Redbooks
ibm.com/redbooks

IBM SONAS Best Practices

Megan Gilge

Balazs Benyovszky

David Denny

Mary Lovelace

Bill Marshall

Gabor Penzes

Ravikumar Ramaswamy

Joe Roa

John Sing

John Tarella

Michael Taylor

Shradha Nayak Thakare

Storage

IBM

Redbooks

International Technical Support Organization

IBM SONAS Best Practices

September 2015

SG24-8051-00

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

First Edition (September 2015)

This edition applies to IBM Scale Out Network Attached Storage 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.

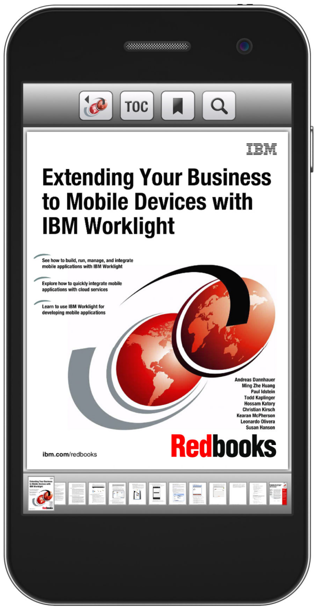
COPYRIGHT LICENSE:

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

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


TOC
IBM
Extending Your Business to Mobile Devices with IBM Worklight
Redbooks
ibm.com/redbooks

Get the latest version of the Redbooks Mobile App

iOS Download Now Android

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

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

As IBM® Scale Out Network Attached Storage (SONAS) is adopted, it is important to provide information about planning, installation, and daily administration. This IBM Redbooks® publication also describes leading tuning practices information gained by those who implement and support SONAS.

These preferred practices are based on hands-on experience from the field. Monitoring of the SONAS system is included. This IBM Redbooks publication provides information about IBM SONAS features and function at the 1.5.1 level.

This book is the companion to the *IBM SONAS Implementation Guide*, SG24-7962 IBM Redbooks publication. It is intended for readers who have implemented SONAS and are responsible for daily administration and monitoring.

Authors

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

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

Balazs Benyovszky is a Storage Technical Specialist working for IBM Systems and Technology Group (STG) in Hungary. He is responsible for IBM System Storage® presales technical support within STG. Balazs has worked for IBM since 2004 in various divisions. He worked for IBM Innovation Center for five years as a Technical Consultant on Storage and IBM Power Systems™ in the virtualization technical field. He graduated from the Technical University of Budapest in Budapest, Hungary.

David Denny is a Solutions Architect with Storage Solutions Engineering under IBM STG. David has over 20 years of experience in the information technology (IT) field, ranging from systems administration to enterprise storage architect. David's focus is the SONAS and IBM XIV® product lines, and he is the lead corporate resource for data migrations with XIV. Before joining IBM, David was a Lead Architect of the Enterprise storage area network (SAN) for the Department of Defense Disaster Recovery Program at the Pentagon following the events of September 11, 2001. He holds a Bachelor of Arts degree as well a Bachelor of Science degree in computer science from Lynchburg College.

Mary Lovelace was a Consulting IT Specialist at the ITSO. 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. Mary wrote many Redbooks publications about IBM SONAS, IBM Tivoli® Storage Productivity Center, Tivoli Storage Manager, and IBM z/OS® storage products.

Bill Marshall is a Linux subject matter expert (SME) in Rochester, MN in Strategic Outsourcing, IBM Global Technology Services® (GTS). He has a master's 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 Microsoft 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 (DFS), and kernel-based virtual machine (KVM) virtualization.

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

Ravikumar Ramaswamy is an Advisory Software Engineer with IBM India software labs in Pune, India, and holds a bachelor's degree in computer engineering from the University of Mumbai. Ravikumar has 14 years of experience in IT, most of which has been in storage and networking. He is currently working as a SONAS L3 support engineer, and is involved in various customer engagements in the Growth Markets Unit (GMU). Ravikumar also serves as a Lab advocate for SONAS customers. Before his current assignment, Ravikumar worked in IBM General Parallel File System (IBM GPFS™) functional verification test (FVT) for Linux on IBM System z®, and SONAS regression testing. He has experience with networking technologies, including Network management, and domain name server (DNS) and Dynamic Host Configuration Protocol (DHCP) implementation.

Joe Roa is an XIV and SONAS Solutions Architect working for IBM STG from upstate New York. Joe has over 25 years of experience in UNIX Server environments and Enterprise-level Storage and Data Protection. He has extensive experience in real world applications on 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 with UNIX and Storage systems. Joe holds a degree in Electronics Engineering and works primarily on SONAS and XIV platforms. He is also helping companies worldwide to solve IT storage problems with IBM storage solutions.

John Sing is an Executive IT Consultant with IBM STG, with 30 years in the IT industry. John is a world-recognized IBM expert, speaker, and strategist in the areas of large SONAS, big data and modern analytics, IT strategy and planning, and IT high availability (HA) and business continuity. Since 2001, John has been an integral member of the IBM System Storage worldwide planning and support organizations. He brings his field experience in large-scale IT from 18 years in IBM sales, marketing, and engineering. His experience includes four years in overseas growth markets, IBM Hong Kong S.A.R. of the PRC, and IBM China from 1994 to 1998. Returning to the US in 1998, John joined the IBM Enterprise Storage Server® Planning team for Peer-to-Peer Remote Copy (PPRC), Extended Remote Copy (XRC, now known as IBM z/OS Global Mirror), and IBM FlashCopy®. He was the marketing manager for these products. In 2002, he began working in business continuity and IT strategy and planning. In 2009, John took on an additional role as an IBM Storage Strategist and Technical Lead in the IBM SONAS, big data, and cloud storage arenas.

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. At present, he is working on storage infrastructures and data protection for IBM managed cloud environments. He has written extensively on z/OS Data Facility Storage Management Subsystem (DFSMS), Tivoli Storage Manager, storage area networks (SANs), storage business continuity solutions, content management, information lifecycle 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's 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®, IBM TotalStorage Virtual Tape Server (VTS), IBM System Storage TS7740, 3584 tape library, 3592 tape drives with encryption, DCS3700, IBM TS7650G 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's degree in computer science engineering and has eight years of experience. She has worked in the storage domain, with expertise in Scale out File Service (SoFS) and SONAS. She currently works as a Level-3 developer for SONAS, and also assists many client engagements for SONAS. Shradha is interested in storage products and cloud storage. She is currently focusing on SONAS authentication and authorization, assisting clients 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:

Karen Orlando
ITSO, Tucson Center

Jason Avenshine
Christian Ambach
Pratap Banthia
Tom Bish
Tom Beglin
Dave Bennin
Mathias Dietz
Mark Doumas
Greg Kishi
Doug Ledden
Andreas Luengen
Thomas Luther
Todd Neville
Chuck Quinn
Fred Stock
Mark Taylor
Renu Tewari
IBM

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 two - six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Learn more about the residency program, browse the residency index, and apply online:

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:

  ibm.com/redbooks

- Send your comments in an email:

  redbooks@us.ibm.com

- Mail your comments:

  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

Planning preferred practices

This chapter describes the type of information to consider as you plan your IBM Scale Out Network Attached Storage (SONAS) solution implementation. It is important to consider critical solution requirements.

Record this information early in the SONAS solution planning process by using the SONAS survey questionnaire. This questionnaire is provided to you by your IBM representative to begin your cluster planning. This IBM Redbooks publication is meant to highlight the key points, but it does not describe all aspects of the questionnaire.

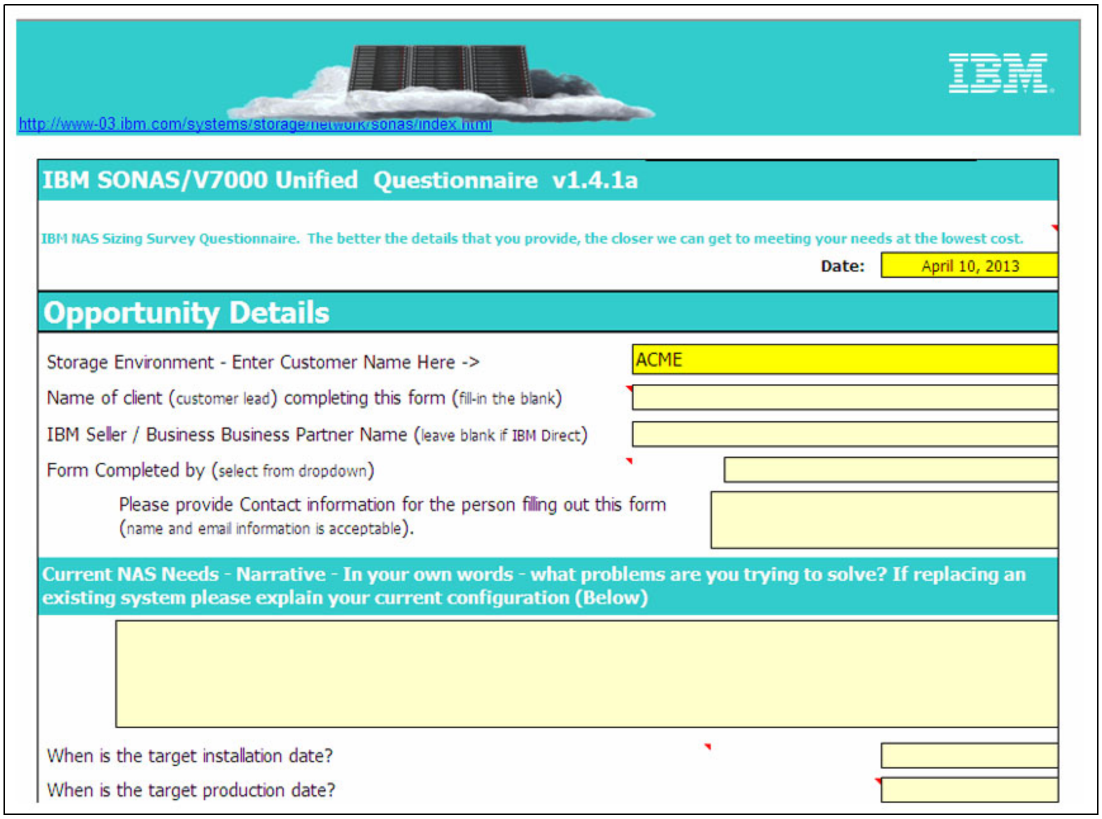
This chapter includes the following information:

- Worksheet-oriented qualification and sizing review
- Staff and training
- Environment
- Sizing

1.1 Worksheet-oriented qualification and sizing review

When you are planning a new SONAS solution, be aware of some important considerations for the requirements of your current environment. These considerations include settings, expectations, needs, and common working habits. Considering all critical NAS solution requirements is key to a successful cluster planning project. With proper planning, you can avoid performance problems and functionality defects and build a system that meets your needs.

As a pre-sales activity, your IBM Account team presents you with a SONAS questionnaire, which is similar to the one in Figure 1-1. The information that you provide helps you and IBM to understand the important information and requirements of your particular solution. Take the time that is required to respond as completely and as accurately as possible to ensure that the cluster design is optimal.



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 1-1 SONAS and IBM Storwize V7000 Unified questionnaire

If your organization is considering SONAS and your IBM Sales representative has not reviewed the questionnaire with your staff, request that review as a component of preferred practice. It is better to do it late than not to do it, and it helps you to capture the full scope of your needs in a single worksheet.

The purpose of the solution preparation questionnaire is to record as much information as possible to help clarify the requirements for the environment on which SONAS is to be installed. The questionnaire also serves as a single point of reference to capture current usage patterns within your environment. It gives you an appreciation of what needs to be considered to facilitate a successful implementation and migration.

This chapter provides information about the following topics, which can help planners understand the key targets for solution planning:

1.1.1 Environment details

Unless this installation is a net-new service, you are asked to provide information about the current NAS storage environment that will be replaced. This information helps the IBM team accurately size, compare, and qualify an appropriate solution. This section describes many key points that are considered in a full NAS solution assessment.

> **Important:** Always ensure that you are using the current survey that is available from your IBM Solution sales team.

The following information is included:

- Opportunity details

  The evaluation begins by identifying source data collection resources and generally describing the goals of the project.

  It is helpful to make sure that the people who provide the details for the solution planning survey can be identified. In some cases, the feedback details require further discussion with solution subject matter experts (SMEs). The questionnaire serves as the initial solution planning guide. Sizing and solution qualification and services are based on it.

  > **Remember:** Not every question applies directly to your needs. Answer as completely and accurately as you can. Select only the boxes for protocols that you need explicitly or immediately. If you might use options in the future, describe those requirements in the narrative explanation boxes that are provided throughout the survey.

- Capacity, growth rates, and storage tier planning

  Always express capacities in real sizes. Declare capacity requirements without compression or data deduplication. Space reduction features are often not reliable enough to predict explicit capacity goals. Data profiles must be considered explicitly to see value from space reduction techniques.

  Growth rates are often difficult to predict, but try to provide an expectation for one-year growth at the minimum. One value of the SONAS product is the ability to scale non-disruptively to very high capacities.

  Storage tiers are a good way to segment data storage behind SONAS or Storwize V7000 Unified systems. Try to predict your tiering needs with the growth rate in your planning.

  > **Remember:** The typical sizing for metadata is roughly 3% - 5%, depending on whether the metadata is replicated.

- Tape-based storage tiers

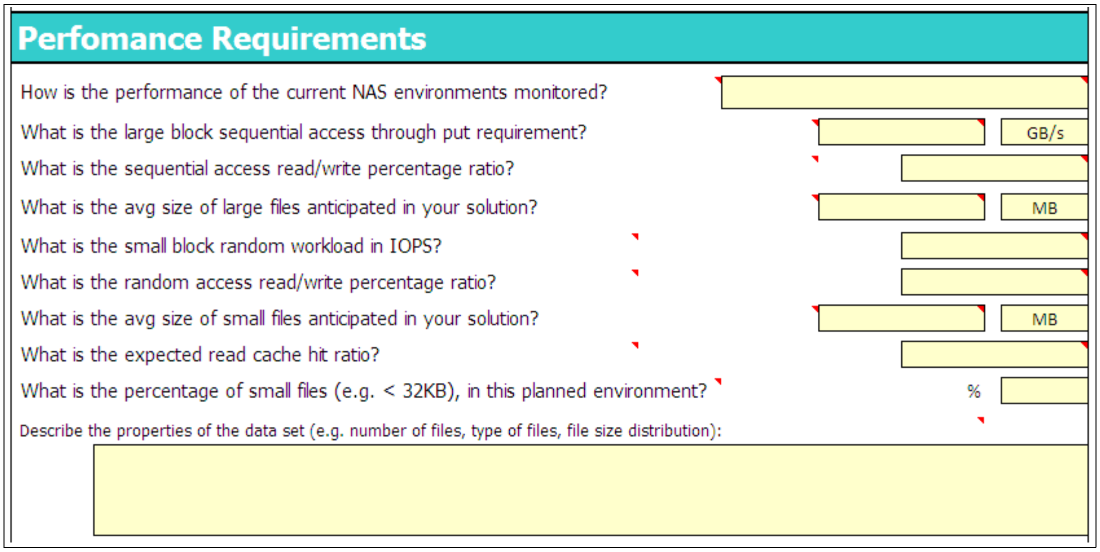
  Hierarchical storage management (HSM) space management is a popular archive solution that can be managed simply with efficient HSM and information lifecycle management (ILM) policies. However, as this tape-drive-managed access is passed via Transmission Control Protocol/Internet Protocol (TCP/IP) through the cluster and the IBM Tivoli Storage Manager server, there can be high latencies that are associated with file access and data recall (from tape back to disk).

  For this reason, it is critical to plan targeting data that is not likely to be recalled in a large scale from these archives. Retrieval of a few files here or there does not have a major effect on the system. However, retrieving millions of files at a time can create resource challenges for the SONAS and for the correlating Tivoli Storage Manager environment.

> **Tip:** Plan use for HSM tiers for true archive, and not an inexpensive tier of storage for use with data that can change or be subject to frequent recall.

1.1.2 Performance requirements

This section describes some elements of the Performance Requirements section of the questionnaire, as shown in Figure 1-2.



Perfomance Requirements

How is the performance of the current NAS environments monitored?	
What is the large block sequential access through put requirement?	GB/s
What is the sequential access read/write percentage ratio?	
What is the avg size of large files anticipated in your solution?	MB
What is the small block random workload in IOPS?	
What is the random access read/write percentage ratio?	
What is the avg size of small files anticipated in your solution?	MB
What is the expected read cache hit ratio?	
What is the percentage of small files (e.g. < 32KB), in this planned environment?	%

Describe the properties of the data set (e.g. number of files, type of files, file size distribution):

Figure 1-2 Image of Performance Requirements section of the questionnaire

When you are planning a new SONAS solution, remember the following key performance requirements:

- How performance is monitored in current NAS environments

  If this installation is a net-new NAS environment, you might not have a plan for performance monitoring yet. However, in many cases, clients do have existing solutions that they are replacing, for one reason or another. In this case, provide the performance metrics from those environments to clearly map the new system configurations.

