monitoring, Files, Copy Services, Access, and Settings. Each of these fields is then divided in to subcategories. The icons at the bottom are Allocated Capacity, Running Tasks, and Health Status. These menus are available any time in the GUI. They are independent throughout GUI navigation.

When you are logged in, on the main page, in the upper right bar, you can see the CLI user name that is logged in to the GUI. At the right corner, you also see the link to log out of the GUI.

7.1.2 Context-sensitive help

Context-sensitive help is a kind of online help that is available in a GUI window to help you understand features that are associated with that window.

Access to context-sensitive help

You can access context-sensitive help by clicking the question mark at the upper right of the window, as shown in Figure 7-3.

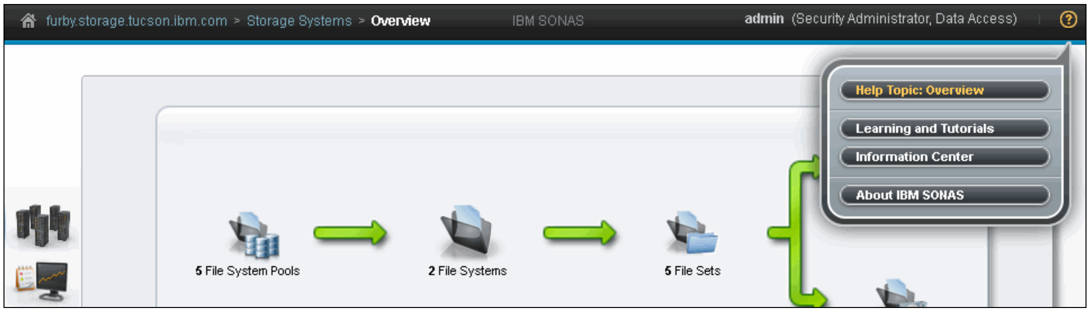


Figure 7-3 Access to context-sensitive help

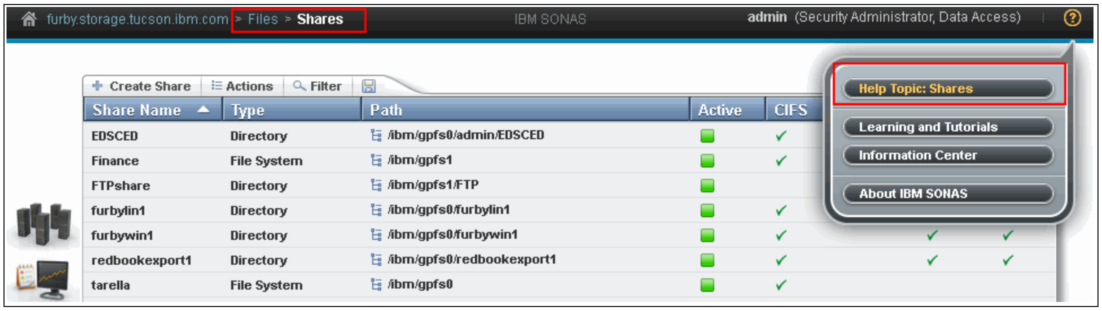
> **Note:** No context-sensitive help was available for IBM SONAS before Version 1.5.1.

There are four sections in the context-sensitive help menu:

- Help Topic: Shows the current help information.
- Learning and Tutorials: Provides a link to the internal SONAS information center.
- Information Center. Provides a link to internal SONAS information center.
- About IBM SONAS.: Shows the current SONAS version information.

Using context-sensitive help

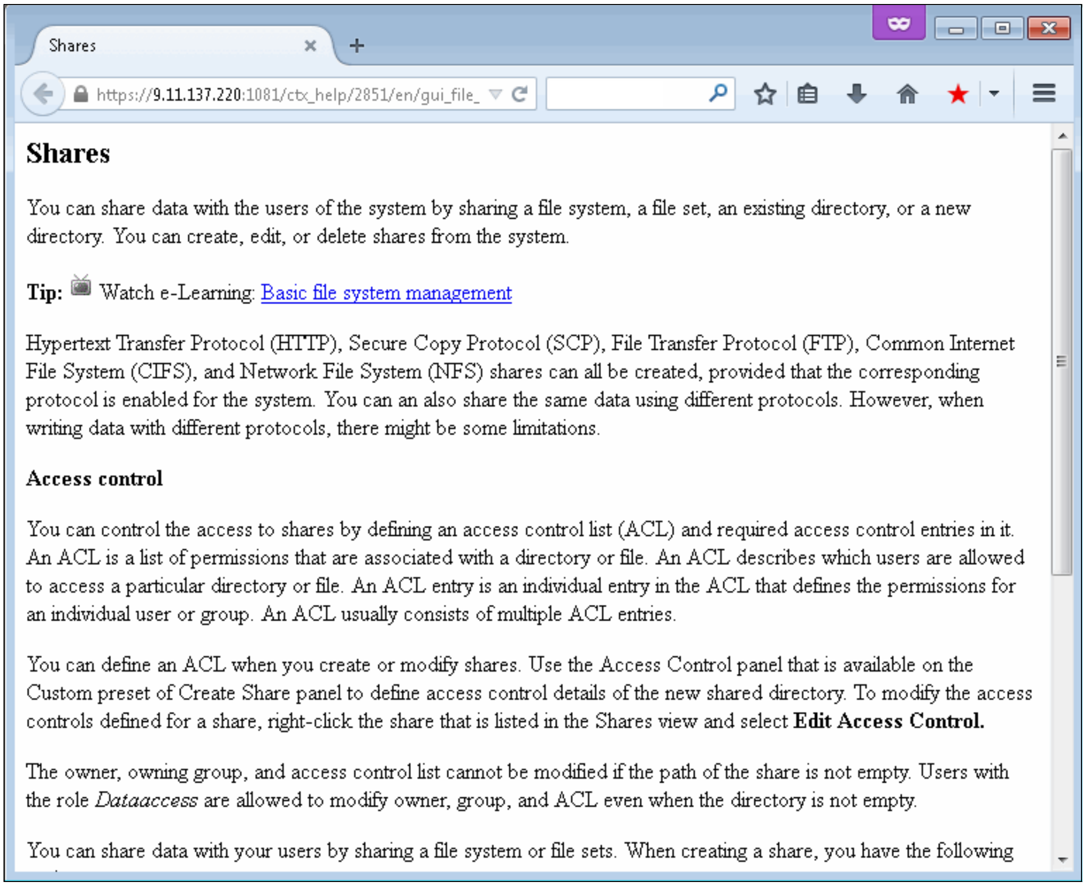
The context-sensitive help “Help” topic shows you the help for the current location. For example, if the current window shows the Shares window (accessed by clicking **Files** → **Shares**), the help topic also shows help for shares, as shown in Figure 7-4.



Share Name	Type	Path	Active	CIFS		
EDSCED	Directory	/ibm/gpfs0/admin/EDSCED	🟩	✓		
Finance	File System	/ibm/gpfs1	🟩	✓		
FTPshare	Directory	/ibm/gpfs1/FTP	🟩			
furbylin1	Directory	/ibm/gpfs0/furbylin1	🟩	✓		
furbywin1	Directory	/ibm/gpfs0/furbywin1	🟩	✓	✓	✓
redbookexport1	Directory	/ibm/gpfs0/redbookexport1	🟩	✓	✓	✓
tarella	File System	/ibm/gpfs0	🟩	✓		

Figure 7-4 Help topic shares

After you click **Help Topic** → **Shares**, you can see detailed descriptions in a new browser window, as shown in Figure 7-5 on page 497.



Shares

You can share data with the users of the system by sharing a file system, a file set, an existing directory, or a new directory. You can create, edit, or delete shares from the system.

Tip: Watch e-Learning: Basic file system management

Hypertext Transfer Protocol (HTTP), Secure Copy Protocol (SCP), File Transfer Protocol (FTP), Common Internet File System (CIFS), and Network File System (NFS) shares can all be created, provided that the corresponding protocol is enabled for the system. You can an also share the same data using different protocols. However, when writing data with different protocols, there might be some limitations.

Access control

You can control the access to shares by defining an access control list (ACL) and required access control entries in it. An ACL is a list of permissions that are associated with a directory or file. An ACL describes which users are allowed to access a particular directory or file. An ACL entry is an individual entry in the ACL that defines the permissions for an individual user or group. An ACL usually consists of multiple ACL entries.

You can define an ACL when you create or modify shares. Use the Access Control panel that is available on the Custom preset of Create Share panel to define access control details of the new shared directory. To modify the access controls defined for a share, right-click the share that is listed in the Shares view and select **Edit Access Control.**

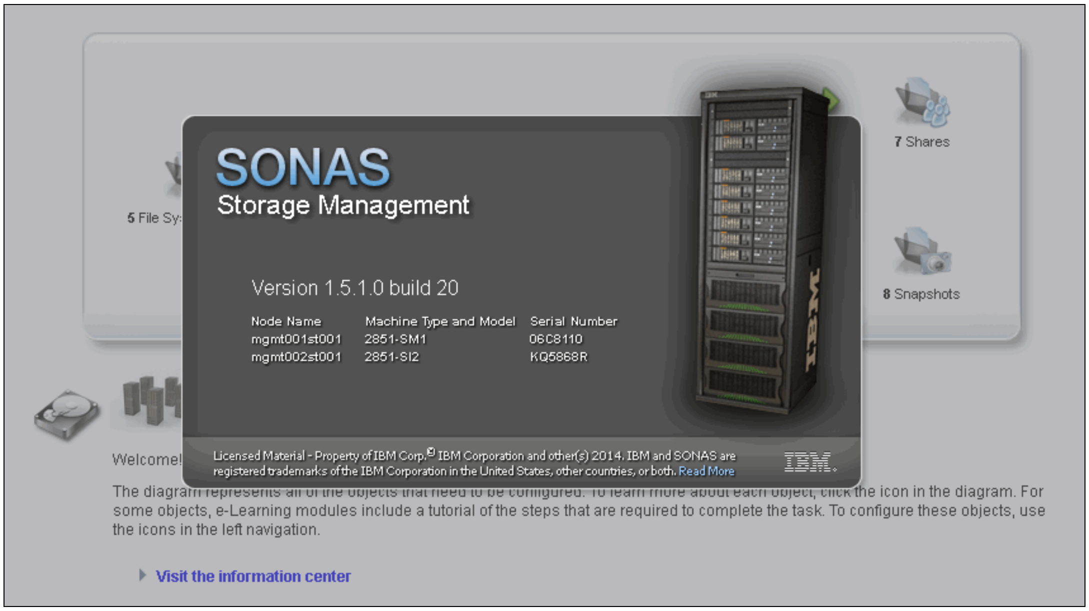
The owner, owning group, and access control list cannot be modified if the path of the share is not empty. Users with the role *Dataaccess* are allowed to modify owner, group, and ACL even when the directory is not empty.

You can share data with your users by sharing a file system or file sets. When creating a share, you have the following

Figure 7-5 Detail description of Shares

The Learning and Tutorial and Information Center links point to the SONAS internal information center page.

The About IBM SONAS link shows the SONAS version and Management node information, as shown in Figure 7-6.



Version 1.5.1.0 build 20

Node Name	Machine Type and Model	Serial Number
mgmt001st001	2851-SM1	06C8110
mgmt002st001	2851-SI2	KQ5868R

Licensed Material - Property of IBM Corp.© IBM Corporation and other(s) 2014. IBM and SONAS are registered trademarks of the IBM Corporation in the United States, other countries, or both. Read More

The diagram the icon in the diagram. For some objects, e-Learning modules include a tutorial of the steps that are required to complete the task. To configure these objects, use the icons in the left navigation.

Figure 7-6 Context-sensitive help about IBM SONAS

7.1.3 Navigational pane

The left pane shows the navigational bar of categories that provide links to any task on the cluster. Click the icons on the left to view the corresponding pane in a window.

The GUI navigational pane is divided into several fields:

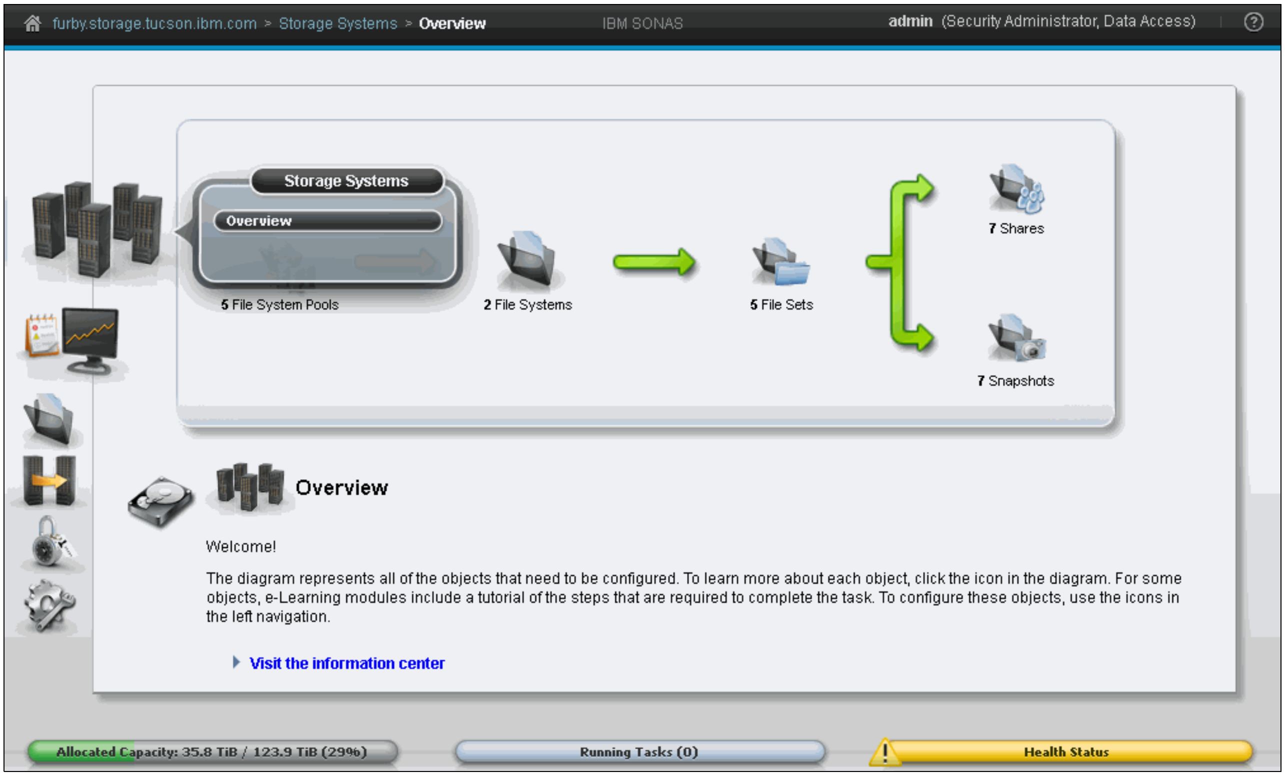
1. Storage Systems: View and manage the current system configuration.
2. Monitoring: View hardware components, manage and troubleshoot events, and monitor capacity and performance metrics on your system.
3. Files: Create and manage file systems, file sets, snapshots of file systems, and quotas.
4. Copy Services: Create and manage replication between two file systems for data recovery purposes.
5. Access: Manage and create user access to system functions and monitor audit records to determine user activity on the system.
6. Settings: Configure and manage event notifications, authentication, networks, and management GUI options.

The next section describes each of the categories.

Storage Systems

This category allows you to see basic usage of the SONAS system. It consists of two sections:

Overview: This window (Figure 7-7 on page 499) displays logical SONAS usage information, including number of file systems, number of file sets, and number of shares and snapshots. There are also links that go to e-Learning videos and to the internal information center.


furby.storage.tucson.ibm.com > Storage Systems > Overview
IBM SONAS
admin (Security Administrator, Data Access)
Storage Systems
Overview
5 File System Pools
2 File Systems
5 File Sets
7 Shares
7 Snapshots
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Overview

Welcome!

The diagram represents all of the objects that need to be configured. To learn more about each object, click the icon in the diagram. For some objects, e-Learning modules include a tutorial of the steps that are required to complete the task. To configure these objects, use the icons in the left navigation.

Visit the information center

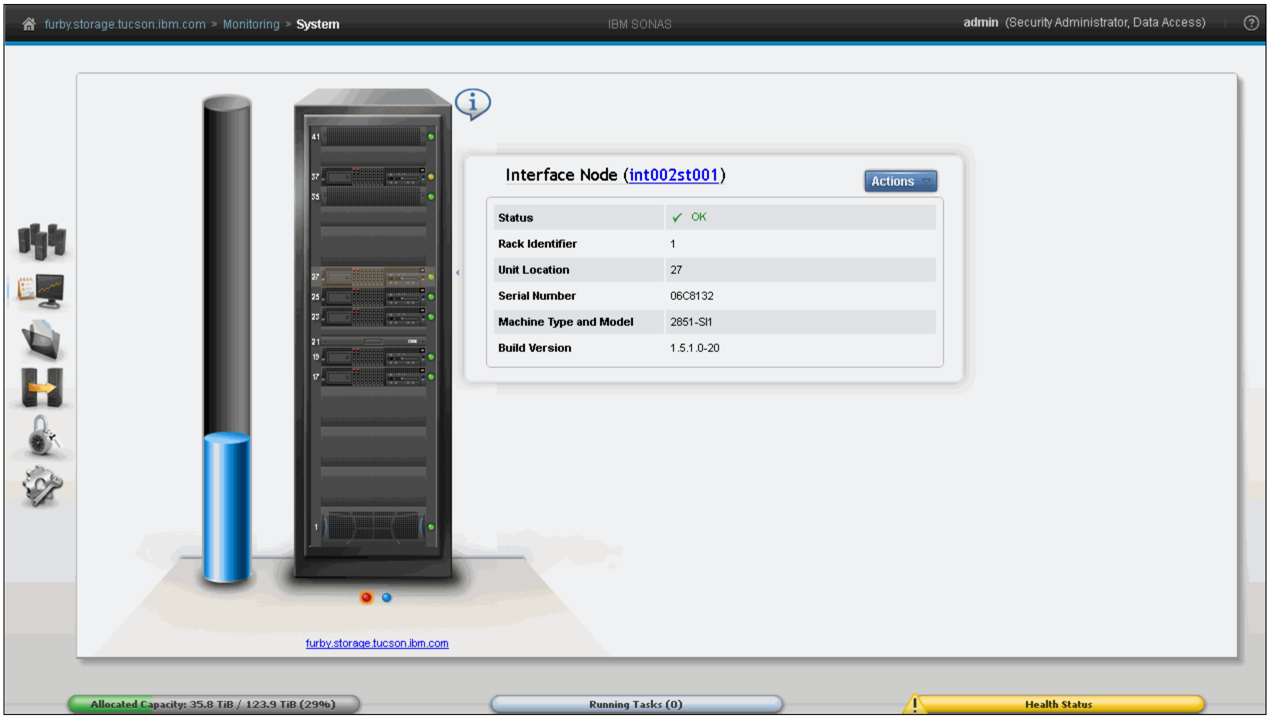
Figure 7-7 Storage Systems view of the SONAS cluster GUI

Monitoring

In this category and its menus, you can check various logs, basic system health with information for InfiniBand and Ethernet switches, and disk state, in addition to basic performance graphs. It consists of five sections:

1. System: This view shows animation of disk space on the left side and the entire SONAS rack on the right side with installed components: InfiniBand switches, Ethernet switches, Interface nodes, Storage nodes, and storage subsystem. Next to each component on the left side, a number represents the rack position, and on the right side, a light represents node health status. A green light means that all is OK. A yellow light means minor warnings. A red light means that an issue occurred. Red and yellow lights require the attention of the SONAS administrator.

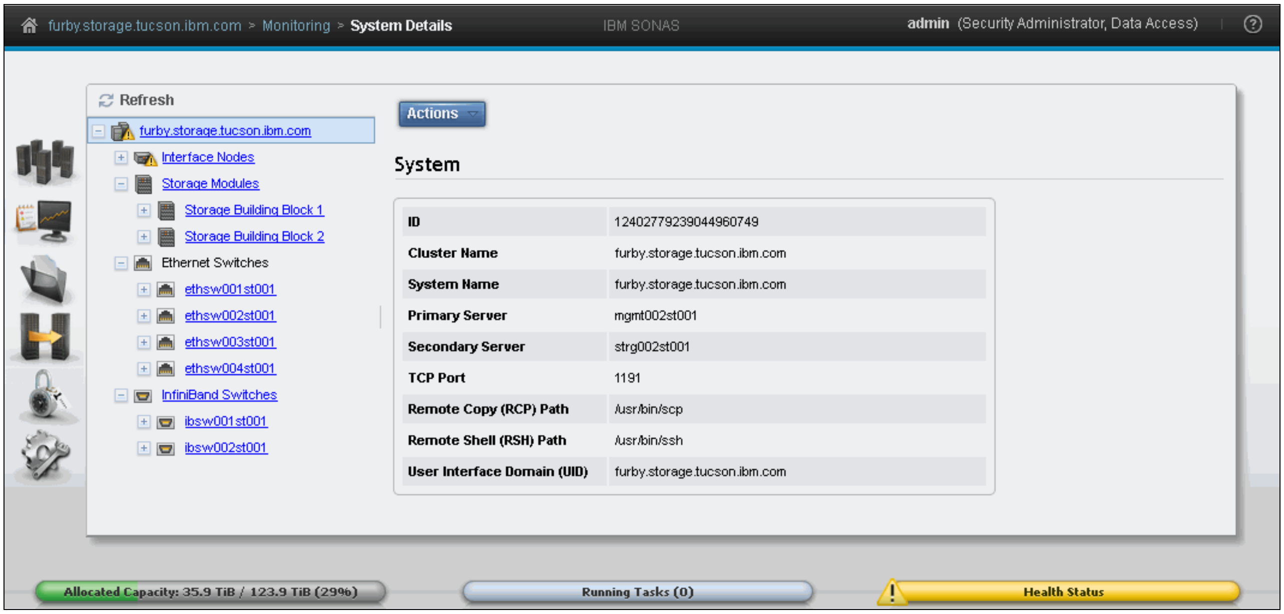
When you click one of the components, a new window opens, as shown in Figure 7-8. In this window, basic information about the selected component is shown. Clicking a **Warning** message redirects you to the next section of the Monitoring menu.


furby.storage.tucson.ibm.com > Monitoring > System
IBM SONAS
admin (Security Administrator, Data Access)
Interface Node (int002st001)
Actions
furby.storage.tucson.ibm.com
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Status	✓ OK
Rack Identifier	1
Unit Location	27
Serial Number	06C8132
Machine Type and Model	2851-SI1
Build Version	1.5.1.0-20

Figure 7-8 Main System view under the Monitoring section in the GUI

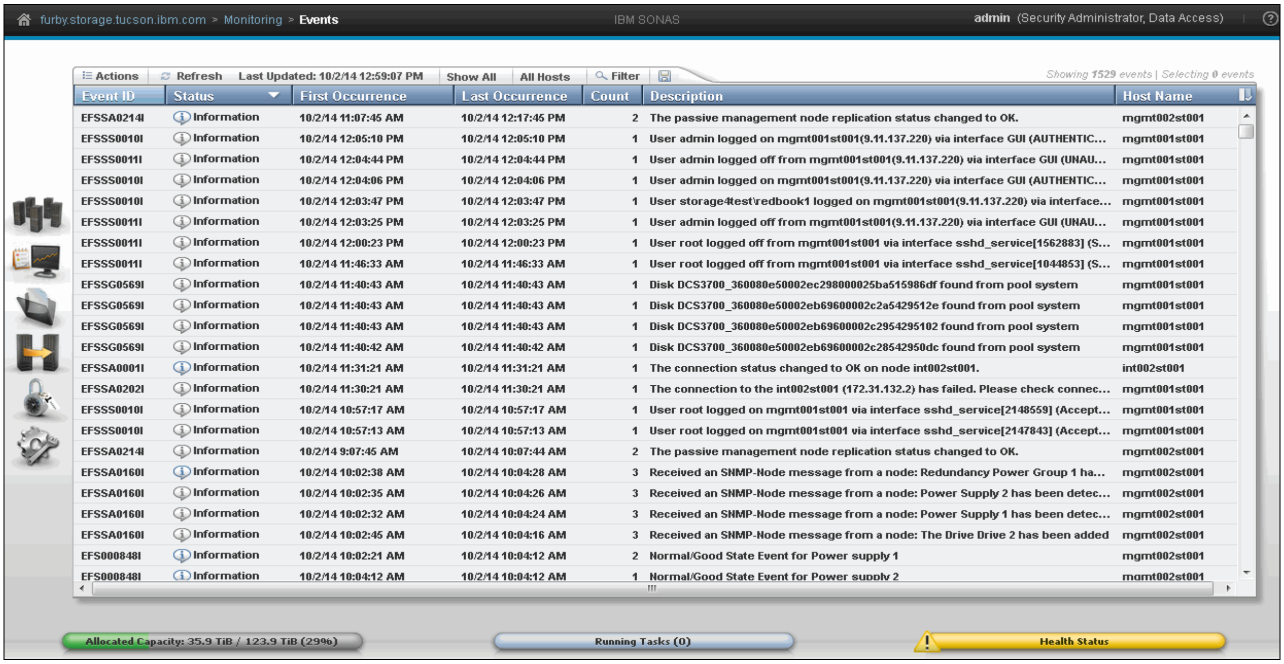
2. System Details: In this view, you can monitor the system health in great detail. You can select each node separately and view the hardware and software state summary, as shown in Figure 7-9.



ID	12402779239044960749
Cluster Name	furby.storage.tucson.ibm.com
System Name	furby.storage.tucson.ibm.com
Primary Server	mgmt002st001
Secondary Server	strg002st001
TCP Port	1191
Remote Copy (RCP) Path	/usr/bin/scp
Remote Shell (RSH) Path	/usr/bin/ssh
User Interface Domain (UID)	furby.storage.tucson.ibm.com

Figure 7-9 System details view under the monitoring section

3. Events: Here you can check all events on your SONAS system. You can select between these choices: Show All, Critical/Warning, and Current Critical/Warning. You can also filter by origin of the device, such as a specific node. See Figure 7-10.


furby.storage.tucson.ibm.com > Monitoring > Events
IBM SONAS
admin (Security Administrator, Data Access)
Actions
Refresh
Show All
All Hosts
Filter
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Last Updated: 10/2/14 12:59:07 PM

Showing 1529 events | Selecting 0 events

Event ID	Status	First Occurrence	Last Occurrence	Count	Description	Host Name
EFSSA0214I	Information	10/2/14 11:07:45 AM	10/2/14 12:17:45 PM	2	The passive management node replication status changed to OK.	mgmt002st001
EFSSS0010I	Information	10/2/14 12:05:10 PM	10/2/14 12:05:10 PM	1	User admin logged on mgmt001st001(9.11.137.220) via interface GUI (AUTHENTIC...	mgmt001st001
EFSSS0011I	Information	10/2/14 12:04:44 PM	10/2/14 12:04:44 PM	1	User admin logged off from mgmt001st001(9.11.137.220) via interface GUI (UNAU...	mgmt001st001
EFSSS0010I	Information	10/2/14 12:04:06 PM	10/2/14 12:04:06 PM	1	User admin logged on mgmt001st001(9.11.137.220) via interface GUI (AUTHENTIC...	mgmt001st001
EFSSS0010I	Information	10/2/14 12:03:47 PM	10/2/14 12:03:47 PM	1	User storage4test\redbook1 logged on mgmt001st001(9.11.137.220) via interface...	mgmt001st001
EFSSS0011I	Information	10/2/14 12:03:25 PM	10/2/14 12:03:25 PM	1	User admin logged off from mgmt001st001(9.11.137.220) via interface GUI (UNAU...	mgmt001st001
EFSSS0011I	Information	10/2/14 12:00:23 PM	10/2/14 12:00:23 PM	1	User root logged off from mgmt001st001 via interface sshd_service[1562883] (S...	mgmt001st001
EFSSS0011I	Information	10/2/14 11:46:33 AM	10/2/14 11:46:33 AM	1	User root logged off from mgmt001st001 via interface sshd_service[1044853] (S...	mgmt001st001
EFSSG0569I	Information	10/2/14 11:40:43 AM	10/2/14 11:40:43 AM	1	Disk DCS3700_360080e50002ec298000025ba515986df found from pool system	mgmt001st001
EFSSG0569I	Information	10/2/14 11:40:43 AM	10/2/14 11:40:43 AM	1	Disk DCS3700_360080e50002eb69600002c2a5429512e found from pool system	mgmt001st001
EFSSG0569I	Information	10/2/14 11:40:43 AM	10/2/14 11:40:43 AM	1	Disk DCS3700_360080e50002eb69600002c2954295102 found from pool system	mgmt001st001
EFSSG0569I	Information	10/2/14 11:40:42 AM	10/2/14 11:40:42 AM	1	Disk DCS3700_360080e50002eb69600002c28542950dc found from pool system	mgmt001st001
EFSSA0001I	Information	10/2/14 11:31:21 AM	10/2/14 11:31:21 AM	1	The connection status changed to OK on node int002st001.	int002st001
EFSSA0202I	Information	10/2/14 11:30:21 AM	10/2/14 11:30:21 AM	1	The connection to the int002st001 (172.31.132.2) has failed. Please check connec...	mgmt001st001
EFSSS0010I	Information	10/2/14 10:57:17 AM	10/2/14 10:57:17 AM	1	User root logged on mgmt001st001 via interface sshd_service[2148559] (Accept...	mgmt001st001
EFSSS0010I	Information	10/2/14 10:57:13 AM	10/2/14 10:57:13 AM	1	User root logged on mgmt001st001 via interface sshd_service[2147843] (Accept...	mgmt001st001
EFSSA0214I	Information	10/2/14 9:07:45 AM	10/2/14 10:07:44 AM	2	The passive management node replication status changed to OK.	mgmt002st001
EFSSA0160I	Information	10/2/14 10:02:38 AM	10/2/14 10:04:28 AM	3	Received an SNMP-Node message from a node: Redundancy Power Group 1 ha...	mgmt002st001
EFSSA0160I	Information	10/2/14 10:02:35 AM	10/2/14 10:04:26 AM	3	Received an SNMP-Node message from a node: Power Supply 2 has been detec...	mgmt002st001
EFSSA0160I	Information	10/2/14 10:02:32 AM	10/2/14 10:04:24 AM	3	Received an SNMP-Node message from a node: Power Supply 1 has been detec...	mgmt002st001
EFSSA0160I	Information	10/2/14 10:02:45 AM	10/2/14 10:04:16 AM	3	Received an SNMP-Node message from a node: The Drive Drive 2 has been added	mgmt002st001
EFS000848I	Information	10/2/14 10:02:21 AM	10/2/14 10:04:12 AM	2	Normal/Good State Event for Power supply 1	mgmt002st001
EFS000848I	Information	10/2/14 10:04:12 AM	10/2/14 10:04:12 AM	1	Normal/Good State Event for Power supply 2	mgmt002st001

Figure 7-10 Events view under the monitoring section

4. Performance: In the System view, graphs show client network throughput, cluster throughput, cluster latency, and cluster operations data for the system. You can select the time frame (minute, hour, day, week, month, quarter, or year) in which data is displayed. See Figure 7-11.

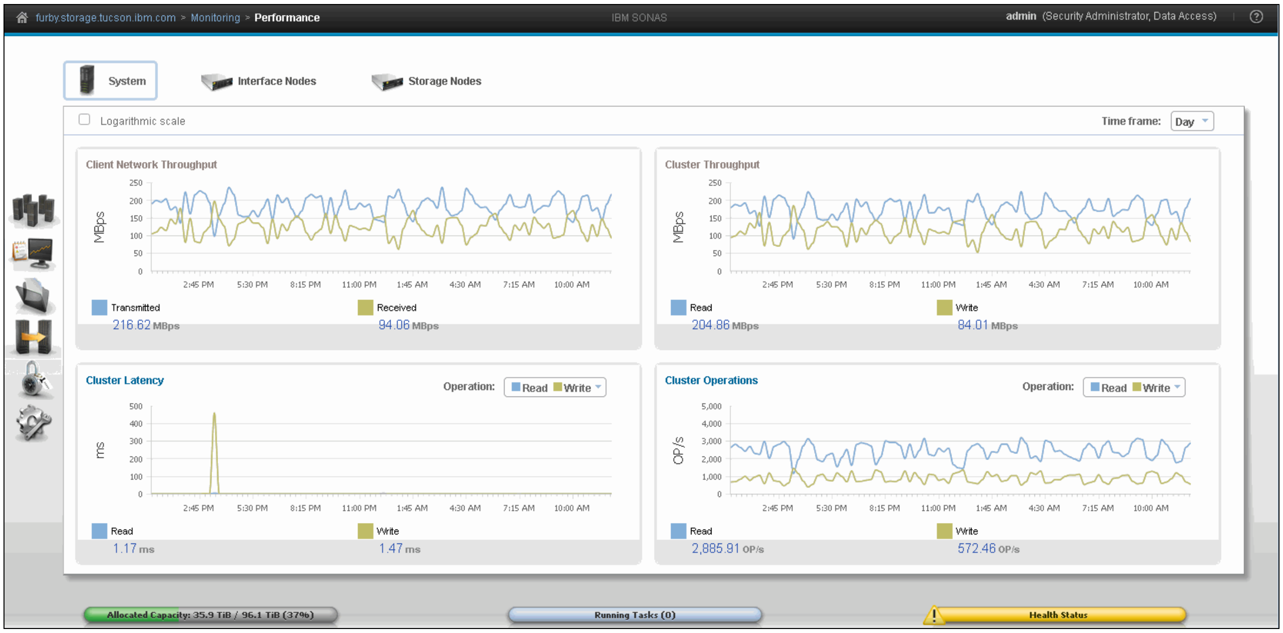
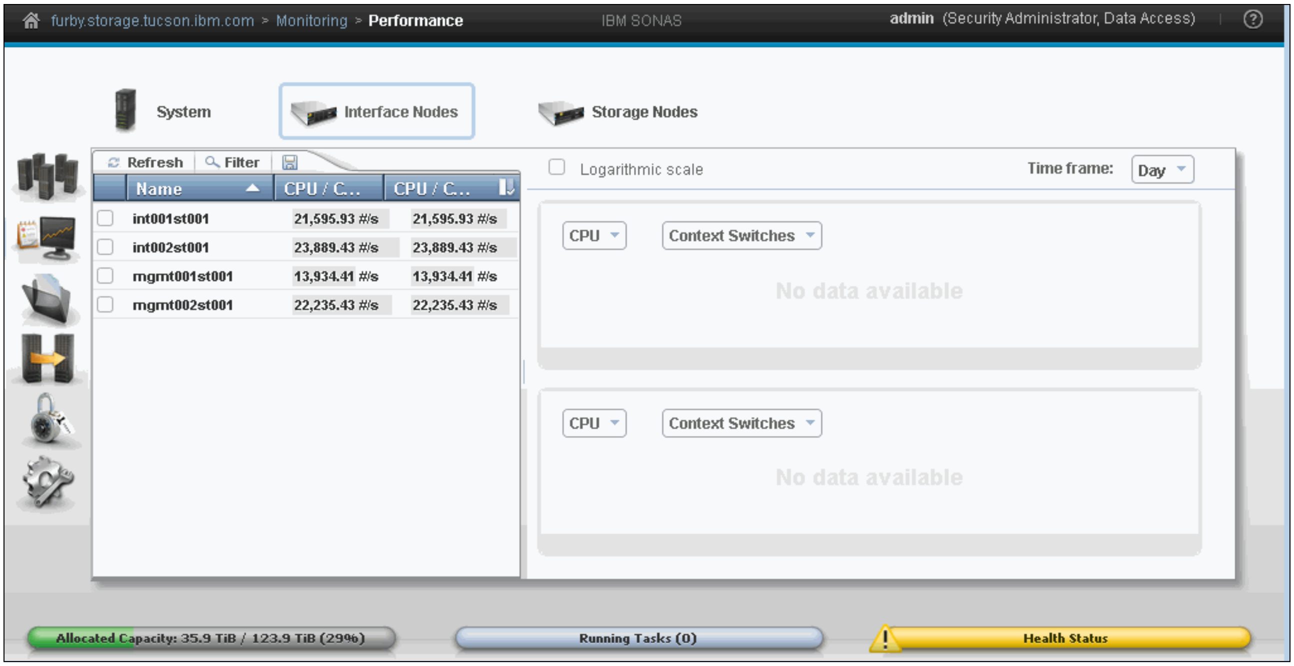


Figure 7-11 Performance view of the Monitoring section

The Interface Nodes and Storage Nodes icons displays graphs that show processor, memory, and public network data for each Interface node and Storage node in the system. You select the metrics for which you want to view data. You can select the time frame (minute, hour, day, week, month, quarter, or year) in which data is displayed. See Figure 7-12.



Name	CPU / C...	CPU / C...
□ int001st001	21,595.93 #/s	21,595.93 #/s
□ int002st001	23,889.43 #/s	23,889.43 #/s
□ mgmt001st001	13,934.41 #/s	13,934.41 #/s
□ mgmt002st001	22,235.43 #/s	22,235.43 #/s

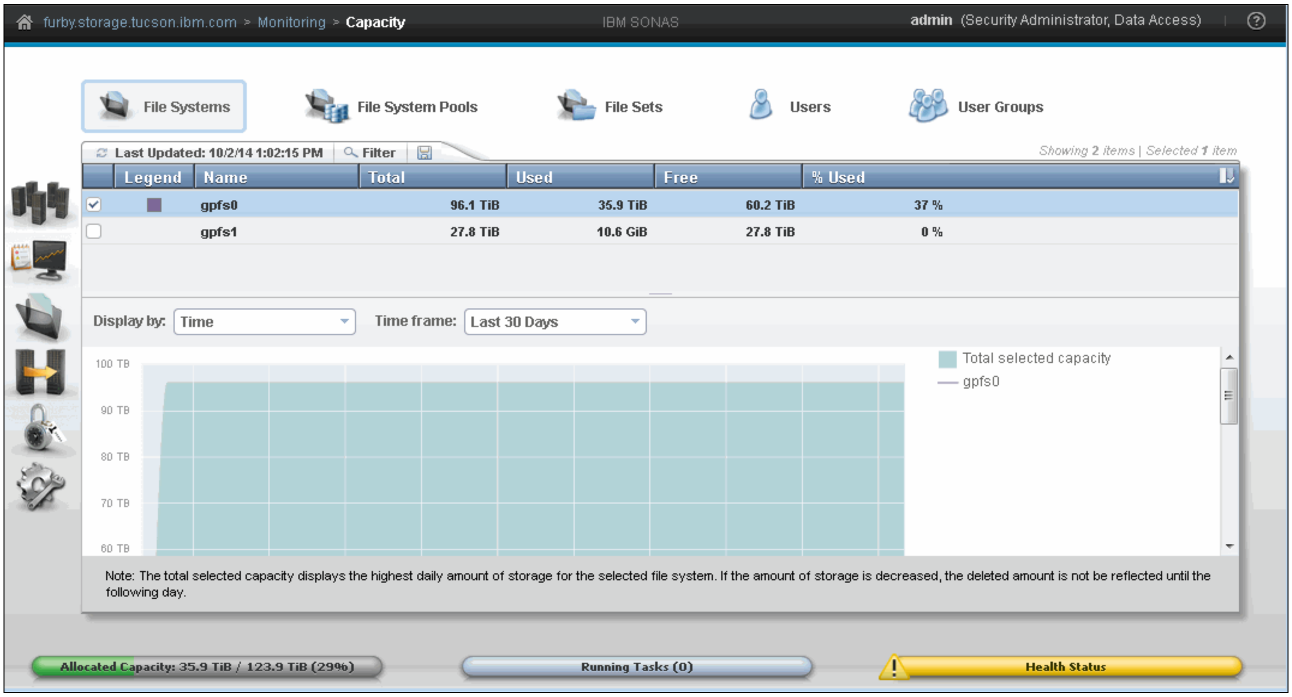
□ Logarithmic scale

Time frame: Day

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)

Figure 7-12 Interface Nodes tab graphs

5. Capacity: The Capacity window is used for checking graphs of used space. You can check for a specific GPFS file system, file system pool and specific file set, user, or user groups for a selected file system. Charts can be displayed in specific time or percentages, as shown in Figure 7-13 on page 503.


furby.storage.tucson.ibm.com > Monitoring > Capacity
IBM SONAS
admin (Security Administrator, Data Access)
File Systems
File System Pools
File Sets
Users
User Groups
Last Updated: 10/2/14 1:02:15 PM
Filter
Showing 2 items | Selected 1 item
Total selected capacity
gpfs0
100 TB
90 TB
80 TB
70 TB
60 TB
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

	Legend	Name	Total	Used	Free	% Used
☑	■	gpfs0	96.1 TiB	35.9 TiB	60.2 TiB	37 %
□		gpfs1	27.8 TiB	10.6 GiB	27.8 TiB	0 %

Display by: Time
Time frame: Last 30 Days

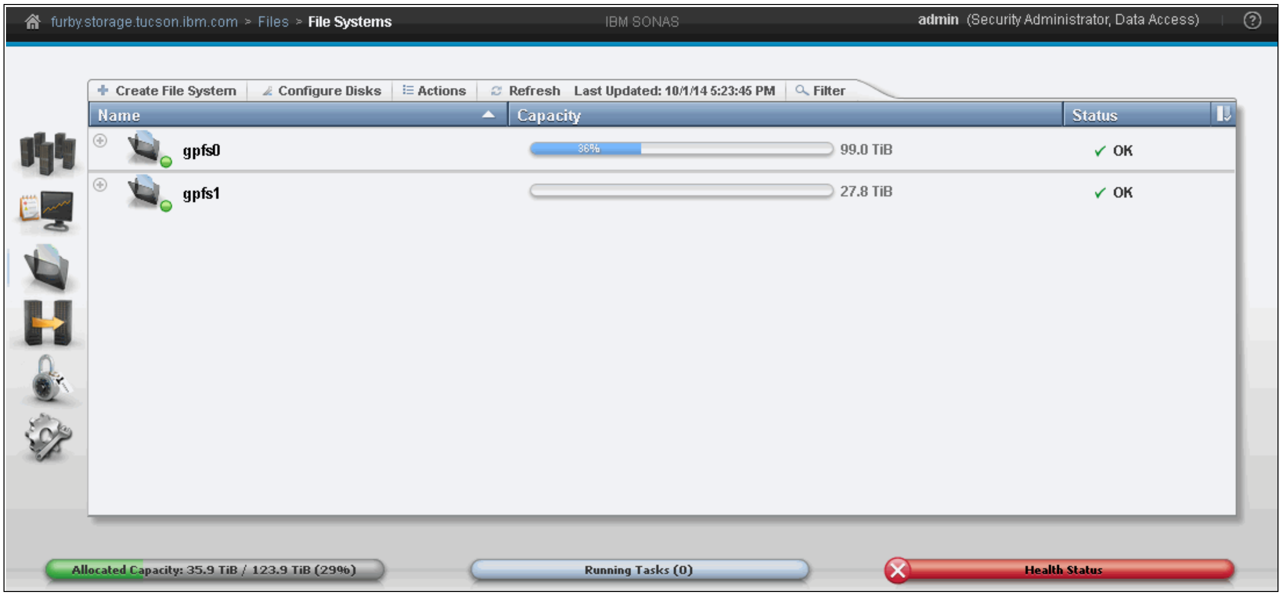
Note: The total selected capacity displays the highest daily amount of storage for the selected file system. If the amount of storage is decreased, the deleted amount is not be reflected until the following day.

Figure 7-13 Capacity view of the Monitoring section

Files

This category allows you to do file-system-related tasks. You can create file systems, exports, file sets, snapshots, and more. Each of the tasks that can be done are described here.

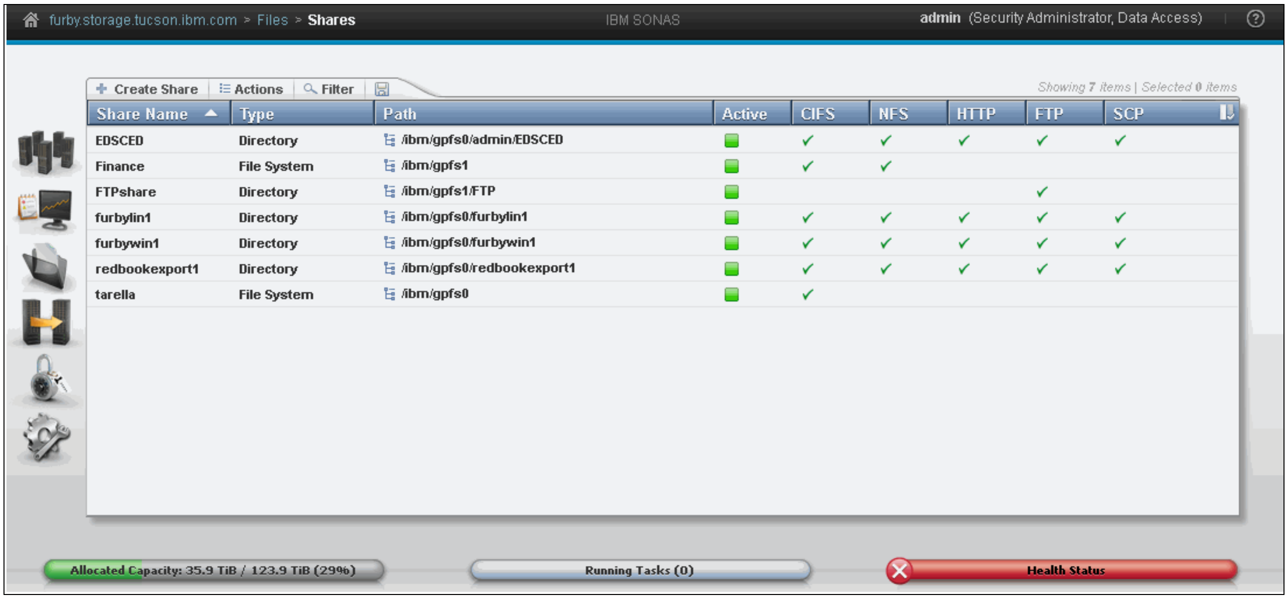
1. File Systems: In this window, you can create a file system, or edit or delete a file system (see Figure 7-14).


furby.storage.tucson.ibm.com > Files > File Systems
IBM SONAS
admin (Security Administrator, Data Access)
Create File System
Configure Disks
Actions
Refresh
Last Updated: 10/1/14 5:23:45 PM
Filter
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Name	Capacity	Status
gpfs0	36% 99.0 TiB	✓ OK
gpfs1	27.8 TiB	✓ OK

Figure 7-14 File Systems view of the Files section

2. Shares: In this window, you can create, edit, or delete shares on your cluster. In the primary view, you can see all created shares and their basic characteristics, such as type of share (see Figure 7-15).

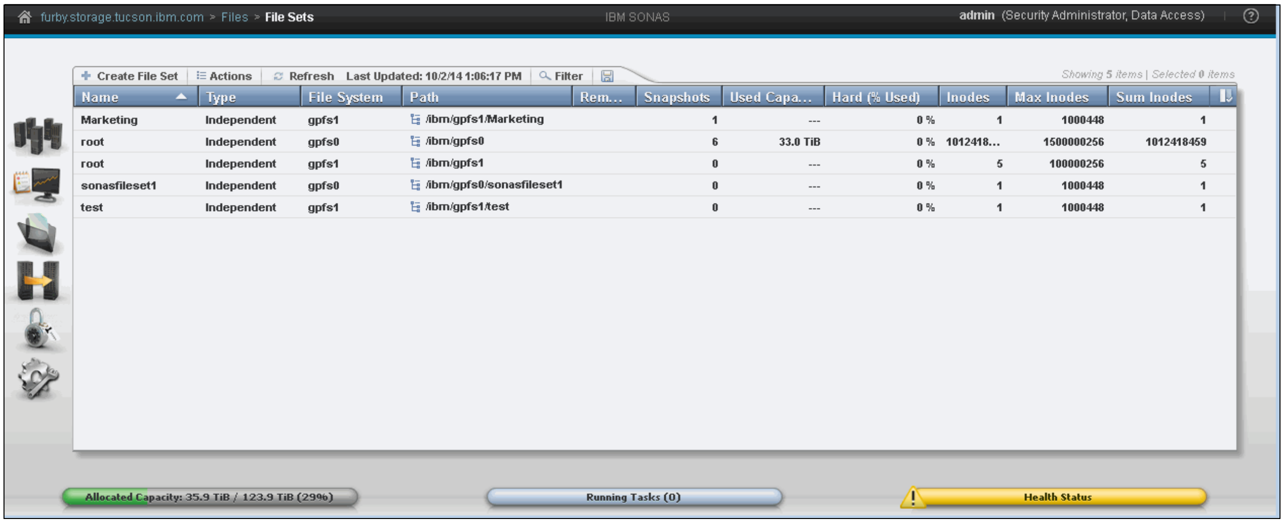

furby.storage.tucson.ibm.com > Files > Shares
IBM SONAS
admin (Security Administrator, Data Access)
Create Share
Actions
Filter
Showing 7 items | Selected 0 items

Share Name	Type	Path	Active	CIFS	NFS	HTTP	FTP	SCP
EDSCED	Directory	/ibm/gpfs0/admin/EDSCED	🟩	✓	✓	✓	✓	✓
Finance	File System	/ibm/gpfs1	🟩	✓	✓			
FTPshare	Directory	/ibm/gpfs1/FTP	🟩				✓	
furbylin1	Directory	/ibm/gpfs0/furbylin1	🟩	✓	✓	✓	✓	✓
furbywin1	Directory	/ibm/gpfs0/furbywin1	🟩	✓	✓	✓	✓	✓
redbookexport1	Directory	/ibm/gpfs0/redbookexport1	🟩	✓	✓	✓	✓	✓
tarella	File System	/ibm/gpfs0	🟩	✓				

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Figure 7-15 Shares view of the Files section

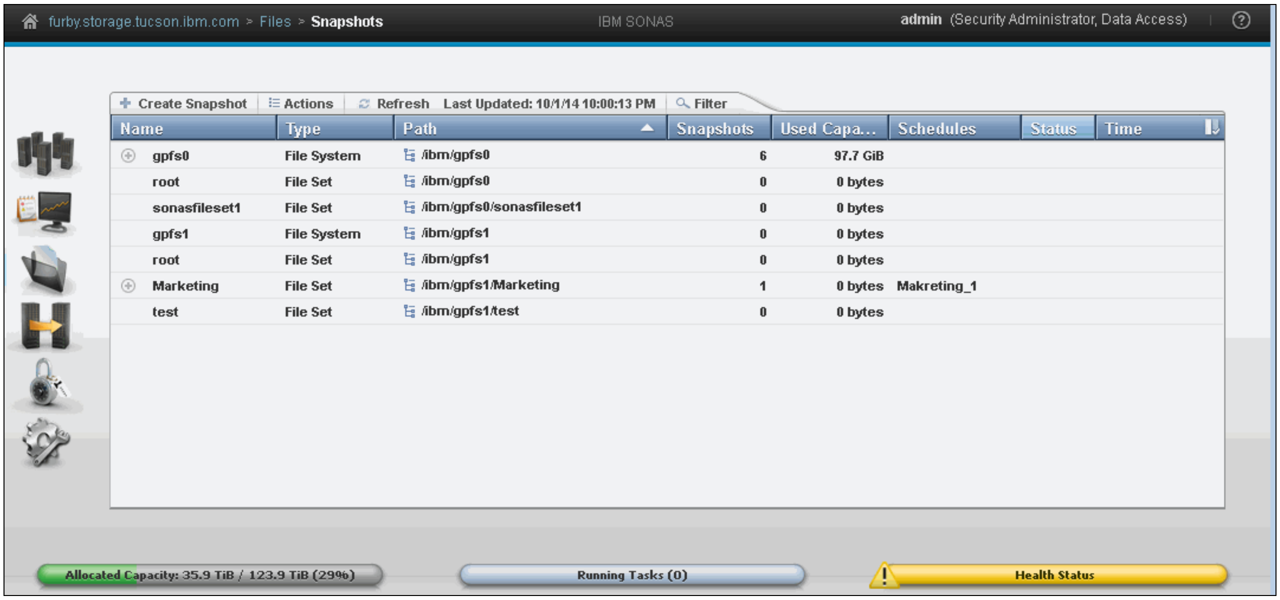
3. File Sets: This window displays basic information and also allows you to create, delete, or modify file sets, as shown in Figure 7-16.


furby.storage.tucson.ibm.com > Files > File Sets
IBM SONAS
admin (Security Administrator, Data Access)
Create File Set
Actions
Refresh
Last Updated: 10/2/14 1:06:17 PM
Filter
Showing 5 items | Selected 0 items

Name	Type	File System	Path	Rem...	Snapshots	Used Capa...	Hard (% Used)	Inodes	Max Inodes	Sum Inodes
Marketing	Independent	gpfs1	/ibm/gpfs1/Marketing		1	---	0 %	1	1000448	1
root	Independent	gpfs0	/ibm/gpfs0		6	33.0 TiB	0 %	1012418...	1500000256	1012418459
root	Independent	gpfs1	/ibm/gpfs1		0	---	0 %	5	100000256	5
sonasfileset1	Independent	gpfs0	/ibm/gpfs0/sonasfileset1		0	---	0 %	1	1000448	1
test	Independent	gpfs1	/ibm/gpfs1/test		0	---	0 %	1	1000448	1

Figure 7-16 File Sets view of the Files section

4. Snapshots: In this window, you can create, delete, and modify a snapshot. In the first view, the existing snapshots can be seen with detailed information (see Figure 7-17).

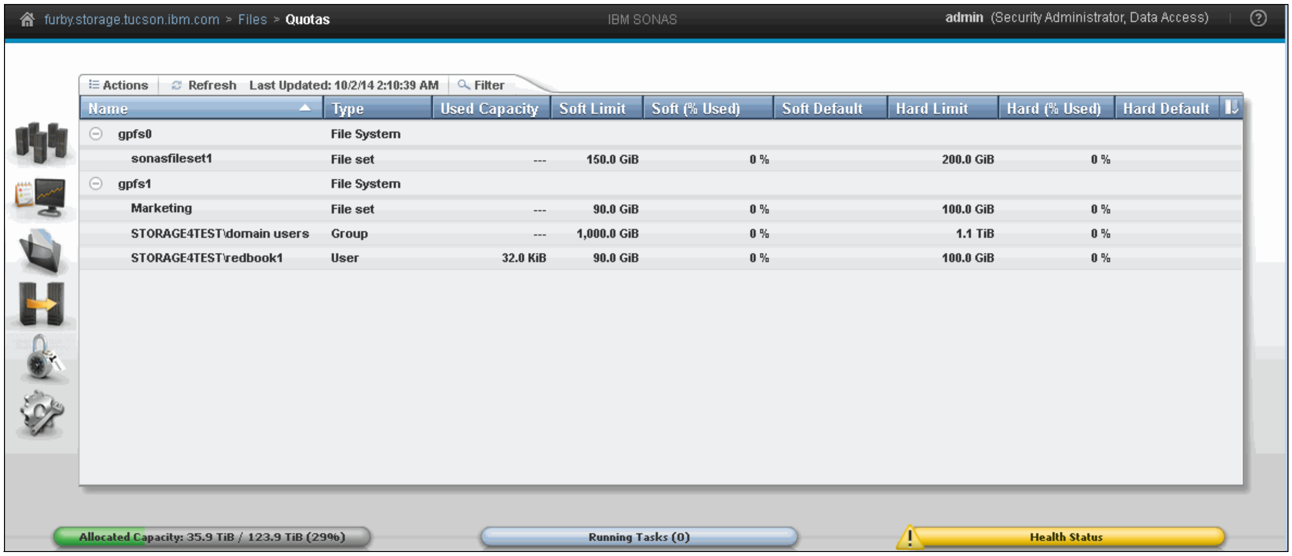

furby.storage.tucson.ibm.com > Files > Snapshots
IBM SONAS
admin (Security Administrator, Data Access)
Create Snapshot
Actions
Refresh
Last Updated: 10/1/14 10:00:13 PM
Filter

Name	Type	Path	Snapshots	Used Capa...	Schedules	Status	Time
gpfs0	File System	/ibm/gpfs0	6	97.7 GiB			
root	File Set	/ibm/gpfs0	0	0 bytes			
sonasfileset1	File Set	/ibm/gpfs0/sonasfileset1	0	0 bytes			
gpfs1	File System	/ibm/gpfs1	0	0 bytes			
root	File Set	/ibm/gpfs1	0	0 bytes			
Marketing	File Set	/ibm/gpfs1/Marketing	1	0 bytes	Makreting_1		
test	File Set	/ibm/gpfs1/test	0	0 bytes			

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Figure 7-17   Snapshots view of the Files section

5. Quotas: This window displays information about quotas. You can also create a quota or modify and delete existing ones. See Figure 7-18.

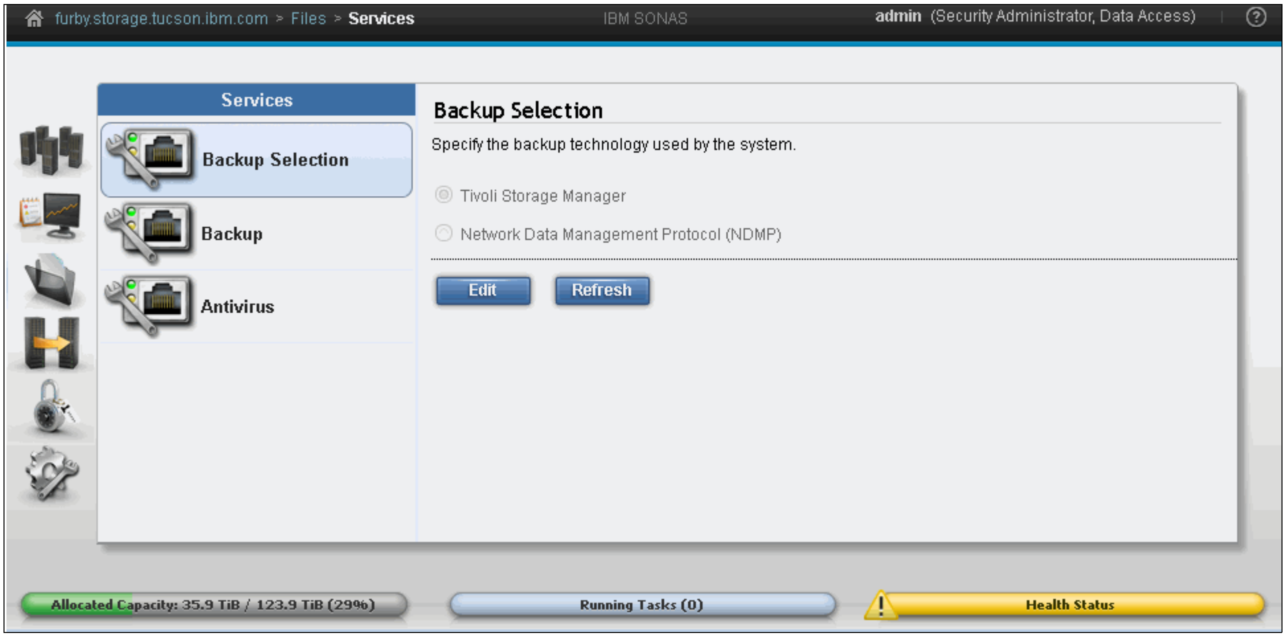

furby.storage.tucson.ibm.com > Files > Quotas
IBM SONAS
admin (Security Administrator, Data Access)
Actions
Refresh
Last Updated: 10/2/14 2:10:39 AM
Filter

Name	Type	Used Capacity	Soft Limit	Soft (% Used)	Soft Default	Hard Limit	Hard (% Used)	Hard Default
gpfs0	File System							
sonasfileset1	File set	---	150.0 GiB	0 %		200.0 GiB	0 %	
gpfs1	File System							
Marketing	File set	---	90.0 GiB	0 %		100.0 GiB	0 %	
STORAGE4TEST\domain users	Group	---	1,000.0 GiB	0 %		1.1 TiB	0 %	
STORAGE4TEST\redbook1	User	32.0 KiB	90.0 GiB	0 %		100.0 GiB	0 %	

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Figure 7-18   Quotas view of the Files section

6. Services: In this window, you manage backups and antivirus scans. For antivirus or Tivoli Storage Manager use, you must configure a separate server with these capabilities. See Figure 7-19.


furby.storage.tucson.ibm.com > Files > Services
IBM SONAS
admin (Security Administrator, Data Access)
Services
Backup Selection
Backup
Antivirus
Edit
Refresh
Running Tasks (0)
Health Status

Backup Selection

Specify the backup technology used by the system.

○ Tivoli Storage Manager

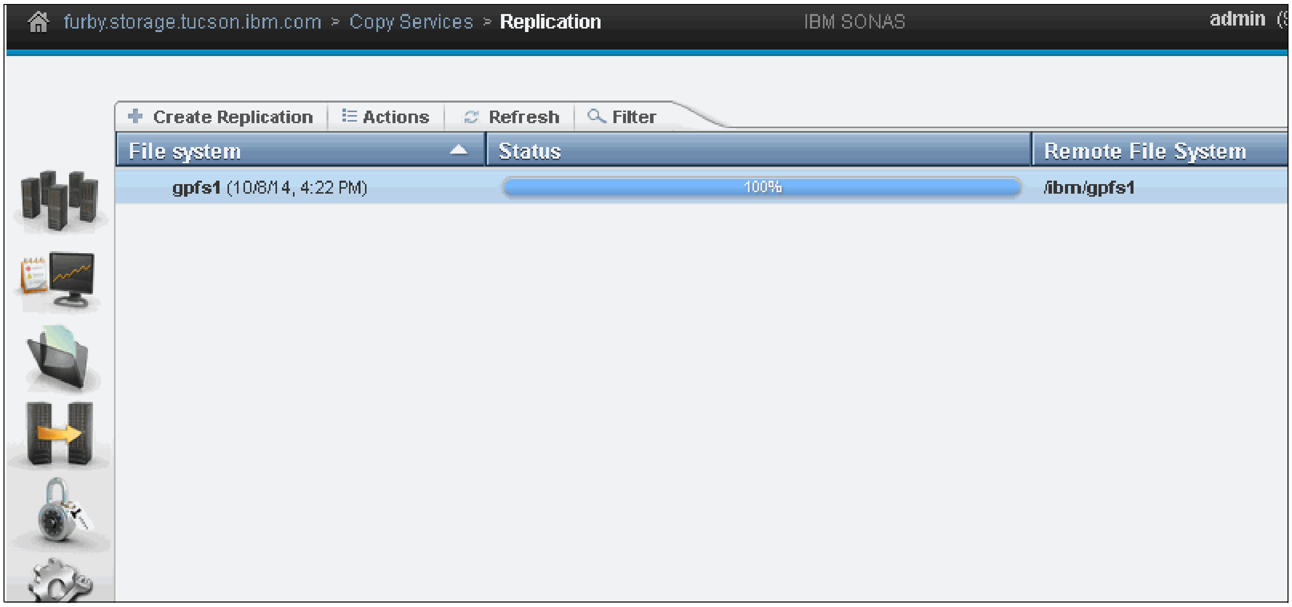
○ Network Data Management Protocol (NDMP)

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)

Figure 7-19 Services view of the Files section

Replication

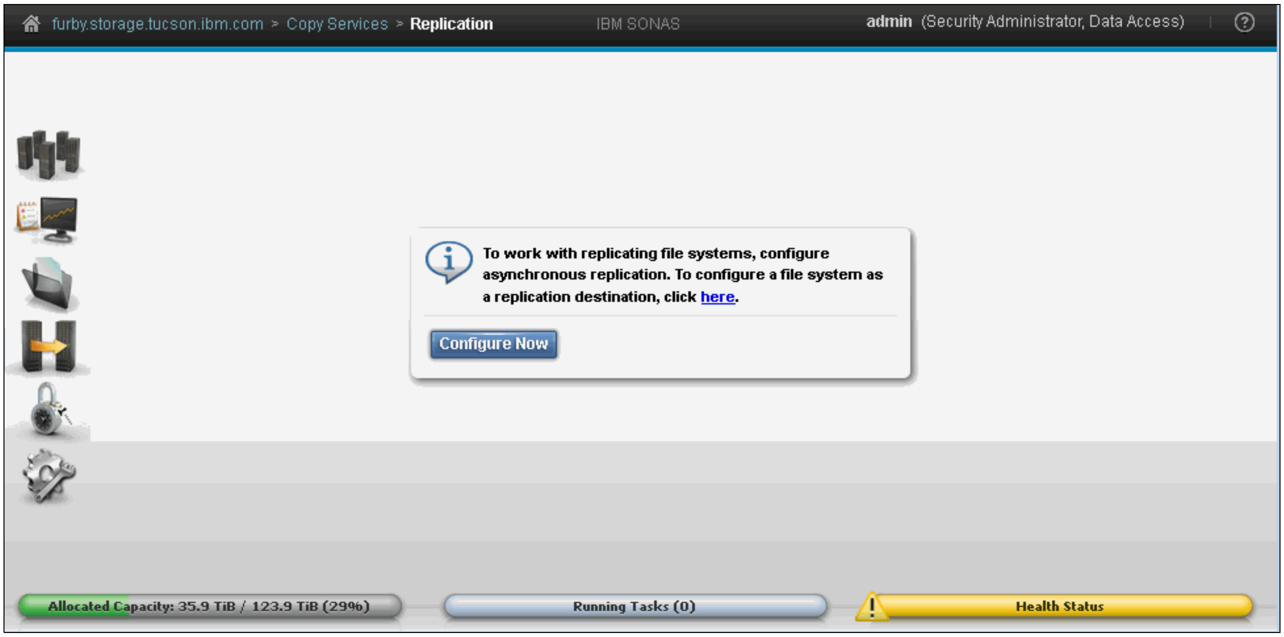
Here you can configure replication services between two SONAS systems. This configuration requires actions on both SONAS systems. Chapter 6, “Backup and recovery, availability, and resiliency functions” on page 367 guides you through the process of replication of SONAS data. In Figure 7-20, you can see configured replication on the example system. If no configuration exists, a window opens and shows the configuration option, as shown in Figure 7-21 on page 507. Figure 7-20 shows one configured replication task.



File system	Status	Remote File System
gpfs1 (10/8/14, 4:22 PM)	100%	/ibm/gpfs1

Figure 7-20 Configured replication between two SONAS systems for file system gpfs0

Figure 7-21 shows replication between two SONAS systems.


furby.storage.tucson.ibm.com > Copy Services > Replication
IBM SONAS
admin (Security Administrator, Data Access)

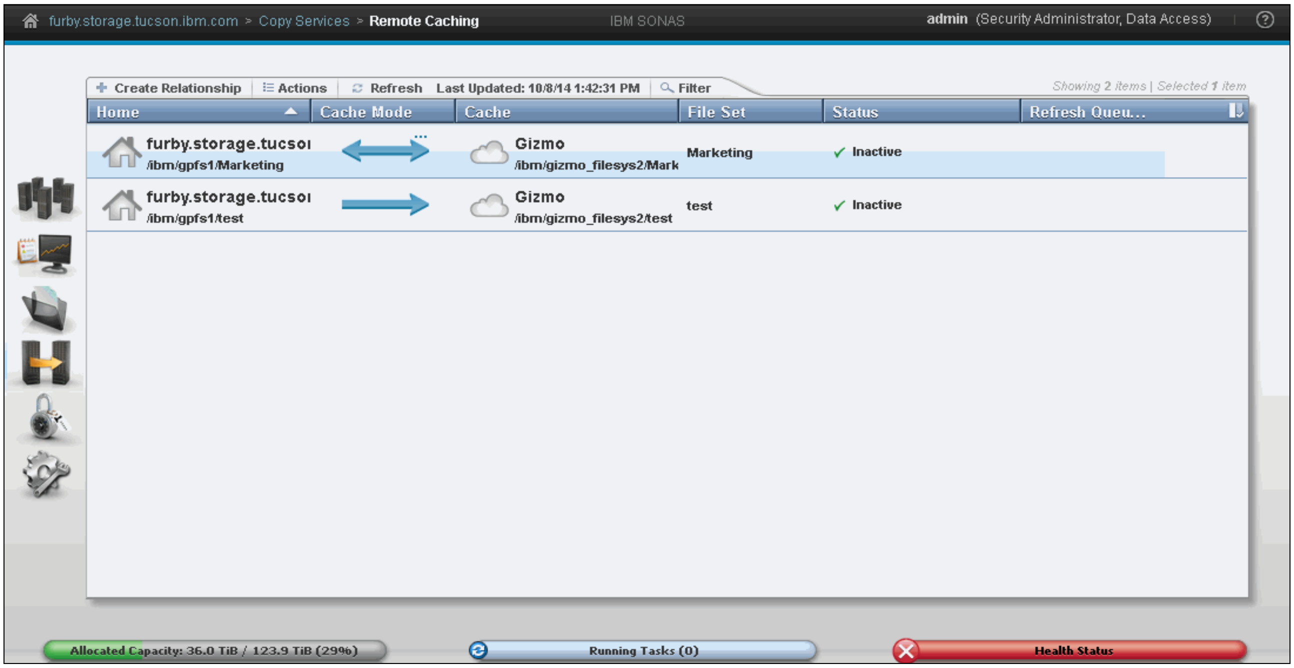
To work with replicating file systems, configure asynchronous replication. To configure a file system as a replication destination, click here.

Configure Now
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Figure 7-21 Window to start configuring a replication between two SONAS systems

Remote Caching

Here you can configure remote caching services between two or more SONAS systems. Chapter 6, "Backup and recovery, availability, and resiliency functions" on page 367 guides you through the process of configuring remote caching. In Figure 7-22, you can see configured remote caching on the example system.


furby.storage.tucson.ibm.com > Copy Services > Remote Caching
IBM SONAS
admin (Security Administrator, Data Access)
Create Relationship
Actions
Refresh
Last Updated: 10/8/14 1:42:31 PM
Filter
Showing 2 items | Selected 1 item

Home	Cache Mode	Cache	File Set	Status	Refresh Queu...
furby.storage.tucsoı /ibm/gpfs1/Marketing		Gizmo /ibm/gizmo_filesys2/Mark	Marketing	✓ Inactive	
furby.storage.tucsoı /ibm/gpfs1/test		Gizmo /ibm/gizmo_filesys2/test	test	✓ Inactive	

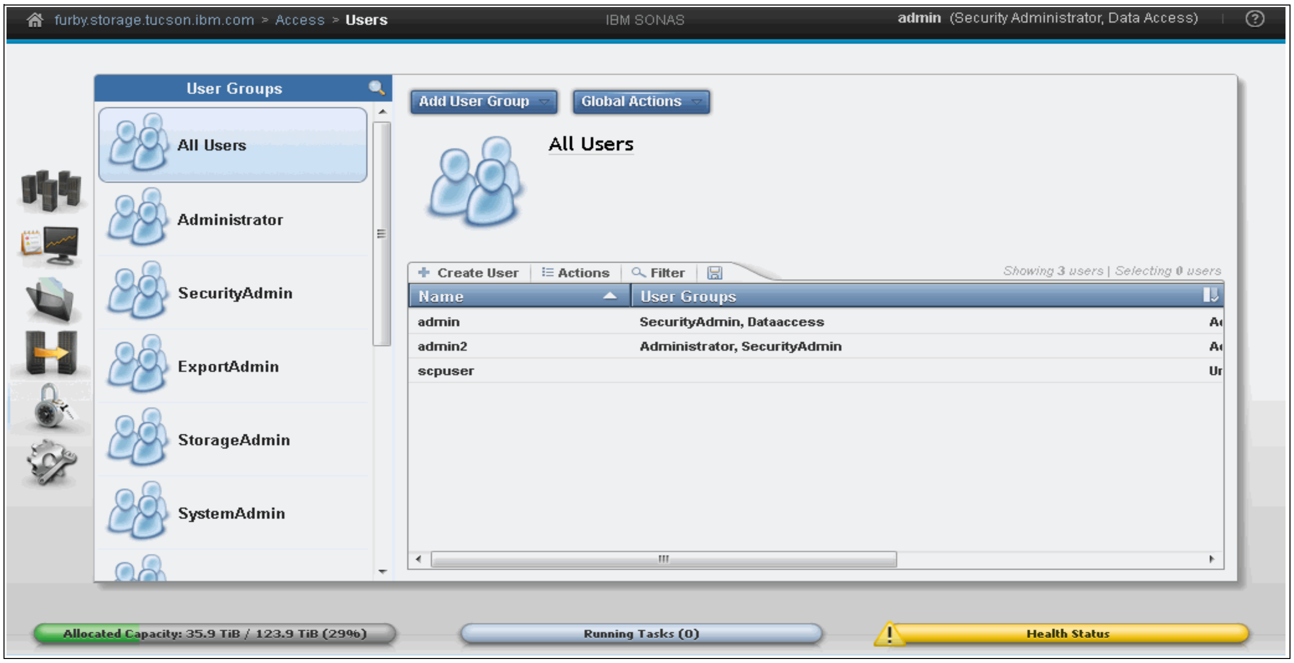
Allocated Capacity: 36.0 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Figure 7-22 Configured remote caching between two SONAS systems

Access

This category allows you to create users, user groups with different permissions, and check the audit log. It is consists of four sections:

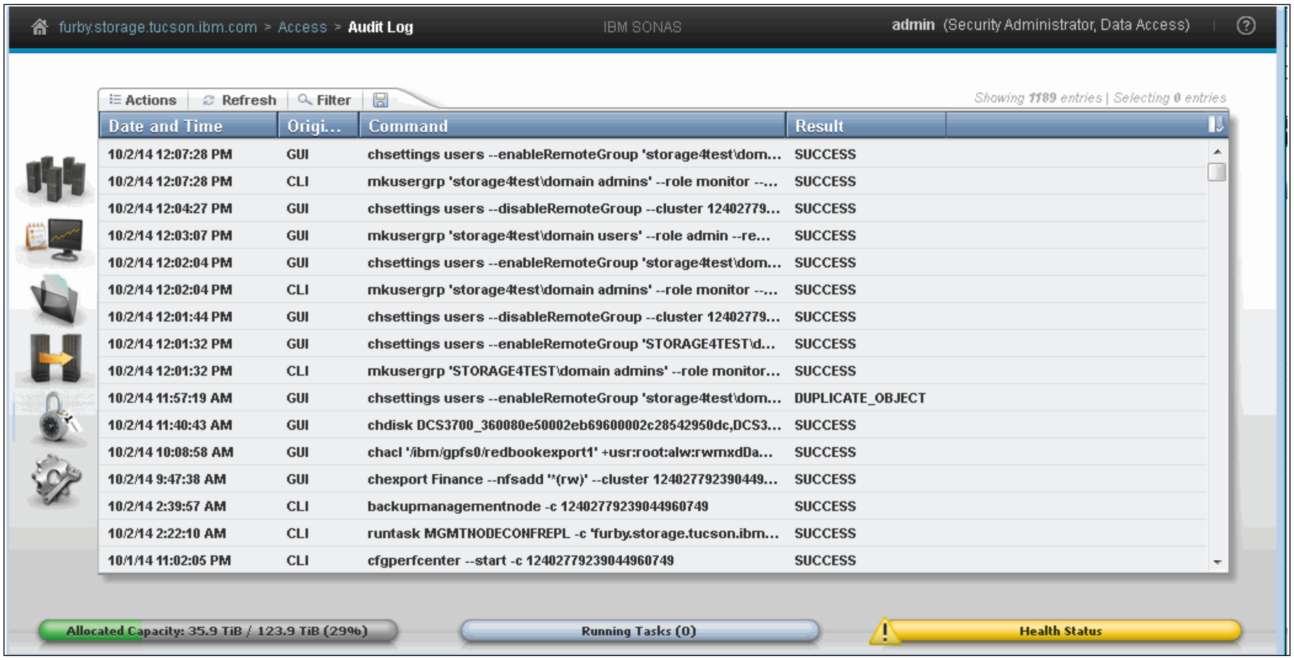
1. Users: In this section, you can create users, user groups, and change privileges for different users or groups. See Figure 7-23.



Name	User Groups	
admin	SecurityAdmin, Dataaccess	Ac
admin2	Administrator, SecurityAdmin	Ac
scpuser		Ur

Figure 7-23 Users view in Access section

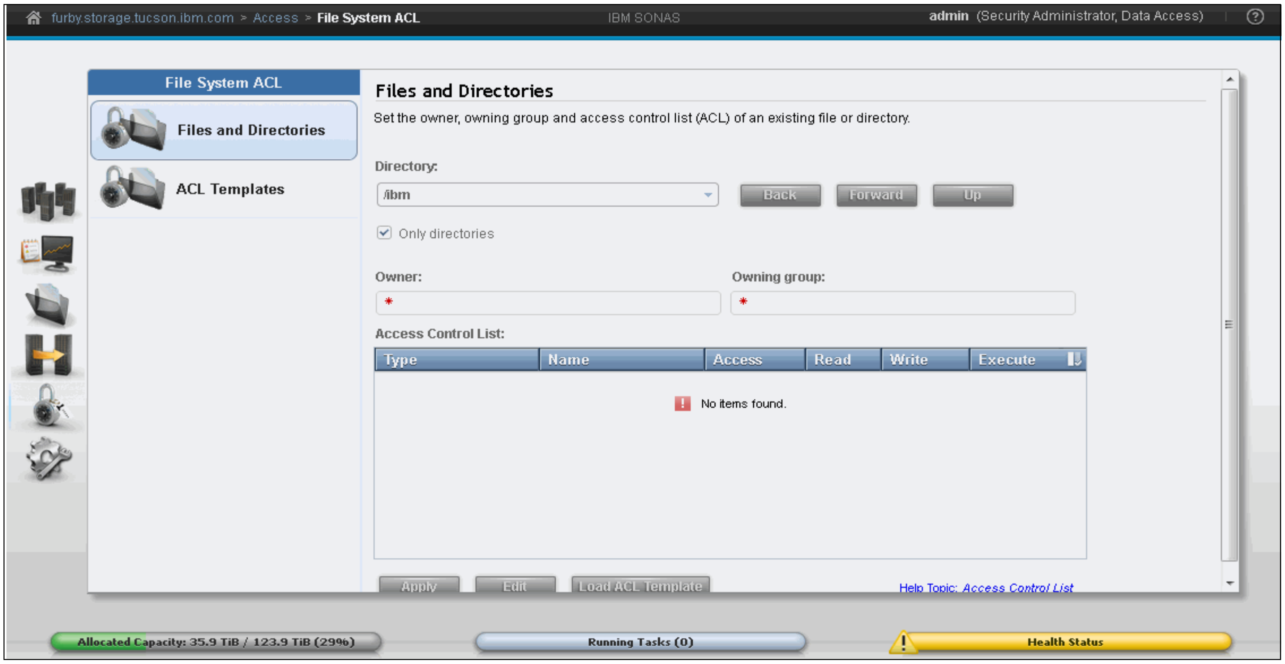
2. Audit Log: The audit log shows all commands that are run on the SONAS system by all users. You can also filter the view by the origin of commands, such as CLI or GUI. See Figure 7-24.


furby.storage.tucson.ibm.com > Access > Audit Log
IBM SONAS
admin (Security Administrator, Data Access)
Actions
Refresh
Filter
Showing 1189 entries | Selecting 0 entries
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Date and Time	Origi...	Command	Result
10/2/14 12:07:28 PM	GUI	chsettings users --enableRemoteGroup 'storage4test\dom...	SUCCESS
10/2/14 12:07:28 PM	CLI	mkusergrp 'storage4test\domain admins' --role monitor --...	SUCCESS
10/2/14 12:04:27 PM	GUI	chsettings users --disableRemoteGroup --cluster 12402779...	SUCCESS
10/2/14 12:03:07 PM	GUI	mkusergrp 'storage4test\domain users' --role admin --re...	SUCCESS
10/2/14 12:02:04 PM	GUI	chsettings users --enableRemoteGroup 'storage4test\dom...	SUCCESS
10/2/14 12:02:04 PM	CLI	mkusergrp 'storage4test\domain admins' --role monitor --...	SUCCESS
10/2/14 12:01:44 PM	GUI	chsettings users --disableRemoteGroup --cluster 12402779...	SUCCESS
10/2/14 12:01:32 PM	GUI	chsettings users --enableRemoteGroup 'STORAGE4TEST\d...	SUCCESS
10/2/14 12:01:32 PM	CLI	mkusergrp 'STORAGE4TEST\domain admins' --role monitor...	SUCCESS
10/2/14 11:57:19 AM	GUI	chsettings users --enableRemoteGroup 'storage4test\dom...	DUPLICATE_OBJECT
10/2/14 11:40:43 AM	GUI	chdisk DCS3700_360080e50002eb69600002c28542950dc,DCS3...	SUCCESS
10/2/14 10:08:58 AM	GUI	chacl '/ibm/gpfs0/redbookexport1' +usr:root:alw:rwmxdDa...	SUCCESS
10/2/14 9:47:38 AM	GUI	chexport Finance --nfsadd '*(rw)' --cluster 124027792390449...	SUCCESS
10/2/14 2:39:57 AM	CLI	backupmanagementnode -c 12402779239044960749	SUCCESS
10/2/14 2:22:10 AM	CLI	runtask MGMTNODECONFREPL -c 'furby.storage.tucson.ibm...	SUCCESS
10/1/14 11:02:05 PM	CLI	cfgperfcenter --start -c 12402779239044960749	SUCCESS

Figure 7-24 Audit log view in Access section

3. File System ACL: In this section, you can set the ACL for directories and files. You can also configure ACL templates. See Figure 7-25.


furby.storage.tucson.ibm.com > Access > File System ACL
IBM SONAS
admin (Security Administrator, Data Access)
File System ACL
Files and Directories
ACL Templates
Back
Forward
Up
Type
Name
Access
Read
Write
Execute
Apply
Edit
Load ACL Template
Help Topic: Access Control List
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Files and Directories

Set the owner, owning group and access control list (ACL) of an existing file or directory.

Directory: /ibm

☑ Only directories

Owner: *

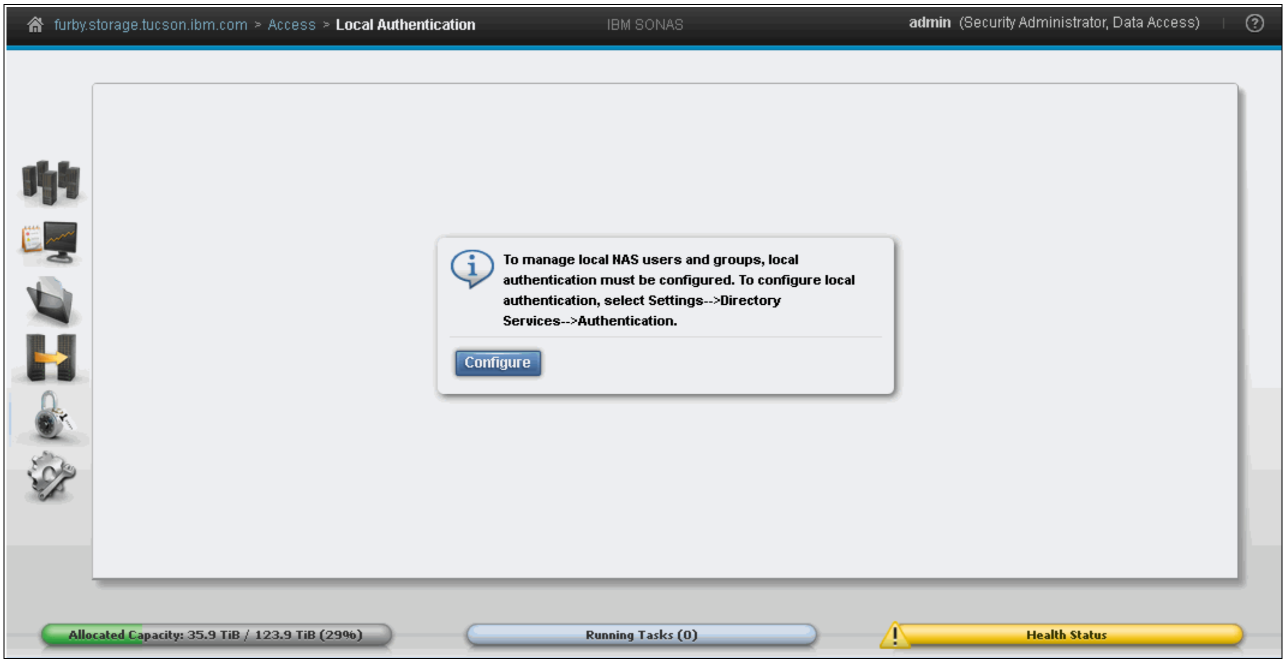
Owning group: *

Access Control List:

No items found.

Figure 7-25 File System ACL view in the Access section

4. Local Authentication: This section allows you to add and configure NAS user groups in a local authentication environment. You must use only one authentication method for SONAS. This requirement means that if you configure local authentication, you cannot use external AD or LDAP users. The preferred practice for SONAS authentication is to configure external authentication, such as AD, NIS, or LDAP. Figure 7-26 shows that the local authentication is not configured currently.


furby.storage.tucson.ibm.com > Access > Local Authentication
IBM SONAS
admin (Security Administrator, Data Access)
Configure
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

To manage local NAS users and groups, local authentication must be configured. To configure local authentication, select Settings-->Directory Services-->Authentication.