> **Tip:** There is little value in providing solution sizing requirements that are based solely on advertised “high watermark” specifications. Typically, this method adds cost to solutions by providing functions that clients probably will not use. There is no way to size for realistic performance goals when data and input/output (I/O) workload characteristics are not clearly understood. Small file random workloads perform differently than large file sequential workloads.

- Workload descriptions

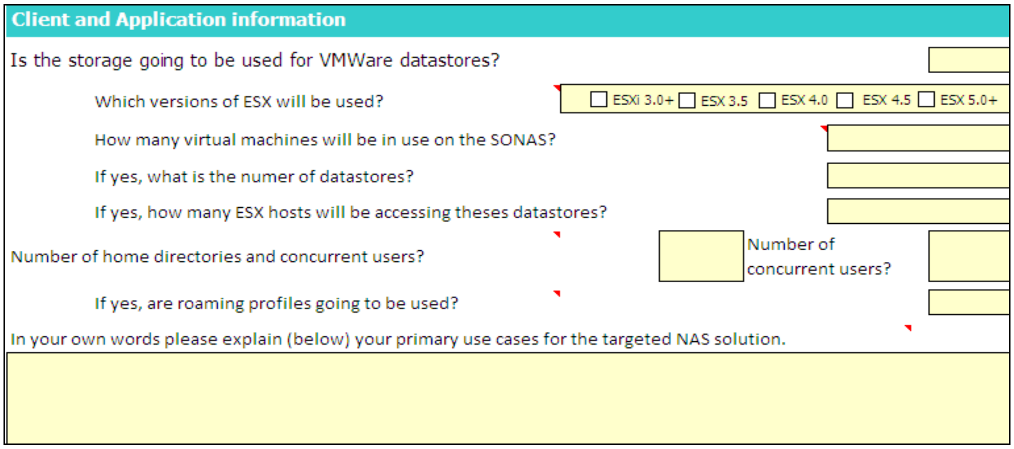
  Take the time to explain the workload characteristics that you do and do not know. For example, explain whether it is primarily for random or sequential access. Specify whether there are mostly large or small files. Explain whether there are peaks and valleys in client use. Specify whether files are reread often (cache hit ratios). That information helps with sizing the front and back-end storage requirements, and helps the team that is designing the implementation to prepare to meet the performance goals at the lowest possible cost.

- Description of your storage needs (in your own words)

  Space is provided for you to elaborate the workload and performance characteristics. It is provided in case the typical questions do not exactly cover the requirements.

1.1.3 Client applications and use cases

This section describes the planned use case for the network-attached storage (NAS). Figure 1-3 shows the Client and Application section of the survey.



Client and Application information

Is the storage going to be used for VMWare datastores?

Which versions of ESX will be used? □ ESXi 3.0+ □ ESX 3.5 □ ESX 4.0 □ ESX 4.5 □ ESX 5.0+

How many virtual machines will be in use on the SONAS?

If yes, what is the numer of datastores?

If yes, how many ESX hosts will be accessing theses datastores?

Number of home directories and concurrent users? Number of concurrent users?

If yes, are roaming profiles going to be used?

In your own words please explain (below) your primary use cases for the targeted NAS solution.

Figure 1-3 Image of Client and Application section of the questionnaire

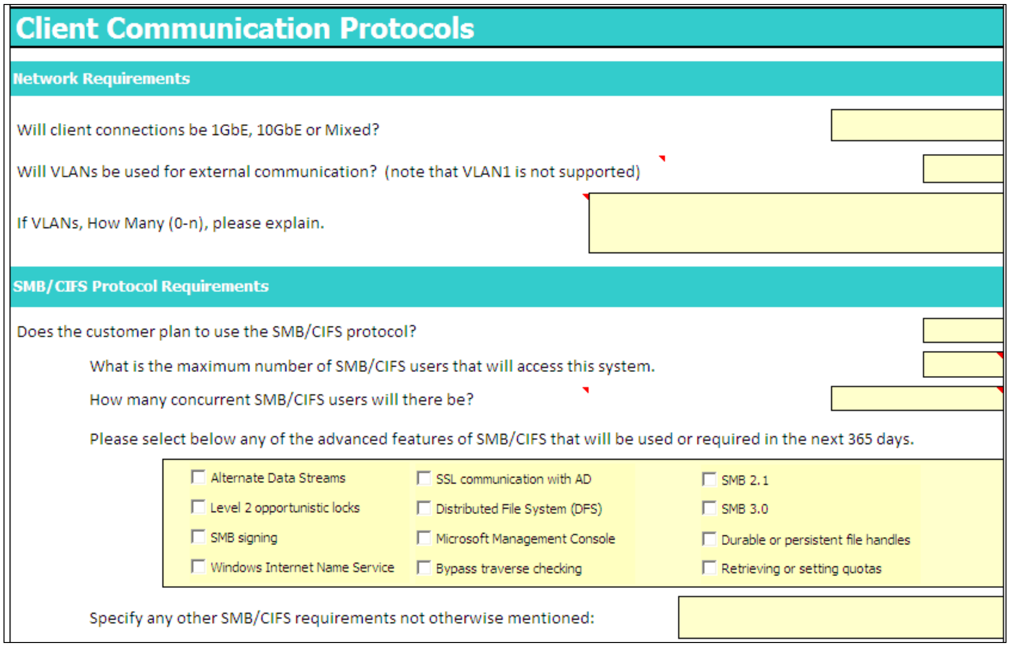
Be prepared to consider the following topics:

- Specify whether you plan to use the NAS solution for VMware storage. If so, describe what things you plan to support on the shares. If it is not otherwise obvious, describe the planned predominant use case for the solution.
- The number of concurrent users and home directory use cases might require special planning or extra interface nodes to manage the number of concurrent active users.

> **Important:** SONAS can support multiple use cases. However, one size does not fit all. In some cases, using separate file systems or network groups makes sense for providing a more multi-tenant approach to I/O isolation. Early planning for this situation prevents surprises and inevitable scaling.

1.1.4 Client communications protocols

Client connection speeds, TCP/IP and port requirements, and specific protocol requirements are required to qualify SONAS as a solution. Figure 1-4 shows a sample of the Client Communication Protocols section of the questionnaire.



Client Communication Protocols

Network Requirements

Will client connections be 1GbE, 10GbE or Mixed?

Will VLANs be used for external communication? (note that VLAN1 is not supported)

If VLANs, How Many (0-n), please explain.

SMB/CIFS Protocol Requirements

Does the customer plan to use the SMB/CIFS protocol?

What is the maximum number of SMB/CIFS users that will access this system.

How many concurrent SMB/CIFS users will there be?

Please select below any of the advanced features of SMB/CIFS that will be used or required in the next 365 days.

☐ Alternate Data Streams
☐ Level 2 opportunistic locks
☐ SMB signing
☐ Windows Internet Name Service
☐ SSL communication with AD
☐ Distributed File System (DFS)
☐ Microsoft Management Console
☐ Bypass traverse checking
☐ SMB 2.1
☐ SMB 3.0
☐ Durable or persistent file handles
☐ Retrieving or setting quotas

Specify any other SMB/CIFS requirements not otherwise mentioned:

Figure 1-4 Image of the Client Communications Protocol section of the questionnaire

SONAS is developed to meet the requirements that clients request, and the roadmap (for development) has largely followed that focus of evolution. IBM delivers a SONAS implementation only when it can be delivered well.

In some cases, advanced solutions are approved on a per-client, case-by-case basis, using a request for price quotation (RPQ) to ensure that client deployments are clearly configured and therefore supported.

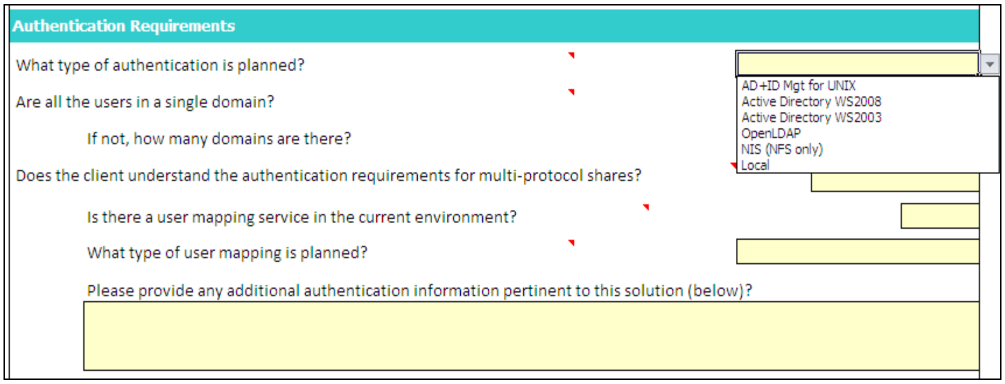
There is little value in becoming dependent on protocol models that are still in a state of flux (from a standards development perspective) or are not ready from a client stability perspective. For this reason, the questionnaire asks that you remain honest about protocols you need for day one and describe the protocols that you are interested in using in the future.

If you are not running Server Message Block 3 (SMB3) or Network File System version 4 (NFSv4) today, but plan to in the future, be sure to make that plan clear in your requirements summary.

Remember: Some protocols and check boxes in the survey might refer to points of interest that are not in the current release of SONAS or Storwize V7000 Unified, but are available in other forms of IBM-supported NAS solutions. If you want to use them in your solution, be clear about when you need them, why, and how, so that the most appropriate solution for your business requirements and applications is provided.

1.1.5 Authentication requirements

Figure 1-5 shows the Authentication Requirements section of the questionnaire.



Authentication Requirements

What type of authentication is planned?

Are all the users in a single domain?

If not, how many domains are there?

Does the client understand the authentication requirements for multi-protocol shares?

Is there a user mapping service in the current environment?

What type of user mapping is planned?

Please provide any additional authentication information pertinent to this solution (below)?

Figure 1-5 Image of the Authentication Section of the questionnaire

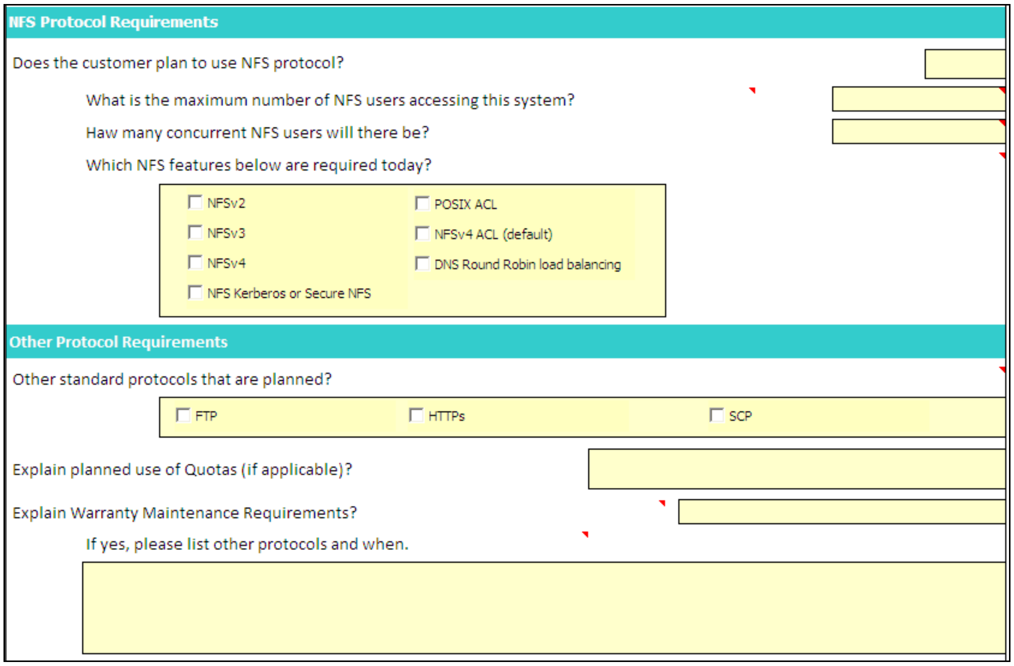
The preferred practice is generally to use Active Directory and Services for UNIX (AD + SFU) or Lightweight Directory Access Protocol (LDAP) for authentication. However, several forms of authentication are supported.

Remember: Authentication and SONAS need expanded visibility and cross-team understanding. This survey asks the questions that help to qualify the solutions that fit your business needs. In many situations, migrations from existing platforms also need consolidation of authentication along with storage.

In this case, IBM can help you design solutions that simplify the long-term goals of your enterprise storage. In any case, ensure that your authentication personnel clearly communicate with solution authentication experts to understand authentication requirements and configurations.

1.1.6 NFS protocol requirements

NFS and other protocol requirements are gathered as part of the questionnaire. This section is shown in Figure 1-6.



NFS Protocol Requirements

Does the customer plan to use NFS protocol?

What is the maximum number of NFS users accessing this system?

Haw many concurrent NFS users will there be?

Which NFS features below are required today?

□ NFSv2 □ POSIX ACL
□ NFSv3 □ NFSv4 ACL (default)
□ NFSv4 □ DNS Round Robin load balancing
□ NFS Kerberos or Secure NFS

Other Protocol Requirements

Other standard protocols that are planned?

□ FTP □ HTTPs □ SCP

Explain planned use of Quotas (if applicable)?

Explain Warranty Maintenance Requirements?

If yes, please list other protocols and when.

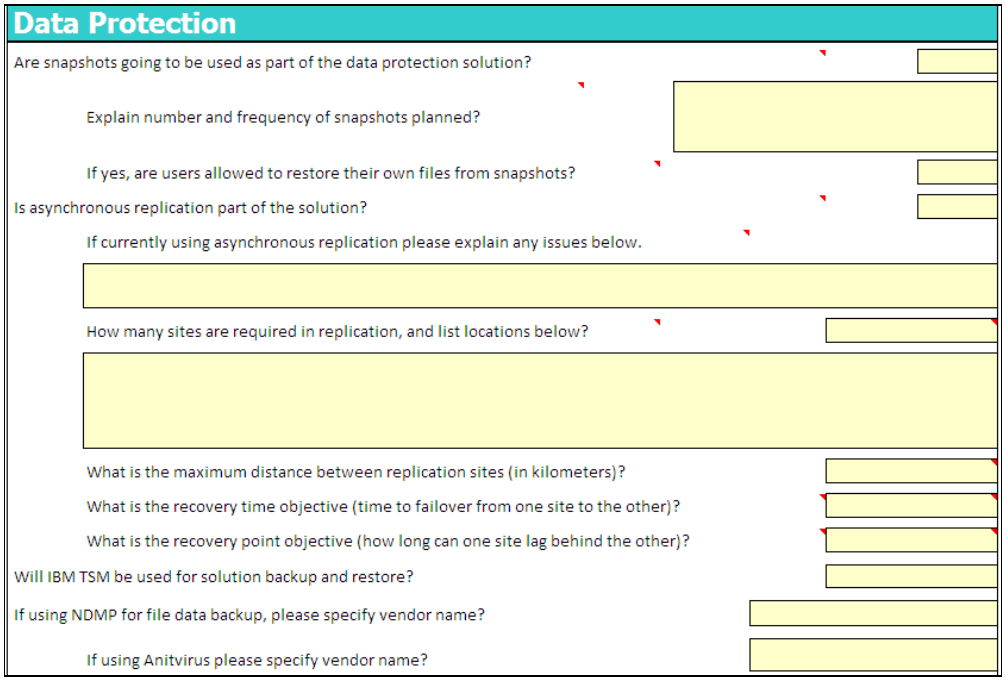
Figure 1-6 Image of the NFS and Other Protocol Requirements section of the questionnaire

For more information about supported NFS versions and file systems, see the IBM SONAS Support Matrix in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/ovr_nfssupportmatrix.html

1.1.7 Data protection

Data protection covers several features and options with scale-out NAS products. The questionnaire includes a short list of questions to consider to help you plan for implementing a SONAS or Storwize V7000 Unified solution. Figure 1-7 shows the Data Protection section of the questionnaire.



Data Protection

Are snapshots going to be used as part of the data protection solution?

Explain number and frequency of snapshots planned?

If yes, are users allowed to restore their own files from snapshots?

Is asynchronous replication part of the solution?

If currently using asynchronous replication please explain any issues below.

How many sites are required in replication, and list locations below?

What is the maximum distance between replication sites (in kilometers)?

What is the recovery time objective (time to failover from one site to the other)?

What is the recovery point objective (how long can one site lag behind the other)?

Will IBM TSM be used for solution backup and restore?

If using NDMP for file data backup, please specify vendor name?

If using Anitvirus please specify vendor name?

Figure 1-7 Image of the Data Protection section of the questionnaire

Snapshots are a common form of protecting data from unwanted changes, and allowing customer-recoverable snapshots is common. The questionnaire is designed to show your planned use of snapshots. The remaining chapters in this Redbooks publication include extensive information about snapshot management.

This book also refers to preferred practice information for asynchronous replication. A clear understanding of replication goals is critical for advanced sizing and performance planning. For more information about replication, see Chapter 7, “Data protection” on page 209 and the Planning for data protection topic in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/plan_data_protect.html

Also see the IBM Active Cloud Engine® information in *IBM SONAS Implementation Guide*, SG24-7962.

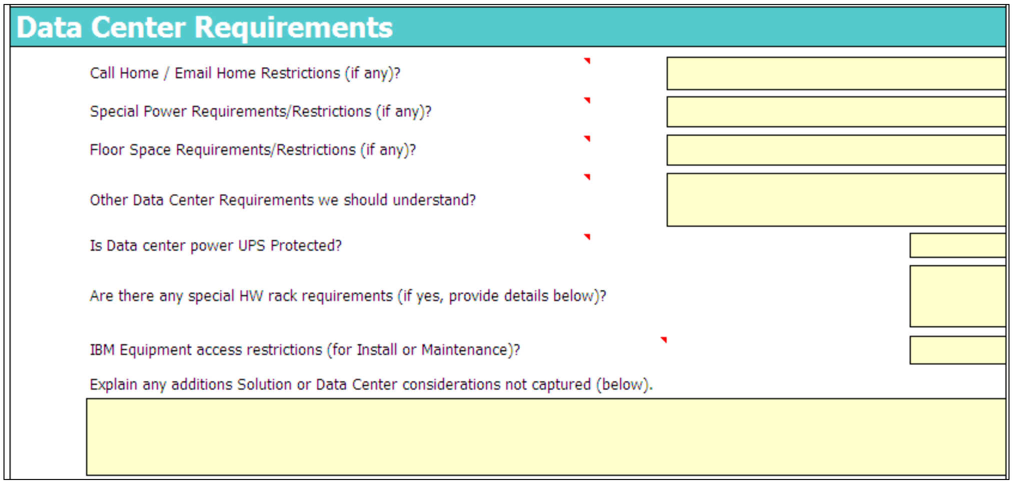
Data backup strategies are the obvious consideration for data protection. SONAS and Storwize V7000 Unified have built-in integration for Tivoli Storage Manager-based backup solutions or Network Data Management Protocol (NDMP). The questionnaire challenges the design team to consider the solution backup requirements along with the sizing information.

In many cases, this is one aspect of your environment that will grow along with your scale-out NAS. Although it is not typically installed for the explicit purpose of disaster recovery (DR), it is extremely useful for aspects of data protection and file-based restore.

Because there are many components to consider in the planning of the SONAS solution implementation, it is a preferred practice to plan this thoroughly, along with your backup expansion plan, when planning your scale out solution.

1.1.8 Data center requirements

The data center requirements can be important for many reasons. You need to identify issues that might affect installation, power, cooling, racking, delivery, or ongoing maintenance from IBM service and support resources. The Data Center Requirements section of the questionnaire is shown in Figure 1-8.



Data Center Requirements

Call Home / Email Home Restrictions (if any)?

Special Power Requirements/Restrictions (if any)?

Floor Space Requirements/Restrictions (if any)?

Other Data Center Requirements we should understand?

Is Data center power UPS Protected?

Are there any special HW rack requirements (if yes, provide details below)?

IBM Equipment access restrictions (for Install or Maintenance)?

Explain any additions Solution or Data Center considerations not captured (below).

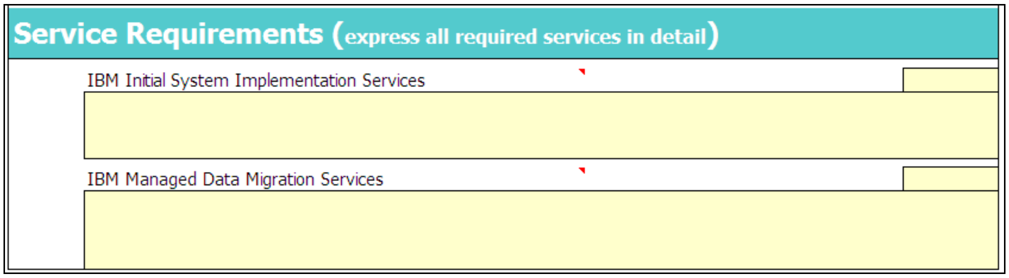
Figure 1-8 Image of the Data Center requirements section of the questionnaire

Be sure to list all hardware or data center access concerns in the narrative box that is provided in the questionnaire.

1.1.9 Service requirements

The services section is a place to list any services with which you might need assistance in any phase of your SONAS solution implementation project. You can also list preliminary services that might help you prepare and make decisions, such as knowledge transfer for advanced skills, proof of concepts, cluster installation, or data migration.

Narrative boxes are provided for free-form explanations of what you want to discuss or request for services requirements. The questionnaire leads to additional discussions with subject matter experts. Brief explanations typically suffice in this section of the questionnaire. See Figure 1-9.



Service Requirements (express all required services in detail)

IBM Initial System Implementation Services

IBM Managed Data Migration Services

Figure 1-9 The Service Requirements section of the questionnaire

1.2 Staff and training

SONAS and Storwize V7000 Unified are simple platforms to manage from a graphical user interface (GUI) and command-line interface (CLI) perspective. However, SONAS can be a complex environment that involves many external variables.

Knowledge of the following topics is critical for supporting a SONAS environment:

- Networking
- Block storage
- Network file share protocols, such as Network File System (NFS), Common Internet File System (CIFS), and Hypertext Transfer Protocol (HTTP)
- File access control lists (ACLs), and ACL management practices
- Network share exports and options that affect shares and share access controls
- Client platforms and how clients tune, manage, and control share access
- Client platform security
- Data Access Authentication
- Data backup and Restore
- Domain name server (DNS)
- Network Time Protocol (NTP) general troubleshooting skills for all of the preceding items

These topics are critical skills for NAS administration. Also consider requirements to work across teams; to articulate, document, and communicate issue descriptions; troubleshoot processes and progress; and manage advanced NAS topics.

SONAS implementations require many skills and can affect personnel in many roles. Staffing the SONAS or Storwize V7000 Unified solution is of critical importance to success as you plan a SONAS project. Plan training, backup resources, or staff augmentation to address skills gaps.

Maintaining detailed run books is a common preferred practice. These books can be used to provide instructional guidance and process management. Proper change control is critical to tracking events and issue resolution for all pertinent complex SONAS tasks.

1.2.1 Staffing considerations

This section describes requirements for SONAS staffing resources. In in some cases, multiple roles can be shared by qualified resources.

NAS management

This resource is a dedicated resource to decide, track, manage, and communicate all status, requirements, issues, and resources that are assigned to SONAS. This resource is also responsible for scale out purchase decisions.

NAS project manager

This resource is a dedicated resource to document cluster requirements across internal teams and external vendors. This resource organizes project schedules resource requirements and event timing. They also serve as a point of contact for recording and organizing all team event management and coordination of resources that are required to establish NAS policy, protocol, and maintenance.

NAS technical lead

This resource is responsible for the technical skills that are required in all stations of support. They are responsible for the accurate capture of all aspects of cluster design and layout, and can communicate up and down on all aspects of the community of NAS resource issues and requirements at an in-depth technical level for all the preceding points.

This resource must be able to work across teams with collaboration, and possess analytical expertise to help solve and coordinate all efforts that are related to scale out NAS as deployed in the client environment.

Network resource

This resource is an advanced-level network engineer who is assigned to client networking technology that can be used when network issues are causing problems. This resource must be able to articulate, communicate, and document across teams. These skills are used to help manage and troubleshoot issues with network-attached storage for all server and client networks, openly and constructively as needed.

Authentication resource

This resource is an advanced-level authentication expert who is familiar with all client data authentication environments used for SONAS. This resource must be familiar with and able to articulate, document, and communicate all aspects of client and SONAS authentication requirements and assist with all authentication challenges and troubleshooting as needed.

Backup and restore resource

This resource must understand all technical aspects of the backup and replication solution in place for SONAS. This resource must also understand all disaster recovery (DR) or business continuance requirements of the client, and how that process can be managed, validated, and implemented when necessary. This resource must know and help articulate all requirements for both the SONAS and the backup service in scale with solution requirements.

Client platform specialists

For any platform that is supported in the client SONAS environment, the client should have a technical expert that can help verify client requirements for access and control of shared data on those platforms. The expert should also verify patch, driver, and operating-system level requirements for client platform readiness, security, and data integrity in the specified platforms.

Client application experts

Each advanced application that uses shared data from the SONAS solution must include a representative application expert. This expert can help articulate, document, and communicate application requirements, conditions, status, and troubleshooting for NAS service, as needed.

1.2.2 Education

The *Scale Out Network Attached Storage (SONAS) technical training* course is available in the IBM Training catalog. Search for `SONAS` on the following website:

http://www.ibm.com/training

This course covers details and usage of the SONAS product. Hands-on exercises include configuration aspects of the product using the CLI and the GUI. The participants create shares on the SONAS and verify file access from Windows and Linux clients.

They also learn how to configure the SONAS storage and file systems, back up with Tivoli Storage Manager, automated storage tiering with ILM and HSM, along with asynchronous replication between two SONAS systems.

1.3 Environment

It is important to prepare the environment for your SONAS before delivery and deployment of the solution. This section describes the environmental requirements for your data center:

- Space requirements
- Power requirements
- Cooling requirements
- Network requirements

1.3.1 Space requirements

Space clearance is important. When you prepare the environment for the SONAS, consider the distance limitations between nodes. The nodes connect to each other by InfiniBand cables through InfiniBand switches. The location of the SONAS frames and the distance between them is limited due to InfiniBand technology.

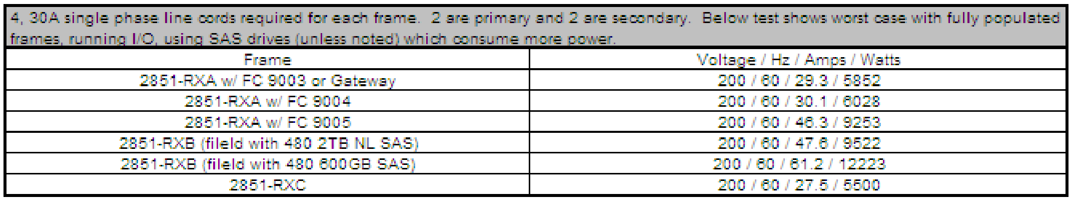
Another consideration is future growth. Remember that, for future growth and expansion, you need to plan space close to the base frame. A typical configuration is one row for your SONAS hardware, but the main concern is to keep the hardware in a concentrated location.

Your location must meet floor weight load requirements. You can find detailed information about floor load requirements, rack dimensions, and individual weights in the *IBM SONAS Implementation Guide*, SG24-7962 publication.

1.3.2 Power requirements

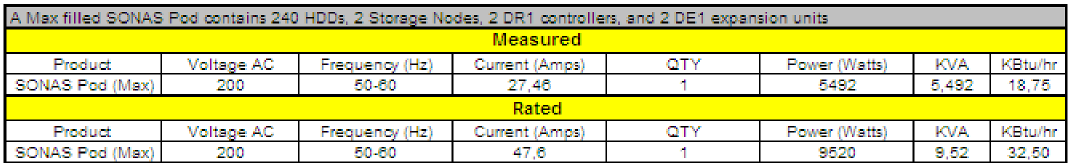
Ensure that the SONAS environment meets the proper ac power and voltage requirements.

SONAS has redundant construction for power. Every frame uses two primary and two secondary power cords (four per rack). You must plan the power environment so that one circuit can handle the full load. The full consumption of the SONAS frame especially depends on the populated nodes and the types of disk drives, in case of the storage pod, as shown in Figure 1-10, Figure 1-11, and Figure 1-12. 15,000 revolutions per minute (RPM) or 10,000 RPM SAS drives use more power than the Near-Line SAS drives.



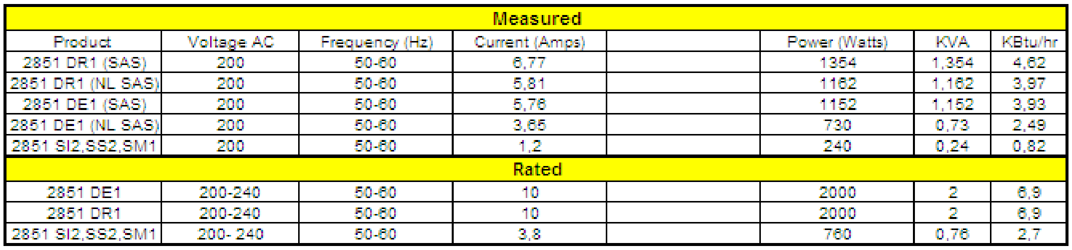
4, 30A single phase line cords required for each frame. 2 are primary and 2 are secondary. Below test shows worst case with fully populated frames, running I/O, using SAS drives (unless noted) which consume more power.	
Frame	Voltage / Hz / Amps / Watts
2851-RXA w/ FC 9003 or Gateway	200 / 60 / 29.3 / 5852
2851-RXA w/ FC 9004	200 / 60 / 30.1 / 6028
2851-RXA w/ FC 9005	200 / 60 / 46.3 / 9253
2851-RXB (fileld with 480 2TB NL SAS)	200 / 60 / 47.6 / 9522
2851-RXB (fileld with 480 600GB SAS)	200 / 60 / 61.2 / 12223
2851-RXC	200 / 60 / 27.5 / 5500

Figure 1-10 Power use for SONAS frames



A Max filled SONAS Pod contains 240 HDDs, 2 Storage Nodes, 2 DR1 controllers, and 2 DE1 expansion units							
Measured							
Product	Voltage AC	Frequency (Hz)	Current (Amps)	QTY	Power (Watts)	KVA	KBtu/hr
SONAS Pod (Max)	200	50-60	27,46	1	5492	5,492	18,75
Rated							
Product	Voltage AC	Frequency (Hz)	Current (Amps)	QTY	Power (Watts)	KVA	KBtu/hr
SONAS Pod (Max)	200	50-60	47,6	1	9520	9,52	32,50

Figure 1-11 Power use for SONAS storage pods



Measured							
Product	Voltage AC	Frequency (Hz)	Current (Amps)		Power (Watts)	KVA	KBtu/hr
2851 DR1 (SAS)	200	50-60	6,77		1354	1,354	4,62
2851 DR1 (NL SAS)	200	50-60	5,81		1162	1,162	3,97
2851 DE1 (SAS)	200	50-60	5,76		1152	1,152	3,93
2851 DE1 (NL SAS)	200	50-60	3,65		730	0,73	2,49
2851 SI2,SS2,SM1	200	50-60	1,2		240	0,24	0,82
Rated							
2851 DE1	200-240	50-60	10		2000	2	6,9
2851 DR1	200-240	50-60	10		2000	2	6,9
2851 SI2,SS2,SM1	200- 240	50-60	3,8		760	0,76	2,7

Figure 1-12 Power use for SONAS individual components

Measured power has many variables that can change the value of the power usage, and it needs to be used and interpreted with caution. Depending on what input voltage is used in your lab, what data I/O is running, and your lab infrastructure, all can play a role on measured data results. So, it is imperative to note that the measured data is an estimate, and is subject to change dependent on hardware upgrades, lab setup, and I/O choice.

If the hardware that is used is upgraded or the technology changes, the measured power data changes. Measured data is a good reference point, but set up your lab to accommodate the maximum rack power consumption numbers. With the 30A power cords, each rack can draw 9600 W. With 60A power cords, each rack can draw 19200 W. Measured data is not published.

> **Important:** The whole redundant architecture is ineffective if the SONAS is not connected to a separate power circuit. In special cases, if you require a high-availability (HA) solution, you must provide a separate power circuit from a different power provider.

Redundancy and power distribution

Redundancy might not be intuitive. Ensure that your power specialists have a complete understanding of the wired redundancy plan for your SONAS implementation.

SONAS power distribution

The SONAS RXA cabinet has four power-distribution units that help distribute the redundancy of power from all components in the SONAS (see Figure 1-13).

Each rack has either four intelligent power-distribution units (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.

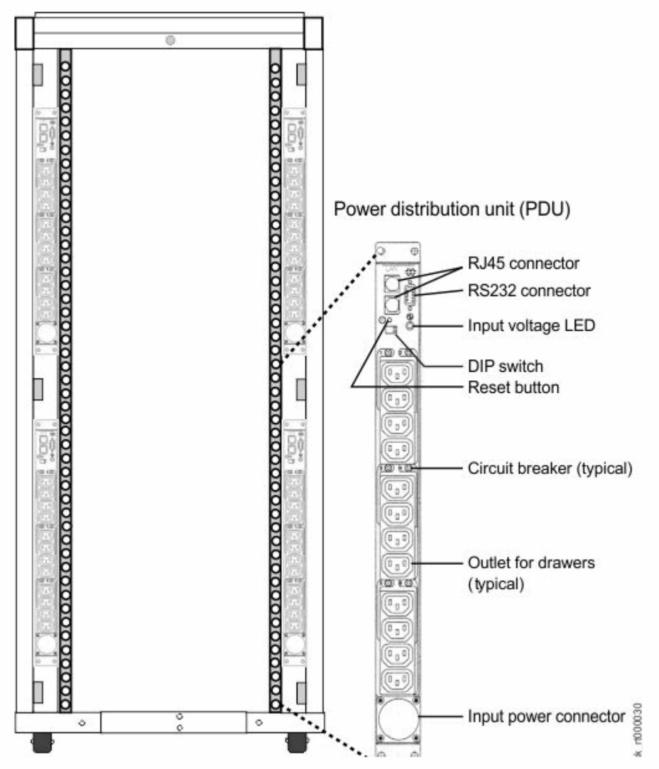


Figure 1-13 Power Distribution Unit (PDU) in a SONAS RXA frame

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

Four power-distribution units (PDUs) in an IBM SONAS rack each require a separate power source. Each of the PDUs contains twelve 200 - 240 V ac outlets that provide power to the drawers and devices in the rack, as shown in Figure 1-14 on page 16.