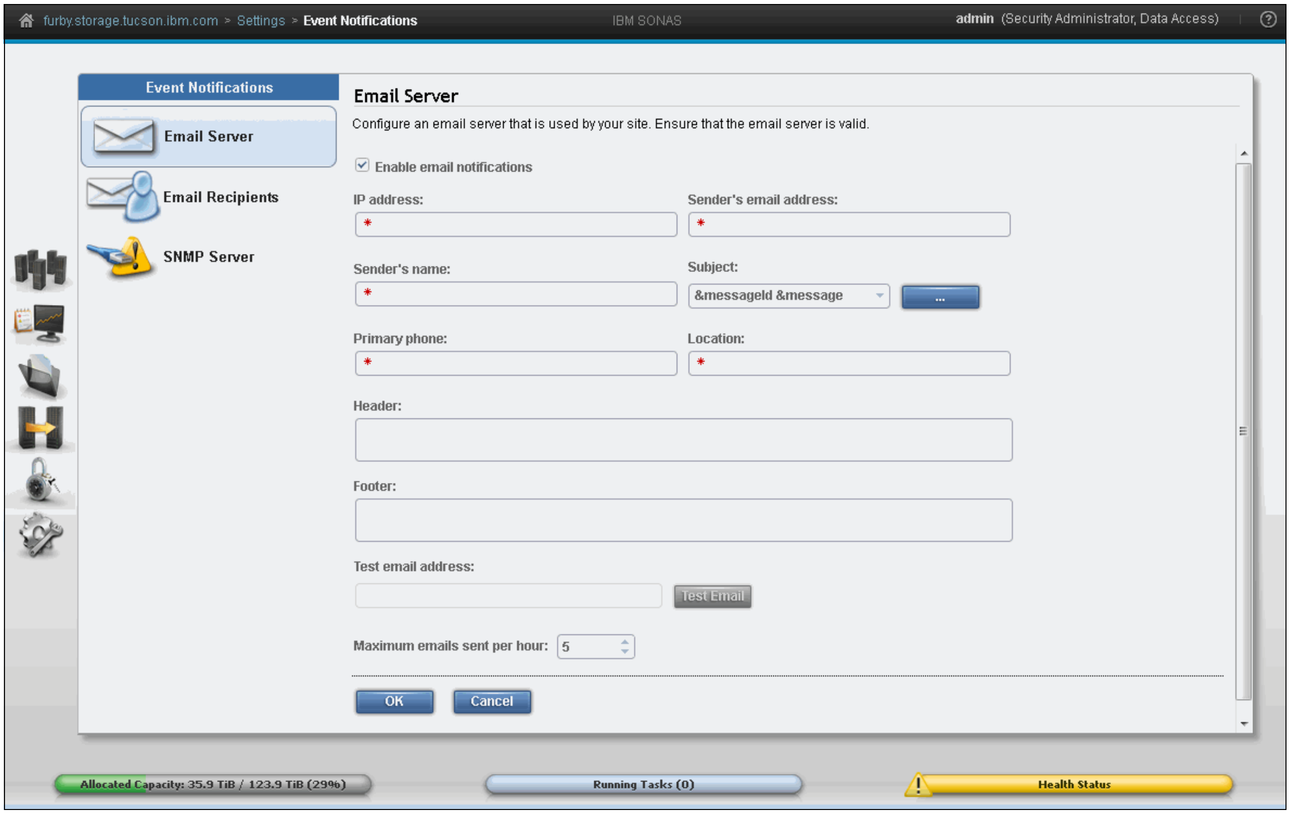
Figure 7-26 Local Authentication view in the Access section

Note: Local Authentication has several limitations. SONAS local authentication provides 1000 users and 100 user groups. You can use it in both the GUI and the CLI. The **`mknasuser`** and **`mknasgroup`** commands help create NAS users groups. These commands and the GUI function are supported only when the local authentication is configured.

Settings

In this category, you can view or modify primary SONAS parameters. The window is divided into the following sections:

1. Event Notifications: Here you can view and configure notifications in a software and hardware component failure. You can configure email servers, email recipients, and an SNMP server. See Figure 7-27.



furby.storage.tucson.ibm.com > Settings > Event Notifications

IBM SONAS

admin (Security Administrator, Data Access)

Email Server

Configure an email server that is used by your site. Ensure that the email server is valid.

☑ Enable email notifications

IP address: *

Sender's email address: *

Sender's name: *

Subject: &messageId &message

Primary phone: *

Location: *

Header:

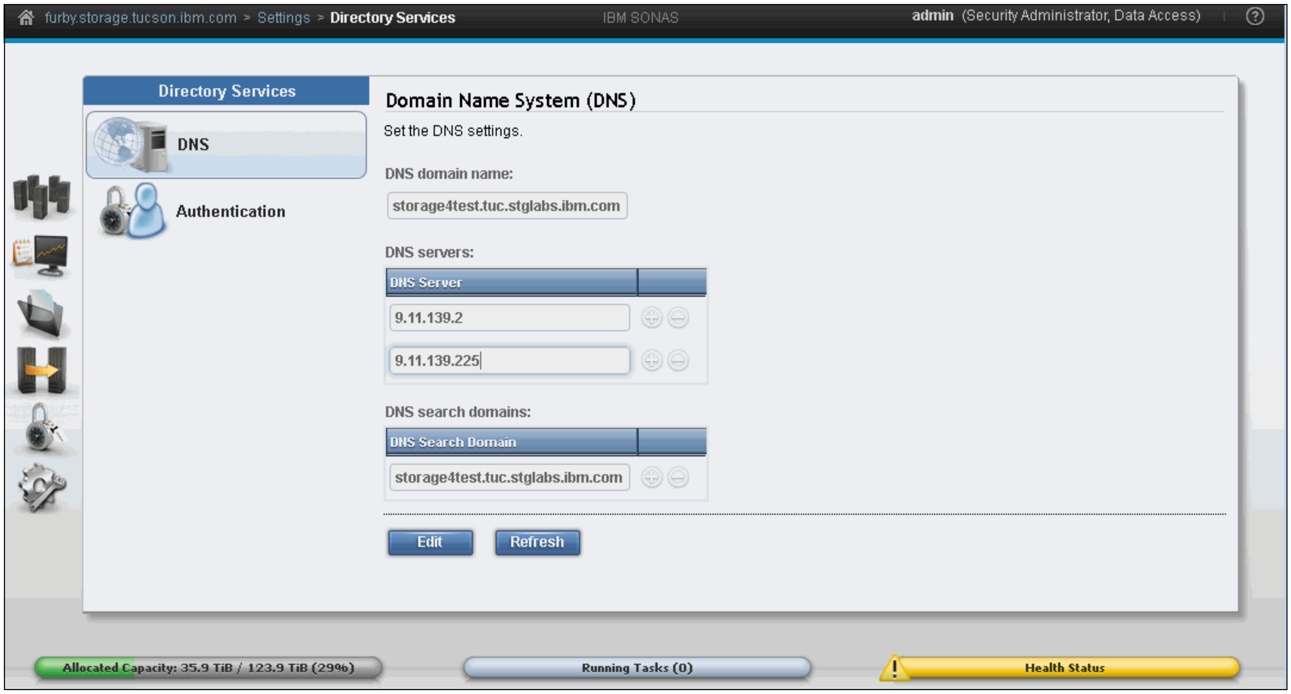
Footer:

Test email address:

Maximum emails sent per hour: 5

Figure 7-27 Event Notification submenu

2. Directory Services: Here you can configure Domain Name Systems and authentication methods, such as AD, LDAP, Samba primary domain controller (PDC), NIS, or AD with NIS. See Figure 7-28.


furby.storage.tucson.ibm.com > Settings > Directory Services
IBM SONAS
admin (Security Administrator, Data Access)
Directory Services
DNS
Authentication
Edit
Refresh
Running Tasks (0)
Health Status

Domain Name System (DNS)

Set the DNS settings.

DNS domain name: storage4test.tuc.stglabs.ibm.com

DNS servers:

DNS Server
9.11.139.2
9.11.139.225

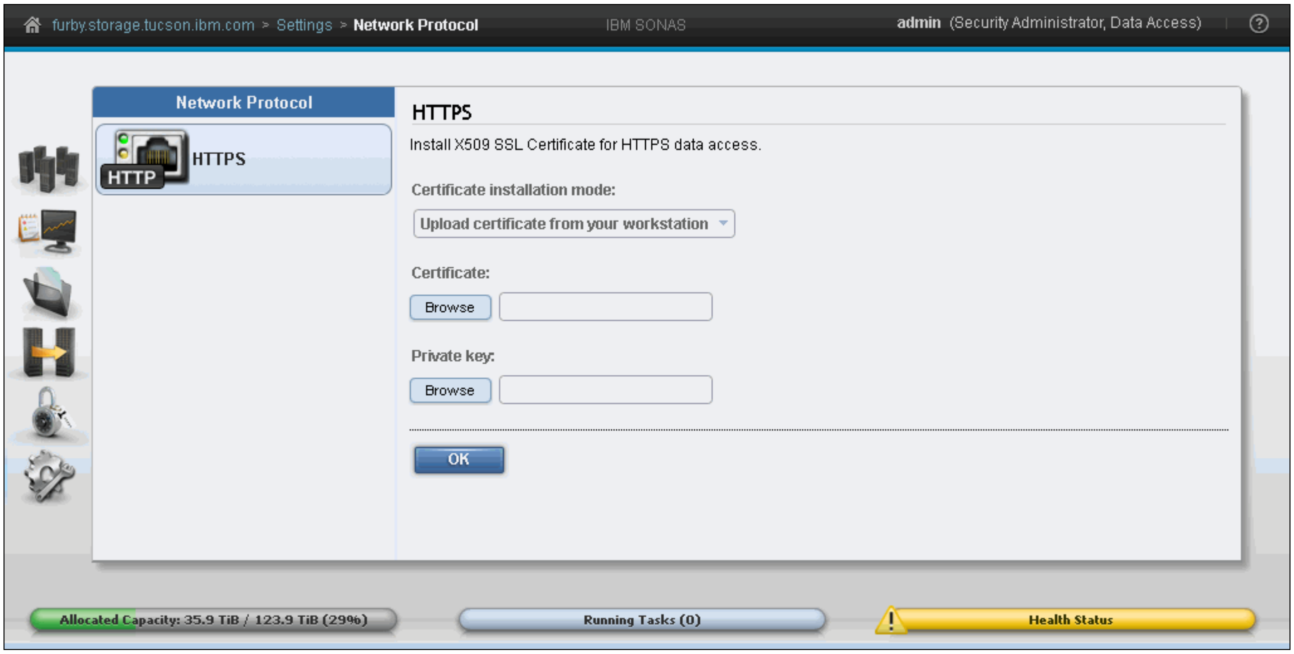
DNS search domains:

DNS Search Domain
storage4test.tuc.stglabs.ibm.com

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)

Figure 7-28 Example of Active Directory configuration in the Setting menu

3. Network Protocols: Here you can configure HTTPS, upload certificates, and private keys. See Figure 7-29.


furby.storage.tucson.ibm.com > Settings > Network Protocol
IBM SONAS
admin (Security Administrator, Data Access)
Network Protocol
HTTP
HTTPS
Browse
Browse
OK
Running Tasks (0)
Health Status

HTTPS

Install X509 SSL Certificate for HTTPS data access.

Certificate installation mode: Upload certificate from your workstation

Certificate:

Private key:

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)

Figure 7-29 View of Network Protocols window under the Setting menu

4. Network: In this section, you can view or configure internal and external network parameters :

   – Network Groups: This group is made up from a minimum of one node, but because of redundancy concerns, a group must consist of at least two nodes.

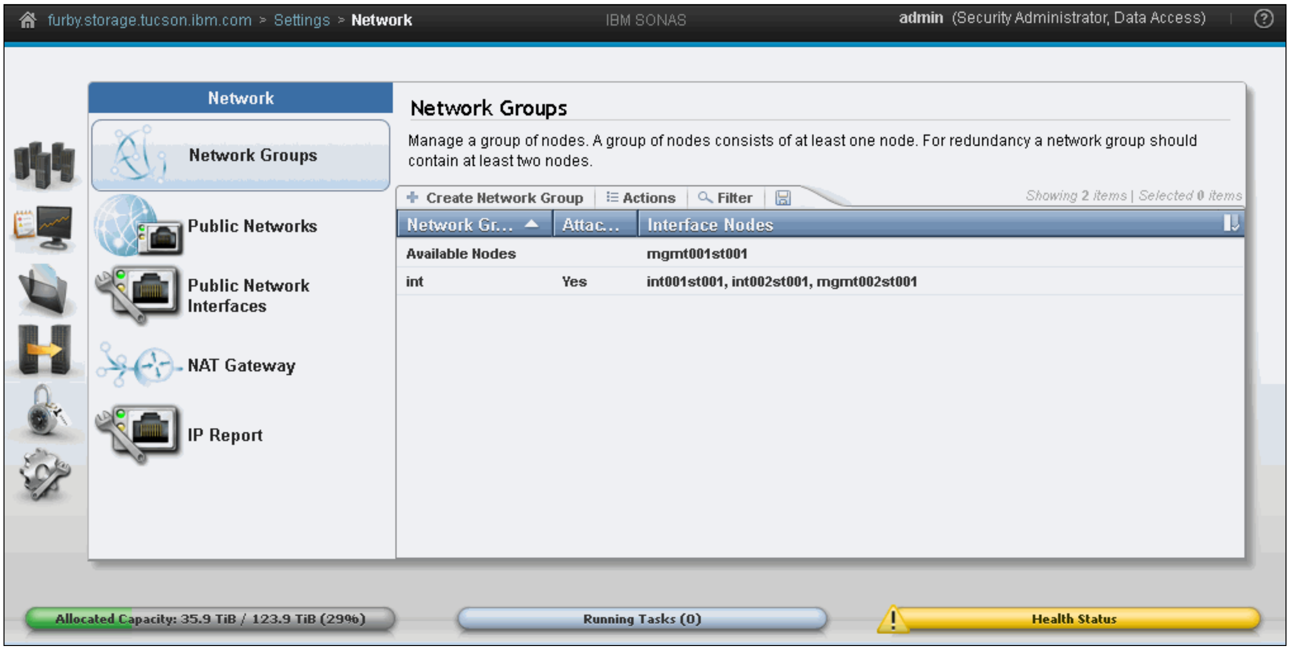
   – Public Networks: Configure the public network that external clients can use to reach the system. It includes all network interfaces that are used to provide NAS services to external clients in the network.

   – Public Network Interfaces: Displays the public data network interfaces, bonding mode, speed, and their status.

   – NAT Gateway: Specify a public address that can be used as a virtual gateway to access the Interface nodes on the private network.

   – IP report: Displays an overview of IP addresses that are used across the system.

See Figure 7-30.


furby.storage.tucson.ibm.com > Settings > Network
IBM SONAS
admin (Security Administrator, Data Access)
Network
Network Groups
Public Networks
Public Network Interfaces
NAT Gateway
IP Report
Create Network Group
Actions
Filter
Showing 2 items | Selected 0 items
Running Tasks (0)
Health Status

Network Groups

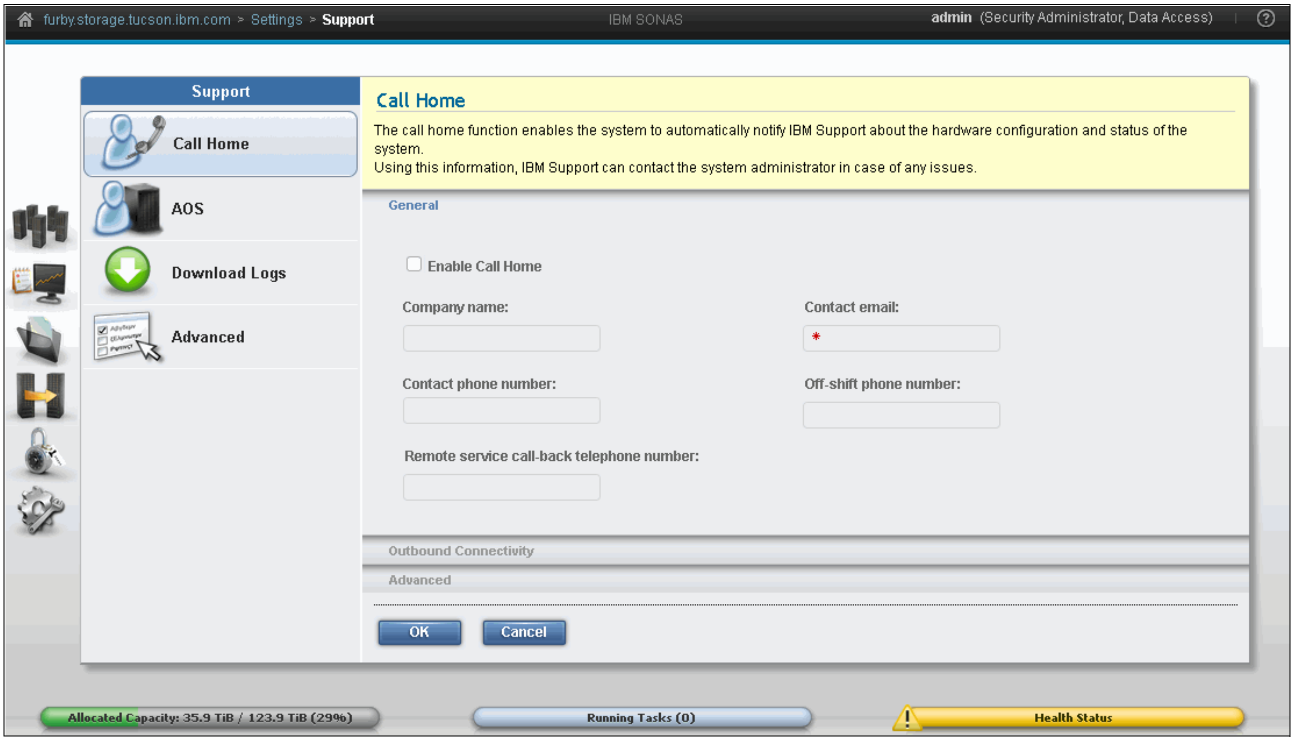
Manage a group of nodes. A group of nodes consists of at least one node. For redundancy a network group should contain at least two nodes.

Network Gr...	Attac...	Interface Nodes
Available Nodes		mgmt001st001
int	Yes	int001st001, int002st001, mgmt002st001

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)

Figure 7-30 View of the Network Group section in the Settings menu

5. Support: In the support section, you can configure SONAS functions in case there is a problem with the system. IBM Support personnel can do repair actions quickly and more effectively. See Figure 7-31 on page 513.


furby.storage.tucson.ibm.com > Settings > Support
IBM SONAS
admin (Security Administrator, Data Access)
Support
Call Home
AOS
Download Logs
Advanced
General
Outbound Connectivity
Advanced
OK
Cancel
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Call Home

The call home function enables the system to automatically notify IBM Support about the hardware configuration and status of the system.
Using this information, IBM Support can contact the system administrator in case of any issues.

☐ Enable Call Home

Company name:

Contact email: *

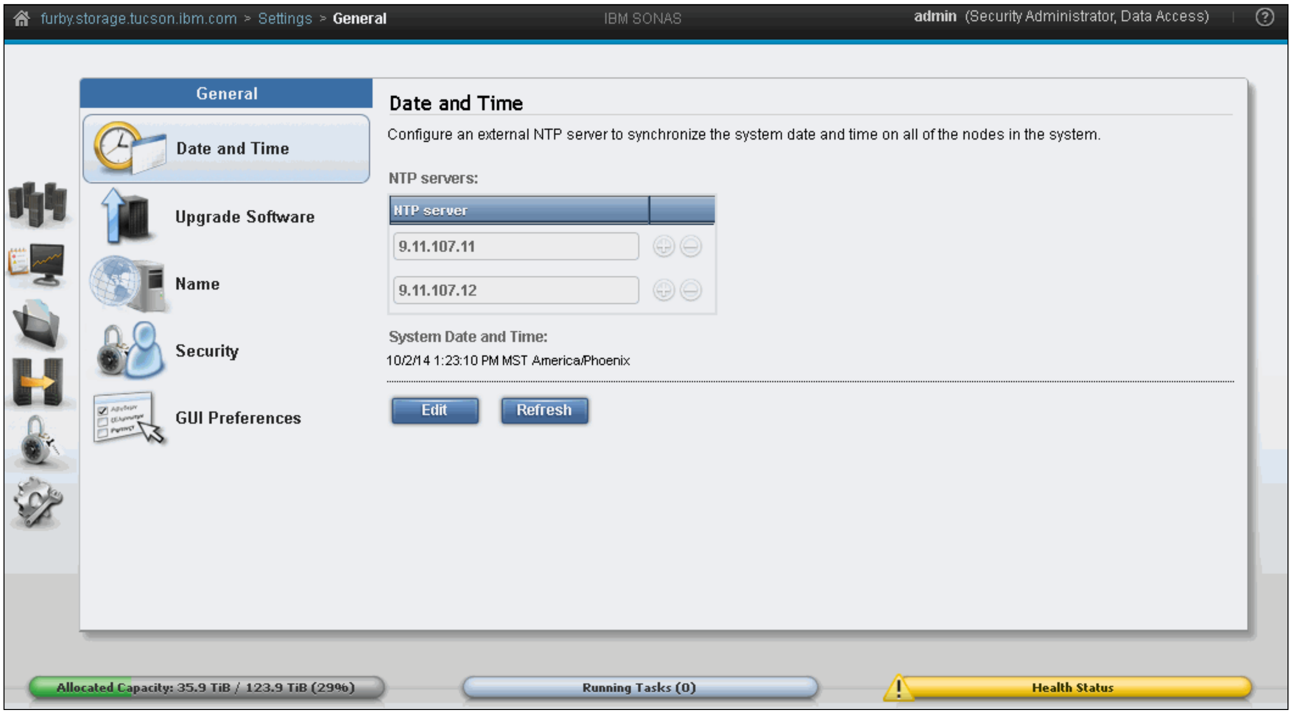
Contact phone number:

Off-shift phone number:

Remote service call-back telephone number:

Figure 7-31 View of a Support submenu under the Settings menu

6. General: Under the General settings, you can set the primary and secondary NTP server, which provides clock synchronization. See Figure 7-32.


furby.storage.tucson.ibm.com > Settings > General
IBM SONAS
admin (Security Administrator, Data Access)
General
Date and Time
Upgrade Software
Name
Security
GUI Preferences
Edit
Refresh
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Date and Time

Configure an external NTP server to synchronize the system date and time on all of the nodes in the system.

NTP servers:

NTP server
9.11.107.11
9.11.107.12

System Date and Time:
10/2/14 1:23:10 PM MST America/Phoenix

Figure 7-32 Date and Time setting under the General menu

7.1.4 Accessing the CLI

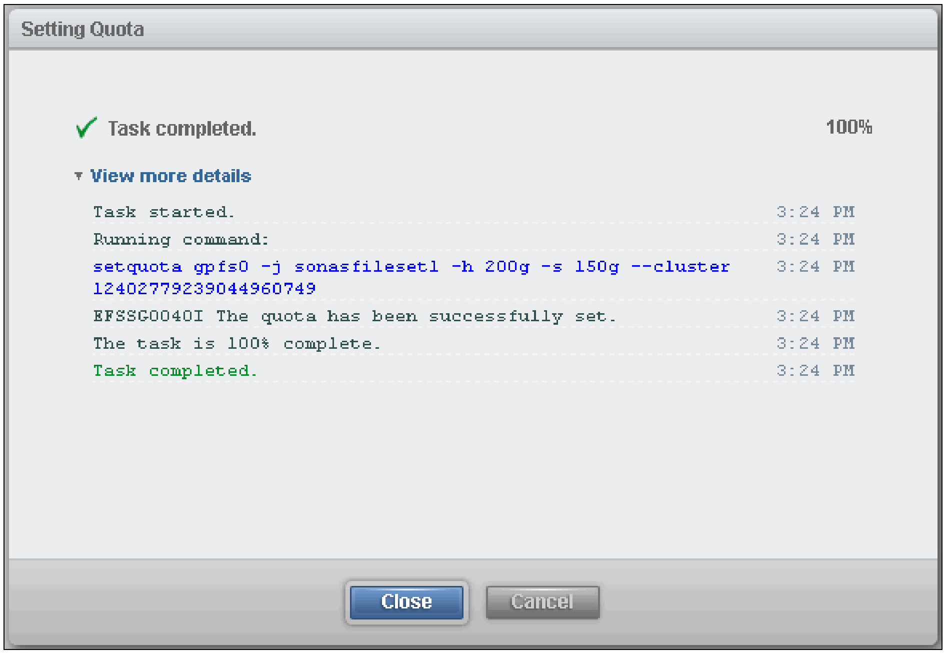
To access the CLI, run **ssh** to open the Management node and log in with your user name and password. You are taken to a restricted shell that allows you to run only CLI commands.

For example, consider the example where the Management node host name is `sonas1.ibm.com`. You log in to the CLI by running either of the following commands:

- **#ssh user_name@sonas1.ibm.com**
- **#ssh user_name@sonas_management_IP_address**

You are asked to enter a password for the user name. Then, you are taken to a CLI prompt. In the CLI, you can run **help** to see a full list of commands that can be run on the system.

The GUI shows the commands that are run. In Figure 7-33, a new file system that is named `gpfs1` is created. Figure 7-33 shows all the commands that the system ran in the background. This window appears every time that you do an action in the GUI. It reflects the corresponding CLI commands.



```
Task started.                                                  3:24 PM
Running command:                                               3:24 PM
setquota gpfs0 -j sonasfileset1 -h 200g -s 150g --cluster      3:24 PM
12402779239044960749
EFSSG0040I The quota has been successfully set.                3:24 PM
The task is 100% complete.                                     3:24 PM
Task completed.                                                3:24 PM
```

Figure 7-33 CLI command that is shown after you apply changes in the GUI

Example 7-1 contains a list of commands that are available for a user in the CLI.

Example 7-1 CLI command list

```
[furby.storage.tucson.ibm.com]$ help
command                   description
addcluster                Adds an existing cluster to the Management node.
addnode                   Adds an Interface node to the cluster.
applysoftware             Applies a new software package to the system.
attachnw                  Attaches a specified network to a specified network group.
backupmanagementnode      Backs up a Management node.
catxmlspec                Return an XML representation of the command set
cfgad                     Configures the Active Directory server-based authentication for the cluster.
cfgaos                    Sets up and queries AOS (Assist On-site) remote access service.
cfgav                     Modifies the antivirus configuration options.
cfgbackupfs               Defines the Tivoli Storage Manager server on which a file system should be backed up and by which nodes.
cfgcallhome               Configures or sets the Call Home options on the current system.
cfgcluster                Starts the postinstallation steps that are required before the cluster can be configured.
cfghsmfs                  Configures the file systems to be enabled for hierarchical storage management (HSM).
cfghsmnodes               Configures nodes to be enabled for hierarchical storage management (HSM).
cfgidmap                  Configure the cluster with ID mapping services
```

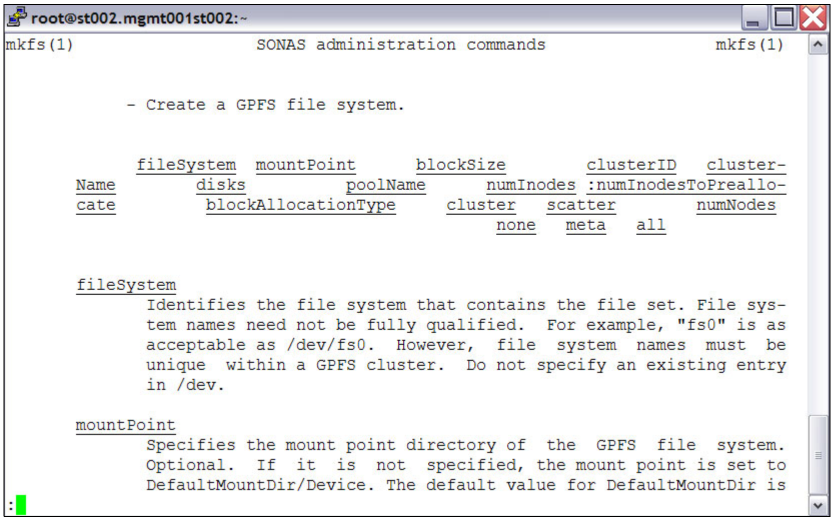
```
cfgldap                 Configures an LDAP server or an LDAP with Kerberos server on all the nodes present in the cluster with the input
values.
cfglocalauth            Configures the Local Auth.
cfgndmp                 Configures the NDMP server on a group of nodes of a cluster.
cfgndmpprefetch         Configures the NDMP prefetch settings on a group of nodes of a cluster.
cfgnis                  Configures the Network Information Service (NIS).
cfgnt4                  Configures the Microsoft Windows NT 4.0 (NT4) server on all the nodes present in the cluster by using the input
values.
cfgperfcenter           Configures Performance Center services on nodes.
cfgrepl                 Configures asynchronous replication for nodes.
cfgreplfs               Defines a relationship between a local file system and a target directory on a remote cluster for replication.
cfgsfu                  Configures the cluster with Services For UNIX (SFU) user mapping services.
cfgtsmnode              Configures the Tivoli Storage Manager node by defining the node name, node password, and by adding a Tivoli
Storage Manager server stanza.
chacl                   Change the ACL for a specified file or directory.
chbackupfs              Modifies the list of backup nodes for a file system.
chbanner                Modifies the banner for all the nodes of the system.
chcfg                   Changes the configuration data for a cluster.
chcluster               Change the cluster attributes
chcurrnode              Changes a current node.
chdisk                  Changes a disk.
chemail                 Changes email notification configuration.
chemailserver           Changes email server definition.
chemailuser             Changes the settings of an email user.
chexport                Modifies the protocols and their settings of an existing share.
chfs                    Changes the properties of the file system.
chfset                  Changes a file set.
chkauth                 Checks authentication settings of a cluster.
chkdept                 Checks departments status.
chkfs                   Checks and repairs a file system.
chkpolicy               Checks a policy on the nodes of a specified cluster.
chkquota                Checks file system user, group, and file set quotas.
chmgr                   Change the file system / cluster manager node.
chnasuser               Modify a local authentication nas user.
chnfsserver             Changes the NFS server stack.
chnode                  Changes the quorum nodes configuration.
chnw                    Modifies a network configuration for a subnet.
chnwgroup               Adds or removes nodes to or from a specified network group.
chnwmgt                 Changes basic network configuration settings.
chnwsdg                 Change or delete the explicit declaration for a system default gateway.
chowner                 Change owner for a specified file or directory.
chpasswordpolicy        Changes password policy applicable to the whole cluster.
chpolicy                Changes an existing policy, adds or removes rule declarations.
chpsnaptask             Changes the psnap tasks.
chrootpwd               Changes the root password on cluster nodes at the same time.
chservice               Changes the configuration of a protocol service.
chsessionpolicy         Changes session policy applicable to the whole cluster.
chsettings              Change settings applicable to the whole cluster.
chsnapassoc             Changes the snapshot rule that is associated with a file set or file system.
chsnaprule              Changes a snapshot rule.
chsnmpserver            Changes an existing SNMP server definition.
chsyslogserver          Changes an external syslog server configuration.
chuser                  Modifies an administrative user ID for the system.
chusergrp               Changes the attributes of a user group.
chwcache                Modifies a caching file set.
chwcachesource          Modifies an existing wan caching share. Administrator can add or remove clients
cleanupauth             Clean up the authentication and ID map configuration
cleanuprepl             Deletes replication logs.
ctlavbulk               Runs bulk scan.
ctlwcache               Performs operations on cached file sets.
detachnw                Detaches a network from an interface of a network group.
dumptrace               Capture in memory trace of the file system and restart in memory tracing.
enablelicense           Enables the license agreement flag. After the license is enabled, user access to the GUI windows is restricted
until the user accepts the license.
help                    Displays the help screen.
initnode                Stops or restarts the node.
linkfset                Links a file set.
locatenode              Enables, disables, or flashes the system locator LEDs on any SONAS node.
lookupname              Translates a user ID or group ID into a NAS user or group name.
lsacl                   Displays the access control list (ACL) of a file or directory.
lsaudit                 List the audit log entries.
lsauth                  Lists authentication settings of a cluster.
lsav                    Displays the current antivirus configuration.
lsbackupfs              Lists file system to the Tivoli Storage Manager server and backup node associations.
lsbanner                Lists the banner for SONAS systems.
lscallhome              Lists the actual Call Home configuration.
lscallhomelog           List the call home event log entries.
lscfg                   Displays the current configuration data for a cluster.
lsclone                 Lists the parents of the clone file, both the source and the parent file.
lscluster               Lists all the clusters of the system.
lscurrentuser           Lists details of the user who is logged in.
lsdisk                  Lists all the disks.
lsemailserver           Lists the email-server definitions.
```

```
lsemailuser             Lists all the email users.
lsexport                Lists exports.
lsfs                    Lists all the file systems on a given device in a cluster.
lsfset                  Lists file sets for a given device in a cluster.
lshealth                Lists the overall status of the system.
lshsm                   Lists all the hierarchical storage management (HSM)-enabled file systems in the cluster.
lshsmlog                Lists hierarchical storage management (HSM) log messages.
lshsmstatus             Lists the status of the hierarchical storage management (HSM)-enabled nodes in the cluster.
lsidmap                 Display ID mappings that are configured on cluster
lsjobstatus             Shows the status of currently running or already finished jobs.
lslog                   Lists the event log entries.
lsmgr                   Displays the node that is the file system manager for specified file systems or the node that is the cluster
manager.
lsmount                 Lists the mounted file systems that belong to the cluster and file system.
lsnasgroup              List all the NAS groups.
lsnasuser               List all the NAS users.
lsndmp                  Lists NDMP data server settings of a subgroup of existing Interface nodes of a preconfigured cluster.
lsndmplog               Shows NDMP logs of existing Interface nodes of a preconfigured cluster.
lsndmpprefetch          Shows the status of prefetch on the Interface nodes of the NDMP node group.
lsndmpsession           Lists NDMP data sessions on existing Interface nodes of a preconfigured cluster.
lsnfsserver             List the active NFS server stack of a cluster.
lsnode                  Lists all Nodes.
lsnw                    Lists public network configurations for the current cluster.
lsnwdns                 Lists DNS configurations for the current cluster.
lsnwgroup               Lists network group configurations for the current cluster.
lsnwinterface           Lists the network interfaces.
lsnwmgt                 Lists the service IP address configuration of the Management nodes.
lsnwnatgateway          Lists NAT gateway configurations for the current cluster.
lsnwntp                 Lists NTP configurations for the current cluster.
lsnwsdg                 List the effective system default gateways and the explicit declarations (overwrite setting)
lsowner                 Show owner for a specified file or directory.
lspasswordpolicy        Lists the password policy applicable to the whole cluster.
lsperfdata              Retrieves historical performance data as CSV output.
lspolicy                Lists the policies and rules that belong to the cluster and file system.
lspool                  Lists all pools.
lsprepop                Shows the status of pre-population of files on wcache.
lspsnap                 Lists peer snapshots.
lspsnaptask             Lists all psnap tasks.
lsquota                 Lists all quotas.
lsreconciletask         Lists the tasks that are scheduled for reconcile.
lsrepl                  Lists the result of asynchronous replications.
lsreplcfg               Lists configuration of asynchronous replications.
lsreplfs                Lists file system that is configured for asynchronous replication.
lsrepltarget            Lists target of asynchronous replications.
lsrepltask              Lists the tasks that are scheduled for asynchronous replication.
lsservice               Lists all the services of a cluster.
lssessionpolicy         Lists the session policy applicable to the whole cluster.
lssettings              List the settings applicable to the whole cluster.
lssnapassoc             Lists the snapshot associations.
lssnapnotify            Lists the snapshot event notification settings.
lssnapops               Displays a list of queued and running snapshot operations.
lssnaprule              Lists the snapshot rules.
lssnapshot              Lists all snapshots.
lssnmpserver            Lists all configured SNMP servers.
lssoftwareupgradestatus Lists status of system upgrade to a new level of software.
lssyslogserver          Lists existing syslog server definitions.
lstask                  Lists the scheduled tasks that belong to a Management node for the selected cluster.
lstime                  Shows time.
lstrace                 Lists active traces.
lstsmnode               Lists Tivoli Storage Manager nodes in the cluster.
lsuser                  Lists all the command-line interface (CLI) users of the Management node.
lsusergrp               Displays a list of user groups that have been created on the cluster.
lsvpd                   Displays VPD Information.
lswcache                Lists all the caching file sets for a given device in a cluster.
lswcachesource          Lists the WAN-caching sources on the home cluster.
lswcachestate           Lists all the caching file sets for a device in a cluster.
mkclone                 Creates a clone file of a specified archive.
mkdisk                  Creates storage system NAS volumes. Command is supported only on SONAS Gateway configuration and Storwize V7000
Unified.
mkemailserver           Creates email server definition.
mkemailuser             Creates an email user.
mkexport                Creates a new share by using one or more protocols.
mkfs                    Creates a file system.
mkfset                  Creates a file set.
mknasgroup              Create a local authentication NAS Group.
mknasuser               Creates a nas user for local authentication server.
mknw                    Defines a new network configuration for a subnet, and assigns multiple IP addresses and routes.
mknwbond                Creates a bond from the specified subordinate group.
mknwgroup               Creates a group of nodes to which a network configuration can be attached. See also the commands mknw and
attachnw.
mknwnatgateway          Creates a clustered trivial database (CTDB) network address translation (NAT) gateway.
mkpolicy                Makes entries of the policy and rules in the database.
mkpolicytask            Schedules a task for data placement and data movement with the aid of a GPFS policy.
```

```
mkpsnap                 Creates the peer snapshot for a file set.
mkpsnaptask             Creates the psnap tasks to be run periodically.
mkreconciletask         Schedules a reconcile task.
mkrepltarget            Declares source cluster and target path for replication.
mkrepltask              Schedules a task for asynchronous replication.
mksnapassoc             Associates a snapshot rule with a file set or file system.
mksnaprule              Creates a snapshot rule.
mksnapshot              Creates a file system snapshot.
mksnmpserver            Creates a new SNMP server definition.
mksyslogserver          Creates a new external syslog server configuration.
mktask                  Schedules a GUI or a cron task on the selected cluster that belongs to the management system.
mkuser                  Creates an administrative user ID for the Management node.
mkusergrp               Creates a new user group.
mkwcache                Creates a WAN cache on the client cluster.
mkwcachenode            Configures cache features on the cluster. The administrator can later use this feature by creating a cached file
set.
mkwcachesource          Creates a new WAN-caching share.
mountfs                 Mounts a file system.
querybackup             Queries backup summary for the specified file pattern.
restripefs              Rebalances or restores the replication of files in a file system.
resumenode              Resumes a list of nodes.
rmbackupfs              Removes file system to Tivoli Storage Manager server association.
rmcluster               Removes a cluster from the Management node.
rmdisk                  Removes storage system NAS volumes.
rmemailserver           Removes email server definition.
rmemailuser             Removes email user definition.
rmexport                Removes the given share.
rmfs                    Removes an existing file system from the cluster.
rmfset                  Removes a file set.
rmidmapcacheentry       Removes the ID map cache entry.
rmjobstatus             Removes old logs and corresponding information.
rmlock                  Releases orphan locks that are acquired by crashed clients.
rmlog                   Removes all the log entries that are stored in the database.
rmnasgroup              Removes a local authentication NAS group.
rmnasuser               Removes a NAS user.
rmndmpcfg               Removes the NDMP data server configuration from sub group of existing Interface nodes of a preconfigured cluster.
rmnw                    Deletes a network configuration.
rmnwbond                Deletes a regular bond interface.
rmnwdns                 Removes the name servers.
rmnwgroup               Removes a network group.
rmnwnatgateway          Removes the configuration of a clustered trivial database (CTDB) network address translation (NAT) gateway.
rmnwntp                 Removes one or more external NTP servers.
rmpolicy                Removes a policy and all the rules that are associated with it.
rmpolicytask            Removes a scheduled task for data placement and data movement.
rmpsnap                 Removes a peer snapshot.
rmpsnaptask             Removes a psnap task.
rmreconciletask         Removes a scheduled reconcile task from a file system.
rmreplfs                Removes the replication file system association.
rmrepltarget            Removes a replication target that is created by mkrepltarget.
rmrepltask              Removes a scheduled task from asynchronous replication.
rmsnapassoc             Removes a snapshot rule and file set or file-system association.
rmsnapnotify            Disables notification for an event type.
rmsnaprule              Removes the snapshot rule.
rmsnapshot              Removes the snapshot.
rmsnmpserver            Removes an SNMP server definition.
rmsyslogserver          Removes an external syslog server.
rmtask                  Removes a scheduled task that belongs to the Management node on the selected cluster.
rmtsmnode               Removes Tivoli Storage Manager server stanza for the node.
rmuser                  Removes an administrative user ID for SoFS.
rmusergrp               Deletes a given user group that was created on the cluster.
rmwcache                Removes a caching file set.
rmwcachenode            Removes configuration of cache node/s of cache cluster.
rmwcachesource          Removes the specified WAN-caching source on the home cluster.
rpldisk                 Replaces a current disk with a specified disk.
runpolicy               Applies the policy on the nodes of a specified cluster for a specified device.
runprepop               Runs the prepop command to pre-populate the files on wcache.
runreplrecover          Recovers inactive asynchronous replications processes.
runtask                 Runs a scheduled task that belongs to the Management node directly on the selected cluster.
setnwdns                Sets the name servers.
setnwntp                Sets one or more external Network Time Protocol (NTP) servers on the Management node.
setpolicy               Sets placement policy rules for a GPFS file system.
setquota                Sets the quota settings.
setsnapnotify           Sets snapshot notification conditions.
settime                 Sets the time and date.
settz                   Sets the time zone.
showerrors              Displays the error log of a given jobcategory or jobID.
showlog                 Displays the log of a given jobcategory.
showreplresults         Displays replication errors and logs.
srvdump                 Manage dump files.
startbackup             Starts the backup process.
startemail              Enables email notifications.
startreconcile          Starts reconcile process.
startrepl               Starts asynchronous replication.
```

```
startrestore            Starts the restore process.
starttrace              Starts tracing of network traffic.
stopbackup              Stops a running Tivoli Storage Manager backup session.
stopcluster             Performs controlled shutdown of a cluster or node.
stopemail               Disables email notifications.
stopndmpsession         Stops NDMP Data sessions on existing Interface nodes of a preconfigured cluster.
stoppolicy              Stops running policy jobs depending on the parameters specified.
stopreconcile           Stops a reconcile session.
stoprepl                Stops asynchronous replication.
stoprestore             Stops a restore session.
stoptrace               Stops a previously defined trace.
suspendnode             Suspends a list of nodes.
testemail               Sends a test email to all or a specified user.
unlinkfset              Delinks a file set.
unmountfs               Unmounts a file system.
updatefs                Updates a file system to be used with a new version.
```

Figure 7-34 shows the man page of a SONAS administration command (in this example, **mkfs**).



```
mkfs(1)                    SONAS administration commands                  mkfs(1)


            - Create a GPFS file system.


             fileSystem  mountPoint      blockSize        clusterID   cluster-
       Name        disks           poolName       numInodes :numInodesToPreallo-
       cate         blockAllocationType      cluster   scatter          numNodes
                                                 none   meta   all


       fileSystem
              Identifies the file system that contains the file set. File sys-
              tem names need not be fully qualified.  For example, "fs0" is as
              acceptable as /dev/fs0.  However,  file  system  names  must  be
              unique  within a GPFS cluster.  Do not specify an existing entry
              in /dev.

       mountPoint
              Specifies the mount point directory of  the  GPFS  file  system.
              Optional.  If  it  is  not  specified, the mount point is set to
              DefaultMountDir/Device. The default value for DefaultMountDir is
:
```

Figure 7-34 Man page for mkfs

Similarly, you can run **help** for each of the commands that are available in the CLI and run **man page** for each one. Another way to see parameters that are needed for a command to run successfully is by attaching the **<command_name> --help** option. See Example 7-2 for **help** use.

Example 7-2 Help or usage for the mkfs command

```
[furby.storage.tucson.ibm.com]$ mkfs --help
usage: mkfs filesystem  [mountpoint] [-b <blocksize>] {-F <disks> | --pool <poolName> | --discoverdisks | --createdisks
<size,mdiskgrp1[,mdiskgrp2][,N=numDisks]>} [-i <numInodes:numInodesToPreallocate>] [-j <blockAllocationType>] [-N <numNodes>] [--dmapi |
--nodmapi] [--noverify] [-R <policy>] [--logplacement <logPlacement>] [-q <quota>] [-c < clusterID | clusterName >]
Creates a file system.
Parameter           Description
filesystem          Identifies the file system that contains the file set.
mountpoint          Specifies the mount point of the new file system.
-b                  Specifies the size of the data blocks. The size must be 256 KB (default size), 1 MB, or 4 MB.
-F                  Specifies the disks on which the file system is to be created.
    --pool          Specifies the file system pool on which the file system is to be created.
    --discoverdisks (IFS only) Detects and uses all free GPFS NSDs that are tagged for the specified file system but not yet included.
    --createdisks   (IFS only) Creates disks implicitly and adds them to the file system.
-i                  Specifies the maximum number of files for this file system.
-j                  Specifies the block allocation map type (cluster/scatter).
-N                  Specifies the estimated number of nodes that will mount this file system.
    --dmapi         Enables DMAPI support. For example, HSM and Antivirus.
    --nodmapi       Disables DMAPI support. For example, HSM and Antivirus.
    --noverify      Skips the verification of the disk descriptor, so that disks that contain on old descriptor can be reused.
-R                  Specifies the replication policy (none/meta/all) for this file system.
    --logplacement  Sets whether the logs will be stored striped across the disks or not.
```

```
-q, --quota         Specifies the quota option (yes/no/perfileset) for this file system.
-c, --cluster       The cluster scope for this command
```

7.2 SONAS administrator tasks list

In the SONAS V1.3 code release, there are small differences between GUI and CLI usage capabilities. As already mentioned, in the GUI, every action that is done shows you the corresponding CLI command that is being run.

In the GUI, performance data is shown in a more readable form. In the CLI, you can shut down nodes or disable certain services. For daily routine checks and configuration changes, you can use the interface that you prefer.

Here are some typical daily administrator tasks:

- Checking the system health:
  - The export state
  - The file system state
  - The inode state
  - The ctdb/gpfs state
  - The basic event log checks
- Checking the system performance
- Checking the backup process HSM/TSM if used
- Checking the replication state (if it is used)
- Checking for any hardware errors or failures
- Checking the snapshots
- Quota management
- Checking the scheduled tasks

7.3 Cluster management

Cluster-related commands are the commands that are used to view or modify the cluster configuration. The commands include configuration of Management nodes, Interface nodes, Storage nodes, or the cluster as a whole. The next section describes some of the common cluster tasks in detail.

7.3.1 Adding a cluster to the Management node

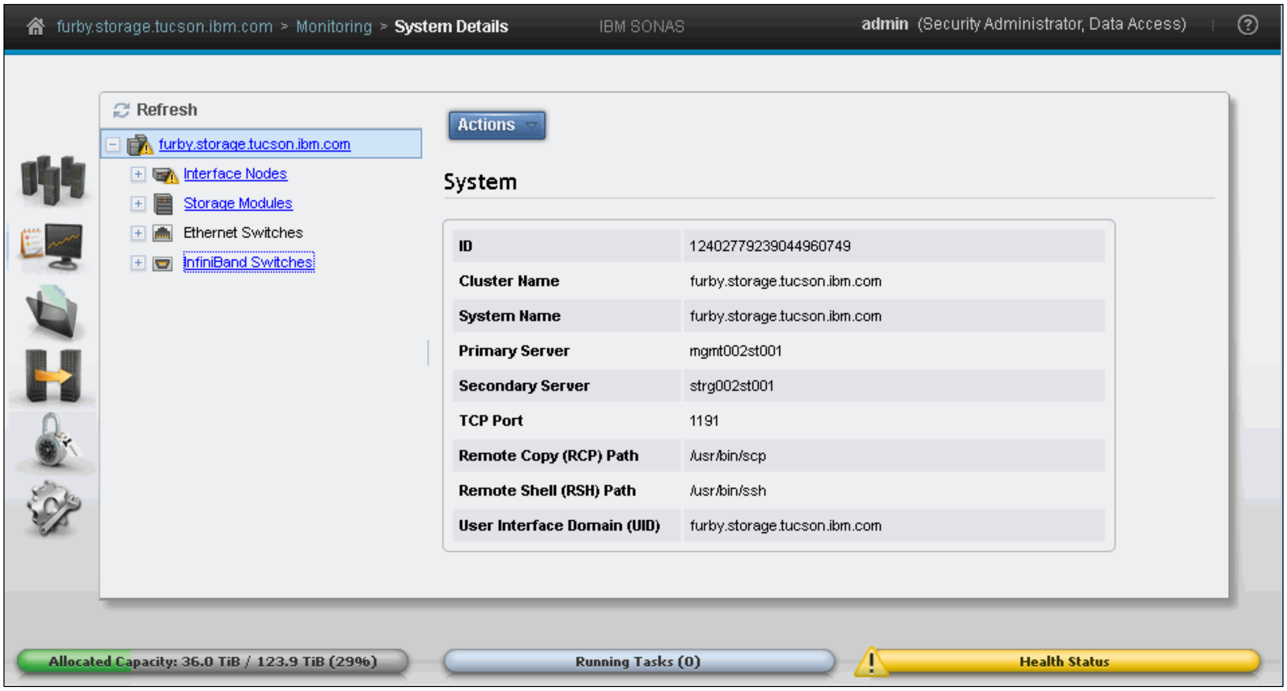
You can add the cluster to the CLI by **addcluster**. Example 7-3 shows the command and its output. You add the cluster only one time after the first-time installation. For more information, see “Post first-time installation procedures” on page 159.

Example 7-3 Usage and command output for the addcluster command

```
[furby.storage.tucson.ibm.com]$ addcluster --help
usage: addcluster  [-h <host>] [-p <pwd>]
Adds an existing cluster to the Management node.
Parameter      Description
-h, --host     Specifies the cluster host.
-p, --password Specifies the root password to access the cluster hosts.
```

7.3.2 Viewing the cluster status

Using the GUI: You can find cluster details by clicking **Monitoring** → **System details**. See Figure 7-35.


furby.storage.tucson.ibm.com > Monitoring > System Details
IBM SONAS
admin (Security Administrator, Data Access)
Refresh
furby.storage.tucson.ibm.com
Interface Nodes
Storage Modules
Ethernet Switches
InfiniBand Switches
Actions
System
Allocated Capacity: 36.0 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

ID	12402779239044960749
Cluster Name	furby.storage.tucson.ibm.com
System Name	furby.storage.tucson.ibm.com
Primary Server	mgmt002st001
Secondary Server	strg002st001
TCP Port	1191
Remote Copy (RCP) Path	/usr/bin/scp
Remote Shell (RSH) Path	/usr/bin/ssh
User Interface Domain (UID)	furby.storage.tucson.ibm.com

Figure 7-35 Viewing cluster details in the GUI

Using the CLI: You can view the cluster details by running the `lscluster`. See Example 7-4 for the command output.

Example 7-4 Command output for the lscluster command

```
[furby.storage.tucson.ibm.com]$ lscluster
Cluster id           Name                         Primary server Secondary server Profile
12402779239044960749 furby.storage.tucson.ibm.com mgmt002st001   strg002st001     SONAS
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$
```

7.3.3 Viewing the Interface node and Storage node status

Using the GUI: You can check specific node details by clicking **Monitoring** → **System Details,** as shown in Figure 7-35, and then selecting the node that you want. Each node is also divided into subcomponents, such as Operating System, Hardware, Services, Network, and Status.

Using the CLI: You can view the status of the nodes by running `lsnode`. Example 7-5 shows the usage and command output. You can get more information by using the `-v` option or see output that is formatted with delimiters by using the `-Y` option.

Example 7-5 Usage and command output for the lsnode CLI command

```
[furby.storage.tucson.ibm.com]$ lsnode --help
usage: lsnode  [-v | -x] [-r] [-Y] [-c < clusterID | clusterName >]
Lists all Nodes.
```

```
Parameter      Description
-v, --verbose  Shows additional columns.
-x, --extended Shows extended output.
-r, --refresh  Forces a refresh of the exports data in the database by scanning cluster exports before retrieving the data from the
database.
-Y             Shows parseable output.
-c, --cluster  The cluster scope for this command
[furby.storage.tucson.ibm.com]$ lsnode
Hostname     IP           Description            Role                 Product version Connection status GPFS status CTDB status Last
updated
int001st001  172.31.132.1                        interface            1.5.1.0-14      OK active      active      9/29/14 9:56 AM
int002st001  172.31.132.2                        interface            1.5.1.0-14      OK active      active      9/29/14 9:56 AM
mgmt001st001 172.31.136.2 active Management node  management,interface 1.5.1.0-14      OK active      active      9/29/14 9:56 AM
mgmt002st001 172.31.136.3 passive Management node management,interface 1.5.1.0-14      OK active      active      9/29/14 9:56 AM
strg001st001 172.31.134.1                        storage              1.5.1.0-14      OK active                  9/29/14 9:47 AM
strg002st001 172.31.134.2                        storage              1.5.1.0-14      OK active                  9/29/14 9:47 AM
strg003st001 172.31.134.3                        storage              1.5.1.0-14      OK active                  9/29/14 9:47 AM
strg004st001 172.31.134.4                        storage              1.5.1.0-14      OK active                  9/29/14 9:47 AM
EFSSG1000I The command completed successfully.
```

7.3.4 Modifying the status of Interface nodes and Storage nodes

You can also modify the status of Interface nodes and Storage nodes, as described in this section.

Interface nodes

This section describes the Interface node commands:

1. **`initnode`**: You can use this command to shut down or restart an Interface node. Before you run this command, you must suspend the node.
2. **`suspendnode`**: This command suspends the Interface node, bans the CTDB on the node, and disables the node. A banned node does not participate in the cluster and does not host any records for the CTDB. Its IP address is taken over by another node and no services are hosted.

   Suspending a node is only possible by using the CLI. This action cannot be done by the GUI.

   Using the CLI: Run **`suspendnode`**. Example 7-6 shows the syntax and command output.

Example 7-6 Command usage and output for the suspendnode command

```
[furby.storage.tucson.ibm.com]$ suspendnode --help
usage: suspendnode nodeNames  [--reason <reason>] [--hsm-grace-period <hsm-grace-period>] [--hsm-force] [-c < clusterID | clusterName >]
Suspends a list of nodes.

Parameter              Description
nodeNames              Lists the nodes that will be suspended. Define them with the fully qualified domain name, short name, or IP
address.
    --reason           Specifies the reason for suspending a node. If no reason is provided, maintenance is used.
    --hsm-grace-period Specifies the grace period in seconds for HSM shutdown. The default 60 seconds.
    --hsm-force        Forces an HSM shutdown.
-c, --cluster          The cluster scope for this command
[furby.storage.tucson.ibm.com]$
```

3. **`resumenode:`** The **`resumenode`** command resumes the suspended Interface node. It “unbans” the CTDB on that node and enables the node. The resumed node participates in the cluster and hosts records for the CTDB. It takes back its IP address and starts hosting services.

   Resuming a node is only possible by using the CLI. This action cannot be done by the GUI.

Using the CLI: Run **resumenode**. Example 7-7 shows the syntax and command output.

Example 7-7 Command usage and output for resumenode

```
[furby.storage.tucson.ibm.com]$ resumenode --help
usage: resumenode nodeNames  [-c < clusterID | clusterName >]
Resumes a list of nodes.

Parameter     Description
nodeNames     Lists the nodes that will be resumed. Define them with a fully qualified
domain name, short name, or IP address.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$
```

Storage nodes

With Storage nodes, you can use the **initnode** command. With this command, you can shut down or restart a node. See Example 7-8.

Example 7-8 Help output for the initnode command

```
[furby.storage.tucson.ibm.com]$ initnode --help
usage: initnode {-r | -s} [-n <node>]
-r, --reboot    restart
-s, --shutdown  shutdown
-n <node>       specify target node

[furby.storage.tucson.ibm.com]$
```

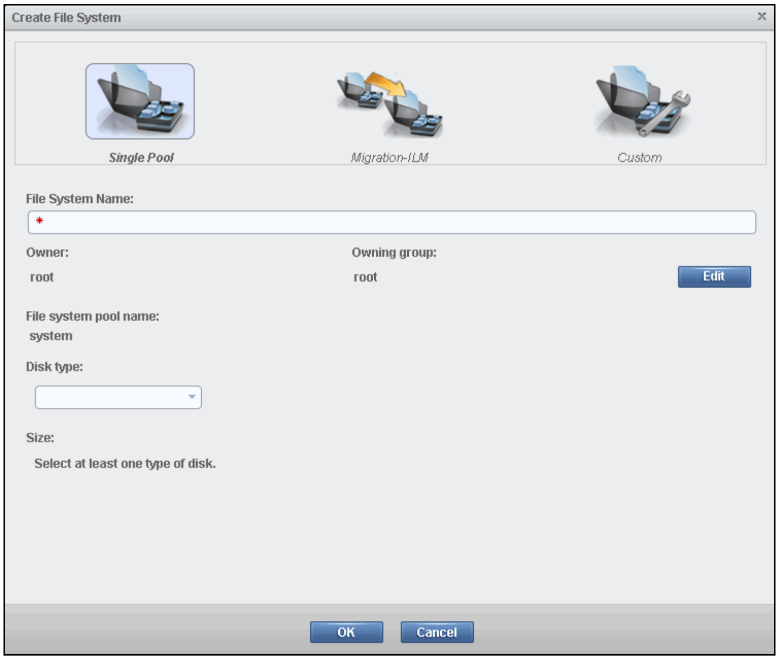
7.4 File system management

File system management is one of the essential tasks in SONAS. The file system that is created is the GPFS file system. Under this category, you can do man tasks, including creating, mounting, unmounting, deleting, changing file system details, adding disks, and more. This section describes some of the important and commonly used file system tasks in detail.

7.4.1 Creating a file system

Using the GUI: You can create the file system by using the GUI (click **Files** → **File Systems)**. The window that is shown in Figure 7-14 on page 503 opens.

Click the **New File System** option. The wizard starts, as shown in Figure 7-36.



File System Name: *

Owner: root

Owning group: root

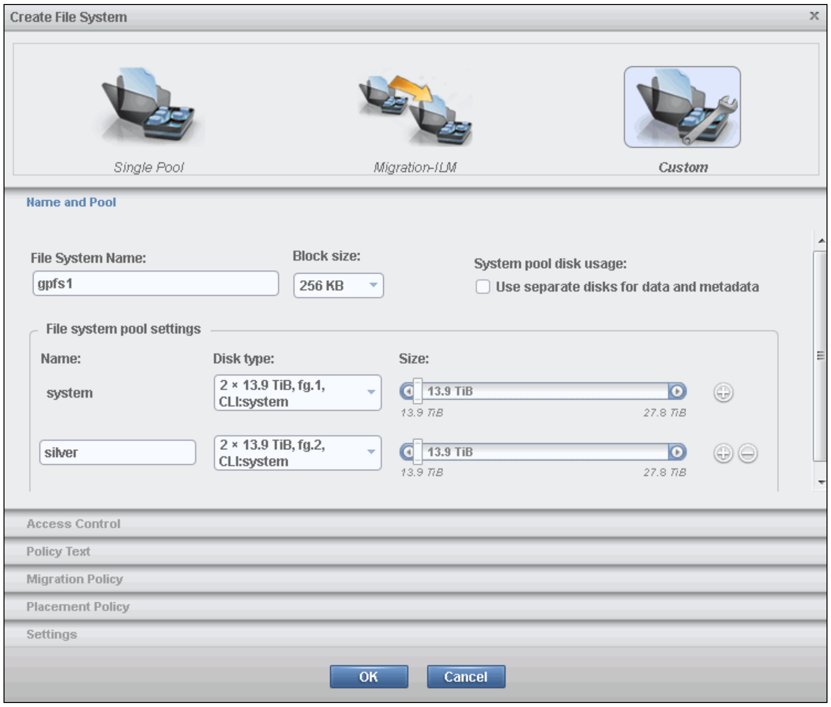
File system pool name: system

Disk type:

Size: Select at least one type of disk.

Figure 7-36 New File System wizard

You can select between multiple options for what type of file system to create. For the purposes of this book, a Custom file system that is named `gpfs1` is created. From this menu, click **Custom**. More options are shown as in Figure 7-37. In this case, if you select multiple pools, you receive two more options during creation. You can immediately configure a Migration and Placement policy. In this case, you must select more than one pool. The first pool is `system` and the second pool is `silver`.



File System Name: gpfs1

Block size: 256 KB

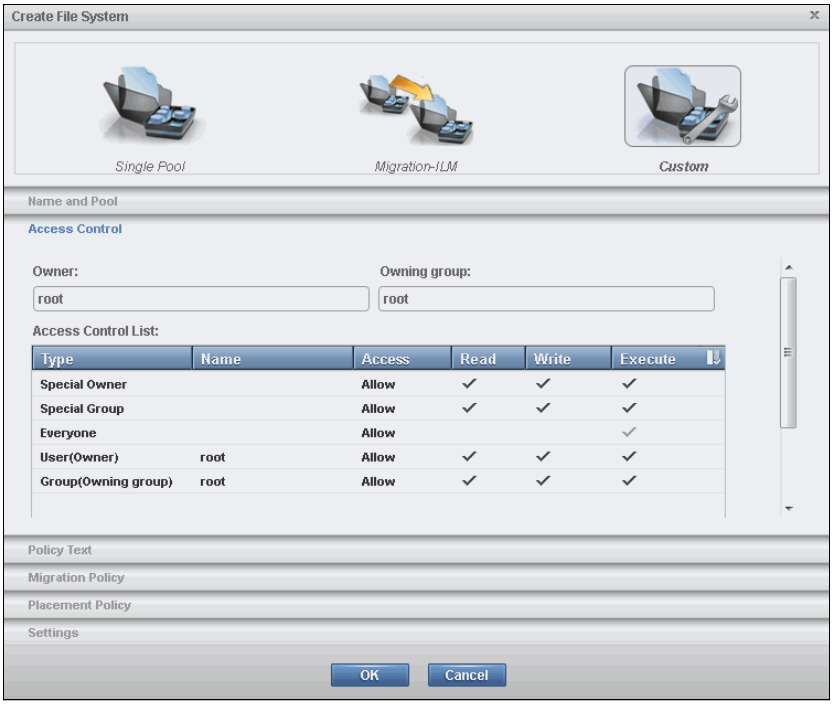
System pool disk usage: □ Use separate disks for data and metadata

File system pool settings

Name:	Disk type:	Size:
system	2 × 13.9 TiB, fg.1, CLI:system	13.9 TiB (13.9 TiB – 27.8 TiB)
silver	2 × 13.9 TiB, fg.2, CLI:system	13.9 TiB (13.9 TiB – 27.8 TiB)

Figure 7-37 Create a file system with custom options

After all the required information is provided, you can configure the ACL on the file system level. Select the **Access Control** option to configure access authority. You can also skip this step. In this case, the system uses the default setting. See Figure 7-38.



Owner: root

Owning group: root

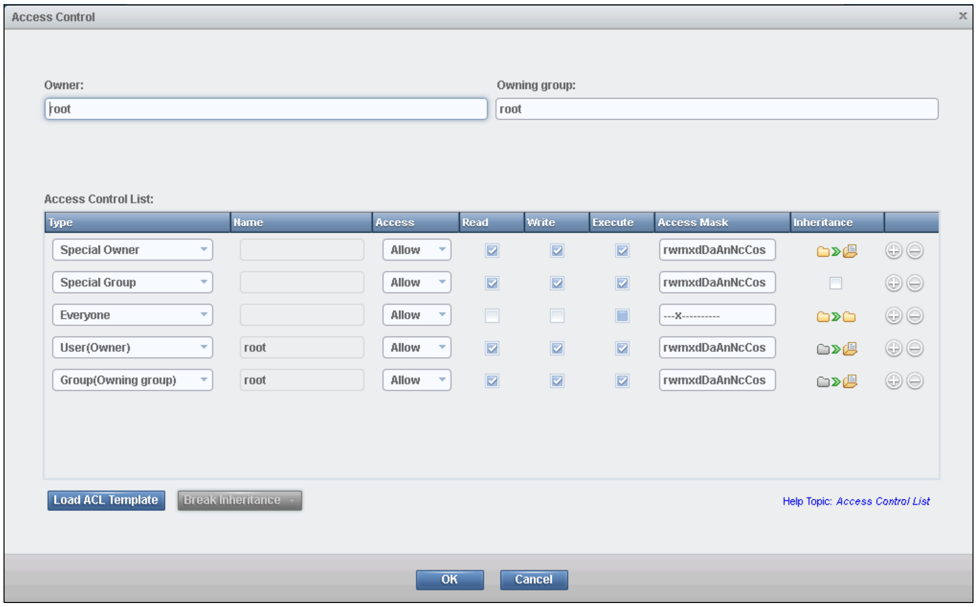
Access Control List:

Type	Name	Access	Read	Write	Execute
Special Owner		Allow	✓	✓	✓
Special Group		Allow	✓	✓	✓
Everyone		Allow			✓
User(Owner)	root	Allow	✓	✓	✓
Group(Owning group)	root	Allow	✓	✓	✓

Figure 7-38 Access Control options

If you want to change or add an ACL, click **Edit** or **Load ACL Template** on the Access Control options. For more information, see 7.5.6, “Managing access control lists” on page 559.

Figure 7-39 shows the Access Control window options.



Owner: root

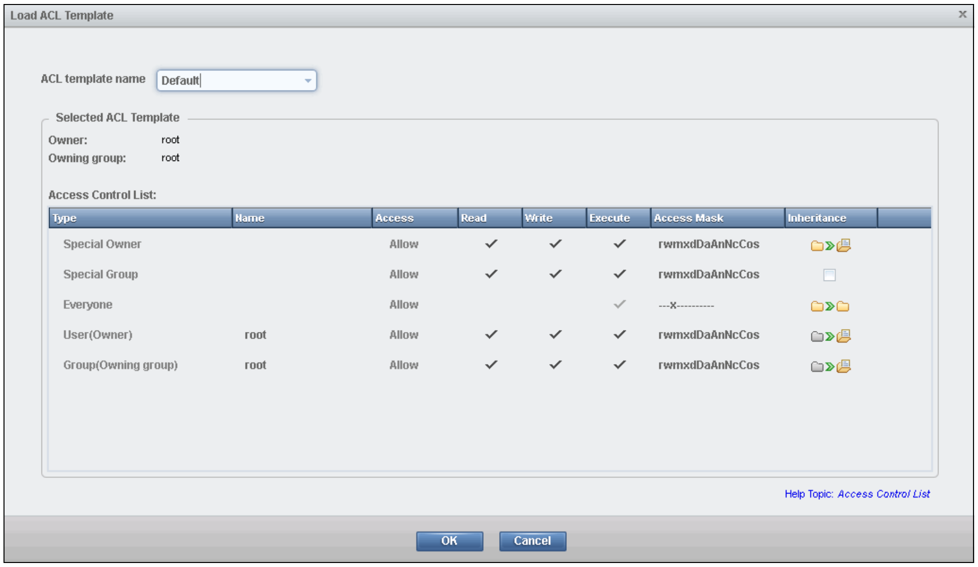
Owning group: root

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance
Special Owner		Allow	☑	☑	☑	rwmxdDaAnNcCos	
Special Group		Allow	☑	☑	☑	rwmxdDaAnNcCos	☐
Everyone		Allow	☐	☐	■	---x----------	
User(Owner)	root	Allow	☑	☑	☑	rwmxdDaAnNcCos	
Group(Owning group)	root	Allow	☑	☑	☑	rwmxdDaAnNcCos	

Figure 7-39 Access Control window options

Figure 7-40 shows the Load ACL Template options.



ACL template name: Default

Owner: root
Owning group: root

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance
Special Owner		Allow	✓	✓	✓	rwmxdDaAnNcCos	📁»📄
Special Group		Allow	✓	✓	✓	rwmxdDaAnNcCos	□
Everyone		Allow			✓	---x---------	📁»📁
User(Owner)	root	Allow	✓	✓	✓	rwmxdDaAnNcCos	📁»📄
Group(Owning group)	root	Allow	✓	✓	✓	rwmxdDaAnNcCos	📁»📄

Figure 7-40 Load ACL Template options

The Policy Text option allows you to add or modify migration policies manually. See Figure 7-41. This option also shows current configured migration and placement policy as text. However, the Migration Policy and Placement Policy options are easier to configure than the Policy Text option.

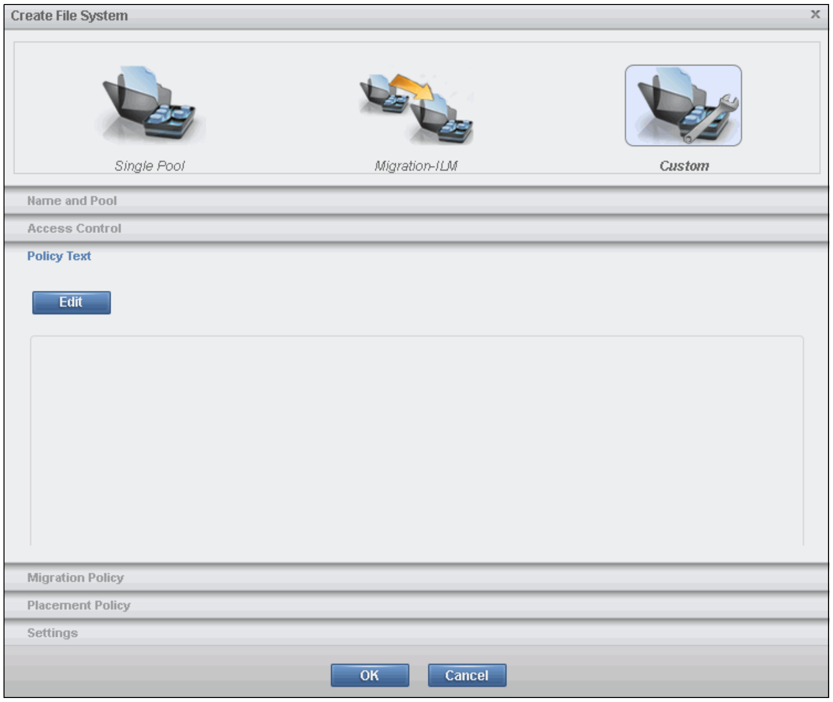
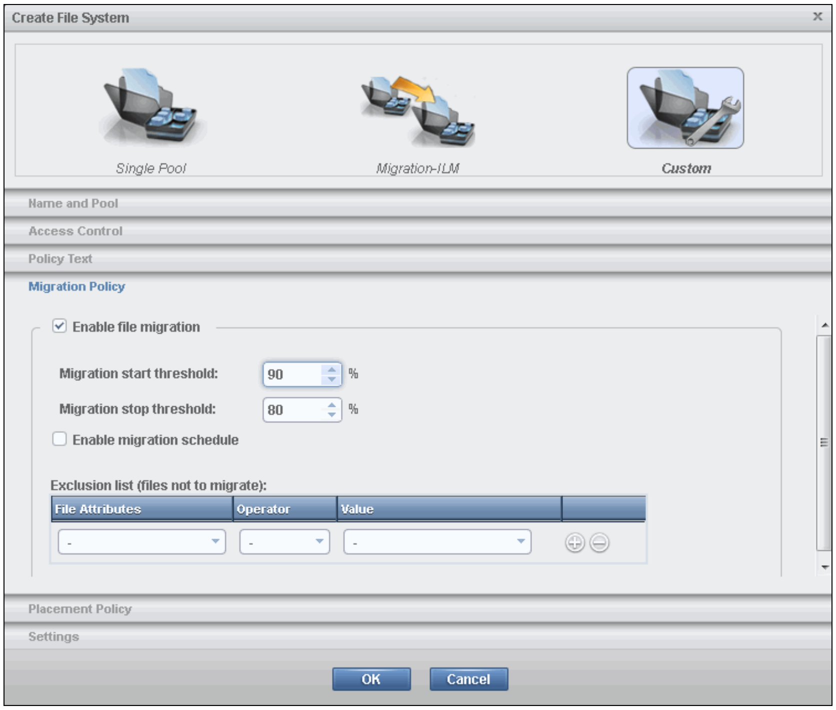


Figure 7-41 Policy Text option

Click **Migration Policy**. Select the **Enable file migration** check box, as shown in Figure 7-42.



☑ Enable file migration

Migration start threshold: 90 %

Migration stop threshold: 80 %

□ Enable migration schedule

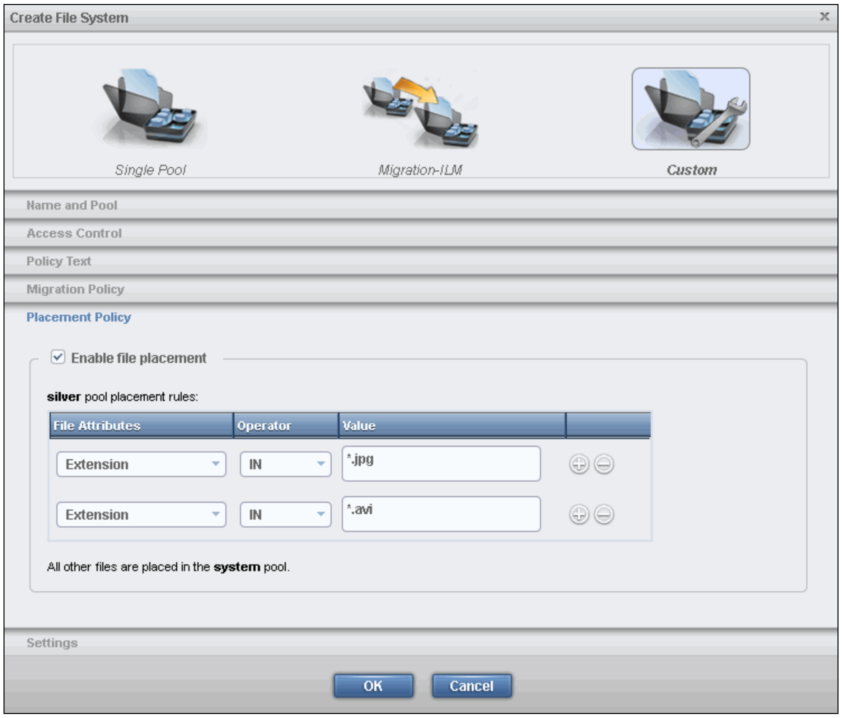
Exclusion list (files not to migrate):

File Attributes	Operator	Value
-	-	-

Figure 7-42   Migration policy options

In this scenario, you configure migration to start when threshold reaches 90% and stop when it reaches 80%. The migration start time is set to after the threshold is reached. No create exclusion lists are created. The migration start can be set to a specific time, for example, once per day at 8.00 PM.

When all information is entered, select **Placement Policy** if it is to be used. As with a migration policy, select the check box to use a placement policy. In this example, you create a placement policy where all files with extensions `.jpg` and `.avi` are moved to the `silver` pool. All others go to the `system` pool. See Figure 7-43. You can also set a placement policy for a specific user or group.



☑ Enable file placement

silver pool placement rules:

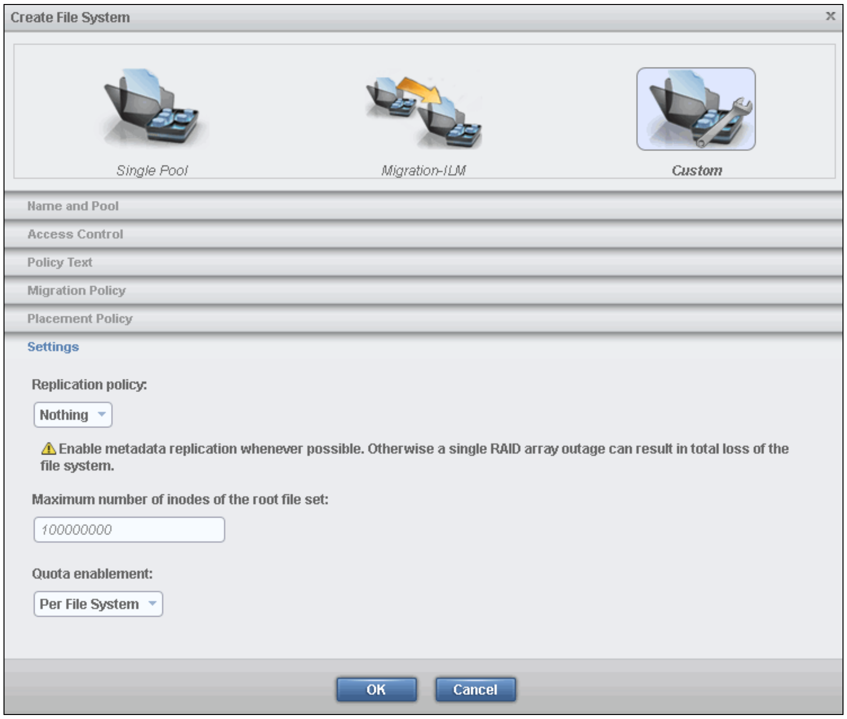
File Attributes	Operator	Value
Extension	IN	*.jpg
Extension	IN	*.avi

All other files are placed in the **system** pool.

Figure 7-43 Placement policy options

The last options are settings. Here you can configure follow tree items and each item has several options. See Figure 7-44.

- Replication policy
  - Nothing
  - MetaData
  - Everything
- Maximum number of inodes of the root file set
- Quota enablement
  - Per file system
  - Per file set
  - Disabled



Replication policy: Nothing

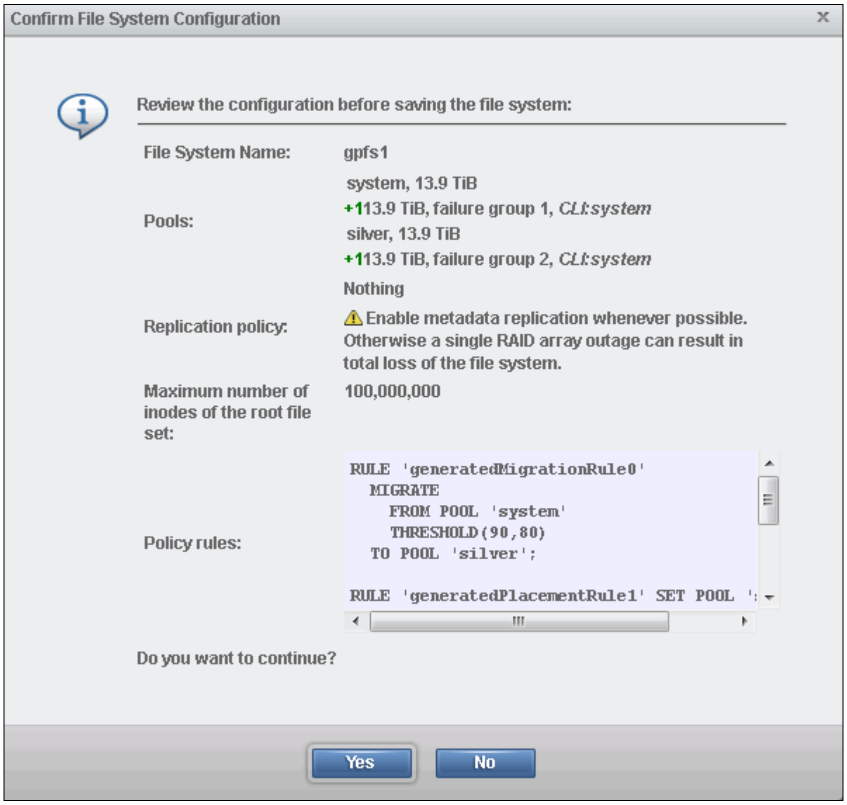
⚠ Enable metadata replication whenever possible. Otherwise a single RAID array outage can result in total loss of the file system.

Maximum number of inodes of the root file set: 100000000

Quota enablement: Per File System

Figure 7-44 Settings options

After you check all the tabs, you can select **OK**. The summary windows open. See Figure 7-45 on page 529.



Review the configuration before saving the file system:

File System Name: gpfs1

Pools: system, 13.9 TiB
+113.9 TiB, failure group 1, *CLI:system*
silver, 13.9 TiB
+113.9 TiB, failure group 2, *CLI:system*

Replication policy: Nothing
⚠ Enable metadata replication whenever possible. Otherwise a single RAID array outage can result in total loss of the file system.

Maximum number of inodes of the root file set: 100,000,000

Policy rules:

```
RULE 'generatedMigrationRule0'
  MIGRATE
    FROM POOL 'system'
    THRESHOLD(90,80)
  TO POOL 'silver';

RULE 'generatedPlacementRule1' SET POOL ':
```

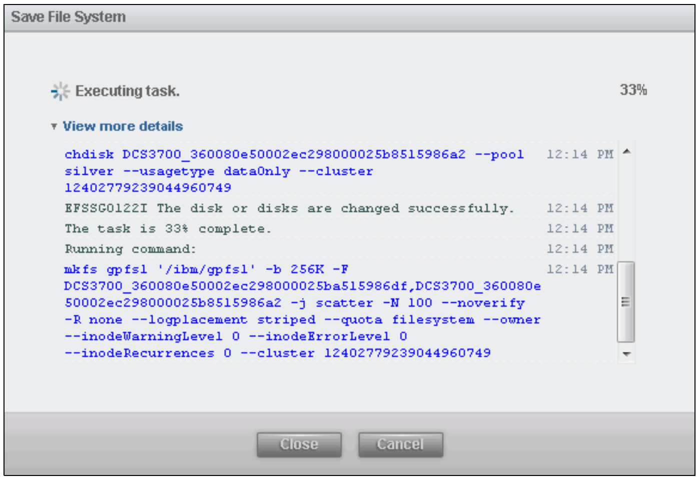
Do you want to continue?

Figure 7-45 Confirm File System Configuration window

Click **Yes** to proceed. The GUI shows the command that is used. However, in this example, multiple commands are run.

In the following screen captures (Figure 7-46, Figure 7-47 on page 530, and Figure 7-48 on page 530), you can see the entire process of creating the file system.

As shown in Figure 7-46, multiple CLI commands are run to achieve the task. These commands include **`chdisk`** and **`mkfs`**.

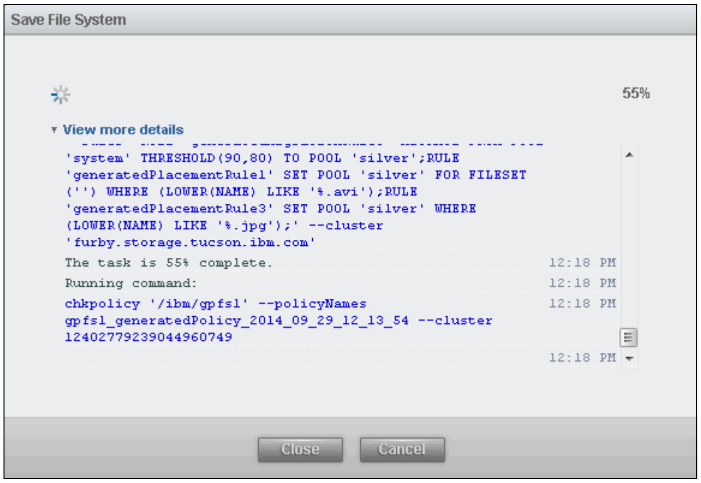


Executing task. 33%

```
chdisk DCS3700_360080e50002ec298000025b8515986a2 --pool   12:14 PM
silver --usagetype dataOnly --cluster
12402779239044960749
EFSSG0122I The disk or disks are changed successfully.    12:14 PM
The task is 33% complete.                                 12:14 PM
Running command:                                          12:14 PM
mkfs gpfs1 '/ibm/gpfs1' -b 256K -F                        12:14 PM
DCS3700_360080e50002ec298000025ba515986df,DCS3700_360080e
50002ec298000025b8515986a2 -j scatter -N 100 --noverify
-R none --logplacement striped --quota filesystem --owner
--inodeWarningLevel 0 --inodeErrorLevel 0
--inodeRecurrences 0 --cluster 12402779239044960749
```

Figure 7-46 Save File System procedure shown 1

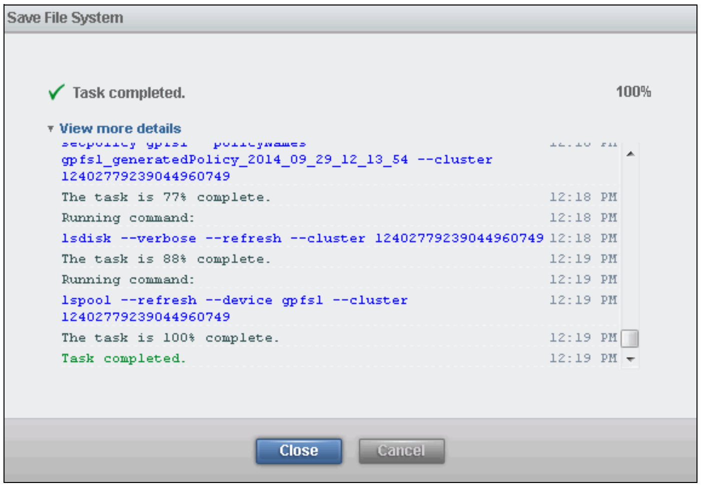
In Figure 7-47, you can see the intermediate steps that occur while file system `gpfs1` with migration and placement is being created. In this case, the procedure was at 55%.



```
'system' THRESHOLD(90,80) TO POOL 'silver';RULE
'generatedPlacementRule1' SET POOL 'silver' FOR FILESET
('') WHERE (LOWER(NAME) LIKE '%.avi');RULE
'generatedPlacementRule3' SET POOL 'silver' WHERE
(LOWER(NAME) LIKE '%.jpg');' --cluster
'furby.storage.tucson.ibm.com'
The task is 55% complete.                                  12:18 PM
Running command:                                           12:18 PM
chkpolicy '/ibm/gpfs1' --policyNames                       12:18 PM
gpfs1_generatedPolicy_2014_09_29_12_13_54 --cluster
12402779239044960749
                                                           12:18 PM
```

Figure 7-47 Save File System procedure shown 2

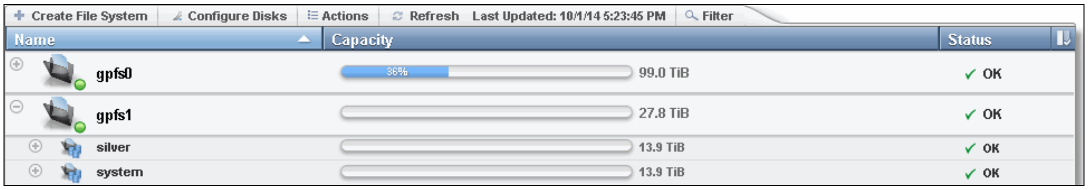
In Figure 7-48, you can see that the procedure of creating the `gpfs1` file system finished.



```
gpfs1_generatedPolicy_2014_09_29_12_13_54 --cluster
12402779239044960749
The task is 77% complete.                                  12:18 PM
Running command:                                           12:18 PM
lsdisk --verbose --refresh --cluster 12402779239044960749 12:18 PM
The task is 88% complete.                                  12:19 PM
Running command:                                           12:19 PM
lspool --refresh --device gpfs1 --cluster                  12:19 PM
12402779239044960749
The task is 100% complete.                                 12:19 PM
Task completed.                                            12:19 PM
```

Figure 7-48 Save File System ends

When file system is created, you can see it in the File System navigational window, as shown in Figure 7-49.