> **Note:** All four PDUs in the rack *must* be configured.

The IBM SONAS PDUs are split vertically in the rack with upper two PDUs as the secondary PDUs and lower two as the primary PDUs. To provide power redundancy to all the components in the rack, all four PDUs must be plugged in.

If either the upper two (secondary) PDUs or the lower two (primary) PDUs have power, the rack can still function, but without redundancy. However, if only one of the upper PDUs and one of the lower PDUs have power, that is, if only the right two PDUs, or the left two PDUs have power, the rack cannot function and some of the components in the rack will be without power.

You need to plug the upper two (secondary) PDUs into one power bus and the lower two (primary) PDUs into another power bus. So, even if the power to one bus is lost or disconnected, the rack still has power. If you connect the two right PDUs to one bus and the two left PDUs to another bus (splitting the power horizontally), and if one bus goes down or is disconnected, the rack cannot function because some components will not have power. See Figure 1-14.

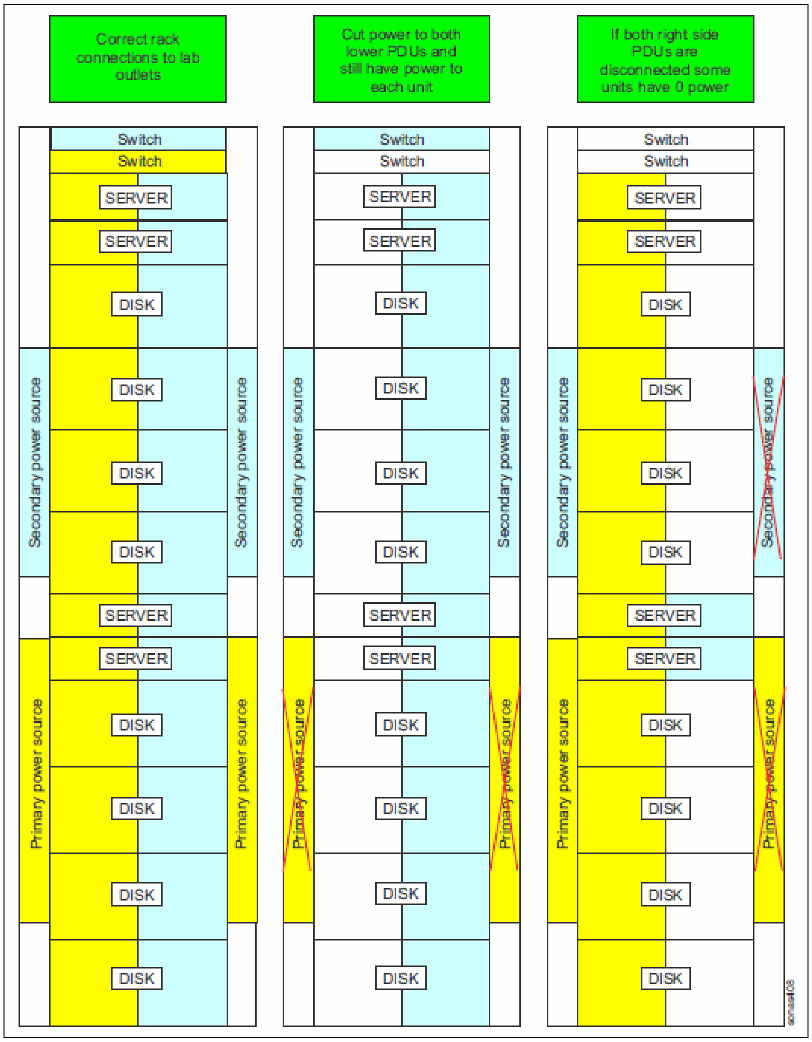

Correct rack connections to lab outlets
Cut power to both lower PDUs and still have power to each unit
If both right side PDUs are disconnected some units have 0 power
Switch
Switch
SERVER
SERVER
DISK
DISK
DISK
DISK
SERVER
SERVER
DISK
DISK
DISK
DISK
Secondary power source
Primary power source
sonas408

Figure 1-14 Power redundancy (graphical review)

The power redundancy layout is important to understand before you connect your SONAS frames to the data center Uninterruptible Power Supply Units (UPSs), and redundant power source feeds.

Cooling requirements

For the best performance and proper operation, you must optimize the cooling system of your SONAS storage solution. Use a raised floor to increase air circulation, in combination with perforated tiles. The air flow enters at the front of the rack and leaves at the back as shown in Figure 1-15.

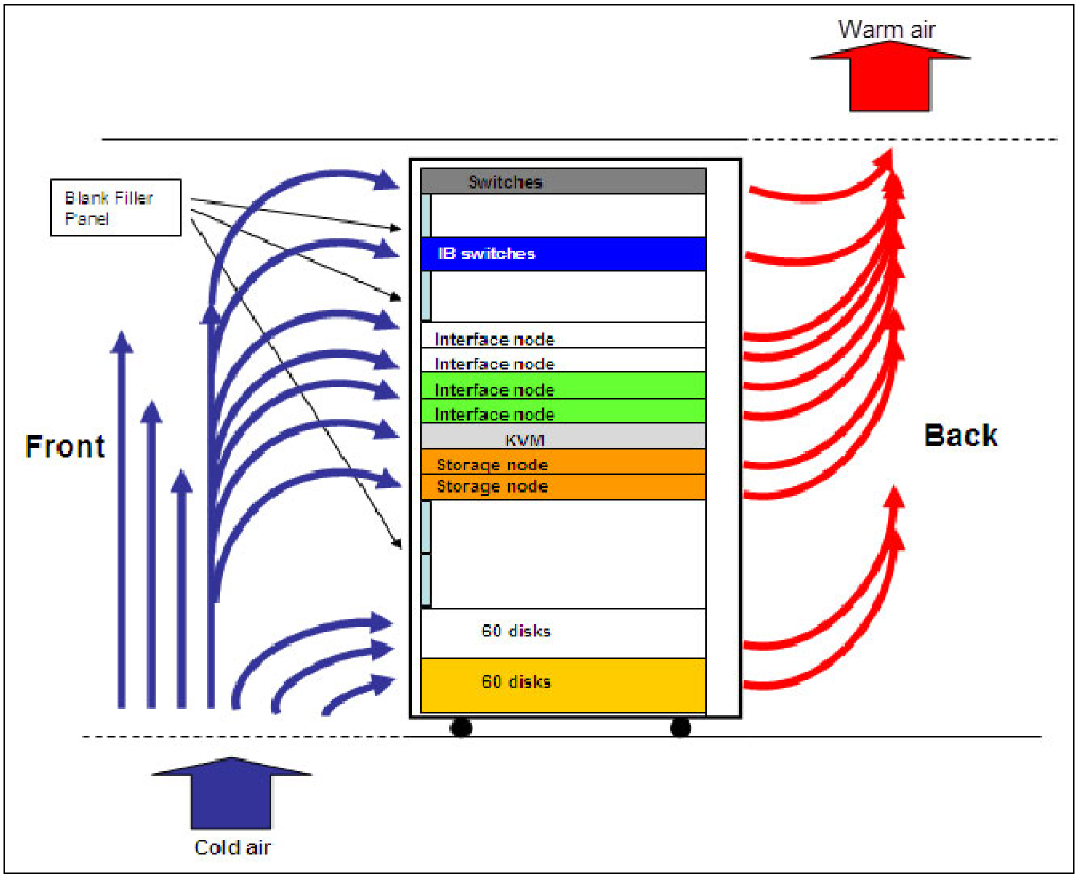


Figure 1-15 Air flow

To prevent the air that is leaving the rack from entering the intake of another piece of equipment, racks should be positioned in alternating rows, back-to-back and front-to-front, as shown in Figure 1-16. The front of racks should be positioned on floor-tile seams, with a full line of perforated tiles immediately in front of the racks, and the air temperature in front of the rack at less than 27 degrees C.

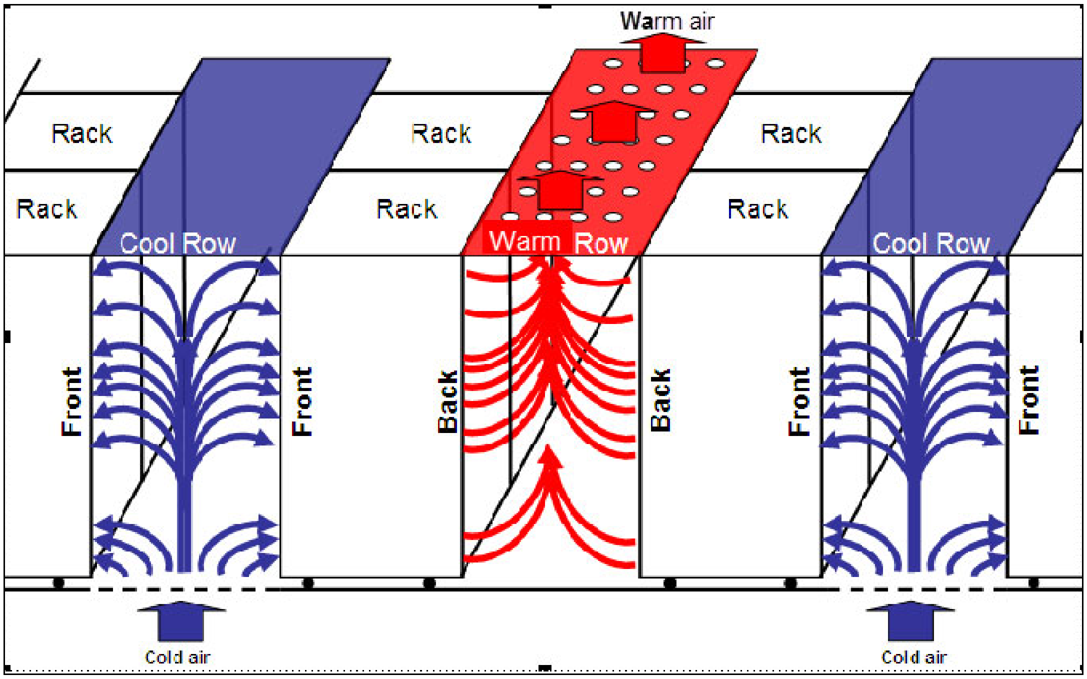
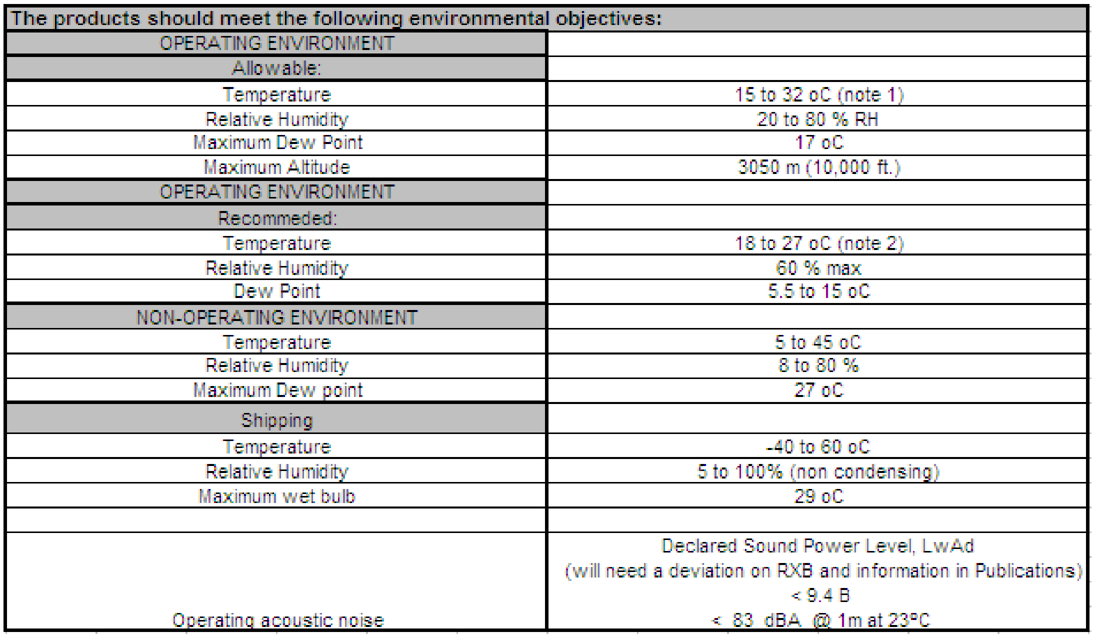


Figure 1-16 Rack positioning for proper airflow

Often forgotten in an installation, rack filler panels are an important addition. Blank panels cover up unused rack space and prevent unwanted access to equipment. Blank panels are also crucial to thermal management by controlling and restricting airflow through a rack enclosure. All unused rack space requires the usage of a filler panel.

Figure 1-17 shows the temperature and humidity information.



The products should meet the following environmental objectives:	
OPERATING ENVIRONMENT	
Allowable:	
Temperature	15 to 32 oC (note 1)
Relative Humidity	20 to 80 % RH
Maximum Dew Point	17 oC
Maximum Altitude	3050 m (10,000 ft.)
OPERATING ENVIRONMENT	
Recommeded:	
Temperature	18 to 27 oC (note 2)
Relative Humidity	60 % max
Dew Point	5.5 to 15 oC
NON-OPERATING ENVIRONMENT	
Temperature	5 to 45 oC
Relative Humidity	8 to 80 %
Maximum Dew point	27 oC
Shipping	
Temperature	-40 to 60 oC
Relative Humidity	5 to 100% (non condensing)
Maximum wet bulb	29 oC
Operating acoustic noise	Declared Sound Power Level, LwAd (will need a deviation on RXB and information in Publications) < 9.4 B ≤ 83 dBA @ 1m at 23°C

Figure 1-17 Temperature and humidity information

SONAS hardware that runs continuously must be within the recommended operating environment. Operation of the SONAS hardware at the maximum allowable temperature is only intended for short durations such as can occur during a hard drive or power supply replacement. Continuous operation above the recommended maximum temperature increases the probability of component failure.

For more information about this type of configuration, see the *SONAS Introduction and Planning Guide,* GA32-0716.

1.3.3 Network requirements

SONAS means Scale Out *Network* Attached Storage, and the network preparation and configuration are important for overall performance. For fast and correct implementation, follow these basic principles:

- Network administration

  In most cases, the network and storage administration are separated. Storage configurations are internal to SONAS. However, you must deploy SONAS external networking into the current client network environment. This deployment requires a careful review of all network requirements:

  - VLAN configurations
  - Switch configurations

    For an active-active network connection (as previously described) you must set up current switches with appropriate settings.

– Firewall configurations

– Free IP addresses

The planning guide excerpt in Table 1-1 helps to identify your networking plan for SONAS interface nodes. For example, in many cases, assigning multiple IP addresses to each interface node helps to distribute workloads evenly across the remaining interface nodes if one node fails.

Table 1-1 Interface node planning information

Number of Interface nodes	Number of bonds per Interface node	Number of IP addresses for management	Number of IP addresses for client connections	Summary of IP addresses
2	1	3	1	**4**
	2		2	**5**
	3		3	**6**
	4		4	**7**
3	1		3	**6**
	2		6	**9**
	3		9	**12**
	4		12	**15**
n	k	3	(n - 1) * k	**(n - 1) * k + 3**

Note: Every frame is configured with two mandatory, 50-port SMC gigabit Ethernet (GbE) switches that are used for the internal network. Every node, Interface (2851-SI1), Management (2851-SM1), and Storage (2851-SS1), is connected to the SMC switches with three separate connections, a primary, secondary, and maintenance Ethernet cable.

The InfiniBand switches and the iPDUs also have an Ethernet connection to these SMC switches. The disk storage units are not connected to the SMC switches. Rather, each Redundant Array of Independent Disks (RAID) controller is connected to each of the storage nodes, or, in the case of gateway attached storage, they are managed separately from the SONAS internal network infrastructure.

1.4 Sizing

Sizing is the most important component for successful solution planning. Conducting a complete and thorough review of the current environment, including all requirements and expectations, along with a full Technical Delivery Assurance (TDA) review, helps to close any gaps. The basic task is to determine what is the most appropriate SONAS configuration for your needs.

The purpose of the primary exercise is to avoid over sizing (to save on cost), but also to avoid under sizing. An undersized system might fail early on required capacity or performance and therefore require scaling growth early in the implementation cycle. The best way to begin answering the tough questions is to fully understand your client and application behavior. This includes I/O and access pattern requirements.

IBM offers advanced help in this pre-sales evaluation process. So, it is common to expect a complete review with some interaction with component experts while you size your solution for production use.

For detailed information about SONAS solution sizing, see *IBM Scale Out Network Attached Storage Introduction and Planning Guide*, GA32-0716 and the SONAS planning information in the IBM Knowledge Center:

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

The following section provides several high-level, typical sizing considerations to help to provide an overview of topic considerations. Here are some key points for a successful sizing.