Name	Capacity	Status
gpfs0	36% 99.0 TiB	✓ OK
gpfs1	27.8 TiB	✓ OK
silver	13.9 TiB	✓ OK
system	13.9 TiB	✓ OK

Figure 7-49 Newly created file system gpfs1

Using the CLI: You can create a file system by running **mkfs**. The NSD name is mandatory and you must enter at least one NSD. The NSD names can be found by running **lsdisk** or by clicking **Monitoring** → **System details** and then selecting disks from the menu. Set **-R** (replication) to `none` if you do not want to enable replication. If you do enable replication, you must enter at least two NSDs, where both of these NSDs belong to different failure groups. The block size and replication factors that are chosen affect file system performance. Example 7-9 shows the help options for the **mkfs** command.

Example 7-9 mkfs help command example

```
[furby.storage.tucson.ibm.com]$ mkfs --help
usage: mkfs filesystem  [mountpoint] [-b <blocksize>] {-F <disks> | --pool <poolName> | --discoverdisks | --createdisks
<size,mdiskgrp1[,mdiskgrp2][,N=numDisks][,compressed:rsize]>} [-i <maxNumInodes:numInodesToPreallocate>] [-j <blockAllocationType>] [-N
<numNodes>] [--dmapi | --nodmapi] [--noverify] [-R <policy>] [--logplacement <logPlacement>] [-q <quota>] [--snapdir | --nosnapdir]
[--owner [owner][:[group][:template]]] [--inodeWarningLevel <percentage>] [--inodeErrorLevel <percentage>] [--inodeRecurrences
<recurrences>] [-c < clusterID | clusterName >]
Creates a file system.

Parameter           Description
filesystem          Name of the file system to be created.
mountpoint          Specifies the mount point of the new file system.
-b                  Specifies the size of the data blocks. The size must be 256 KB (default size), 1 MB, or 4 MB.
-F                  Specifies the disks on which the file system is to be created.
    --pool          Specifies the file system pool on which the file system is to be created.
    --discoverdisks (IFS only) Detects and uses all free GPFS NSDs that are tagged for the specified file system but not yet included.
    --createdisks   (IFS only) Creates disks implicitly and adds them to the file system.
-i                  Specifies the maximum number of files for this file system.
-j                  Specifies the block allocation map type (cluster/scatter).
-N                  Specifies the estimated number of nodes that will mount this file system.
    --dmapi         Enables DMAPI support. For example, HSM and Antivirus.
    --nodmapi       Disables DMAPI support. For example, HSM and Antivirus.
    --noverify      Skips the verification of the disk descriptor, so that disks that contain on old descriptor can be reused.
-R                  Specifies the replication policy (none/meta/all) for this file system.
    --logplacement  Sets whether the logs will be stored striped across the disks or not.
-q, --quota         Specifies the quota option (yes/no/perfileset) for this file system.
    --snapdir       Adds a snapshots directory to all subdirectories in the file system.
    --nosnapdir     All invisible snapshot directories are removed.
    --owner         Sets the directory owner.
    --inodeWarningLevel Sets a custom percentage of inodes to be used before generating a warning message. Specify 0 to use the system
default.
    --inodeErrorLevel   Sets a custom percentage of inodes to be used before generating an error message. Specify 0 to use the system
default.
    --inodeRecurrences  Sets a custom number of recurrences of inode usage messages to log. Specify 0 to use the system default.
-c, --cluster       The cluster scope for this command
```

7.4.2 Listing the file system status

This section shows how to list the file system status.

Using the GUI: Click **Files** → **File Systems**. In this view, all created file systems are listed. See Figure 7-49 on page 530. You can check the properties for the selected file system by clicking **Action** → **Properties**.

Using the CLI: You can view the status of the file systems by running **lsfs**. The command displays the file system names, mount point, Quota, Blocksize ACL Types, Replication details, and more. See the usage and command output in Example 7-10.

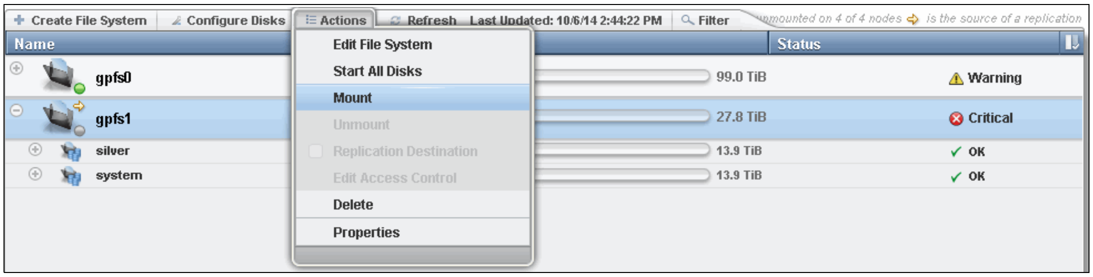
Example 7-10 Command usage and output for the lsfs command

```
[furby.storage.tucson.ibm.com]$ lsfs --help
usage: lsfs  [-v] [-r] [-Y] [-d <name>] [-c < clusterID | clusterName >]
Lists all the file systems on a given device in a cluster.

Parameter     Description
-v, --verbose Shows additional columns.
-r, --refresh Forces a refresh of the exports data in the database by scanning all cluster exports before retrieving data from the
database.
-Y            Shows parseable output.
-d, --device  Defines device.
-c, --cluster The cluster scope for this command
```

7.4.3 Mounting the file system

This section shows how to mount the file system.

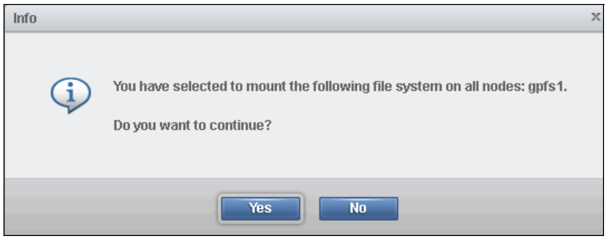
Using the GUI: Click **Files** → **File Systems**. Select the file system that you want, and click **Actions** → **Mount**. If it is disabled, the file system is mounted. If it is enabled, the selected file system is not mounted. See Figure 7-50.



Name		Status
gpfs0	99.0 TiB	⚠ Warning
gpfs1	27.8 TiB	✖ Critical
silver	13.9 TiB	✓ OK
system	13.9 TiB	✓ OK

Figure 7-50 Actions menu for the file system

To mount a selected file system, click **Mount**. A new window opens for confirmation, as shown in Figure 7-51.



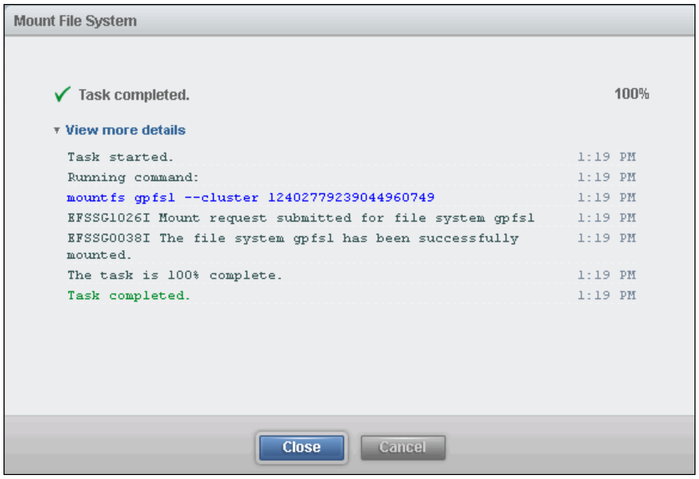
Info

You have selected to mount the following file system on all nodes: gpfs1.

Do you want to continue?

Figure 7-51 Confirmation window

When you click **Yes**, the GUI shows the command that is run for mounting this file system. See Figure 7-52.



Mount File System

✓ Task completed. 100%

▾ View more details

```
Task started.                                              1:19 PM
Running command:                                           1:19 PM
mountfs gpfs1 --cluster 12402779239044960749               1:19 PM
EFSSG1026I Mount request submitted for file system gpfs1   1:19 PM
EFSSG0038I The file system gpfs1 has been successfully     1:19 PM
mounted.
The task is 100% complete.                                 1:19 PM
Task completed.                                            1:19 PM
```

Figure 7-52 Mount of the file system gpfs1 finished successfully

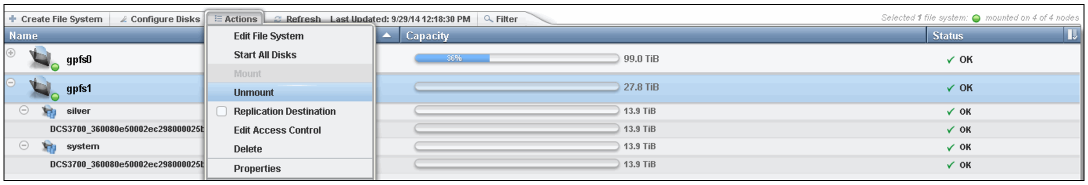
Using the CLI: You can mount the file system by running **mountfs**. The command help output is displayed in Example 7-11.

Example 7-11 Command usage and output for the mountfs command

```
[furby.storage.tucson.ibm.com]$ lsfs --help
usage: lsfs  [-v] [-r] [-Y] [-d <name>] [-c < clusterID | clusterName >]
Lists all the file systems on a given device in a cluster.

Parameter        Description
-v, --verbose    Shows additional columns.
-r, --refresh    Forces a refresh of the exports data in the database by scanning all cluster exports before retrieving data from the
database.
-Y               Shows parseable output.
-d, --filesystem The file system
-c, --cluster    The cluster scope for this command
[furby.storage.tucson.ibm.com]$ ^C
[furby.storage.tucson.ibm.com]$ mountfs --help
usage: mountfs fileSystem  [-n <nodeNames>] [-c < clusterID | clusterName >]
Mounts a file system.

Parameter     Description
fileSystem    Specifies the name of the file system to be mounted. File system names do not need to be fully qualified, but they must be
unique within a GPFS cluster.
-n            Lists the nodes on which the file system is to be mounted. The list is comma-separated. Specify only interface or Management
nodes. If omitted, the file system is mounted on all interface or Management nodes.
-c, --cluster The cluster scope for this command
```

7.4.4 Unmounting the file system

Using the GUI: To unmount a file system in the GUI, click **Files** → **File Systems**. All file systems that are configured are shown. To unmount one, you must select it, and click **Actions** → **Unmount**. See Figure 7-53. In this case, file system `gpfs0` is unmounted. When you click this option, a new window opens as confirmation for unmounting.



Name	Capacity		Status
gpfs0	36%	99.0 TiB	✓ OK
gpfs1		27.8 TiB	✓ OK
silver		13.9 TiB	✓ OK
DCS3700_360080e50002ec298000025t		13.9 TiB	✓ OK
system		13.9 TiB	✓ OK
DCS3700_360080e50002ec298000025t		13.9 TiB	✓ OK

Figure 7-53 Unmount a file system

Using the CLI: You can unmount the file system by running **unmountfs**. The command output is shown in Example 7-12.

Example 7-12 Command usage and output for the unmountfs command

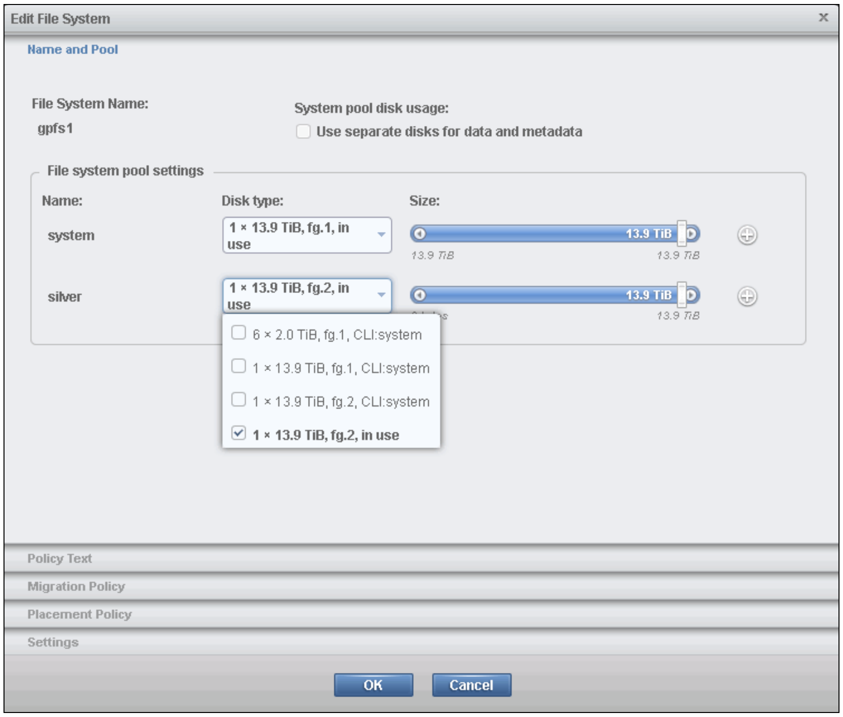
```
[furby.storage.tucson.ibm.com]$ unmountfs --help
usage: unmountfs fileSystem  [-n <nodeNames>] [-w] [-c < clusterID | clusterName >]
Unmounts a file system.

Parameter              Description
fileSystem             Specifies the name of the file system to be unmounted. File system names do not need to be fully qualified, but
they must be unique within a GPFS cluster.
-n                     Lists the nodes to unmount the file system on, in a comma-separated list. If this option is omitted, the file
system is unmounted on all nodes.
-w, --wait-for-complete Indicates that the system should wait until the file system is unmounted from all the nodes. Timeout with error
after 3 minutes.
-c, --cluster          The cluster scope for this command
```

7.4.5 Modifying the file system configuration

This section describes how to modify the file system configuration.

Using the GUI: You can use the SONAS GUI to modify the file system configuration of a file system. Some of the parameters require that the file system is unmounted, and some actions can be done while it is still mounted.

Using the GUI: To change file system configuration parameter, click **Files** → **File Systems**. From the list of file systems, select the one for which you want to change a few parameters. In this example, you change a few parameters to the file system that you previously created, that is, `gpfs1`. First, select file system `gpfs1`, and then click **Actions** → **Edit File System**. See Figure 7-53 on page 533. When you select **Edit File System**, a window that is similar to the one for creating a file system opens. The wizard is shown in Figure 7-54.



File System Name:
gpfs1

System pool disk usage:
□ Use separate disks for data and metadata

File system pool settings

Name:	Disk type:	Size:
system	1 × 13.9 TiB, fg.1, in use	13.9 TiB (13.9 TiB – 13.9 TiB)
silver	1 × 13.9 TiB, fg.2, in use	13.9 TiB (– 13.9 TiB)

□ 6 × 2.0 TiB, fg.1, CLI:system
□ 1 × 13.9 TiB, fg.1, CLI:system
□ 1 × 13.9 TiB, fg.2, CLI:system
☑ 1 × 13.9 TiB, fg.2, in use

Figure 7-54 Edit file system gpfs1

In this case, add a pool named *gold*. After you make all of the changes and modifications that you want, select **OK**.

Using the CLI: You can change the file system parameters by running **`chfs`**. Example 7-13 shows the help output for the **`chfs`** command. In the CLI, you have many more options for changing file system configuration parameters than in the GUI. You can also set a maximum number of inodes in the file system.

Example 7-13 Command usage and output by adding disk to change properties

```
[furby.storage.tucson.ibm.com]$ chfs --help
usage: chfs fileSystem  [--atime <accessTime>] [--mtime <modTime>] [-i <maxNumInodes:numInodesToPreallocate>] [-q <quota>] [-R
<replication>] [--pool <poolName> | --add <disks>] [--remove <disks>] [--force] [--noverify] [--dmapi | --nodmapi] [--logplacement
<logPlacement>] [--snapdir | --nosnapdir] [--inodeWarningLevel <percentage>] [--inodeErrorLevel <percentage>] [--inodeRecurrences
<recurrences>] [-c < clusterID | clusterName >]
Changes the properties of the file system.

Parameter              Description
fileSystem             Specifies the name of the file system to be changed. File system names do need not be fully qualified, but they
must be unique within a GPFS cluster.
```

```
    --atime             Stamps access times on every access to a file or directory if the 'exact' variable is selected. If the 'suppress'
variable is selected, access times will not be recorded.
    --mtime            Updates the modification time immediately to files and directories if the 'exact' variable is selected. Otherwise,
modification times will be updated after a delay of several seconds.
-i                      Sets the maximum number of inodes in the file system.
-q                      Changes the quota option (yes, no, perfileset).Changing this setting requires the file system to be in unmounted
state.
-R                      Sets the level of replication in this file system.
    --pool              Adds a set of free Network Shared Disks (NSDs), which have the pool name set as pool, to the file system.
    --add               Adds disks to the file system. The disks contain a comma-separated list of disk names.
    --remove            Removes disks from the file system. The disks variable contains a comma-separated list of disk names.
    --force             Does not prompt for manual confirmation.
    --noverify          Suppresses the verification that specified disks do not belong to an existing file system.
    --dmapi             Enables external file system pool support, such as Tivoli Storage Manager.
    --nodmapi           Disables external file system pool support, such as Tivoli Storage Manager.
    --logplacement      Sets whether the logs will be stored striped across the disks or not.
    --snapdir           Adds a snapshots subdirectory to all subdirectories in the file system.
    --nosnapdir         All invisible snapshot directories are removed.
    --inodeWarningLevel Sets a custom percentage of inodes to be used before generating a warning message. Specify 0 to use the system
default.
    --inodeErrorLevel   Sets a custom percentage of inodes to be used before generating an error message. Specify 0 to use the system
default.
    --inodeRecurrences  Sets a custom number of recurrences of inode usage messages to log. Specify 0 to use the system default.
-c, --cluster           The cluster scope for this

command
```

7.4.6 Deleting a file system

This section describes how to delete a file system.

Using the GUI: To delete the `gpfs1` file system, click **Files** → **File Systems**. Then, click **Actions** → **Delete**, as shown in Figure 7-55.

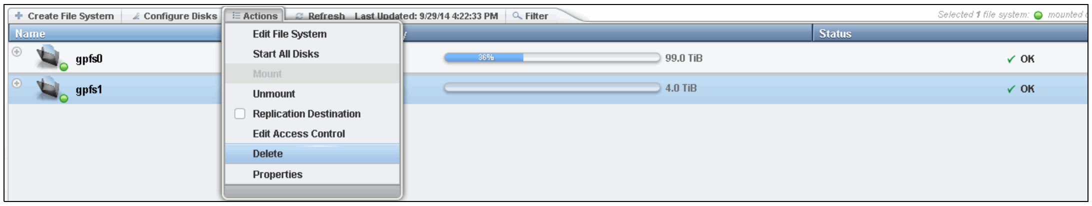
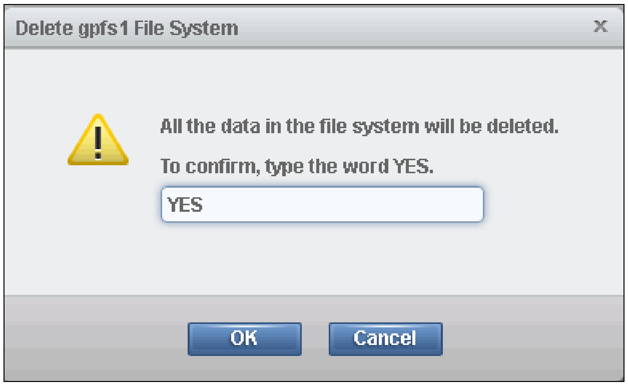


Figure 7-55 Delete file system gpfs1

As you can see, the file system is unmounted. When you select the **Delete** option, a new window opens, as shown in Figure 7-56. Enter YES to confirm.



Delete gpfs1 File System

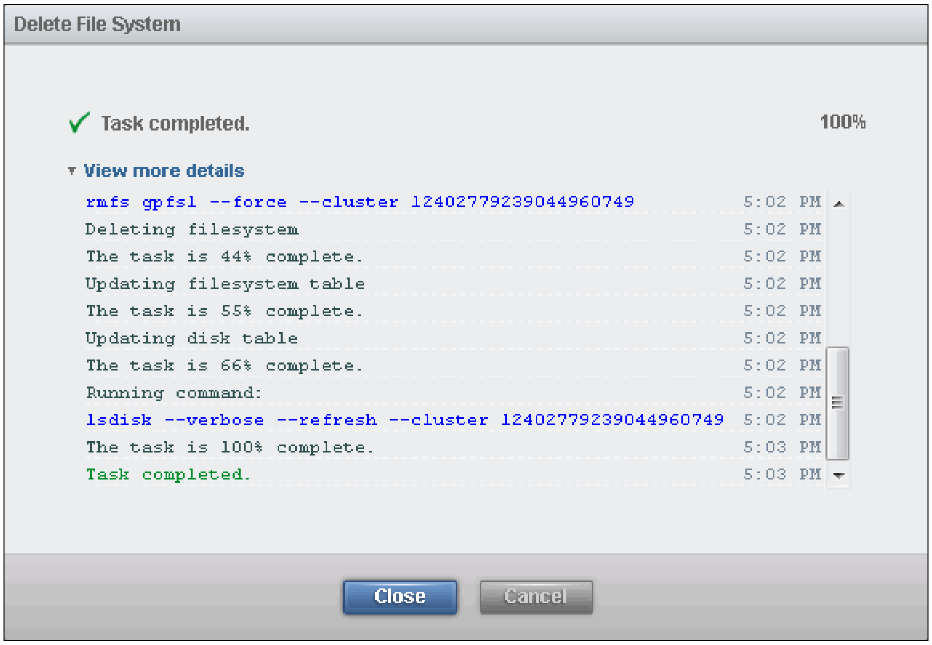
All the data in the file system will be deleted.

To confirm, type the word YES.

YES

Figure 7-56 Confirmation window before you delete a file system

After you enter YES, click **OK**. Because of operations that are running in the background that are connected with NSDs, this command might take a few seconds to run. Next, the window that is shown in Figure 7-57 opens.



```
rmfs gpfsl --force --cluster 12402779239044960749        5:02 PM
Deleting filesystem                                      5:02 PM
The task is 44% complete.                                5:02 PM
Updating filesystem table                                5:02 PM
The task is 55% complete.                                5:02 PM
Updating disk table                                      5:02 PM
The task is 66% complete.                                5:02 PM
Running command:                                         5:02 PM
lsdisk --verbose --refresh --cluster 12402779239044960749 5:02 PM
The task is 100% complete.                               5:03 PM
Task completed.                                          5:03 PM
```

Figure 7-57 Deletion of the file system is complete

Using the CLI: You can delete an existing file system from the cluster by running **rmfs**. The command help output is shown in Example 7-14.

Example 7-14 Command usage and output for removing the file system

```
[furby.storage.tucson.ibm.com]$ rmfs --help
usage: rmfs fileSystem  [--force] [-c < clusterID | clusterName >]
Removes an existing file system from the cluster.

Parameter     Description
fileSystem    Specifies the name of the file system to be removed. File system names do not need
to be fully qualified, but they must be unique within a GPFS cluster.
    --force   Do not prompt for manual confirmation.
-c, --cluster The cluster scope for this command
```

7.4.7 Quota management for file systems

You can use the SONAS file system add quotas to the file systems and to the users and groups that exist on the system. You can set the quota for a user, a group, or a file set. Soft limits are subject to reporting, and hard limits are enforced by the file system or until the grace period is exceeded. To use quotas for file system, you must create a file set in that file system.

Using the GUI: There are three options for creating quotas. The options are User, Group, and File Set. Here is an example for user quota:

1. To create a user quota in the GUI, click **Files** → **Quotas**. Select one file system and click **Actions**. You can see three activated items for Create Quota. See Figure 7-58.

.

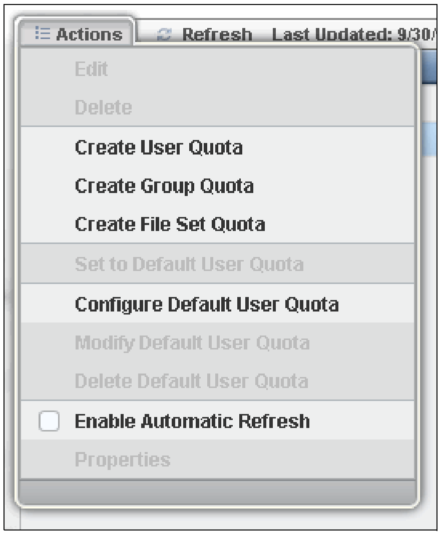
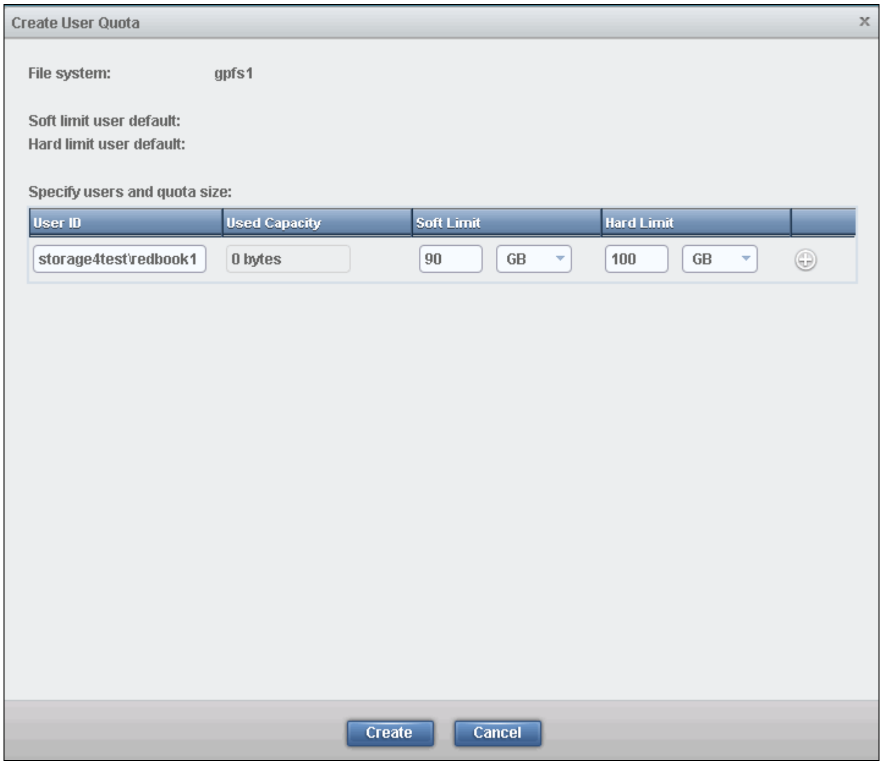

Actions
Refresh
Last Updated: 9/30/
Edit
Delete
Create User Quota
Create Group Quota
Create File Set Quota
Set to Default User Quota
Configure Default User Quota
Modify Default User Quota
Delete Default User Quota
□ Enable Automatic Refresh
Properties

Figure 7-58 Actions menu in the Quota section

2. To create the user quota, click **Create User Quota.** The wizard starts, as shown in Figure 7-59.



File system: gpfs1

Soft limit user default:

Hard limit user default:

Specify users and quota size:

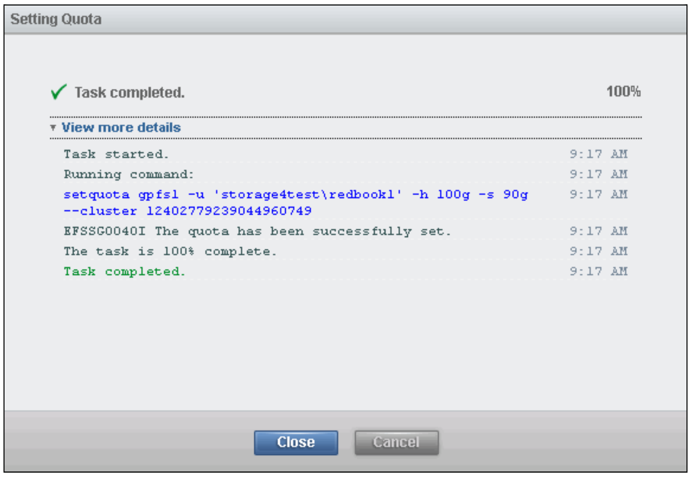
User ID	Used Capacity	Soft Limit		Hard Limit		
storage4test\redbook1	0 bytes	90	GB	100	GB	⊕

Figure 7-59 New quota wizard

Note: SONAS V1.5 provides a default user quota for file systems. To use the default, select a file system and click **Actions** → **Configure Default User Quota**.

Note: You can also create a file set quota during file set creation. For more information, see 7.4.8, “File set management” on page 539.

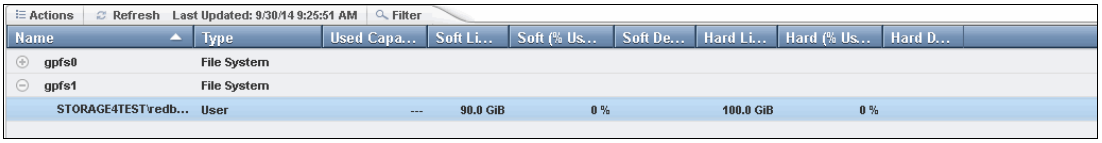
3. The quota is for user *redbook1* in the *storage4test* domain, with the soft limit set to 90G, and the hard limit set to 100 G. Enter the values and click **Create**. The window that is shown in Figure 7-60 opens.



```
Task started.                                             9:17 AM
Running command:                                          9:17 AM
setquota gpfs1 -u 'storage4test\redbook1' -h 100g -s 90g  9:17 AM
--cluster 12402779239044960749
EFSSG0040I The quota has been successfully set.           9:17 AM
The task is 100% complete.                                9:17 AM
Task completed.                                           9:17 AM
```

Figure 7-60 Setting the quota wizard finished

The configured user quota is listed under the file system. See Figure 7-61.



Name	Type	Used Capa...	Soft Li...	Soft (% Us...	Soft De...	Hard Li...	Hard (% Us...	Hard D...
⊕ gpfs0	File System							
⊖ gpfs1	File System							
STORAGE4TEST\redb...	User	---	90.0 GiB	0 %		100.0 GiB	0 %	

Figure 7-61 User Quota list

Using the CLI: To create a quota in the CLI, run **`setquota`**. Example 7-15 shows the **`help`** output for the **`setquota`** command. Figure 7-60 shows the command that is used in the GUI.

Example 7-15 Help output for the setquota command

```
[furby.storage.tucson.ibm.com]$ setquota --help
usage: setquota filesystem  {-u <user> | -g <group> | -j <fileset> | --default} [-h <hardLimit>] [-s <softLimit>] [-H <iHardLimit>] [-S
<iSoftLimit>] [-c < clusterID | clusterName >]
Sets the quota settings.

Parameter     Description
filesystem    Specifies the name of the file system. File system names do not need to be fully qualified.
-u            Specifies the name of the user.
-g            Specifies the name of the group.
-j            Specifies the name of the file set.
-h            Specifies the hard limit of disk usage.
-s            Specifies the soft limit of disk usage.
-H            Specifies the hard limit of inodes.
-S            Specifies the soft limit of inodes.
    --default Specify default quota
-c, --cluster The cluster scope for this command

Accepted multiplicative suffixes for hardLimit and softLimit:
  'k' : kiloByte, 'm' : MegaByte, 'g' : GigaByte, 't' : TeraByte, 'p' : PetaByte
Accepted multiplicative suffixes for iHardLimit and iSoftLimit:
  'k' : kiloByte, 'm' : MegaByte
```

```
These suffixes are case insensitive.

A quota limit can be removed by setting it to 0
```

To list already configured quotas in the CLI, run **`lsquota`**. In Example 7-16, the help option and command result are shown.

Example 7-16 Help option for the lsquota command and output for lsquota

```
[furby.storage.tucson.ibm.com]$ lsquota --help
usage: lsquota  [-r] [-Y] [-d <name>] [-j <fileset> | -u <userName> | -g <groupName>] [-v] [--default] [-c < clusterID | clusterName >]
Lists all quotas.

Parameter        Description
-r, --refresh    Forces a refresh of the quota data in the database by scanning the cluster before retrieving data from the database.
-Y               Shows parseable output.
-d, --filesystem The file system
-j, --fileset    Lists quota information for a given file set, or all file sets if none is given.
-u, --user       Lists quota information for a given user, or all users if none is given.
-g, --group      Lists quota information for a given group, or all groups if none is given.
-v               Show additional information.
    --default    Show only default quotas.
-c, --cluster    The cluster scope for this command
```

7.4.8 File set management

You can use SONAS to create file sets.

Overview

A file set is a group of files. They are created inside an existing file system. They are similar to file systems in some ways because you can perform file system operations on them. You can replicate, set quotas, and also create snapshots. File sets are not mounted but are linked or unlinked. You can link a file set to a directory. This function creates a junction or a link. The directory to which you link the file set should not be an existing directory. It is created when you link and deleted when you unlink. You also can view, create, remove, link, and unlink on a file set.

Viewing or listing file sets

Using the GUI: To view file sets, click **Files** → **File Sets**, as shown in Figure 7-16 on page 504. All the file sets are listed.

Using the CLI: You can view the file sets in the cluster by running **`lsfset`**. This command lists all the file sets along with the details.

Example 7-17 shows the command usage and output of the **`lsfset`** command.

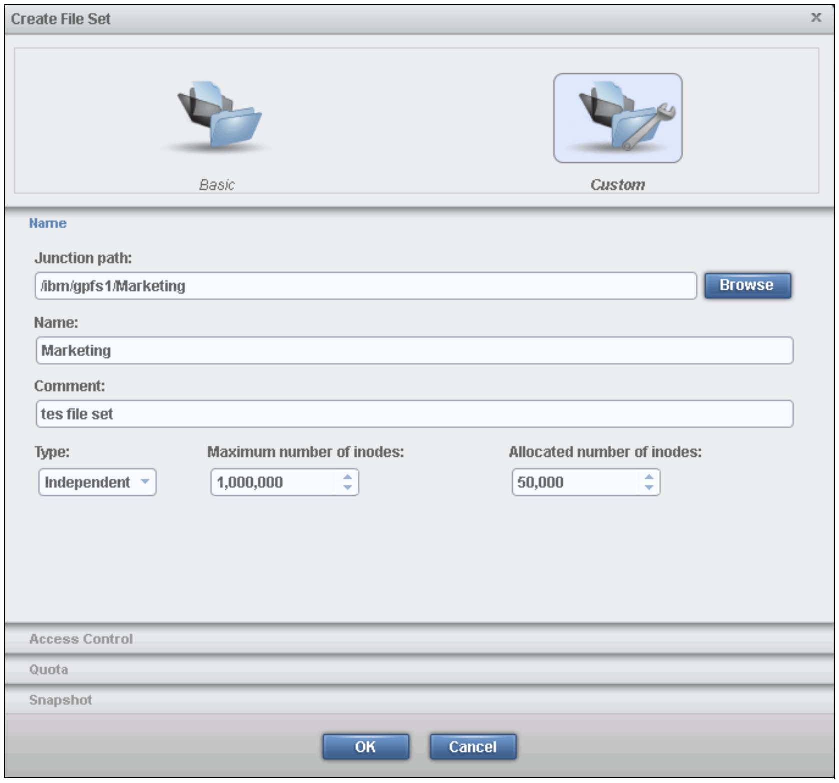
Example 7-17 Usage and output for the lsfset command

```
[furby.storage.tucson.ibm.com]$ lsfset --help
usage: lsfset fileSystem  [-v] [-r] [-u] [-Y] [-c < clusterID | clusterName >]
Lists file sets for a given device in a cluster.

Parameter     Description
fileSystem    Specifies the device name of the file system to contain the file set. File system names do need not be fully qualified.
-v, --verbose Shows additional columns.
-r, --refresh Forces a refresh of the file set data in the database by scanning all the cluster file sets before retrieving data from the
database.
-u, --usage   Forces a refresh of the file set and file set usage data in the database, by scanning all cluster file sets before
retrieving data from the database.
-Y            Shows parseable output.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$ lsfset gpfs1
ID Name Status Path           Is independent Creation time   Comment               Timestamp
0  root Linked /ibm/gpfs1     yes            9/29/14 5:06 PM root file set          9/30/14 9:26 AM
```

```
1  test Linked /ibm/gpfs1/test yes           9/29/14 5:11 PM test file set creating 9/30/14 9:26 AM
EFSSG1000I The command completed successfull/11/11 1:02 PM             11/11/11 1:26 PM
```

Creating file sets

Using the GUI: To create a file set, click **Files** → **File Sets.** All configured file sets are listed here, as shown in Figure 7-16 on page 504. To create a file set, click **Create File Set** in the right pane to start the Create File Set wizard. See Figure 7-62.



Junction path: /ibm/gpfs1/Marketing

Name: Marketing

Comment: tes file set

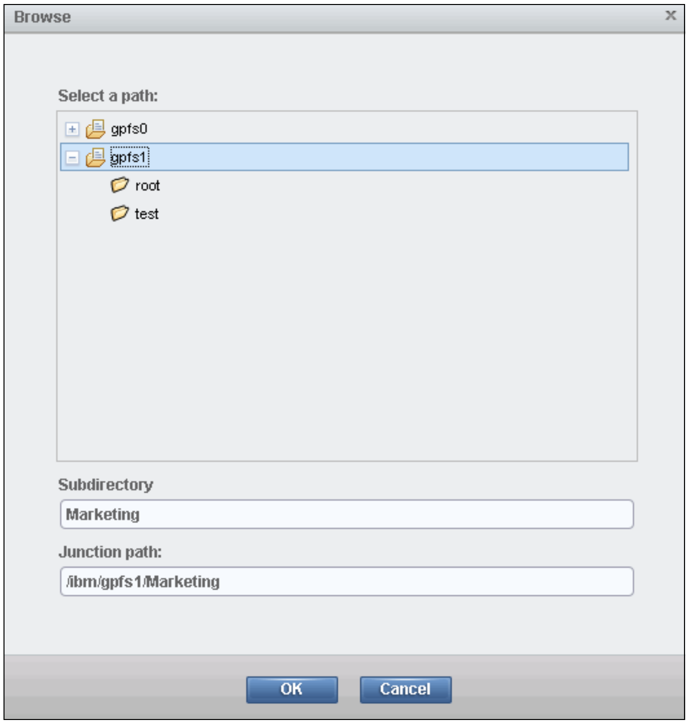
Type: Independent

Maximum number of inodes: 1,000,000

Allocated number of inodes: 50,000

Figure 7-62   New File Set wizard

Create a file set named `Marketing` with custom options. It is on the file system `gpfs1`. To configure the file set name and junction path, click **Browse**, select **gpfs1**, and enter the subdirectory name. Configure the junction path with the same name and a subdirectory name. Click **OK**. See Figure 7-63.



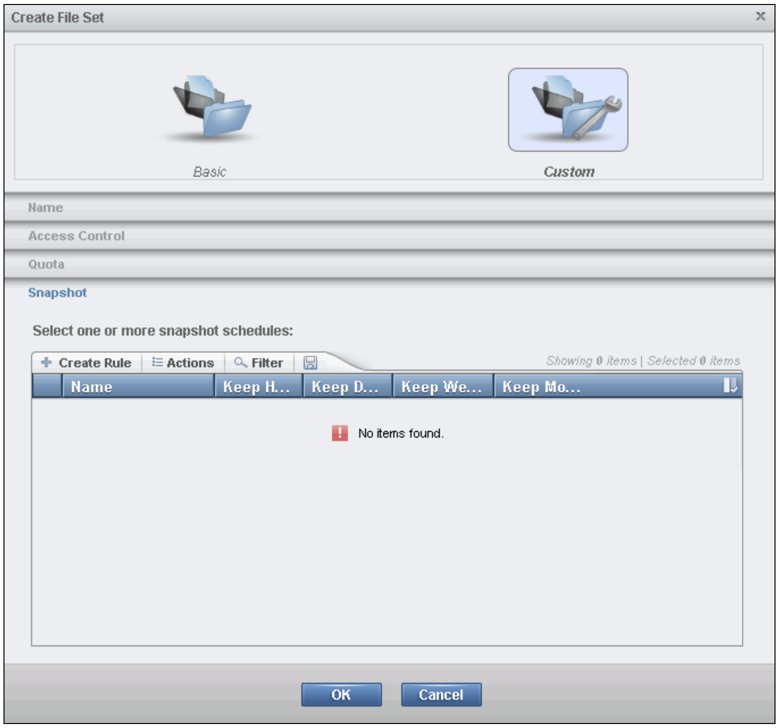
Subdirectory: Marketing

Junction path: /ibm/gpfs1/Marketing

Figure 7-63 Configure the junction path and file set name

Note: You can configure the junction path name and file set name without selecting **Browse**. You can enter the file set name manually. This option means that the file set and junction path can have different names. For example, you can configure the file set name as `Marketing` and junction path name as `/ibm/gpfs1/Marketingteam` for the file set `Marketing`. SONAS allows this configuration. However, in many cases, administrators can be confused by the differences between the file set name and the directory (junction path) names. This confusion might lead to misconfiguration. The preferred practice is to use the same name for the file set and junction path.

With custom options, you can also configure access control, quotas, and snapshots for your new file set. For this file set, create a snapshot rule. To create a snapshot, select the **Snapshot** tab. See Figure 7-64.

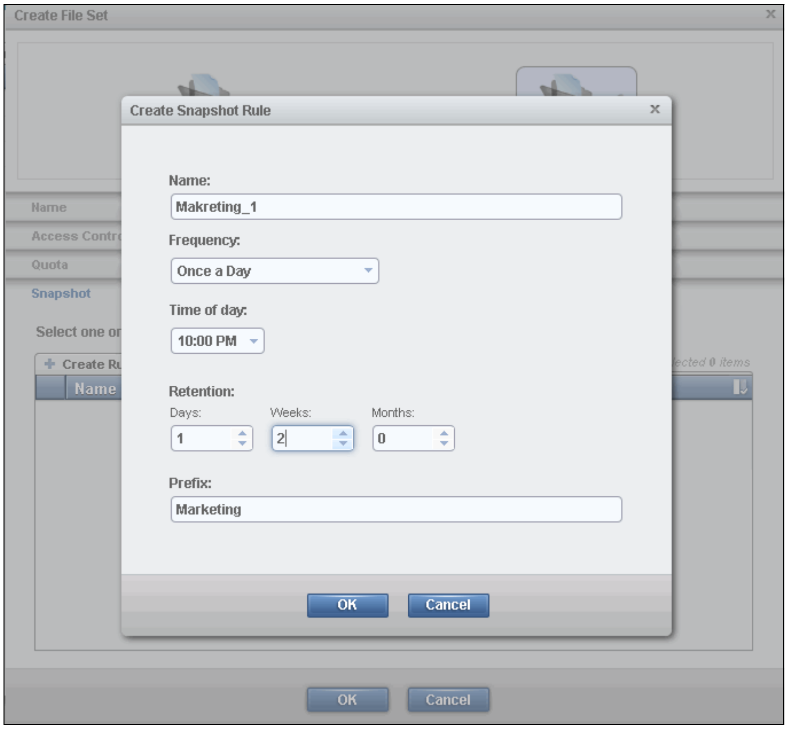


Select one or more snapshot schedules:

No items found.

Figure 7-64 Create a snapshot for the Marketing file set

Select the **Create Rule** option, which opens a snapshot menu, as shown in Figure 7-65. A snapshot is taken every day at 10.00 p.m. The prefix of the snapshot file is Marketing with a retention of 2 weeks and 1 day.



Create Snapshot Rule

Name: Makreting_1

Frequency: Once a Day

Time of day: 10:00 PM

Retention:

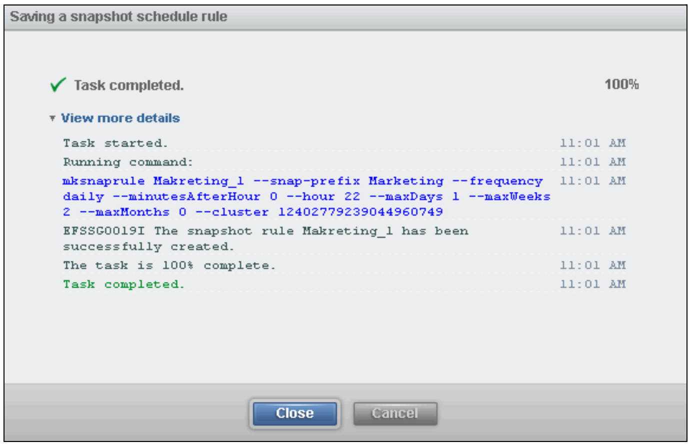
Days: 1 Weeks: 2 Months: 0

Prefix: Marketing

Figure 7-65 Create a snapshot rule

Note: You can leave the Prefix value. The Prefix option enables you to create specific snapshot name for each file set. However, if a prefix is defined on a snapshot, Volume Shadow Copy Service (VSS) is not supported on Windows clients. For more information, see 6.6.9, “VSS snapshot integration” on page 428.

Click **OK** to create a snap rule for this file set. The result is shown in Figure 7-66.



Saving a snapshot schedule rule

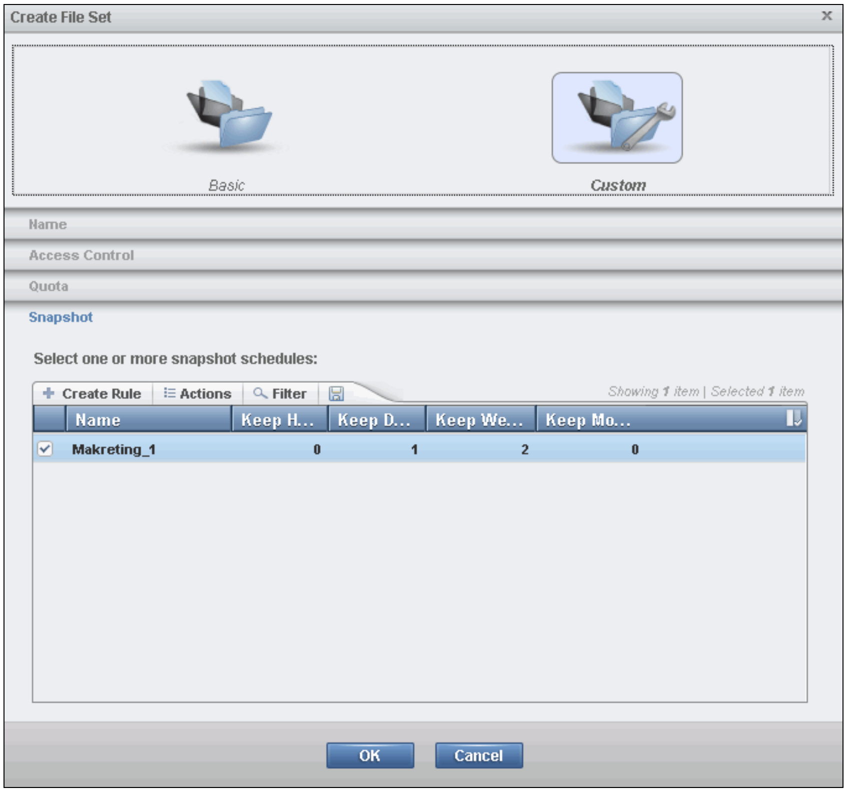
✓ Task completed. 100%

▾ View more details

```
Task started.                                           11:01 AM
Running command:                                        11:01 AM
mksnaprule Makreting_1 --snap-prefix Marketing --frequency  11:01 AM
daily --minutesAfterHour 0 --hour 22 --maxDays 1 --maxWeeks
2 --maxMonths 0 --cluster 12402779239044960749
EFSSG0019I The snapshot rule Makreting_1 has been       11:01 AM
successfully created.
The task is 100% complete.                              11:01 AM
Task completed.                                         11:01 AM
```

Figure 7-66 Save a snapshot schedule rule on Create File set wizard

Click **Close** to return to the Create File Set wizard. The wizard shows a created snapshot rule list for the Marketing file set. See Figure 7-67.



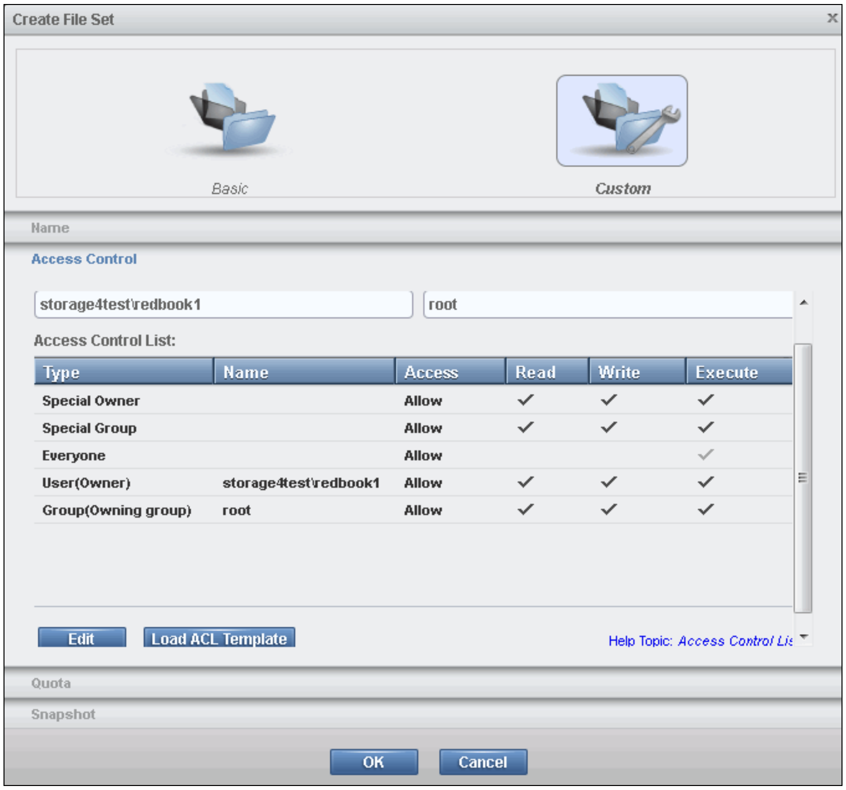
Select one or more snapshot schedules:

Showing 1 item | Selected 1 item

	Name	Keep H...	Keep D...	Keep We...	Keep Mo...
☑	Makreting_1	0	1	2	0

Figure 7-67   Created snap rule list

To define the access control, select **Access Control**. See Figure 7-68. Enter the names in the Owner and Owning group fields. If you want to add or modify the ACL list for this file system, click **Edit**. For more information, see 7.5.6, “Managing access control lists” on page 559.



storage4test\redbook1

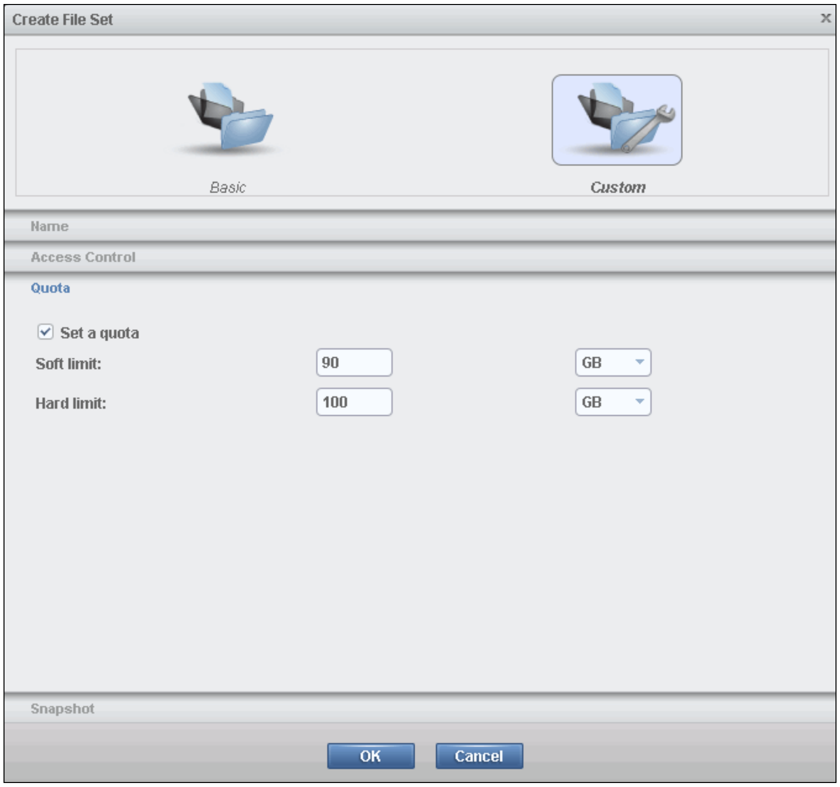
root

Access Control List:

Type	Name	Access	Read	Write	Execute
Special Owner		Allow	✓	✓	✓
Special Group		Allow	✓	✓	✓
Everyone		Allow			✓
User(Owner)	storage4test\redbook1	Allow	✓	✓	✓
Group(Owning group)	root	Allow	✓	✓	✓

Figure 7-68 Configuring Access Control in the Create file set wizard

Select **Quota** to configure a file set quota. Enter the soft and hard limit. See Figure 7-69.



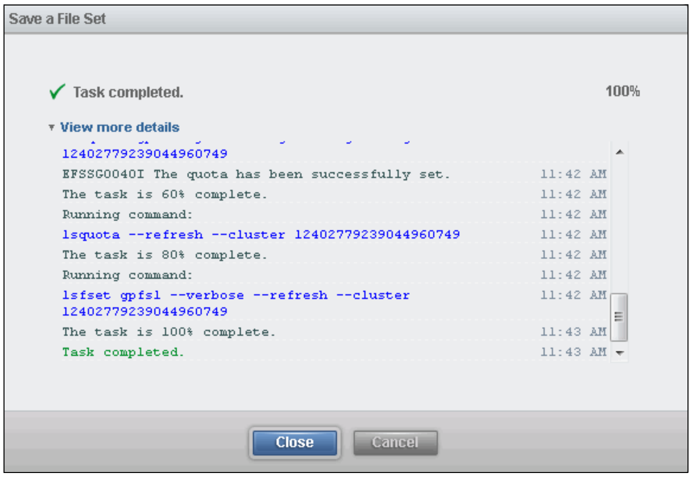
☑ Set a quota

Soft limit: 90 GB

Hard limit: 100 GB

Figure 7-69 Configure Quota with the Create File Set wizard

After you verify all parameters, click **OK**. For this task, three commands are run. The first one is creates a file set, the second one links the file set to the file system, and the third one creates a snapshot. See Figure 7-70.



✓ Task completed. 100%

```
12402779239044960749
EFSSG0040I The quota has been successfully set.        11:42 AM
The task is 60% complete.                              11:42 AM
Running command:                                       11:42 AM
lsquota --refresh --cluster 12402779239044960749       11:42 AM
The task is 80% complete.                              11:42 AM
Running command:                                       11:42 AM
lsfset gpfs1 --verbose --refresh --cluster             11:42 AM
12402779239044960749
The task is 100% complete.                             11:43 AM
Task completed.                                        11:43 AM
```

Figure 7-70 File set creation finished

The newly created file set can now be seen in the File Set menu.

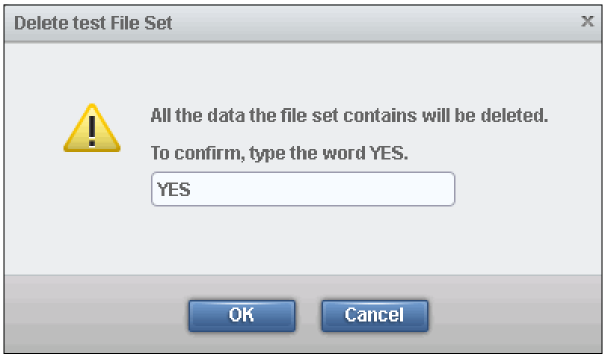
Using the CLI: You can create the file set by running **mkfset**. This command creates a file set that uses the specified name. When you use the CLI to create a file set, you also must run a command to link the new file set. To do so, run **lnkfset**. Example 7-18 shows the help option for the **mkfset**, **linkfset**, and **unlinkfset** commands.

Example 7-18   Help output for mkfset, linkfset, and unlinkfset

```
[furby.storage.tucson.ibm.com]$ mkfset --help
usage: mkfset fileSystem filesetName  [-t <comment>] [--inodeSpaceOwner <ownerFilesetName> | -n] [-i
<maxNumInodes:numInodesToPreallocate>] [--link] [--junction <path>] [--owner [owner][:[group][:template]]] [--chmod | --nochmod]
[--inodeWarningLevel <percentage>] [--inodeErrorLevel <percentage>] [--inodeRecurrences <recurrences>] [-c < clusterID | clusterName >]
Creates a file set.

Parameter               Description
fileSystem              Specifies the device name of the file system to contain the new file set. File system names do not need to be
fully qualified.
filesetName             Specifies the name of the new file set.
-t, --comment           Specifies an optional comment that appears in the output of the "lsfset" command.
    --inodeSpaceOwner   It specifies the independent file set where the new file set is going to be allocated
-n, --independent       Creates an independent file set with its own inode space.
-i                     Specifies the number of inodes to be created initially and the maximum number of inodes to be allowed in this file
set.
    --link              Specifies that the file set is automatically linked at the default junction
    --junction          Specifies that the file set is automatically linked at the given junction. The path must not refer to an existing
file or directory.
    --owner             Sets the directory owner. File set will be automatically linked if this option is specified.
    --chmod             Enables the use of chmod on the file set
    --nochmod           Disables the use of chmod on the file set
    --inodeWarningLevel Sets a custom percentage of inodes to be used before generating a warning message. Specify 0 to use the system
default.
    --inodeErrorLevel   Sets a custom percentage of inodes to be used before generating an error message. Specify 0 to use the system
default.
    --inodeRecurrences  Sets a custom number of recurrences of inode usage messages to log. Specify 0 to use the system default.
-c, --cluster           The cluster scope for this command
[furby.storage.tucson.ibm.com]$ linkfset --help
usage: linkfset fileSystem filesetName  [junctionPath] [-c < clusterID | clusterName >]
Links a file set.

Parameter     Description
fileSystem    Specifies the device name of the file system to contain the new file set. File system names do not need to be fully
qualified.
filesetName   Specifies the name of the file set for identification.
junctionPath  Specifies the name of the junction. The name must not refer to an existing file system object.
-c, --cluster The cluster scope for this command
```

Removing file sets

Using the GUI: To remove a file set, highlight the one that you want and click **Actions** → **Delete**. A new window opens and prompts you for confirmation. Enter YES, as shown in Figure 7-71. Click **OK** to delete the file set.



Delete test File Set

All the data the file set contains will be deleted.

To confirm, type the word YES.

YES

Figure 7-71   Delete file set confirmation window

Using the CLI: You can delete a file set by running **rmfset**. The command asks for confirmation, and then on confirmation, deletes the file set specified. The **rmfset** command fails if the file set is linked into the namespace. By default, the **rmfset** command fails if the file set contains any contents except for an empty root directory. The root file set cannot be deleted.

Example 7-19 shows the command usage and output for deleting a file set. In this example, the file set that is used is “test”, which is created on the “gpfs1” file system.

Example 7-19 Example rmfset command and help output

```
[furby.storage.tucson.ibm.com]$ rmfset --help
usage: rmfset fileSystem filesetName  [--force] [-c < clusterID | clusterName >]
Removes a file set.

Parameter     Description
fileSystem    Specifies the device name of the file system to contain the new file set. File system names do not need to be fully
qualified.
filesetName   Specifies the name of the file set for identification.
    --force   Do not prompt for manual confirmation.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$ rmfset gpfs1 test
EFSSG0388W All data in the file set will be deleted.
EFSSG0621I This operation might take a long time to complete.
Do you really want to perform the operation (yes/no - default no):yes
EFSSG0073I File set test removed successfully.
EFSSG1000I The command completed successfully.
```

7.5 Creating and managing shares

Data that is stored in the directories, file sets, and file system can be accessed by using data access protocols such as CIFS, NFS, FTP, HTTPS, and SCP. For this purpose, you must configure the services and also create shares or exports on the GPFS file system with which you can then access the data by using any of the mentioned protocols.

Services are configured during the installation and configuration of the SONAS. After the services are configured, it is possible for you to share your data with any of the protocols by creating exports with the command-line option or the GUI.

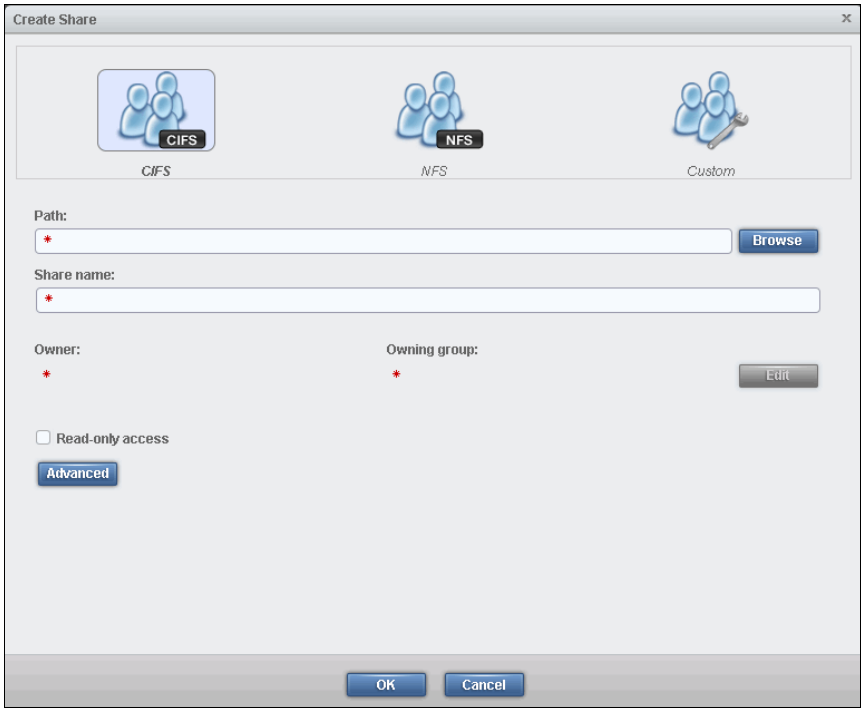
You can add more protocols. You can also remove protocols from the export if you do not want to export the data with a service or protocol. You can also activate and deactivate the export. Finally, you can delete the existing export.

To view and manage all the exports in SONAS, click **Files** → **Shares** in the SONAS GUI. A table that lists all existing exports is shown. See Figure 7-15 on page 504.

7.5.1 Creating shares

This section describes how to create shares.

Using the GUI: To create a share, click **Files** → **Share**. There, select **Create Share**, which starts the New Share wizard, as shown in Figure 7-72 on page 549. Here, you can select between CIFS, NFS, and custom share options. HTTP, FTP, and SCP are under the custom view. For the purposes of this book, you create one CIFS, one NFS share, and one FTP share.



Path: *

Share name: *

Owner: *

Owning group: *

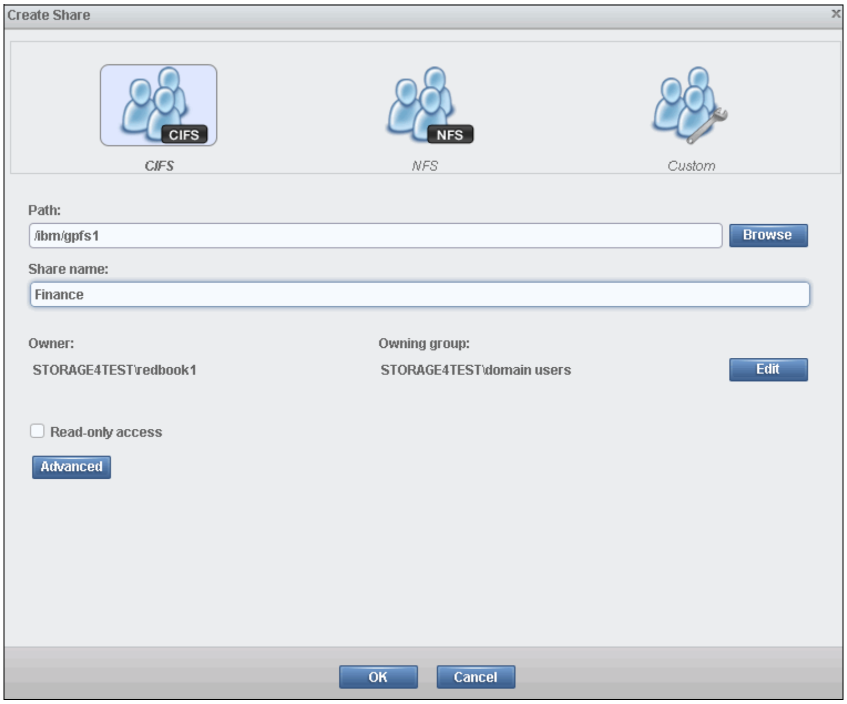
□ Read-only access

Figure 7-72   Create Share wizard

Creating a CIFS share

To create a CIFS share, you can select the **CIFS** option directly or go through the Custom menu. Using the CIFS menu, create a share that is named Finance.

To create a CIFS share, you must enter a path and share name. In this case, the path is `/ibm/gpfs1` and the share name is Finance, as shown in Figure 7-73.



Path: /ibm/gpfs1

Share name: Finance

Owner: STORAGE4TEST\redbook1

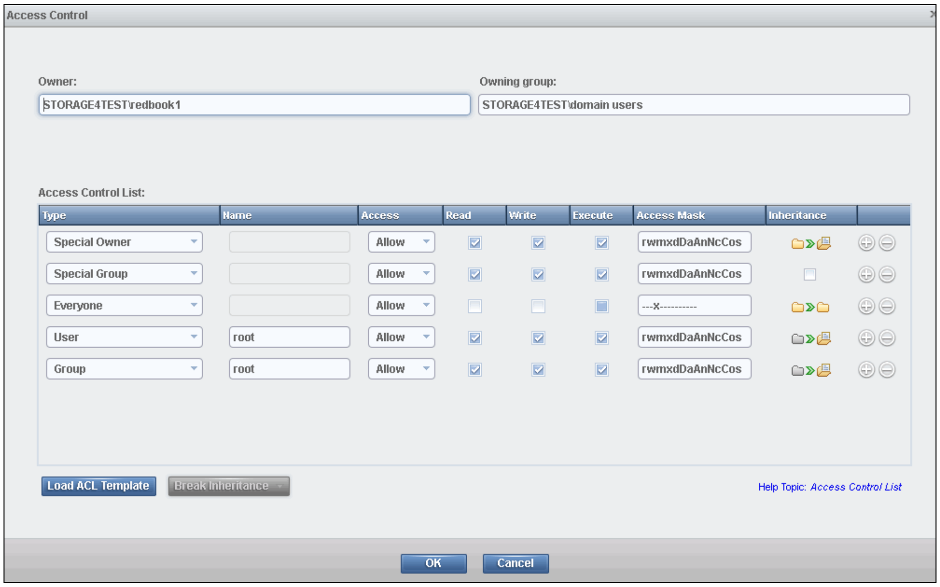
Owning group: STORAGE4TEST\domain users

□ Read-only access

Figure 7-73   Create a CIFS share named Finance

You can also configure the Access Control and the advance CIFS setting by creating an ACL list for user, group, or SID. However, it is not used in this section. For more information about Access Control, see 7.5.6, “Managing access control lists” on page 559. The advanced option menu for CIFS shares is to define default access authority (see Figure 7-75).

To configure an access control list (ACL), click **Edit**. See Figure 7-74.



Access Control

Owner: STORAGE4TEST\redbook1

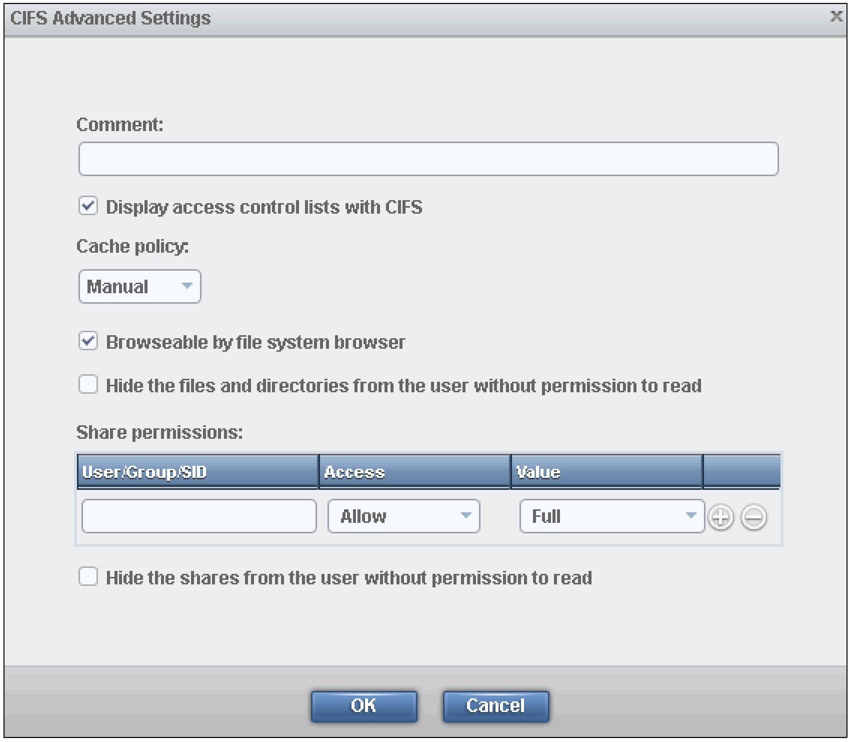
Owning group: STORAGE4TEST\domain users

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance
Special Owner		Allow	☑	☑	☑	rwmxdDaAnNcCos	
Special Group		Allow	☑	☑	☑	rwmxdDaAnNcCos	□
Everyone		Allow	□	□	■	---x----------	
User	root	Allow	☑	☑	☑	rwmxdDaAnNcCos	
Group	root	Allow	☑	☑	☑	rwmxdDaAnNcCos	

Figure 7-74 Configuring access control with the Create Share wizard

Figure 7-75 shows the CIFS Advanced Setting menu.



CIFS Advanced Settings

Comment:

☑ Display access control lists with CIFS

Cache policy: Manual

☑ Browseable by file system browser

□ Hide the files and directories from the user without permission to read

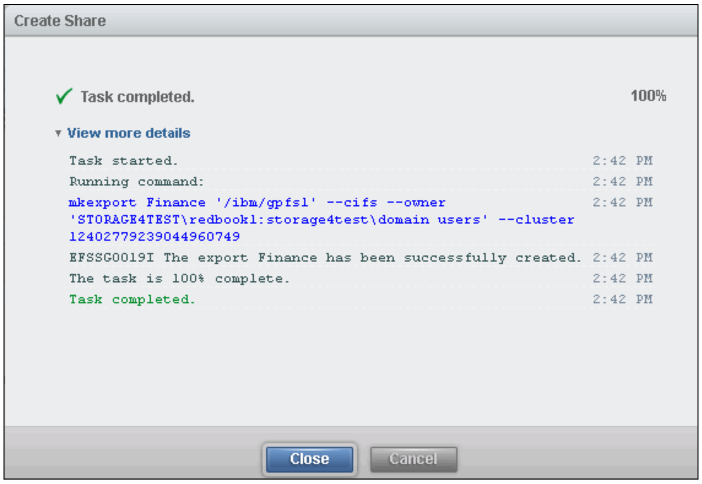
Share permissions:

User/Group/SID	Access	Value
	Allow	Full

□ Hide the shares from the user without permission to read

Figure 7-75 CIFS Advanced Setting menu

When all information is entered, click **OK**. You see a confirmation window that the share was created. See Figure 7-76.



✓ Task completed. 100%

▾ View more details

```
Task started.                                              2:42 PM
Running command:                                           2:42 PM
mkexport Finance '/ibm/gpfsl' --cifs --owner               2:42 PM
'STORAGE4TEST\redbookl:storage4test\domain users' --cluster
12402779239044960749
EFSSG0019I The export Finance has been successfully created. 2:42 PM
The task is 100% complete.                                 2:42 PM
Task completed.                                            2:42 PM
```

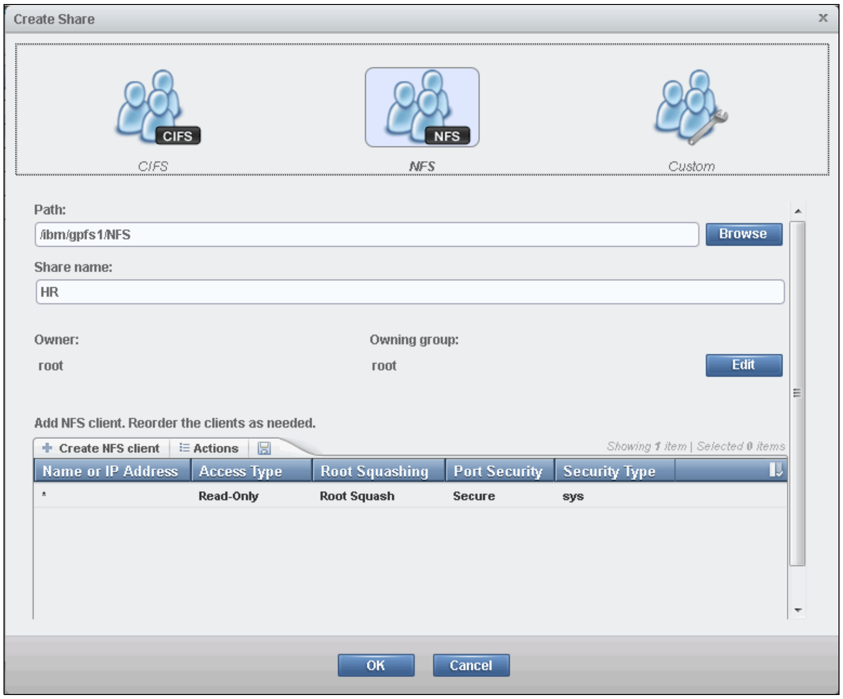
Figure 7-76 Finish the creation of a share named Finance

Creating an NFS share

To create an NFS share, you can select the NFS option or go through the Custom menu. In this scenario, the NFS menu is used to create a share that is named HR.

To create an NFS share, you need to enter the path name and share name. In this scenario, the Share name is HR and the path is `/ibm/gpfs1`, as shown in Figure 7-77.

To create an NFS client and permission, select **Create NFS Client** in the Add NFS client section. You must add at least one client. When all the information is entered, click **OK** to continue.



Path: /ibm/gpfs1/NFS

Share name: HR

Owner: root

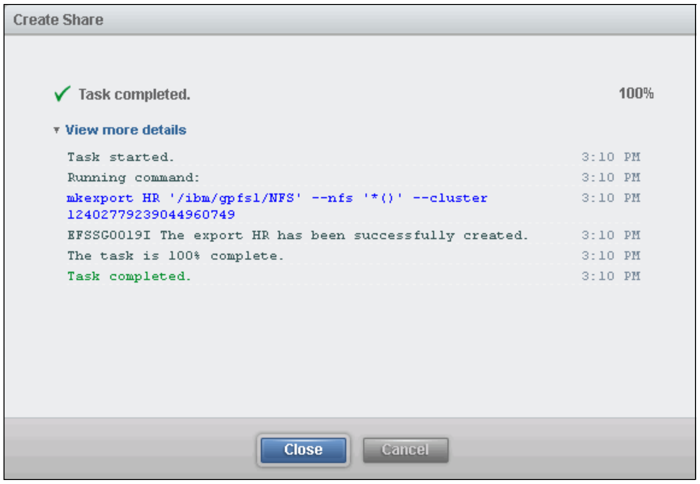
Owning group: root

Add NFS client. Reorder the clients as needed.

Name or IP Address	Access Type	Root Squashing	Port Security	Security Type
*	Read-Only	Root Squash	Secure	sys

Figure 7-77 Create the NFS share wizard

When the system finishes creating the share, a confirmation window opens, as shown in Figure 7-78.



✓ Task completed. 100%

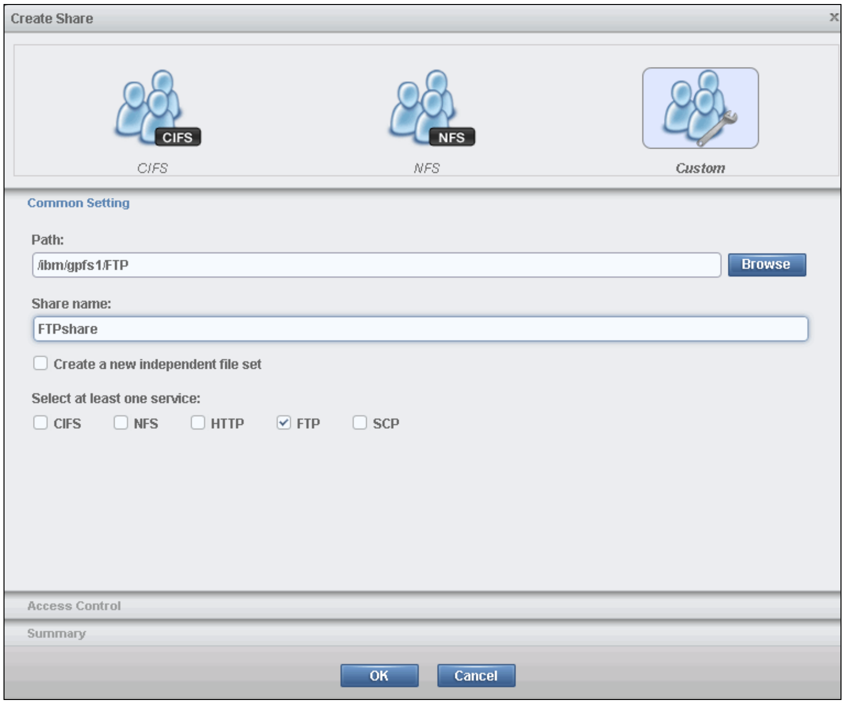
▾ View more details

```
Task started.                                              3:10 PM
Running command:                                           3:10 PM
mkexport HR '/ibm/gpfs1/NFS' --nfs '*()' --cluster         3:10 PM
12402779239044960749
EFSSG0019I The export HR has been successfully created.    3:10 PM
The task is 100% complete.                                 3:10 PM
Task completed.                                            3:10 PM
```

Figure 7-78 NFS share creation is finished

Creating an FTP share

To create an FTP share, select the **Custom** option in the Create Share wizard. You must enter the path and share name. Creating an FTP share is shown in Figure 7-79.



Common Setting

Path: /ibm/gpfs1/FTP

Share name: FTPshare

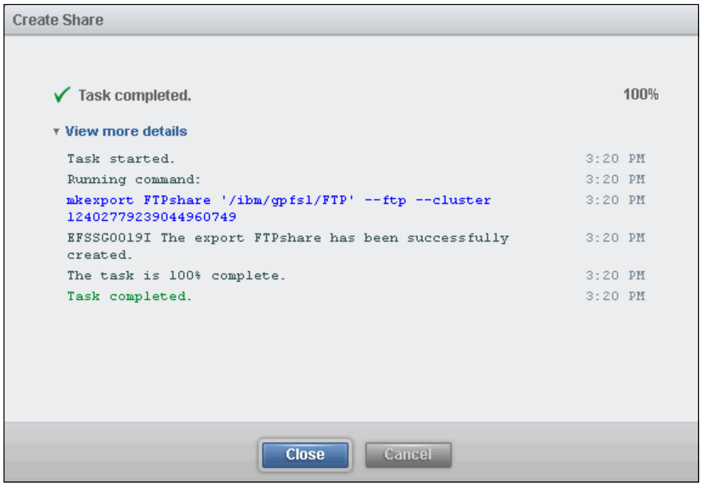
□ Create a new independent file set

Select at least one service:

□ CIFS □ NFS □ HTTP ☑ FTP □ SCP

Figure 7-79 Create an FTP share

After all the information is entered, click **OK** to continue. When the system finishes creating a new share, a confirmation window opens, as shown in Figure 7-80.


Create Share

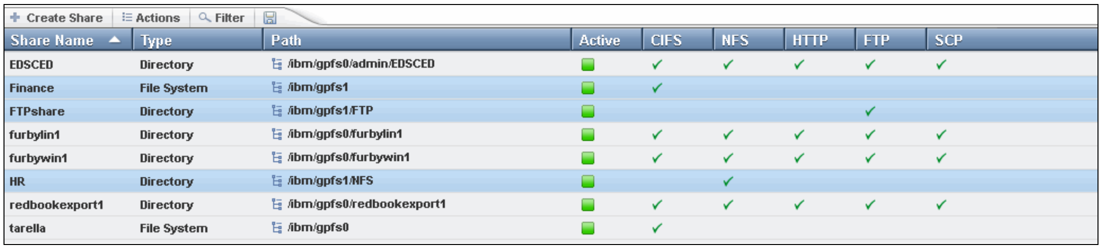
✓ Task completed. 100%

▾ View more details

```
Task started.                                               3:20 PM
Running command:                                            3:20 PM
mkexport FTPshare '/ibm/gpfs1/FTP' --ftp --cluster          3:20 PM
12402779239044960749
EFSSG0019I The export FTPshare has been successfully        3:20 PM
created.
The task is 100% complete.                                  3:20 PM
Task completed.                                             3:20 PM
```

Figure 7-80   FTP share creation is finished

The main window for shares now shows all the newly created shares. See Figure 7-81.



Share Name	Type	Path	Active	CIFS	NFS	HTTP	FTP	SCP
EDSCED	Directory	/ibm/gpfs0/admin/EDSCED	🟩	✓	✓	✓	✓	✓
Finance	File System	/ibm/gpfs1	🟩	✓				
FTPshare	Directory	/ibm/gpfs1/FTP	🟩				✓	
furbylin1	Directory	/ibm/gpfs0/furbylin1	🟩	✓	✓	✓	✓	✓
furbywin1	Directory	/ibm/gpfs0/furbywin1	🟩	✓	✓	✓	✓	✓
HR	Directory	/ibm/gpfs1/NFS	🟩		✓			
redbookexport1	Directory	/ibm/gpfs0/redbookexport1	🟩	✓	✓	✓	✓	✓
tarella	File System	/ibm/gpfs0	🟩	✓				

Figure 7-81   Shares view under the Shares menu

Creating an export by using the CLI

You can create an export by running **`mkexport`**. This command takes the name of the sharename and the directory path of the share you want to create. You can create an FTP, CIFS, and NFS share with this command. Using the command, you can also create an *inactive* share. Inactive shares exist when the creation of the share is complete, but the share cannot be used by the users. By default, the share is active. You can also add “owner,” which gives the required ACLs to the user to access the share.

The command usage and output is shown in Example 7-20. In this example, FTP and CIFS shares are created.

Example 7-20   Example mkexport command output

```
[furby.storage.tucson.ibm.com]$ mkexport --help
usage: mkexport exportName path  {--http | --ftp | --scp | --nfs <clientDefs> | --cifs <cifsOptions>} [--inactive] [--owner
[owner][:[group][:template]] | --reference <reference>] [-c < clusterID | clusterName >]
Creates a share by using one or more protocols.

Parameter       Description
exportName      Specifies the name of the newly created export. The name is limited to 80 characters.
path            Specifies the path for the share.
    --http      Configures the HTTP protocol for this share.
    --ftp       Configures the FTP protocol for this share.
    --scp       Configures the SCP protocol for this share.
    --nfs       Configures the NFS protocol for this share and defines the clients along with their options.
    --cifs      Defines options for the CIFS protocol.
    --inactive  Marks the created share as inactive.
    --owner     Sets the directory owner.
    --reference Sets the directory owner to the same owner the reference has.
-c, --cluster   The cluster scope for this command
```

You can also create an inactive share by **`mkexport --inactive`**.

7.5.2 Listing and viewing the status of the created exports

Using the GUI: You can view the exports that were created by clicking **Files** → **Shares**. See Figure 7-81 on page 553.

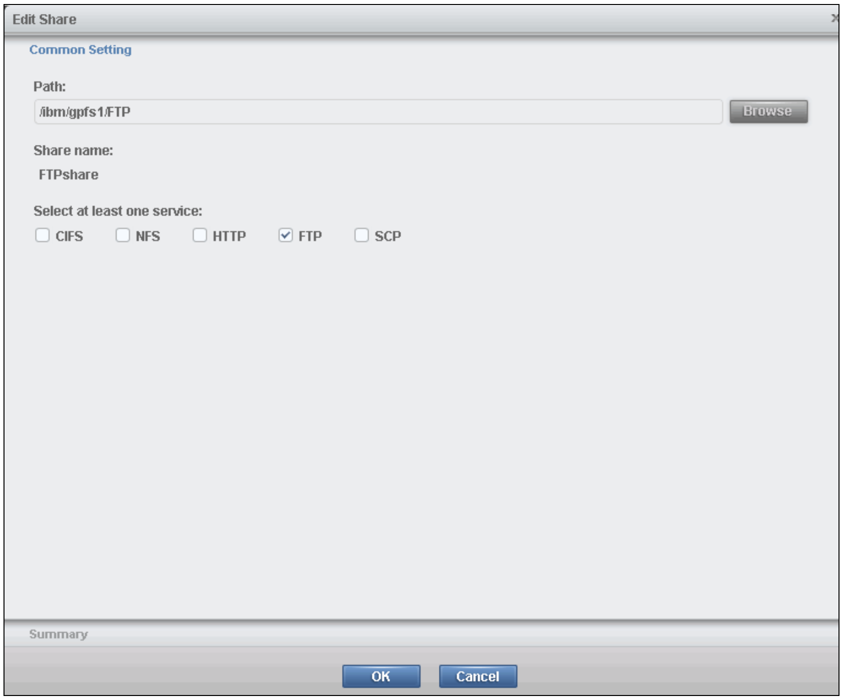
Using the CLI: You can list the exports or shares by running **`lsexport`**. This command lists all the exports as a list for each protocol for which they are created. Example 7-21 shows the command usage and output.

Example 7-21 Command usage and output for listing exports by running lsexport

```
[furby.storage.tucson.ibm.com]$ lsexport --help
usage: lsexport  [-v] [-r] [-Y] [--nfsDefs <nfsShareName>] [-c < clusterID | clusterName >]
Lists exports.

Parameter     Description
-v, --verbose Shows additional columns.
-r, --refresh Forces a refresh of the exports data in the database by scanning all the cluster exports before retrieving data from the
database.
-Y            Shows parseable output.
    --nfsDefs Specifies the name of the NFS share.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$ lsexport
Name           Path                        Protocol Active Timestamp
EDSCED         /ibm/gpfs0/admin/EDSCED   FTP      true   9/30/14 3:27 PM
FTPshare       /ibm/gpfs1/FTP            FTP      true   9/30/14 3:27 PM
furbylin1      /ibm/gpfs0/furbylin1      FTP      true   9/30/14 3:27 PM
furbywin1      /ibm/gpfs0/furbywin1      FTP      true   9/30/14 3:27 PM
redbookexport1 /ibm/gpfs0/redbookexport1 FTP      true   9/30/14 3:27 PM
EDSCED         /ibm/gpfs0/admin/EDSCED   HTTP     true   9/30/14 2:34 PM
furbylin1      /ibm/gpfs0/furbylin1      HTTP     true   9/30/14 2:34 PM
furbywin1      /ibm/gpfs0/furbywin1      HTTP     true   9/30/14 2:34 PM
redbookexport1 /ibm/gpfs0/redbookexport1 HTTP     true   9/30/14 2:34 PM
EDSCED         /ibm/gpfs0/admin/EDSCED   NFS      true   9/30/14 2:41 PM
furbylin1      /ibm/gpfs0/furbylin1      NFS      true   9/30/14 2:41 PM
furbywin1      /ibm/gpfs0/furbywin1      NFS      true   9/30/14 2:41 PM
HR             /ibm/gpfs1/NFS            NFS      true   9/30/14 3:10 PM
redbookexport1 /ibm/gpfs0/redbookexport1 NFS      true   9/30/14 2:41 PM
EDSCED         /ibm/gpfs0/admin/EDSCED   CIFS     true   9/30/14 2:48 PM
Finance        /ibm/gpfs1                CIFS     true   9/30/14 2:48 PM
furbylin1      /ibm/gpfs0/furbylin1      CIFS     true   9/30/14 2:48 PM
furbywin1      /ibm/gpfs0/furbywin1      CIFS     true   9/30/14 2:48 PM
redbookexport1 /ibm/gpfs0/redbookexport1 CIFS     true   9/30/14 2:48 PM
tarella        /ibm/gpfs0                CIFS     true   9/30/14 2:48 PM
EDSCED         /ibm/gpfs0/admin/EDSCED   SCP      true   9/30/14 2:55 PM
furbylin1      /ibm/gpfs0/furbylin1      SCP      true   9/30/14 2:55 PM
furbywin1      /ibm/gpfs0/furbywin1      SCP      true   9/30/14 2:55 PM
redbookexport1 /ibm/gpfs0/redbookexport1 SCP      true   9/30/14 2:55 PM
EFSSG1000I The command completed successfully.
```

7.5.3 Modifying exports

Using the GUI: To modify a share in the GUI, select a share from share list and click **Actions** → **Edit**. The Edit Share wizard starts, as shown in Figure 7-82 on page 555.



Path:

/ibm/gpfs1/FTP

Share name:

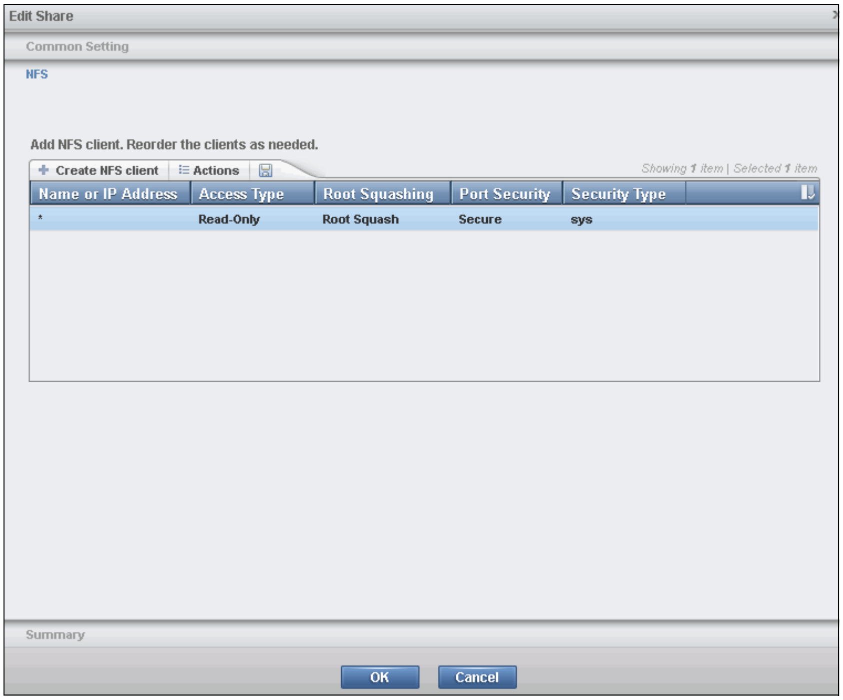
FTPshare

Select at least one service:

□ CIFS □ NFS □ HTTP ☑ FTP □ SCP

Figure 7-82   Edit Share wizard

You can modify, add new, or remove unused protocols. In this case, you want to add the HTTP and NFS protocols to the `cli_example` share. To add them, select **FTP** and **HTTP**. Click the **NFS** tab and select the **Enable NFS** check box, as shown in Figure 7-83. To remove a protocol from the share, clear the check box in front of the protocol name.



Add NFS client. Reorder the clients as needed.

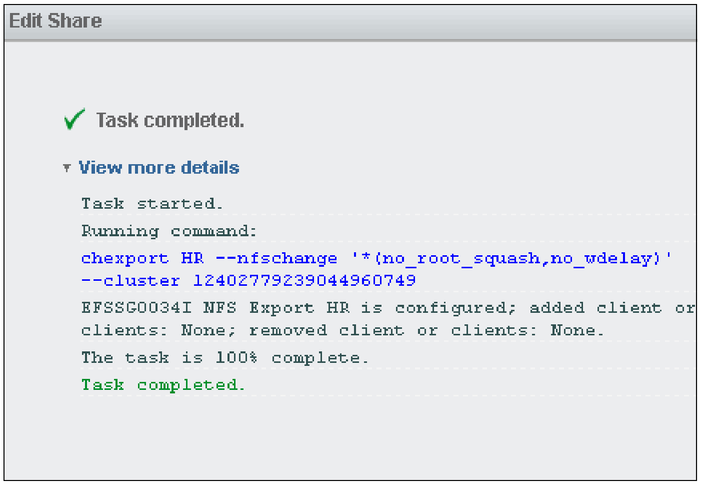
Showing 1 item | Selected 1 item

Name or IP Address	Access Type	Root Squashing	Port Security	Security Type
*	Read-Only	Root Squash	Secure	sys

Figure 7-83   Add the NFS protocol to the share

For NFS shares, you can also add or remove client names or IDs. Click the **NFS** tab and then select **Create NFS client** to add a client, or click **Actions** → **Delete** to remove the client. You can edit current client properties by clicking **Actions** → → **Edit**.

After all modifications are done, click **OK**. The system processes these changes and opens the window that is shown in Figure 7-84.



✓ Task completed.

▾ View more details

```
Task started.
Running command:
chexport HR --nfschange '*(no_root_squash,no_wdelay)'
--cluster 12402779239044960749
EFSSG0034I NFS Export HR is configured; added client or
clients: None; removed client or clients: None.
The task is 100% complete.
Task completed.
```

Figure 7-84 The share is successfully edited

Using the CLI: You can modify an existing share or export by running **`chexport`**. Unlike the GUI, which requires separate steps, you can remove or add protocols by using a single command. Each method has different options. This section explains how to add new protocols.

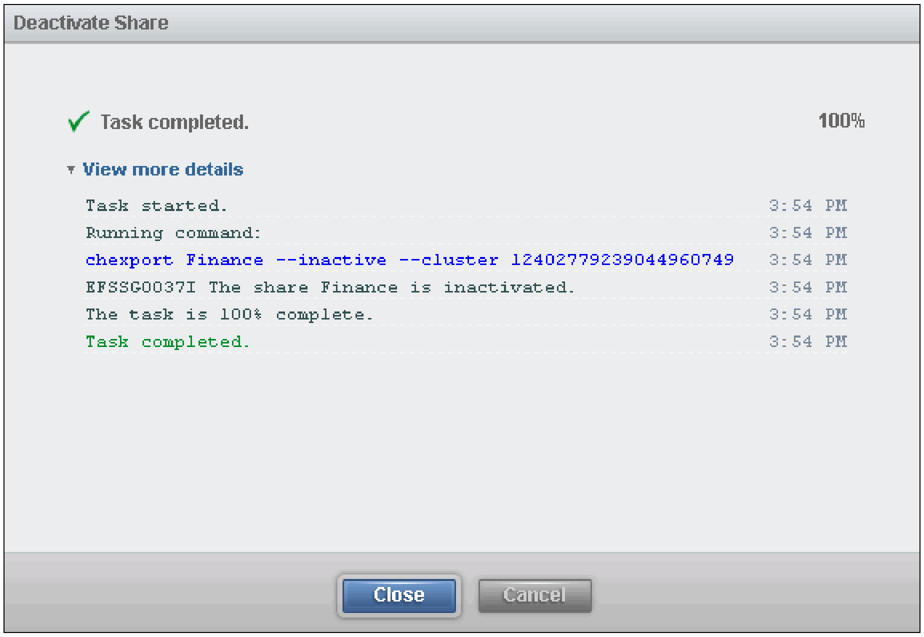
You can add new protocols by adding the **`--cifs`**, **`--ftp`**, and **`--nfs`** options and the protocol definitions. The command usage and output are shown in Example 7-22. For this example, the existing export is a CIFS export. FTP and NFS are added by running **`chexport`**.

Example 7-22 Command usage and output for adding new protocols to an existing share

```
[furby.storage.tucson.ibm.com]$ chexport --help
usage: chexport exportName  [--active | --inactive] [--nfsadd <clientdefs> | --nfschange <clientdefs>] [--nfsPosition <nfsPosition>]
[--nfsremove <clientdefs>] [--http | --httpoff] [--ftp | --ftpoff] [--scp | --scpoff] [--nfs <clientDefs> | --nfsoff] [--cifs
<cifsOptions> | --cifsoff] [--force] [-c < clusterID | clusterName >]
Modifies the protocols and their settings of an existing share.

Parameter         Description
exportName        Specifies the name of the share.
    --active      Sets share to an active state.
    --inactive    Sets share to an inactive state.
    --nfsadd      Adds NFS clients.
    --nfschange   Changes NFS clients.
    --nfsPosition Specifies the absolute position or client name where the entry is to be set.
    --nfsremove   Removes NFS clients.
    --http        Adds the HTTP protocol to the share identified by the exportName argument. Use the --httpoff option to remove the HTTP
protocol from the share.
    --httpoff     Removes the HTTP protocol from the share that is identified by the exportName argument.
    --ftp         Adds the FTP protocol to the share identified by the exportName argument. Use the --ftpoff option to remove FTP protocol
from the share.
    --ftpoff      Removes the FTP protocol from the share that is identified by the exportName argument.
    --scp         Adds the SCP protocol to the share identified by the exportName argument. Use the --scpoff option to remove SCP protocol
from the share.
    --scpoff      Removes the SCP protocol from the share that is identified by the exportName argument.
    --nfs         Configures the NFS protocol for this share and defines the clients along with their options. Use the --nfsoff option to
remove NFS protocol from the share.
    --nfsoff      Removes the NFS protocol from the share that is identified by the exportName argument.
    --cifs        Configures the CIFS protocol for this share and defines the corresponding options. Use the --cifsoff option to remove
CIFS protocol from the share.
    --cifsoff     Removes the CIFS protocol from the share that is identified by the exportName argument.
    --force       enforce operation without calling back the user
-c, --cluster     The cluster scope for this command
```

7.5.4 Deactivating exports

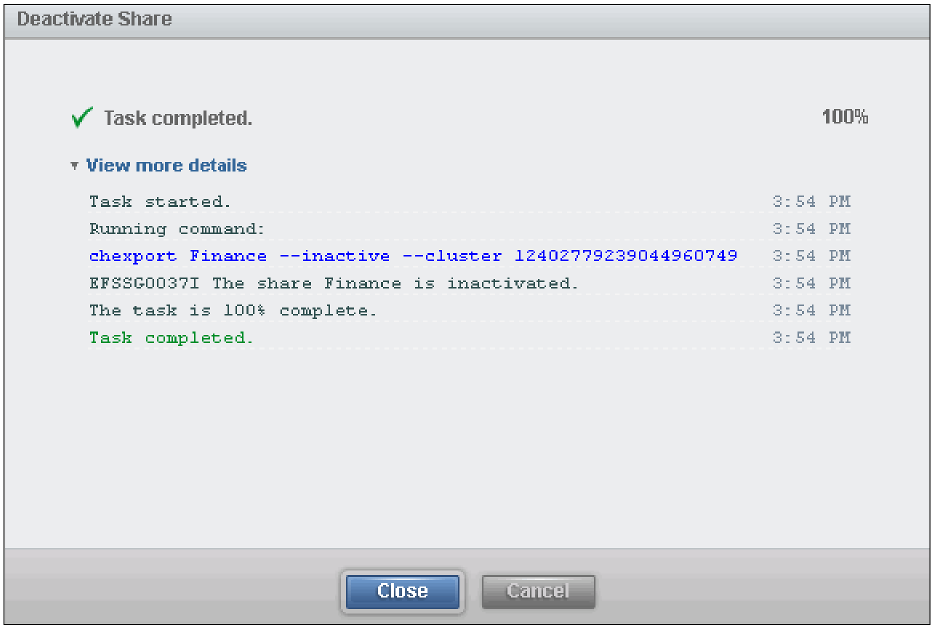
Using the GUI: You can deactivate an existing export that is active by clearing the check box that represents the share status. Select the share that you want from the **Actions** menu and clear the **Active** check box. See Figure 7-85. A confirmation message window opens when you clear this box. Click **OK** to confirm deactivating the share.



```
Task started.                                                     3:54 PM
Running command:                                                  3:54 PM
chexport Finance --inactive --cluster 12402779239044960749        3:54 PM
EFSSG0037I The share Finance is inactivated.                      3:54 PM
The task is 100% complete.                                        3:54 PM
Task completed.                                                   3:54 PM
```

Figure 7-85 Edit Share window with Active share option

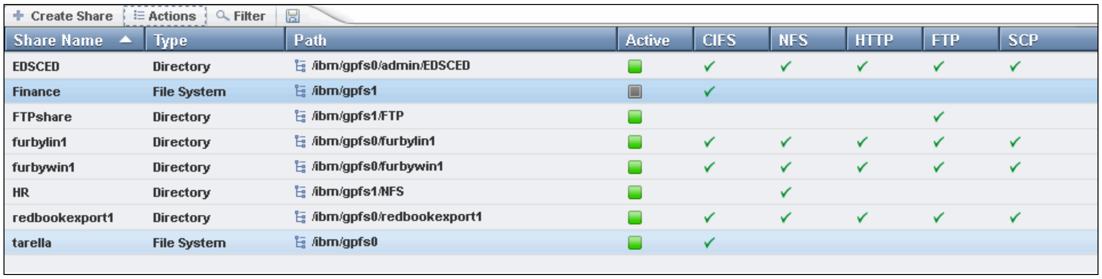
This change opens the window that is shown in Figure 7-86.



```
Task started.                                                     3:54 PM
Running command:                                                  3:54 PM
chexport Finance --inactive --cluster 12402779239044960749        3:54 PM
EFSSG0037I The share Finance is inactivated.                      3:54 PM
The task is 100% complete.                                        3:54 PM
Task completed.                                                   3:54 PM
```

Figure 7-86 Deactivate share result message

In Figure 7-87, the share that is named Finance is inactive. To set it as active again, select the share that you want. Click the **Actions** menu, and then select the **Active** check box. The share is activated without a confirmation process.



Share Name	Type	Path	Active	CIFS	NFS	HTTP	FTP	SCP
EDSCED	Directory	/ibm/gpfs0/admin/EDSCED	🟩	✓	✓	✓	✓	✓
Finance	File System	/ibm/gpfs1	▨	✓				
FTPshare	Directory	/ibm/gpfs1/FTP	🟩				✓	
furbylin1	Directory	/ibm/gpfs0/furbylin1	🟩	✓	✓	✓	✓	✓
furbywin1	Directory	/ibm/gpfs0/furbywin1	🟩	✓	✓	✓	✓	✓
HR	Directory	/ibm/gpfs1/NFS	🟩		✓			
redbookexport1	Directory	/ibm/gpfs0/redbookexport1	🟩	✓	✓	✓	✓	✓
tarella	File System	/ibm/gpfs0	🟩	✓				

Figure 7-87 Share status in the GUI

Using the CLI: You can activate a share by running **`chexport --active`**. You can use the **`--inactive`** option to deactivate a share. The command usage and output are shown in Example 7-23.

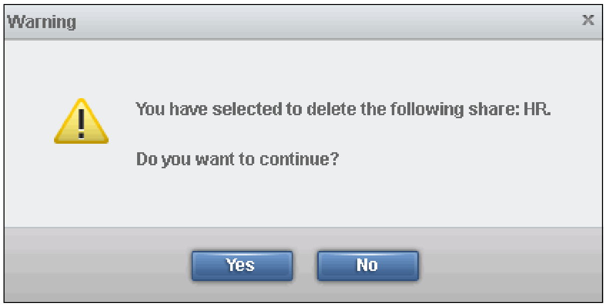
> **Exports:** You can also create an export that is inactive by using the **`--inactive`** option with the **`mkexport`** command. For more information, see 7.5.1, “Creating shares” on page 548.

Example 7-23 Command usage and output to deactivate and activate an existing share

```
[furby.storage.tucson.ibm.com]$ chexport Finance --inactive
Do you really want to perform the operation (yes/no - default no):yes
EFSSG0037I The share Finance is inactivated.
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ chexport Finance --active
EFSSG0036I The share Finance is activated.
EFSSG1000I The command completed successfully.
```

7.5.5 Removing exports

Using the GUI: To remove a share, select a share, and from the **Actions** menu, select **Delete**. A confirmation window opens, as shown in Figure 7-88.



You have selected to delete the following share: HR.

Do you want to continue?

Figure 7-88 Confirmation window before deleting a share

Using the CLI: You can remove an existing export by running **`rmexport`**, which removes all the configuration details of the export from all nodes. See the command usage and output in Example 7-24 on page 559.

Example 7-24   Command usage and output to remove an existing export

```
[furby.storage.tucson.ibm.com]$ rmexport --help
usage: rmexport shareName  [--force] [-c < clusterID | clusterName >]
Removes the given share.

Parameter     Description
shareName     Specifies the name of the share for identification.
    --force   Do not prompt for manual confirmation.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$ rmexport HR
Do you really want to perform the operation (yes/no - default no):yes
EFSSG0021I The export HR has been successfully removed.
EFSSG1000I The command completed successfully.
```

7.5.6 Managing access control lists

An access control list (ACL) is a list of permissions that is associated with a resource. An access control entry (ACE) is an individual entry in an access control list, and describes the permissions for an individual user or group. An ACL usually consists of multiple ACEs. An ACL describes which identities are allowed to access a particular resource. ACLs are the built-in access control mechanism of UNIX and Windows operating systems. The IBM SONAS system uses the built-in Linux ACL mechanism for access control to files that are stored on the IBM SONAS system. Types of access include read, write, and execute. A file that can be accessed by a user is within an IBM SONAS file system that is created by using the General Parallel File System (GPFS).

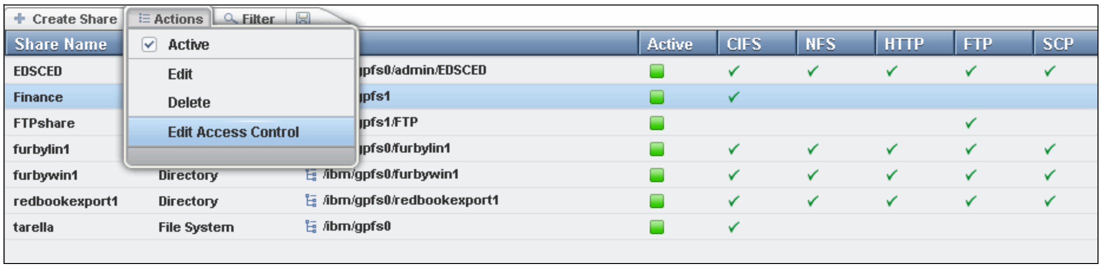
SONAS provides ACL and ACE configuration for file systems, file sets, and shares. This section describes how to configure ACLs with the GUI.

> **Note:** An ACL management function is provided in the SONASV 1.5.1 release. The previous versions of SONAS do not provide ACL management functions. To manage ACLs for versions before Version 1.5, you must create a CIFS export so you can edit ACLs with a Windows client.
>
> **Note:** To edit an ACL that has existing files, the user must have the Data Access role. You must create a user group with the Data Access role, then add the user group to the user. For more information, see 7.7, “User management” on page 575.

Editing access control

You can edit the access control of the File systems, File Sets, and Shares sections. The same features are available for each section. In this example, an ACL is edited in the Share section.

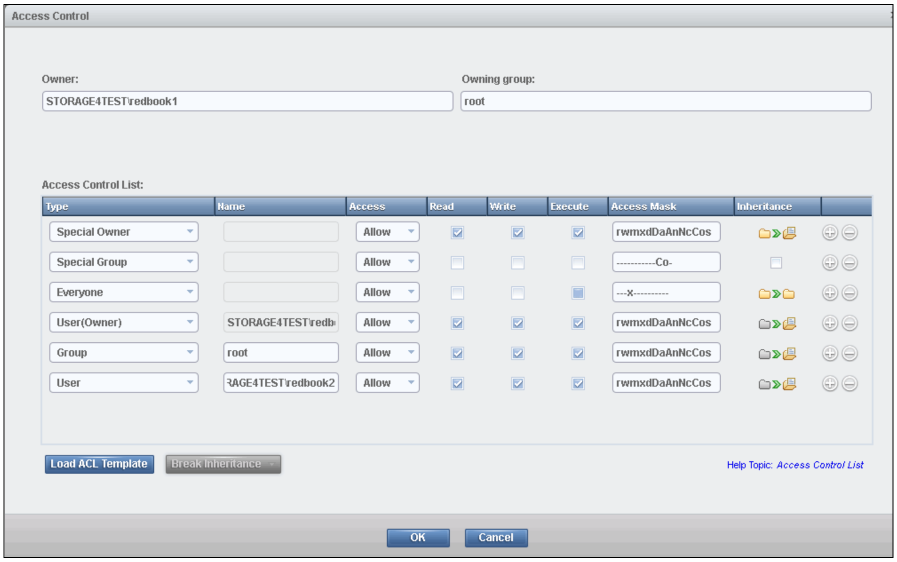
To edit the access control of a share, click **Files** → **Shares** and select the share that you want. Click **Actions** → **Edit Access Control** to edit the ACL. See Figure 7-89.



Share Name			Active	CIFS	NFS	HTTP	FTP	SCP
EDSCED		pfs0/admin/EDSCED	🟩	✓	✓	✓	✓	✓
Finance		pfs1	🟩	✓				
FTPshare		pfs1/FTP	🟩				✓	
furbylin1		pfs0/furbylin1	🟩	✓	✓	✓	✓	✓
furbywin1	Directory	/ibm/gpfs0/furbywin1	🟩	✓	✓	✓	✓	✓
redbookexport1	Directory	/ibm/gpfs0/redbookexport1	🟩	✓	✓	✓	✓	✓
tarella	File System	/ibm/gpfs0	🟩	✓				

Figure 7-89   Edit Access Control menu

You can edit the owner and the group for this share. You can add and modify the ACE and ACLs. See Figure 7-90.



Access Control

Owner: STORAGE4TEST\redbook1

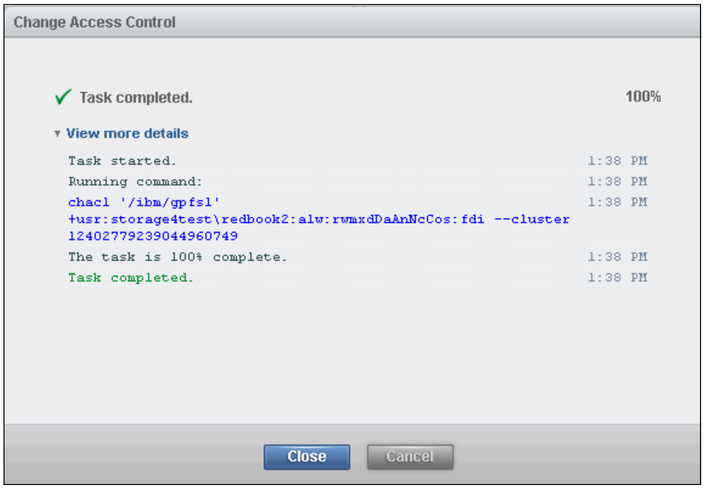
Owning group: root

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance	
Special Owner		Allow	☑	☑	☑	rwmxdDaAnNcCos		⊕ ⊖
Special Group		Allow	□	□	□	----------Co-	□	⊕ ⊖
Everyone		Allow	□	□	■	---x----------		⊕ ⊖
User(Owner)	STORAGE4TEST\redb	Allow	☑	☑	☑	rwmxdDaAnNcCos		⊕ ⊖
Group	root	Allow	☑	☑	☑	rwmxdDaAnNcCos		⊕ ⊖
User	RAGE4TEST\redbook2	Allow	☑	☑	☑	rwmxdDaAnNcCos		⊕ ⊖

Figure 7-90 Access Control window

You can see the current ACL setting for the *Finance* share. Add one user whose name is *redbook2* on the *storage4test* domain to access this share. To add the user, select the plus (**+**) button at the end of the column. The new entry is added. Click a new entry of **Type** column and select **User**. Input the user name *storage4test\redbook2* and configure the access control that you want. After all entry inputs are done, click **OK**. The result of editing the access control is shown in the Change Access Control window. See Figure 7-91.



Change Access Control

✓ Task completed. 100%

▾ View more details

```
Task started.                                               1:38 PM
Running command:                                            1:38 PM
chacl '/ibm/gpfs1'                                          1:38 PM
+usr:storage4test\redbook2:alw:rwmxdDaAnNcCos:fdi --cluster
12402779239044960749
The task is 100% complete.                                  1:38 PM
Task completed.                                             1:38 PM
```

Figure 7-91 Change Access Control result

Note: The preferred method to manage access is per group instead of per individual user. This way, users can be easily added to or removed from the group and the access permissions follow. If individual users are added directly to ACLs, a change means that the ACLs of all corresponding directories and files must be updated. On the authentication server, such as Active Directory or LDAP, you can create groups and add users as members. Provide access to this group. If you have local authentication that is configured, you can create groups and add users to it too. Providing ACLs to groups has an added advantage of managing inheritance easily for the whole group of users simultaneously.

Using the CLI: You can edit ACLs by running **chacl**. You can see the current ACL settings for each file system, file set, and share by running **lsacl**. See the command usage and output in Example 7-25.

Example 7-25 Command usage and output to edit and list Access Control

```
[furby.storage.tucson.ibm.com]$ chacl --help
usage: chacl path  [operations] [--copy-inherited | --delete-inherited] [-c < clusterID | clusterName >]
Change the ACL for a specified file or directory.

Parameter             Description
path                  Specifies the path name of the file or directory for which the ACL is manipulated.
operations            Specifies the operations against the ACL.
    --copy-inherited   Converts inherited ACEs into explicit ones
    --delete-inherited Deletes all inherited ACEs
-c, --cluster         The cluster scope for this command
usage: lsacl  [path] [--templates] [-v | -Y | -p] [-c < clusterID | clusterName >]
Displays the access control list (ACL) of a file or directory.

Parameter       Description
path            Specifies the path name of the file or directory for which the ACL is to be displayed.
    --templates List all ACL templates
-v, --verbose   Shows verbose output
-Y              Shows parseable output
-p, --pipe      Output format for piping into chacl
-c, --cluster   The cluster scope for this command
[furby.storage.tucson.ibm.com]$ lsacl /ibm/gpfs1
Wty Who                    Type Access mask    Flags
spl owner@                 alw  rwmxdDaAnNcCos fd----
spl group@                 alw  -----------Co- ------
spl everyone@              alw  ---x---------- -d----
usr root                   alw  rwmxdDaAnNcCos fdi---
usr STORAGE4TEST\redbook2 alw  rwmxdDaAnNcCos fdi---
grp root                   alw  rwmxdDaAnNcCos fdi---
EFSSG1000I The command completed successfully.
```

Note: Edited ACLs are not propagated to the existing subdirectory in SONAS V1.5.1. If there are already files in the file system, file set, or share, you cannot edit ACLs unless the user has the Data Access role. Because of these limitations, the preferred practice is to configure the ACL when you create file systems, file sets, or shares. However, you can enable the user to edit an ACL that has existing files by adding the Data Access role to the user. For more information, see 7.7, “User management” on page 575.

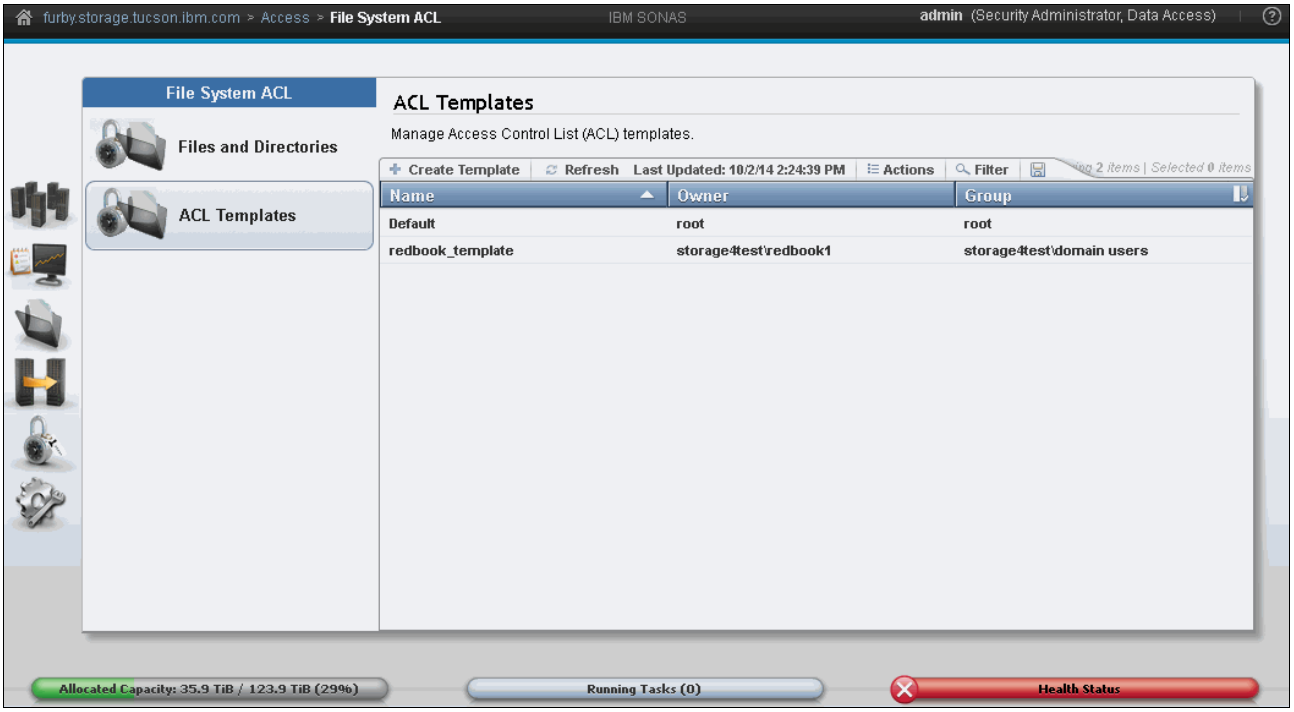
Editing ACL templates

SONAS supports the following type of ACL templates:

- Default template
- Department template
- User template

The default and department templates are applied automatically to the file systems and file sets. However, user templates must always be applied when they are created with the GUI and CLI.

To edit the ACL templates, click **Access** → **File System ACL** and select **ACL templates** in the right pane. See Figure 7-92.



furby.storage.tucson.ibm.com > Access > File System ACL
IBM SONAS
admin (Security Administrator, Data Access)
File System ACL
Files and Directories
ACL Templates
Create Template
Refresh
Actions
Filter
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

ACL Templates

Manage Access Control List (ACL) templates.

Last Updated: 10/2/14 2:24:39 PM

ing 2 items | Selected 0 items

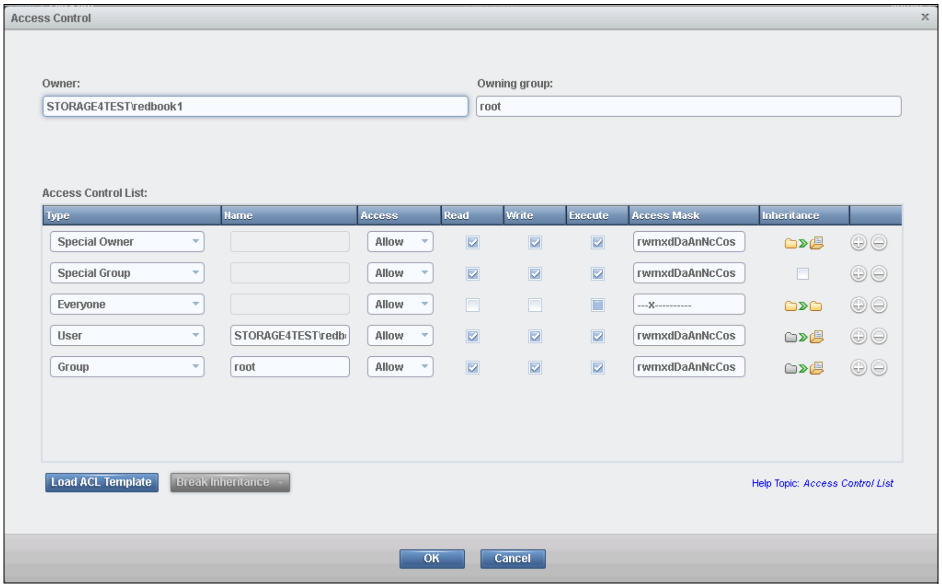
Name	Owner	Group
Default	root	root
redbook_template	storage4test\redbook1	storage4test\domain users

Figure 7-92 ACL Templates pane

Only the Default template is shown. The Department template is not shown and you cannot remove a default template. You can only modify it. You can also create a User template to apply a specific file system, file set, or share. To modify the Default template, click **Actions** → **Edit**. To create a User template, click **Create template**.

Loading ACL templates

The created User templates can be loaded when the file system, file set, or share are created, and can replace an existing template. To apply this template to an existing file set, click **Files** → **File Sets** and select the file set that you want. Then, click **Actions** → **Edit Access Control**. The Access Control window opens. See Figure 7-93 on page 563.



Owner: STORAGE4TEST\redbook1

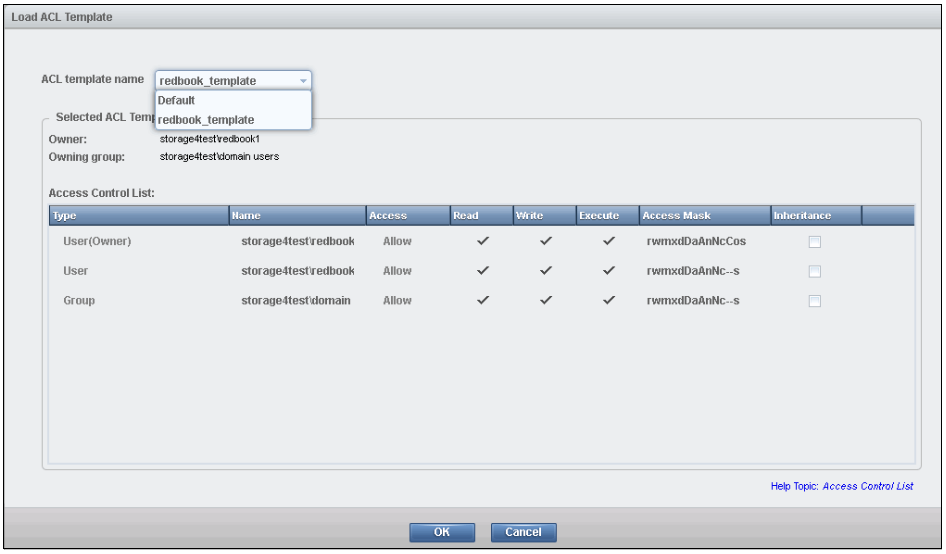
Owning group: root

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance	
Special Owner		Allow	☑	☑	☑	rwmxdDaAnNcCos	📁»📄	⊕ ⊖
Special Group		Allow	☑	☑	☑	rwmxdDaAnNcCos	☐	⊕ ⊖
Everyone		Allow	☐	☐	■	---x----------	📁»📁	⊕ ⊖
User	STORAGE4TEST\redb	Allow	☑	☑	☑	rwmxdDaAnNcCos	📁»📄	⊕ ⊖
Group	root	Allow	☑	☑	☑	rwmxdDaAnNcCos	📁»📄	⊕ ⊖

Figure 7-93 Access Control window

Click **Load ACL Template** to load a user-defined ACL template. Select the user template at the ACL template field. The *redbook_template* template is shown in Figure 7-94.



ACL template name: redbook_template

Owner: storage4test\redbook1

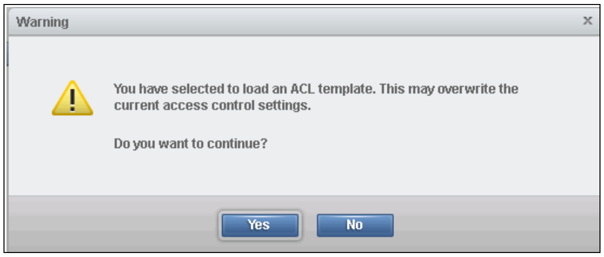
Owning group: storage4test\domain users

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance
User(Owner)	storage4test\redbook	Allow	✓	✓	✓	rwmxdDaAnNcCos	☐
User	storage4test\redbook	Allow	✓	✓	✓	rwmxdDaAnNc--s	☐
Group	storage4test\domain	Allow	✓	✓	✓	rwmxdDaAnNc--s	☐

Figure 7-94 Load ACL Template window

Click **OK** to apply on that file set. If you load and apply a user template, the existing ACL is overwritten by the loaded template. A warning window opens. See Figure 7-95.

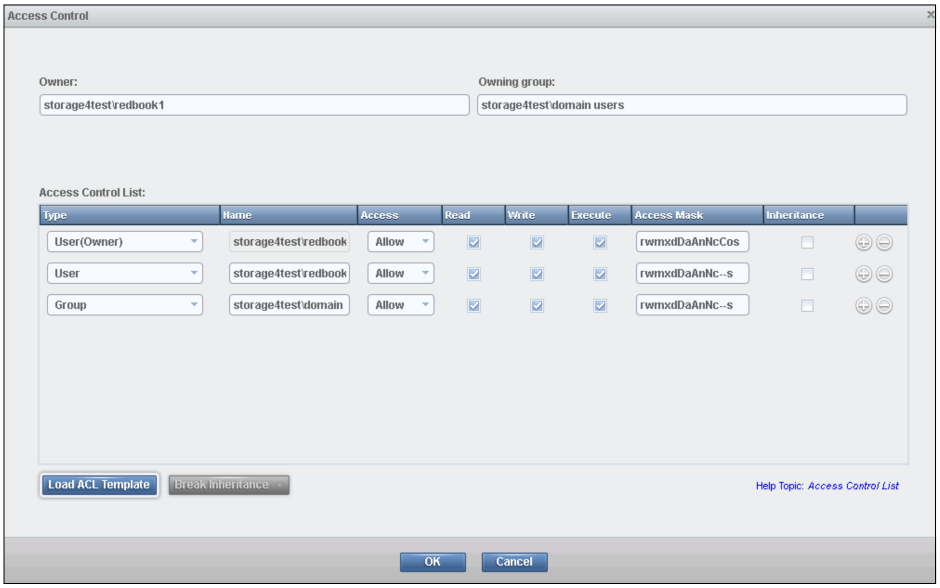


You have selected to load an ACL template. This may overwrite the current access control settings.

Do you want to continue?

Figure 7-95 Warning window

Click **Yes** to confirm. The loaded ACL is shown in the Access Control window. See Figure 7-96.



Owner: storage4test\redbook1

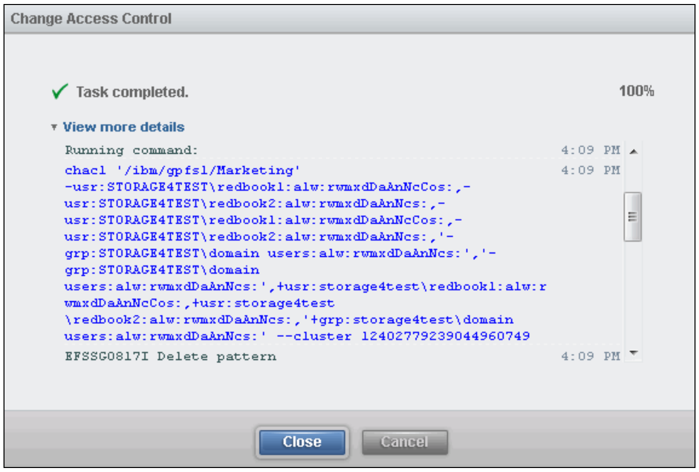
Owning group: storage4test\domain users

Access Control List:

Type	Name	Access	Read	Write	Execute	Access Mask	Inheritance	
User(Owner)	storage4test\redbook	Allow	☑	☑	☑	rwmxdDaAnNcCos	□	⊕ ⊖
User	storage4test\redbook	Allow	☑	☑	☑	rwmxdDaAnNc--s	□	⊕ ⊖
Group	storage4test\domain	Allow	☑	☑	☑	rwmxdDaAnNc--s	□	⊕ ⊖

Figure 7-96 Access Control window after the template is loaded

Click **OK** to confirm. The result is shown in the Change Access Control window. See Figure 7-97.



✓ Task completed. 100%

```
Running command:                                        4:09 PM
chacl '/ibm/gpfsl/Marketing'                            4:09 PM
-usr:STORAGE4TEST\redbook1:alw:rwmxdDaAnNcCos:,-
usr:STORAGE4TEST\redbook2:alw:rwmxdDaAnNcs:,-
usr:STORAGE4TEST\redbook1:alw:rwmxdDaAnNcCos:,-
usr:STORAGE4TEST\redbook2:alw:rwmxdDaAnNcs:,'-
grp:STORAGE4TEST\domain users:alw:rwmxdDaAnNcs:','-
grp:STORAGE4TEST\domain
users:alw:rwmxdDaAnNcs:',+usr:storage4test\redbook1:alw:r
wmxdDaAnNcCos:,+usr:storage4test
\redbook2:alw:rwmxdDaAnNcs:,'+grp:storage4test\domain
users:alw:rwmxdDaAnNcs:' --cluster 12402779239044960749
EFSSG0817I Delete pattern                               4:09 PM
```

Figure 7-97 Change Access Control result window

Note: The SONAS GUI provides a window that you can use to display and edit the Owner, Owning Group, and ACL of any file or directory within a file system. To access this window, click **Access** → **File System ACL** → **Files and Directories**. Only users with the Data Access role have permission to use this window.

7.5.7 Testing access to the exports

This section explains how to access the shares. It examines how NFS and CIFS can be accessed by mounting the exports.

CIFS

CIFS exports must be mounted before they can be accessed. A CIFS share can be accessed by using both Windows and UNIX systems.

- Accessing CIFS from the Windows operating system:

  To mount a CIFS share in the Windows operating system, right-click **My computer** and click **Map a Network Drive**, as shown in Figure 7-98.

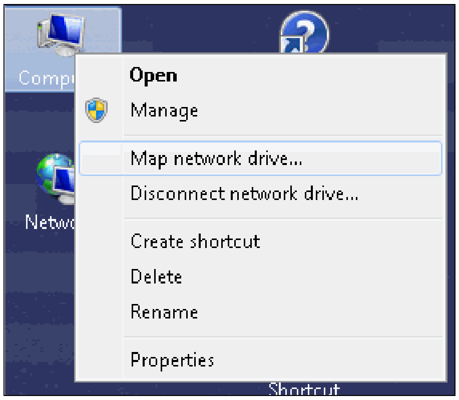


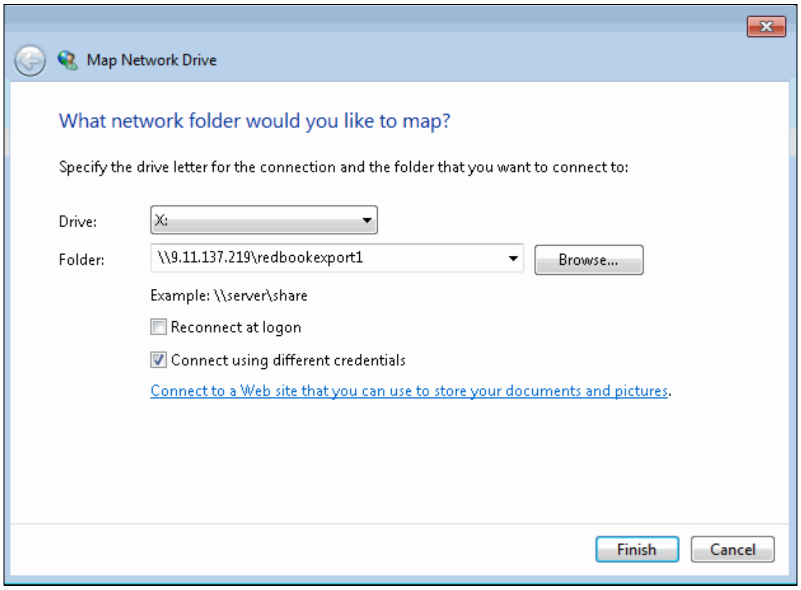
Figure 7-98 Mapping a drive on Windows to access a CIFS share

A new window opens with fields for the drive and path details. Choose a drive letter from the list. Enter the path for the share you want to access in the following format:

`\\cluster_name\sharename`

cluster_name is the name of the cluster that you want to access and *sharename* is the name of the share that you want to mount.

In this example, as shown in Figure 7-99, specify the IP as 9.11.137.219 and the share name as redbookexport1. Mount the share on the X drive. Select the **Connect using different credentials** check box.



Map Network Drive

What network folder would you like to map?

Specify the drive letter for the connection and the folder that you want to connect to:

Drive: X:

Folder: \\9.11.137.219\redbookexport1

Example: \\server\share

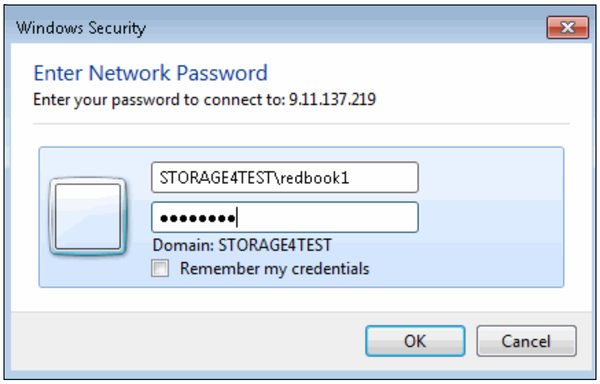
□ Reconnect at logon

☑ Connect using different credentials

Browse...
Connect to a Web site that you can use to store your documents and pictures.
Finish
Cancel

Figure 7-99 Choose the drive letter and path to mount

Click **Finish** and enter the user name and password in the new window. This user must have access or ACLs set to access this share. In this example, the user is “STORAGE4TEST\redbook1” in the “STORAGE4TEST” domain. See Figure 7-100.



Windows Security

Enter Network Password

Enter your password to connect to: 9.11.137.219

STORAGE4TEST\redbook1

••••••••

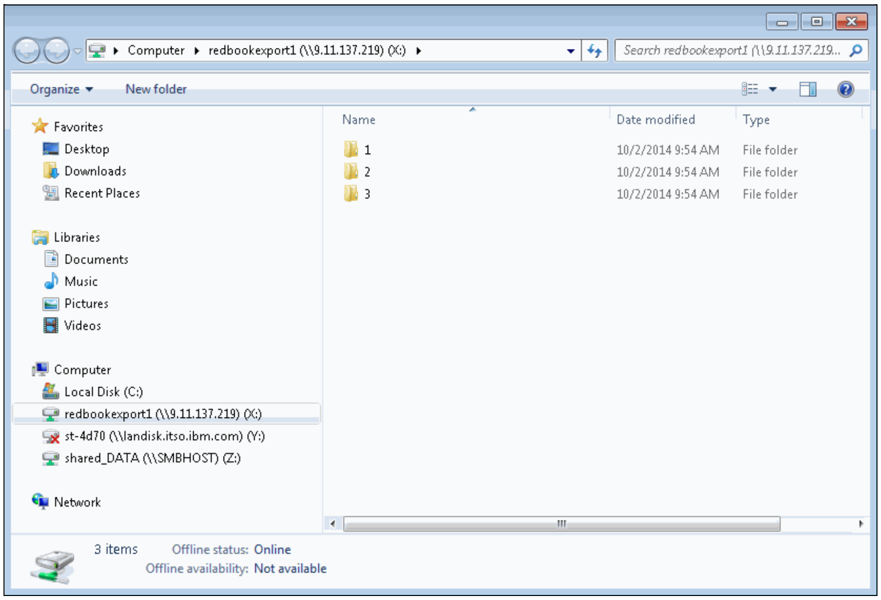
Domain: STORAGE4TEST

□ Remember my credentials

Figure 7-100 Add a user name and password

Click **OK**. The share mounts successfully. You can then access the share by accessing My Computer and the X drive, which you just mounted.

Double-click the drive and you can see the contents of the share, as shown in Figure 7-101.


Computer ▸ redbookexport1 (\\9.11.137.219) (X:) ▸
Search redbookexport1 (\\9.11.137.219...
Organize ▾
New folder
Favorites
Desktop
Downloads
Recent Places
Libraries
Documents
Music
Pictures
Videos
Computer
Local Disk (C:)
redbookexport1 (\\9.11.137.219) (X:)
st-4d70 (\\landisk.itso.ibm.com) (Y:)
shared_DATA (\\SMBHOST) (Z:)
Network

Name	Date modified	Type
1	10/2/2014 9:54 AM	File folder
2	10/2/2014 9:54 AM	File folder
3	10/2/2014 9:54 AM	File folder

3 items Offline status: Online
Offline availability: Not available

Figure 7-101 Data that is seen from a mounted share

- Accessing CIFS from a Linux client:

  Mount the CIFS share by running **`mount.cifs`**, as shown in Example 7-26.

  In this example, the Linux client *lincli* is used. You create a directory, `cifsmount`, in the `/mnt` directory, where you mount the share. The SONAS cluster is `furby.storage.tucson.ibm.com` and the share is redbookexport1. The STORAGE4TEST\redbook1 user is used for access. It belongs to the STORAGE4TEST domain.

Example 7-26 Command to mount and access the CIFS share from a Linux client

```
[root@lincli mnt]# mkdir cifsmount
[root@lincli mnt]# ls
boot  cdrom  cifsmount
[root@lincli mnt]# mount.cifs furby.storage.tucson.ibm.com:redbookexport1 /mnt/cifsmount -o
user='storage4test\redbook1' pass=Passw0rd dom=storage4test
WARNING: using NFS syntax for mounting CIFS shares is deprecated and will be removed in cifs-utils-6.0. Migrate
to UNC syntax.
Password:
[root@lincli mnt]# df
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/mapper/vg_oc0404016450-lv_root
                     298389212 159633240 135724724  55% /
tmpfs                  3889368      3848   3885520   1% /dev/shm
/dev/sda1              1007896    108100    848596  12% /boot
/dev/sdb1            488383528 306336896 182046632  63% /media/DATA
furby.storage.tucson.ibm.com:redbookexport1 104857600        32 104857568   1% /mnt/cifsmount
[root@lincli mnt]# cd /mnt/cifsmount
[root@lincli cifsmount]# ls
1 2 3
```

> **Note:** A CIFS connection is typically used by Windows clients. UNIX clients do not support a CIFS connection without more commands or applications. In Example 6-26, the **`mount.cifs`** command is not allowed on all UNIX systems. However, most Linux clients support the **`mount.cifs`** command. Ensure that the **`mount.cifs`** command is supported on the current UNIX system before you connect to a CIFS share from a UNIX client.

NFS

NFS shares are mounted so that you can access data. This example shows how to mount UNIX clients. In this example, the Linux client lincli is used. You create a directory, `nfs_export`, in the `/mnt` directory, where you mount the NFS export. The cluster is `furby.storage.tucson.ibm.com` and the share is “redbookexport1”. See Example 7-27.

Example 7-27   NFS share mount

```
[root@lincli mnt]# pwd
/mnt
[root@lincli mnt]# mkdir nfsmount
[root@lincli mnt]# showmount -e furby.storage.tucson.ibm.com
Export list for 9.11.137.219:
/ibm/gpfs1                *
/ibm/gpfs0/redbookexport1 *
/ibm/gpfs0/admin/EDSCED   *
/ibm/gpfs0/furbywin1      *
/ibm/gpfs0/furbylin1      *
[root@lincli mnt]# mount furby.storage.tucson.ibm.com:/ibm/gpfs0/redbookexport1 /mnt/nfsmount
[root@lincli mnt]# df
Filesystem              1K-blocks        Used   Available Use% Mounted on
/dev/mapper/vg_oc0404016450-lv_root
                        298389212   157866324   137491640  54% /
tmpfs                     3889368        3848     3885520   1% /dev/shm
/dev/sda1                 1007896      108100      848596  12% /boot
/dev/sdb1               488383528   306336892   182046636  63% /media/DATA
furby.storage.tucson.ibm.com:/ibm/gpfs0/redbookexport1
                     103150518272 38492766208 64657752064  38% /mnt/nfsmount
[root@lincli ~]# ls /mnt/nfsmount
1  2  3
```

FTP

FTP shares can be accessed by both Windows and UNIX clients. To access the export, run **`ftp`**. You can also use external FTP client applications on Windows clients to access the share. This section explains access from both Windows and UNIX.

- Accessing FTP from Windows clients:

  You can use any FTP client to access data from the FTP export. Use the CLI to display the information. In this example, the cluster is `furby.storage.tucson.ibm.com` and the share is redbookexport1.

  When you run FTP, you are prompted to enter the user ID and password. In this example, the user is STORAGE4TEST\\redbook1, which is in the STORAGE4TEST domain. See Example 7-28. You then must run **`cd`** at the FTP prompt to the sharename that you want to access. As shown next, run **`ftp> cd shared`** to access the redbookexport1 FTP export.

  Example 7-28   Access an FTP Share from the Windows operating system

  ```
  C:\Users\IBM_ADMIN>ftp furby.storage.tucson.ibm.com
  Connected to Furby.storage.tucson.ibm.com.
  220 (vsFTPd 2.2.2)
  User (furby.storage.tucson.ibm.com:(none)): storage4test\redbook1
  331 Specify the password.
  Password:
  230 Login successful.
  ftp> cd redbookexport1
  250 Directory successfully changed.
  ftp> ls
  200 PORT command successful. Consider using PASV.
  ```

```
150 Here comes the directory listing.
1
2
3
226 Directory send OK.
ftp: 9 bytes received in 0.00Seconds 9000.00Kbytes/sec.
```

- Accessing FTP from UNIX:

  You can access the FTP data by running **`ftp`** from the UNIX client. In this example, the cluster is `furby.storage.tucson.ibm.com`, the share is redbookexport1, and the Linux client is lincli.

  When you run FTP, you are prompted to enter the user ID and password. In this example, the user is STORAGE4TEST\\redbook1, which belongs to the STORAGE4TEST domain. See Example 7-29. You then must run **`cd`** at the FTP prompt to the share name that you want to access. As shown next, run **`ftp> cd shared`** to access the redbookexport1 FTP export.

  Example 7-29 Access the FTP share from a Linux client

```
[root@lincli ~]# ftp furby.storage.tucson.ibm.com
Connected to furby.storage.tucson.ibm.com.
220 (vsFTPd 2.2.2)
Name (furby.storage.tucson.ibm.com:root): storage4test\redbook1
331 Specify the password.
Password:
230 Login successful.
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> cd redbookexport1
250 Directory successfully changed.
ftp> ls
227 Entering Passive Mode (9,11,137,219,99,213).
150 Here comes the directory listing.
drwx-----x    2 STORAGE4TEST\redbook1 STORAGE4TEST\domain users      512 Oct 02       09:54 1
drwx-----x    2 STORAGE4TEST\redbook1 STORAGE4TEST\domain users      512 Oct 02       09:54 2
drwx-----x    2 STORAGE4TEST\redbook1 STORAGE4TEST\domain users      512 Oct 02       09:54 3
226 Directory send OK.
```

7.6 Disk management

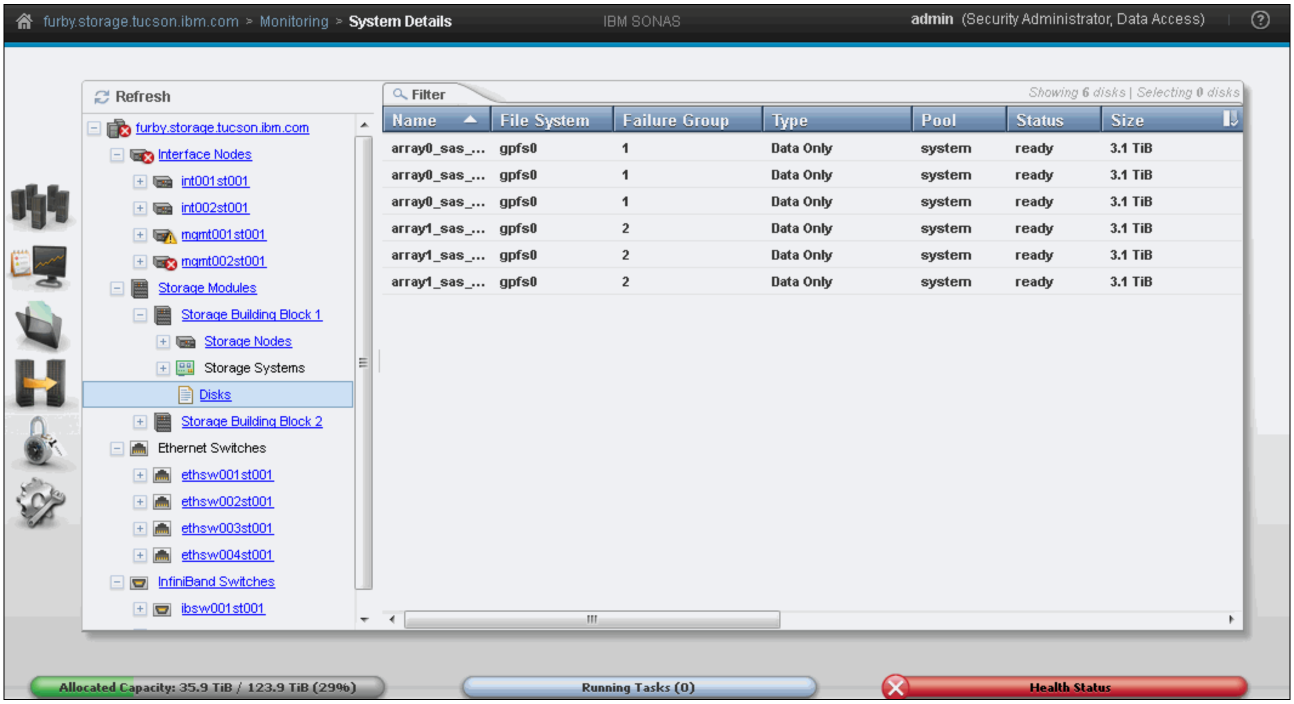
Each of the disks that exists in the SONAS can be managed. You can view the status of the disks and do actions such as suspend and resume. You can also start disks.

7.6.1 Listing disks and viewing status

This section describes how to list disks and view their status.

Using the GUI: Hover your cursor over the monitoring icon and select **System Details**. From there, go to Storage Building Block 1. The system can be made from multiple storage building blocks. In this case, two building blocks are attached to the system.

It displays a table with all the disks and the information about each one. You can see the name, the file system it is attached to, usage details, failure group, storage pool, and more. See Figure 7-102.



Name	File System	Failure Group	Type	Pool	Status	Size
array0_sas_...	gpfs0	1	Data Only	system	ready	3.1 TiB
array0_sas_...	gpfs0	1	Data Only	system	ready	3.1 TiB
array0_sas_...	gpfs0	1	Data Only	system	ready	3.1 TiB
array1_sas_...	gpfs0	2	Data Only	system	ready	3.1 TiB
array1_sas_...	gpfs0	2	Data Only	system	ready	3.1 TiB
array1_sas_...	gpfs0	2	Data Only	system	ready	3.1 TiB

Figure 7-102 List disks in the GUI

Using the CLI: You can list the disks in the cluster by running **lsdisk**. This command lists the existing disks along with information such as the file system it is attached to, the failure group, the storage pool, the type of disk, and many more things. The command usage and output are shown in Example 7-30.

Example 7-30 Command usage and help to list the disks in the cluster

```
[furby.storage.tucson.ibm.com]$ lsdisk
Name                                      File system Failure group Type            Pool     Status Availability Timestamp
array0_sas_60001ff07735017890b0001        gpfs0       1             dataOnly        system   ready  up           10/2/14 3:03 AM
array0_sas_60001ff07735019890d0003        gpfs0       1             dataOnly        system   ready  up           10/2/14 3:03 AM
array0_sas_60001ff0773501b890f0005        gpfs0       1             dataOnly        system   ready  up           10/2/14 3:03 AM
array1_sas_60001ff07735016890a0000        gpfs0       2             dataOnly        system   ready  up           10/2/14 3:03 AM
array1_sas_60001ff07735018890c0002        gpfs0       2             dataOnly        system   ready  up           10/2/14 3:03 AM
array1_sas_60001ff0773501a890e0004        gpfs0       2             dataOnly        system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002e9dfa0000290b51519eb1 gpfs0       3             dataOnly        system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002e9dfa0000290d51519ed7 gpfs0       3             dataOnly        system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002eb696000026ed515983e1 gpfs0       1             dataOnly        nearline ready  up           10/2/14 3:03 AM
DCS3700_360080e50002eb696000026f15159841c gpfs0       1             dataOnly        3TBNL    ready  up           10/2/14 3:03 AM
DCS3700_360080e50002eb696000026f251598438 gpfs0       1             dataOnly        nearline ready  up           10/2/14 3:03 AM
DCS3700_360080e50002eb696000026f35159845b gpfs0       1             dataOnly        nearline ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ec298000025b751598685 gpfs0       1             dataOnly        nearline ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d28512e2be7 gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d2b512e2be8 gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d2d512e2bea gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d2f512e2beb gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d31512e2bec gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d33512e2bed gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d35512e2bef gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d37512e2bef gpfs0       5             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d3a512e2bf1 gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d3d512e2bf2 gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d3f512e2bf5 gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d41512e2bf6 gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d43512e2bf8 gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d45512e2bf9 gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d47512e2bfb gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002d49512e2bfc gpfs0       6             metadataOnly    system   ready  up           10/2/14 3:03 AM
```

```
DCS3700_360080e50002ed6d600002dda51519e61 gpfs0       4             dataOnly        system   ready  up          10/2/14 3:03 AM
DCS3700_360080e50002ed6d600002ddc51519ed1 gpfs0       4             dataOnly        system   ready  up          10/2/14 3:03 AM
DCS3700_360080e50002ec298000025b8515986a2 gpfs1       2             dataAndMetadata system   ready  up          10/2/14 3:03 AM
EFSSG1000I The command completed successfully.
```

7.6.2 Suspending disks

This section explains how to suspend disks.

Using the GUI: There is no GUI option to suspend a disk.

Using the CLI: To suspend a disk, run **chdisk**. In Example 7-31, a procedure for suspending and resuming a disk that is named `array0_sas_60001ff07975001890901` is shown, in addition to the help option. To suspend a disk, the disk must be a part of a file system. In this case, it is part of file system `gpfs0`.

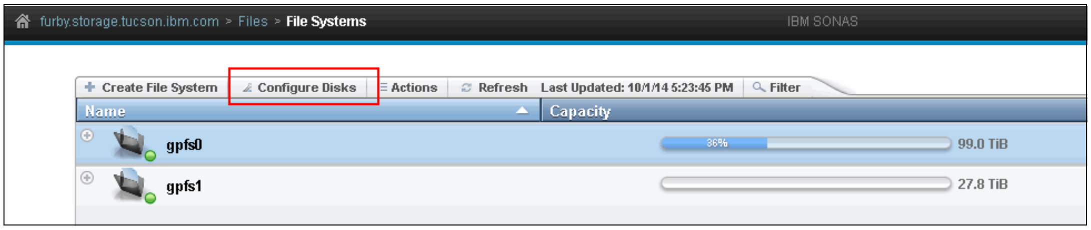
Example 7-31 Example of suspending a disk and resuming it

```
[furby.storage.tucson.ibm.com]$ lsdisk
Name                               File system Failure group Type            Pool   Status Availability Timestamp
array0_sas_60001ff07975001890901   gpfs0       1             dataAndMetadata system ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890902   gpfs0       1             dataAndMetadata system ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890901   gpfs0       2             dataAndMetadata system ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890902   gpfs0       2             dataAndMetadata system ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890903               1             dataAndMetadata system ready               10/21/11 8:56 PM
array1_sas_60001ff07975001890903               2             dataAndMetadata system ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890901             3             dataAndMetadata system ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890902             3             dataAndMetadata system ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890901             4             dataAndMetadata system ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890902             4                             system ready               10/21/11 8:56 PM
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ chdisk --help
usage: chdisk disks  {--failuregroup <failureGroup> | --pool <pool> | --usagetype <usageType> | --action <action>} [-c < clusterID |
clusterName >]
Changes a disk.

Parameter          Description
disks              Specifies either a stand-alone disk or a comma-separated list of disks that should be changed. The only disks that can
be changed cannot belong to a file system yet.
    --failuregroup Specifies a failure group for the specified disks.
    --pool         Specifies a pool for the specified disks.
    --usagetype    Specifies the usage type for the specified disks. Valid usage types are dataAndMetadata, dataOnly, metadataOnly, and
descOnly.
    --action       Changes the state for the specified disks. Valid values are suspend, resume, and start.
-c, --cluster      The cluster scope for this command
[furby.storage.tucson.ibm.com]$ chdisk array0_sas_60001ff07975001890901 --action suspend
EFSSG0122I The disk or disks are changed successfully.
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsdisk
Name                               File system Failure group Type            Pool   Status    Availability Timestamp
array0_sas_60001ff07975001890901   gpfs0       1             dataAndMetadata system suspended up           10/21/11 8:56 PM
array0_sas_60001ff07975001890902   gpfs0       1             dataAndMetadata system ready     up           10/21/11 8:56 PM
array1_sas_60001ff07975001890901   gpfs0       2             dataAndMetadata system ready     up           10/21/11 8:56 PM
array1_sas_60001ff07975001890902   gpfs0       2             dataAndMetadata system ready     up           10/21/11 8:56 PM
array0_sas_60001ff07975001890903               1             dataAndMetadata system ready                  10/21/11 8:56 PM
array1_sas_60001ff07975001890903               2             dataAndMetadata system ready                  10/21/11 8:56 PM
array2_nlsas_60001ff07975001890901             3             dataAndMetadata system ready                  10/21/11 8:56 PM
array2_nlsas_60001ff07975001890902             3             dataAndMetadata system ready                  10/21/11 8:56 PM
array3_nlsas_60001ff07975001890901             4             dataAndMetadata system ready                  10/21/11 8:56 PM
array3_nlsas_60001ff07975001890902             4                             system ready                  10/21/11 8:56 PM
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ chdisk array0_sas_60001ff07975001890901 --action resume
EFSSG0122I The disk or disks are changed successfully.
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsdisk
Name                               File system Failure group Type            Pool   Status Availability Timestamp
array0_sas_60001ff07975001890901   gpfs0       1             dataAndMetadata system ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890902   gpfs0       1             dataAndMetadata system ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890901   gpfs0       2             dataAndMetadata system ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890902   gpfs0       2             dataAndMetadata system ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890903               1             dataAndMetadata system ready               10/21/11 8:56 PM
array1_sas_60001ff07975001890903               2             dataAndMetadata system ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890901             3             dataAndMetadata system ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890902             3             dataAndMetadata system ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890901             4             dataAndMetadata system ready               10/21/11 8:56 PM
```

```
array3_nlsas_60001ff07975001890902           4                            system ready              10/21/11 8:56 PM
EFSSG1000I The command completed successfully.
```

7.6.3 Changing the properties of disks

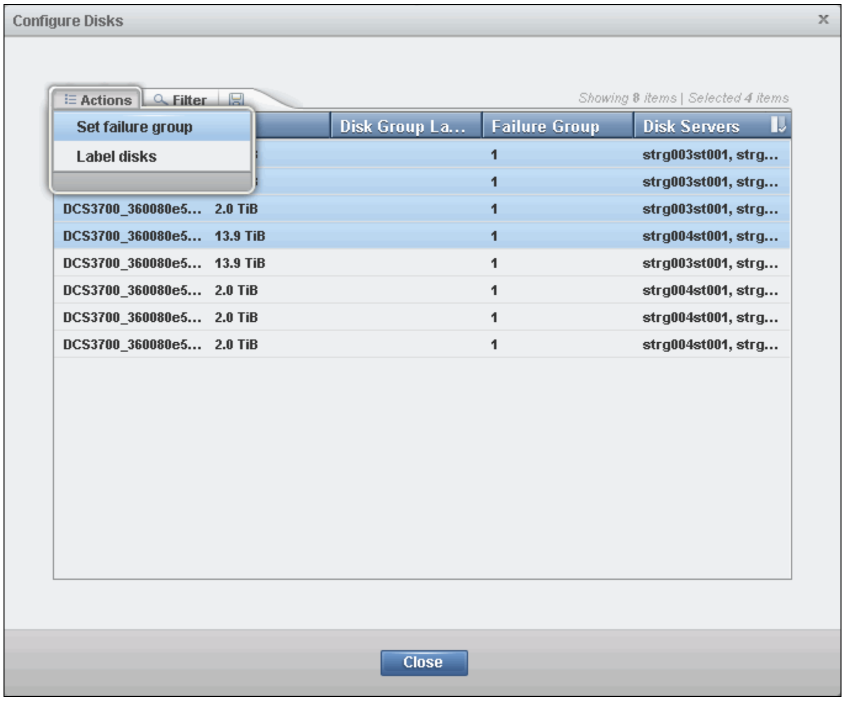
Using the GUI: Click **Files** → **File Systems** and click **Configure Disks**. See Figure 7-103.


furby.storage.tucson.ibm.com > Files > File Systems
IBM SONAS
Create File System
Configure Disks
Actions
Refresh
Last Updated: 10/1/14 5:23:45 PM
Filter
36%

Name	Capacity
gpfs0	99.0 TiB
gpfs1	27.8 TiB

Figure 7-103 Configure Disks menu in the File Systems window

A list of the disks is shown. Choose one or more disks from the list and click the **Set failure group** option from the **Actions** menu. See Figure 7-104.

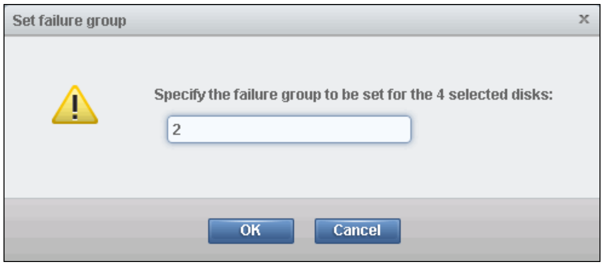


Showing 8 items | Selected 4 items

		Disk Group La...	Failure Group	Disk Servers
			1	strg003st001, strg...
			1	strg003st001, strg...
DCS3700_360080e5...	2.0 TiB		1	strg003st001, strg...
DCS3700_360080e5...	13.9 TiB		1	strg004st001, strg...
DCS3700_360080e5...	13.9 TiB		1	strg003st001, strg...
DCS3700_360080e5...	2.0 TiB		1	strg004st001, strg...
DCS3700_360080e5...	2.0 TiB		1	strg004st001, strg...
DCS3700_360080e5...	2.0 TiB		1	strg004st001, strg...

Figure 7-104 Disk configuration view

Enter the selected failure group number in the Set failure group window and click **OK** to save the settings. See Figure 7-105 on page 573.



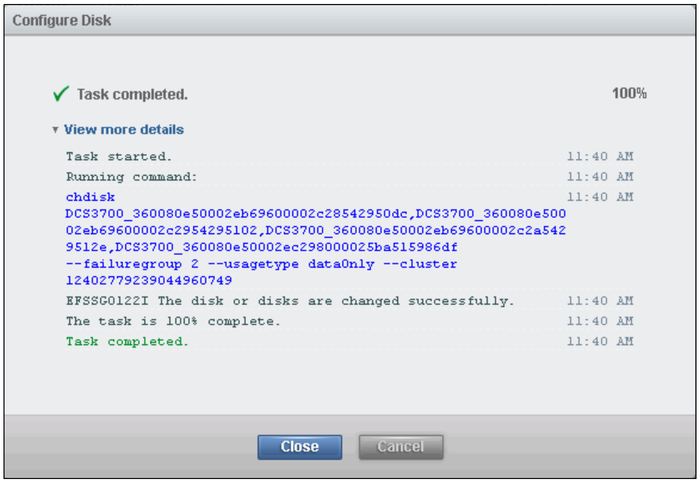
Set failure group

Specify the failure group to be set for the 4 selected disks:

2

Figure 7-105 Set failure group window

Click **Close** to exit the settings, as shown in Figure 7-106.



Configure Disk

✓ Task completed. 100%

▾ View more details

```
Task started.                                               11:40 AM
Running command:                                            11:40 AM
chdisk                                                      11:40 AM
DCS3700_360080e50002eb69600002c28542950dc,DCS3700_360080e500
02eb69600002c2954295102,DCS3700_360080e50002eb69600002c2a542
9512e,DCS3700_360080e50002ec298000025ba515986df
--failuregroup 2 --usagetype dataOnly --cluster
12402779239044960749
EFSSG0122I The disk or disks are changed successfully.      11:40 AM
The task is 100% complete.                                  11:40 AM
Task completed.                                             11:40 AM
```

Figure 7-106 Task window

Using the CLI: You can change the properties of a disk by running **chdisk**. The properties that you can modify for a disk are the Failure Group, Storage Pool, and Usage Type. The command usage is shown in Example 7-32.

Example 7-32 Command usage for changing the properties of a disk

```
[furby.storage.tucson.ibm.com]$ chdisk --help
usage: chdisk disks  {--failuregroup <failureGroup> | --pool <pool> | --usagetype <usageType> | --action <action>} [-c < clusterID |
clusterName >]
Changes a disk.
Parameter          Description
disks              Specifies either a stand-alone disk or a comma-separated list of disks that should be changed. The only disks that can
be changed cannot belong to a file system yet.
    --failuregroup Specifies a failure group for the specified disks.
    --pool         Specifies a pool for the specified disks.
    --usagetype    Specifies the usage type for the specified disks. Valid usage types are dataAndMetadata, dataOnly, metadataOnly, and
descOnly.
    --action       Changes the state for the specified disks. Valid values are suspend, resume, and start.
-c, --cluster      The cluster scope for this command
```

Here is a detailed explanation of each of the parameters that can be changed:

- Failure Group: You can change the Failure Group of a disk by using the **--failuregroup** option with **chdisk**.
- Storage Pool: You can change the Failure Group of a disk by using the **--storagepool** option with **chdisk**.

- Usage Type: You can change the Failure Group of a disk by using the **`--usagetype`** option with **`chdisk`**.
- Action: You can change a disk state for a disk by using the **`--action suspend/resume/start`** option. See Example 7-31 on page 571.

In Example 7-33, each of the parameters for one of the disks, `array0_sas_60001ff07975001890903`, are changed. The example also shows the state of the disk before changing and the disk whose information is changed, in **bold**.

Example 7-33 Command output for CLI command lsdisk and using chdisk to change failure group of disk

```
[furby.storage.tucson.ibm.com]$ lsdisk
Name                               File system Failure group Type            Pool   Status Availability Timestamp
array0_sas_60001ff07975001890901   gpfs0       1             dataAndMetadata system ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890902   gpfs0       1             dataAndMetadata system ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890901   gpfs0       2             dataAndMetadata system ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890902   gpfs0       2             dataAndMetadata system ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890903               1             dataAndMetadata system ready               10/21/11 8:56 PM
array1_sas_60001ff07975001890903               2             dataAndMetadata system ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890901             3             dataAndMetadata system ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890902             3             dataAndMetadata system ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890901             4             dataAndMetadata system ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890902             4                             system ready               10/21/11 8:56 PM
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ chdisk array0_sas_60001ff07975001890903 --failuregroup 100 --pool newpool --usagetype dataOnly
EFSSG0122I The disk or disks are changed successfully.
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsdisk
Name                               File system Failure group Type            Pool    Status Availability Timestamp
array0_sas_60001ff07975001890901   gpfs0       1             dataAndMetadata system  ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890902   gpfs0       1             dataAndMetadata system  ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890901   gpfs0       2             dataAndMetadata system  ready  up           10/21/11 8:56 PM
array1_sas_60001ff07975001890902   gpfs0       2             dataAndMetadata system  ready  up           10/21/11 8:56 PM
array0_sas_60001ff07975001890903               100           dataOnly        newpool ready               10/21/11 8:56 PM
array1_sas_60001ff07975001890903               2             dataAndMetadata system  ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890901             3             dataAndMetadata system  ready               10/21/11 8:56 PM
array2_nlsas_60001ff07975001890902             3             dataAndMetadata system  ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890901             4             dataAndMetadata system  ready               10/21/11 8:56 PM
array3_nlsas_60001ff07975001890902             4                             system  ready               10/21/11 8:56 PM
EFSSG1000I The command completed successfully.
```

7.6.4 Starting disks

Using the GUI: Click **Files** → **File Systems** and highlight one file system. Click **Actions** → **Start All Disks**, as shown in Figure 7-107.

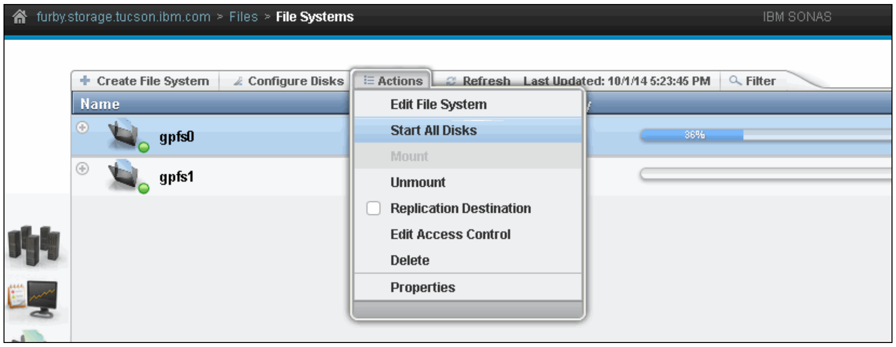


Figure 7-107 Start All disks menu

Using the CLI: Run **`chdisk --action start`**. See Example 7-31 on page 571, where help is listed.

7.6.5 Removing disks

Using the GUI: As of now, there is no GUI option to remove the disk.

Using the CLI: Run **chfs --remove**. See Example 7-34 for the **chfs** command help option.

Example 7-34 Help option for the chfs command

```
[furby.storage.tucson.ibm.com]$ chfs --help
usage: chfs fileSystem  [--atime <accessTime>] [--mtime <modTime>] [-i <maxInodes>] [-q <quota>] [-R <replication>] [--pool <poolName> |
--add <disks>] [--remove <disks>] [--force] [--noverify] [--dmapi | --nodmapi] [--logplacement <logPlacement>] [-c < clusterID |
clusterName >]
Changes the properties of the file system.

Parameter          Description
fileSystem         Specifies the name of the file system to be changed. File system names do not need to be fully qualified, but they must
be unique within a GPFS cluster.
    --atime        Stamps access times on every access to a file or directory if the 'exact' variable is selected. If the 'suppress'
variable is selected, access times will not be recorded.
    --mtime        Updates the modification time immediately to files and directories if the 'exact' variable is selected. Otherwise,
modification times will be updated after a delay of several seconds.
-i                 Sets the maximum number of inodes in the file system.
-q                 Changes the quota option (yes, no, perfileset).Changing this setting requires the file system to be in unmounted state.
-R                 Sets the level of replication in this file system.
    --pool         Adds a set of free Network Shared Disks (NSDs), which have the pool name set as pool, to the file system.
    --add          Adds disks to the file system. The disks contain a comma-separated list of disk names.
    --remove       Removes disks from the file system. The disks variable contains a comma-separated list of disk names.
    --force        Does not prompt for manual confirmation.
    --noverify     Suppresses the verification that specified disks do not belong to an existing file system.
    --dmapi        Enables external file system pool support, such as Tivoli Storage Manager.
    --nodmapi      Disables external file system pool support, such as Tivoli Storage Manager.
    --logplacement Sets whether the logs will be stored striped across the disks or not.
-c, --cluster      The cluster scope for this command
```

7.7 User management

Users who can access SONAS can be of multiple types, from system operators to specific backup or replication administrators to export administrators, and more. This section explains all user group roles.