Capacity sizing

The capacity sizing question is probably the easiest question to answer. The most important thing is not to build it at 100% utilization of capacity, because that choice restricts the growth availability within the solution that is purchased. Take time to consider the effect of snapshots, snapshot retention requirements, and file clones.

Make sure to plan for expanded file capacity if the data you plan to migrate is stored in tape archives, or compressed or de-duplicated file stores. Calculate enough capacity for the early growth. It might be a good starting point if you understand the needed data capacity and add 30 - 40%.

Storage tiers

The type of disks, number of spindles, and the RAID configuration can dramatically affect performance and capacity in addition to the cost of the overall solution.

Solid-state drives (SSDs) offer transaction speed benefits but also cost the most (financially). For this reason, it might make complete sense to put metadata on SSDs, while you plan the tier 1 data on high-speed serial-attached SCSI (SAS) and tier 2 data on Near-Line SAS (NLSAS), and perhaps even plan for some long-term archive to a tape-based tier. Discuss and evaluate this planning before you make your initial purchase.

Back-end sizing

Each storage node pair can drive up to about 3 gigabytes per second (GBps) in aggregate performance today, with peaks that spike well above that. Some clients drive that range of bandwidth. Note however, that this estimate is a typical estimate and I/O and workload patterns can deviate from that estimate). Consider the most common *mixed workload performance* to average 1.5 - 2 GBps per storage node pair.

This means that if the back-end storage configurations are inferior to supporting that performance, your number will be lower, and if the back-end storage is twice that fast, it will not likely be fully used through the storage node pair workload ceilings (today). So, planning back-end storage behind each storage node pair to enable using sequential performance somewhere between 3 - 4 GBps is a reasonable target for maximizing the storage channels.

Front-end sizing

Each interface node can drive up to 3.5 GBps from cache, with 4 x 10 GbE teamed converged network adapter (CNA) ports. However, do not expect all your I/O to come from cache. The pass-through expectation for each interface node with optimal networking remains somewhere in the vicinity of approximately 2 - 2.5 GBps for mixed loads (with about a 15% - 20% data reuse rate) with multiple 10 GbE ports teamed.

You might also find a performance limit of concurrent user activity for each interface node. What is meant by that is that there is a soft limit of approximately 2500 - 3000 active concurrent users per interface node today. Where users do often behave differently in many circumstances, you might consider this a typical pattern for common client use cases with concurrent, active users per interface node.

Therefore, if you expected to have 7500 active concurrent users constantly using your data shares, you might want three interface nodes in your environment to support that activity. To provide adequate protection of performance capacity if an interface node fails, you might even consider having four (for added redundancy). This limit is a soft limit because it depends on how active the users are.

Performance sizing

This is the most complex part of the sizing. You must collect much information for a fully appreciated performance sizing exercise. However, always remember that the nature of scale out NAS is that it (SONAS) enables you to grow the environment when expansion is required. So, when you do approach the ceiling or a bottleneck, evaluate where it is carefully, and expand your cluster to move that mark.

Sizing workflow

Figure 1-18 on page 23 shows a high-level sizing workflow to help you understand the phases of sizing.

> **Note:** Sizing seems complex, but scientific review of planned requirements is important. SONAS solutions include subject matter expert (SME) review with every purchase. If you feel that this review has not been offered, ask your storage representative to speak to a storage performance expert. IBM can help size storage solution requirements.

Figure 1-18 is a diagram to help you appreciate some of the considerations in review sizing requirements for scale out NAS.

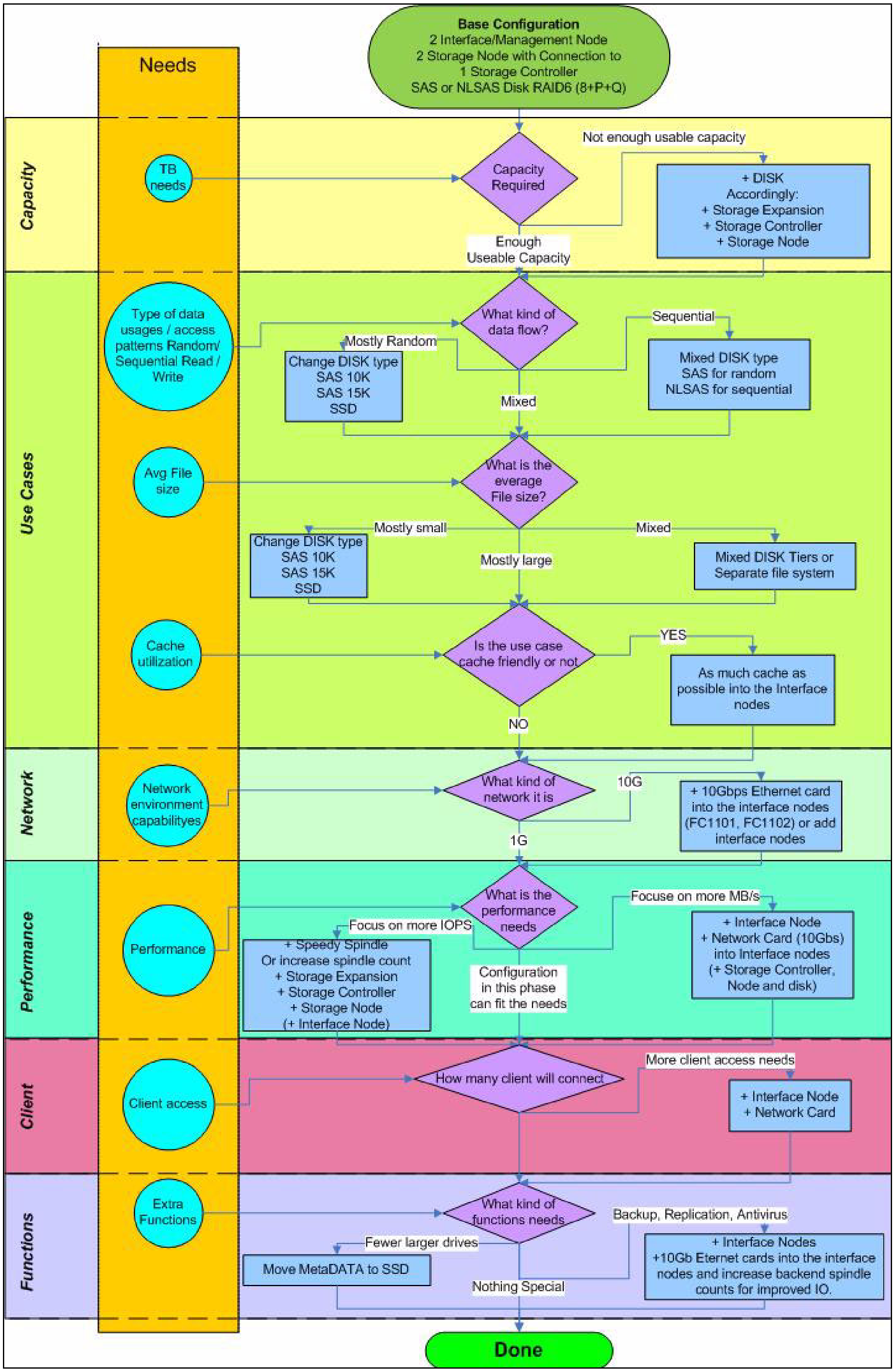


Figure 1-18 Sizing workflow

2

Authentication

Configuring authentication is one of the most important tasks after successfully installing and configuring the IBM Scale Out Network Attached Storage (SONAS) system. Setting up authentication correctly is essential, because it is the entry point to data that is stored on the SONAS system for users. Authentication and authorization ensure data security.

This chapter includes the following information:

- Introduction to authentication preferred practices
- Authentication configuration with asynchronous replication
- Preferred practices for migration of data by using the Robocopy tool
- Preferred practices for plain Active Directory (AD)
- Preferred practices for AD and Services for UNIX (SFU)
- Preferred practices for Lightweight Directory Access Protocol (LDAP)
- Preferred practices for Network Information Service (NIS)
- Preferred practices for local authentication
- Common authentication issues
- Troubleshooting authentication issues

2.1 Introduction to authentication preferred practices

It is important to plan for the type of authentication to be configured in the SONAS. Planning for authentication is based on any or all of the following considerations:

- Whether replication is configured on the SONAS system
- What type of clients access SONAS data (Microsoft Windows clients, UNIX clients, or a combination of clients)
- What authentication method was used before adding the SONAS
- Whether multiple domains are available
- How the existing data is being migrated to SONAS
- Whether the authentication servers have been upgraded recently
- Whether an external authentication server is required

> **Important:** Do not change authentication after SONAS is configured, because reconfiguration might lead to data access loss. Migration of authentication is unsupported. When it is configured, the only way to change authentication is by cleaning up all previously configured authentication and starting again.

The following sections describe the various preferred practices to be implemented for authentication to work well.

For a detailed overview of SONAS authentication, see *IBM SONAS Implementation Guide*, SG24-7962 and the Managing authentication and ID mapping SONAS product documentation in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_auth_srv_topic_welcome.html

2.2 Authentication configuration with asynchronous replication

This section explains how to configure and correct authentication with asynchronous replication.

2.2.1 Authentication requirements for replication

For proper management of files that are copied by using asynchronous replication, ensure that correct user access control information is captured with the file data. File permissions must also be copied, as-is, from source to destination.

For this copying, consistent authentication mappings between the source and target SONAS are required. The users and groups, and therefore UID and GID information, on both source and target must be identical to have the access permissions replicated correctly.

To ensure consistent user and group mappings, authentication configuration must be identical on both the source and target SONAS clusters.

2.2.2 Setting up authentication for asynchronous replication

To configure authentication to have consistent mappings or identical authentication configuration between source and target, make the user and group information, such as UID and GID, available external to SONAS and accessible by both the source and destination.

For asynchronous replication, authentication management must be done by an AD with SFU extension or by an LDAP server. Before SONAS version 1.5.1, AD without the SFU extension was not supported.

SONAS version 1.5.1 supports deterministic auto ID mapping. With deterministic auto ID mapping, the SONAS cluster is configured with AD authentication or Samba Primary Domain Controller (PDC) authentication. This ID mapping method uses a reserved ID range to allocate IDs.

For SONAS 1.5.1, asynchronous replication is also supported by the deterministic auto ID mapping mechanism. For more information about how to use deterministic auto ID mapping, see the Authentication chapter in *IBM SONAS Implementation Guide*, SG24-7962.

Authentication mechanisms (such as standard AD) that do not hold the Windows security identifier (SID) to UID/GID mapping external to the SONAS, require the mapping to be done internally to the SONAS. Before SONAS 1.5.1, there was no mechanism supported by SONAS that enabled a coordinated agreement between multiple SONAS systems to ensure that the user mapping was consistent between the discrete systems. However, with deterministic auto ID mapping with AD support in SONAS 1.5.1, the preceding restriction is no longer valid.

To summarize, the authentication service that is configured should be an LDAP, 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. For SONAS 1.5.1 and later, async replication also works with clusters that are configured with AD if the ID map configuration on source and target clusters is in sync. For more information, see the Authentication chapter in *IBM SONAS Implementation Guide*, SG24-7962.

Another important requirement is that the time on the SONAS systems must be synchronized, preferably with a common Network Time Protocol (NTP) server.

2.2.3 Consequences of incorrect authentication setup

During replication, when data is copied, asynchronous replication looks up the user and group names in the access control lists (ACLs) of the file on the target site. It then applies the respective UID and GID on the target site.

If it does not find the user and group name on the target, the UID and GID information from the source is copied. This scenario can typically occur when the authentication that is configured on the source and target are not identical. This scenario might lead to incorrect ACLs being applied with a different UID and GID set for the replicated files. As a result, data on the target might be inaccessible.

2.2.4 Correcting an incorrect configuration

It is important that, at the time of setting up the replication site or target site, even before setting up asynchronous replication between the two sites, that the authentication that is configured on the target is identical to that of source. Alternatively, it can use the deterministic auto ID mapping that is supported in SONAS 1.5.1 and later and its associated tools to maintain same ID map configuration between the source and target sites.

For more information, see the Authentication chapter in *IBM SONAS Implementation Guide*, SG24-7962.

However, if the authentication that is configured on the target is not identical to that of the source, the only way to correct it is to clean up and reconfigure authentication on the target. If there is already data on the target cluster, it might be inaccessible after reconfiguration of authentication.

2.2.5 Preferred practices for an asynchronous replication configuration

Asynchronous replication is a preferred practice to ensure that data is always available, especially when the primary site is down. Just as you prepare for setting up an alternative SONAS cluster, you must also plan to set up a secondary or alternative authentication server to service the site.

This means that you must have a secondary authentication server at the same location as the secondary SONAS site. The server is part of the same domain.

You can use this configuration to ensure that the following conditions are true:

- There is a minimum latency for data access on the secondary site.
- If both the primary site and the authentication server go down, the secondary site faces no issues to authenticate clients because it has its own secondary authentication server.
- The time synchronization across both sites is sufficient to enable successful authentication with SONAS systems.

2.3 Preferred practices for migration of data by using the Robocopy tool

The Robocopy tool is a Microsoft software tool that is provided with the Microsoft Resource Kit for Windows 2003 or as a standard tool in Windows 2008. You can use the Robocopy tool to migrate groups of Common Internet File System (CIFS) shares from any Source Filer to the IBM SONAS. In general, install the Robocopy tool on a dedicated Windows server that is attached to the network, and then copy a group of shares at a time. Figure 2-1 on page 29 is an overview of how the Robocopy tool works.

The Robocopy tool has the capability to continue to compare the source and the target files, and recopy any files that have been changed since the last copy. It also provides granularity to copy data and metadata.

After all the files are copied onto IBM SONAS, you can change the mapped shares in clients to point to the SONAS shares, and reboot and reconnect these clients.

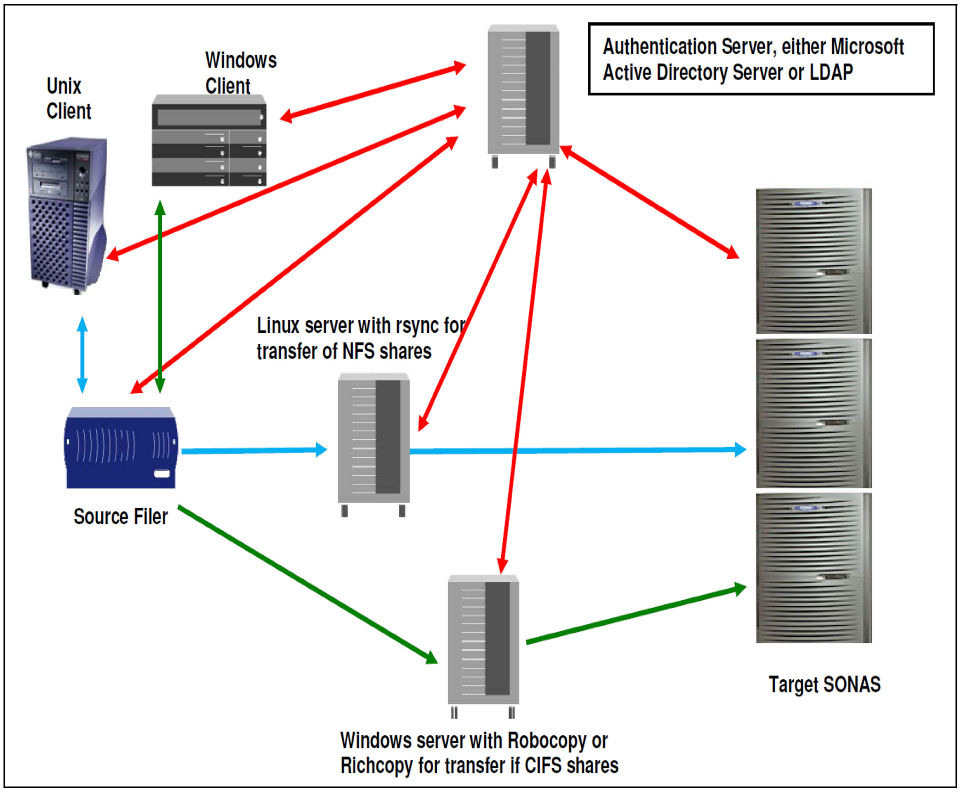

Authentication Server, either Microsoft Active Directory Server or LDAP
Windows Client
Unix Client
Linux server with rsync for transfer of NFS shares
Source Filer
Target SONAS
Windows server with Robocopy or Richcopy for transfer if CIFS shares

Figure 2-1 How the Robocopy tool works

2.3.1 Authentication requirements for replication

For proper management of files that are copied by using asynchronous replication, along with the file data, user access control information must also be correctly copied. It is essential that file permissions are also copied without alteration from source to destination.

For this copying, consistent authentication mappings between the source and target SONAS are required. The users and groups and therefore, UID and GID information on both source and target must be identical to have the access permissions replicated correctly.

To ensure consistent user and group mappings, it is mandatory that the authentication configuration is identical on both the source and target SONAS clusters.