7.7.1 User groups

The following user group roles are already available in the SONAS system:

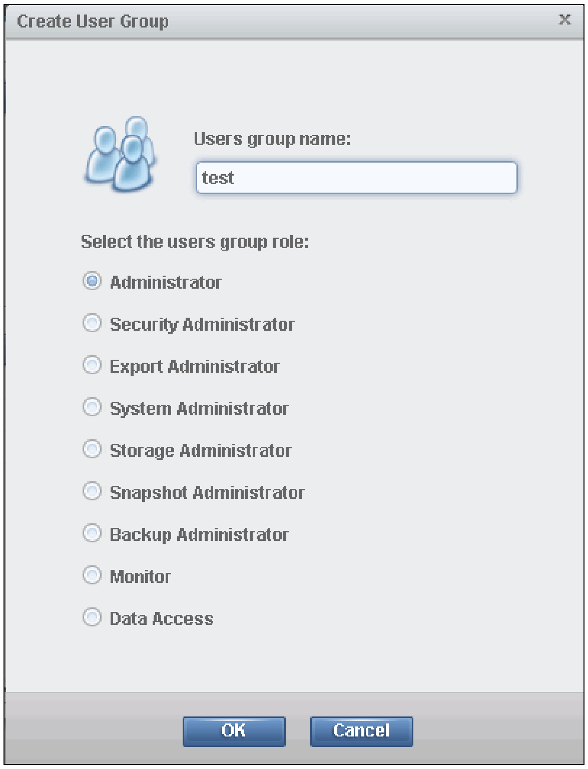
- Administrator: Manages the system, but cannot create or change new user groups or users.
- Security administrator: Manages all commands, including user management commands.
- Export administrator: Manages the export definitions for all supported protocols. Responsibility starts at the file system content level and services such as antivirus.
- System administrator: Manages clusters, nodes, syslogs, and authentication. Can access commands, which can take down the whole cluster.
- Storage administrator: Manages the disks, file system, pools, and file sets. Can also define ILM policies.
- Snapshot administrator: Manages snapshots for file system, file sets, and peer snapshots. Can configure snapshots associations and notifications.
- Backup administrator: Manages Tivoli Storage Manager and NDMP backup and replication.
- Operator: Has read-only access to the system.

In this case, a user who is called admin is added to the Security administrator group, so it has the privileges of creating and modifying user settings such as the password. The next section demonstrates how to create a user group.

7.7.2 Creating a user group

Here is an example of how to create a user group with the GUI. The CLI commands are also listed. You can create user groups and delegate roles from the list.

Create a user group that is named test with system administrator role privileges. When you click **Add User Group** → **Create New User Group,** as shown in Figure 7-23 on page 508, a window opens, as shown in Figure 7-108. Here, you enter the name and role of the new user group, in this case, the user group name *test* with the administrator role. When all the fields are filled, click **OK**.



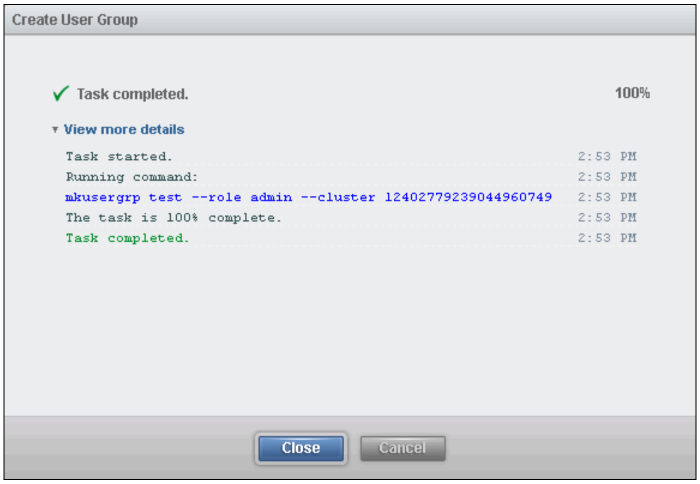
Users group name: test

Select the users group role:

● Administrator
○ Security Administrator
○ Export Administrator
○ System Administrator
○ Storage Administrator
○ Snapshot Administrator
○ Backup Administrator
○ Monitor
○ Data Access

Figure 7-108 New User Group window

Upon clicking **OK**, a new window opens, where you can see an appropriate CLI command that is used to achieve same task. A new window can be seen in Figure 7-109 on page 577. Here you see that the `mkusergrp test --role admin` command can be used in the CLI to achieve the same result as you did with the GUI.



```
Task started.                                                    2:53 PM
Running command:                                                 2:53 PM
mkusergrp test --role admin --cluster 12402779239044960749       2:53 PM
The task is 100% complete.                                       2:53 PM
Task completed.                                                  2:53 PM
```

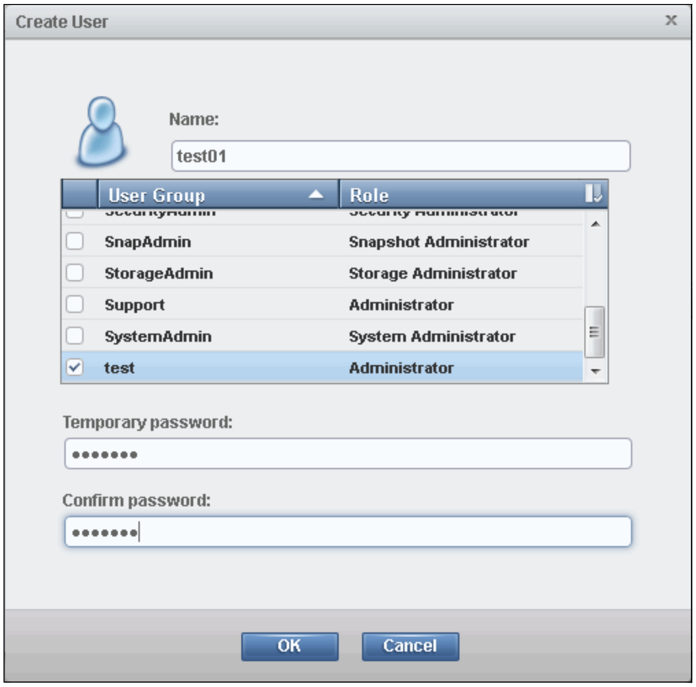
Figure 7-109 Completion of group creation and the corresponding CLI command

After you create a user group, create a user for this user group, as described in 7.7.3, “Creating a user” on page 577.

7.7.3 Creating a user

This section shows how to create a user with the GUI. The corresponding CLI command is also listed. For all actions that are done in the GUI, the corresponding CLI command is shown.

In the GUI, click **Access** → **Users**. A window opens, as shown in Figure 7-23 on page 508. Select **Create User**. A window opens, as shown in Figure 7-110. Select the user name, select the user group, and enter a password for the new user. Later in this section, you see how to manipulate it with the password policy for users. When all fields are correctly filled, click **OK**.



Name: test01

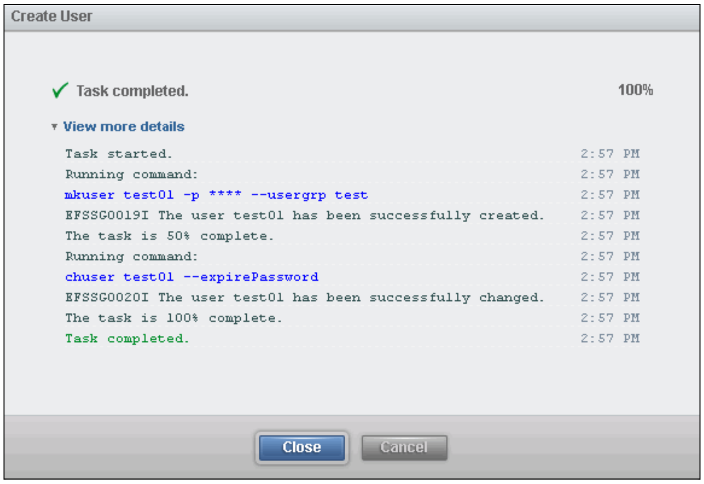
	User Group	Role
□	SnapAdmin	Snapshot Administrator
□	StorageAdmin	Storage Administrator
□	Support	Administrator
□	SystemAdmin	System Administrator
☑	test	Administrator

Temporary password: •••••••

Confirm password: •••••••

Figure 7-110 New User options

In this case, you create a user called *test01* in the test group. When you select **OK**, a window opens, as shown in Figure 7-111. Again, you can see the appropriate CLI command.



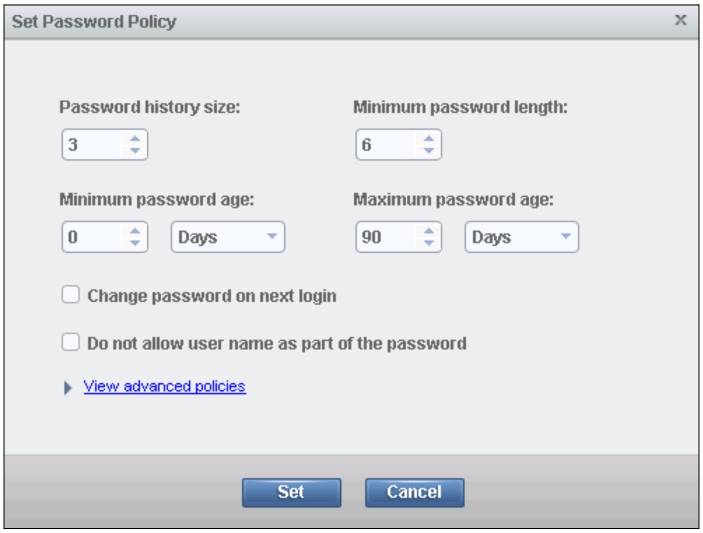
```
Task started.                                                   2:57 PM
Running command:                                                2:57 PM
mkuser test01 -p **** --usergrp test                            2:57 PM
EFSSG0019I The user test01 has been successfully created.       2:57 PM
The task is 50% complete.                                       2:57 PM
Running command:                                                2:57 PM
chuser test01 --expirePassword                                  2:57 PM
EFSSG0020I The user test01 has been successfully changed.       2:57 PM
The task is 100% complete.                                      2:57 PM
Task completed.                                                 2:57 PM
```

Figure 7-111 Final create user window

Previously, you created a new user group with the System Administrator role. Then, you created a user for this group. Now, you must set a password policy and session policy.

7.7.4 User password and session policy

After the creation of a new user, it is a good idea to set a password policy for this user so that the new user must change it at first logon. To use this function, click **Global Actions** → **Set Password Policy** in the window that is shown in Figure 7-23 on page 508. A window opens, as shown in Figure 7-112. You can click **View advanced polices** to see more polices.



Password history size: 3

Minimum password length: 6

Minimum password age: 0 Days

Maximum password age: 90 Days

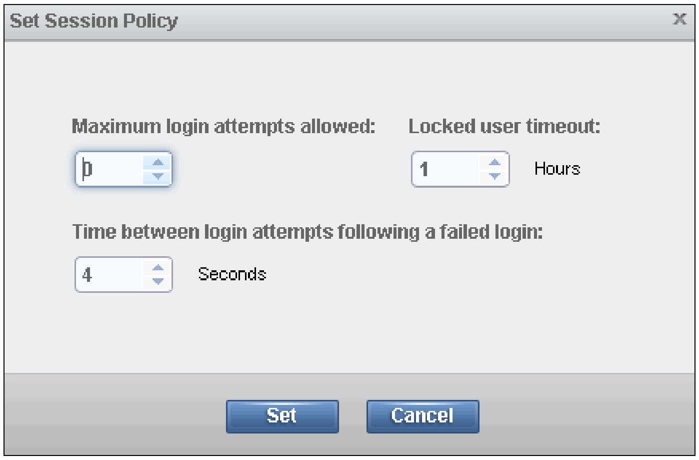
□ Change password on next login

□ Do not allow user name as part of the password

Figure 7-112 Password policy window

Here you can force a new or old user to change the password on next logon. You can also set a minimum password length and minimum and maximum password age.

From the **Global Actions** menu, if you select **Session policy**, you can configure how many login attempts the user can have, and you can also set a locked user timeout. This menu is shown in Figure 7-113.



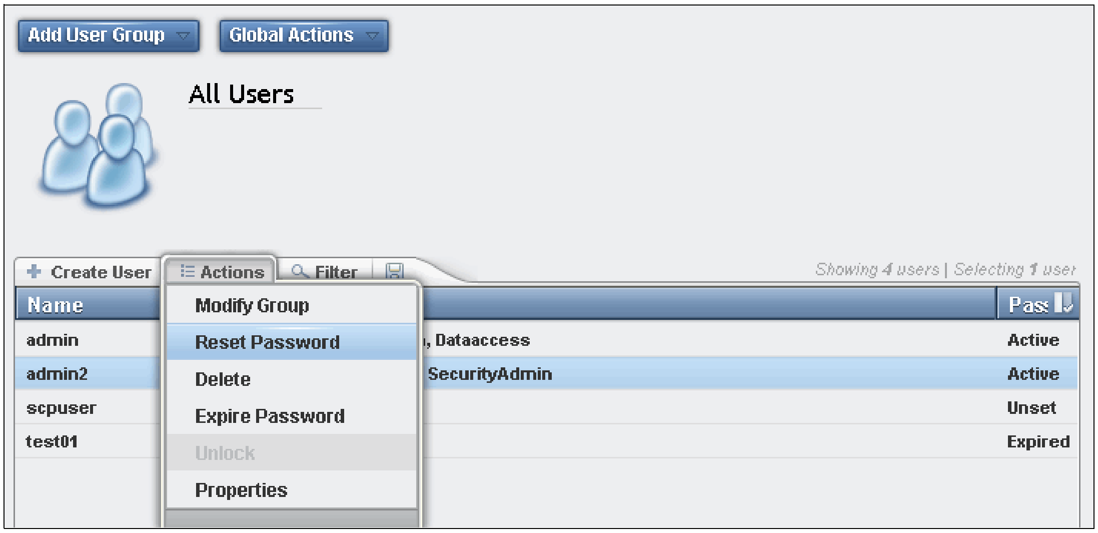
Maximum login attempts allowed: 0

Locked user timeout: 1 Hours

Time between login attempts following a failed login: 4 Seconds

Figure 7-113 Session Policy menu

If the user cannot remember their password, you can change the password or mark the current password as expired. You can do these tasks only if you have a role that enables you to do so. Select the user that you want and click **Actions** → **Expire Password** or **Reset Password**, as shown in Figure 7-114.



Showing 4 users | Selecting 1 user

Name		Pas
admin	, Dataaccess	Active
admin2	SecurityAdmin	Active
scpuser		Unset
test01		Expired

Figure 7-114 Set a user's password as expired or using the Actions menu to change it

7.7.5 Enabling remote users as SONAS CLI and GUI administrators

SONAS can enable remote user to access SONAS GUI and CLI environments in the remote authentication environment. This task consists of two functions: Enable Remote Users as Admins and Use Remote User Group. The Enable remote Users as Admins function allows the user in the specific remote group to log in to the GUI and CLI with the *monitor* role. Use Remote User Group allows the user in the specific remote group to log in to the GUI and CLI with the specific role that you want. This section explains the configuration procedure.

Note: To use this function, three conditions must be satisfied.

1. External authentication must be configured (AD, LDAP, and so on.)
2. You must configure the Enable remote User as Admins function first. If not, the Use Remote User Group function is deactivated.
3. SONAS V1.5.1 or higher must be installed.

Enabling remote users as admins

Using the GUI: Click **Access** → **Users** and click **Global Actions** → **Enable Remote Users as Admins**. This menu is visible only if LDAP or AD is already configured. Otherwise, it is disabled. See Figure 7-115.

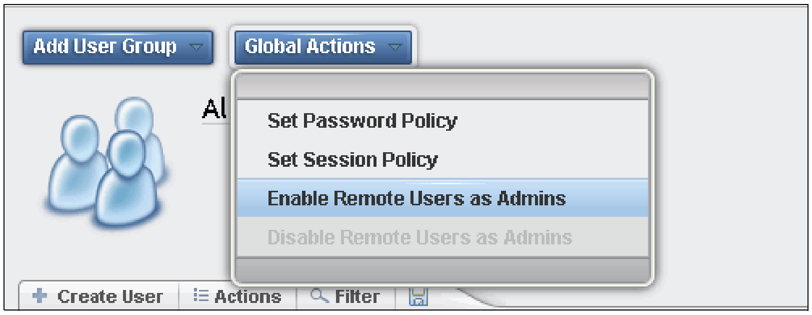
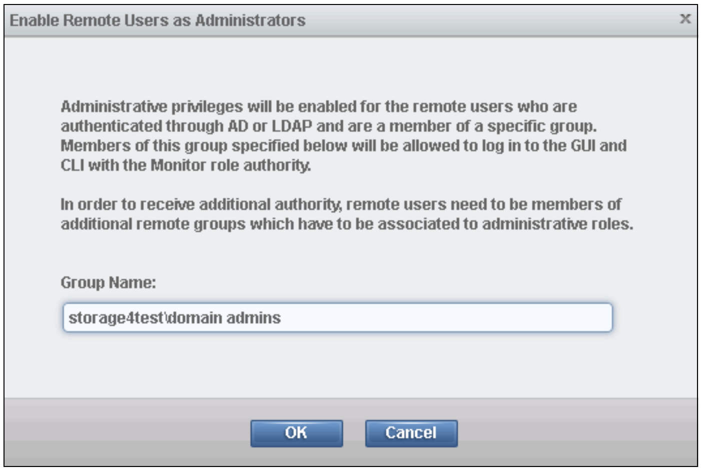


Figure 7-115 Enable Remote Users as Admins menu

Enter the remote user group, In this case, use *t*he storage4test domain and the domain admins group. The users in the storage4test\domain admins group can log in to GUI and CLI with only the Monitor role. See Figure 7-116. Click **OK** to proceed.


Enable Remote Users as Administrators

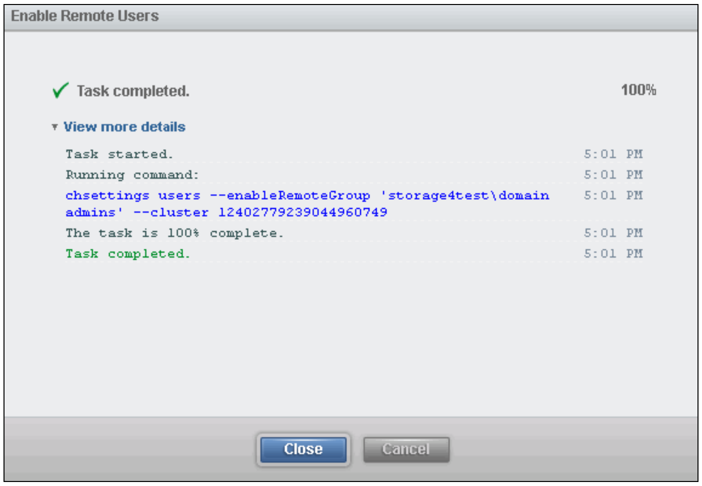
Administrative privileges will be enabled for the remote users who are authenticated through AD or LDAP and are a member of a specific group. Members of this group specified below will be allowed to log in to the GUI and CLI with the Monitor role authority.

In order to receive additional authority, remote users need to be members of additional remote groups which have to be associated to administrative roles.

Group Name: storage4test\domain admins

Figure 7-116 Enable Remote User as Administrators window

The result is shown in Figure 7-117 on page 581.



```
Task started.                                                5:01 PM
Running command:                                             5:01 PM
chsettings users --enableRemoteGroup 'storage4test\domain    5:01 PM
admins' --cluster 12402779239044960749
The task is 100% complete.                                   5:01 PM
Task completed.                                              5:01 PM
```

Figure 7-117 Enable Remote Users result

Using the CLI: You can configure this setting by running **`chsettings`**. See Example 7-35.

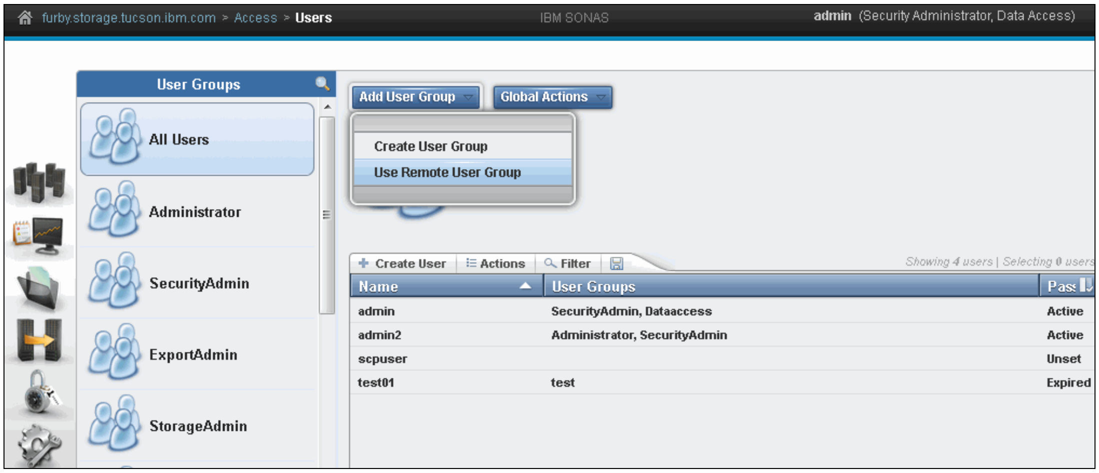
Example 7-35 The chsettings commands

```
[root@furby.mgmt001st001 ~]# chsettings --help
usage: chsettings component  [--bootLoaderPassword <bootLoaderPassword>] [--disableBootLoaderPassword]
[--bootFromRemovableMedia <yes|no>] [--sshHardening <yes|no>] [--allowLegacyEncryption] [--disAllowLegacyEncryption]
[--enableRemoteGroup <groupName> | --disableRemoteGroup] [-c < clusterID | clusterName >]
Change settings applicable to the whole cluster.

Parameter                       Description
component                      Specifies the component for which the settings are applied / changed for the cluster.
Valid components [security, kerberos].
    --bootLoaderPassword        Sets boot loader password for all the nodes within the cluster
    --disableBootLoaderPassword Disable boot loader password for all the nodes within the cluster
    --bootFromRemovableMedia    Specifies if we can boot from the removable media or not.
    --sshHardening              Specifies whether to enable or disable the SSH hardening settings to all the ssh
configuration files on the nodes in the SONAS cluster.
    --allowLegacyEncryption     Allow clients using legacy encryption.
    --disAllowLegacyEncryption  Disallow clients using legacy encryption.
    --enableRemoteGroup         Enable remote logging for remote users, members of the given remote group
    --disableRemoteGroup        Disable remote logging for remote users.
-c, --cluster                   The cluster scope for this command
```

Using a remote user group

You can use a remote user as a SONAS GUI and CLI user with a specific role.

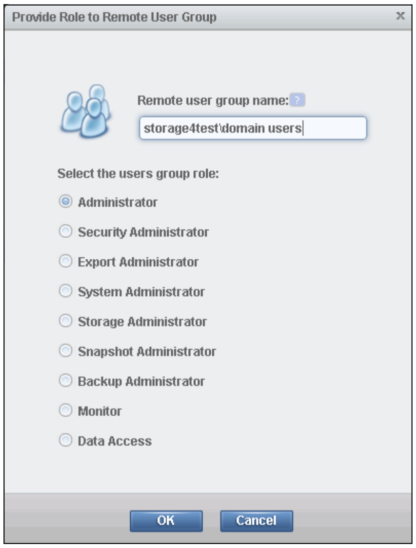
Using the GUI: Click **Access** → **Users** and click **Add User Group** → **Use Remote User Group**. See Figure 7-118.



Name	User Groups	Pass
admin	SecurityAdmin, Dataaccess	Active
admin2	Administrator, SecurityAdmin	Active
scpuser		Unset
test01	test	Expired

Figure 7-118 Select Use Remote User Group window

The Provide Role to Remote user Group window opens. Enter the storage4test\domain users group in the Remote user group name field and select the **Administrator** role. After you enter the information, click **OK** to proceed. See Figure 7-119.



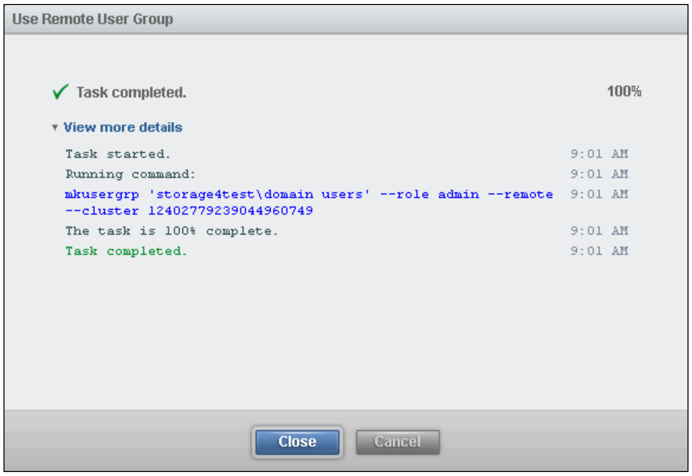
Remote user group name: storage4test\domain users

Select the users group role:

● Administrator
○ Security Administrator
○ Export Administrator
○ System Administrator
○ Storage Administrator
○ Snapshot Administrator
○ Backup Administrator
○ Monitor
○ Data Access

Figure 7-119 Provide Role to Remote Group window

The result of changing the Use Remote User Group setting is shown in Figure 7-120 on page 583.



✓ Task completed. 100%

▾ View more details

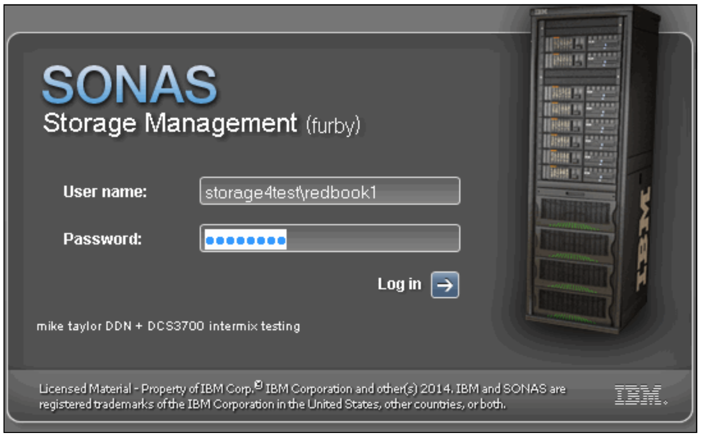
```
Task started.                                                9:01 AM
Running command:                                             9:01 AM
mkusergrp 'storage4test\domain users' --role admin --remote  9:01 AM
--cluster 12402779239044960749
The task is 100% complete.                                   9:01 AM
Task completed.                                              9:01 AM
```

Figure 7-120 Use Remote User Group result messages

Logging in to the GUI and CLI

You have a redbook1 user in the storage4test domain and this user is in the domain users group. Because all configuration was done in “Using a remote user group” on page 582, you can log in to the SONAS GUI and CLI with the redbook1 user.

Using the GUI: Connect to the management IP address by using a web browser. Enter storage4test\redbook1 in to the User name field and the password in to the Password field. See Figure 7-121.



User name: storage4test\redbook1

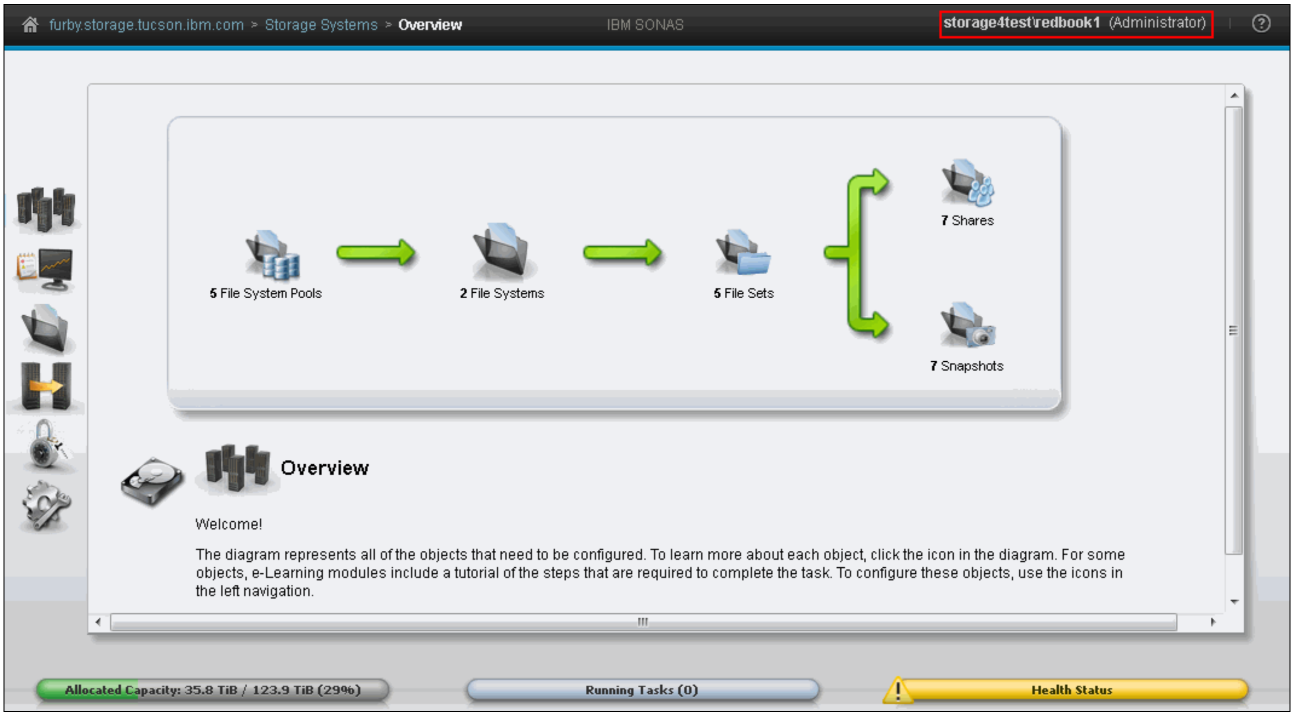
Password: ●●●●●●●●

mike taylor DDN + DCS3700 intermix testing

Licensed Material - Property of IBM Corp.© IBM Corporation and other(s) 2014. IBM and SONAS are registered trademarks of the IBM Corporation in the United States, other countries, or both.

Figure 7-121 SONAS GUI login window

You are connected as the storage4test\redbook1 user. The current user is shown at the upper right of the window. See Figure 7-122.


furby.storage.tucson.ibm.com > Storage Systems > Overview
IBM SONAS
storage4test\redbook1 (Administrator)
5 File System Pools
2 File Systems
5 File Sets
7 Shares
7 Snapshots
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Overview

Welcome!

The diagram represents all of the objects that need to be configured. To learn more about each object, click the icon in the diagram. For some objects, e-Learning modules include a tutorial of the steps that are required to complete the task. To configure these objects, use the icons in the left navigation.

Figure 7-122 Remote User redbook1 login window

Using the CLI: You can log in to the CLI with the same user name and password that you use for the GUI.

7.7.6 SONAS users

The SONAS users are the users who are accessing the data that is stored in the file system. They can write and read data. Data from the cluster can be accessed by the users only through the data exports. SONAS supports the CIFS, FTP, and NFS protocols.

To access the data by using the protocols, the users must authenticate. SONAS supports the Windows Active Directory authentication server and the LDAP server. To learn more about integrating the authentication server into the SONAS, see the SONAS information in the IBM Knowledge Center, found at the following website:

http://www-01.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/chooseauthenticationmethod.html?lang=en

NFS, as an exception, does not need users to authenticate because it checks for the authenticity of the client or hosts. The other protocols, such as FTP and CIFS, require that the users authenticate. CIFS authenticates with the Windows AD server while FTP works for both Windows AD users and LDAP.

Authentication is the process of verifying the identity of the user. Users confirm that they are indeed the users they are claiming to be. It is typically accomplished by verifying the user ID and password from the authentication server.

Authorization is the process of determining whether the users are allowed to access files. The users might have permission to access certain files but not others. This control is typically managed with ACLs.

The file system ACLs that are supported in SONAS are GPFS ACLs, which are NFSV4 ACLs. The directories and exports must be given the correct ACLs for the users to gain access. You can give the owner the rights or permissions to an export by specifying the owner option when create an export from both the GUI and the CLI. If you want to give other users access, modify the ACL file in GPFS for the directory or export by running **`mmeditacl`**. You can view ACLs by running **`mmgetacl`**.

> **ACLs:** SONAS V1.5 provides the **`lsacl`** and **`chacl`** ACL commands. You do not need to use GPFS commands to verify and modify ACLs. For more information, see 7.5.6, "Managing access control lists" on page 559.

Example 7-36 shows how you can provide ACLs to a directory or export.

Example 7-36 View current ACLs for an export by using the GPFS command mmgetacl

```
export EDITOR=/bin/vi

$ mmgetacl /ibm/gpfs0/Sales
#NFSv4 ACL
#owner:root
#group:root
special:owner@:rwxc:allow
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR (-)WRITE_NAMED

special:group@:r-x-:allow
 (X)READ/LIST (-)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (X)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR (-)WRITE_NAMED

special:everyone@:r-x-:allow
 (X)READ/LIST (-)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (X)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR (-)WRITE_NAMED
```

Example 7-37 adds ACLs for another user.

In the example, you are giving "READ/WRITE" access to the Windows AD user "David" for an existing export that is named "Sales" in the `/ibm/gpfs0` file system.

Example 7-37 Add an ACL for giving user DAVID access to the export

```
$ mmeditacl /ibm/gpfs0/Sales
#NFSv4 ACL
#owner:root
#group:root
special:owner@:rwxc:allow
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR (-)WRITE_NAMED

user:STORAGE3\david:rwxc:allow
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-READ_NAMED
 (-)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR (-)WRITE_NAMED

special:group@:r-x-:allow
 (X)READ/LIST (-)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (X)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR (-)WRITE_NAMED
```

```
special:everyone@:r-x-:allow
 (X)READ/LIST (-)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (X)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR (-)WRITE_NAMED
```

Save the file, and when you quit, click **Yes** when prompted to confirm the ACLs. The new ACLs are written for the user and the export.

Depending on the users to whom you want to give access, you can add them in the ACLs file. You can also give group access in a similar way as before and add users to the group.

7.8 Services management

This section describes the Management Service function and administration.

7.8.1 Overview

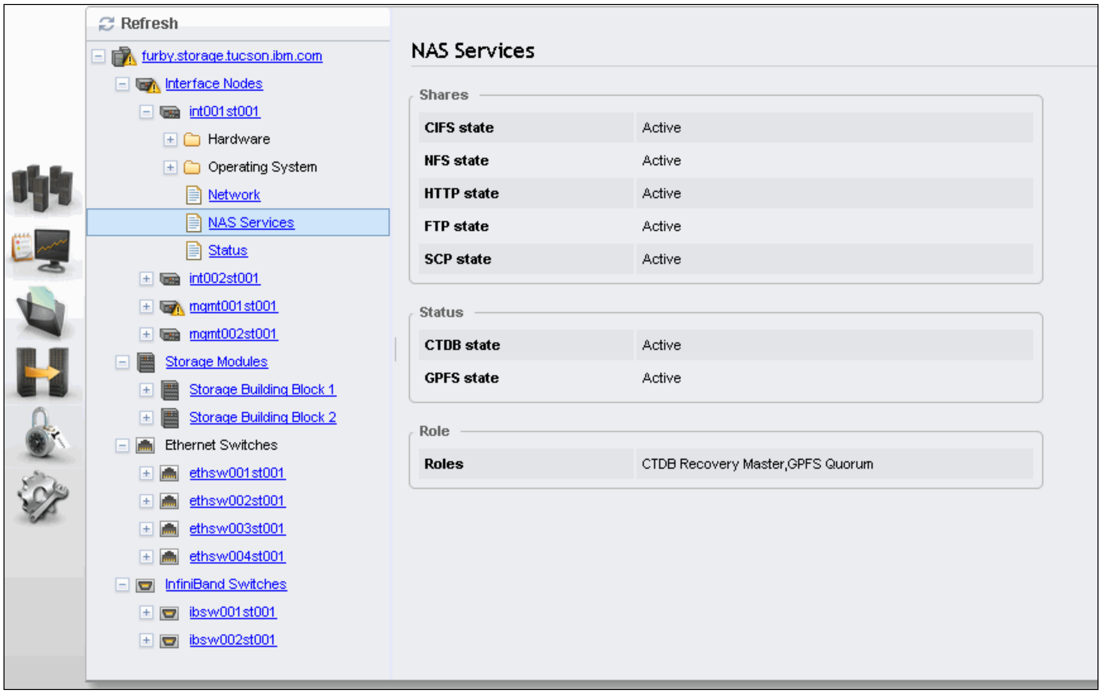
The services that are running on SONAS are CIFS, FTP, HTTP, NFS, and SCP. These services are needed for clients to access the SONAS data exports.

7.8.2 Managing services on the cluster

You can view the status of the services that are configured. You can also enable and disable them. They must be configured for you to do any operations on them. This section describes each task that can be carried out on the services.

- List the service status.

  Using the GUI: You can view running services in the GUI. Click **System Details**, select one of the Interface nodes, and click the **NAS Services** option. See Figure 7-123.



NAS Services

Shares

CIFS state	Active
NFS state	Active
HTTP state	Active
FTP state	Active
SCP state	Active

Status

CTDB state	Active
GPFS state	Active

Role

Roles	CTDB Recovery Master,GPFS Quorum

Figure 7-123 View services through the GUI

Using the CLI: You can list the services by running **lservice**. This command lists all of the services, the state of each service, and also whether it is configured. The command usage and output are shown in Example 7-38.

Example 7-38 Example for usage and command output for lsservice

```
[furby.storage.tucson.ibm.com]$ lsservice --help
usage: lsservice  [-r] [-Y] [-c < clusterID | clusterName >]
Lists all the services of a cluster.
Parameter     Description
-r, --refresh Forces a refresh of the exports data in the database by scanning cluster exports
before retrieving data from the database.
-Y            Shows parseable output.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$ lsservice
Name Description   Is active Is configured
FTP  FTP protocol  yes       yes
HTTP HTTP protocol yes       yes
NFS  NFS protocol  yes       yes
CIFS CIFS protocol yes       yes
SCP  SCP protocol  yes       yes
EFSSG1000I The command completed successfully.
```

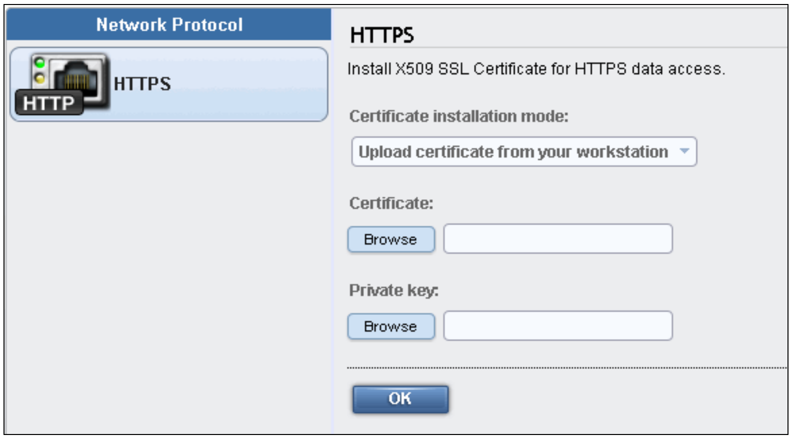
In the example, you can see that all the services are configured. This configuration means that all the configuration files for the services are up-to-date on each node of the cluster. Under the column “Is Active”, you can see whether the service is active or inactive. Active denotes that the service is up and running. Exports can be accessed by using that service. Users or clients can access the data that is exported by using that protocol or service. Inactive means that the service is not running and therefore all data connections break.

- Change the service configuration.

  Using the GUI: You can change the configuration for each configured service by using the GUI. As seen in step 1 on page 499, you can see the table that contains the list of services. Each of these services is a link that you can click. When you click a link, a new window opens, which you can use to change configuration parameters for the service.

  – FTP: FTP does not have any configuration parameters to modify.

  – HTTP: HTTP requires you to install an HTTP certificate or private key. You can install an existing certificate or generate one that is the same as the private key. To upload a certificate or private key, click **Settings** and select **Network Protocol**. See Figure 7-124.



HTTPS

Install X509 SSL Certificate for HTTPS data access.

Certificate installation mode: Upload certificate from your workstation

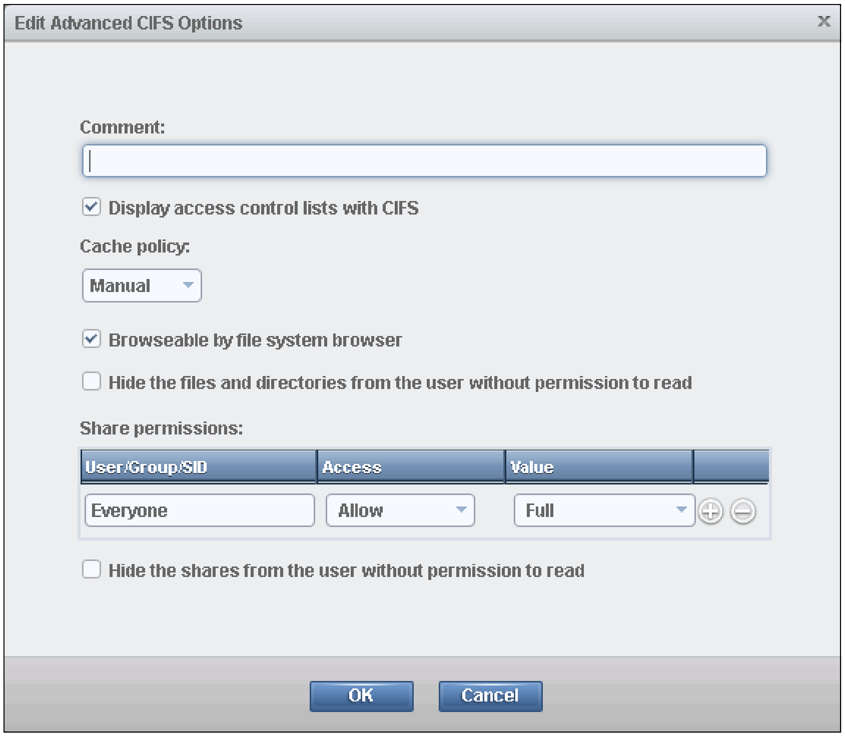
Certificate:

Private key:

Figure 7-124 HTTP Configuration window

  – NFS: The NFS menu is shown after you run **chnfsserver** with the **NFSv4.0** option. For more information, see 7.8.3, “Managing the NFS service” on page 589.

  – CIFS: As shown in Figure 7-125 on page 589, you can see the different parameters that you can change for CIFS. As you can see in the figure, you can change some common parameters and also some advanced options. Click **Apply** when you are done. The configuration is successfully written on all nodes.



Edit Advanced CIFS Options

Comment:

☑ Display access control lists with CIFS

Cache policy: Manual

☑ Browseable by file system browser

□ Hide the files and directories from the user without permission to read

Share permissions:

User/Group/SID	Access	Value
Everyone	Allow	Full

□ Hide the shares from the user without permission to read

Figure 7-125 CIFS configuration parameters

Using the CLI: You cannot change the configuration parameters from the SONAS CLI.

7.8.3 Managing the NFS service

SONAS V1.5 supports NFSv3 and NFSv4. The default setting is NFSv3, and this option supports only NFSv3 clients. You must change the option to connect NFSv4 clients. The default NFSv3 option uses kernel space. Otherwise, the NFSv4 option uses user space and supports NFSv3 clients at the same time. NFSv4 has more improvements for performance than NFSv3.

Enabling the NFS option

You need to enable the NFSv4 option to use NFSv4. However, there is no menu to enable this option on the GUI. Therefore, you must run **`chnfsserver`**. To find the current NFS server version, run **`lsnfsserver`**. See Example 7-39.

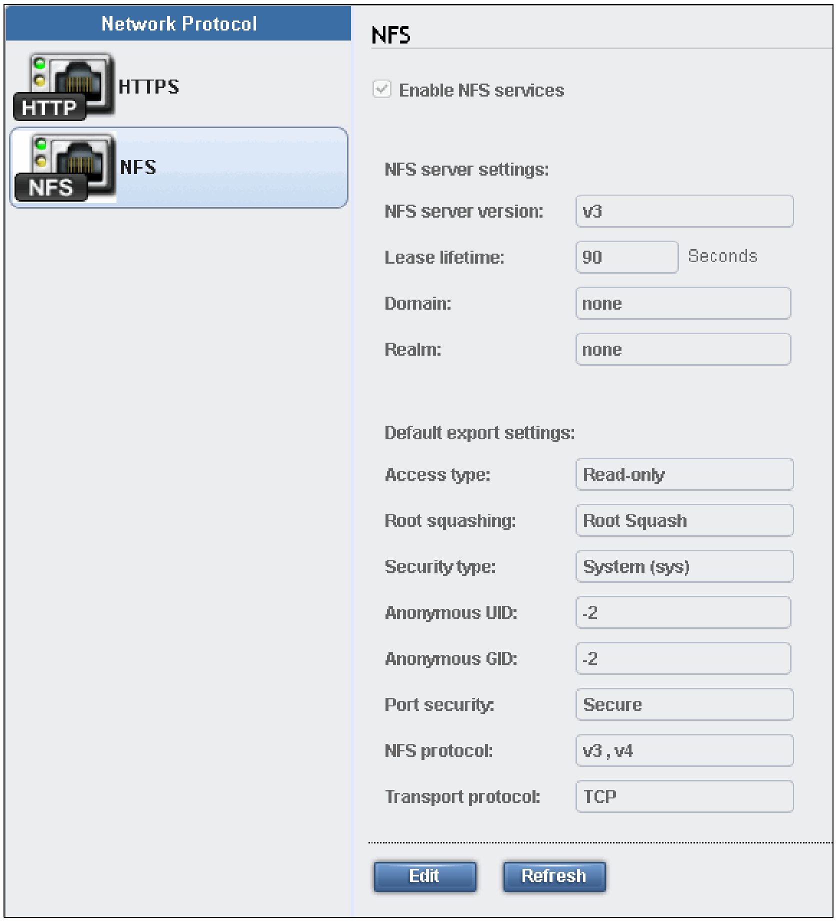
Example 7-39 Change the NFS options with commands

```
[furby.storage.tucson.ibm.com]$ lsnfsserver
NFS Server Stack
kNFSv3
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ chnfsserver NFSv4.0
EFSSG1108W WARNING: The exports created with current NFS stack will not be available with
new NFSv4.0 stack. It will be required to re-create the exports once the NFS stack is
switched. All NFS clients will have to remount the exports. It is recommended to delete any
existing ACE relationships before switching to new stack. It will be required to re-create
the ACE relationships once the NFS stack is switched.
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsnfsserver
```

```
NFS Server Stack
NFSv4.0
EFSSG1000I The command completed successfully.
```

If you want to change back to NFSv3, run **`chnfsserver kNFSv3`**.

After the NFSv4 option is enabled, the NFS menu is shown in the GUI. Click **Settings** → **Network Protocol**. See Figure 7-126.



NFS

☑ Enable NFS services

NFS server settings:

NFS server version:	v3	
Lease lifetime:	90	Seconds
Domain:	none	
Realm:	none	

Default export settings:

Access type:	Read-only
Root squashing:	Root Squash
Security type:	System (sys)
Anonymous UID:	-2
Anonymous GID:	-2
Port security:	Secure
NFS protocol:	v3 , v4
Transport protocol:	TCP

Figure 7-126 NFS service option menu

NFS settings

In NFSv4.0 mode, you can change many options, including the following ones:

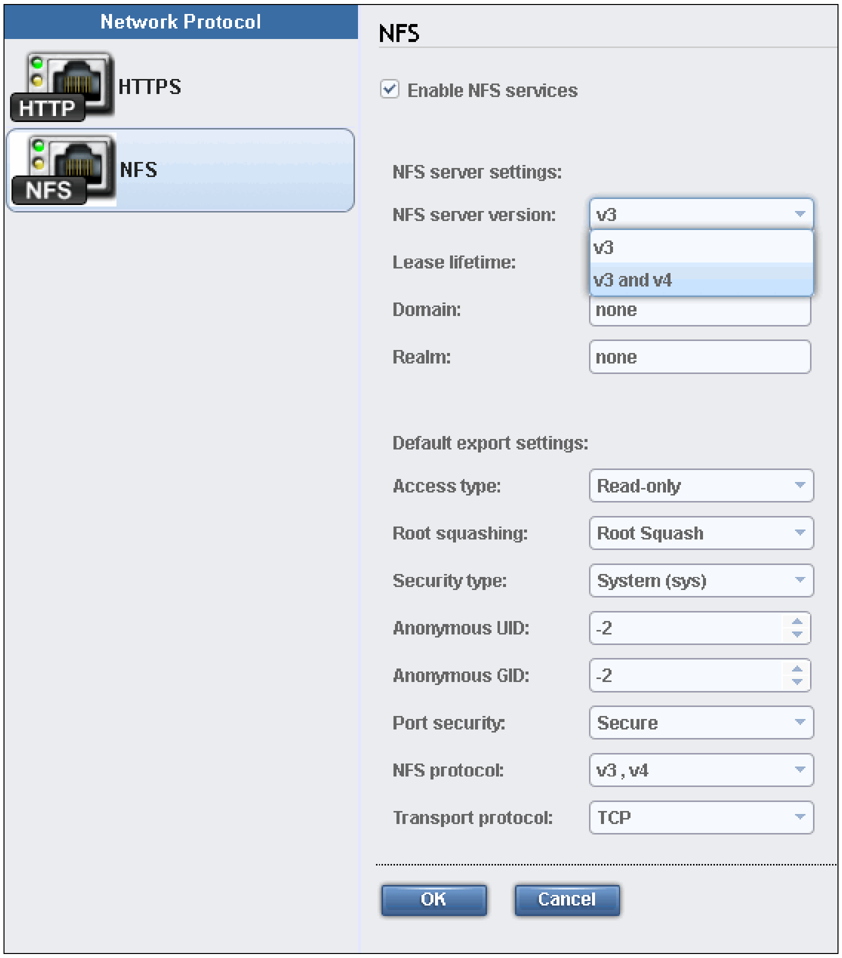
- Enable NFS services: This setting enables or disables the whole NFS service.
- NFS server version: Defines which service supports which clients.
- Lease lifetime: The grace period to reclaim the ownership of file locks by the clients during failover events and maximum idle time for a client. The valid range is 20 – 180. The default value is 90.
- Domain: The local NFSv4 domain name. The default value is none.
- Realm: The Kerberos realm name.

- Default export settings: When you create a share, these values are loaded as the default settings:
  - Access type: Read-only and Read-write.
  - Root Squashing: Root Squash, No Root Squash, and All Squash. For more information, see Table 1-9 on page 31.
  - Security type: Specify the type of security to use to authenticate an NFS connection. The default value is System. When System is selected, the UNIX UIDs and GIDs are used to authenticate users. Kerberos can also be used for stricter authentication and data protection
  - Anonymous UID: The anonymous UID for the root user when its host does not have a root access. The default value is -2.
  - Anonymous GID: The anonymous GID for the root user group when its host does not have a root access. The default value is -2.
  - Port Security: With this feature enabled, the system prevents access to requests that originate from ports where the port number is greater than the hardcoded threshold value of 1024.
  - NFS protocol: Version of the NFS protocol that can be used to access exports. The default value is v3.
  - Transport protocol: Network protocols that can be used for accessing the exports.

Changing the NFS settings

You must enable NFSv4 mode before completing the steps in this section. For more information, see “Enabling the NFS option” on page 589.

Using the GUI: Click **Settings** → **Network Protocol**. Click **Edit** at the lower right of the window to change the options. See Figure 7-126 on page 590. Change the **NFS server version** option from v3 to v3 and v4. See Figure 7-127.



NFS

☑ Enable NFS services

NFS server settings:

NFS server version: v3 (v3 / v3 and v4)

Lease lifetime:

Domain: none

Realm: none

Default export settings:

Access type: Read-only

Root squashing: Root Squash

Security type: System (sys)

Anonymous UID: -2

Anonymous GID: -2

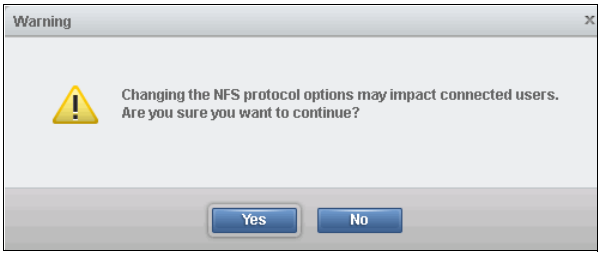
Port security: Secure

NFS protocol: v3 , v4

Transport protocol: TCP

Figure 7-127 Edit the NFS options

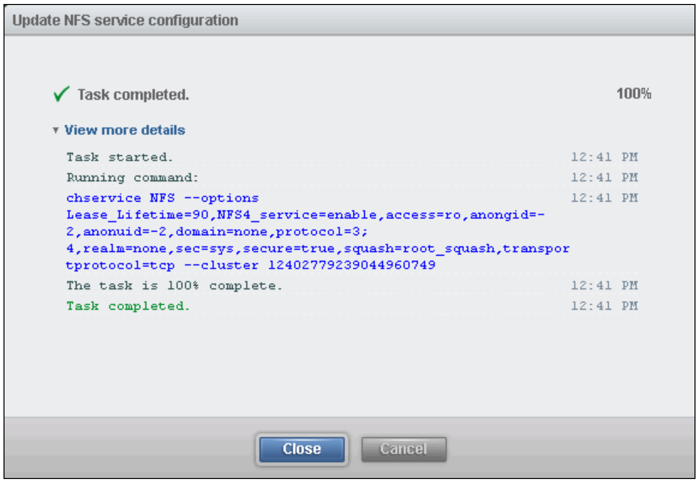
After you change the options, click **OK** to apply the changed values, as shown in Figure 7-128.



Changing the NFS protocol options may impact connected users.
Are you sure you want to continue?

Figure 7-128 Change the NFS server warning message

The result is shown in the Update NFS service configuration window. See Figure 7-129 on page 593.



```
Task started.                                              12:41 PM
Running command:                                           12:41 PM
chservice NFS --options                                    12:41 PM
Lease_Lifetime=90,NFS4_service=enable,access=ro,anongid=-
2,anonuid=-2,domain=none,protocol=3;
4,realm=none,sec=sys,secure=true,squash=root_squash,transpor
tprotocol=tcp --cluster 12402779239044960749
The task is 100% complete.                                 12:41 PM
Task completed.                                            12:41 PM
```

Figure 7-129 Edit NFS Result

Using the CLI: You can also change and check these NFS options with the **chservice** and **lsservice** commands. Change the **NFS4_service** option to disable it, as shown in Example 7-40. This is the same task as for changing the NFS server version from v3 and v4 to v3 with the GUI.

Example 7-40 Change the NFS settings with commands

```
[furby.storage.tucson.ibm.com]$ lsservice --protocoloptions
CIFS
=====
serverDescription : "IBM NAS"
diskFreeQuota : yes
NFS
=====
Lease_Lifetime : 90
domain : none
realm : none
NFS4_service : enable
anongid : -2
anonuid : -2
access : ro
squash : root_squash
sec : sys
secure : true
transportprotocol : tcp
protocol : 3;4
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ chservice NFS --options NFS4_service=disable
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsservice --protocoloptions
CIFS
=====
serverDescription : "IBM NAS"
diskFreeQuota : yes
NFS
=====
Lease_Lifetime : 90
domain : none
realm : none
NFS4_service : disable
```

```
anongid : -2
anonuid : -2
access : ro
squash : root_squash
sec : sys
secure : true
transportprotocol : tcp
protocol : 3;4
EFSSG1000I The command completed successfully.
```

7.9 Scheduling tasks in SONAS

SONAS allows you to schedule some tasks that can be done without any manual intervention. The administrative tasks are issued on the Management node by using either the GUI or the CLI. Cron jobs are scheduled by the cron job scheduler of the underlying operating system. Cron job tasks are used to run administrative operations, such as creating snapshots or triggering an asynchronous replication process. The cron jobs are pre-configured as task templates. An existing cron job task template must be used to create a new task.

Here are the cron job task specifics:

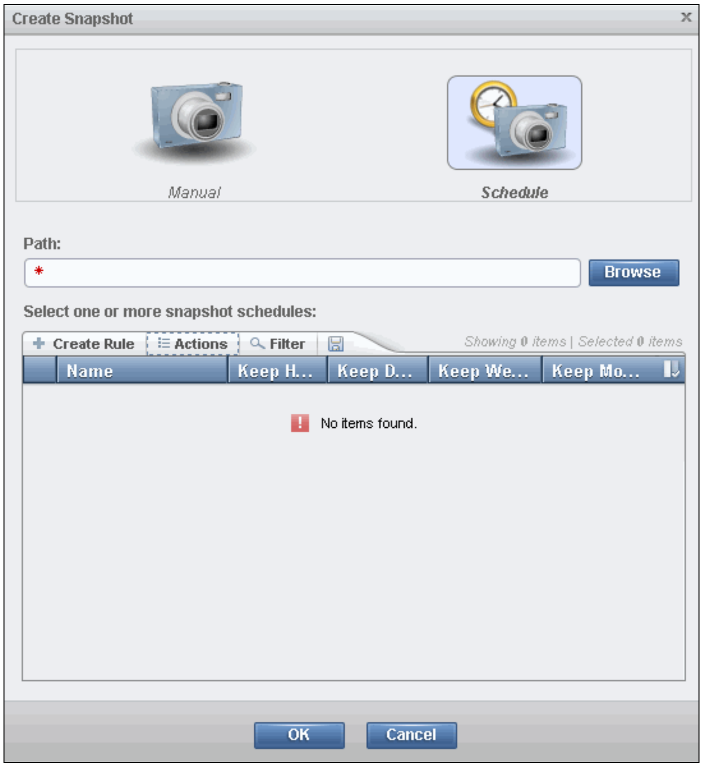
- A cron job task can be run either on all nodes or on one node only, that is, the recovery master node.
- A cron job task can be configured multiple times with different parameters.

The execution of a scheduled task can be displayed by the CLI or GUI.

Here are two examples for configuring a cron job in the GUI:

- Creating a snapshot task

  Click **Files** → **Snapshots** and select **Create Snapshot**. A window opens, as shown in Figure 7-130 on page 595.



Path:

Select one or more snapshot schedules:

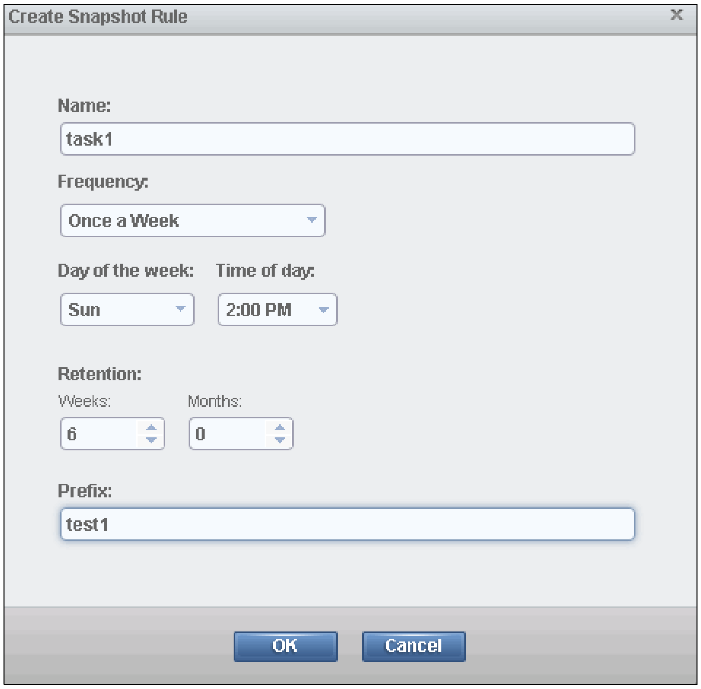
Showing 0 items | Selected 0 items

Name	Keep H...	Keep D...	Keep We...	Keep Mo...

No items found.

Figure 7-130   Create a snapshot

Select **Schedule** and enter the required path. Then, select **Create Rule**. A new window opens, as shown in Figure 7-131. Enter information about the snapshot rule, including the name of the rule, and how often the rule applies (for example, once per week or once per day). You can also specify day of the week to run and what time of day. The retention field provides information about how long this rule applies.



Name: task1

Frequency: Once a Week

Day of the week: Sun

Time of day: 2:00 PM

Retention:

Weeks: 6

Months: 0

Prefix: test1

Figure 7-131   Create a snapshot rule

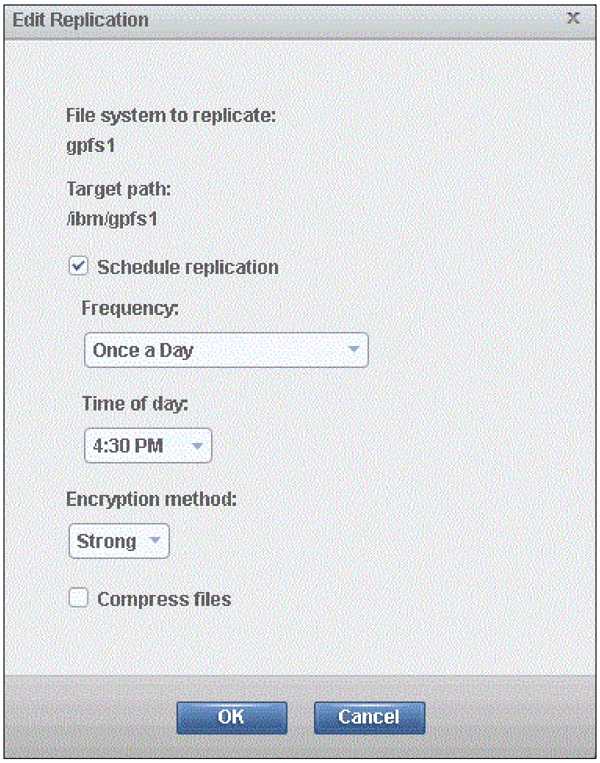
For this example, create a rule that is named task1. The frequency is once per week, and it is run every Sunday at 2:00 p.m. with a retention policy of six weeks.

The same process can be done with the CLI by running the following command:

```
mksnaprule task1 --frequency weekly --minutesAfterHour 0 --hour 14 --dayOfWeek
0 --maxWeeks 6 --maxMonths 0 --cluster 12402779243267445246
```

- Creating a replication rule

  This step can be done during the creation of replication task, or you can modify an existing replication task. Because replication is addressed elsewhere, this section shows only how to create a task for an existing replication. Under Copy Services, click **Replication**. A list of already created replications is shown. Select one and click **Actions** → **Edit**. The window that is shown in Figure 7-132 opens.



File system to replicate:
gpfs1

Target path:
/ibm/gpfs1

☑ Schedule replication

Frequency:
Once a Day

Time of day:
4:30 PM

Encryption method:
Strong

□ Compress files

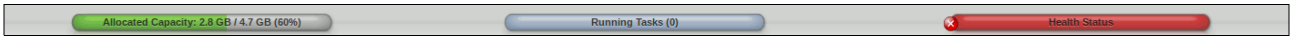
Figure 7-132 Edit Replication window.

As shown in Figure 7-132, you created a replication task that is run every day at 3:00 AM. You also set an encryption method to *strong*.

7.9.1 Listing tasks

This section describes how to list tasks by using both the GUI and the CLI.

Using the GUI: For now, you cannot see the defined tasks in the GUI. You can see tasks only when they are running. You can also configure a notification for when tasks finish. The Running tasks icon is at the bottom center of the GUI, as shown in Figure 7-133.



Allocated Capacity: 2.8 GB / 4.7 GB (60%)

Running Tasks (0)

Figure 7-133 Bottom pane of the GUI where the center icon represents running tasks

Using the CLI: The tasks can be scheduled by running **mktask**. The command takes some input values, such as cluster name, seconds, minutes, hours, and other time values for the task to run. There is an option that is called **parameter**. The **parameter** option is optional and is valid only for a cron task. The GUI tasks do not have any parameters. An error is returned to the caller if this option is denoted for a GUI task. The **parameter** variable is a space-separated parameter. Example 7-41 shows the help for the **cron** commands. You can also use this information when you use the GUI.

The following cron tasks are available in SONAS:

- **MkSnapshotCron**:

  Parameter: The cron job expects two parameters in the following order:

  - **clusterName**: The name of the cluster to which the file system belongs.
  - **filesystem**: The file system description (for example, `/gpfs/office`).

- **StartReplCron**:

  Parameter: The cron job expects two parameters in the following order:

  - **source_path**: The directory that is replicated.
  - **target_path**: The directory to which the data is copied.

- **StartBackupTSM**:

  Parameter: The cron job expects one parameter:

  clusterName: The cluster of the file systems that must be backed up.

- **StartReconcileHSM**:

  Parameter: The cron job expects three parameters in the following order:

  - **clusterName**: The cluster of the file systems that must be backed up.
  - **filesystem**: The file system to be reconciled.
  - **node**: The node on which the file system is to be reconciled.

- BackupTDB:

  Parameter: The cron job expects one parameter:

  target_path: The directory to which the backup is copied

For more information about how to add these parameters for these cron tasks, see the man page for the **mktask** command.

Example 7-41 shows how to add the **MkSnapshotCron** task. This task is a cron task, which takes two parameters: **clustername** and **filesystem**. For this example, you have a cluster that is named `furby.storage.tucson.ib.com` and a file system that is named `gpfs0`.

In the second example in Example 7-41, you add a task that is a GUI task. The last example is the help output for the **lstask** command, which is used to list tasks in the CLI.

Example 7-41 Command usage and output in adding cron and GUI tasks by running mktask

```
[furby.storage.tucson.ibm.com]$ mktask --help
usage: mktask taskName  [--second <secondDef> | --month <monthDef>] [--minute <minuteDef>] [--hour <hourDef>] [--dayOfWeek <dayOfWeekDef>]
[--dayOfMonth <dayOfMonthDef>] [-p <parameterDef>] [-c < clusterID | clusterName >]
Schedules a GUI or a cron task on the selected cluster that belongs to the management system.

Parameter        Description
taskName         Names the task to be run.
    --second     Defines at what second the scheduled task will be run.
    --minute     Defines at what minute the scheduled task will be run.
    --hour       Defines at what hour the scheduled task will be run.
    --dayOfWeek  Defines a day of the week on which the scheduled task will be run.
    --dayOfMonth Defines a day of the month on which the scheduled task will be run.
    --month      Defines in which month the scheduled task will be run.
-p, --parameter  Defines the parameter for the task to be scheduled.
```

```
-c, --cluster    The cluster scope for this command

[furby.storage.tucson.ibm.com]$ mktask MkSnapshotCron --parameter "furby.storage.tucson.ibm.com gpfs0" --minute 10 --hour 2 --dayOfMonth
*/3
EFSSG0019I The task MkSnapshotCron has been successfully created.

[furby.storage.tucson.ibm.com]$ mktask FTP_REFRESH  --minute 2 --hour 5 --second 40
EFSSG0019I The task FTP_REFRESH has been successfully created.

[furby.storage.tucson.ibm.com]$ lstask --help
usage: lstask  [-t <taskType>] [-s] [-v] [-r] [-Y] [-c < clusterID | clusterName >]
Lists the scheduled tasks that belong to a Management node for the selected cluster.
Parameter     Description
-t, --type    Selects the type tasks to be listed.
-s, --short   Prints out a list with these column details: Name, Status, Last run and Runs on.
-v, --verbose Shows additional columns.
-r, --refresh Forces a refresh of the exports data in the database by scanning cluster exports before retrieving data from the database.
-Y            Shows parseable output.
-c, --cluster The cluster scope for this command
```

7.9.2 Removing tasks

Using the GUI: You can remove the task by selecting the task from the table of tasks and clicking **Delete**. The operation opens a window, which asks for confirmation, as shown in Figure 7-134.

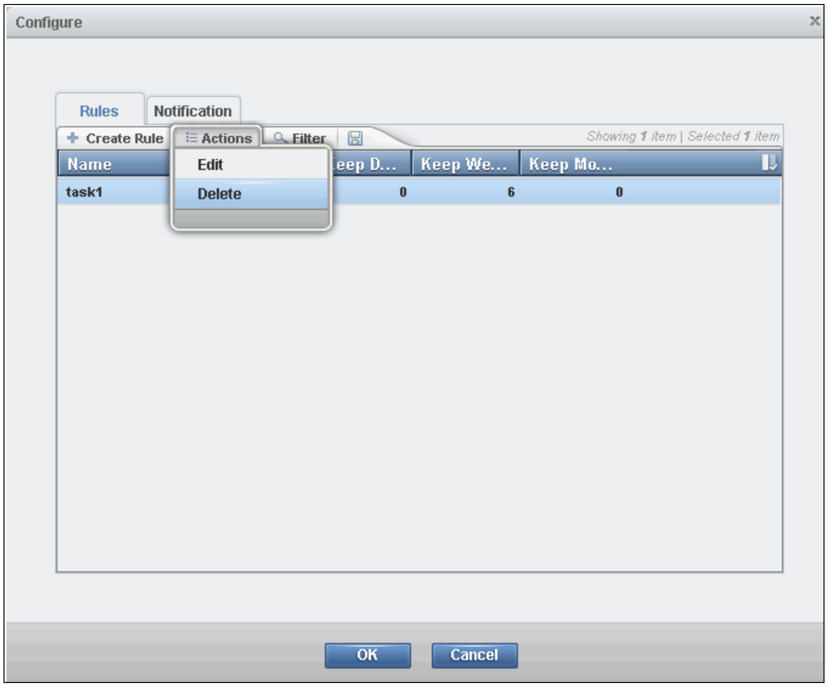


Figure 7-134 Removing snapshot tasks in the GUI

When you click **Delete**, a window opens for confirmation. Click **OK** to confirm.

Using the CLI: You can remove the task by running **`rmtask`**. This command deletes the command from the list of tasks to be scheduled by the system. An error is returned to the caller if a task that does not exist is denoted. The command usage and output is shown in Example 7-42 on page 599. In the first example, you delete a cron task that you added that is called MkSnapshotCron. In the second example, you delete the GUI task FTP_REFRESH.

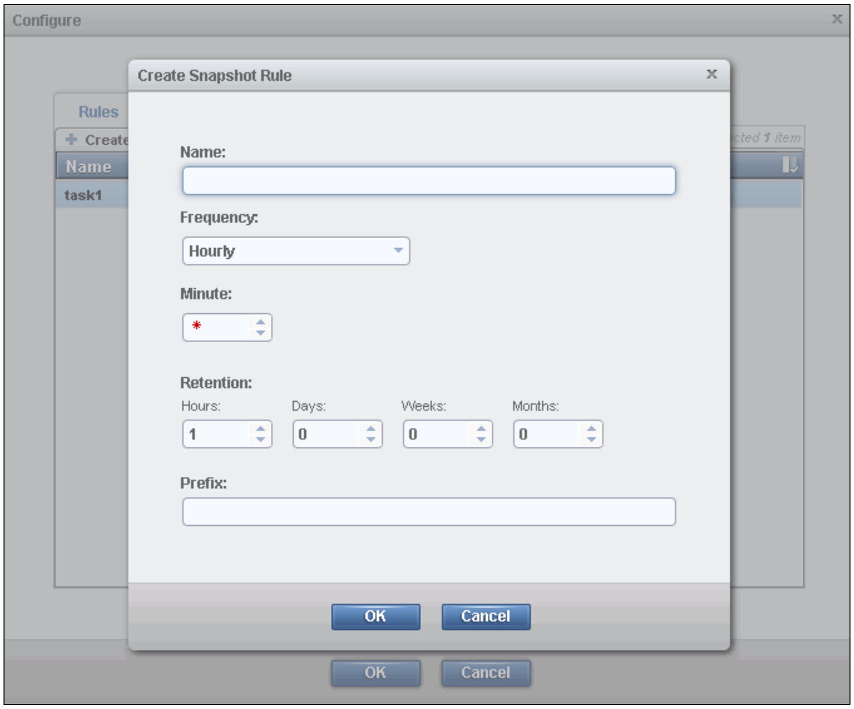
Example 7-42 Example of the rmtask command

```
[furby.storage.tucson.ibm.com]$ rmtask --help
usage: rmtask taskName  [-c < clusterID | clusterName >]
Removes a scheduled task that belongs to the Management node on the selected cluster.

Parameter     Description
taskName      Identifies a scheduled task that will be removed from the list of scheduled tasks.
-c, --cluster The cluster scope for this command
```

7.9.3 Modifying the schedule tasks

This section describes how to modify the scheduled tasks.

Using the GUI: You can modify a task from the GUI. Select the task from the appropriate navigational menu. To modify a snapshot task, click **Files** → **Snapshots**, select the snapshot that you want, click **Actions**, and select **Configure**. A new Configure window opens, as shown in Figure 7-135. You can modify the periods of taking snapshots, such as creating a new one, and you can add multiple rules to one snapshot. Figure 7-135 shows how to create a snapshot rule.



Frequency: Hourly

Minute: *

Retention:
Hours: 1
Days: 0
Weeks: 0
Months: 0

Figure 7-135 Modify snapshots rules

You can also create notifications for a task to receive an email when a task finishes, and to know whether it finished successfully.

7.10 Health Center

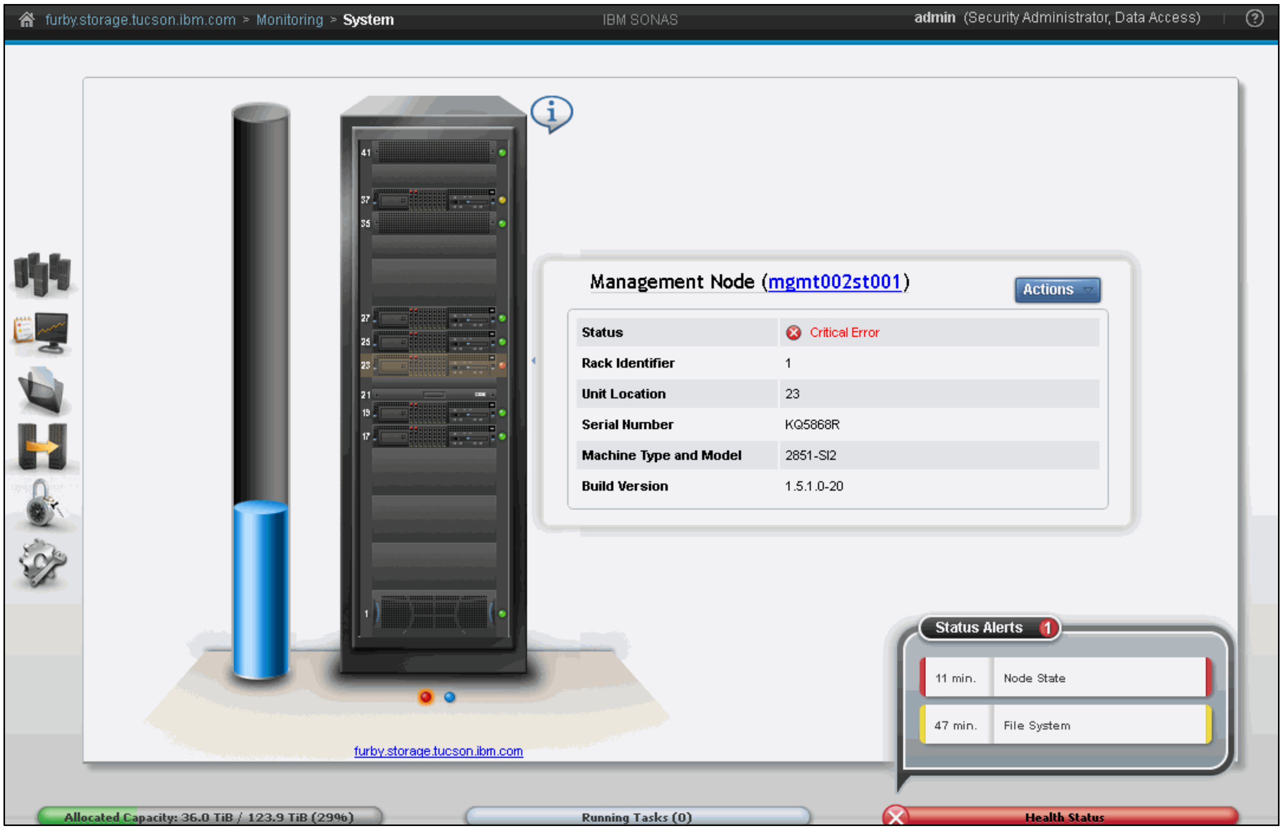
On the Monitoring menu in the GUI, you can see basic details about the SONAS system. You can check for system health, node health, or a specific component in the SONAS cluster.

7.10.1 Topology

From the GUI, in the monitoring section, click **System** and a model of the SONAS system appears. Next to the system model on the left side, a bar represents the allocated capacity. In the lower right corner, a health status is available. This health status is CLI command visible through the GUI navigation.

Overview

The overview view, as shown in Figure 7-136, provides a basic picture of your system. For more details about a specific area, you can use this window or click **Monitoring** → **System Details**.


furby.storage.tucson.ibm.com > Monitoring > System
IBM SONAS
admin (Security Administrator, Data Access)
Actions
furby.storage.tucson.ibm.com
Status Alerts 1
Allocated Capacity: 36.0 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Management Node (mgmt002st001)

Status	Critical Error
Rack Identifier	1
Unit Location	23
Serial Number	KQ5868R
Machine Type and Model	2851-SI2
Build Version	1.5.1.0-20

11 min.	Node State
47 min.	File System

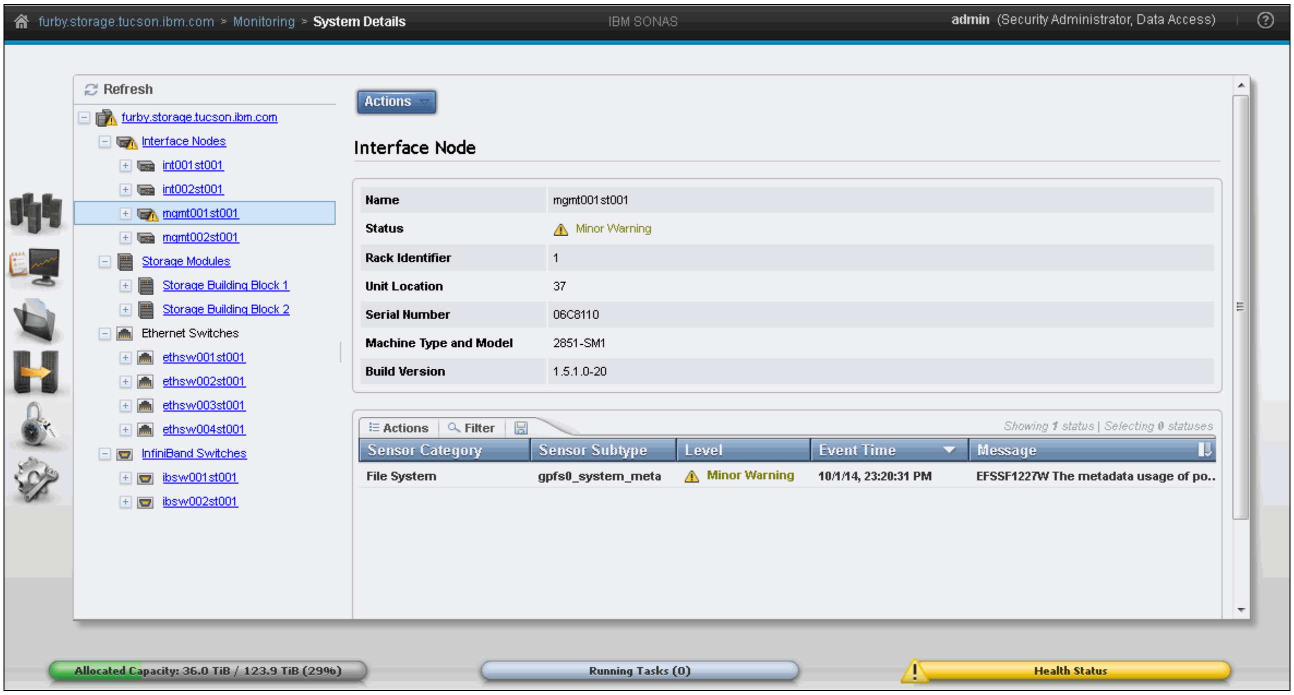
Figure 7-136 SONAS basic health check view

In this view, all components of your SONAS system can be seen. When you move the cursor over the selected component or click an item, you see a window with a more detailed view of the health for a selected item. For more information about one of these components, click the appropriate link. In lower right corner, the health status is showing errors and warnings. When you click these messages, you are redirected to the event view. When you click a hardware component, basic information is shown. Again, when you click an error message, you are redirected to a more detailed health view.

Basic health checks

To display a detailed health state window, click **Monitoring** → **System Details**. From this window, you can check the detailed health status for each cluster component. It applies to hardware and software cluster components. When you click a node of any type and click **Actions** → **Identify**, the LED on the selected node starts to flash (see Figure 7-137).

For more information about health checks and troubleshooting, see Chapter 9, “Troubleshooting, hints, and tips” on page 655.



furby.storage.tucson.ibm.com > Monitoring > System Details   IBM SONAS   admin (Security Administrator, Data Access)

Refresh

- furby.storage.tucson.ibm.com
  - Interface Nodes
    - int001st001
    - int002st001
    - mgmt001st001
    - mgmt002st001
  - Storage Modules
    - Storage Building Block 1
    - Storage Building Block 2
  - Ethernet Switches
    - ethsw001st001
    - ethsw002st001
    - ethsw003st001
    - ethsw004st001
  - InfiniBand Switches
    - ibsw001st001
    - ibsw002st001

Actions

Interface Node

Name	mgmt001st001
Status	⚠ Minor Warning
Rack Identifier	1
Unit Location	37
Serial Number	06C8110
Machine Type and Model	2851-SM1
Build Version	1.5.1.0-20

Actions | Filter   *Showing 1 status | Selecting 0 statuses*

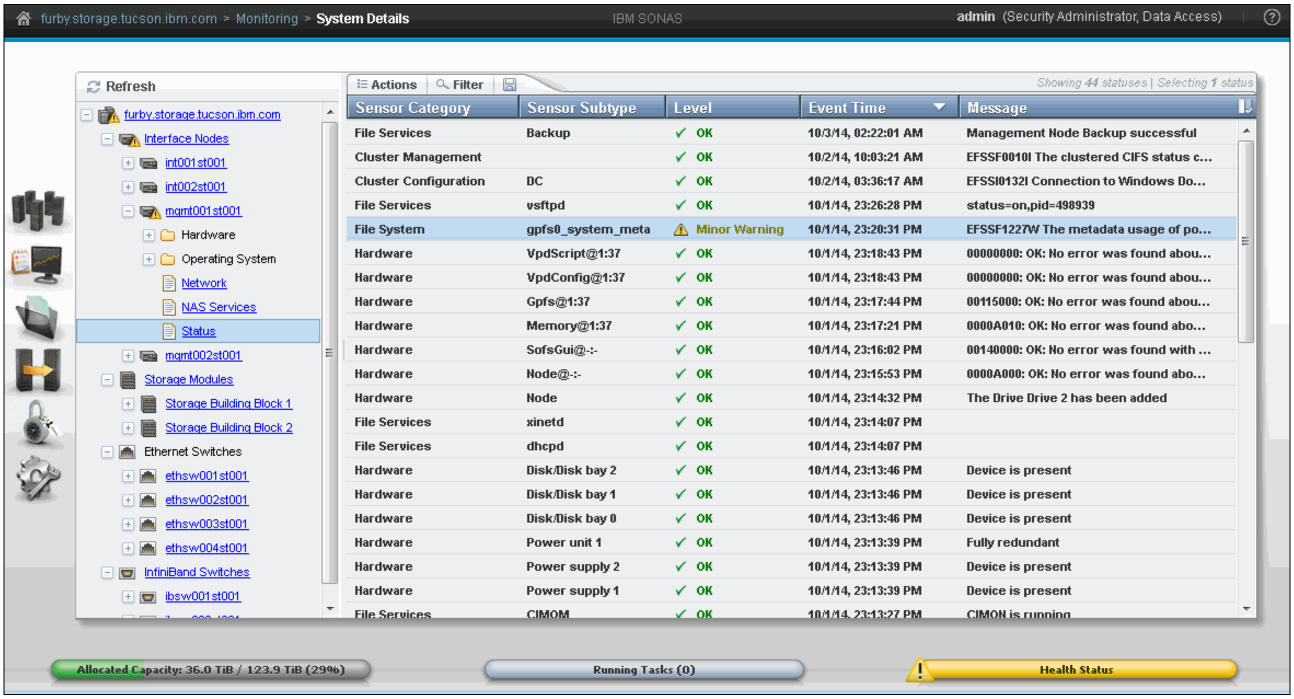
Sensor Category	Sensor Subtype	Level	Event Time	Message
File System	gpfs0_system_meta	⚠ Minor Warning	10/1/14, 23:20:31 PM	EFSSF1227W The metadata usage of po..

Allocated Capacity: 36.0 TiB / 123.9 TiB (29%)   Running Tasks (0)   Health Status

Figure 7-137 Basic node view with an option to identify in System Details

Interface nodes and Management nodes

The Status view displays the health of the selected node, as shown in Figure 7-138. The CTDB component runs only on the Interface and Management node types.



Sensor Category	Sensor Subtype	Level	Event Time	Message
File Services	Backup	✓ OK	10/3/14, 02:22:01 AM	Management Node Backup successful
Cluster Management		✓ OK	10/2/14, 10:03:21 AM	EFSSF0010I The clustered CIFS status c...
Cluster Configuration	DC	✓ OK	10/2/14, 03:36:17 AM	EFSSI0132I Connection to Windows Do...
File Services	vsftpd	✓ OK	10/1/14, 23:26:28 PM	status=on,pid=498939
File System	gpfs0_system_meta	⚠ Minor Warning	10/1/14, 23:20:31 PM	EFSSF1227W The metadata usage of po...
Hardware	VpdScript@1:37	✓ OK	10/1/14, 23:18:43 PM	00000000: OK: No error was found abou...
Hardware	VpdConfig@1:37	✓ OK	10/1/14, 23:18:43 PM	00000000: OK: No error was found abou...
Hardware	Gpfs@1:37	✓ OK	10/1/14, 23:17:44 PM	00115000: OK: No error was found abou...
Hardware	Memory@1:37	✓ OK	10/1/14, 23:17:21 PM	0000A010: OK: No error was found abo...
Hardware	SofsGui@-:-	✓ OK	10/1/14, 23:16:02 PM	00140000: OK: No error was found with ...
Hardware	Node@-:-	✓ OK	10/1/14, 23:15:53 PM	0000A000: OK: No error was found abo...
Hardware	Node	✓ OK	10/1/14, 23:14:32 PM	The Drive Drive 2 has been added
File Services	xinetd	✓ OK	10/1/14, 23:14:07 PM	
File Services	dhcpd	✓ OK	10/1/14, 23:14:07 PM	
Hardware	Disk/Disk bay 2	✓ OK	10/1/14, 23:13:46 PM	Device is present
Hardware	Disk/Disk bay 1	✓ OK	10/1/14, 23:13:46 PM	Device is present
Hardware	Disk/Disk bay 0	✓ OK	10/1/14, 23:13:46 PM	Device is present
Hardware	Power unit 1	✓ OK	10/1/14, 23:13:39 PM	Fully redundant
Hardware	Power supply 2	✓ OK	10/1/14, 23:13:39 PM	Device is present
Hardware	Power supply 1	✓ OK	10/1/14, 23:13:39 PM	Device is present
File Services	CIMOM	✓ OK	10/1/14, 23:13:27 PM	CIMON is running

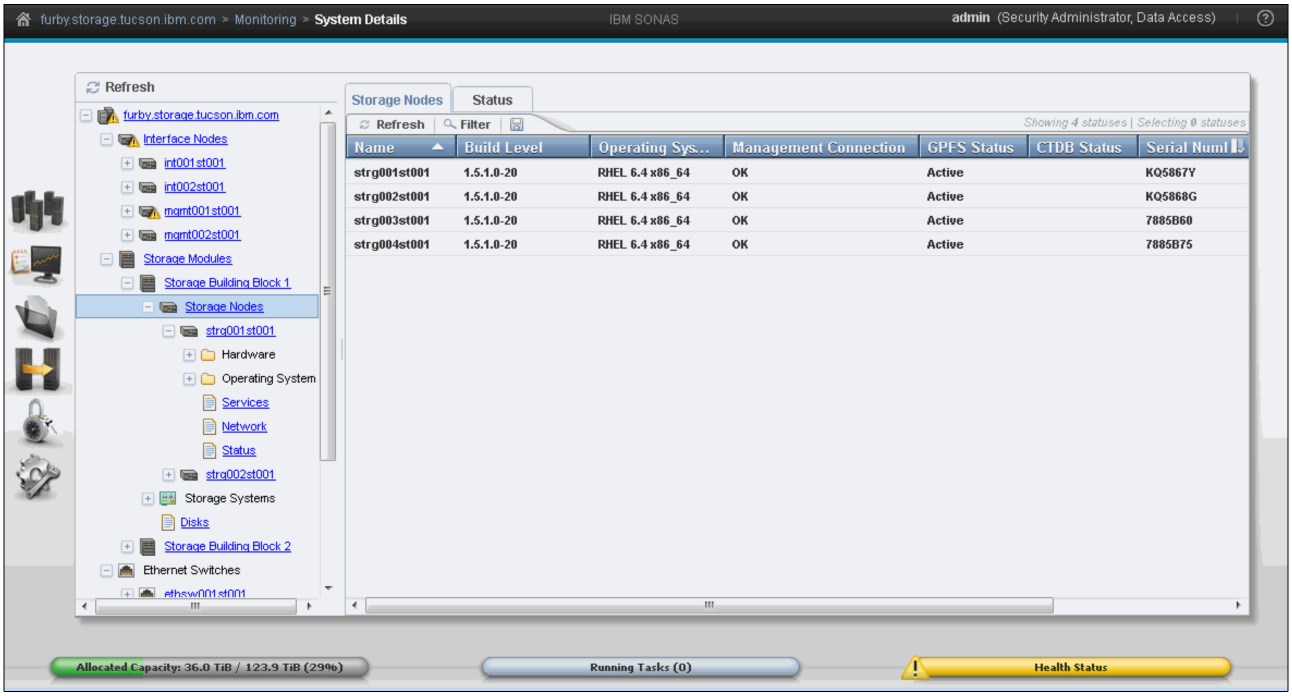
Figure 7-138 Detailed health status view for Management node 1

Views are categorized into different areas:

- Hardware:
  - Motherboard
  - CPU
  - Fan
  - HDD
  - Memory Modules
  - Power
  - Network Cards
- Operating System:
  - Computer System Details
  - Operating System Details
  - Local File System
- Network
- NAS Services
- Status

Storage nodes

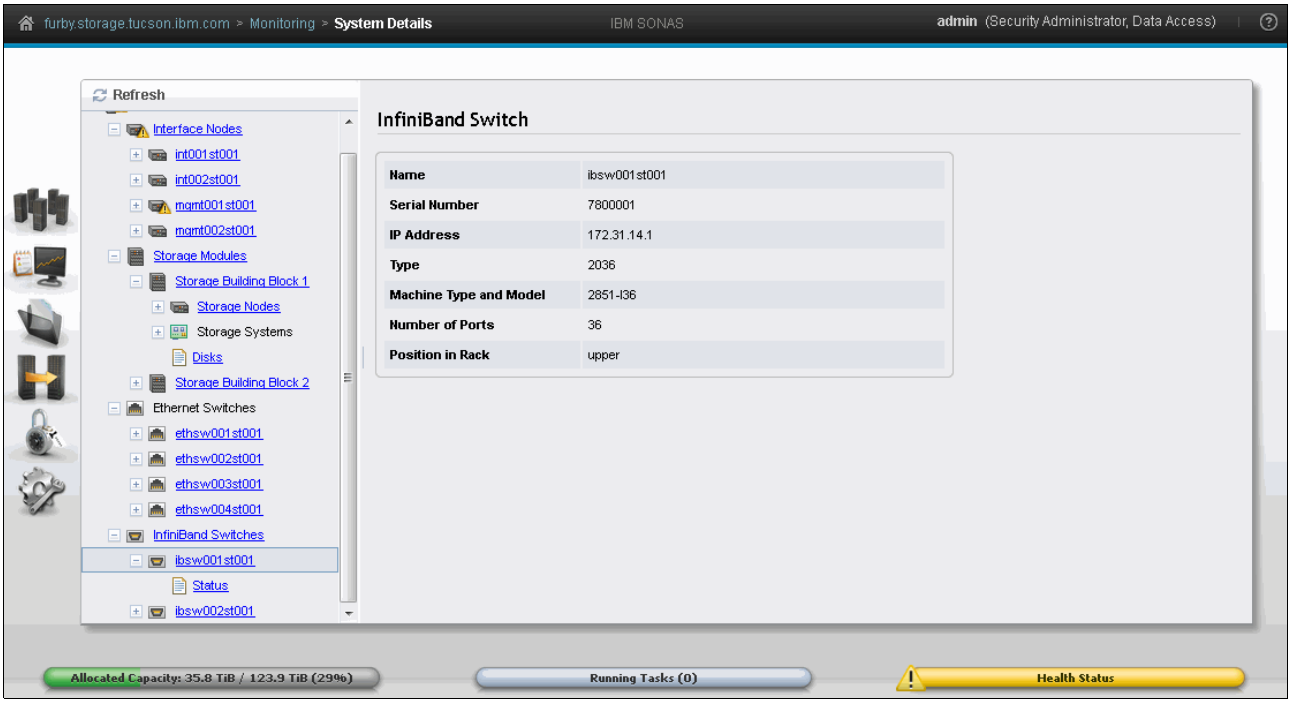
Different from the Interface node and Management node views, this view is categorized in Storage building blocks. It depends on the SONAS system size. There is a minimum of one storage building block that is present in all SONAS installations. Another difference is the status of the back-end storage subsystem and health view of the disks, as shown in Figure 7-139. In this case, SONAS has two building blocks.



Name	Build Level	Operating Sys...	Management Connection	GPFS Status	CTDB Status	Serial Numl
strg001st001	1.5.1.0-20	RHEL 6.4 x86_64	OK	Active		KQ5867Y
strg002st001	1.5.1.0-20	RHEL 6.4 x86_64	OK	Active		KQ5868G
strg003st001	1.5.1.0-20	RHEL 6.4 x86_64	OK	Active		7885B60
strg004st001	1.5.1.0-20	RHEL 6.4 x86_64	OK	Active		7885B75

Figure 7-139 Storage Nodes section

In this view, you can also check disk states and basic information about internal Ethernet and InfiniBand switches (see Figure 7-140).



InfiniBand Switch

Name	ibsw001st001
Serial Number	7800001
IP Address	172.31.14.1
Type	2036
Machine Type and Model	2851-I36
Number of Ports	36
Position in Rack	upper

Figure 7-140 InfiniBand switch basic health status

From the information that is collected here, you can do basic troubleshooting if there are any errors. For more detailed problem determination, you must collect a cndump for IBM Support.

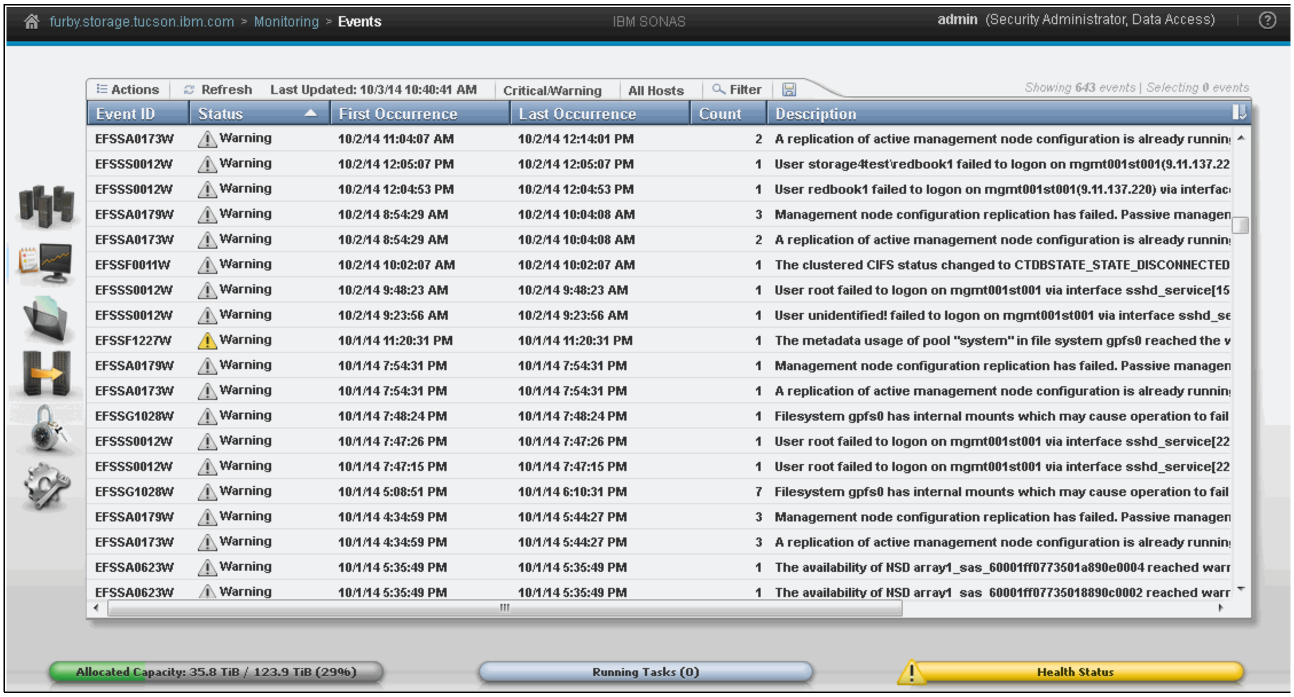
7.10.2 Event logs

Event logs are composed of two kinds of logs: Alert and System logs. The Red Hat Linux operating system reports all internal information, events, issues, or failures. Each Interface and Storage node has its own syslog file. In SONAS, all nodes send their syslog files to the Management node, which consolidate all these files and displays them in the event log, which is available from the GUI. It is a *raw* display of these files with some filtering tools. as shown in Figure 7-141 on page 605 in the upper right corner.

You can filter the view based on three different types of event logs:

- Show all
- Current Critical/Warning
- Critical/Warning.

You can also filter the event log to show events from a specific node or for the entire cluster.


furby.storage.tucson.ibm.com > Monitoring > Events
IBM SONAS
admin (Security Administrator, Data Access)
Actions
Refresh
Last Updated: 10/3/14 10:40:41 AM
Critical/Warning
All Hosts
Filter
Showing 643 events | Selecting 0 events
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Event ID	Status	First Occurrence	Last Occurrence	Count	Description
EFSSA0173W	Warning	10/2/14 11:04:07 AM	10/2/14 12:14:01 PM	2	A replication of active management node configuration is already runnin
EFSSS0012W	Warning	10/2/14 12:05:07 PM	10/2/14 12:05:07 PM	1	User storage4test\redbook1 failed to logon on mgmt001st001(9.11.137.22
EFSSS0012W	Warning	10/2/14 12:04:53 PM	10/2/14 12:04:53 PM	1	User redbook1 failed to logon on mgmt001st001(9.11.137.220) via interfac
EFSSA0179W	Warning	10/2/14 8:54:29 AM	10/2/14 10:04:08 AM	3	Management node configuration replication has failed. Passive managen
EFSSA0173W	Warning	10/2/14 8:54:29 AM	10/2/14 10:04:08 AM	2	A replication of active management node configuration is already runnin
EFSSF0011W	Warning	10/2/14 10:02:07 AM	10/2/14 10:02:07 AM	1	The clustered CIFS status changed to CTDBSTATE_STATE_DISCONNECTED
EFSSS0012W	Warning	10/2/14 9:48:23 AM	10/2/14 9:48:23 AM	1	User root failed to logon on mgmt001st001 via interface sshd_service[15
EFSSS0012W	Warning	10/2/14 9:23:56 AM	10/2/14 9:23:56 AM	1	User unidentified! failed to logon on mgmt001st001 via interface sshd_se
EFSSF1227W	Warning	10/1/14 11:20:31 PM	10/1/14 11:20:31 PM	1	The metadata usage of pool "system" in file system gpfs0 reached the v
EFSSA0179W	Warning	10/1/14 7:54:31 PM	10/1/14 7:54:31 PM	1	Management node configuration replication has failed. Passive managen
EFSSA0173W	Warning	10/1/14 7:54:31 PM	10/1/14 7:54:31 PM	1	A replication of active management node configuration is already runnin
EFSSG1028W	Warning	10/1/14 7:48:24 PM	10/1/14 7:48:24 PM	1	Filesystem gpfs0 has internal mounts which may cause operation to fail
EFSSS0012W	Warning	10/1/14 7:47:26 PM	10/1/14 7:47:26 PM	1	User root failed to logon on mgmt001st001 via interface sshd_service[22
EFSSS0012W	Warning	10/1/14 7:47:15 PM	10/1/14 7:47:15 PM	1	User root failed to logon on mgmt001st001 via interface sshd_service[22
EFSSG1028W	Warning	10/1/14 5:08:51 PM	10/1/14 6:10:31 PM	7	Filesystem gpfs0 has internal mounts which may cause operation to fail
EFSSA0179W	Warning	10/1/14 4:34:59 PM	10/1/14 5:44:27 PM	3	Management node configuration replication has failed. Passive managen
EFSSA0173W	Warning	10/1/14 4:34:59 PM	10/1/14 5:44:27 PM	3	A replication of active management node configuration is already runnin
EFSSA0623W	Warning	10/1/14 5:35:49 PM	10/1/14 5:35:49 PM	1	The availability of NSD array1_sas_60001ff0773501a890e0004 reached warr
EFSSA0623W	Warning	10/1/14 5:35:49 PM	10/1/14 5:35:49 PM	1	The availability of NSD array1 sas 60001ff07735018890c0002 reached warr

Figure 7-141 System Logs window

The Event Log window displays specific information warning and critical events from the syslog and displays them in a summarized view. SONAS administrators should look first at this log when they check for problems. The Event Log window displays system log events that are generated by the SONAS software, which include management console messages, system utilization incidents, status changes, and syslog events.

7.11 Call Home

The SONAS Storage Solution is designed to provide full support. The previous sections describe how to use the SONAS GUI and find information in the Monitoring navigation pane or directly from the event logs.

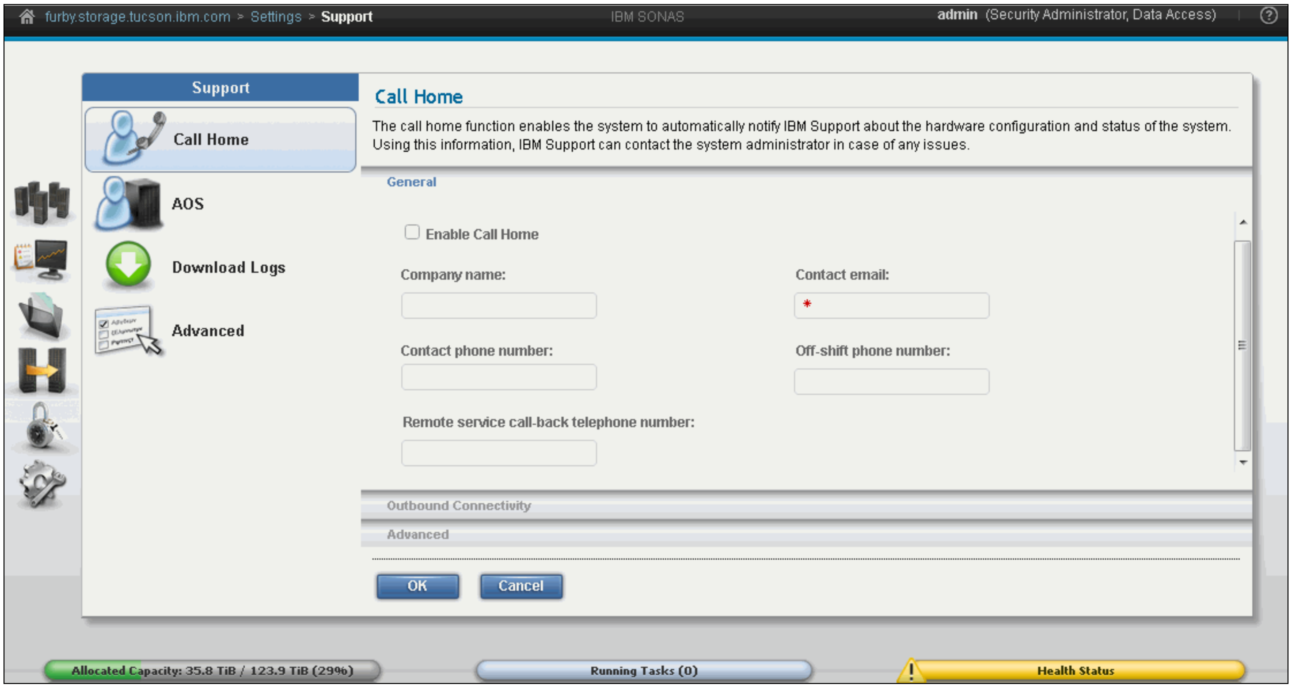
Each SONAS hardware component has at least one Error Detection Code method. This method can be the Denali code, which is a Director API module for checking and monitoring Interface, Storage, and Management nodes. Alternatively, it can be the System Checkout code, which is based on tape products, which monitors components such as InfiniBand switches, Ethernet switches, Fibre Channel connection, or Storage Controller. Last is the SNMP mechanism, which is used inside SONAS only, which monitors every component, server, switch, and Storage Controller.

The Denali method uses CIM providers, which are also used by the System Checkout method, and SNMP traps are converted into CIM providers. All these methods provide inputs to the GUI Health Center Event Log. Depending on the severity of this issue, it can raise an Electronic Customer Care (ECC) Call Home call. The Call Home feature was designed to start first with hardware events based on unique error codes. This Call Home feature is configured as part of first-time installation. It is used to send hardware events to IBM Support. Call Home is based only on Denali and System Checkout errors, but SNMP traps do not initiate Call Home.

The valid machine models that can use Call Home are as follows:

- 2851-SI1 – Interface Nodes
- 2851-SM1 – Management Nodes
- 2851-SS1 – Storage Nodes
- 2851-DR1 – Storage Controller
- 2851-I36 – 36-Port InfiniBand Switch
- 2851-I96 – 96-Port InfiniBand Switch

There are no Call Home calls against a 2851-DE1 Storage Expansion unit because any errors from it can Call Home against its parent 2851-DR1 Storage Controller Unit. Similarly, any errors against the Ethernet switches can Call Home against the 2851-SM1 Management node. To set Call Home, go to **Settings** and click **Support**. Complete the required fields to configure Call Home. See Figure 7-142. If you are using a proxy server, this information must be also entered. For a successful Call Home, you also must make small modifications to your Ethernet environment, such as opening a specific port. More details about Call Home can be found in 9.5, “Call Home” on page 669.


furby.storage.tucson.ibm.com > Settings > Support
IBM SONAS
admin (Security Administrator, Data Access)
Support
Call Home
AOS
Download Logs
Advanced
Outbound Connectivity
Advanced
OK
Cancel
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Call Home

The call home function enables the system to automatically notify IBM Support about the hardware configuration and status of the system. Using this information, IBM Support can contact the system administrator in case of any issues.

General

☐ Enable Call Home

Company name:

Contact email: *

Contact phone number:

Off-shift phone number:

Remote service call-back telephone number:

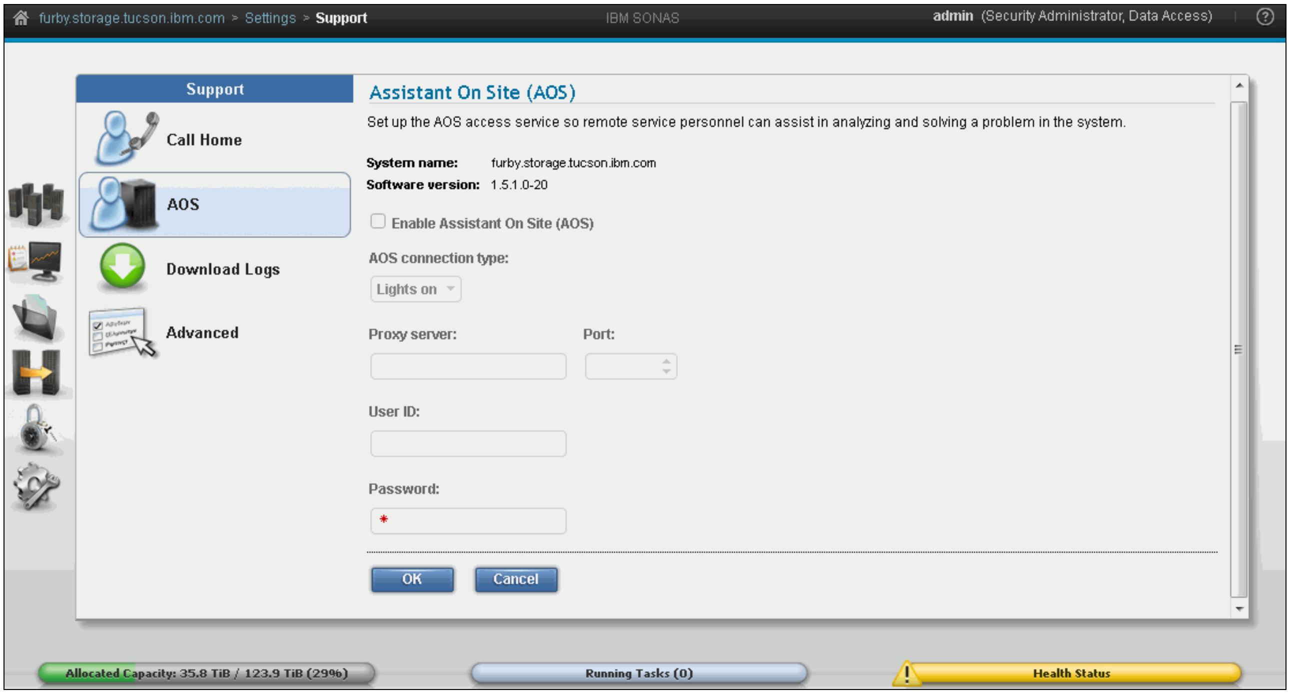
Figure 7-142 Call Home window in the GUI

7.12 Assist On-site

This feature is frequently used by IBM Support personnel to access remotely your system. You can set the authority level in four ways:

- Chat Only mode
- View Only mode
- Guidance mode
- Shared Control mode

These setting changes are made by personnel onsite. At all times, those personnel can regain full control or change the authority level to any level wanted. To work with Assist On-site (AOS) settings, go to the Settings window and click **Support**. In the newly opened menu, click the **Assist On-site** tab, as shown in Figure 7-143. For more information about AOS, see 9.4, “Assist On-site” on page 667.


furby.storage.tucson.ibm.com > Settings > Support
IBM SONAS
admin (Security Administrator, Data Access)
Support
Call Home
AOS
Download Logs
Advanced
OK
Cancel
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Assistant On Site (AOS)

Set up the AOS access service so remote service personnel can assist in analyzing and solving a problem in the system.

System name: furby.storage.tucson.ibm.com
Software version: 1.5.1.0-20

☐ Enable Assistant On Site (AOS)

AOS connection type: Lights on

Proxy server: Port:

User ID:

Password: *

Figure 7-143 AOS menu in the GUI

7.13 Logs: Uploading and downloading

If a system problem is reported to IBM Support by a support procedure or Call Home call, you might be asked to collect logs. The SONAS **`cndump`** utility collects memory dumps across a cluster and sends the memory dumps to IBM Support. They can be collected by using the GUI or CLI. The **`cndump`** utility creates a package with logs from the entire SONAS system.

The **`cndump`** utility in SONAS is organized in the following manner:

- Management node logs:
  - Basic state information from InfiniBand switches
  - Basic state information from Ethernet switches
  - CTDB logs
  - GPFS logs
- Interface nodes:
  - CTDB logs
  - GPFS logs
- Storage nodes:
  - Back-end storage controller logs
  - GPFS logs
  - Multipath state information

All nodes include operating system logs.

7.13.1 Collecting logs with the CLI

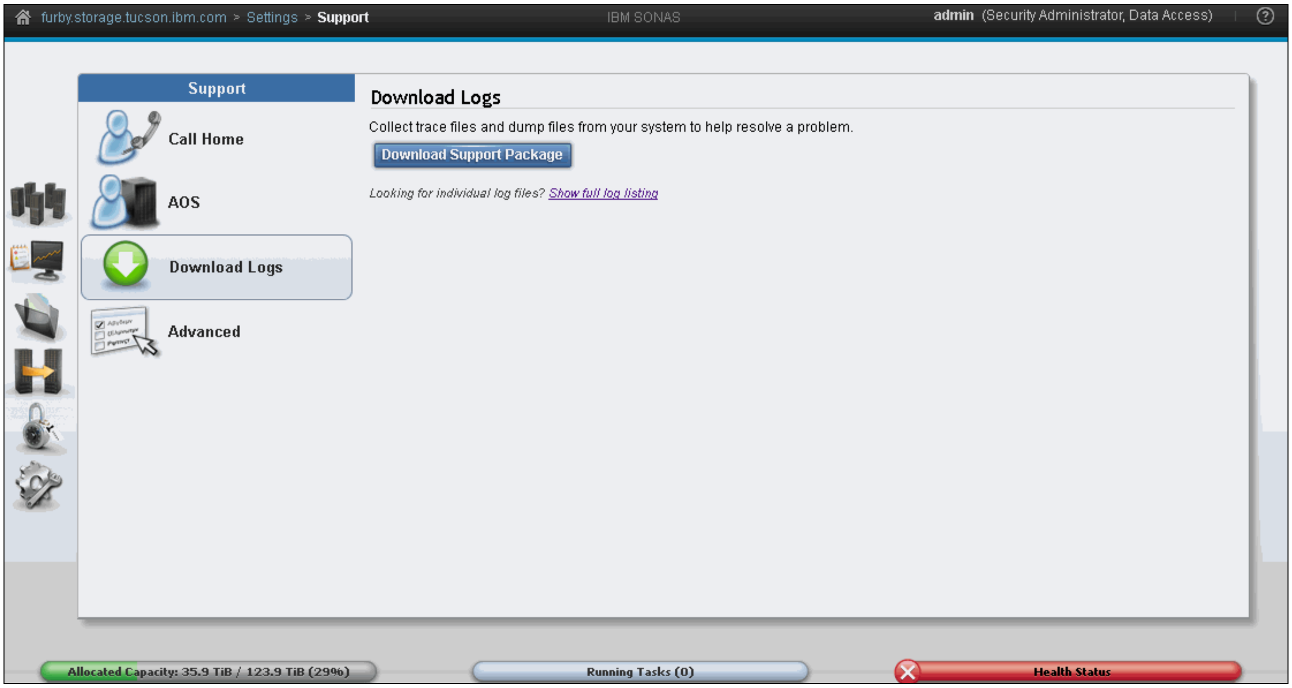
To collect logs by using the CLI, run **srvdump**. In Example 7-43, you see the **--help** option and log collection procedure.

Example 7-43 Help option for the srvdump command to generate logs

```
[furby.storage.tucson.ibm.com]$ srvdump --help
usage: srvdump  {-g <generateOption> | -d <deleteOption> | -l <listOption> | --get <getOption> |
-s <sendOption>} [-X] [-c < clusterID | clusterName >]
Manage dump files.

Parameter      Description
-g, --generate Generates dump files.
-d, --delete   Deletes dump files.
-l, --list     Lists dump files.
    --get      Extracts specific parts from a dump file.
-s, --send     Sends dump files to the specified destination.
-X, --Xtended  Collects extended dump files. Use this option in conjunction with the generate
option.
-c, --cluster  The cluster scope for this command
[fstsonas01.storage.tucson.ibm.com]$srvdump -g
2011.10.24-13:40:16 (cndump) ================================
2011.10.24-13:40:16 (cndump) =========== STARTED ============
2011.10.24-13:40:16 (cndump) ================================
2011.10.24-13:40:16 (cndump) Checking to see whether another copy of cndump is running
2011.10.24-13:40:16 (cndump) Seems safe to run now
2011.10.24-13:40:16 (cndump) Running on mgmt001st001
2011.10.24-13:40:16 (cndump) Host is type M
2011.10.24-13:40:16 (cndump) Checking for space in the /ftdc file system
2011.10.24-13:40:16 (cndump) Seems to be enough available space in /ftdc
2011.10.24-13:40:17 (cndump) Created /ftdc/cndump_fstsonas01_all_20111024-134016_MN/
2011.10.24-13:40:17 (cndump) Target directory is /ftdc/cndump_fstsonas01_all_20111024-134016_MN/
2011.10.24-13:40:17 (cndump) Using /opt/IBM/sonas/etc/cngetlogs.default for instructions
2011.10.24-13:40:17 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:17 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:17 (cndump) Working on node mgmt001st001
2011.10.24-13:40:17 (cndump) background process for mgmt001st001 has PID 1019926
2011.10.24-13:40:17 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:17 (cndump) Working on node int001st001
2011.10.24-13:40:19 (cndump) background process for int001st001 has PID 1020796
2011.10.24-13:40:19 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:19 (cndump) Working on node int002st001
2011.10.24-13:40:21 (cndump) background process for int002st001 has PID 1021229
2011.10.24-13:40:21 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:21 (cndump) Working on node strg001st001
2011.10.24-13:40:23 (cndump) background process for strg001st001 has PID 1021697
2011.10.24-13:40:23 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:23 (cndump) Working on node strg002st001
2011.10.24-13:40:25 (cndump) background process for strg002st001 has PID 1021867
2011.10.24-13:40:45 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:45 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:45 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:40:45 (cndump) Waiting for the background process(es) to finish
2011.10.24-13:43:51 (cndump_node) ===== strg002st001  FINISHED (rc=0) =====
2011.10.24-13:43:59 (cndump_node) ===== strg001st001  FINISHED (rc=0) =====
2011.10.24-13:44:09 (cndump_node) ===== int001st001  FINISHED (rc=0) =====
2011.10.24-13:44:15 (cndump_node) ===== int002st001  FINISHED (rc=0) =====
2011.10.24-13:45:57 (cndump_node) ===== mgmt001st001  FINISHED (rc=0) =====
2011.10.24-13:45:57 (cndump) All the background process(es) seem to be finished
2011.10.24-13:45:57 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:45:57 (cndump) *-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2011.10.24-13:45:57 (cndump) Create tar file
/ftdc/cndump_fstsonas01_all_20111024-134016_MN.tar.gz
2011.10.24-13:46:01 (cndump) Now delete the extra copy of the saved files
2011.10.24-13:46:01 (cndump)
**********************************************************************
2011.10.24-13:46:01 (cndump)
**********************************************************************
2011.10.24-13:46:01 (cndump) The files are stored in
/ftdc/cndump_fstsonas01_all_20111024-134016_MN.tar.gz
2011.10.24-13:46:01 (cndump) -rwxrwxrwx 1 root root 71M Oct 24 13:46
/ftdc/cndump_fstsonas01_all_20111024-134016_MN.tar.gz
2011.10.24-13:46:01 (cndump) For now move the tar file somewhere else before running cndump
again
2011.10.24-13:46:01 (cndump)
**********************************************************************
2011.10.24-13:46:01 (cndump)
**********************************************************************
EFSSG1000I The command completed successfully.
```

7.13.2 Collecting logs with the GUI

To collect logs by using the GUI, click **Settings**, click **Support**, and click **Logs**. For collecting new sets of logs, click **Download Support Package**, as shown in Figure 7-144.


furby.storage.tucson.ibm.com > Settings > Support
IBM SONAS
admin (Security Administrator, Data Access)
Support
Call Home
AOS
Download Logs
Advanced
Download Logs
Download Support Package
Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

Collect trace files and dump files from your system to help resolve a problem.

Looking for individual log files? Show full log listing

Figure 7-144 Download a support package with the GUI

7.14 Network settings

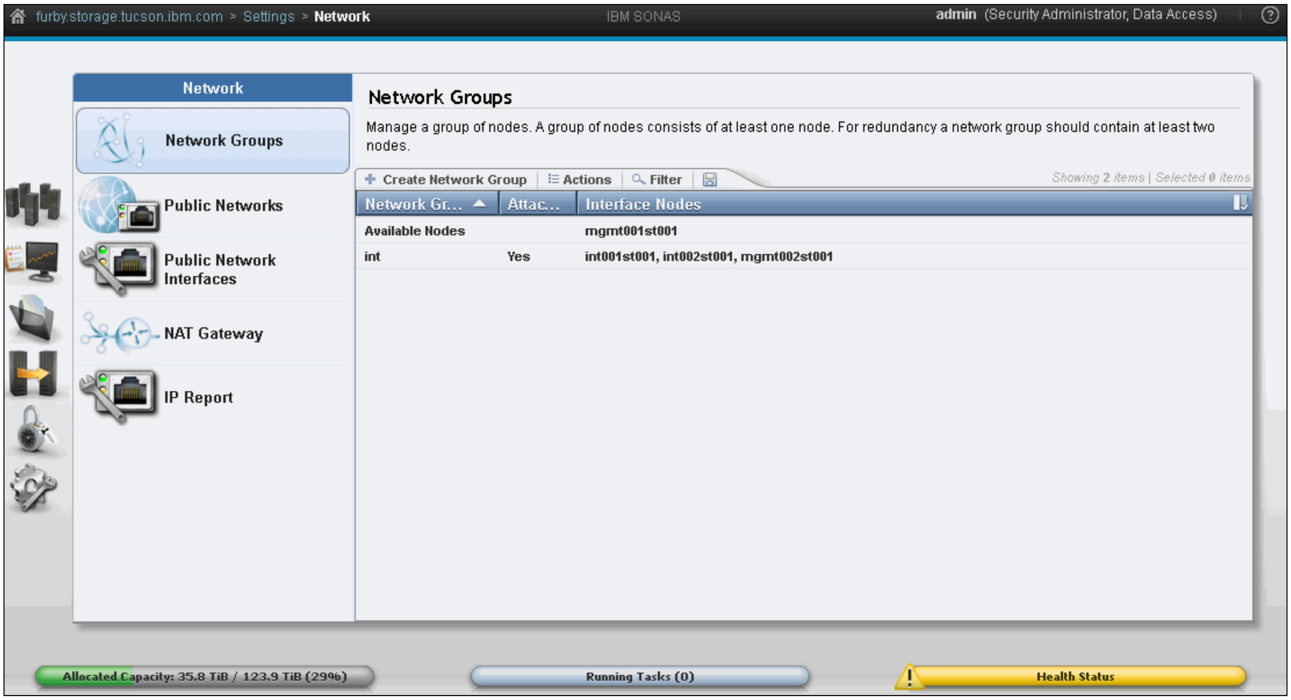
This section describes network settings tasks for both the GUI and CLI. To reach an IBM SONAS system with external clients, a public IP network must be configured and available. This network includes all network interfaces that are used to provide NAS services to external clients in the customer network.

7.14.1 Network groups

This section explains how to work with network groups.

Creating a network group by using the GUI

Click **Settings** → **Network** → **Network Group** and click **Create Network Group**, as shown in Figure 7-145.



furby.storage.tucson.ibm.com > Settings > Network
IBM SONAS
admin (Security Administrator, Data Access)
Network
Network Groups
Public Networks
Public Network Interfaces
NAT Gateway
IP Report
Create Network Group
Actions
Filter
Showing 2 items | Selected 0 items
Allocated Capacity: 35.8 TiB / 123.9 TiB (29%)
Running Tasks (0)
Health Status

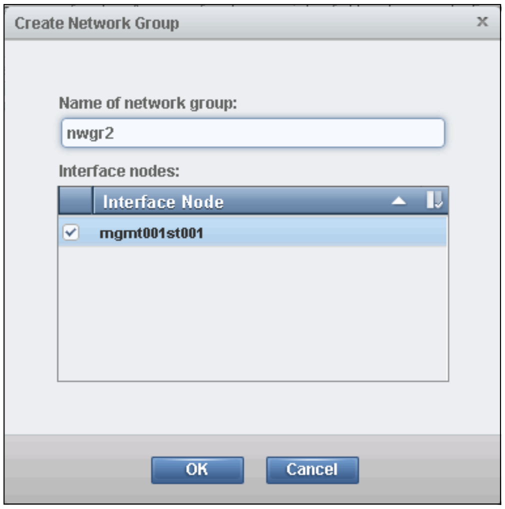
Network Groups

Manage a group of nodes. A group of nodes consists of at least one node. For redundancy a network group should contain at least two nodes.

Network Gr...	Attac...	Interface Nodes
Available Nodes		mgmt001st001
int	Yes	int001st001, int002st001, mgmt002st001

Figure 7-145 Network Groups settings view

When you create a network group from a list of specified node names, you cannot use the name of an existing network group. All of the specified nodes must be a member of the default network group. See Figure 7-146 on page 611.



Name of network group: nwgr2

Interface nodes:

☑ mgmt001st001

Figure 7-146 New network group configuration window

Click **OK** to save the settings. For more information about network groups, see 2.5.8, “Configuring the Data Path IP address group” on page 80.

> **Note:** While a node is a member of the DEFAULT network group, you cannot move that node out of the DEFAULT network group into any other network group while a network is attached to the DEFAULT network group. This limitation can require you to remove (detach) the network from all of the nodes that are members of the DEFAULT network group when you want to do maintenance on only one node. Custom (non-default) network groups do not have this limitation; therefore, to avoid this limitation, create at least one custom network group during initial system configuration.

Creating a network group by using the CLI

To create a network group, run **`mknwgroup`**. In Example 7-44, the network group that is named `nwgr2` is created with the nodes `mgmt001st003` and `int002st003` as members of this group.

Example 7-44 Network group CLI configuration example

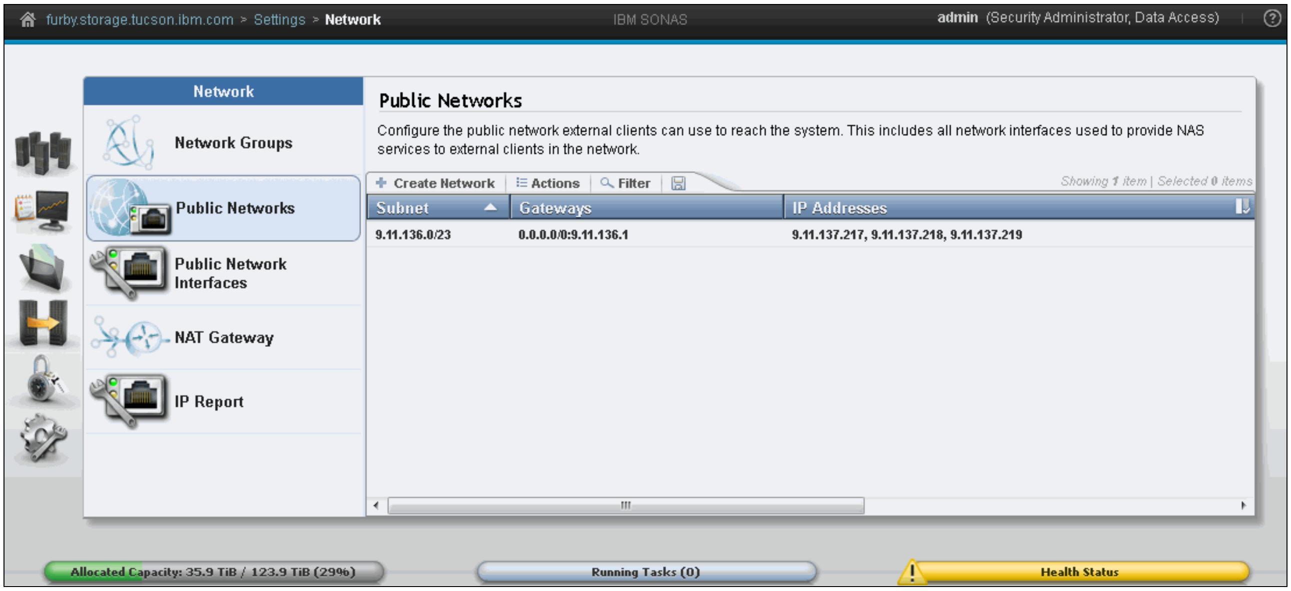
```
[furby.storage.tucson.ibm.com]$ mknwgroup --help
usage: mknwgroup groupname nodelist  [-c < clusterID | clusterName >]
Creates a group of nodes to which a network configuration can be attached. See also the commands mknw and attachnw.

Parameter     Description
groupname     Specifies the name of the group-created task.
nodelist      Displays a comma-separated list of nodes that will be used to build this NAT gateway group.
-c, --cluster The cluster scope for this command

[furby.storage.tucson.ibm.com]$ mknwgroup nwgr2 mgmt001st003,int002st003
EFSSG0015I Refreshing data.
Reconfiguring NAT gateway 10.0.0.31/24
EFSSG0087I NAT gateway successfully removed.
EFSSG0086I NAT gateway successfully configured.
EFSSG1000I The command completed successfully.

[furby.storage.tucson.ibm.com]$ lsnwgroup
Network Group Nodes                   Interfaces
DEFAULT
int           int001st003,mgmt002st003 ethX0
nwgr2         mgmt001st003,int002st003
EFSSG1000I The command completed successfully.
```

7.14.2 Public networks

There are a number of network adapters in the SONAS Interface nodes. Each is configured with IP addresses, and these addresses can be changed if required. This section provides information about the adapters and how to configure and reconfigure them.

Adding a network by using the GUI

These addresses are on the 10 Gb Ethernet ports and also the 1 Gb ports. Hosts and clients use these addresses to access the SONAS systems and open file shares. Click the **Settings** icon and select the **Network** option. This option displays the window that is shown in Figure 7-147. Click **Public Networks**.


furby.storage.tucson.ibm.com > Settings > Network
IBM SONAS
admin (Security Administrator, Data Access)
Network
Network Groups
Public Networks
Public Network Interfaces
NAT Gateway
IP Report
Create Network
Actions
Filter
Showing 1 item | Selected 0 items
Running Tasks (0)
Health Status

Public Networks

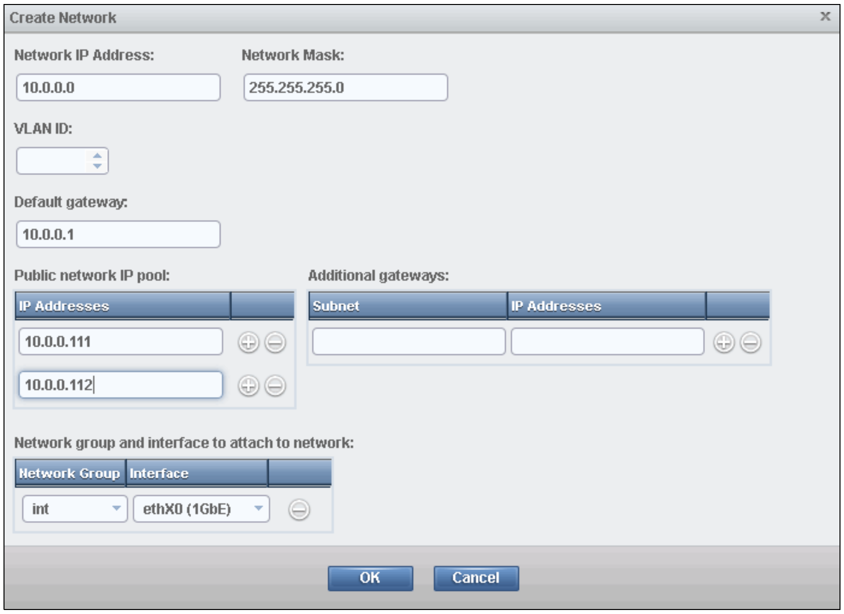
Configure the public network external clients can use to reach the system. This includes all network interfaces used to provide NAS services to external clients in the network.

Subnet	Gateways	IP Addresses
9.11.136.0/23	0.0.0.0/0:9.11.136.1	9.11.137.217, 9.11.137.218, 9.11.137.219

Allocated Capacity: 35.9 TiB / 123.9 TiB (29%)

Figure 7-147 Public Networks configuration view

From here, you can add new network definitions and delete existing ones. Each definition is defined to one of the virtual adapters. Select **Create Network** to add a network, as shown in Figure 7-148 on page 613.



Network IP Address: 10.0.0.0

Network Mask: 255.255.255.0

VLAN ID:

Default gateway: 10.0.0.1

Public network IP pool:

IP Addresses
10.0.0.111
10.0.0.112

Additional gateways:

Subnet	IP Addresses

Network group and interface to attach to network:

Network Group	Interface
int	ethX0 (1GbE)

Figure 7-148   Create Network window

Here are descriptions of the fields that are shown in Figure 7-148:

Subnet The IP subnet for this definition. The format of this value uses the CIDR syntax *xxx.xxx.xxx.xxx/yy*, where *yy* is the decimal mask that is given as the number of left-aligned bits in the mask.

VLAN ID If VLANs are being used, enter the VLAN number here. Otherwise, leave the field empty. Valid values are 2 - 4095. VLAN 1 is not supported for security reasons.

Default Gateway The default gateway (or router) within this subnet for routing. This field is not required if all devices that are connected to this interface are in the same subnet.

Interface Pool Use the **+** button to add IP addresses to the pool for use on this logical interface. These addresses must be in the same subnet as entered previously. A minimum of one address is required.

Additional Gateways If more than one gateway (router) exists in the subnet, add the IP addresses here.

Interface Select the logical interface to which this definition is assigned.

A progress window opens. Wait for the completion message. Click **OK** when it is complete. Repeat the process for each network by clicking **New Network** until all the required networks are defined.

Creating a network with the CLI

To configure the public network from the CLI, run **mknw** and **attachnw**. In Example 7-45, you see the **--help** option and the setup procedure.

Example 7-45   Add a network by using the CLI example

```
[furby.storage.tucson.ibm.com]$ mknw --help
usage: mknw subnet  [routes] [--add <ips>] [--vlanid <vlanID>] [-c < clusterID | clusterName >]
Defines a new network configuration for a subnet, and assigns multiple IP addresses and routes.

Parameter     Description
subnet        Specifies the subnet to be modified.
```

```
routes        Lists the subnets and gateways in a colon-separated pair and in a comma-separated list.
    --add     Lists the IP addresses to add to this network configuration by using a comma-separated list.
    --vlanid  Specifies the virtual LAN (VLAN) ID to be used. The VLAN ID must be in the range of 1 to 4094.
-c, --cluster The cluster scope for this command

[furby.storage.tucson.ibm.com]$ mknw '10.0.0.0/24' 0.0.0.0/0:10.0.0.1 --add 10.0.0.121,10.0.0.122
EFSSG1000I The command completed successfully.
[st003.virtual.com]$ attachnw '10.0.0.0/24' ethX0 -g int
EFSSG0015I Refreshing data.
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsnw
Network     VLAN ID Network Groups IP-Addresses          Routes
10.0.0.0/24         int            10.0.0.131,10.0.0.132 0.0.0.0/0:10.0.0.1
10.1.0.0/24         int            10.1.0.121,10.1.0.122 0.0.0.0/0:10.1.0.1
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ lsnwinterface
Node         Interface MAC               Master/Subordinate Bonding mode      Transmit hash policy Up/Down Speed
IP-Addresses          MTU
int001st003  ethX0     02:b8:dd:02:03:00 MASTER active-backup (1) UP      1000  10.0.0.131            1500
mgmt001st003 ethX0     02:b8:dd:00:03:00 MASTER active-backup (1) UP      1000                        1500
mgmt002st003 ethX0     02:b8:dd:01:03:00 MASTER active-backup (1) UP      1000  10.0.0.132,10.1.0.121 1500
EFSSG1000I The command completed successfully.
```

7.15 Event notifications

You can configure the IBM SONAS system so that you receive automated notifications when certain events occur in the system, such as a file system approaching its size limit, a quota being exceeded, or a processor becoming overloaded. You can use the GUI or CLI commands to configure notification settings. Only an Administrator, Security Administrator, or System Administrator can configure and change the event notification settings.

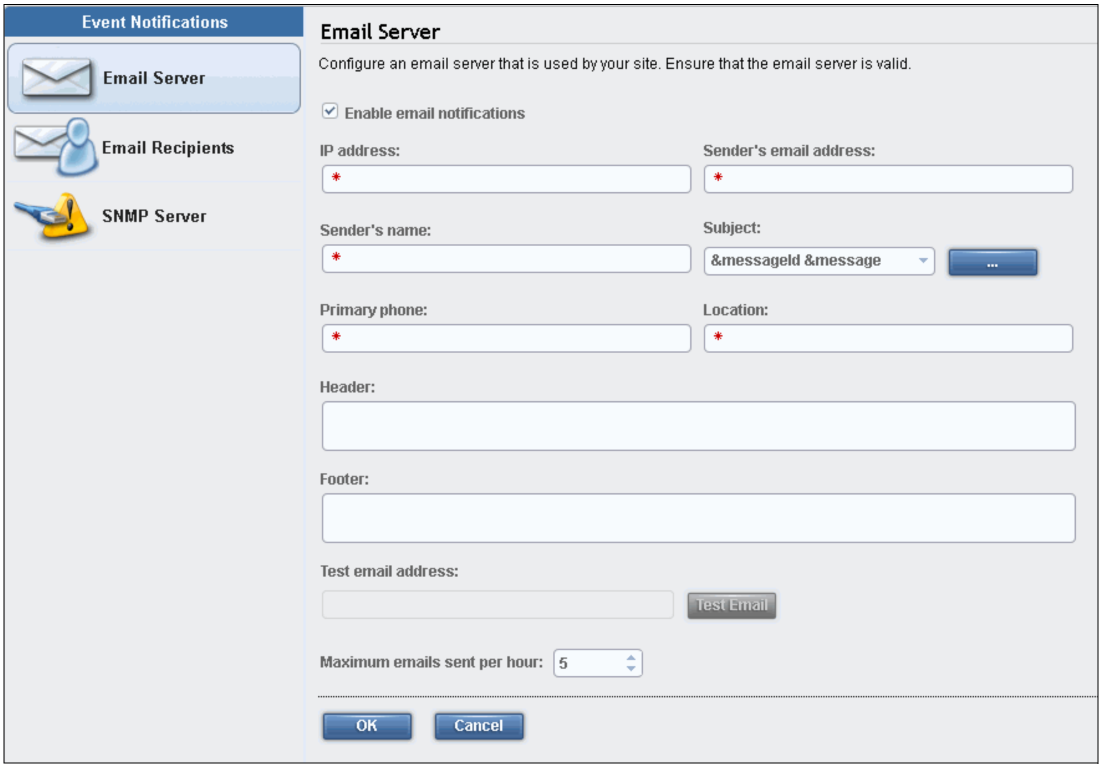
7.15.1 Configuring event notifications settings

This section provides information about configuring notification settings. To manage event notifications in the management GUI, click **Settings** → **Event Notifications** (see Figure 7-27 on page 510). You can also configure event notification with the CLI.

Configuring the email server

To use the Event Notifications function, you must configure the SMTP server first. You can configure the email server by using the GUI or CLI.

Using the GUI: Click **Settings** → **Event Notifications** and click **Email Server** in the right pane. See Figure 7-149 on page 615.



Email Server

Configure an email server that is used by your site. Ensure that the email server is valid.

☑ Enable email notifications

IP address: *

Sender's email address: *

Sender's name: *

Subject: &messageId &message

Primary phone: *

Location: *

Header:

Footer:

Test email address:

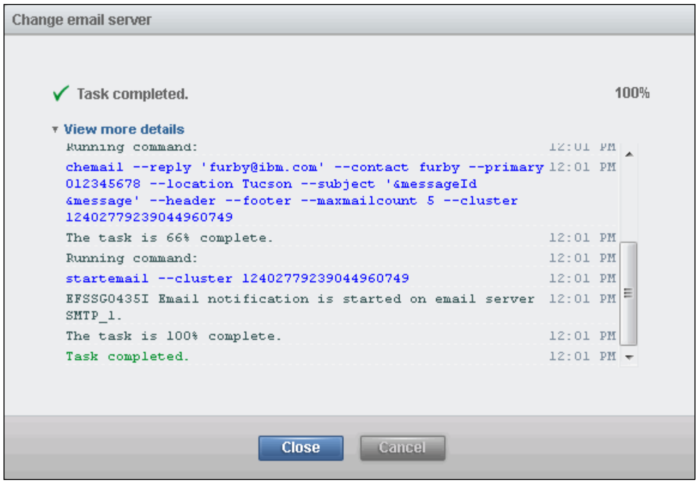
Maximum emails sent per hour: 5

Figure 7-149 Configure Email Server window

Configure the following settings:

- IP address: The SMTP server IP address.
- Sender's email address: Enter the sender's email address that is sent with email notifications. This mail address can be contacted by the recipient of the email notifications.
- Sender's name: Enter the sender's name. This name that is sent with email notifications.
- Primary phone: Enter the phone number, which the recipient can contact.
- Location: Enter the system's location.

After all the required information is entered, click **OK** to apply the settings. The result is shown in the Change email server window. See Figure 7-150.



```
Running command:                                                 12:01 PM
chemail --reply 'furby@ibm.com' --contact furby --primary        12:01 PM
012345678 --location Tucson --subject '&messageId
&message' --header --footer --maxmailcount 5 --cluster
12402779239044960749
The task is 66% complete.                                        12:01 PM
Running command:                                                 12:01 PM
startemail --cluster 12402779239044960749                        12:01 PM
EFSSG0435I Email notification is started on email server         12:01 PM
SMTP_1.
The task is 100% complete.                                       12:01 PM
Task completed.                                                  12:01 PM
```

Figure 7-150 Change the email server and the resultant messages

Using the CLI: To configure the email server, run **mkemailserver**, **chemailserver**, and **chemail**. For details, see Example 7-46.

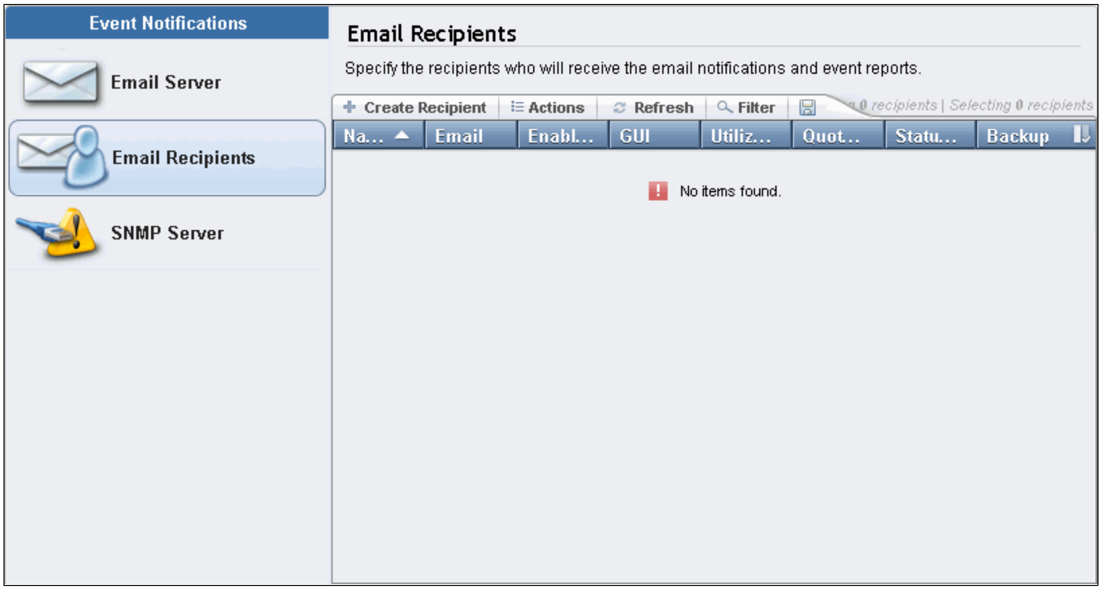
Example 7-46 Configure the email server with the CLI

```
[furby.storage.tucson.ibm.com]$ mkemailserver --help
usage: mkemailserver serverName  --ip <serverIP> [--port <portNumber>] [-c < clusterID |
clusterName >]
Creates email server definition.

Parameter     Description
serverName    Specifies the name of the email server.
    --ip      Specifies the IP address of the email server.
    --port    Specifies the port number of the email server.
-c, --cluster The cluster scope for this command
[furby.storage.tucson.ibm.com]$ chemailserver --help

usage: chemailserver serverName  {--ip <serverIP> | --port <portNumber>} [-c < clusterID |
clusterName >]
Changes email server definition.

Parameter     Description
serverName    Specifies the name of the email server.
    --ip      Specifies the IP address of the email server.
    --port    Specifies the port number of the email server.
-c, --cluster The cluster scope for this command

[furby.storage.tucson.ibm.com]$ chemail --help
usage: chemail  {--reply <reply_email_address/sender_address> | --contact
<contact_name/sender_name> | --primary <primary_telephone_number> | --alternate
<alternate_telephone_number> | --location <machine_location> | --subject <email subject> |
--header <headertext> | --footer <footertext> | --maxmailcount <count_per_hour>} [-c <
clusterID | clusterName >]
Changes email notification configuration.

Parameter          Description
    --reply        Designates the response email address or sender's address.
    --contact      Specifies the contact's name or sender's name.
    --primary      Specifies the primary telephone number.
    --alternate    Specifies the alternate telephone number.
    --location     Specifies location.
    --subject      Specifies the subject of the email.
    --header       Designates the header of the email.
    --footer       Designates the footer of the email.
    --maxmailcount Specifies the maximum number of emails sent per hour. The default is 5.
-c, --cluster      The cluster scope for this command
```

Configuring email recipients

This section describes how to configure the recipients for email notifications and event reports.

Using the GUI: Click **Settings** → **Event Notifications** and click **Email Recipients** in the right pane. See Figure 7-151.

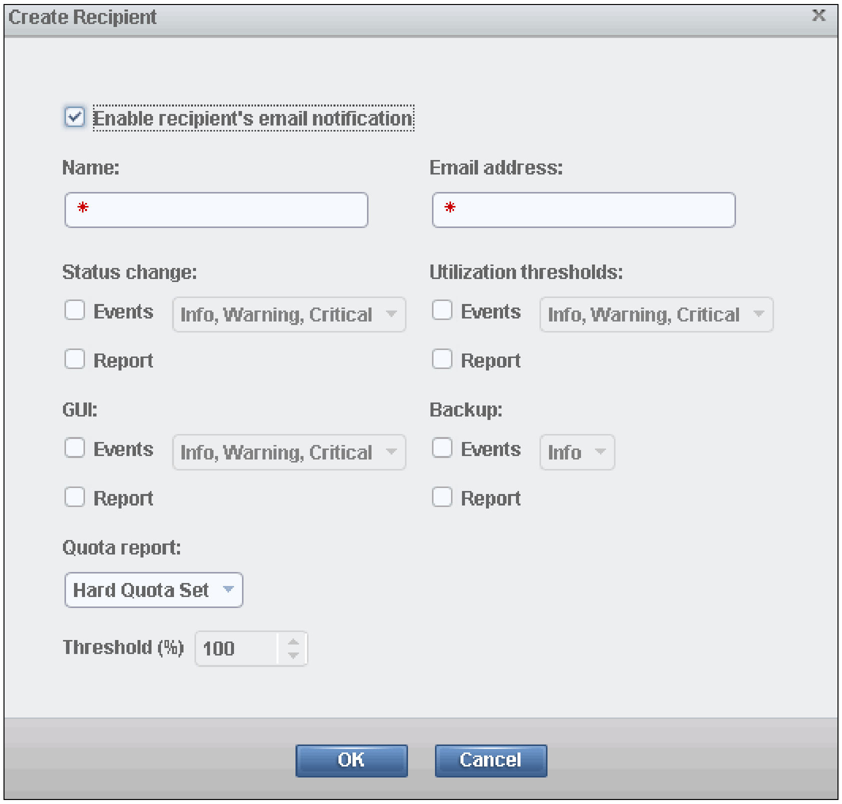


Specify the recipients who will receive the email notifications and event reports.

No items found.

Figure 7-151 Configure the Email Recipients window

Click **Create Recipient** in the right pane. The Create Recipient window opens, as shown in Figure 7-152.



Create Recipient

☑ Enable recipient's email notification

Name: *

Email address: *

Status change:

□ Events Info, Warning, Critical

□ Report

Utilization thresholds:

□ Events Info, Warning, Critical

□ Report

GUI:

□ Events Info, Warning, Critical

□ Report

Backup:

□ Events Info

□ Report

Quota report: Hard Quota Set

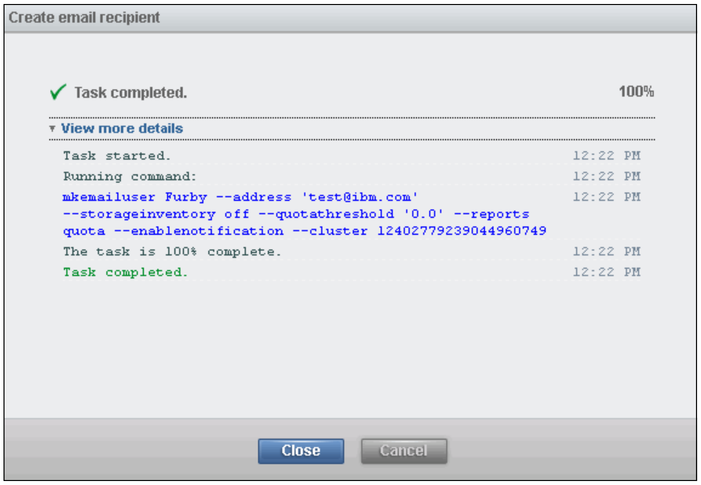
Threshold (%) 100

Figure 7-152 Create Recipient window

Enter all of the required information and click **OK** to apply the changes. A Task completed window opens, as shown in Figure 7-153 on page 619.

Note: The Quota report men in Figure 7-152 has the following five options.

- **All Data** reports all capacity data for users, groups, and file sets.
- **Exceeding Soft Quota** reports only the values where the soft quota is exceeded, even if no hard quota is set.
- **Hard Quota Set** includes all entries where the hard quota is set.
- **Exceeding Hard Quota Threshold** reports all values that are equal or greater than the defined threshold of percentage of the hard quota.
- **Exceeding Soft Quota (Hard Set)** reports only the values where the hard quota is set and the soft quota is exceeded.



Create email recipient

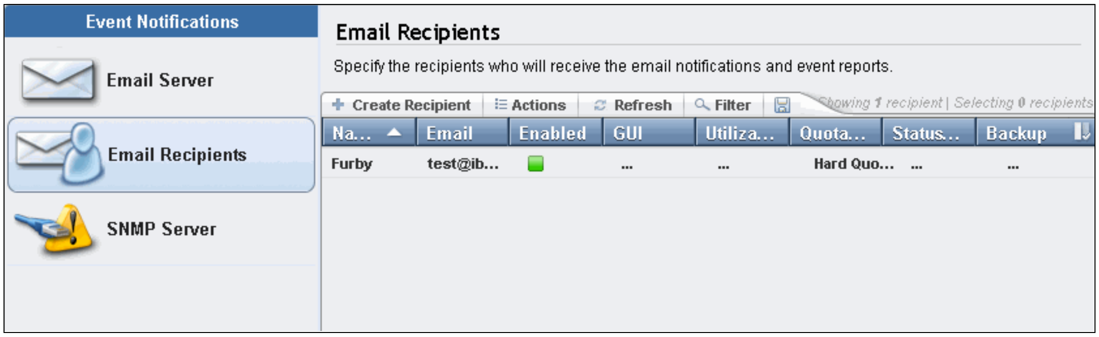
✓ Task completed. 100%

▾ View more details

```
Task started.                                              12:22 PM
Running command:                                           12:22 PM
mkemailuser Furby --address 'test@ibm.com'                 12:22 PM
--storageinventory off --quotathreshold '0.0' --reports
quota --enablenotification --cluster 12402779239044960749
The task is 100% complete.                                 12:22 PM
Task completed.                                            12:22 PM
```

Figure 7-153 Create email recipient result window

After that, the newly created email recipient is shown in the right pane. See Figure 7-154.



Email Recipients

Specify the recipients who will receive the email notifications and event reports.

Showing 1 recipient | Selecting 0 recipients

Na...	Email	Enabled	GUI	Utiliza...	Quota...	Status...	Backup
Furby	test@ib...	🟩	...	...	Hard Quo...	...	...

Figure 7-154 Recipient list in the Email Recipients pane

Using the CLI: To create a recipient, run **mkemailuser**. For details, see Example 7-47.

Example 7-47 Create an email recipient with the mkemailuser command

```
[furby.storage.tucson.ibm.com]$ mkemailuser --help
usage: mkemailuser userName  --address <userAddress> [--utilization <level>]
[--gui <level>] [--status <level>] [--systemlog <level>] [--storagealerts <level>]
[--storageinventory <on/off>] [--quotathreshold <percentage>] [--reports
<componentList>] [--backup <level>] [--enablenotification] [-c < clusterID |
clusterName >]
Creates an email user.

Parameter                Description
userName                 Specifies the name of the email user that is going to be
changed.
    --address            Specifies the address of the email user.
    --utilization        Specifies the severity level of utilization events.
    --gui                Specifies the severity level of GUI events.
    --status             Specifies the severity level of status events.
    --systemlog          Specifies the severity level of systemlog events.
    --storagealerts      Specifies the severity level of storage alerts.
    --storageinventory   Designates the storage inventory switch.
    --quotathreshold     Sets the quota threshold. The default value is 100.
    --reports            Specifies the component to be reported
(utilization/status/gui/syslog/quota/backup).
    --backup             The severity level of backup (info/logs/full/none).
    --enablenotification Specifies the state of the email user.
-c, --cluster            The cluster scope for this command
```

Configuring the SNMP server

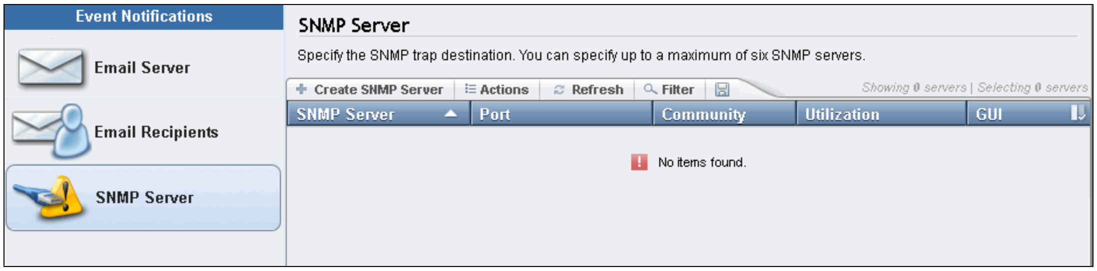
To configure the SNMP server, you must provide SONAS MIB files. SONAS provides two MIB files that include hardware and software information to provide to the SNMP server:

- `IBM-SONAS-NOTIFICATION-MIB.txt`
- `SONAS-TC-MIB.txt`

These two MIB files can be downloaded from the SONAS IBM Knowledge Center website:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/pln_snmp.html

Using the GUI: Click **Settings** → **Event Notifications** and select **SNMP Server**. See Figure 7-155.



Specify the SNMP trap destination. You can specify up to a maximum of six SNMP servers.

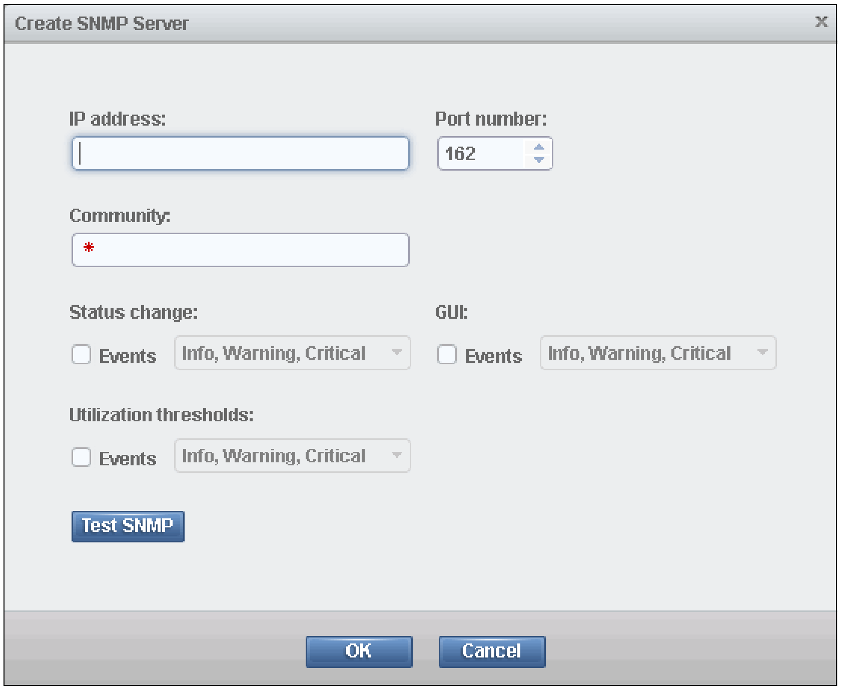
Showing 0 servers | Selecting 0 servers

SNMP Server	Port	Community	Utilization	GUI

No items found.

Figure 7-155 SNMP Server window

To create an SNMP server, click **Create SNMP Server** in the SNMP Server window. See Figure 7-156 on page 621.


Create SNMP Server

IP address:

Port number: 162

Community: *

Status change:

□ Events Info, Warning, Critical

GUI:

□ Events Info, Warning, Critical

Utilization thresholds:

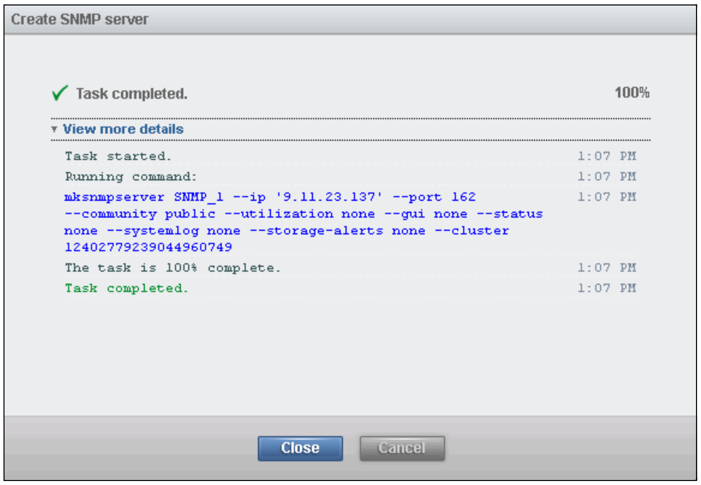
□ Events Info, Warning, Critical

Figure 7-156 Create SNMP Server window

Complete the fields, After all information is entered, click **OK** to apply your changes.

> **Note:** The IP address, Port number, and Community strings are provided by the customer.

The task completes, as shown in Figure 7-157.


Create SNMP server

✓ Task completed. 100%

▾ View more details

```
Task started.                                           1:07 PM
Running command:                                        1:07 PM
mksnmpserver SNMP_1 --ip '9.11.23.137' --port 162       1:07 PM
--community public --utilization none --gui none --status
none --systemlog none --storage-alerts none --cluster
12402779239044960749
The task is 100% complete.                              1:07 PM
Task completed.                                         1:07 PM
```

Figure 7-157 Create SNMP Server result window

Using the CLI: To create and modify the SNMP server, run **mksnmpserver** and **chsnmpserver**. The **lssnmpserver** command shows the configured SNMP server list. For details, see Example 7-48.

Example 7-48 Configure and manage the SNMP server with commands

```
[furby.storage.tucson.ibm.com]$ lssnmpserver
Server name IP          Port Community Utilization GUI  Status Syslog Storage alert
SNMP_1      9.11.23.137 162  public    none        none none   none   none
EFSSG1000I The command completed successfully.
[furby.storage.tucson.ibm.com]$ mksnmpserver --help
usage: mksnmpserver snmpServerName  --ip <ip> [--port <port_number>] [--community <community_name>] [--utilization
<utilization>] [--gui <gui_level>] [--status <status_level>] [--systemlog <systemlog_level>] [--storage-alerts
<storageAlerts_level>] [--testOnly] [-c < clusterID | clusterName >]
Creates a new SNMP server definition.

Parameter           Description
snmpServerName      Specifies the name of the SNMP server.
    --ip            Specifies the  IP of the SNMP server.
    --port          Specifies the port number of the SNMP server. The default is 162.
    --community     Specifies the community name (default name of public is used if omitted).
    --utilization   Specifies the information level for utilization (default is none).
    --gui           Specifies the information level for the GUI (By default, none).
    --status        Specifies the information level for the status (default is none).
    --systemlog     Specifies the information level for the system log (default is none).
    --storage-alerts Specifies the information level for the storage alerts (default is none).
    --testOnly      Sends a test SNMP trap.
-c, --cluster       The cluster scope for this command
[furby.storage.tucson.ibm.com]$ chsnmpserver --help
usage: chsnmpserver snmpServerName  {--ip <ip> | --port <port_number> | --community <community_name> | --utilization
<utilization_level> | --gui <gui_level> | --status <status_level> | --systemlog <systemlog_level> | --storage-alerts
<storageAlerts_level>} [-c < clusterID | clusterName >]
Changes an existing SNMP server definition.

Parameter           Description
snmpServerName      Specifies the name of the SNMP server.
    --ip            Specifies the IP of the SNMP server
    --port          Specifies the port number of the SNMP server.
    --community     Specifies the community name.
    --utilization   Specifies the information level for utilization.
    --gui           Specifies the information level for the GUI.
    --status        Specifies the information level for the status.
    --systemlog     Specifies the information level for the system log.
    --storage-alerts Specifies the information level for the storage alerts.
-c, --cluster       The cluster scope for this command
```

8

Monitoring

This chapter shows how to monitor the status and the performance of your SONAS system with the built-in monitoring functions and the Performance Center that are provided in the GUI and the CLI.

This chapter describes the following topics:

- Introduction to monitoring
- Monitoring the SONAS system
- Performance
- Capacity
- Other important cluster monitoring considerations

8.1 Introduction to monitoring

You can use the monitoring capabilities in SONAS to perform the following tasks:

- Monitor the system.
- Monitor the logs:
  - System logs
  - Alert logs
- Monitor the events.
- Monitor the system details (that is, nodes and disks).
- Monitor the capacity.
- Monitor the performance.

For more information about Monitoring Logs, Events, and System Details, see 9.7, "Overview of logs" on page 673, where you can see how to view logs from the GUI and the CLI. You can also see information about how to monitor the nodes, disks, and more.

Monitoring functions of SONAS system are the second item in the left pane of the GUI. The Monitoring icon provides you with five different categories of monitoring, as shown in Figure 8-1.

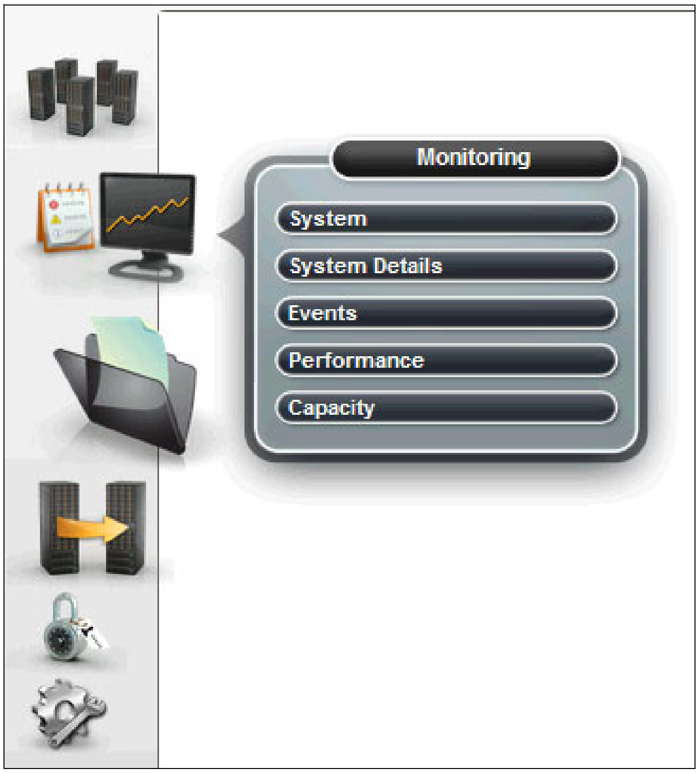


Figure 8-1 Monitoring menu

8.2 Monitoring the SONAS system

You can use the monitoring functions of SONAS to check the SONAS system and its status. At the time of writing, the monitoring functions are accessible only through the GUI. The GUI provides a pictorial view of the SONAS rack. The rack slots are marked where there is a node, expansion unit, or switch.

To see the system overview, click the **Monitoring** icon and then click **System**, as shown in Figure 8-2.

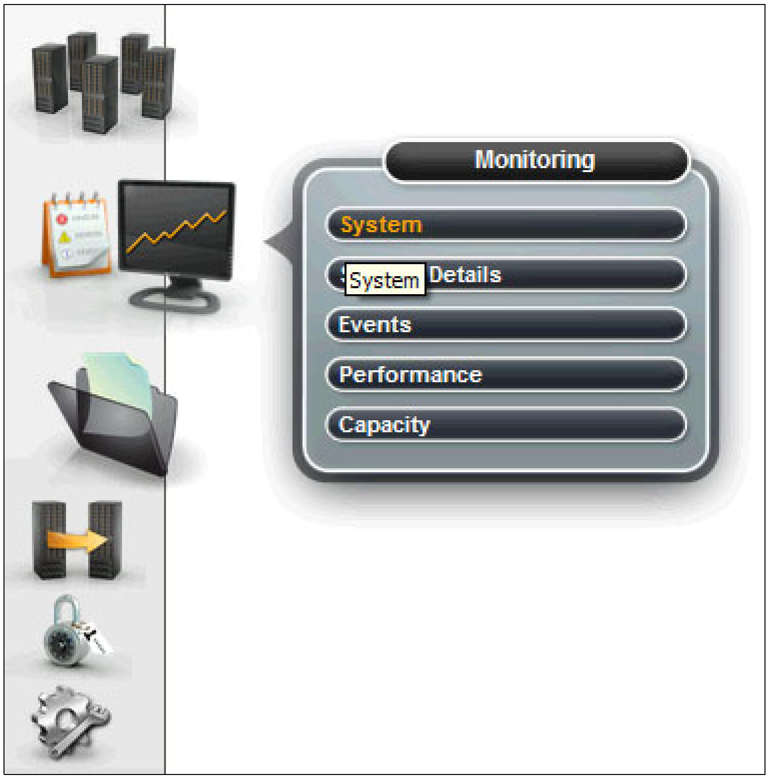


Figure 8-2 Monitoring System option

For step-by-step recommendations for health and system component checks, see “Monitoring as a daily administration task” in *Scale Out Network Attached Storage Monitoring*, SG24-8207.

8.3 Performance

Performance monitoring of your SONAS systems is done by Performance Center through the GUI or the CLI, and they use the same status log data that the system collects. Performance Center collects the status data of your SONAS system every second, and keeps it for up to one year. The data collection task has little impact on your system. You can view performance graphs of a specific part of your system in real time for a specific period.

8.3.1 GUI interface for the SONAS Performance Center

The GUI Performance Center is the fourth item under the Monitoring menu. You can view Client Network Throughput, Cluster Throughput, Cluster Latency, and Cluster Operations in each quadrant, which are cluster-wide metrics. Click **Interface Nodes** or **Storage Nodes** at the top to view node-level metrics. See Figure 8-3. The graphs refresh themselves every five seconds for near-real-time display.

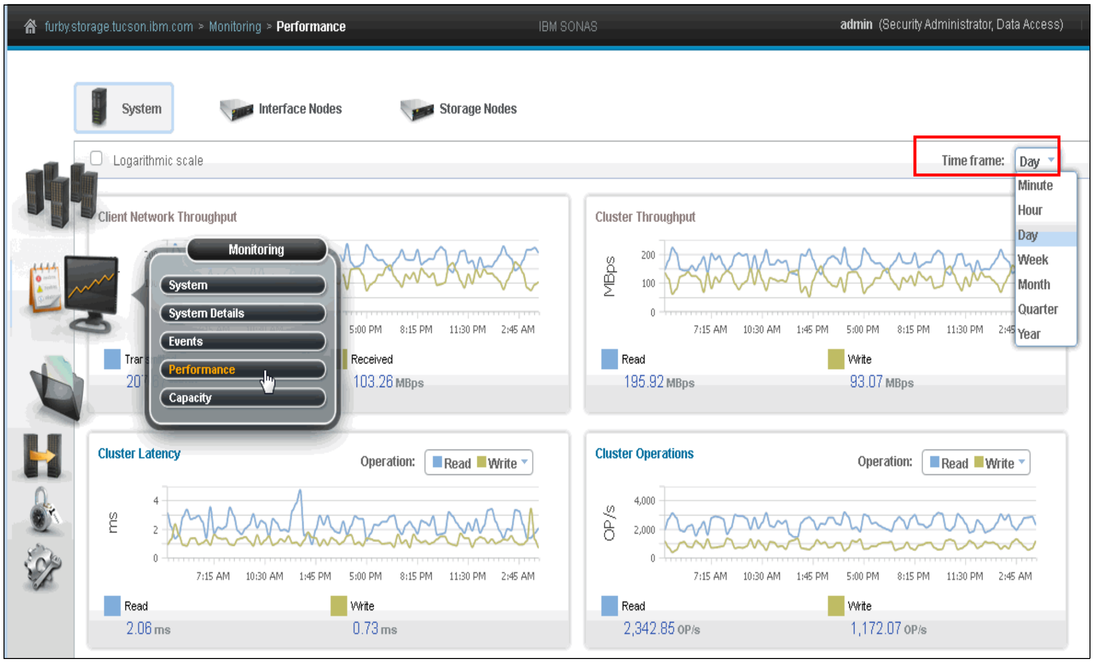


Figure 8-3 Performance monitoring in the SONAS GUI

You can also specify the period for which you want to collect the data (for a minute, an hour, a day, a week, a month, a quarter, or a year) by selecting **Time frame**, which is marked red in Figure 8-3. Your selection becomes the starting point of the period, and the ending point is always the current time. Changing the Time Frame setting affects all graphs.