2.3.2 Setting up SONAS for migration by using the Robocopy tool

Complete the following steps to set up SONAS for migration by using the Robocopy tool:

1. Sign in to the SONAS Management node as the root user and run the following command:

```
net conf setparm global "admin users" "[DOMAIN]\\administrator"
```

> **Important:** The **`admin users`** parameter must to be set with the user name that is used to run the Robocopy tool.
>
> After all CIFS migrations are finished, remove the settings with the following command:
>
> ```
> net conf delparm global "admin users" "[DOMAIN]\\administrator"
> ```

2. It is important to check whether the existing NAS environment uses Active Directory. If it is used, check if any AD Domain migration was done in the past. If migration was done, it results in the introduction of a new parameter, **`SIDHistory`**, with every user and group in AD.
3. Ensure that you understand security identifier (SID) history.

   SID history means that the object is assigned a new SID when it is migrated into the new AD domain, but it also retains a record of its old SID so that it can access resources on its old domain.

   Implementing SID histories works as a temporary solution, because the SID histories are the mechanisms that enable migrated users to continue to be able to access resources from the member servers in the old domain. Because of the way that SID histories work, if you were to migrate the member servers and then take the old AD domain offline, all of the SID histories that are related to the old AD domain would stop working and users would no longer be able to access resources on the member servers.

   To make the SIDs continue to work even after the old AD domain is removed, re-create the ACLs of all of the member servers as a part of the migration.

> **Important:** SONAS does not support **`SIDHistory`** (this means the SID that belongs to the old domain). Therefore, before running the Robocopy tool, replace all the ACLs with historic SIDs with new SIDs for every file and directory on the existing NAS.
>
> After this is done, the Robocopy tool can be used to copy security information from heritage NAS to SONAS.

2.3.3 Prerequisites for starting migration with the Robocopy tool

This section provides a scenario with the Source NAS filer *`A`* and the Target SONAS cluster *`B`*.

Complete the following steps to migrate data from `A` to `B`:

1. Identify Common Internet File System (CIFS) shares on A whose data and metadata needs to be migrated.
2. At the target (SONAS), export the respective file sets as CIFS shares (this step is required for the Robocopy tool).
3. Log in to the SONAS Management node as the root user and run the following command:

```
net conf setparm global "admin users" "[DOMAIN]\\administrator"
```

> **Requirement**: The **`admin users`** parameter must be set with the user name that is used to run the Robocopy tool. After all CIFS migrations are finished, run the following command to remove the settings:
>
> ```
> net conf delparm global "admin users" "[DOMAIN]\\administrator"
> ```

4. Map the source and the target on the migration server. Map the CIFS shares from `A` and `B` on the Windows server:

```
net use x: \\A\test
net use z: \\B\fset1
```

5. Migrate CIFS DATA and ACLs:

   a. Run the Robocopy tool to copy only metadata (ACL information):

   ```
   robocopy \\A\test \\B\fset1 /S /E /IS /COPY:ATSO /log:c:\robocopylog.txt /V
   ```

   This command includes the following information:

   - `/E` = Copy all subdirectories including empty ones
   - Copy file information (`/COPY:ATSO`):

     `A` = Attributes
     `T` = Timestamps
     `S` = Security = New Technology File System (NTFS) ACLs
     `O` = Owner info

   b. Run the Robocopy tool to copy data and metadata (ACL information):

   ```
   robocopy \\A\test \\B\fset1 /S /E /IS /COPY:DATSO /log:c:\robocopylog.txt /V
   ```

   This command includes the following information:

   - `/E` = Copy all subdirectories including empty ones
   - Copy file information (`/COPY:DATSO`)

     `D` = Data
     `A` = Attributes
     `T` = Timestamps
     `S` = Security = NTFS ACLs
     `O` = Owner info

   c. After the completion, the log file can be used to review errors.

   > **Important:** Make sure that the administrator or the user with which the Robocopy tool is run has read access on all files and folders from the source files that need to be migrated. If this user does not have explicit access, the user must be given backup rights and Robocopy must be run with the `/B` flag.

   > **Attention:** Migration of additional information, such as audit information, might fail if the SONAS does not permit storing this information.

2.3.4 Common issues for Robocopy configuration

This section describes common issues that are encountered with the Robocopy setup.

Issue: Data is copied but security information (ACLs) is not copied

To debug this issue, check if the files and folders have ACLs with old SIDs.

If the heritage NAS contains files and folders that have ACLs with old SIDs (historic SIDs), these ACLs are not copied.

Complete the following steps to correct this issue:

1. Remove the SID history on the heritage NAS and re-ACL all shares and re-run the Robocopy tool.
2. Check for old SIDs in the ACLs and correct them as described in step 3 on page 30.

Issue: Permission denied error

To debug this issue, check if the user has the required permissions to run the Robocopy tool.

This error typically occurs when the user with which Robocopy runs does not have permission to either read and write file data or read and write file metadata.

To correct this issue, provide the user who will run Robocopy with the correct privileges. Read point 2 from 2.3.3, “Prerequisites for starting migration with the Robocopy tool” on page 30.

Robocopy must be run with backup user privileges, and the **`admin users`** parameter from the Samba registry on SONAS must be set to the user who runs Robocopy.

Issue: ACLs are not copied for well-known and built-in groups

SONAS does not support ACLs with SIDs from well-known and built-in groups. Therefore the Robocopy tool shows errors while trying to copy ACLs for well-known and built-in groups, if any. The Windows local administrators group is an example of a built-in group.

Issue: Copy fails for files and folders with a group set as owner

SONAS does not support a group as a file owner. The Robocopy tool gives an error for the copy of owner information for such files and directories.

2.4 Preferred practices for plain Active Directory

As described in the Authentication chapter of *IBM SONAS Implementation Guide*, SG24-7962, plain Active Directory supports only Windows client access to SONAS. Avoid using this method for mixed environments where the clients accessing SONAS are Windows in addition to UNIX.

After plain AD is configured, the only way to change to another authentication method is to clean up authentication and reconfigure the new method. With this process, there might be loss of access to existing data.

2.4.1 ID mapping and range size

When you configure plain AD, planning the correct ID Mapping range and range size is important. As described in the *IBM SONAS Implementation Guide*, SG24-7962 Redbooks publication, the ID mapping that is generated, meaning the UID and GID for every Windows user and group, are created as follows:

```
ID = RANGE_LOWER_VALUE + RANGE_SIZE * DOMAIN_NUMBER + RID
```

This statement uses the following values:

- `RANGE_LOWER_VALUE`, `RANGE_SIZE` = These are specified on SONAS while running the **`cfgad`** command.
- `RID` = This is fetched from the authentication source.
- `DOMAIN_NUMBER` = This is assigned to each domain that SONAS recognizes.

The lower ID of the range must be at least 1000, and the range size must be at least 2000.

The range size (number of IDs per domain) defines the available number of UIDs and GIDs per domain. When a user or group is defined in Active Directory, it is identified by an SID that includes a component that is called a *relative identifier* (RID).

Whenever a user or group from an Active Directory domain accesses Storwize V7000 Unified, a range is allocated per Active Directory domain. A UID or GID is then allocated depending upon this range and the RID of the user and group.

For SONAS 1.5.1 and later, 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. For more information, see the Authentication chapter in the *IBM SONAS Implementation Guide*, SG24-7962 IBM Redbooks publication.

For SONAS 1.5.1 and later, 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)
```

This statement uses the following values:

- `RANGE_LOWER_VALUE`, `RANGE_SIZE` = These are specified on SONAS while running the **`cfgad`** command. See the man page of **`cfgad`** in the SONAS 1.5.1.0 IBM Knowledge Center for more details.
- `RID` = This is fetched from the authentication source.
- `DOMAIN_NUMBER` = This is assigned to each domain that SONAS recognizes, and starts with `0`.
- `MULTIPLIER` = This is generated internally by SONAS.

The default value for `RANGE` is `10000000 - 299999999` and the default value for `RANGE_SIZE` is `1000000`. Therefore, with default values, 290 domains each of size `1000000` can be mapped. The lowerID of the range must be at least 1000 and the range size must be at least 2000.

2.4.2 Configuring plain AD

When you configure authentication for the first time, you must do it correctly before you create data on SONAS.

There is no way to correct misconfiguration, except by cleaning up the authentication that is configured. However, it is not advisable to clean up and reconfigure AD.

The domain number that is generated might not be the same every time if there are multiple domains. Because the domain number is used to calculate the UID and GID, if a new value is generated, the assigned UID and GID for a user might change and they might lose access to existing data.

Choose the range lower value wisely

It is not possible to modify the lower range value after authentication is configured unless you clean up and rerun the configuration. Choose this value carefully.

Make sure that it does not clash with any existing range. If you are attempting to rerun configuration with existing data, if there is a single domain, the domain number is the same. If you choose an incorrect lower range, this selection might create different UID and GID values for users and groups. Users might lose access to existing data.

Choose the range size correctly to accommodate future growth

For SONAS versions earlier than 1.5.1, consider future expansion where the number of users and groups increases when you choose the range size. The `RID` value depends on the number of users and groups. If the `RID` of any user is greater than the range size, that user cannot access SONAS exports. So, choose the range size to enable the highest possible RID of users and groups.

The preceding restriction is removed in SONAS version 1.5.1, because multiple ranges can be allocated to domains as a result of the changes to the autorid ID generation algorithm.

Choose the number of domains to be supported

The number of domains that are supported can be calculated with the following formula:

```
Number of Domains = (Integer) <HigherID of the range> - <LowerID of the range>/<rangesize>
```

It is important to choose the lower range, higher range, and range size effectively to be able to support the future number of domains.

Multiple IBM SONAS systems that are configured against the same AD server

If there is more than one IBM SONAS system to be configured against the same AD authentication server, ensure that each IBM SONAS system has a unique netBIOS name. The netBIOS name of the SONAS system and AD server domain name must not conflict.

Migrating files

When you are migrating files from other NAS servers to SONAS, make sure that the users and groups to whom the files belong and have access to are also available on the SONAS domain controller. If the users and groups are missing, access is denied.

If users and groups who are part of another domain need access to these files, make sure that the domain is in two-way trust with the domain controller with which you configure SONAS.

If files have ACLs for users and groups who are in a domain that has been removed or upgraded, these ACLs are not copied. SONAS does not resolve those users and therefore, even if it reads the ACLs at source, at the time of applying ACLs, it skips those ACLs. This behavior is expected.

Important: The **`cfgad`** command restarts the file services:

- File Transfer Protocol (FTP)
- Secure Copy Protocol (SCP)
- Hypertext Transfer Protocol (HTTP)
- Network File System (NFS)
- Common Internet File System (CIFS)

This restart is disruptive for connected clients. The connected clients lose their connection and the file operations are interrupted. File services resume a few seconds after the command finishes successfully.

When Active Directory authentication is configured against the SONAS system, rerunning the **`cfgad`** command overwrites the earlier configuration with the new one.

2.4.3 Common issues for plain AD configurations

This section lists issues that are commonly encountered in the plain AD configuration, and explains how to correct them.

Issue: NFS users on UNIX clients are unable to access data because plain AD does not support UNIX clients

After configuring plain AD, NFS users on UNIX clients try to access data. They are denied access to data that they are able to access from a Windows client.

How to debug this issue

Complete the following steps to debug this issue:

1. Check for the UID or GID for the AD user or group that has access to the file by using the **`chkauth`** command on SONAS.
2. Check the UID or GID for the UNIX user that is denied access.

   Typically, the UID and GID are not the same. In this case, access is denied and this is expected behavior.

   The UID and GID for users on the UNIX clients is typically a smaller value like less than 1024 as compared to the UID or GID that is automatically created by SONAS.

Conclusion

If you have UNIX users who want to access data, plain AD is not the correct authentication for SONAS. Implement AD + SFU or AD + NIS or LDAP. With SONAS 1.5.1, there is support for using the local authentication server, which can help support NFS users on a UNIX client.

How to correct this issue

The only way to correct this issue is to clean up authentication by running the **`cleanupauth`** command. Use the **`--idmapDelete`** option to delete the ID mapping that was created.

Rerun the configuration command after you select the correct method for your environment.

> **Attention:** Cleaning up authentication and rerunning the authentication command might lead to data being inaccessible.

Preferred practice to be followed

Have all the information about which clients need access to data. Based on the client data access, decide on the best solution for your environment, which is typically AD + SFU or AD + NIS or LDAP.

Configure authentication based on the study. Do not copy, migrate, or create data until you are certain that the authentication that is configured is the correct one for your setup.

Issue: Users from another domain cannot access data

Even after plain AD is configured successfully, Windows users from another trusted domain are unable to access data.

How to debug this issue

If data is inaccessible, the first thing you need to check is if the user that is trying to access it has sufficient access through the file ACLs. If not, update the ACLs and try again. Make sure that users from the other domain are added in the format *`<DOMAIN_NAME>`*`\\`*`<username>`* so that it is resolved successfully.

If ACLs are sufficient, and data is still inaccessible, check if the UID and GID for that user is resolvable. Use the following command to check whether the user or group has a UID or GID assigned:

```
$chkauth -i -u "<DOMAIN\\username>"
```

As seen in Example 2-1, if the command does not show a UID or GID for that user or group, you need to check whether the SONAS cluster can resolve any of the users on the other domain controller. Check for the direction of trust.

Example 2-1 Failed to get IDMapping for user

```
$ chkauth -i -u "SONAS\user1"
EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
```

The ideal case is that there is two-way trust between both domains. If there is not a two-way trust, the direction of trust must be correct.

In a one-way trust between Domain A and Domain B, users in Domain A (trusted domain) can access resources in Domain B (trusting domain). However, users in Domain B cannot access resources in Domain A. For example, if the SONAS system is joined to Domain X, and user accounts exist in Domain Y, Domain Y must be trusted by Domain X for the users of Domain Y to gain access to SONAS shares.

Conclusion

When multiple domains exist, make sure that the trusts are configured such that the users can access resources in the SONAS domain.

How to correct this issue

The only way to fix this issue is to fix the trust direction.

Preferred practice to be followed

Have all the other domains in two-way trust with the domain to which SONAS is configured.

Issue: All AD users are denied access to SONAS even though authentication is configured correctly because of a permissions issue on AD

AD is successfully configured with the correct parameters for ID mapping. However, SONAS access fails for all users.

How to debug this issue

Check whether SONAS is inaccessible to a particular user or all users.

If it is unavailable for a particular user, check if the password is correct and retry. If the password is correct, check for ACLs and provide sufficient ACLs if they are not set yet.

If it is for all users, check if the UID and GID are assigned to every user and group. Run the following command to check for UID and GID:

```
$chkauth -i -u "<DOMAIN\\username>"
```

As shown in Example 2-2 on page 37, if the command does not show a UID or GID for that user or group, you need to check whether the SONAS cluster can resolve any of the users on the other domain controller.

Example 2-2 Failed to get IDMapping for user

```
$ chkauth -i -u "SONAS\user1"
EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
```

If there are multiple domain controllers in setup, check for the direction of trust. See "Issue: Users from another domain cannot access data" on page 35 for more information about trusts.

If you do not have multiple domain controllers in your setup, or if the trust directions are set up correctly, you need to check whether SONAS has sufficient permissions to access user and group attributes on the AD server.

Conclusion

SONAS uses a Samba component that is known as Winbind to map Windows user information into a UNIX format. When systems are joined to a Windows domain, a *machine account* that is similar to a user account is created. Winbind internally uses the machine account for user and group attribute lookup. If the machine account has insufficient privileges to read these attributes, SONAS cannot read user and group information and therefore is unable to create the UID and GID mapping that is essential to access the SONAS system.

How to correct this issue

To rectify this issue, you need to *delegate control* for the SONAS computer account to `Read all user information`.

A Windows administrator must delegate control to read user attributes to the machine account as follows:

1. In the Active Directory console tree, right-click the domain, select **Delegate Control**, click **Next**, click **Add**, and select the object type **Computers**.
2. In the object name field, enter the IBM SONAS system's machine account (the account that is created with the netBIOS name under the Computers container, which is the cluster name that is used with the **`cfgcluster`** command).
3. Click **Next**, and select **Delegate the following common tasks**. From the displayed list, select **Read all user information**.
4. Click **Next**, and then click **Finish**.

If you have multiple IBM SONAS systems, you can create a group in Active Directory, add each IBM SONAS system machine account to that group, and delegate control to that group.

Preferred practice to be followed

Check and confirm that the SONAS computer account can read all user information. Provide explicit permissions to read the user attributes by delegating control for the SONAS computer account to read all user information if not already set.

Issue: AD is successfully configured, but some users cannot access data because the RID value is out of the Range set

Plain AD was configured successfully, but some windows users are unable to access data.

> **Remember:** With SONAS version 1.5.1 and later this issue no longer exists, because the new autorid algorithm supports multiple ranges for domains.

How to debug this issue

If data is inaccessible, check if the user who is trying to access it has sufficient access using file ACLs. If not, update the ACLs and try again. For users from another domain, ensure that the trusts are configured correctly. See “Issue: Users from another domain cannot access data” on page 35 for more information about trusts.

If the ACLs are sufficient, and data is still inaccessible, check if the UID and GID for that user is resolvable. Use the following command to check if the user or group has a UID or GID assigned:

```
$chkauth -i -u "<DOMAIN\\username>"
```

As seen in Example 2-3, if the UID and GID are assigned, check for the RID of the user or group. The RID might be greater than the range size. If the RID for that user is indeed greater than the range size, those users and groups are denied any access by default.

Example 2-3 Failed to get IDMapping for user

```
$ chkauth -i -u "SONAS\user2"
Command_Output_Data     UID      GID      Home_Directory            Template_Shell
FETCH USER INFO SUCCEED 10061795 10000513 /var/opt/IBM/sofs/scproot /usr/bin/rssh
```

In Example 2-4, the **lsauth** command displays the range as 5000. The maximum value of the ID that is allowed is 10005000. The UID for the user is seen as 10061795*,* which is greater and therefore is out of range. Because it is out of range, access is denied.

Example 2-4 lsauth displaying range for ID mapping

```
$ lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,5000
domain = SONAS
idMappingMethod = auto
clusterName = bhandar
userName = administrator
adHost = SONAS-PUNE.SONAS.COM
passwordServer = *
realm = SONAS.COM
EFSSG1000I The command completed successfully.
```

Conclusion

RID for all users and groups must always be less than the rangesize specified in the **cfgad** command. It is important to consider expansion in the future, and anticipate that the number of users and groups will grow, and therefore RID will grow.

How to correct this issue

The only way to correct this issue is to provide a range size that is high enough to anticipate future expansion of the number of users and groups. However, this configuration cannot be done on the current configuration.

Attention: Cleaning up authentication and rerunning the authentication command might lead to data being inaccessible. Use this as a last resort.

Run the **cleanupauth** command to clean up authorizations that were configured previously. Rerun the command with the **--idMapDelete** option so that all UIDs and GIDs that were previously created are deleted.

Select a new range size that is feasible and rerun the **cfgad** command.

Preferred practice to be followed

Have all ID mapping-related parameters that need to be passed to the **cfgad** command revisited and recalculated to make sure that it is the minimum possible value to work as expected considering future expansion of users and groups.

Issue: Move to AD + SFU from plain AD

After they use SONAS with plain AD to store and access data, the administrator wants to migrate from plain AD to AD + SFU.

Conclusion

SONAS does not support migration of the authentication method. The only way to change configuration from one type of authentication to another is to clean up authentication by using the **cleanupauth** command and rerun the required authentication command. You might also need to run the **cleanupauth** command again with the **--idMapDelete** option before reconfiguring authentication.

Attention: Cleaning up authentication and rerunning the authentication command might lead to data being inaccessible. Use this command as a last resort.

Preferred practice to be followed

Take time to understand the possible clients that might be access data. Think of the future and anticipate the requirement before you select an authentication type. Do not change authentication because access to data might be lost in the process.

Issue: Even after successful migration of data, some ACLs are not copied onto the files because older domain SID or SID history values are not copied

SONAS was successfully configured with AD. Migration of data by using the Robocopy tool was done to copy all data from other NAS servers onto SONAS. When users try to access data, it is observed that some ACLs are copied and some are not. Some users are denied access when they try to access files.

How to debug this issue

Start with a quick check and make sure that on the source, the user or group does have ACLs on that file or folder. If ACLs are available on the source, check if the user or group has a UID or GID created. Run the following command to check for UID and GID:

```
$chkauth -i -u "<DOMAIN\\username>"
```

As seen in Example 2-5, if the UID or GID is available, check if the SID for the ACLs set on the source is the same as that on the target SONAS.

Example 2-5 UID GID mapping correct for user

```
$ chkauth -i -u "SONAS2\user2"
Command_Output_Data     UID      GID      Home_Directory          Template_Shell
FETCH USER INFO SUCCEED 10061795 10000513 /var/opt/IBM/sofs/scproot /usr/bin/rssh
```

Determine whether the domain controller was recently moved. Check the SID for the user on SONAS. If the SID History value is being referred to on the source, that can be the root cause.

Example 2-6 shows one method of obtaining the SID of a local user. The **wmic** command can also be used to query domain users.

Example 2-6

```
>wmic useraccount where name="marshall" get name,sid
Name      SID
marshall  S-1-5-21-955979511-3378639333-3486311995-1000
```

Conclusion

Typically, upon moving from one domain controller to another, the users and groups are assigned a new SID. This new SID is mapped to a UID on SONAS. However, the older SID which is in SID history is not read by SONAS. Therefore, this SID is not mapped to a UID or GID. Due to this, ACLs for the old SIDs are not copied on SONAS by using the Robocopy utility.

If files were created before data is moved, they have the older SID in the ACLs. It can happen that inheritance has been set and, therefore, many child folders might have this ACL set. Upon upgrade, the SID history is used.

When the file ACLs are being read, the SID that is sent by the source is a value that SONAS does not recognize. Because it is unknown to SONAS, a corresponding UID or GID is not created. Without a UID or GID, SONAS is not able to write the ACLs for the files on the file system.

Therefore, these files are skipped. And due to this skipping, those users will not be able to access the file system because their ACLs are not copied.

How to correct this issue

There are two ways to correct this issue:

- Re-create new folders with new inheriting ACLs on SONAS, and then copy the data without copying the ACLs. This option has the drawback that it destroys manually created ACLs where a user has been added to the access list.
- Replace all old SIDs of the SID history with the current SIDs on the source before copying. Basically, the algorithm iterates over all users. Then, for each user, query the old and new SID and then recursively replace the old SID with the new one on the file system.

  The Microsoft SubInACL tool enables recursive manipulation of ACLs. In the worst case, it must be called once for each user.

Preferred practice to be followed

Check for the SID for the ACLs on the files before you consider migration. If there has been a change in the domain and files have an older SID assigned, it might be a good idea to first replace the old SID with the new SID and then copy the data onto SONAS.

2.5 Preferred practices for AD and Services for UNIX

You can configure AD + SFU to support both Windows clients and UNIX clients to access SONAS.

Before you configure authentication on SONAS, study what is the best configuration for your environment because SONAS does not support migrating authentication from one type to another. The only way to change it is to clean up and reconfigure.

Here, SFU works as an external ID mapping server. The authentication only happens using Active Directory. External ID mapping enables the ID mapping to remain external to the SONAS system. This enables multiple SONAS systems to access the same UID and GID information for users and groups. So, for asynchronous replication, both the source and the target are able to see the same user and group information, and therefore apply ACLs correctly.

2.5.1 Range and Schema Mode parameters in SFU

This list describes the range and schema mode parameters:

- Range. The range for the range parameter must be in the format `LowerID - UpperID`; for example, `20000 - 30000`. The allowed range is from `1 - 4294967295`.

  Make the lower range greater that 1024 to avoid conflict with the management command-line interface (CLI) users. When you run the command with a lower range that is less than 1024, a warning is generated, which also asks for confirmation. You can use the **`--force`** option to override it.

  > **Important:** The specified range must not intersect with the range that is specified with the **`--idMapConfig`** option of Active Directory server authentication configuration. The default range for the Active Directory server authentication is `10000000 - 299999999`.

  Users and groups that have a UID or GID that does not fit in the range are denied access.

- SchemaMode. This parameter value 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`.

2.5.2 Configuring AD + SFU correctly before data is created on SONAS

When configuring authentication for the first time, it is essential you do it immediately before creating or migrating data on SONAS. There is no way to fix any misconfiguration, except by cleaning up the authentication configuration. However, it is not advisable to clean up and reconfigure AD.

The domain number that is generated might not be the same every time if there are multiple domains. Because this number is used to calculate UIDs and GIDs, if a new value is generated, the assigned UID and GID for a user might change, and they might lose access to existing data.

Also, check for the schema mode and make sure that you choose the correct value. SONAS does not verify if you have chosen the correct schema mode.

Choose the User and Group Range value wisely

It is not possible to modify the range for the user and group for SFU after authentication is configured unless you clean up and rerun the configuration. Therefore, you need to choose this value with great care.

The value should be such that the UNIX user ID (UID) and group ID (GID) which will be mapped in SFU for the respective Windows user and group are in that range.

Make sure that it does not clash with the existing range that is provided for AD in the **`--idMapConfig`** command because this range will be used for the Windows user whose UID and GID will be automatically created. This value should not clash with the existing SFU UID and GID.

Check that every user has the UID and GID for its primary Windows group

After AD + SFU is configured with the correct parameters, set up the users and groups. Make sure that all the users have the UID set. Also, the Primary Windows Group for each user must have a valid GID using the Active Directory administration tools.

If other groups that the user is a member of do not have a GID, the group does not have access to data. However, if the user's Primary Group does not have a valid GID, that user is denied access to data.

Migrating files

Make sure that when you are migrating files from other NAS servers to SONAS, the users and groups to whom the files belong and have access to, are also available on the SONAS domain controller.

If the users and groups are missing, access is denied. These users must also have the UNIX attributes updated in the SFU so that the UID or GID is not auto created, and it is set correctly when files are copied onto SONAS. This way the Windows users and UNIX users can access the data.

If users and groups who are part of another domain need access to these files, make sure that the domain is in two-way trust with the domain to which you have joined SONAS.

If files have ACLs for users and groups who are in a domain that has been removed or upgraded, these ACLs are not copied. SONAS does not resolve those users and groups, even if it reads the ACLs at source. At the time of applying ACLs on SONAS, it skips those ACLs. This behavior is expected.

2.5.3 Common issues for AD + SFU configurations

This section lists common issues that you might encounter in an AD + SFU configuration, and describes how to correct them.

Issue: AD + SFU successfully configured, but some users cannot access data - UID or GID not set correctly

Even after AD + SFU is configured successfully, some Windows users are unable to access data.

How to debug this issue

If data is inaccessible, the first thing you need to check is whether the user trying to access the data is successfully resolved by the system. For users from another domain, make sure that they are also resolved by having the trusts that are configured correctly. For more information about trusts, see “Issue: Users from another domain cannot access data” on page 35.

If, after sufficient ACLs, data is inaccessible, verify that the UID for that user is set. Use the following command to check if the user or group has a UID or GID assigned:

```
$chkauth -i -u "<DOMAIN\\username>"
```

Example 2-7 shows that the **chkauth** command does not show the UID or GID for user1. If the **chkauth** command does not show UID or GID, even though the SONAS cluster can resolve the users on the domain controller, the UID or Primary group might not have a valid GID set in SFU for that user.

Example 2-7 Failed to get ID mapping for user

```
$ chkauth -i -u "SONAS\user1"
EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
```

In Example 2-7, you can see that the user information cannot be fetched, which means that it is not yet present on the SONAS system.

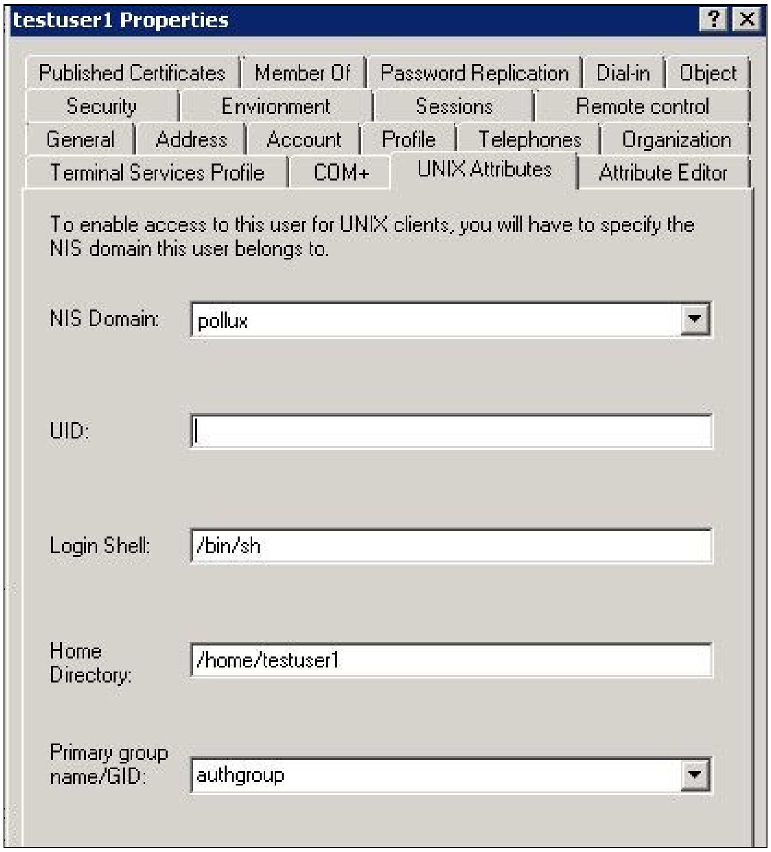
Conclusion

Access for those users and groups is denied if UID or GID are not set correctly.

How to correct this issue

Complete the following steps to correct this issue:

1. Open the User or Group details on the domain controller by right-clicking Properties on that user or group. Open the UNIX Attributes tab and update the UID or GID for every user or group.
2. As shown in Figure 2-2, the UID for the user is missing. Complete this value with the valid UID and save. The user now has a valid UID set and the user should be resolved.



To enable access to this user for UNIX clients, you will have to specify the NIS domain this user belongs to.

NIS Domain: pollux

UID:

Login Shell: /bin/sh

Home Directory: /home/testuser1

Primary group name/GID: authgroup

Figure 2-2 UID to be completed for the user in the UNIX Attributes tab in the properties for the user

> **Note:** Make sure that the Primary Group that is set for the user in the **Member Of** tab also has a valid GID for the user to be resolved successfully.

Preferred practice to be followed

It is mandatory that you set up SFU correctly before you store data. Assign a UID and GID to users and groups. The preferred practice is to verify that the setup is correct before you store data or access the cluster.

Issue: UID successfully configured, yet some users cannot access data: User’s Primary Windows Group does not have a valid GID

Even after AD + SFU is configured successfully, some Windows users are unable to access data.

How to debug this issue

Complete these steps to debug user access:

1. If data is inaccessible, and the UID is set correctly, the first thing you might want to check is if the user that is trying to access data has sufficient ACLs. If not, provide the ACLs and try again. Make sure that users from another domain are also resolved by configuring the trusts correctly. For more information about trusts, see “Issue: Users from another domain cannot access data” on page 35.
2. If ACLs are sufficient, and data is still inaccessible, check if the UID for that user is set. Use the following command to check whether the user or group has a UID or GID assigned:

   ```
   $chkauth -i -u "<DOMAIN\\username>"
   ```

3. If the output from the **chkauth** command does show the UID, check whether the user’s Primary Windows Group has a valid GID.

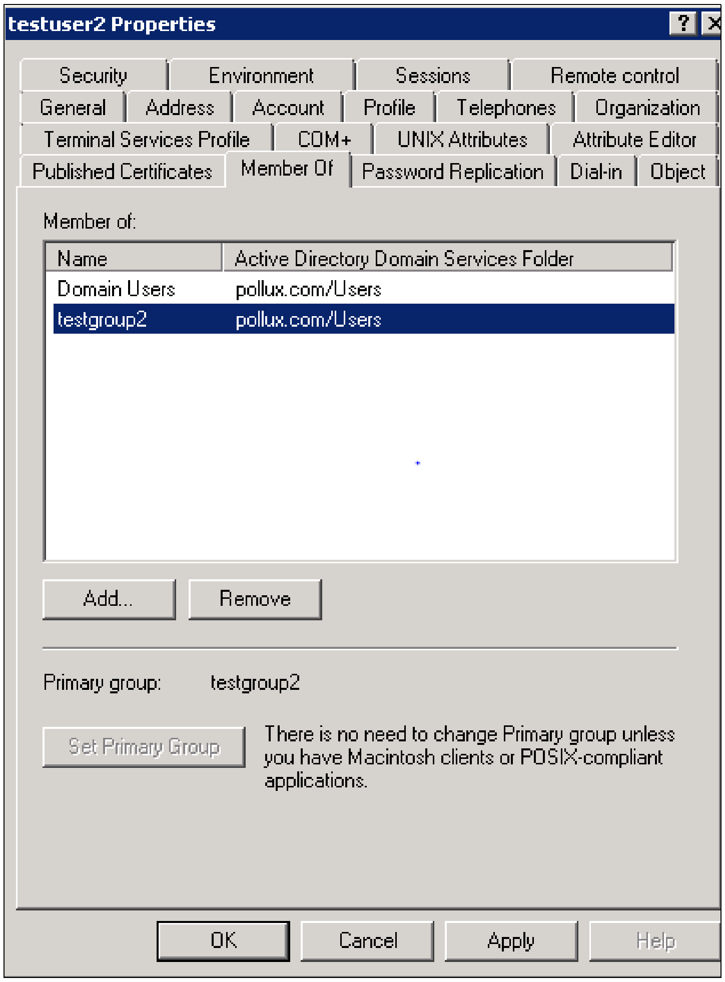
   Example 2-8 shows that the user’s UID is correctly mapped. However, access is still denied for the user. This means that the user is not set up correctly.

   Example 2-8 Failed to get IDMapping for user

   ```
   $ chkauth -i -u "SONAS\user1"
   EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
   ```

4. Next, check whether the Primary Group that is assigned to the user has a valid GID set. If this setting is missing, the user is denied access even if the UID has been set.

5. Check which Primary Windows Group is set by accessing the user on the AD server. Right-click **Properties** and select **Member Of**, as shown in Figure 2-3.



Member of:

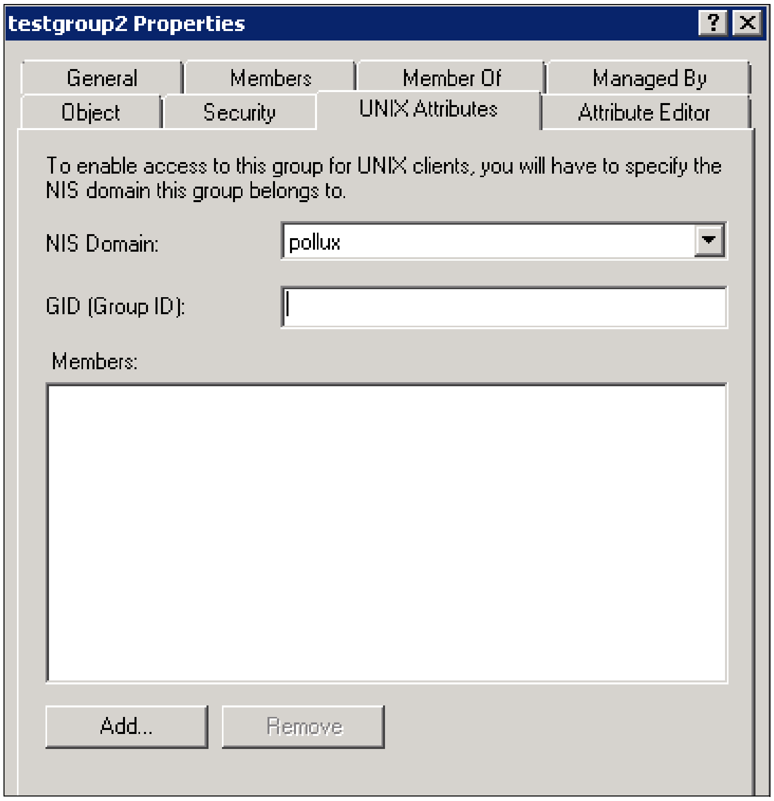
Name	Active Directory Domain Services Folder
Domain Users	pollux.com/Users
testgroup2	pollux.com/Users

Primary group: testgroup2

There is no need to change Primary group unless you have Macintosh clients or POSIX-compliant applications.

Figure 2-3 Check for Primary Group

6. Now, check for the Group properties for this group. Right-click **group,** select **Properties,** and then **UNIX Attributes.** See Figure 2-4.



testgroup2 Properties

To enable access to this group for UNIX clients, you will have to specify the NIS domain this group belongs to.

NIS Domain: pollux

GID (Group ID):

Members:

Figure 2-4 Primary Group has a missing GID

Here, the GID is missing. This missing value is the cause for access denied.

Groups that do not have a valid GID are denied access. If a Primary group has a missing GID, access is denied for the respective user.

Conclusion

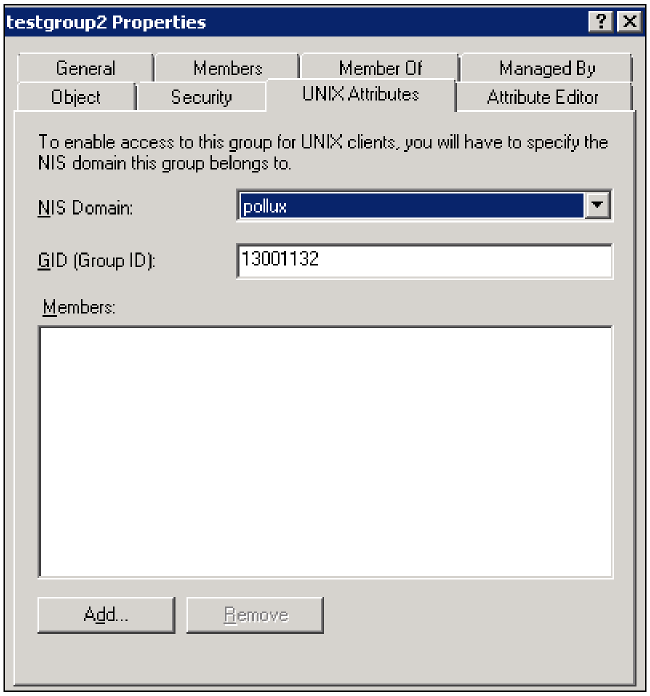
Access for those users Primary Windows group set for user does not have a valid GID.

How to fix this issue

Complete the following steps to fix this issue:

1. Open the user or group details on the domain controller.
2. Right-click **Properties** on that user or group.
3. Open the **Member Of** tab and check which is the Primary Group set for that user.
4. Now, click the corresponding group and check the GID by right-clicking **Properties** on that user or group.

5. Open the UNIX Attributes tab. If the GID is missing, enter a valid GID for the name. See Figure 2-5. Make sure that the GID is in the range that is specified when you configure authentication on SONAS.



To enable access to this group for UNIX clients, you will have to specify the NIS domain this group belongs to.

NIS Domain: pollux

GID (Group ID): 13001132

Figure 2-5 Primary Group now has a valid GID

Preferred practice to be followed

It is mandatory that the user's UID and Primary Group (GID) are correctly set. The preferred practice is to verify that these steps are followed before you try to store data or access data.

Issue: UID and GID details are correctly set in SFU, but some users cannot access data because the UID is out of SFU Range

Even after AD + SFU is configured successfully, where the UIDs for all users and the Primary GID are set and the GID has been set for all groups, some Windows users are unable to access data.

How to debug this issue

Complete the following steps to debug user access:

1. If data is inaccessible, the first thing that you need to check is if the user who is trying to access the data has sufficient ACLs. If not, provide the ACLs and try again. Make sure that users from another domain are also resolved by having the trusts configured correctly. See "Issue: Users from another domain cannot access data" on page 35 for more information that is related to trusts.

2. If ACLs are sufficient, and the data is still inaccessible, check if the UID and GID for that user is set. Use the following command to check if the user or group has a UID or GID assigned:

   ```
   $chkauth -i -u "<DOMAIN\\username>"
   ```

3. As seen in Example 2-9 on page 49, if the **`chkauth`** command does show the UID or GID, check for the range set for each user and group. The UID and GID set for users and groups must be within the User range and Group range.

You can also check this value on the AD side on the UNIX Attributes tab in the properties for that user. See Example 2-9.

Example 2-9 UID out of range for user