Table 8-1 shows the *resolution* of each Time Frame option that you select. The resolution of the time frame refers to the granularity of the data that Performance Center picks for each sample out of the set of data that collected every second during that period.

Table 8-1 Performance monitoring with the Time Frame option

Time Frame setting	Resolution
Minute	Every 1 second
Hour	Every 12 seconds
Day	Every 5 minutes
Week	Every 30 minutes
Month	Every 2 hours
Quarter	Every 8 hours
Year	Every 1 day

Client Network Throughput monitoring

Client Network Throughput monitoring is in the first quadrant of the Performance Center GUI, as shown in Figure 8-4. You can view the sum of the throughput data that occurred each second over the network between all clients and Interface nodes in MBps during the specified time frame.

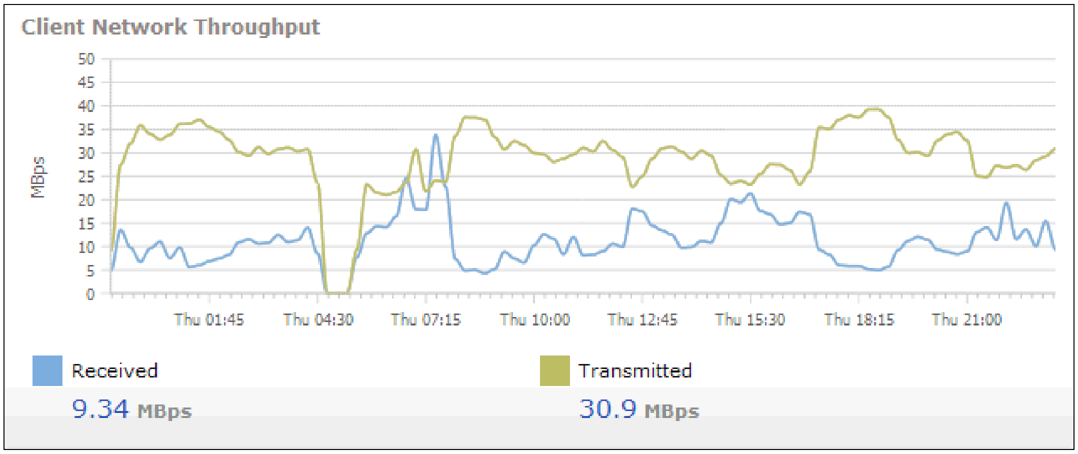


Figure 8-4 Client Network Throughput monitoring

Cluster Throughput monitoring

Cluster Throughput monitoring is in the second quadrant of the Performance Center GUI, as shown in Figure 8-5. You can view the sum of the throughput of read and write data that occurred each second across the cluster in MBps during a specified time frame.

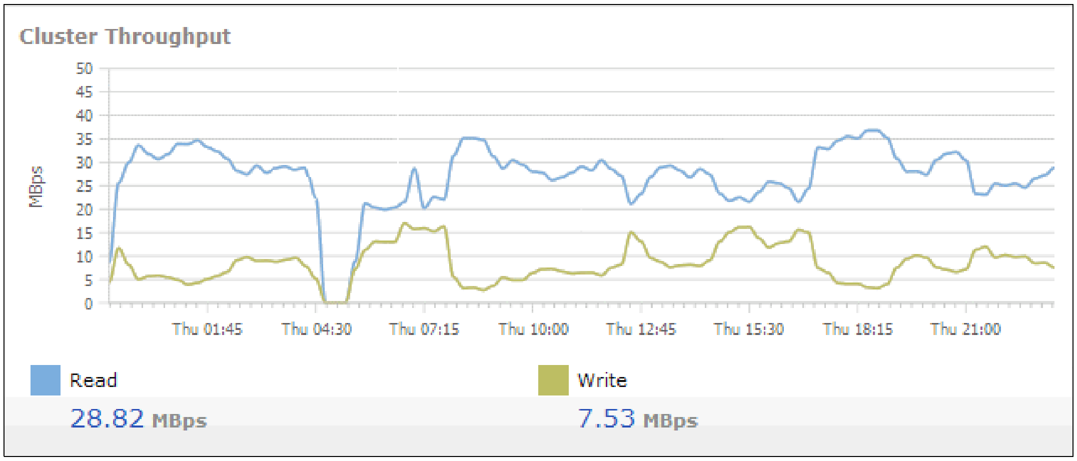


Figure 8-5 Cluster Throughput monitoring

Cluster Latency monitoring

Cluster Latency monitoring is in the third quadrant of the Performance Center GUI, as shown in Figure 8-6. You can view the average latency time between each of the operations across the cluster in ms/OPs during specified time frame.

You can select which operation data to see by selecting **Operation** at the upper right. There are three options: Read/Write, Open/Close, and Create/Delete. Changing this option affects only the display within this quadrant.

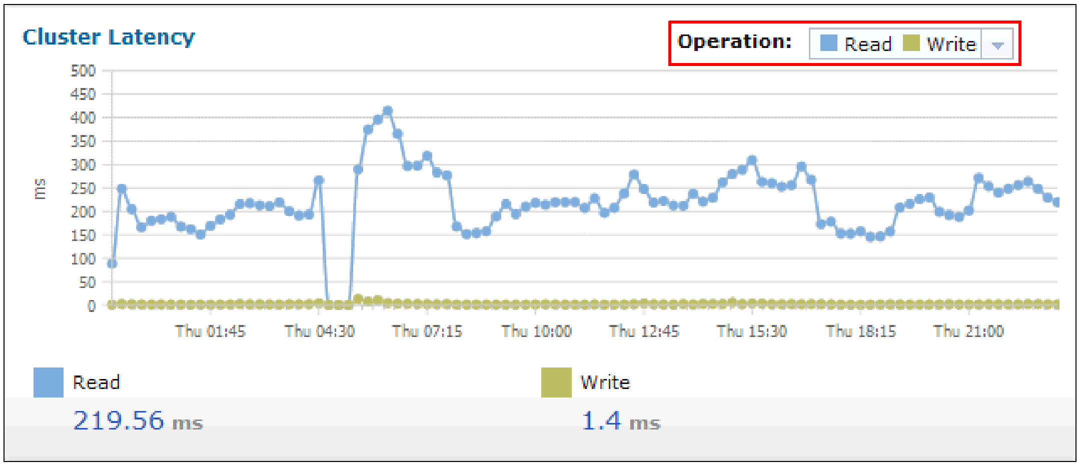


Figure 8-6 Cluster Latency monitoring

Cluster Operations monitoring

Cluster Operations (IOPS) monitoring is in the fourth quadrant of the Performance Center GUI, as shown in Figure 8-7.

You can view the number of operations system handled every second across the cluster in OP/s during the specified time frame.

You can select which operation data to see by clicking **Operation** at the upper right. There are three options: Read/Write, Open/Close, and Create/Delete. Changing this option affects only the display within this quadrant.

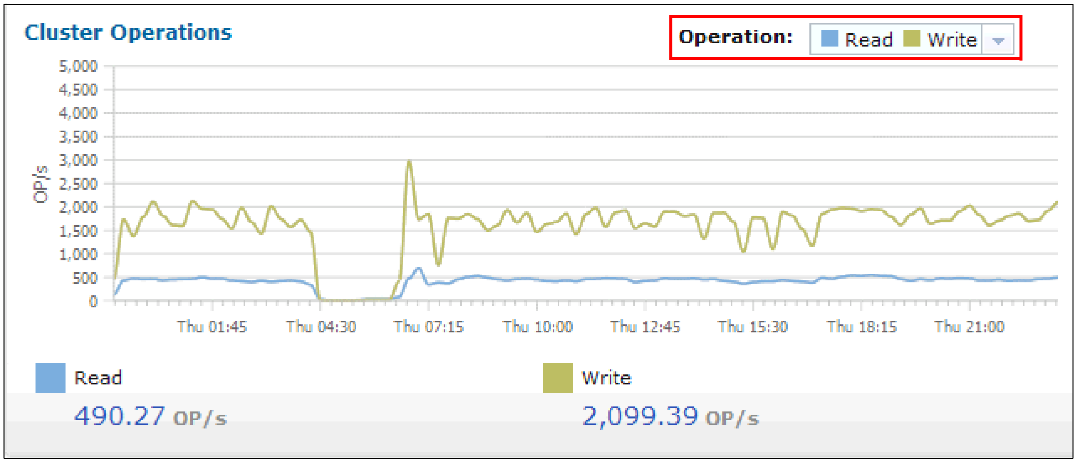


Figure 8-7 Cluster Operations monitoring

Interface and Storage node monitoring

Performance monitoring supports the viewing of performance statistics of CPU, memory, and network for the Interface nodes and the performance statistics of processor and memory for the Storage nodes.

For the processor, it supports the following fields:

- Context Switches
- Hardware Interrupts
- I/O wait
- Idle
- Interrupts
- Nice
- Software interrupts
- System
- User

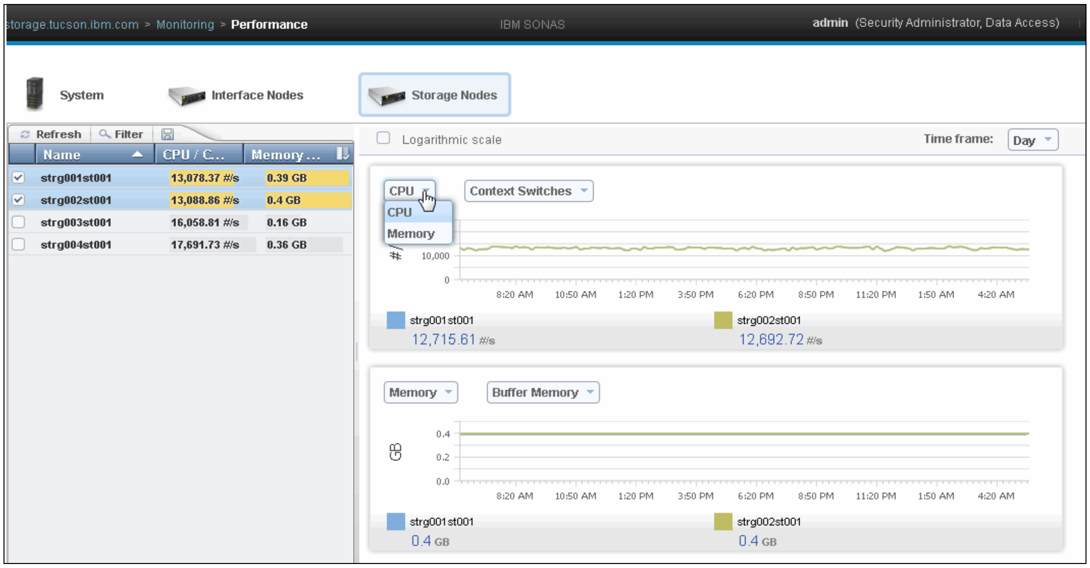
For memory, it supports the following fields:

- Buffer memory
- Cached memory
- Free memory
- Swapped cached memory
- Swap Free memory
- Total memory

For Network, it supports the following fields:

- Bytes received
- Bytes sent
- Collisions
- Drops received
- Drops sent
- Errors received
- Errors sent
- Packets received
- Packets sent

To view the Storage node performance statistics window, click **Storage Nodes**. Figure 8-8 shows the Storage nodes monitoring window. Then, select the Storage nodes whose performance statistics are to be viewed from the left pane. Select either **CPU** or **Memory** from the drop-down list in the right pane and then select the required field from the adjacent drop-down list.



	Name	CPU / C...	Memory ...
☑	strg001st001	13,078.37 #/s	0.39 GB
☑	strg002st001	13,088.86 #/s	0.4 GB
□	strg003st001	16,058.81 #/s	0.16 GB
□	strg004st001	17,691.73 #/s	0.36 GB

□ Logarithmic scale

Time frame: Day

CPU | Context Switches

strg001st001 12,715.61 #/s

strg002st001 12,692.72 #/s

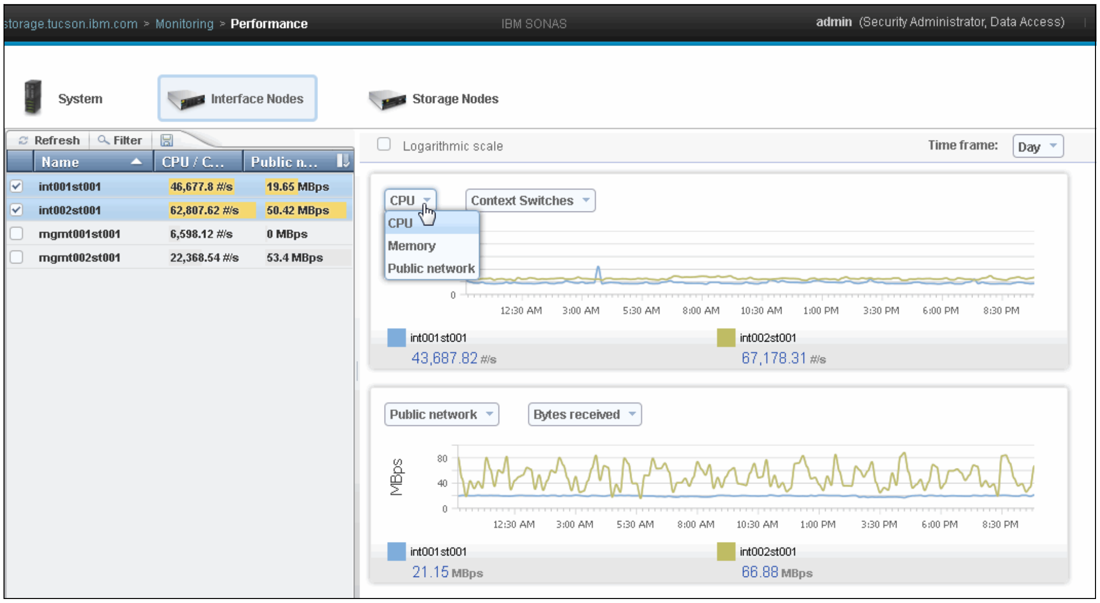
Memory | Buffer Memory

strg001st001 0.4 GB

strg002st001 0.4 GB

Figure 8-8 Storage node performance monitoring

To view the Interface node performance statistics window, click **Interface Nodes**. Figure 8-9 shows the Interface nodes monitoring window. Select the Interface nodes whose performance statistics are to be viewed from the left pane and select either **CPU**, **Memory**, or **Public network** from the drop-down list on the right pane and then select the required field from the adjacent drop-down list.



	Name	CPU / C...	Public n...
☑	int001st001	46,677.8 #/s	19.65 MBps
☑	int002st001	62,807.62 #/s	50.42 MBps
□	mgmt001st001	6,598.12 #/s	0 MBps
□	mgmt002st001	22,368.54 #/s	53.4 MBps

□ Logarithmic scale

Time frame: Day

CPU | Context Switches

int001st001 43,687.82 #/s

int002st001 67,178.31 #/s

Public network | Bytes received

int001st001 21.15 MBps

int002st001 66.88 MBps

Figure 8-9 Interface node performance monitoring

Common tips for viewing graphs

There are a few details that you can get by hovering your cursor over elements of the Performance Center GUI:

- If you hover your cursor over the graph line, it shows you individual samples as connected points. You can see the value of a point by hovering your cursor over that point (Figure 8-10).

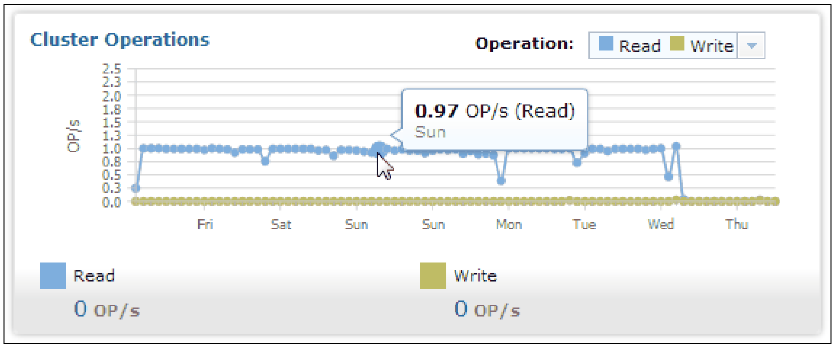


Figure 8-10 Mouse-over graph line

- You can show and hide graph lines by clicking the label of each graph line (Figure 8-11).

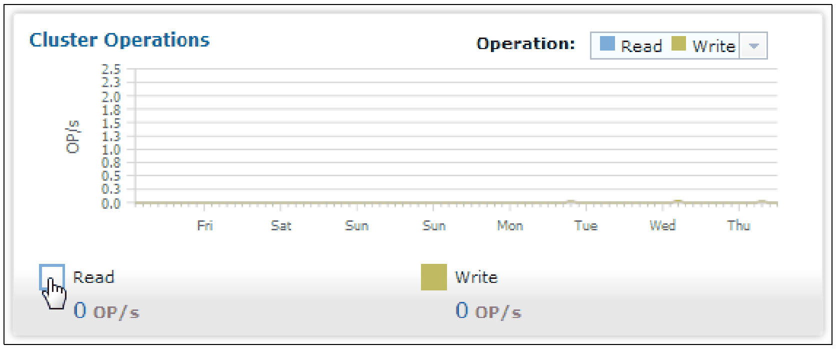
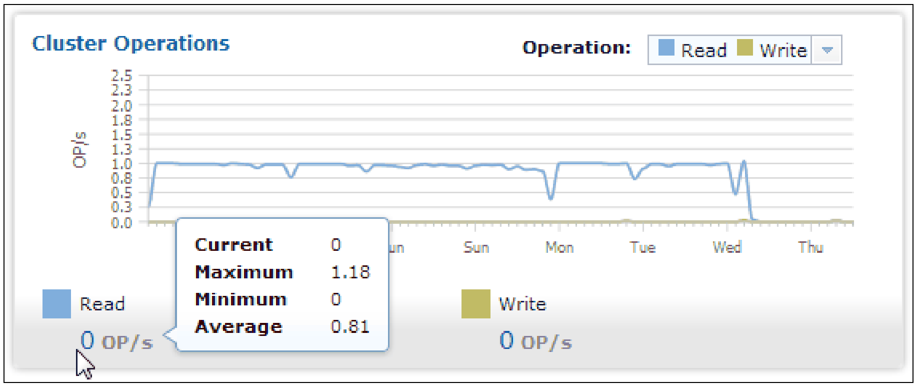


Figure 8-11 Show or hide a graph

- If you hover your cursor over the value (number), a small box opens, which shows Current, Maximum, Minimum, and Average values for that graph line (Figure 8-12).



Current	0
Maximum	1.18
Minimum	0
Average	0.81

Figure 8-12 Mouse-over value

8.3.2 Command-line interface for SONAS Performance Center

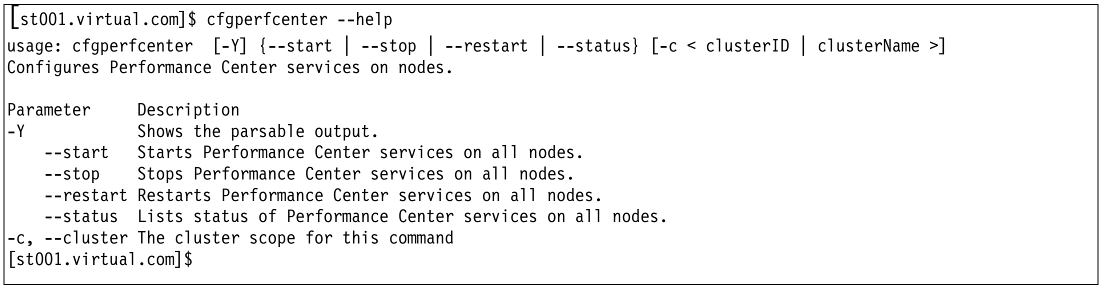
Although the GUI Performance Center provides you graphical information about the client, cluster and file system, the CLI Performance Center provides more detailed information for many components, such as CPU, memory, disk, and the network in numbered values, in addition to what the GUI Performance Center provides. You can save data in a CSV text file, from which you can make a chart or graph.

Commands

The CLI Performance Center consists of two shell commands.

The cfgperfcenter command

This command configures and manages Performance Center services on all nodes. Services are ON by default. You can turn off the services manually only in cases where you do not want the performance impact or the logging space that it takes. Collecting log data causes only a small impact and uses minimal storage space on the system (Figure 8-13).



```
[st001.virtual.com]$ cfgperfcenter --help
usage: cfgperfcenter  [-Y] {--start | --stop | --restart | --status} [-c < clusterID | clusterName >]
Configures Performance Center services on nodes.

Parameter     Description
-Y            Shows the parsable output.
    --start   Starts Performance Center services on all nodes.
    --stop    Stops Performance Center services on all nodes.
    --restart Restarts Performance Center services on all nodes.
    --status  Lists status of Performance Center services on all nodes.
-c, --cluster The cluster scope for this command
[st001.virtual.com]$
```

Figure 8-13 Example of the cfgperfcenter command syntax

The lsperfdata command

This command retrieves data that is collected by Performance Center for various metrics and graphs:

- Name of the graph.
- Nodelist.
- Timeperiod. Each period collects data in a different resolution, as shown in Table 8-2.

Table 8-2 Time period options for the lsperfdata command

Period	Resolution
Minute	Every 1 second
Hour	Every 12 seconds
Day	Every 5 minutes
Week	Every 30 minutes
Month	Every 2 hours
Quarter	Every 8 hours
Year	Every 1 day

Figure 8-14 on page 633 shows the output from the **lsperdata** command with the help option, which shows the available command parameters.

```
[st001.virtual.com]$ lsperfdata --help
usage: lsperfdata {-l | {-g name -t timeperiod | [-n nodelist] | [-f filesystemlist] | [-p poollist] }} [-c {
clusterID | clusterName }]
Retrieves historical performance data as CSV output.

Parameter           Description
-l, --list          Returns a list of graph names whose data can be dumped.
-g, --graph         Specifies the name of the graph for which data is requested.
-t, --timeperiod    Specifies the period for which the data is required. Possible values are minute, hour, week,
day, month, quarter or year.
-n, --nodelist      Set the nodes for which data is required. Possible values are all, interface, storage, or
comma-separated list of node names. Values interface and storage are invalid for an IFS cluster.
-f, --filesystemlist Displays a comma-separated list of file systems.
-p, --poollist      Displays a comma-separated list of pools.
-c, ry--cluster       The cluster scope for this command
[st001.virtual.com]$
```

Figure 8-14 Options for the lsperfdata command

Basic usage

You can specify the kind of base data you want to use to make a graph by using the **-g** (or **--graph**) option. Then, you can set the time period for which you want to collect data by using the **-t** (or **--timeperiod**) option. Here is a sample **lsperfdata** command to make a graph:

```
$ lsperfdata -g name -t timeperiod
```

Here are the measurable metrics of data when you use the **-g** option:

- Client: **client_throughput** retrieves the total bytes received and total bytes sent across the client network interface on all the Interface nodes.
- Cluster:
  - **cluster_create_delete_latency**

    Retrieves the latency of the file create and delete operations across all the file systems on all the nodes of the GPFS cluster.
  - **cluster_create_delete_operations**

    Retrieves the number of file create and delete operations across all the file systems on all the nodes of the GPFS cluster.
  - **cluster_open_close_latency**

    Retrieves the number of file open and close operations across all the file systems on all the nodes of the GPFS cluster.
  - **cluster_read_write_latency**

    Retrieves the latency of file read and write operations across all the file systems on all the nodes of the GPFS cluster.
  - **cluster_read_write_operations**

    Retrieves the number of file read and write operations across all the file systems on all the nodes of the GPFS cluster.
  - **cluster_throughput**

    Retrieves the number of bytes read and written across all the file systems on all the nodes of the GPFS cluster.

- CPU:
  - **`cpu_context_usage`**

    Retrieves the statistics for the CPU context switches for each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_hiq_usage`**

    Retrieves the statistics for the percentage of CPU that is spent on processing hardware interrupts on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_idle_usage`**

    Retrieves the statistics for the percentage of CPU that is spent idle on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_interrupts_usage`**

    Retrieves the statistics for the total interrupts that are processed by the CPU on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_iowait_usage`**

    Retrieves the statistics for the percentage of CPU that is spent waiting for IO to complete on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_nice_usage`**

    Retrieves the statistics for the percentage of CPU that is spent on processing nice processes on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_siq_usage`**

    Retrieves the statistics for the percentage of CPU that is spent on processing software interrupts on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_stats`**

    Retrieves the CPU stats for a node.
  - **`cpu_system_usage`**

    Retrieves the statistics for the percentage of CPU that is spent on processes that are running in kernel mode on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`cpu_user_usage`**

    Retrieves the statistics for the percentage of CPU that is spent of processes running in the user mode on each of the nodes that are specified in the **`nodelist`** parameter.
- Disk:
  - **`disk_reads`**

    Retrieves the number of read operations that are completed on each of the disks on each of the nodes that are specified in the **`nodelist`** parameter.
  - **`disk_stats`**

    Retrieves the number or read and write operations that are completed on each of the disks for a node.
  - **`disk_writes`**

    Retrieves the number of write operations that are completed on each of the disks on each of the nodes that are specified in the **`nodelist`** parameter.

- File system: **filesystem_throughput** retrieves the number of bytes read and written on each of the GPFS file systems that are specified in the **filesystemlist** option.
- Memory:
  - **memory_buffers_usage**

    Retrieves the statistics of the amount of memory that is used as file buffers on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_cached_usage**

    Retrieves the statistics of the amount of memory that is used as the cache on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_dirty_usage**

    Retrieves the statistics of the amount of dirty memory on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_free_usage**

    Retrieves the statistics of the amount of free memory on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_stats**

    Retrieves the memory stats for a node.
  - **memory_swapcached_usage**

    Retrieves the statistics of the amount of swap memory that is used as cache on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_swapfree_usage**

    Retrieves the statistics of the amount of free swap memory on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_swaptotal_usage**

    Retrieves the statistics of the total amount of swap memory on each of the nodes that are specified in the **nodelist** parameter.
  - **memory_total_usage**

    Retrieves the statistics of the total amount of memory on each of the nodes that are specified in the **nodelist** parameter.
- Network:
  - **network_bytes_received**

    Retrieves the statistics of the number of bytes that are received on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.
  - **network_bytes_sent**

    Retrieves the statistics of the number of bytes that are sent on each of the client network interfaces on each of the nodes that are specified in the nodelist parameter
  - **network_collisions**

    Retrieves the statistics of the number of collisions on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.
  - **network_drops_received**

    Retrieves the statistics of the number of drops that are received on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.

- **network_drops_sent**

  Retrieves the statistics of the number of drops that are sent on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.

- **network_errors_received**

  Retrieves the statistics of the number of errors that are received on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.

- **network_errors_sent**

  Retrieves the statistics of the number of errors that are sent on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.

- **network_packets_received**

  Retrieves the statistics of the number of packets that are received on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.

- **network_packets_sent**

  Retrieves the statistics of the number of packets that are sent on each of the client network interfaces on each of the nodes that are specified in the **nodelist** parameter.

- **network_stats**

  Retrieves the network stats for a node.

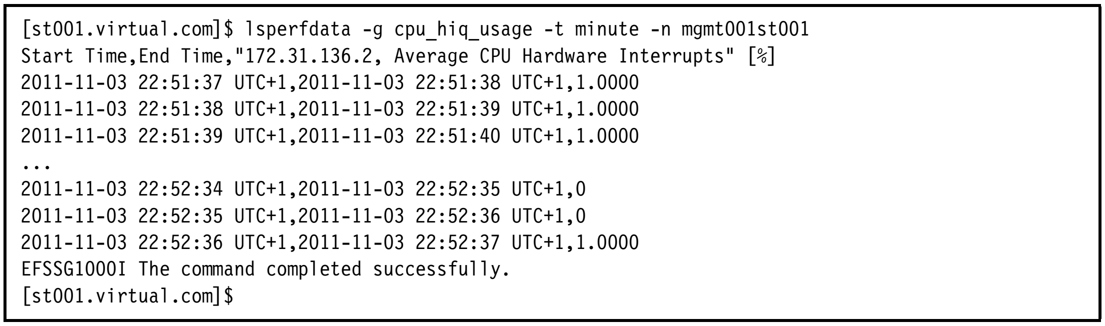
- Pool: **pool_throughput** retrieves the number of bytes read and written on each of the pools that are specified in the **poollist** option.

8.3.3 Importing a saved data text file into MS Excel

You can import **lsperfdata** output data into a spreadsheet editor, such as Microsoft Excel, to organize the data or to make a graph out of it.

Saving output data

You can drag the output screen, then copy and paste it into a text editor and save it (see Figure 8-15). Or, you can save the output directly into a file on client host shell (it must be done on the client host side because you do not have root access to the SONAS system).



```
[st001.virtual.com]$ lsperfdata -g cpu_hiq_usage -t minute -n mgmt001st001
Start Time,End Time,"172.31.136.2, Average CPU Hardware Interrupts" [%]
2011-11-03 22:51:37 UTC+1,2011-11-03 22:51:38 UTC+1,1.0000
2011-11-03 22:51:38 UTC+1,2011-11-03 22:51:39 UTC+1,1.0000
2011-11-03 22:51:39 UTC+1,2011-11-03 22:51:40 UTC+1,1.0000
...
2011-11-03 22:52:34 UTC+1,2011-11-03 22:52:35 UTC+1,0
2011-11-03 22:52:35 UTC+1,2011-11-03 22:52:36 UTC+1,0
2011-11-03 22:52:36 UTC+1,2011-11-03 22:52:37 UTC+1,1.0000
EFSSG1000I The command completed successfully.
[st001.virtual.com]$
```

Figure 8-15 Save screen-copied output data into a text editor

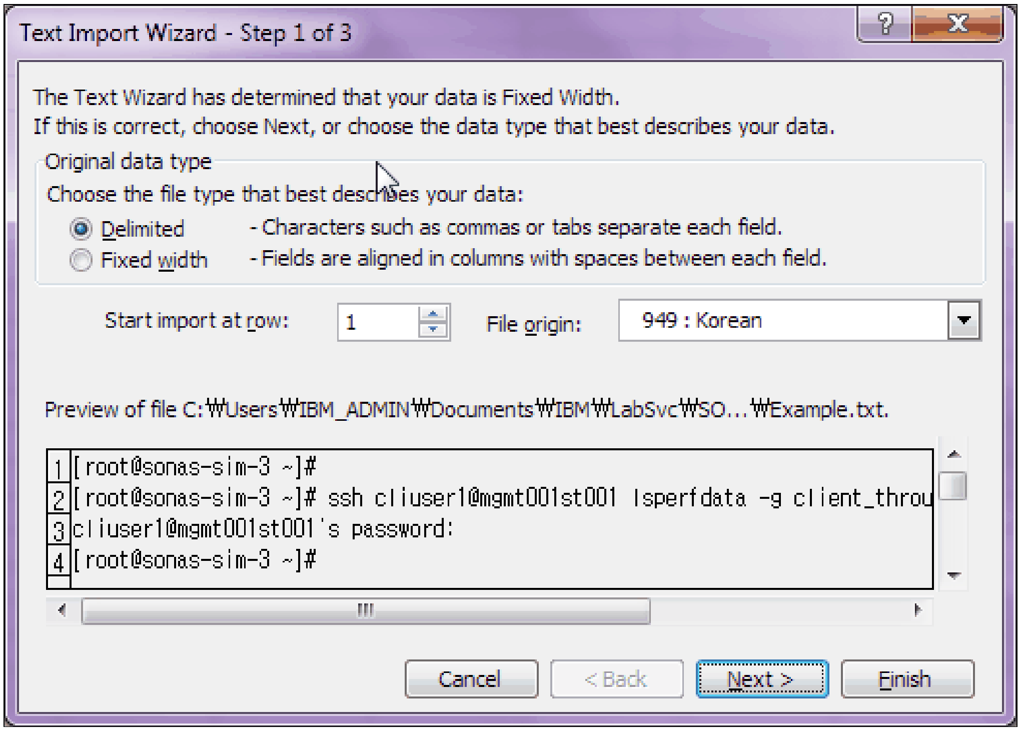
Sample scenario

A sample scenario for importing data into a spreadsheet follows. Complete the following steps:

1. Run the following command:

```
# lsperfdata [-g graph_name] [-t time_period] [-n node_name] > [
output_name.txt]
```

2. Open the saved file that is named `output_name.txt` in Microsoft Excel. You see the Text Import Wizard window. Click **Delimited** and then click **Next** (see Figure 8-16).



Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Fixed Width.
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.
○ Fixed width - Fields are aligned in columns with spaces between each field.

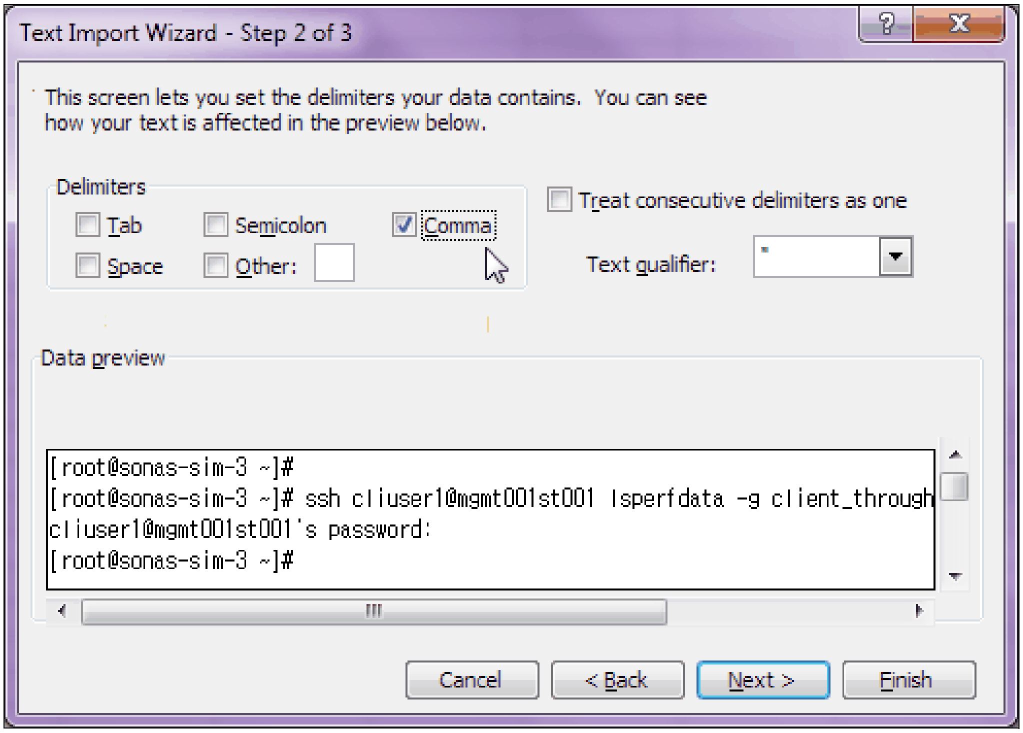
Start import at row: 1 File origin: 949 : Korean

Preview of file C:₩Users₩IBM_ADMIN₩Documents₩IBM₩LabSvc₩SO...₩Example.txt.

```
[root@sonas-sim-3 ~]#
[root@sonas-sim-3 ~]# ssh cliuser1@mgmt001st001 lsperfdata -g client_throu
cliuser1@mgmt001st001's password:
[root@sonas-sim-3 ~]#
```

Figure 8-16 Text Import Wizard - Step 1 of 3

3. The text contents in the saved file are delimited by comma (,), so select **Comma** and click **Next**. See Figure 8-17.



Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab □ Semicolon ☑ Comma

□ Space □ Other:

□ Treat consecutive delimiters as one

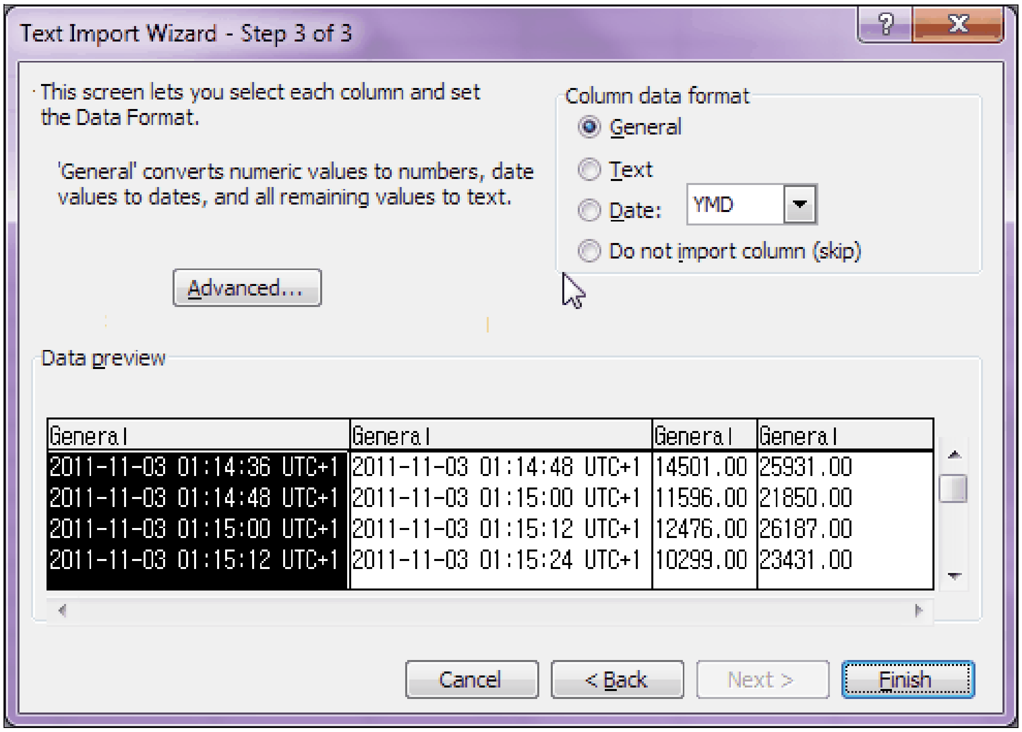
Text qualifier: "

Data preview

```
[root@sonas-sim-3 ~]#
[root@sonas-sim-3 ~]# ssh cliuser1@mgmt001st001 lsperfdata -g client_through
cliuser1@mgmt001st001's password:
[root@sonas-sim-3 ~]#
```

Figure 8-17 Text Import Wizard - Step 2 of 3

4. You can see that the text is divided into columns under the preview pane. Click **Finish**. See Figure 8-18.



Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Column data format

● General

○ Text

○ Date: YMD

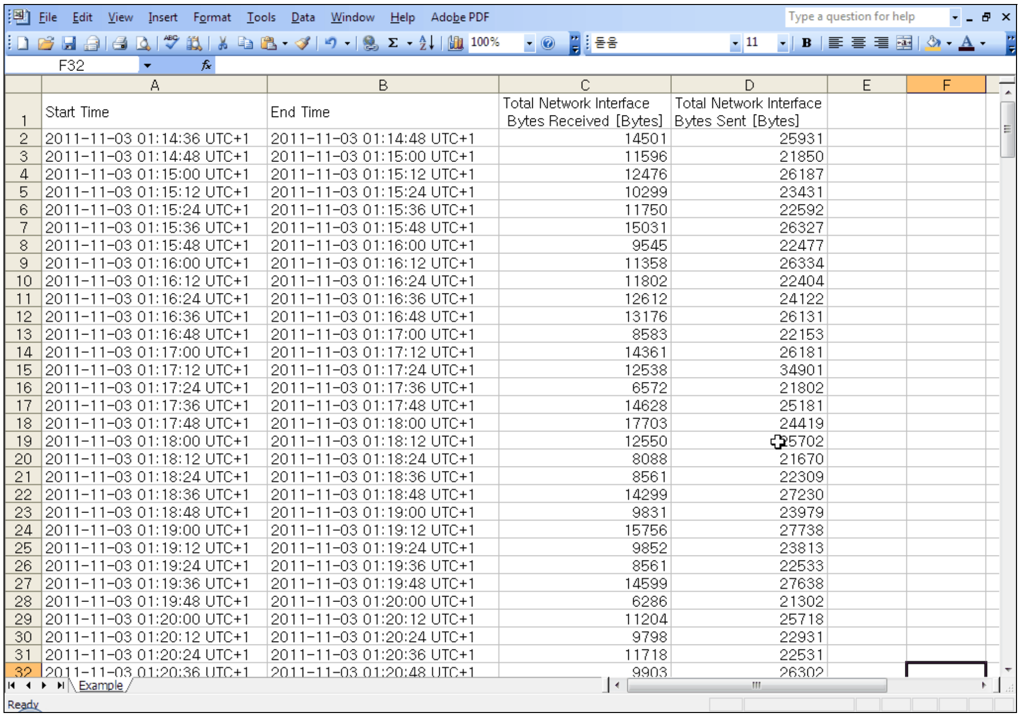
○ Do not import column (skip)

Data preview

General	General	General	General
2011-11-03 01:14:36 UTC+1	2011-11-03 01:14:48 UTC+1	14501.00	25931.00
2011-11-03 01:14:48 UTC+1	2011-11-03 01:15:00 UTC+1	11596.00	21850.00
2011-11-03 01:15:00 UTC+1	2011-11-03 01:15:12 UTC+1	12476.00	26187.00
2011-11-03 01:15:12 UTC+1	2011-11-03 01:15:24 UTC+1	10299.00	23431.00

Figure 8-18 Text Import Wizard - Step 3 of 3

5. Delete unnecessary rows and modify the column width for better viewing (see Figure 8-19). You can make a customized graph out of this data.



	A	B	C	D	E	F
1	Start Time	End Time	Total Network Interface Bytes Received [Bytes]	Total Network Interface Bytes Sent [Bytes]		
2	2011-11-03 01:14:36 UTC+1	2011-11-03 01:14:48 UTC+1	14501	25931		
3	2011-11-03 01:14:48 UTC+1	2011-11-03 01:15:00 UTC+1	11596	21850		
4	2011-11-03 01:15:00 UTC+1	2011-11-03 01:15:12 UTC+1	12476	26187		
5	2011-11-03 01:15:12 UTC+1	2011-11-03 01:15:24 UTC+1	10299	23431		
6	2011-11-03 01:15:24 UTC+1	2011-11-03 01:15:36 UTC+1	11750	22592		
7	2011-11-03 01:15:36 UTC+1	2011-11-03 01:15:48 UTC+1	15031	26327		
8	2011-11-03 01:15:48 UTC+1	2011-11-03 01:16:00 UTC+1	9545	22477		
9	2011-11-03 01:16:00 UTC+1	2011-11-03 01:16:12 UTC+1	11358	26334		
10	2011-11-03 01:16:12 UTC+1	2011-11-03 01:16:24 UTC+1	11802	22404		
11	2011-11-03 01:16:24 UTC+1	2011-11-03 01:16:36 UTC+1	12612	24122		
12	2011-11-03 01:16:36 UTC+1	2011-11-03 01:16:48 UTC+1	13176	26131		
13	2011-11-03 01:16:48 UTC+1	2011-11-03 01:17:00 UTC+1	8583	22153		
14	2011-11-03 01:17:00 UTC+1	2011-11-03 01:17:12 UTC+1	14361	26181		
15	2011-11-03 01:17:12 UTC+1	2011-11-03 01:17:24 UTC+1	12538	34901		
16	2011-11-03 01:17:24 UTC+1	2011-11-03 01:17:36 UTC+1	6572	21802		
17	2011-11-03 01:17:36 UTC+1	2011-11-03 01:17:48 UTC+1	14628	25181		
18	2011-11-03 01:17:48 UTC+1	2011-11-03 01:18:00 UTC+1	17703	24419		
19	2011-11-03 01:18:00 UTC+1	2011-11-03 01:18:12 UTC+1	12550	25702		
20	2011-11-03 01:18:12 UTC+1	2011-11-03 01:18:24 UTC+1	8088	21670		
21	2011-11-03 01:18:24 UTC+1	2011-11-03 01:18:36 UTC+1	8561	22309		
22	2011-11-03 01:18:36 UTC+1	2011-11-03 01:18:48 UTC+1	14299	27230		
23	2011-11-03 01:18:48 UTC+1	2011-11-03 01:19:00 UTC+1	9831	23979		
24	2011-11-03 01:19:00 UTC+1	2011-11-03 01:19:12 UTC+1	15756	27738		
25	2011-11-03 01:19:12 UTC+1	2011-11-03 01:19:24 UTC+1	9852	23813		
26	2011-11-03 01:19:24 UTC+1	2011-11-03 01:19:36 UTC+1	8561	22533		
27	2011-11-03 01:19:36 UTC+1	2011-11-03 01:19:48 UTC+1	14599	27638		
28	2011-11-03 01:19:48 UTC+1	2011-11-03 01:20:00 UTC+1	6286	21302		
29	2011-11-03 01:20:00 UTC+1	2011-11-03 01:20:12 UTC+1	11204	25718		
30	2011-11-03 01:20:12 UTC+1	2011-11-03 01:20:24 UTC+1	9798	22931		
31	2011-11-03 01:20:24 UTC+1	2011-11-03 01:20:36 UTC+1	11718	22531		
32	2011-11-03 01:20:36 UTC+1	2011-11-03 01:20:48 UTC+1	9903	26302		

Figure 8-19 lsperfdata output file imported into MS Excel

8.4 Capacity

You can monitor the capacity of the SONAS system from the GUI. Typically, the capacity for the file system, file set, users, and groups can be monitored. Capacity is commonly known as *quotas*. You can set the quota for a file system, file set, users, and groups. For more information about quotas and how to set them, see Chapter 7, “SONAS administration” on page 493.

After you set quotas, you can monitor whether they are within limits with the Monitoring window in the GUI. You can also set a notification to notify automatically the administrator. This section describes how to monitor quotas and capacity manually by using the GUI and CLI.

8.4.1 Monitoring quotas with the GUI

Quotas can be checked separately for file systems, file sets, users, and groups by using the GUI. Click the **Monitor** icon and then click **Capacity**, as shown in Figure 8-20.

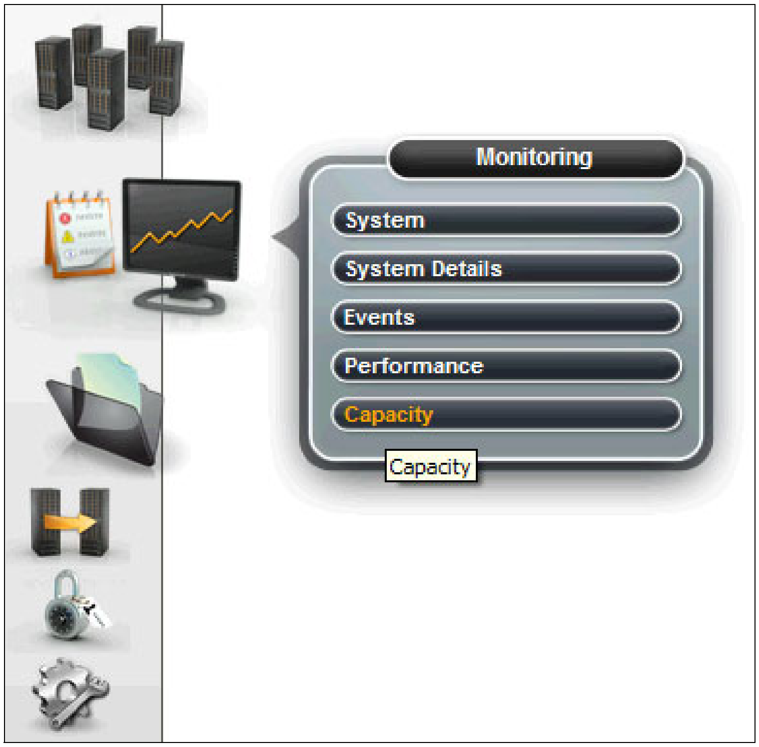
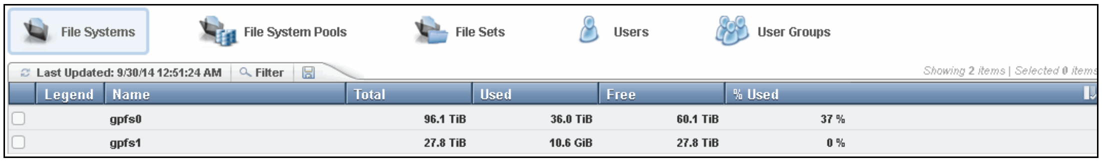


Figure 8-20 Monitoring capacity from the GUI

You see a window with different options from which to choose. They are for file systems, file sets, users, and groups. Clicking any of the tabs displays information about its quota (see Figure 8-21).


File Systems
File System Pools
File Sets
Users
User Groups
Last Updated: 9/30/14 12:51:24 AM
Filter
Showing 2 items | Selected 0 items

	Legend	Name	Total	Used	Free	% Used
☐		gpfs0	96.1 TiB	36.0 TiB	60.1 TiB	37 %
☐		gpfs1	27.8 TiB	10.6 GiB	27.8 TiB	0 %

Figure 8-21 GUI displaying quota options to monitor

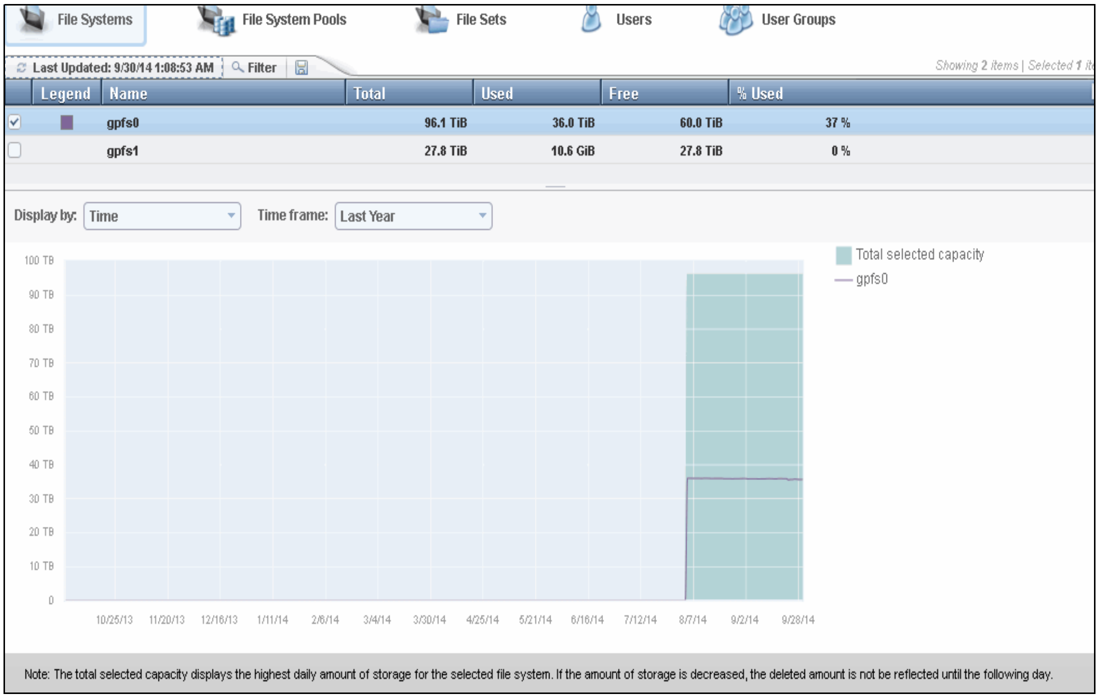
Quotas for file systems

When you click the **File System** tab, you see a table that lists the file systems that exist on the cluster. You can choose the file system whose quota you want to check. Select the check box beside the file system from the table.

You also must choose whether you want the quota and capacity to be displayed by “Time” or “Percentage”. Choosing “Time” displays the quota as a graph with time on the X-axis. However, if you choose “Percentage”, you see the quota as a pie chart.

Figure 8-22 on page 641 shows an example of viewing quota of a “File system” displayed by “Time”. In this example, the gpfs0 file system is chosen. The time frame is “Last year”.

As you can see, the top level of the shaded area is the total capacity that is allotted for “gpfs0”. The dark graph line shows the capacity that is used by “gpfs0”. In this example, the total capacity is 96.1 TB, and the used capacity is close to 36 TB.


File Systems
File System Pools
File Sets
Users
User Groups
Last Updated: 9/30/14 1:08:53 AM
Filter
Showing 2 items | Selected 1 it

	Legend	Name	Total	Used	Free	% Used
☑	▇	gpfs0	96.1 TiB	36.0 TiB	60.0 TiB	37 %
□		gpfs1	27.8 TiB	10.6 GiB	27.8 TiB	0 %

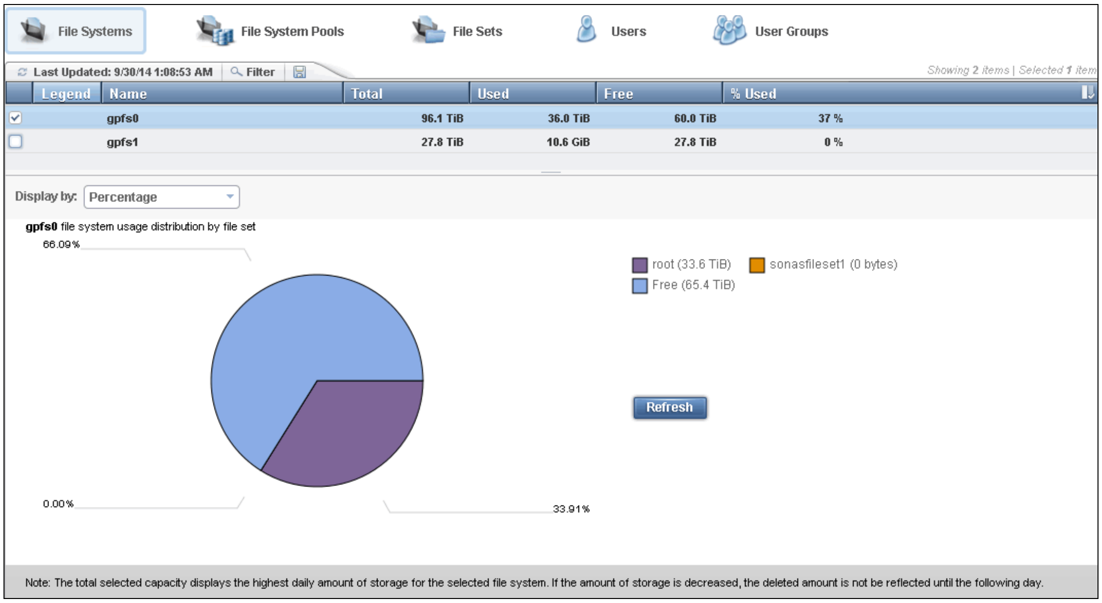
Display by: Time

Time frame: Last Year

Note: The total selected capacity displays the highest daily amount of storage for the selected file system. If the amount of storage is decreased, the deleted amount is not be reflected until the following day.

Figure 8-22 GUI displaying the file system quota by time for the last 30 days

Figure 8-23 shows how the quota for the file system can be monitored as a percentage. A pie chart displays the free capacity and used capacity as a percentage. Together, they display the total capacity for the file system.


File Systems
File System Pools
File Sets
Users
User Groups
Last Updated: 9/30/14 1:08:53 AM
Filter
Showing 2 items | Selected 1 item

	Legend	Name	Total	Used	Free	% Used
☑		gpfs0	96.1 TiB	36.0 TiB	60.0 TiB	37 %
□		gpfs1	27.8 TiB	10.6 GiB	27.8 TiB	0 %

Display by: Percentage

Note: The total selected capacity displays the highest daily amount of storage for the selected file system. If the amount of storage is decreased, the deleted amount is not be reflected until the following day.

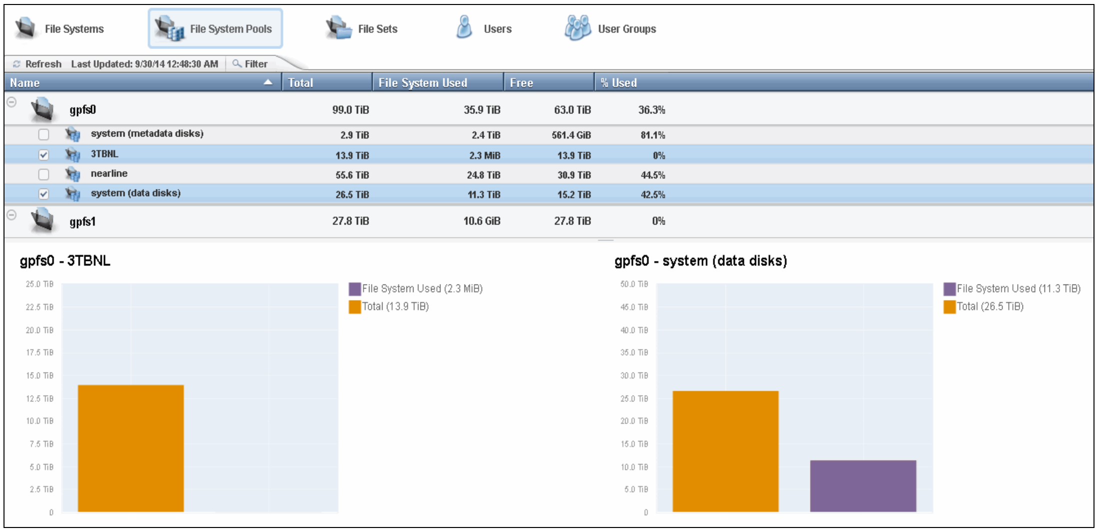
Figure 8-23 GUI displaying the file system quota as a percentage

If you hover your cursor over the total and used capacity, you can see what the color signifies. In this example, the “gpfs0” file system is chosen to check for the capacity by using “Percentage”. The total capacity for the file system is 96.1 TB. The pie chart shows two parts, that is, the free space that is available on that file system and the space that is used by the root file system. In Figure 8-23 on page 641, the amount of free space is 65.4 TB, which is 66.09%, and the amount that is used by root is 33.6 TB, which is about 33.91%.

Usage by file system pool

Click the **File System Pools** tab to see a table that lists the file system. Expand the lists for pools that belong to that file system. You can choose pools for which to check capacity and quota. Select the check box beside the pool name from the table. See Figure 8-24.

In the example in Figure 8-24, two pools, System and 3TBNL, which belong to the gpfs0 file system and the disk space that is used by the pools are displayed along with the total capacity of the pool with a bar pattern. In this example, the 3TBNL pool has a total capacity of 13.9 TB and almost the complete pool is free. The system pool in gpfs0 has a total capacity of 26.5 TB and has a free space of 15.4 TB, which is 42% usage of the total space in that pool.

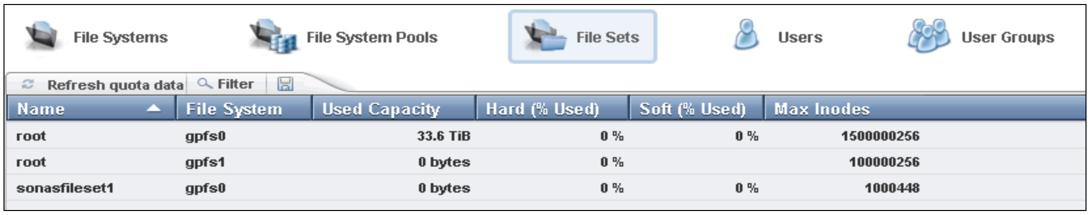


Name	Total	File System Used	Free	% Used
gpfs0	99.0 TiB	35.9 TiB	63.0 TiB	36.3%
□ system (metadata disks)	2.9 TiB	2.4 TiB	561.4 GiB	81.1%
☑ 3TBNL	13.9 TiB	2.3 MiB	13.9 TiB	0%
□ nearline	55.6 TiB	24.8 TiB	30.9 TiB	44.5%
☑ system (data disks)	26.5 TiB	11.3 TiB	15.2 TiB	42.5%
gpfs1	27.8 TiB	10.6 GiB	27.8 TiB	0%

Figure 8-24 File system storage pool view

Quotas for file sets

Click the **File Set** tab to display a table that lists the file sets that exist on the cluster. The table also displays the quota for that file set along with the soft limit on the file set and also the hard limit for the quota on the file set. See Figure 8-25.

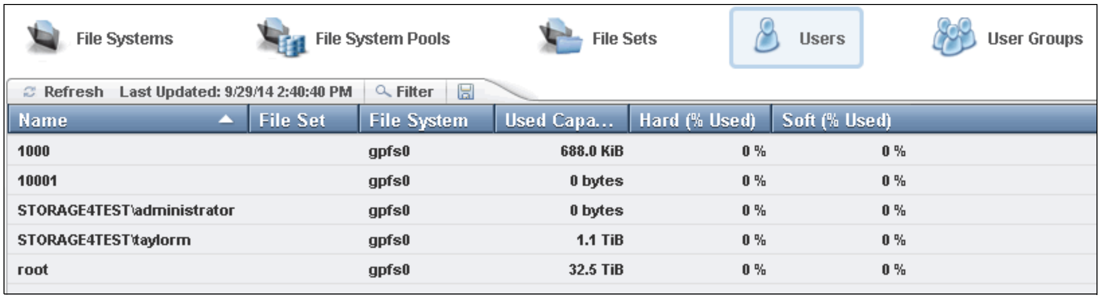


Name	File System	Used Capacity	Hard (% Used)	Soft (% Used)	Max Inodes
root	gpfs0	33.6 TiB	0 %	0 %	1500000256
root	gpfs1	0 bytes	0 %		100000256
sonasfileset1	gpfs0	0 bytes	0 %	0 %	1000448

Figure 8-25 GUI that displays the quota for the file set

Quotas for users

Click the **Users** tab to display the table that lists the users and the file systems and file sets for which the quota is set. Also displayed in the table is the quota for that file set, the soft limit on the file set, and the hard limit of the quota on the file set. See Figure 8-26 on page 643.

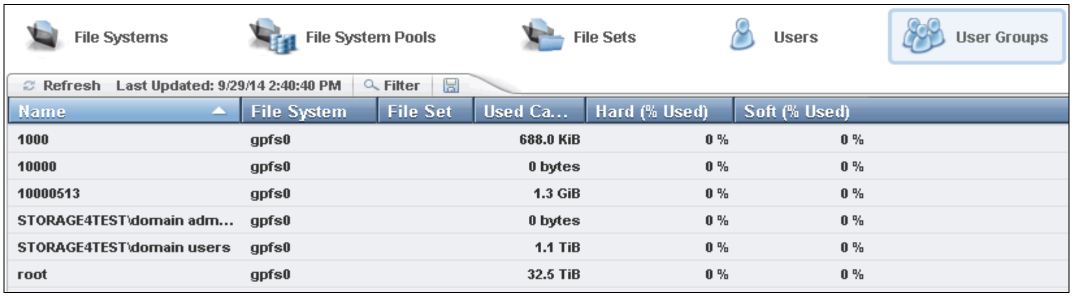


Name	File Set	File System	Used Capa...	Hard (% Used)	Soft (% Used)
1000		gpfs0	688.0 KiB	0 %	0 %
10001		gpfs0	0 bytes	0 %	0 %
STORAGE4TEST\administrator		gpfs0	0 bytes	0 %	0 %
STORAGE4TEST\taylorm		gpfs0	1.1 TiB	0 %	0 %
root		gpfs0	32.5 TiB	0 %	0 %

Figure 8-26 GUI displaying the Users quota

Quotas for groups

Click the **User Groups** to see a table that lists the user groups, and the file systems and file sets for which the quota is set. The quota for that file set, the soft limit on the file set, and the hard limit of quota on the file set are also displayed in the table. See Figure 8-27.



Name	File System	File Set	Used Ca...	Hard (% Used)	Soft (% Used)
1000	gpfs0		688.0 KiB	0 %	0 %
10000	gpfs0		0 bytes	0 %	0 %
10000513	gpfs0		1.3 GiB	0 %	0 %
STORAGE4TEST\domain adm...	gpfs0		0 bytes	0 %	0 %
STORAGE4TEST\domain users	gpfs0		1.1 TiB	0 %	0 %
root	gpfs0		32.5 TiB	0 %	0 %

Figure 8-27 GUI displaying the Group Quota

8.4.2 Monitoring quotas with the CLI

You can monitor quotas from the CLI by running **lsquota**. The command lists all the quotas that belong to the cluster. It retrieves data about the quota that is managed by the Management node from the database and returns a list in either a human-readable format or in a format that can be parsed. Run **chkquota** before running **lsquota** to refresh the quota information in the GPFS file.

The **chkquota** command checks the quota for a user, a group, or a file set. It recounts inode and space usage in a file system by user, group, and file set, and writes the collected data into the database.

The help for the **chkquota** command is shown in Example 8-1.

Example 8-1 Help for the chkquota command

```
# chkquota --help
usage: chkquota device  [--force] [-c < clusterID | clusterName >]
Checks file system user, group, and file set quotas.

Parameter     Description
device        Specifies the mount point or device of the file system.
    --force   Forces quota check without prompting for manual confirmation.
-c, --cluster The cluster scope for this command
```

The help for the **lsquota** command is shown in Example 8-2.

Example 8-2 Help for the lsquota command

```
# lsquota --help
usage: lsquota  [-r] [-Y] [-d <device>] [-j <fileset> | -u <userName> | -g <groupName>] [-c <
clusterID | clusterName >]
Lists all quotas.

Parameter     Description
-r, --refresh Forces a refresh of the quota data in the database by scanning the cluster before
retrieving data from the database.
-Y            Shows parseable output.
-d, --device  Lists quota information for a given device.
-j, --fileset Lists quota information for a given file set, or all file sets if none is given.
-u, --user    Lists quota information for a given user, or all users if none is given.
-g, --group   Lists quota information for a given group, or all groups if none is given.
-c, --cluster The cluster scope for this command
```

> **Tip:** The **chkquota** command is I/O intensive. Run it when the system load is light. Other file-system-related commands, such as **linkfset**, **mkfs**, and **chfs**, might be unable to run while **chkquota** is in progress.

8.5 Other important cluster monitoring considerations

In addition to the standard GUI and CLI monitoring, there are special feature and performance impacting metrics that must be considered to get a full understanding of cluster services across the critical resources at the heart of the cluster and the front and back end. These metric and features include the following items:

- Number of concurrent user connections
- Interface node CPU and network utilization
- Storage workloads
- Replication success
- Snapshots that are collected and consumption
- Storage pool capacity monitoring
- Number of inodes available per file system / file sets

8.5.1 Number of concurrent user connections

In most clusters, ensure that the users remain balanced across all active Interface nodes. A similar number of active concurrent users are on Interface node 1 and on Interface node 2.

The number of concurrent users on all Interface nodes can be determined by running the command that is shown in Figure 8-28 on page 645 as a “Privileged” (root) user on the active Management node.

```
[root@xivsonas.mgmt001st001 ~]# onnode all "ctdb statistics |grep clients"
>> NODE: 172.31.136.2 <<
 num_clients                    11
>> NODE: 172.31.136.3 <<
 num_clients                    11
```

Figure 8-28 Check for the number of concurrent users on each Interface node

In SONAS, concurrent users refer to the number of users on that specific Interface node. So, if you have 100 concurrent users on a cluster and you have four Interface nodes, you can expect to see about 25 concurrent user per Interface node. This amount can differ by a small percentage and not indicate a problem, but if the numbers are vastly different, it might be helpful to disperse connectivity or the possible level of activity.

8.5.2 Interface node CPU and network utilization

The Interface nodes of a cluster should provide balanced workload processing when they are correctly configured. By dispersing client connections evenly, that is the typical result. However, in some environments with fewer clients, some clients can produce a heavier burden on resources than others. Occasionally monitor or measure that effective distribution and analyze opportunities for improvements when the cluster is not balanced.

The front end of the cluster is typically understood as the SONAS Interface nodes. Resources that are affected by client activity on the front end include CPU, memory, and network.

For network performance, it is preferable to see that each Interface node is passing similar workloads, which you can determine by viewing the **sdstat** command output from each node.

In Figure 8-29, which is taken from a quiet lab environment, there is little to see. However, when you watch a production environment, it quickly becomes evident when Sent and Received add up to line speeds, or when one node is not performing at the same network throughput as the other Interface nodes.

```
[root@xivsonas.mgmt001st001 etc]# onnode all "sdstat -n -c -M gpfsops 1 4"

>> NODE: 172.31.136.2 <<
/usr/bin/sdstat:1422: DeprecationWarning: os.popen3 is deprecated.  Use the subprocess
module.
  pipes[cmd] = os.popen3(cmd, 't', 0)
Terminal width too small, trimming output.
-net/total- ----total-cpu-usage---->
 recv  send|usr sys idl wai hiq siq>
   0     0 |  2   4  93   1   0   1>
1253M 1674k|  8  14  74   0   0   4>
 292k  387k|  2   6  92   0   0   0>
 315k  432k|  1   3  97   0   0   0>
 241k  389k|  0   2  97   0   0   0>

>> NODE: 172.31.136.3 <<
/usr/bin/sdstat:1422: DeprecationWarning: os.popen3 is deprecated.  Use the subprocess
module.
  pipes[cmd] = os.popen3(cmd, 't', 0)
Terminal width too small, trimming output.
-net/total- ----total-cpu-usage---->
 recv  send|usr sys idl wai hiq siq>
   0     0 |  2   3  94   1   0   0>
 206k  139k|  0   2  98   0   0   0>
1189k   78k|  1   2  97   0   0   0>
4374k   74k|  5   8  87   0   0   0>
  98k   40k|  0   2  98   0   0   0>
```

Figure 8-29 Example sdstat output from each Interface node

Figure 8-30 is output from two Interface nodes that are dissimilarly configured in a quiet lab (HW). However, in a production environment, you can quickly see whether one node is working harder than the others by running **vmstat**.

```
[root@xivsonas.mgmt001st001 etc]# onnode all "vmstat 1 3"

>> NODE: 172.31.136.2 <<
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 3  0      0 3724424 458748 7345288    0    0     0    10    2    2  2  4 93  1  0
 0  0      0 3725780 458748 7345300    0    0     0   140 14387 15277  1  6 92  1  0
 3  0      0 3721916 458748 7345300    0    0     0   136 31064 29583  3 10 87  0  0

>> NODE: 172.31.136.3 <<
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0  0      0 14841644 401192 5454828    0    0     0     9    0    2  2  3 94  1  0
 1  0      0 14844048 401192 5454832    0    0     0    92 1343 1639  0  2 98  0  0
 4  1      0 14833424 401192 5454832    0    0     0   172 5522 5707  2  3 94  1  0
```

Figure 8-30 Example vmstat command output from each of the Interface nodes

8.5.3 Storage workloads

Back-end resource analysis happens on the Storage node pairs. Two Storage nodes that sit in front of a back-end spindle management device array should always balance that workload fairly evenly.

Because each NSD put into the GPFS configuration behind SONAS is assigned a Storage node preference, it is important to configure NSDs in *pairs*. Each pair must be assigned to file systems in a balanced configuration, where each NSD alternates the Storage node preference in the file system. If a file system is composed of a single NSD, only one Storage node might be assigned to do back-end work, which causes performance problems.

The Storage node preference depicts which Storage node does the work for that NSD (unless that depicted Storage node fails). This behavior occurs because the NSDs are added in an active/passive fashion. The auxiliary storage node preference does not work for that NSD unless the primary Storage node fails. The **mmlsnsd -L** command lists NSDs in their Storage node preference.

Figure 8-31 shows three NSDs that are assigned to the GPFS0 file system. In this scenario, strg001st001 is preferred by two of the NSDs, and strg002st001 is only preferred by one NSD. This scenario creates an unbalanced configuration where strg001st001 works twice as much as strg002st001.

```
[root@xivsonas.mgmt001st001 etc]# mmlsnsd -L

 File system   Disk name    NSD volume ID      NSD servers
---------------------------------------------------------------------------------------------
 gpfs0         DCS3700_360080e50002ec48600000388502104ad AC1F86015048F9D2   strg001st001,strg002st001
 gpfs0         DCS3700_360080e50002ee71e0000030f50210435 AC1F86025048F9CD   strg002st001,strg001st001
 gpfs0         DCS3700_360080e50002ec4860000039150210656 AC1F86015048F9F0   strg001st001,strg002st001
```

Figure 8-31 Output from the mmlsnsd -L command

Each NSD should receive a balanced effort of I/O because of NSD striping patterns and most normal workload processing behavior. NSD1 is expected to work as much as NSD2 and NSD3. In this case, two of the three NSDs are in one Storage node, which causes that node to work harder than the other.

One way to look at NSD workload management success is to log in to the Storage nodes with privileged access and look at the IOSTAT output. See Figure 8-32.

```
onnode -n strg001st001,strg002st001 "iostat -xm /dev/dm* 1 10"

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.36    0.00    0.62    0.06    0.00   98.96

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await  svctm  %util
dm-4              0.00     0.00    0.00    0.00     0.00     0.00     7.00     0.00    7.10   7.10   0.00
dm-5              0.00     0.00    0.00    0.00     0.00     0.00     6.98     0.00    7.06   7.05   0.00
```

Figure 8-32 Command to view NSD performance on both strg001 and strg002 Storage nodes

The command in Figure 8-32 shows 10 instance snapshots on performance, one second apart for all the devices on each of the two Storage nodes. When all the NSDs are pushing constantly at 100% busy, it might be an indication of a poor layout or a need to add spindles to the targeted file system. Adding a second, like-sized storage pod and adding the resources to the SONAS file system targeted reduces the workload of each Pod by two times.

8.5.4 Replication success

Replication affects your ability to recover your critical file systems and performance. Every action against your front- and back-end devices competes with all the others for metadata scan time and resource I/O bandwidth. For this reason, it is helpful to start replication on a modest level and monitor the effects as you tighten the RPO or increase the amount of data to be replicated.

Get a sample of front- and back-end performance during peak hours before you enable replication. That process helps understand the effect of replication in your environment. As you add millions of files, you can expect metadata scan times to increase (because it has more files to scan), and replication completion windows to grow (because you have more data to replicate).

For more information about how to configure and monitor the success of replication, see 6.8, “Managing asynchronous replication” on page 444. The information in this section is provided as a reminder that, when your demands grow, you must monitor your resources to ensure that you understand when that growth rate impacts your cluster success rates and when it might indicate the need to grow your cluster resources.

8.5.5 Snapshots and space usage

A snapshot of an entire file system or of an independent file set can be created to preserve the contents of the file system or the independent file set at a single point in time. You cannot create a snapshot of a dependent file set. The storage that is needed for maintaining a snapshot is because of the required retention of a copy of all of the data blocks that are changed or deleted after the time of the snapshot, and is charged against the file system or independent file set quota.

Snapshots are read-only; changes can be made only to the normal, active files and directories, not to the snapshot.

Collecting snapshots of file systems and independent file sets is a useful feature. However, try not to create too many snapshots, high frequency snapshots, or extremely long retention values. With space-efficient snapshots, the snapshot is captured almost instantaneously and no data is used until the files in the file system or file set snapshot change or are deleted.

With long retention on snapshots and many frequent snapshots, the required storage can grow significantly in space with high delta rate changes. The scan engine response drops, capacity is used, and your cluster competes internally with your client needs. The impact of snapshots is almost impossible to guess if do not have change rate delta information that compares. Therefore, the client must monitor and manage the situation as it grows.

In some cases, resource expansion might be required to increase the success of snapshots and capacity of the cluster. Snapshot cleanup and level setting on expectation might be required. Monitor snapshots and capacity regularly. For more information, see Chapter 6, “Backup and recovery, availability, and resiliency functions” on page 367.

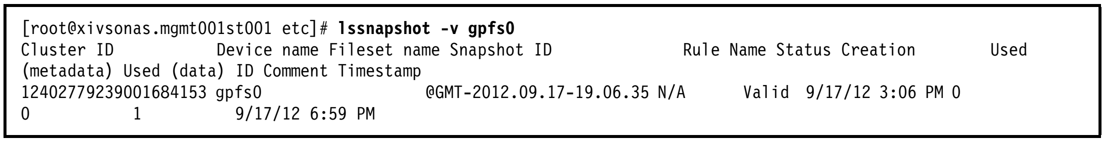
Example 8-3 shows an example **lssnapshot** command.

Example 8-3 Example lssnapshot command

```
lssnapshot gpfs0 -j independent_fset
```

Example 8-3 on page 648 lists snapshots of the “independent_fset” file set on file system “gpfs0”. The following column headers are printed: Device name, Fileset name, Snapshot ID, Rule name, Status, Creation, Used (metadata), Used (data), ID, and Timestamp. The space that is occupied by data and metadata is measured in KB. The Rule column contains the name of the rule that is responsible for creating the snapshot. If the Rule field contains “N/A”, the snapshot was created manually.

Figure 8-33 shows an example **lssnapshot** command output.



```
[root@xivsonas.mgmt001st001 etc]# lssnapshot -v gpfs0
Cluster ID           Device name Fileset name Snapshot ID              Rule Name Status Creation        Used
(metadata) Used (data) ID Comment Timestamp
12402779239001684153 gpfs0                  @GMT-2012.09.17-19.06.35 N/A      Valid  9/17/12 3:06 PM 0
0           1          9/17/12 6:59 PM
```

Figure 8-33 Example lssnapshot command output

The **lssnapops** command displays queued and running snapshot operations in chronological sequence, including invocations of the background process and creation and deletion of snapshot instances. The system administrator must manage the snapshot rules and their associations with file systems and independent file sets to ensure optimal performance. If operations for a rule are still in progress when the current instance is initiated, a warning is logged. The automated background process is queued if the previously initiated process has not completed, and it is serialized with other GPFS create and delete operations and with other instances of itself. You can set thresholds at the process level and at the individual rule level by running **setsnapnotify** to generate a warning if the number of operations threshold is exceeded.

8.5.6 Number of inodes available per file system and file set

The **numInodesToPreallocate** variable specifies the number of inodes that the system immediately preallocates.

For file systems that create multiple files in parallel, if the total number of free inodes is not greater than 5% of the total number of inodes, access to the file system might slow down.

If a file system or independent file set runs out of allocated inodes, new writes to the file system or file set stop.

Consider the inode allocation when you change your file system. If you want to create a smaller file system, use smaller values or you get a GPFS error message. You can specify the values in thousands (k) or millions (M), which stand for decimal values.

GPFS defines a minimum number of inodes, which might be greater than the maximum specified. This definition is done to allow maximum parallel usage of the inode file that is used by GPFS, and it involves calculations of the number of disks and nodes that can be mounted to the file system. It is not allowed to change the maximum number of inodes, if independent file sets are created in the file system. To change the inode limit for the *root file set* if there are independent file sets, run **chfset**.

Inode space

Inode space is inherited from the file system or independent file set to which the file set belongs.

Dependent file sets

For dependent file sets, inode usage is charged against the total number of inodes that are defined for the containing file system or containing independent file set.

Independent file sets

For independent file sets, inode space is separate from, and independent of, the file system that contains the file set. Inode usage is charged against the total number of inodes that are defined for the file set, and not against the containing file system.

Warnings, alerts, and hard limits

Inode consumption warnings, alerts, and hard limits are triggered by consumption thresholds. However, it is a good idea to track inode consumption and use the data as you grow file systems, implement data migrations, or initiate replication into new space allocations.

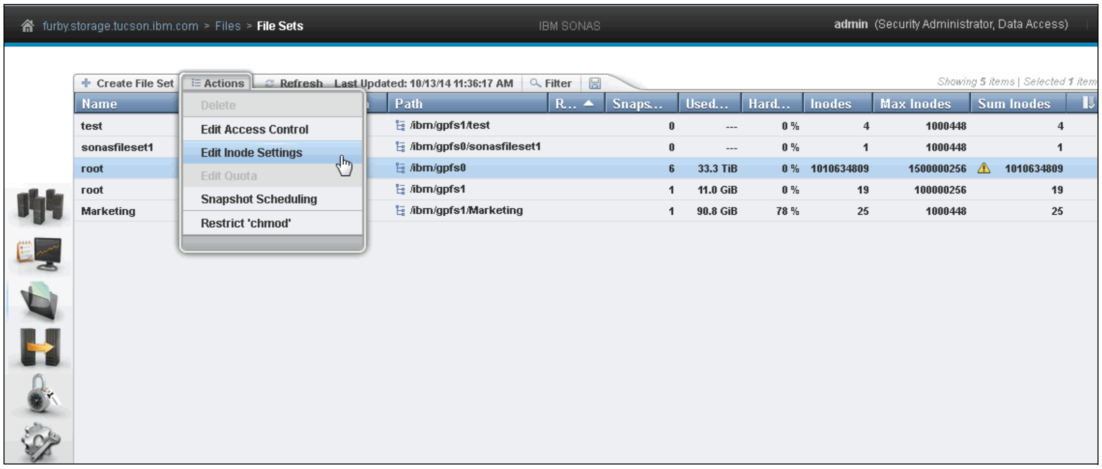
Current SONAS hard inode limit consumption alerts are set to 80% for a yellow warning alert and 90% for a red critical alert. When the inode limitation is exceeded, the GUI health state reflects the correct state of the value that was exceeded until either additional inodes are allocated or until the warning threshold is increased.

For example, a new file set that is created with 1 million max inodes generates a yellow warning alert when the file system uses 80% * 1,000,000, or 800,000 inodes. This alert notifies the SONAS administrator that they have 20% or 200,000 of their inode allocation left. When the file system reaches 900,000 inodes that are used, a red critical alert occurs to let the administrator know that there are only 100,000 inodes left and to formulate a plan for expansion and allocating more inodes by running `chfs`. (This example also applies for a file system.)

SONAS V1.5.1 customization features

With the release of SONAS V1.5.1, the SONAS administrator can customize the threshold settings of file sets, file systems, or both with the GUI or CLI. There are several use cases for adjusting inode utilization thresholds, but the primary reason is for very large file systems. If you consider the file set example in “Warnings, alerts, and hard limits”, 80% utilization of 1,000,000 inodes leaves the user with 200,000 inodes before all the inodes are used. Depending on the usage patterns, this process might take a short or long time. Now, take an entire file system that has grown with time to a maximum of 4,000,000,000 inodes. An 80% warning alert still leaves 800,000,000 inodes that, depending on usage patterns, might be months or years worth of capacity, and it might not be meaningful to have a yellow warning cluster health state, and it might not be valuable to add more inodes to remove the alert. In this case, the SONAS administrator can set the threshold to a more realistic value, which gives them the alert with a more appropriate buffer that works for their environment.

As shown in Figure 8-34 on page 651, to change the inode alert settings, click **Edit Inode Settings**.



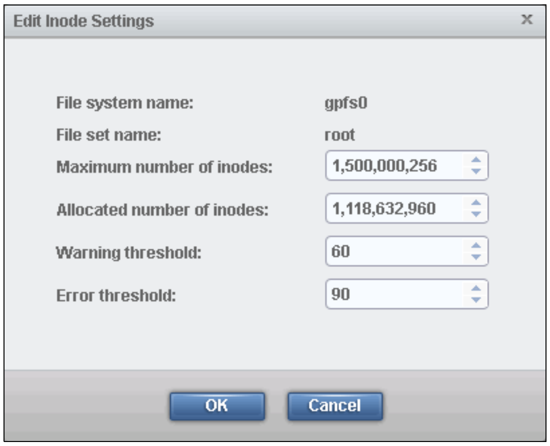
Last Updated: 10/13/14 11:36:17 AM

Showing 5 items | Selected 1 item

Name	Path	R...	Snaps...	Used...	Hard...	Inodes	Max Inodes	Sum Inodes
test	/ibm/gpfs1/test		0	---	0 %	4	1000448	4
sonasfileset1	/ibm/gpfs0/sonasfileset1		0	---	0 %	1	1000448	1
root	/ibm/gpfs0		6	33.3 TiB	0 %	1010634809	1500000256 ⚠	1010634809
root	/ibm/gpfs1		1	11.0 GiB	0 %	19	100000256	19
Marketing	/ibm/gpfs1/Marketing		1	90.8 GiB	78 %	25	1000448	25

Figure 8-34 Open the edit inode settings window

Enter the inode threshold alert settings, as shown in Figure 8-35.



File system name: gpfs0

File set name: root

Maximum number of inodes: 1,500,000,256

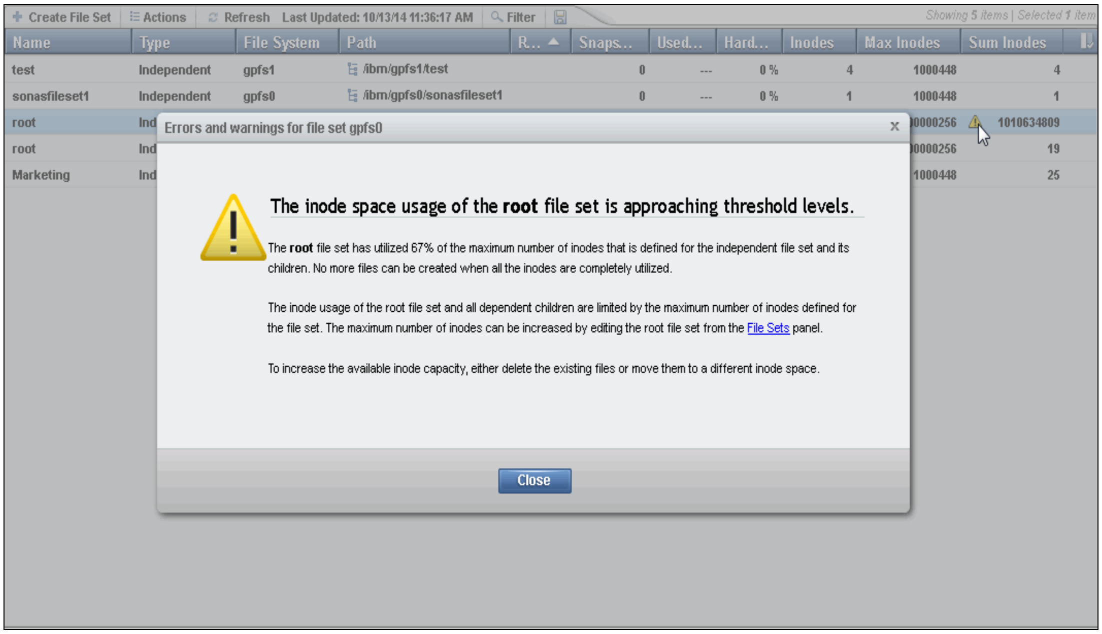
Allocated number of inodes: 1,118,632,960

Warning threshold: 60

Error threshold: 90

Figure 8-35 Edit inode threshold alert settings

You can see the inode threshold alerts and their details, as shown in Figure 8-36.



+ Create File Set
Actions
Refresh
Last Updated: 10/13/14 11:36:17 AM
Filter
Showing 5 items | Selected 1 item

Name	Type	File System	Path	R...	Snaps...	Used...	Hard...	Inodes	Max Inodes	Sum Inodes
test	Independent	gpfs1	/ibm/gpfs1/test		0	---	0 %	4	1000448	4
sonasfileset1	Independent	gpfs0	/ibm/gpfs0/sonasfileset1		0	---	0 %	1	1000448	1
root	Ind								0000256	⚠ 1010634809
root	Ind								0000256	19
Marketing	Ind								1000448	25

Errors and warnings for file set gpfs0

The inode space usage of the root file set is approaching threshold levels.

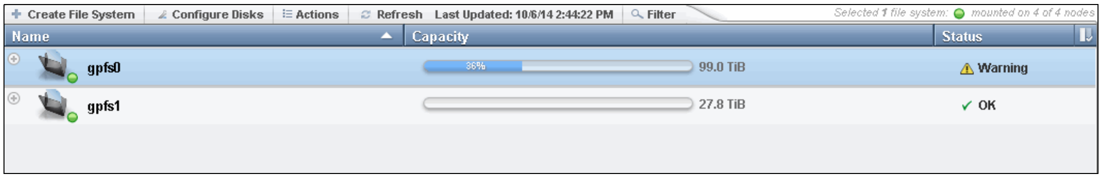
The **root** file set has utilized 67% of the maximum number of inodes that is defined for the independent file set and its children. No more files can be created when all the inodes are completely utilized.

The inode usage of the root file set and all dependent children are limited by the maximum number of inodes defined for the file set. The maximum number of inodes can be increased by editing the root file set from the File Sets panel.

To increase the available inode capacity, either delete the existing files or move them to a different inode space.

Figure 8-36 Details of the inode threshold warning

The inode threshold alert warning can also be seen in the file system window in the GUI, as shown in Figure 8-37.



+ Create File System
Configure Disks
Actions
Refresh
Last Updated: 10/6/14 2:44:22 PM
Filter
Selected 1 file system: mounted on 4 of 4 nodes

Name	Capacity	Status
gpfs0	36% 99.0 TiB	⚠ Warning
gpfs1	27.8 TiB	✓ OK

Figure 8-37 Inode limit alert from the file system window

A simple way to look at inode usage is by running the **lsfs** or **lsfset** commands.

Figure 8-38 lists all file sets for the file system device that is named gpfs0.

```
[root@totem.mgmt001st001 ~]# lsfset gpfs0 -v
ID Name Status Path Is independent CreationTime Comment Timestamp Root inode Parent id Inodes Data Inode space
owner
0 root Linked /ibm/gpfs0 yes 11/28/11 10:15 AM root0 fileset 12/15/11 12:36 PM 3 -- 0  0 kB 0 1.000G  1.000M
1 manuTest Linked /ibm/gpfs0/manuTest no 11/29/11 2:35 PM  12/15/11 12:36 PM 832000  0  0  0 kB 0
```

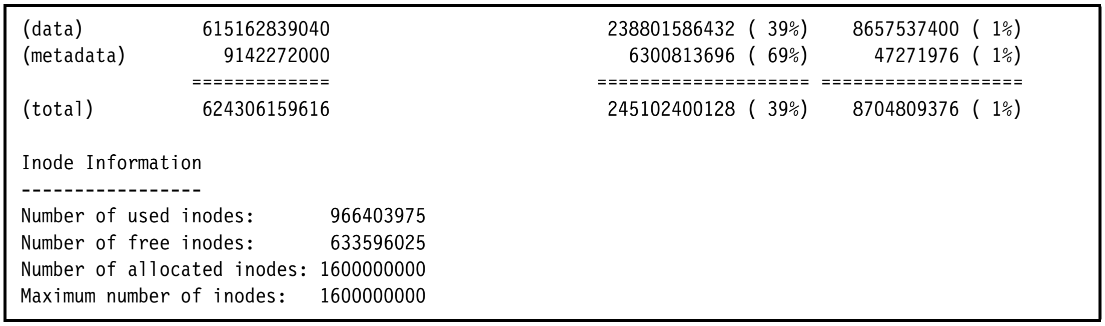
Figure 8-38 Example lsfset -v command output

Another way is to run the **df -i** command for the file system or file set mount point, as shown in Figure 8-39.

```
[root@xiv148 ~]# df -i
File system            Inodes   IUsed   IFree IUse% Mounted on
/dev/mapper/VolGroup00-LogVol00
                     52297728  162672 52135056    1% /
/dev/sda1               26104      34   26070    1% /boot
tmpfs                 4117489       1 4117488    1% /dev/shm
```

Figure 8-39 df -i output when run on a client with mounted file systems

For a deeper evaluation, you can run the **mmdf** command from the Management node as a privileged user. Figure 8-40 shows the output of the **mmdf gpfs0** command.



```
(data)          615162839040                            238801586432 ( 39%)    8657537400 ( 1%)
(metadata)        9142272000                              6300813696 ( 69%)      47271976 ( 1%)
               =============                           ==================== ===================
(total)         624306159616                            245102400128 ( 39%)    8704809376 ( 1%)

Inode Information
-----------------
Number of used inodes:       966403975
Number of free inodes:       633596025
Number of allocated inodes: 1600000000
Maximum number of inodes:   1600000000
```

Figure 8-40   Sample mmdf command output when run on the Management node by a privileged user

9

Troubleshooting, hints, and tips

This chapter contains information about how to troubleshoot your SONAS system. It includes ways to check your system health by using the GUI, and to collect logs by using the CLI.

This chapter describes the following topics:

- Monitoring SONAS system details:
  - Event logs
  - Audit logs
- Troubleshooting the SONAS system
- Assist On-site
- Call Home
- Collecting logs
- Uploading logs

9.1 Introduction to troubleshooting

You can use SONAS to monitor your system's health and do health checks. You can view the logs and collect them for further analysis and problem resolution. The SONAS GUI and CLI display the system logs, which consist of system alerts, warnings, and events. With SONAS V1.3 and later, the audit logs are also included.

Both the GUI and CLI can be used to do a health check. You can monitor the health of the nodes to see whether they have any warning or alert messages. You can also check the logs to debug for more information.

This chapter explains how to dump all the logs in to a single package, which can be uploaded and sent to IBM Support for analysis (see 9.8, "Uploading logs to IBM Support" on page 677). It also describes more troubleshooting methods, such as Assist On-site (AOS) and Call Home.

Before you commit to a diagnostic trail, it is *vital* that you can articulate what you do and do not know about any problem. Be prepared to clearly identify problem indicators, where and when problems happened, and how the problems might relate to the purpose or function of your solution.

Collect the who, what, when, where, how, and why information for each component or source of an issue, alert or warning.

Categorize the issue in one or more of the following ways:

- System or Data Access
- Data Integrity
- Data Replication
- Data Protection
- System Reliability
- System Stability
- System Alerting
- System Redundancy
- System Performance
- Protocol Management
- Feature Functionality
- Networking
- Authentication
- Storage

Early identification of components that are affected by the service issue will lead to a speedy resolution and root cause analysis.

9.2 A detailed description of monitoring SONAS systems

In the SONAS system, there are multiple ways to monitor the system. Most of the monitoring tools can be found through the GUI in the Monitoring window. This chapter describes some of the ways to monitor your SONAS solution.

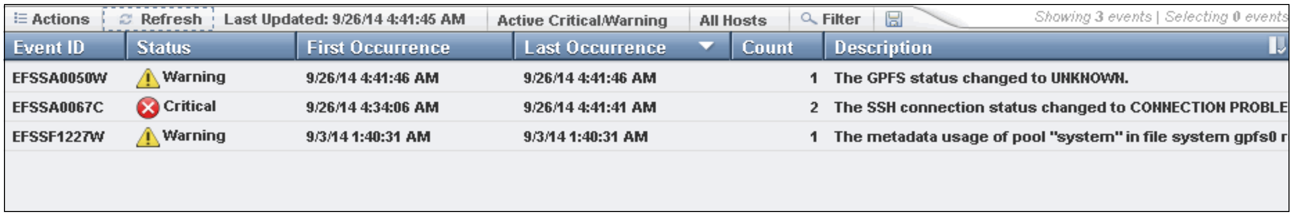
9.2.1 Event logs

The event log displays events that are collected with CIM agents or SNMP traps. The event log collects CIM and SNMP event data from all the components in the system, and stores them on the active Management node. It represents the history of all of the events that have occurred in the system. The log displays the events, which are managed in the health center. Events include successful and failed login and logout attempts by the GUI, external SSH, keyboard, and modem. A filter mask can be used to reduce the number of events that are displayed. The filter attributes can be used to filter by the following things:

- Severity
- Time period
- Source that is accessed by the GUI and CLI

GUI access

In the GUI, click **Monitoring** → **Events** in the upper right corner, where you can find filter options. You can filter by originating device. If you have issues with Storage node 1, you can filter the view to show events only from that node. This option might simplify searching through events. Figure 9-1 shows an event log that shows Current Critical/Warnings events from all nodes.



Actions
Refresh
Last Updated: 9/26/14 4:41:45 AM
Active Critical/Warning
All Hosts
Filter
Showing 3 events | Selecting 0 events

Event ID	Status	First Occurrence	Last Occurrence	Count	Description
EFSSA0050W	⚠ Warning	9/26/14 4:41:46 AM	9/26/14 4:41:46 AM	1	The GPFS status changed to UNKNOWN.
EFSSA0067C	✖ Critical	9/26/14 4:34:06 AM	9/26/14 4:41:41 AM	2	The SSH connection status changed to CONNECTION PROBLE
EFSSF1227W	⚠ Warning	9/3/14 1:40:31 AM	9/3/14 1:40:31 AM	1	The metadata usage of pool "system" in file system gpfs0 r

Figure 9-1 List current Critical/Warning events from all nodes

CLI access

Event logs can also be accessed through the CLI. After you log in to the active Management node, run **lslog**. Example 9-1 shows the help option for the **lslog** command.

Example 9-1 Help option for the lslog command

```
[st001.virtual.com]$ lslog --help
usage: lslog  [--count <n>] [-d <date> | --from <timestamp>] [-l <level>] [-n <node>]
[--reverse] [-v] [-Y] [-c < clusterID | clusterName >]
Lists the event log entries.

Parameter     Description
    --count   Lists the last n entries of the log database.
-d, --date    Lists only log entries for the specified date. Known formats are as follows:
mm/dd/yyyy, dd.mm.yyyy, yyyy-mm-dd, yyyymmdd. Years might be abbreviated for all formats that
contain delimiters. If a time is specified (as with --from) no error will be returned but the
results are undefined.
    --from    Lists only the log entries starting from the specified time stamp. Known formats
are as follows: mm/dd/yyyy hh:mm[:ss[.SSS]] AM|PM, dd.mm.yyyy HH:mm[:ss[.SSS]], yyyy-mm-dd
HH:mm[:ss[.SSS]], yyyymmdd HHmmssSSS. Years might be abbreviated for all formats that contain
delimiters. Seconds and milliseconds are optional for most formats. If no time is specified,
00:00:00.000 is chosen as default.
-l, --level   Lists only logs for the specified log level (FINEST, FINER, FINE, CONFIG, INFO,
SEVERE).
-n, --node    Lists only logs for the specified node.
    --reverse Lists the entries in reversed order (time is increasing).
```

```
-v, --verbose Shows additional columns.
-Y            Shows parseable output.
-c, --cluster The cluster scope for this command
[st001.virtual.com]$
```

The listing for the event log in the CLI can be hard to read. You can use parameters to prevent confusion. All of these parameters are listed in the help example. For example, if there is a problem with Storage node 1, you can look at the event log for that node. To view a specific node, use the **-n** parameter and the node name. Example 9-2 shows the usage of filtering logs to a specific node, in this case, Storage node 1.

Example 9-2 List event log entries for Storage node 1

```
[st001.virtual.com]$ lslog -n strg001st001
Host name       Severity Event id   Date received    Component Message
strg001st001    WARNING  SA0639W    10/27/11 7:34 PM system.ci The CIMOM for the host is not 
sending regular data.
strg001st001    SEVERE   SW0020C    10/27/11 7:30 PM system.lo The ntpd service stopped working.
strg001st001    WARNING  SA0639W    10/25/11 11:19 A system.ci The CIMOM for the host is not 
sending regular data.
strg001st001    SEVERE   SW0020C    10/25/11 11:14 A system.lo The ntpd service stopped working.
EFSSG1000I The command completed successfully.
[st001.virtual.com]$
```

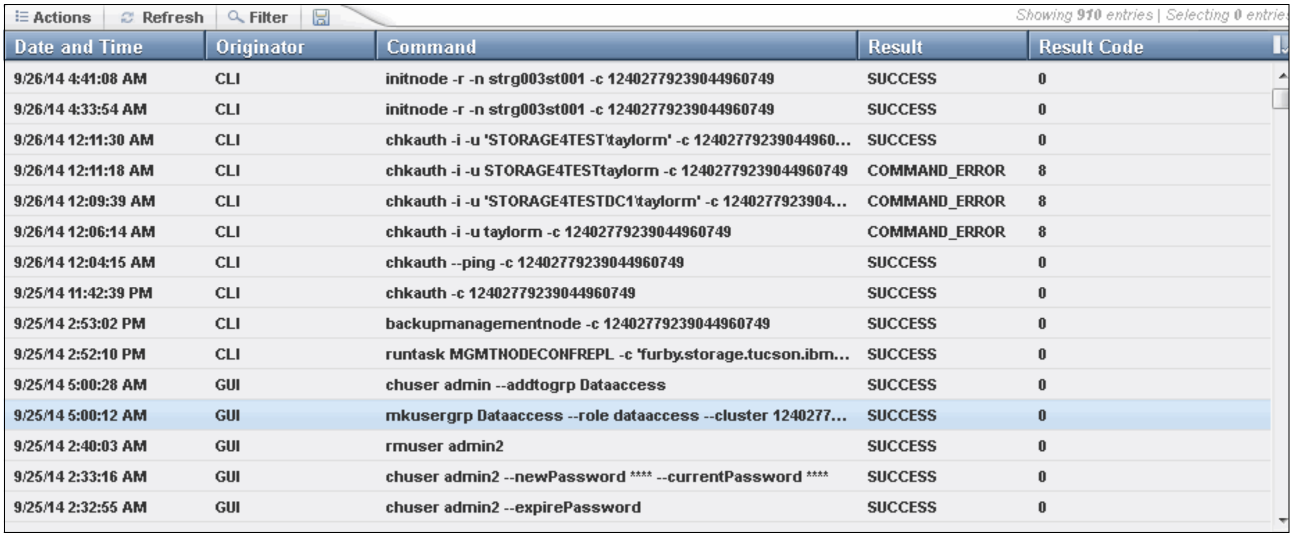
Four entries are displayed for Storage node 1. The event log is useful for directing IBM Support to the root of the problem.

9.2.2 Audit logs

Audit logs provide detailed information about what actions were done on the system and the user type that did them. The logs can be accessed by using the GUI or CLI.

GUI access

Click **Access** and then click **Audit Log**. Figure 9-2 on page 659 shows Audit log entries. Entries can be filtered by time: hour, minute, or days. In the Audit log, you can see commands that are run by users. If you are using the GUI, it shows the corresponding CLI command for actions that are made in the GUI. It also shows the result of the action that is done.



Date and Time	Originator	Command	Result	Result Code
9/26/14 4:41:08 AM	CLI	initnode -r -n strg003st001 -c 12402779239044960749	SUCCESS	0
9/26/14 4:33:54 AM	CLI	initnode -r -n strg003st001 -c 12402779239044960749	SUCCESS	0
9/26/14 12:11:30 AM	CLI	chkauth -i -u 'STORAGE4TEST\taylorm' -c 12402779239044960...	SUCCESS	0
9/26/14 12:11:18 AM	CLI	chkauth -i -u STORAGE4TESTtaylorm -c 12402779239044960749	COMMAND_ERROR	8
9/26/14 12:09:39 AM	CLI	chkauth -i -u 'STORAGE4TESTDC1\taylorm' -c 1240277923904...	COMMAND_ERROR	8
9/26/14 12:06:14 AM	CLI	chkauth -i -u taylorm -c 12402779239044960749	COMMAND_ERROR	8
9/26/14 12:04:15 AM	CLI	chkauth --ping -c 12402779239044960749	SUCCESS	0
9/25/14 11:42:39 PM	CLI	chkauth -c 12402779239044960749	SUCCESS	0
9/25/14 2:53:02 PM	CLI	backupmanagementnode -c 12402779239044960749	SUCCESS	0
9/25/14 2:52:10 PM	CLI	runtask MGMTNODECONFREPL -c 'furby.storage.tucson.ibm...	SUCCESS	0
9/25/14 5:00:28 AM	GUI	chuser admin --addtogrp Dataaccess	SUCCESS	0
9/25/14 5:00:12 AM	GUI	mkusergrp Dataaccess --role dataaccess --cluster 1240277...	SUCCESS	0
9/25/14 2:40:03 AM	GUI	rmuser admin2	SUCCESS	0
9/25/14 2:33:16 AM	GUI	chuser admin2 --newPassword **** --currentPassword ****	SUCCESS	0
9/25/14 2:32:55 AM	GUI	chuser admin2 --expirePassword	SUCCESS	0

Figure 9-2 Audit log entries that are shown in the GUI

CLI access

Log in to the active Management node and run **lsaudit**. In Example 9-3, you can see the help option for the **lsaudit** command.

Example 9-3 Help option for the lsaudit command

```
[st001.virtual.com]$ lsaudit --help
usage: lsaudit  [--count <count>] [-s <startDate>] [-e <endDate>] [--command <command>]
[--originator <originator>] [--filesystem <fileSystem>] [--fileset <fileset>] [--sharename
<shareName>] [--user <user>] [--reverse]
List the audit log entries.

Parameter        Description
    --count      Lists a maximum number of entries for the audit log file.
-s, --startdate  Lists only audit log entries from the specified start date.
-e, --enddate    Lists only the log entries starting until the specified end date.
    --command    Lists the audit log entries containing a specific command.
    --originator Lists the audit log entries containing a specific originator.
    --filesystem Lists the audit log entries containing a specific file system.
    --fileset    Lists the audit log entries containing a specific file set.
    --sharename  Lists the audit log entries containing a specific sharename.
    --user       Lists the audit log entries containing a specific user.
    --reverse    Lists the entries in reversed order (time is increasing).
[st001.virtual.com]$
```

Audit logs can also be deleted. However, it is highly inadvisable because logs can provide valuable clues to IBM Support personnel if there are any problems.

9.3 Troubleshooting: System details

This section describes checking the system health by using the GUI and CLI. You can check for node state details and important cluster parameters, such as CTDB and GPFS.

9.3.1 System details in the GUI

In the GUI, you can monitor the system state through different menus, as shown in Figure 9-3. To see information about the entire SONAS system, window, select the part of the SONAS system that you want to check. Consider the following categories for this view:

- Interface and Management nodes
- Storage nodes
- InfiniBand and Ethernet switches

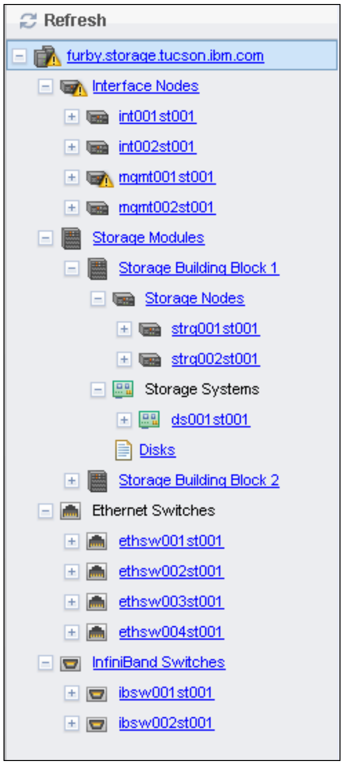
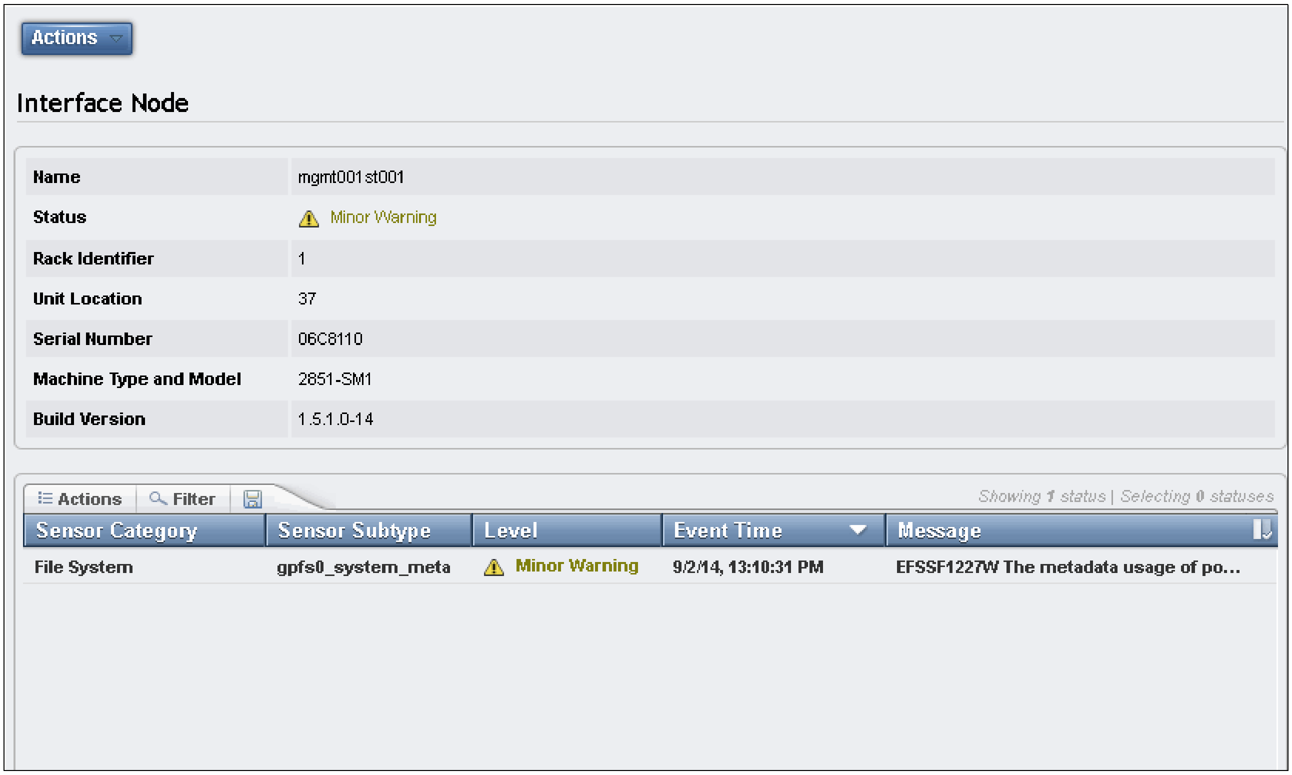


Figure 9-3 Navigational window

9.3.2 Details for Interface nodes and Management nodes

For these nodes, you can monitor their basic hardware state, operating system, and cluster connected services. When you click a node name, as shown in Figure 9-4 on page 661, you see an overview of that node, as shown for Management node 1. In this view, you can see an overview that includes the following information:

- Name: Node name
- Status: Latest status
- Rack identifier: Rack number where the node is
- Unit location: Rack position
- Serial number: Node serial number
- Build version: SONAS code level
- Event view: View of events that are associated with this node



Interface Node

Name	mgmt001st001
Status	⚠ Minor Warning
Rack Identifier	1
Unit Location	37
Serial Number	06C8110
Machine Type and Model	2851-SM1
Build Version	1.5.1.0-14

Sensor Category	Sensor Subtype	Level	Event Time	Message
File System	gpfs0_system_meta	⚠ Minor Warning	9/2/14, 13:10:31 PM	EFSSF1227W The metadata usage of po...

Figure 9-4 Basic view of the mgmt001st001 node view

For each Interface and Management node, you can see details for hardware, operating system, and SONAS-related services.

You can view the following information in the Hardware view:

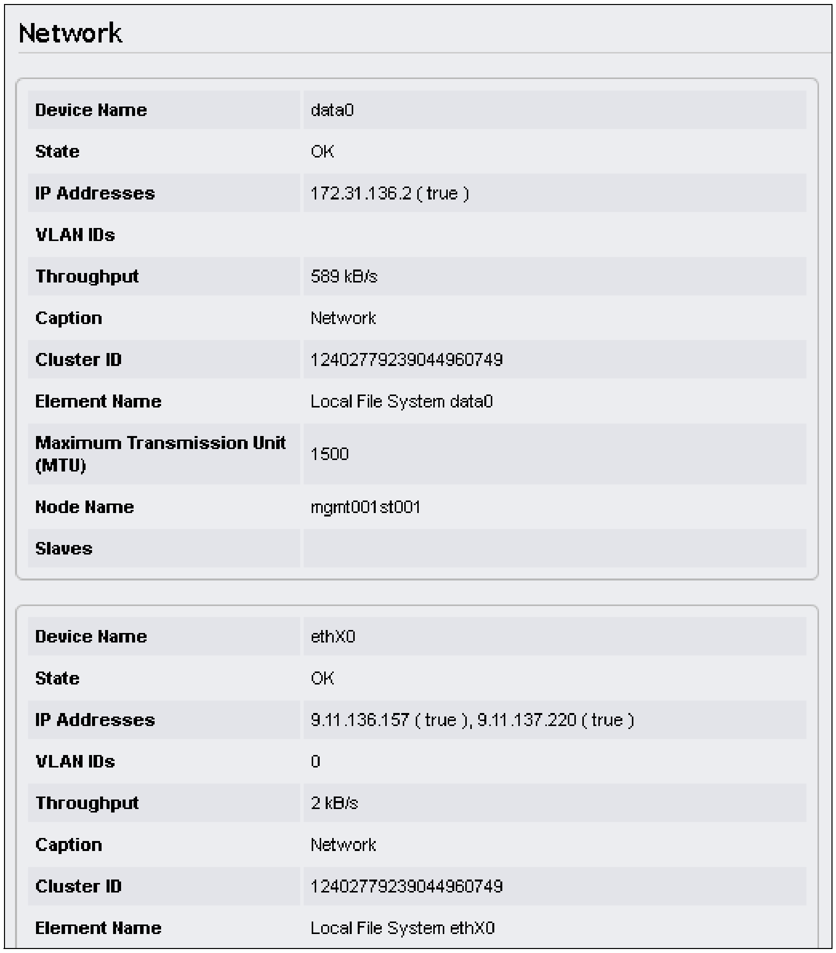
- Motherboard
- CPU
- Fan
- HDD
- Memory Modules
- Power
- Network Card

You can view the following information in the Operating System view:

- Computer System Details: Displays details for the computer system
- Operating System Details: Displays details for the operating system
- Local File System: Displays details for file systems that are local on this node

In SONAS-related services, the view is divided into the following sections:

- Network: Here all network connections are displayed, both internal and external. You can also check current throughput and more. For details, see Figure 9-5.



Network

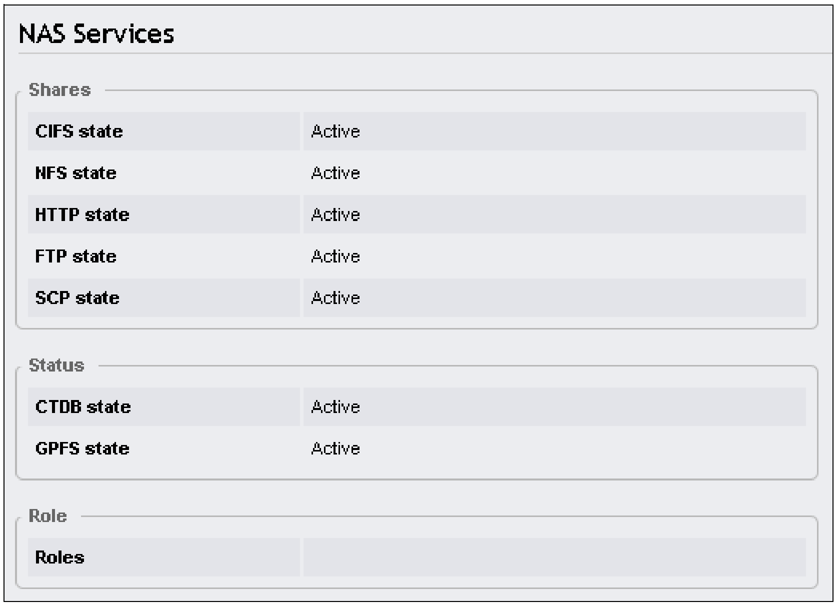
Device Name	data0
State	OK
IP Addresses	172.31.136.2 (true)
VLAN IDs	
Throughput	589 kB/s
Caption	Network
Cluster ID	12402779239044960749
Element Name	Local File System data0
Maximum Transmission Unit (MTU)	1500
Node Name	mgmt001st001
Slaves	

Device Name	ethX0
State	OK
IP Addresses	9.11.136.157 (true), 9.11.137.220 (true)
VLAN IDs	0
Throughput	2 kB/s
Caption	Network
Cluster ID	12402779239044960749
Element Name	Local File System ethX0

Figure 9-5 Network details for internal and external connections

– NAS services:

Figure 9-6 on page 663 shows the details of the listed services.



NAS Services

Shares

CIFS state	Active
NFS state	Active
HTTP state	Active
FTP state	Active
SCP state	Active

Status

CTDB state	Active
GPFS state	Active

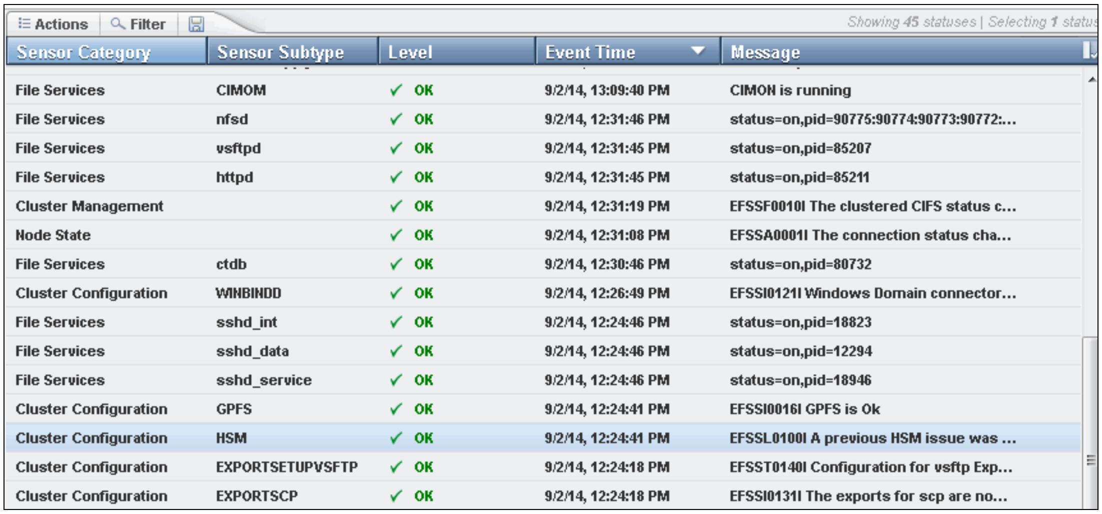
Role

Roles	

Figure 9-6 NAS Services details listed

– Status:

In Figure 9-7, you see the listed services that are needed for normal SONAS operations from the Interface and Management nodes.

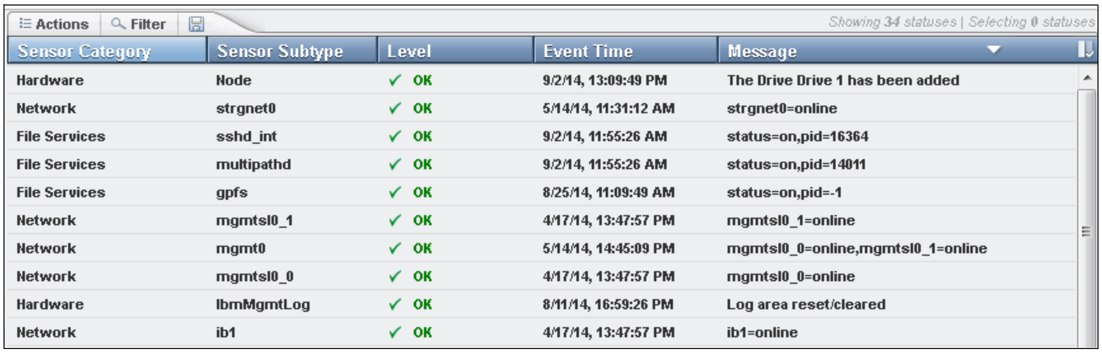


Sensor Category	Sensor Subtype	Level	Event Time	Message
File Services	CIMOM	✓ OK	9/2/14, 13:09:40 PM	CIMON is running
File Services	nfsd	✓ OK	9/2/14, 12:31:46 PM	status=on,pid=90775:90774:90773:90772:...
File Services	vsftpd	✓ OK	9/2/14, 12:31:45 PM	status=on,pid=85207
File Services	httpd	✓ OK	9/2/14, 12:31:45 PM	status=on,pid=85211
Cluster Management		✓ OK	9/2/14, 12:31:19 PM	EFSSF0010I The clustered CIFS status c...
Node State		✓ OK	9/2/14, 12:31:08 PM	EFSSA0001I The connection status cha...
File Services	ctdb	✓ OK	9/2/14, 12:30:46 PM	status=on,pid=80732
Cluster Configuration	WINBINDD	✓ OK	9/2/14, 12:26:49 PM	EFSSI0121I Windows Domain connector...
File Services	sshd_int	✓ OK	9/2/14, 12:24:46 PM	status=on,pid=18823
File Services	sshd_data	✓ OK	9/2/14, 12:24:46 PM	status=on,pid=12294
File Services	sshd_service	✓ OK	9/2/14, 12:24:46 PM	status=on,pid=18946
Cluster Configuration	GPFS	✓ OK	9/2/14, 12:24:41 PM	EFSSI0016I GPFS is Ok
Cluster Configuration	HSM	✓ OK	9/2/14, 12:24:41 PM	EFSSL0100I A previous HSM issue was ...
Cluster Configuration	EXPORTSETUPVSFTP	✓ OK	9/2/14, 12:24:18 PM	EFSST0140I Configuration for vsftp Exp...
Cluster Configuration	EXPORTSCP	✓ OK	9/2/14, 12:24:18 PM	EFSSI0131I The exports for scp are no...

Figure 9-7 Status view for Interface and Management nodes

– Details for Storage nodes:

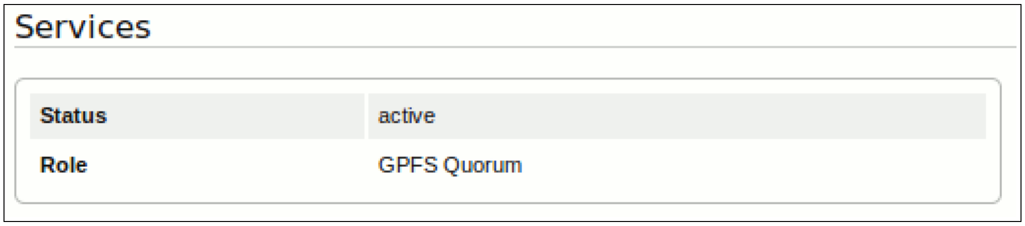
The detailed information for Storage nodes is similar to the information that is presented for the Interface and Management nodes (that is, to the hardware and operating system information). The main difference is in the Status view, as shown in Figure 9-8. No CTDB or other services that are related to shares that are running on the Storage node are needed on Interface and Management nodes.



Sensor Category	Sensor Subtype	Level	Event Time	Message
Hardware	Node	✓ OK	9/2/14, 13:09:49 PM	The Drive Drive 1 has been added
Network	strgnet0	✓ OK	5/14/14, 11:31:12 AM	strgnet0=online
File Services	sshd_int	✓ OK	9/2/14, 11:55:26 AM	status=on,pid=16364
File Services	multipathd	✓ OK	9/2/14, 11:55:26 AM	status=on,pid=14011
File Services	gpfs	✓ OK	8/25/14, 11:09:49 AM	status=on,pid=-1
Network	mgmtsl0_1	✓ OK	4/17/14, 13:47:57 PM	mgmtsl0_1=online
Network	mgmt0	✓ OK	5/14/14, 14:45:09 PM	mgmtsl0_0=online,mgmtsl0_1=online
Network	mgmtsl0_0	✓ OK	4/17/14, 13:47:57 PM	mgmtsl0_0=online
Hardware	IbmMgmtLog	✓ OK	8/11/14, 16:59:26 PM	Log area reset/cleared
Network	ib1	✓ OK	4/17/14, 13:47:57 PM	ib1=online

Figure 9-8 Storage node status view

Under Services, you see only the GPFS status, as shown in Figure 9-9.

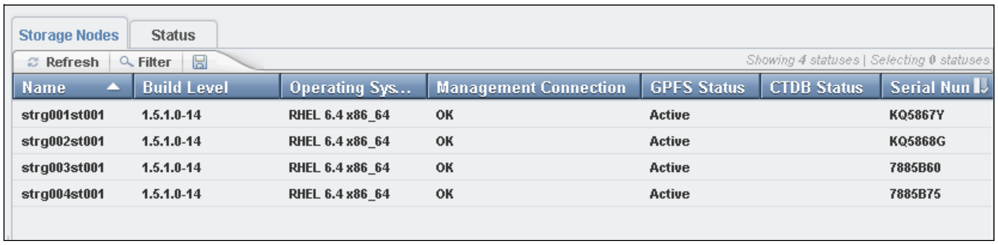


Status	active
Role	GPFS Quorum

Figure 9-9 Storage nodes Services view

Another main difference is that Storage nodes are grouped in Storage Building Blocks. Each Storage Building Block is made from two Storage nodes and storage disk systems that are connected to nodes. Here, you can also check the health of connected storage systems and disk health.

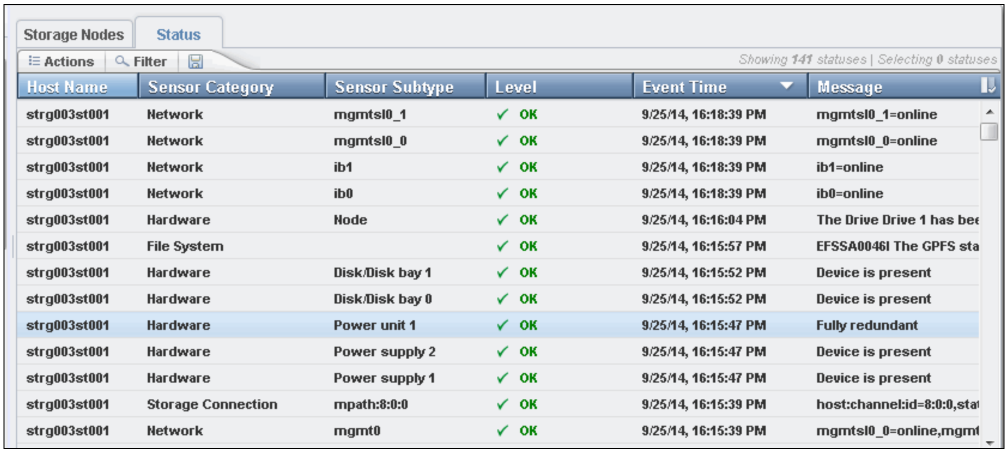
If you select the **Storage nodes** option in the System Details menu, you see the output that is shown in Figure 9-10.



Name	Build Level	Operating Sys...	Management Connection	GPFS Status	CTDB Status	Serial Nun
strg001st001	1.5.1.0-14	RHEL 6.4 x86_64	OK	Active		KQ5867Y
strg002st001	1.5.1.0-14	RHEL 6.4 x86_64	OK	Active		KQ5868G
strg003st001	1.5.1.0-14	RHEL 6.4 x86_64	OK	Active		7885B60
strg004st001	1.5.1.0-14	RHEL 6.4 x86_64	OK	Active		7885B75

Figure 9-10 Storage nodes basic view

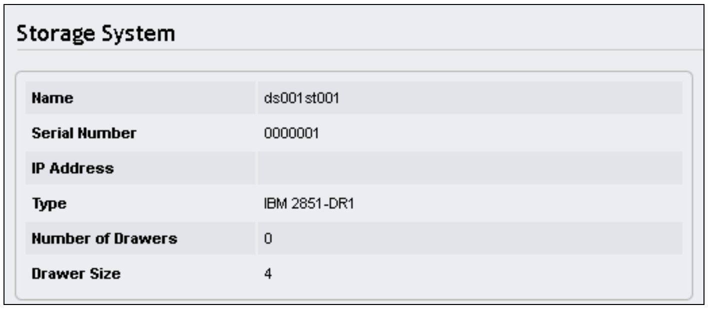
On the Status tab, you see the combined status from the Storage node pair. In this case, it is Storage node 1 and 2, as shown in Figure 9-11 on page 665.



Host Name	Sensor Category	Sensor Subtype	Level	Event Time	Message
strg003st001	Network	mgmtsl0_1	✓ OK	9/25/14, 16:18:39 PM	mgmtsl0_1=online
strg003st001	Network	mgmtsl0_0	✓ OK	9/25/14, 16:18:39 PM	mgmtsl0_0=online
strg003st001	Network	ib1	✓ OK	9/25/14, 16:18:39 PM	ib1=online
strg003st001	Network	ib0	✓ OK	9/25/14, 16:18:39 PM	ib0=online
strg003st001	Hardware	Node	✓ OK	9/25/14, 16:16:04 PM	The Drive Drive 1 has bee
strg003st001	File System		✓ OK	9/25/14, 16:15:57 PM	EFSSA0046I The GPFS sta
strg003st001	Hardware	Disk/Disk bay 1	✓ OK	9/25/14, 16:15:52 PM	Device is present
strg003st001	Hardware	Disk/Disk bay 0	✓ OK	9/25/14, 16:15:52 PM	Device is present
strg003st001	Hardware	Power unit 1	✓ OK	9/25/14, 16:15:47 PM	Fully redundant
strg003st001	Hardware	Power supply 2	✓ OK	9/25/14, 16:15:47 PM	Device is present
strg003st001	Hardware	Power supply 1	✓ OK	9/25/14, 16:15:47 PM	Device is present
strg003st001	Storage Connection	mpath:8:0:0	✓ OK	9/25/14, 16:15:39 PM	host:channel:id=8:0:0,stat
strg003st001	Network	mgmt0	✓ OK	9/25/14, 16:15:39 PM	mgmtsl0_0=online,mgmt

Figure 9-11 Status view for both Storage nodes

In addition to Storage node details, you can also check for details of the connected storage controller and disks that are served. Storage System basic information is shown in Figure 9-12.

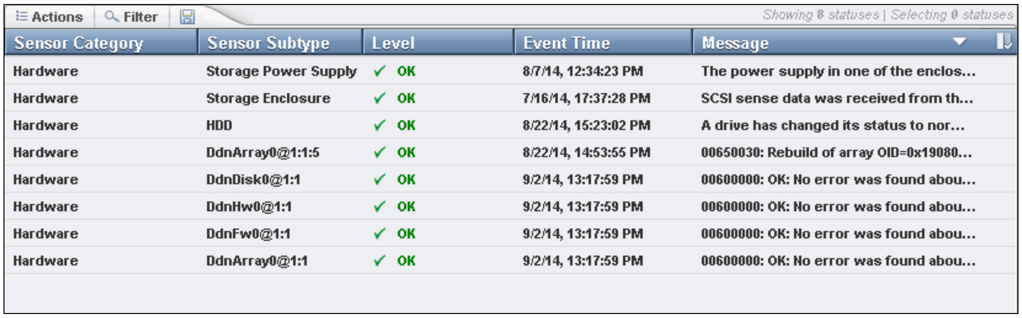


Storage System

Name	ds001st001
Serial Number	0000001
IP Address	
Type	IBM 2851-DR1
Number of Drawers	0
Drawer Size	4

Figure 9-12 Storage System basic information view

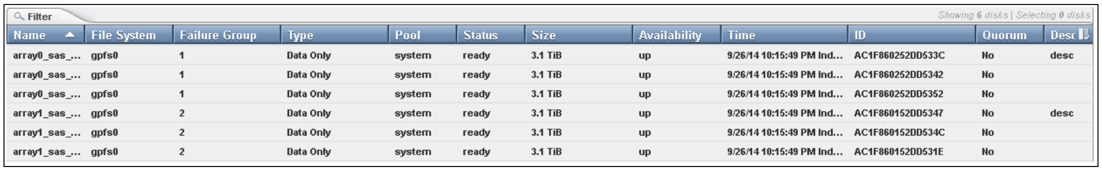
Select **Status** under **Storage System** to show the status of the Storage System, as shown in Figure 9-13.



Sensor Category	Sensor Subtype	Level	Event Time	Message
Hardware	Storage Power Supply	✓ OK	8/7/14, 12:34:23 PM	The power supply in one of the enclos...
Hardware	Storage Enclosure	✓ OK	7/16/14, 17:37:28 PM	SCSI sense data was received from th...
Hardware	HDD	✓ OK	8/22/14, 15:23:02 PM	A drive has changed its status to nor...
Hardware	DdnArray0@1:1:5	✓ OK	8/22/14, 14:53:55 PM	00650030: Rebuild of array OID=0x19080...
Hardware	DdnDisk0@1:1	✓ OK	9/2/14, 13:17:59 PM	00600000: OK: No error was found abou...
Hardware	DdnHw0@1:1	✓ OK	9/2/14, 13:17:59 PM	00600000: OK: No error was found abou...
Hardware	DdnFw0@1:1	✓ OK	9/2/14, 13:17:59 PM	00600000: OK: No error was found abou...
Hardware	DdnArray0@1:1	✓ OK	9/2/14, 13:17:59 PM	00600000: OK: No error was found abou...

Figure 9-13 Status view for Storage System

For a disk view, select **Disks** from the navigational pane in **Monitoring** under **System Details**. For details, see Figure 9-14. This figure is a small portion of the large amount of data that is available in the GUI.

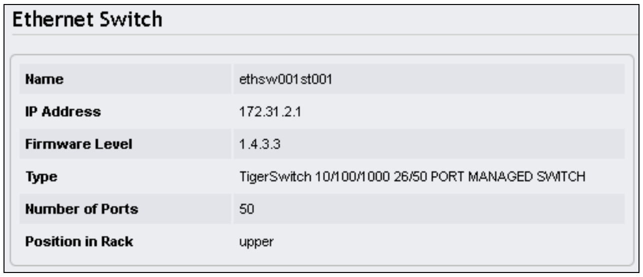


Name	File System	Failure Group	Type	Pool	Status	Size	Availability	Time	ID	Quorum	Desc
array0_sas_...	gpfs0	1	Data Only	system	ready	3.1 TiB	up	9/26/14 10:15:49 PM Ind...	AC1F860252DD533C	No	desc
array0_sas_...	gpfs0	1	Data Only	system	ready	3.1 TiB	up	9/26/14 10:15:49 PM Ind...	AC1F860252DD5342	No	
array0_sas_...	gpfs0	1	Data Only	system	ready	3.1 TiB	up	9/26/14 10:15:49 PM Ind...	AC1F860252DD5352	No	
array1_sas_...	gpfs0	2	Data Only	system	ready	3.1 TiB	up	9/26/14 10:15:49 PM Ind...	AC1F860152DD5347	No	desc
array1_sas_...	gpfs0	2	Data Only	system	ready	3.1 TiB	up	9/26/14 10:15:49 PM Ind...	AC1F860152DD534C	No	
array1_sas_...	gpfs0	2	Data Only	system	ready	3.1 TiB	up	9/26/14 10:15:49 PM Ind...	AC1F860152DD531E	No	

Figure 9-14 Disk view

- Details for InfiniBand and Ethernet switches:

  Here you can check the status for InfiniBand and Ethernet switches. The view is divided into a basic component information part and a detailed view for a specific component. Figure 9-15 shows Ethernet switch information.

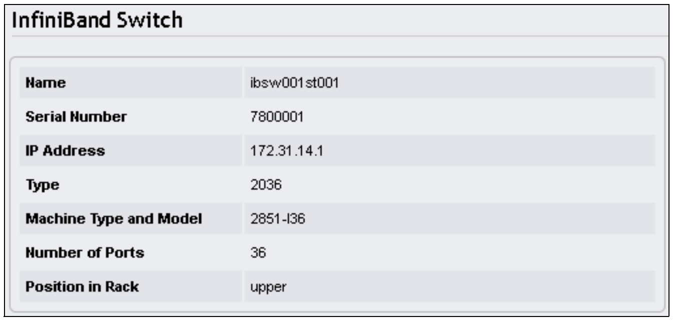


Ethernet Switch

Name	ethsw001st001
IP Address	172.31.2.1
Firmware Level	1.4.3.3
Type	TigerSwitch 10/100/1000 26/50 PORT MANAGED SWITCH
Number of Ports	50
Position in Rack	upper

Figure 9-15 Ethernet switch basic information

Figure 9-16 shows InfiniBand switch basic information.

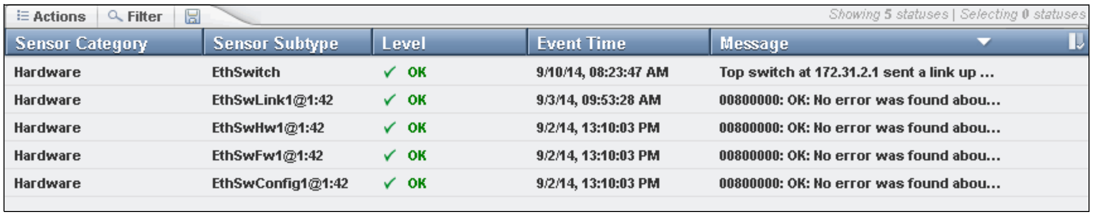


InfiniBand Switch

Name	ibsw001st001
Serial Number	7800001
IP Address	172.31.14.1
Type	2036
Machine Type and Model	2851-I36
Number of Ports	36
Position in Rack	upper

Figure 9-16 InfiniBand switch basic information

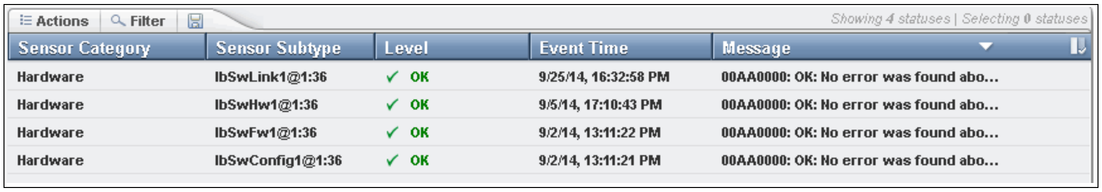
In addition, you can select the **Status** option below the switch to see a detailed view for every switch. Figure 9-17 shows details for the Ethernet switch.



Sensor Category	Sensor Subtype	Level	Event Time	Message
Hardware	EthSwitch	✓ OK	9/10/14, 08:23:47 AM	Top switch at 172.31.2.1 sent a link up ...
Hardware	EthSwLink1@1:42	✓ OK	9/3/14, 09:53:28 AM	00800000: OK: No error was found abou...
Hardware	EthSwHw1@1:42	✓ OK	9/2/14, 13:10:03 PM	00800000: OK: No error was found abou...
Hardware	EthSwFw1@1:42	✓ OK	9/2/14, 13:10:03 PM	00800000: OK: No error was found abou...
Hardware	EthSwConfig1@1:42	✓ OK	9/2/14, 13:10:03 PM	00800000: OK: No error was found abou...

Figure 9-17 Detailed view for the Ethernet switch

Figure 9-18 on page 667 shows details for the InfiniBand switch.

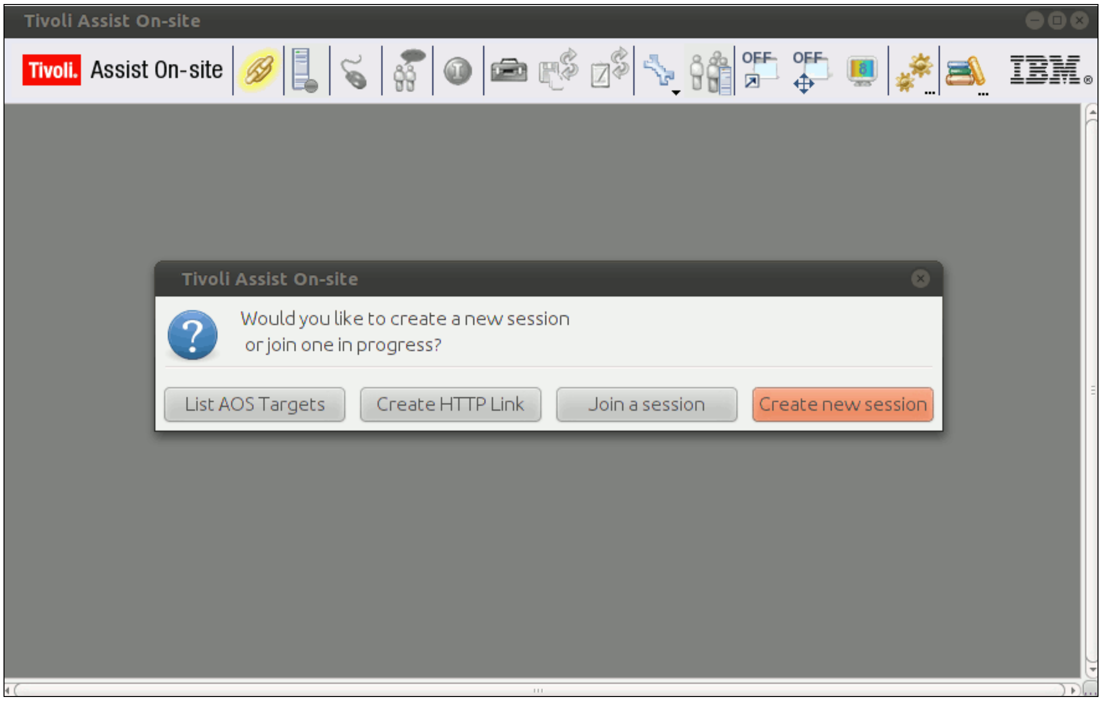


Sensor Category	Sensor Subtype	Level	Event Time	Message
Hardware	IbSwLink1@1:36	✓ OK	9/25/14, 16:32:58 PM	00AA0000: OK: No error was found abo...
Hardware	IbSwHw1@1:36	✓ OK	9/5/14, 17:10:43 PM	00AA0000: OK: No error was found abo...
Hardware	IbSwFw1@1:36	✓ OK	9/2/14, 13:11:22 PM	00AA0000: OK: No error was found abo...
Hardware	IbSwConfig1@1:36	✓ OK	9/2/14, 13:11:21 PM	00AA0000: OK: No error was found abo...

Figure 9-18 Detailed view for the InfiniBand switch

9.4 Assist On-site

Assist On-site (AOS) is a lightweight remote support program that is primarily used by help desks and support engineers to diagnose and fix problems without requiring external dependencies. Assist On-site is based on the IBM Tivoli Remote Control technology. Figure 9-19 shows the main window for AOS. The **Create new session** option is selected.



Would you like to create a new session
or join one in progress?

Figure 9-19 AOS window to create a session

9.4.1 Creating a session

You have multiple options for creating a session:

- List AOS Targets: This option can be used if the system triggers a Call Home. However, it must be configured.
- Create HTTP Link: This option creates an HTTP link and a passcode. Then, this information is provided to person onsite. It connects to the Relay Server.
- Join a session: This option lets you join already running sessions.
- Create new session: This option generates only a passcode. This passcode is provided to client to enter at the AOS Support website.

9.4.2 Assist On-site session modes

AOS can establish remote connections for support sessions in different modes. You choose the session mode after joining the support session or the support engineer can request that you change the mode during the support session. The type of session or the permissions that are associated with the support engineer also determine the session modes that are available during sessions.

With AOS, administrators create a team or a user and they can select the default permissions for that team or user, including the set of session modes that are available. For example, a team might have default permissions to run sessions in View Only and Chat Only session modes. Customers can further restrict the session mode when they consent to sessions:

- Chat Only mode: This session mode allows the support engineer to chat with the customer in the Chat window, but does not allow the support engineer to view the target system or have any control of the target mouse or keyboard.

  The Chat window allows the support engineer to chat with you within another session mode and provides an additional form of contact. The IBM Support engineer can also request, or you can change to, the Chat Only mode during a support session.

- View Only mode: This session mode allows the support engineer to view the target system, but it does not allow the support engineer to have any control over the target mouse or keyboard. In the View Only mode, the support engineer can select and mark areas of the target desktop by using the Remote Support Console tools. The support engineer can also request, or you can change to, the View Only mode during a support session.
- Guidance mode: This session mode allows the support engineer to view the target system and direct the client to perform tasks on the target system, but does not allow the support engineer to have any control of the target mouse or keyboard. The support engineer can use the Guidance mode symbols, Remote Support Console tools, and the chat function to direct you through any task to perform on the target. The Guidance mode is often used in training situations and in workplaces of high sensitivity.
- Shared Control mode: This session mode allows the support engineer to view the target system and to have input control of the target mouse and keyboard. During a support session, the support engineer can turn on local input control to perform actions on the support engineer's machine instead of the target machine. The actions of the customer take precedence over the actions that are performed through the Remote Support Console. When you use the mouse or the keyboard, the input control icon changes to indicate that input control in the Remote Support Console is temporarily blocked until you stop using the mouse or the keyboard. The support engineer can use the Remote Support Console tools, such as the drawing tools, to select and mark areas of the target desktop. The support engineer can request, or you can change to, the Shared Control mode during a support session.

9.5 Call Home

The IBM cluster currently supports automatic, electronic Call Home messaging by using a configured data path and the Management node.

The SONAS cluster initiates Call Home messages against the machine type and model and serial number of the hardware component that triggered the error. The Call Home messages contain error codes that provide specifics about the problem. The following list shows the valid machine types and models for Call Home messaging:

- 2851-SIx - Interface nodes
- 2851-SM1 - Management nodes
- 2851-SSx - Storage node
- 2851-DR1 - Storage controller
- 2851-I36 - 36-port InfiniBand switch
- 2851-I96 - 96-port InfiniBand switch

9.5.1 Call Home caveats

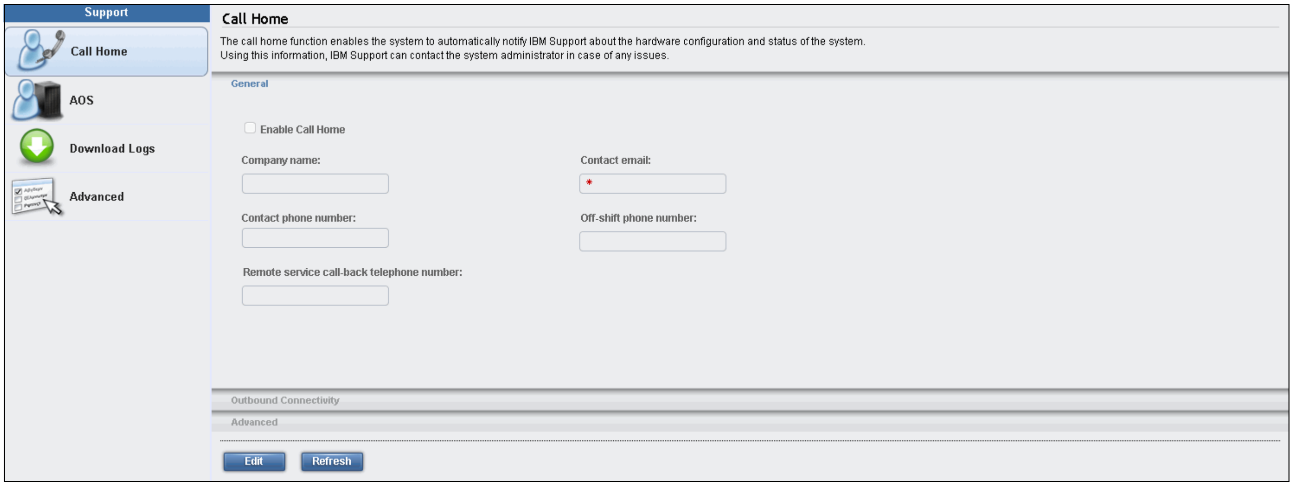
The SONAS Call Home messaging process has the following caveats:

- 2851-DE1 storage expansion unit enclosure errors issue a Call Home message against the parent 2851-DR1 storage controller enclosure. Ethernet switch errors issue a Call Home message against the active Management node's machine type, model, and serial number.
- Frame assembly components lack an interface that supports Call Home messaging. Other than the Ethernet switches, the only frame hardware that can be tied to system issues are power distribution units (PDUs) that are tied to power failures. If there is a PDU failure, a Call Home message is generated against a component that is plugged in to a failing PDU.

9.5.2 Enabling and disabling Call Home

SONAS Call Home messages include an 8-character error code that indicates the problem that occurred. The Call Home feature is `not` mandatory. It can be turned off or on.

Turning on or off Call Home with the GUI: To check and, if needed, change settings for Call Home by using the GUI, go to **Settings** and select **Support**. In the window that is shown in Figure 9-20, you can enable or disable the Call Home feature and enter the necessary parameters that are needed for Call Home.



The call home function enables the system to automatically notify IBM Support about the hardware configuration and status of the system.
Using this information, IBM Support can contact the system administrator in case of any issues.

Figure 9-20 Call Home options in SONAS GUI

Turning on and off Call Home with the CLI: To change the configuration for the Call Home feature, run **`cfgcallhome`**. In Example 9-4, you can see the **`help`** option for this command.

Example 9-4 Help option for cfgcallhome command in CLI

```
[st001.virtual.com]$ cfgcallhome --help
usage: cfgcallhome  [--customer-name <name>] [--voice-phone <number>] [--offshift-phone
<number>] [--modem-phone <number>] [--modem-prefix <number>] [--ign1phone <number>] [--ign2phone
<number>] [--customer-location <location>] [--email <mail>] [--special-instruction
<instruction>] [--enablecallhome | --disablecallhome] [--enablesoftwarecallhome |
--disablesoftwarecallhome] [--callhome-method <method>] [--heart-beat-interval <interval>]
[--remote-service-callback-number <number>] [--enableproxy | --disableproxy] [--proxy-location
<location>] [--proxy-password <password>] [--enablemulticluster | --disablemulticluster]
[--proxy-user <name>] [--enableauthentication | --disableauthentication] [--proxy-port <port>]
[--reset <port>] [-c < clusterID | clusterName >]
Configures or sets the Call Home options on the current system.

Parameter                           Description
    --customer-name                 Specifies the business or company name.
    --voice-phone                   Specifies the customer voice phone number.
    --offshift-phone                Specifies the customer offshift voice phone number.
    --modem-phone                   Specifies the modem phone number.
    --modem-prefix                  Specifies the modem number to access external line.
    --ign1phone                     Specifies the first outbound Call Home phone number.
    --ign2phone                     Specifies the second outbound Call Home phone number.
    --customer-location             Specifies the location of the machine.
    --email                         Specifies the email to be used to contact the customer.
    --special-instruction           Specifies special instructions to IBM Support.
    --enablecallhome                Specifies if Call Home is enabled.
    --disablecallhome               Specifies if Call Home is disabled.
    --enablesoftwarecallhome        Specifies if software Call Home is enabled.
    --disablesoftwarecallhome       Specifies if software Call Home is disabled.
    --callhome-method               Specifies the Call Home method.
    --heart-beat-interval            Specifies the heartbeat interval.
    --remote-service-callback-number Specifies the remote service callback number.
    --enableproxy                    Specifies if proxy server is enabled.
    --disableproxy                   Specifies if a proxy server is disabled.
    --proxy-location                 Specifies the proxy address by IP address or host name.
    --proxy-password                 Specifies the proxy password.
    --enablemulticluster             Specifies if multicluster is enabled.
    --disablemulticluster            Specifies if multicluster is disabled.
    --proxy-user                     Specifies the proxy user name.
    --enableauthentication           Specifies if proxy authentication is enabled.
    --disableauthentication          Specifies if proxy authentication is disabled.
    --proxy-port                     Specifies the proxy port.
    --reset                          Resets the configuration. Type 'reset' to restore
configuration values.
-c, --cluster                        The cluster scope for this command
[st001.virtual.com]$
```

9.6 Collecting logs

You can use SONAS to collect the dump of the logs from all the nodes in the cluster into one single package. This package can be sent to IBM Support for further analysis and debugging.

9.6.1 Collecting logs with the CLI

The **`srvdump`** command is used to manage dump files, including generation, listing, deletion, and sending files to Call Home or media devices. Dump files are used to assist IBM Support with problem determination and resolution.

The help for the **`srvdump`** command is shown in Example 9-5.

Example 9-5 CLI command srvdump to collect dump of logs

```
# srvdump --help
usage: srvdump  {-g [<providers>] | -d <dumpIdentifier> | -l [<dumpIdentifier>[:<node>]] | --get
<dumpIdentifier>:<node>:<sections>[:<dvd|usb>] | -s <dumpIdentifier>:<dvd|usb|callhome>} [-X]
[-c < clusterID | clusterName >]
Manage dump files.

Parameter      Description
-g, --generate Generates dump files.
-d, --delete   Deletes dump files.
-l, --list     Lists dump files.
    --get      Extracts specific parts from a dump file.
-s, --send     Sends dump files to the specified destination.
-X, --Xtended  Collects extended dump files. Use this option in conjunction with the generate
option.
-c, --cluster  The cluster scope for this command
```

The final package, which is a `.tar` file, is created, compressed, and stored in the `/ftdc` directory.

9.6.2 Collecting logs by using the GUI

To collect logs with the GUI, click **Settings** → **Support**, as shown in Figure 9-21.

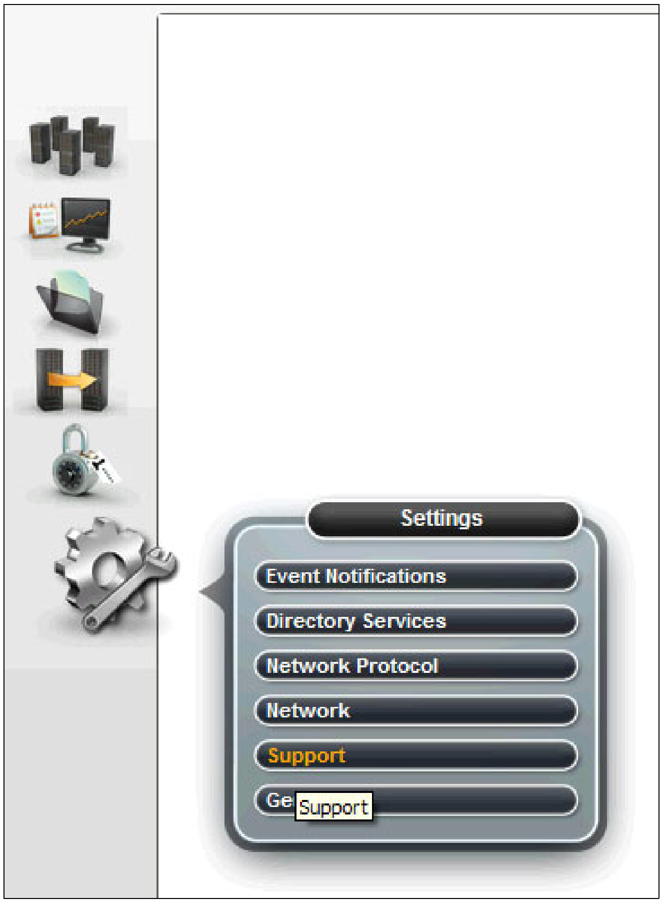
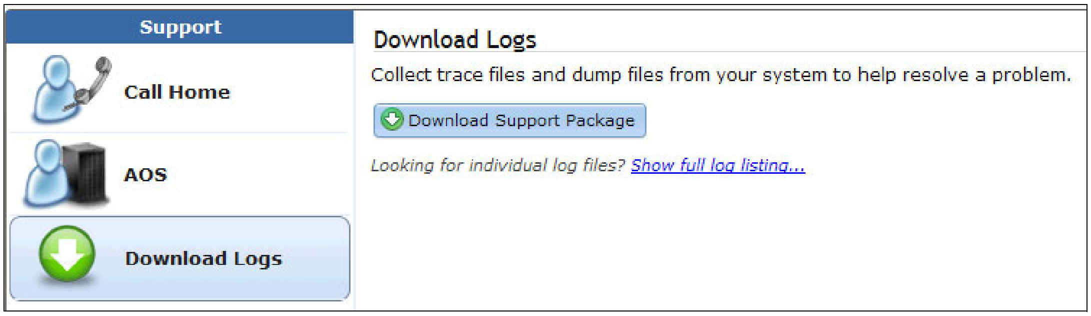


Figure 9-21 Support menu option in the GUI

When you click the link, you see three different options, as shown in Figure 9-22. To collect logs, click **Download Logs**, which is shown in the left pane. Other options in that pane are explained in the previous sections. When you click the link to download logs, a new window opens. Click **Download Support Package** to start downloading the logs.



Download Logs

Collect trace files and dump files from your system to help resolve a problem.

Looking for individual log files? Show full log listing...

Figure 9-22 Download Support Package with the GUI

The collection of logs is initiated on every node. The command is run to gather all the logs. The final package, which is a `.tar` file, is created, compressed, and downloaded on the client.

9.7 Overview of logs

The package that is prepared is a collection of logs from different components of SONAS. The package contains different directories for each of the nodes. Example 9-6 shows the first level of logs, where each node is represented as a directory. All the logs for each of the nodes are collected under this directory. There are some logs that are collected for each node and some that are node-specific. The next sections provide detailed information about these logs and some node-specific logs.

Example 9-6 Directories in the srvdump package

```
cnrscallhomeutil-l
ctdb_scriptstatus
ctdb_statistics
ctdb_status
ctdb_uptime
get_version
mmlscluster
SoNAS.MRPD
uname-a
vpd

z_mgmt001st001
z_mgmt002st001

z_int001st001
z_int002st001
z_int003st001

z_strg001st001
z_strg002st001
z_strg003st001
zz_architeuthis_20111030-133710.node_lis


zz_architeuthis_int001st001_20111030-133710.cndump_node_log
zz_architeuthis_int001st001_20111030-133710.cngetlogs_log

zz_architeuthis_int002st001_20111030-133710.cndump_node_log
zz_architeuthis_int002st001_20111030-133710.cngetlogs_log

zz_architeuthis_int003st001_20111030-133710.cndump_node_log
zz_architeuthis_int003st001_20111030-133710.cngetlogs_log
```

The files in Example 9-7 are node-specific logs. All Interface nodes, Management nodes, and Storage nodes are seen as a directory, which is named `z_<node-name>`.

Each of these directories has the logs in the same directory structure as on the node. You also see the log of the command that was run to collect the logs on each node. They have `*.cndump_node_log` and `*.cngetlogs_log` as part of their names. See Example 9-7.

Example 9-7 List of node-specific logs in the srvdump package

```
z_mgmt001st001
z_mgmt002st001
```

```
z_int001st001
z_int002st001
z_int003st001

z_strg001st001
z_strg002st001
z_strg003st001

zz_architeuthis_int001st001_20111030-133710.cndump_node_log
zz_architeuthis_int001st001_20111030-133710.cngetlogs_log

zz_architeuthis_int002st001_20111030-133710.cndump_node_log
zz_architeuthis_int002st001_20111030-133710.cngetlogs_log

zz_architeuthis_int003st001_20111030-133710.cndump_node_log
zz_architeuthis_int003st001_20111030-133710.cngetlogs_log
```

The files in Example 9-8 are the cluster-wide logs. These logs are not per node. Most of them are cluster or system details and cluster manager details. The next sections provide more detailed information about these logs.

Example 9-8 List of cluster-wide logs in the srvdump package

```
cnrscallhomeutil-l
ctdb_scriptstatus
ctdb_statistics
ctdb_status
ctdb_uptime
get_version
mmlscluster
SoNAS.MRPD
uname-a
vpd

zz_architeuthis_20111030-133710.node_lis
```

9.7.1 Node-specific logs

This section describes the node-specific logs. Some of them are specific to Interface nodes, Storage nodes, or Management nodes. Some can be seen for each type of node.

GPFS logs

The GPFS logs from the cluster are collected. GPFS logs are usually stored in the path `/var/adm/ras/mmfs.log.*` on each node. The same directory structure is used for the package under each node. Here it is stored inside the directory that is created for each node.

GUI logs

The GUI is mostly used by the administrators to do administrative tasks and also for monitoring the system. If there is a failure, you can view GUI logs for problem determination. These logs are SONAS-specific logs.

The GUI logs are a collection of logs for operations or events with the GUI. These logs can be errors or warnings that occurred in the software code base, database, or external commands. You can modify the log level to increase or decrease the log level.

Some of the key elements are SONAS databases, CIM listeners, and business logic that handles administrative tasks.

The GUI logs are found only on the Management node. The GUI logs are found in the /var/log/cnlog/mgtsrv path.

CLI logs

The CLI is used by administrators to perform their routine administrative tasks. In case of any failure in an operation that is run from the CLI, the CLI logs can be looked into for problem determination. These logs are also SONAS-specific logs like the GUI logs.

The CLI logs are a collection of logs for operations or events with the CLI. These logs can be errors or warnings that occurred in the software code base, database, or external commands. You can modify the log level to increase or decrease the log level.

Some of the key elements are the GUI logs, SONAS databases, CIM listeners, and the business logic that handles administrative tasks.

The CLI logs are found in the /var/log/cnlog/mgtsrv path. These logs are only found on the Management node.

CTDB logs

The CTDB logs are logs from the CTDB component. It is a per-node log, so each Interface node and Management node that runs CTDB can log its messages. If there is a problem in the cluster manager, you can analyze these logs. CTDB can become unhealthy or be banned because of issues in other components. The cluster manager tries to manage the cluster and, if any component not working well or has any critical error, CTDB can become unhealthy. It is not always true that CTDB has the problems. However, the logs help you to see what might have be wrong and why CTDB behaved a certain way.

The CTDB logs are part of the /var/log/messages path on each Interface node and Management node. The following logs are additional CTDB-related logs that can be found in the root directory of each node:

- ctdb_diagnostics
- ctdb_natgwlist
- ctdb_scriptstatus
- ctdb_statistics
- ctdb_status
- ctdb_uptime

You cannot find these logs on the Storage nodes because they do not run the CTDB service.

Samba logs

Samba logs are stored in the /var/log/cnlog/samba/* path in the log.smbd file. These logs can have different levels and the “log level” parameter in the Samba configuration file can be modified to get more or less logs. These logs are found on each of the Management nodes and Interface nodes. You can also find some Samba logs in the /var/log/messages file.

Winbind logs

The winbind logs are stored in the /var/log/cnlog/samba/* path. The files have winbindd in their names. Some are related to authentication and ID mapping. These logs are found on each of the Management nodes and Interface nodes. You can also find winbind logs in /var/log/messages on the respective node.

System logs and kernel messages

The system logs or the kernel messages are in the `var/log/messages` file. These logs are on all the nodes, that is, Management nodes, Interface nodes, and Storage nodes.

NFS server logs

The NFS server logs are in the `/var/log/messages` file. These logs are only on the Interface nodes.

HTTPd logs

The HTTPd logs are in the `/var/log/httpd` folder. These logs are found only on the Interface nodes.

SSHD logs

These logs are collected from all the nodes, that is, Management nodes, Interface nodes, and Storage nodes. You can find SSHD logs in `/var/log/messages` and `/var/log/secure`.

vsftpd logs

The vsftpd logs can be found in the `var/log/messages` file. These logs are found only on the Interface nodes.

System check-out logs

The system check-out is run against DDN, the InfiniBand switches, and the Ethernet switches. You can check for warnings, errors, or even status logs. Each time a check is made, it produces a log. Also, each time a state change occurs, it produces a log. This process helps you track status and problem determination of what happened if errors occur. These logs can be found in the `/var/log/cnlog/ras` path.

Call Home modules logs

These modules are used to send Call Home information to IBM RETAIN. These logs are stored in `/var/log/cnlog`.

Installation and upgrade logs

The installation logs are logs that are written when the cluster or node is being installed or upgraded. The logs can be found in the `/var/sonas/platform.log`, `/var/log/cnupgrade.out*`, `/var/log/messages.sml`, and `/var/log/anaconda*` paths.

These logs are written on all nodes, including Management nodes, Interface nodes, and Storage nodes. These logs are all SONAS-specific logs.

CIM servers and providers logs

The CIM servers create SONAS CIM providers. These logs are written on all Interface, Storage, and Management nodes. The log is saved at `/var/log/cnlog/ras/cim/*`. You can find two files, `ibmnas_cim.log` and `cimserver.trc`.

The CIM provider provides the statuses for SONAS hardware and software components and generates indications. Some of the components that it monitors are Network, Multipath, Service, Disk, CPU, Samba, Check-out, SNMP, and VPD.

SNMP logs

The SNMP logs are used to collect traps from all hardware devices, such as IBM System x iMM HW traps, Voltaire Traps, DDN HW Traps, Enet Switch Traps, and netsnmp (trap program). The logs are `/var/log/cnlog/ras/traphandlerlog`, `/var/log/cnlog/ras/snmptraplog`, `/var/log/cnlog/ras/snmp2cim.log`, and `/var/log/cnlog/ras/snmp2cim_backlog.xml`.

These logs are SONAS-specific logs and are collected from Storage nodes and Management nodes.

Tivoli Storage Manager and HSM client logs

For Tivoli Storage Manager Client or HSM client, the logs are in `/var/log/cnlog/dsmerr.log` and `/var/log/cnlog/tsmhsm`. These logs are written on the Interface nodes and Management nodes.

Asynchronous replication logs

Asynchronous replication logs are stored in `/var/log/cnlog/async_repl/`. These logs are written on both the Management nodes on the source and target cluster. Start with the logs that are stored on the active Management node of the source cluster.

Other logs that are in cndump per node

If you log in to the `z_<`*`node-name`*`>` directory in the **`cndump`**, you can see some logs. They give you memory information, cluster manager status and statistics, registry content, some GPFS configuration logs, and more. For more information, see Example 9-7 on page 673.

9.7.2 Cluster-wide logs

The cluster-wide logs that are also contained in **`cndump`** are the cluster manager logs, such as the CTDB status, CTDB statistics, CTDB uptime, and more.

You can also see SONAS-specific information, such as the list of nodes and their roles, and the GPFS cluster and version number of the cluster. For more information, see Example 9-8 on page 674.

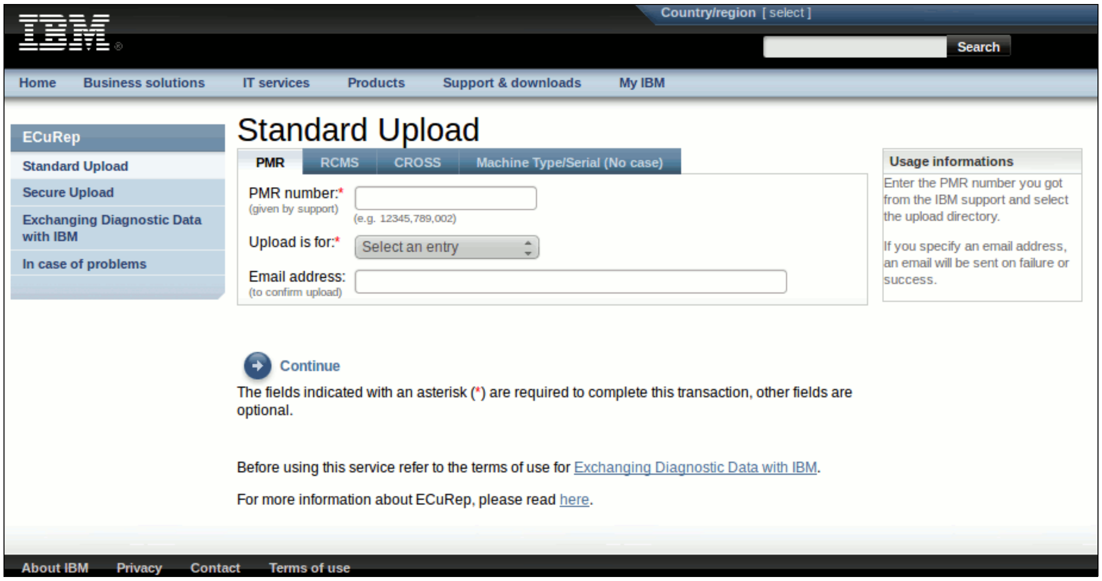
9.8 Uploading logs to IBM Support

When IBM Support requests a **`cndump`**, it must be provided to IBM. One of the ways is to upload it to the IBM ECuRep data portal. ECuRep was established as a data repository in case of system problems. From the ECuRep portal, IBM Support personnel who are working on a reported problem can access the information and logs that are provided by the customer.

To upload data to ECuRep, use your web browser and go to the following website:

`http://www.ecurep.ibm.com/app/upload`

Figure 9-23 shows the ECuRep page where you upload logs to IBM.



Standard Upload

PMR number:* (given by support) (e.g. 12345,789,002)

Upload is for:* Select an entry

Email address: (to confirm upload)

Usage informations

Enter the PMR number you got from the IBM support and select the upload directory.

If you specify an email address, an email will be sent on failure or success.

The fields indicated with an asterisk (*) are required to complete this transaction, other fields are optional.

Before using this service refer to the terms of use for Exchanging Diagnostic Data with IBM.

For more information about ECuRep, please read here.

Figure 9-23 ECuRep page where you upload logs to IBM

The following basic information must be provided at the ECuRep page:

- PMR number: Provided by IBM authorized service
- Upload is for: Other
- Email address: The email address of the person that is doing the upload

You can also select between standard or secure upload.

You can also use the CLI to upload to ECuRep, as shown in Example 9-9.

Example 9-9 Command for CLI upload to ECuRep

```
FTP to "ftp.emea.ibm.com"
login as "anonymous"
enter the email ID as password
enter bin
enter cd toibm
enter cd hw
put file_name
```

All required information about how to upload logs and data to ECuRep can also be provided by local IBM Support personnel.

Related publications

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

IBM Redbooks

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

- *IBM SONAS Best Practices*, SG24-8051
- *IBM System Storage DCS3700 Introduction and Implementation Guide*, SG24-8037
- *IBM XIV Storage System Architecture and Implementation*, SG24-7659
- *Implementing the IBM System Storage SAN Volume Controller V7.4*, SG24-7933
- *Scale Out Network Attached Storage Monitoring*, SG24-8207

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

Other publications

These publications are also relevant as further information sources:

- *IBM Scale Out Network Attached Storage - Installation Guide for IBM SONAS Gateway: Attaching IBM SONAS to IBM XIV Storage System, IBM Storwize V7000, or IBM DCS3700 or IBM SONAS Storage 2,* GA32-2223
- *IBM Scale Out Network Attached Storage Installation Guide,* GA32-0715
- *IBM Scale Out Network Attached Storage Introduction and Planning Guide,* GA32-0716
- *IBM Scale Out Network Attached Storage Troubleshooting Guide,* GA32-0717

Online resources

These websites are also relevant as further information sources:

- IBM Scale Out Network Attached Storage (SONAS) Version 1.5.1 product documentation

  http://www.ibm.com/support/knowledgecenter/STAV45/landing/sonas_151_kc_welcome.html
- IBM Scale Out Network Attached Storage product page

  http://www.ibm.com/systems/storage/network/sonas/
- IBM Support Portal

  https://www.ibm.com/support/entry/portal/support

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM SONAS Implementation Guide

SG24-7962-01

ISBN 0738440612

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

SG24-7962-01

ISBN 0738440612

Printed in U.S.A.