```
$ chkauth -i -u "SONAS\user2"
Command_Output_Data     UID      GID      Home_Directory
Template_Shell
FETCH USER INFO SUCCEED 5403 6789 /var/opt/IBM/sofs/scproot /usr/bin/rssh
```

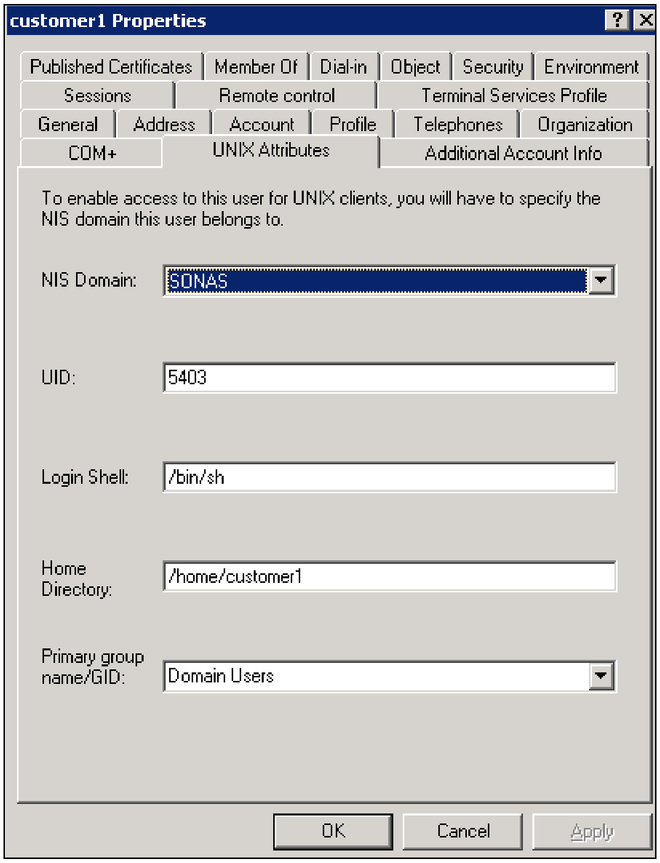
4. You can check for the user range and group range using the **lsauth** CLI command.

   In Example 2-10, the range that is specified is 6000 - 7000. Therefore, every user must have a UID in this range.

Example 2-10 The lsauth command displays the range for the UID and GID

```
$lsauth -c st001.virtual1.com
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
SFU_virtual1 = ad,6000-7000,rfc2307
domain = virtual1
idMappingMethod = sfu
clusterName = st001
userName = administrator
adHost = ad1.virtual1.com
passwordServer = *
realm = virtual1.com
EFSSG1000I The command completed successfully.
```

The user tab shows the UID as 5403 in Figure 2-6. This value does not fit in the range and therefore access is denied. Any user whose UID does not fall in this range is denied access.



To enable access to this user for UNIX clients, you will have to specify the NIS domain this user belongs to.

NIS Domain: SONAS

UID: 5403

Login Shell: /bin/sh

Home Directory: /home/customer1

Primary group name/GID: Domain Users

Figure 2-6 The UNIX Attributes to be set in SFU are out of range

Conclusion

Access is denied for those users whose UID does not fall in the range that was specified when configuring SFU.

How to correct this issue

The only way to correct the issue is to reconfigure SFU. For this step, you need to run the **`cfgsfu`** command again with the correct parameters.

Do not reconfigure SFU because, during this process, other clients might lose access to data because the cluster is in an inconsistent state while authentication is being configured.

Preferred practice to be followed

Set up SFU correctly before you store data. The range must be correctly planned and set to accommodate all users.

Issue: Users are denied access to SONAS even though authentication is configured correctly because of a permissions issue on AD server

AD + SFU is successfully configured with the correct parameters for ID mapping. However, access to SONAS fails for all users.

How to debug this issue

Complete the following steps to debug user access:

1. Check if SONAS is inaccessible to particular user or all users.
2. If for a particular user, check is password is correct and retry. Check if user has sufficient ACLs. Provide with ACLs if not available.
3. If it is for all users, check if the UID and GID are assigned to every user and group. Run the following **chkauth** command to check for UID and GID:

   ```
   $chkauth -i -u "<DOMAIN\\username>"
   ```

4. As shown in Example 2-11, if the command does not show the UID or GID for that user or group, you need to check if the SONAS cluster can resolve any of the users on the other domain controller.

   Example 2-11 Failed to get IDMapping for user

   ```
   $ chkauth -i -u "SONAS\user1"
   EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
   ```

5. If there are multiple domain controllers in the configuration, check for the direction of trust. See “Issue: Users from another domain cannot access data” on page 35 for more information that is related to trusts.
6. If you do not have multiple domain controllers in your setup or if the configured trust directions are also correct, you need to check whether the SONAS has sufficient permissions to access user and group attributes on the AD server.

Conclusion

Winbind internally uses the machine account for user or group attribute lookup. If *machine account* has insufficient privileges to read these attributes, SONAS will not be able to read user and group information, and therefore is unable to create UID and GID that is essential to access the SONAS system. This requires explicit read permissions for the SONAS system machine account to read the user attributes.

How to correct this issue

To rectify this issue, delegate control for the SONAS computer `account[object type]` to `Read all user information`.

To do this, on the OU containing the users and groups, delegate control for the computer account to read user attributes. Complete the following steps:

1. In the Active Directory console tree, right-click the domain, select **Delegate Control,** click **Next,** click **Add,** and select the object type **Computers**.
2. In the object name field, enter the IBM SONAS system’s *machine* account (the account that is created with the netBIOS name under the Computers container), which is the cluster name that is used with **cfgcluster** command.
3. Click **Next**, and select **Delegate the following common tasks**.
4. From the displayed list, select **Read all user information**.

5. Click **Next**.
6. Click **Finish**.

If you have multiple IBM V7000 Unified systems, you can create a group in Active Directory, add each IBM V7000 Unified system machine account to that group, and delegate control to that group.

Preferred practice to be followed

Check and confirm that SONAS machine account can read all user information. Provide explicit permissions to read the user attributes by delegating control for the SONAS computer account to read all user information if it is not already set.

Issue: The UID is correctly set in SFU, but users cannot access data because the SFU Schema mode is incorrectly set

AD + SFU are up correctly, but access is still denied for some users.

How to debug this issue

Complete the following steps to debug user access:

1. If data is inaccessible, the first thing you need to check is if the user who is trying to access data is successfully resolved by the system. Make sure that users other domains are also resolved by having the trusts configured correctly. See “Issue: Users from another domain cannot access data” on page 35 for more information about trusts.
2. Use the following command to check whether the user or group has a UID or GID assigned:

   ```
   $chkauth -i -u "<DOMAIN\\username>"
   ```

   The output in Example 2-12 shows that the UID and GID have been set correctly.

   Example 2-12 Successfully displaying ID mapping for a user

   ```
   $ chkauth -i -u "SONAS\user2"
   Command_Output_Data     UID      GID      Home_Directory
   Template_Shell
   FETCH USER INFO SUCCEED 5403 6789 /var/opt/IBM/sofs/scproot /usr/bin/rssh
   ```

3. In this case, check that the UID and GID are within the range that is provided when configuring authentication with the **`cfgsfu`** command. See “Issue: UID and GID details are correctly set in SFU, but some users cannot access data because the UID is out of SFU Range” on page 48 for more information.
4. If the values are also within range, check if the required permissions are also set correctly. See “Issue: Users are denied access to SONAS even though authentication is configured correctly because of a permissions issue on AD server” on page 51 for more information.
5. If all of the previous settings are correct, verify that the schemas are correct. Check if the schema provided in the **`cfgsfu`** command is the correct one.

   The schemaMode 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`.

Conclusion

If the schema chosen during **cfgsfu** is incorrect, due to the difference in schema variables, authentication does not work as expected. Verify that you chose the correct command parameters before proceeding with the configuration.

How to correct this issue

To rectify the issue, the only method is to reconfigure authentication with the correct values. Run the **cfgsfu** CLI command again with the correct schema mode value.

Preferred practice to be followed

Ensure that you choose the correct values at the time of configuring authentication.

Issue: The UID is set correctly in SFU, but the ID mapping that is displayed shows a different UID, and access is denied because the IDMap cache must be cleared when modifying the SFU UID

AD + SFU is set up correctly, but access is still denied for some users. When the UID is checked, different values are shown.

How to debug this issue

Complete the following steps to debug user access:

1. If you see that the UIDs are shown as different values than what is set in AD under UNIX Attributes, determine where this value is coming from.
2. Check whether the UID for that user or group has been changed recently. If so, that change might be the root cause.

   SONAS usually caches the UID for one week. If any change has been made to the UID set previously, the cached UID is used first until cache is flushed.

Conclusion

UIDs for all users are cached on SONAS. This cache is locally per interface node. For every positive login, the UID is cached for 1 week. For every negative login, the UID is stored for 2 minutes.

How to correct this issue

To rectify the issue, remove the entry for the user from the cache by manually running the **rmidmapcacheentry** command. For more help on this command, see the following information in the SONAS area of the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_t_auth_configure_IDmapping.html

Preferred practice to be followed

Do not change UID and GID values. If it must be done, the administrator should run the **rmidmapcacheentry** command to clear the cache entry.

Issue: SFU is correctly configured but some users cannot access data because the UIDs for some users are auto generated

AD + SFU is set up correctly, but access is denied for some users. When the UID is checked, different values are shown.

How to debug this issue

Complete the following steps to debug user access:

1. If you see that the UIDs are seen as different values than what is set in AD under UNIX Attributes, you need to check from where this value is coming.
2. Check whether the UID for that user or group was changed recently. If so, that might be the root cause.

   SONAS usually caches the UID for one week. If any change has been made to the UID set previously, the cached UID is picked up first until the cache is flushed.
3. If the UID has not been changed recently, check whether **`cfgsfu`** was run immediately after **`cfgad`** was run. In some cases, **`cfgsfu`** is run after data has been copied and accessed. What happens then is that the UID and GID are automatically generated when plain AD is configured. This UID and GID are stored in cache. When, after days or weeks, **`cfgsfu`** is configured, even though a valid UID or GID is set for the users, SONAS first picks up the values from `autorid.tdb` because the value exists there.

Conclusion

UID and GID existing in cache or autorid.tdb is root cause. SONAS picks up the older values and therefore access is denied for SFU users.

How to correct this issue

To correct the issue, complete the following steps:

1. Remove the entry for the user from the cache by manually running the **`rmidmapcacheentry`** command*.*
2. Also, run **`cleanupauth --idMapDelete`** to clean up the existing authentication. The **`--idMapDelete`** option deletes the autorid.tdb database.
3. Now run the **`cfgsfu`** command to reconfigure SFU.

> **Important:** Running **`cleanupauth`** cleans up the configuration. Run this command with caution. If users from other domains are accessing SONAS, they might lose access.

Preferred practice to be followed

Run **`cfgsfu`** immediately after you run the **`cfgad`** command. Run it before any user tries to access exports or data. Upon first login, the UID GID is automatically created if a plain AD setup exists.

2.6 Preferred practices for Lightweight Directory Access Protocol

As described in the Authentication chapter of *IBM SONAS Implementation Guide*, SG24-7962, LDAP on SONAS supports access for both Windows and UNIX clients. Therefore, for mixed environments with both Windows and UNIX clients, this method works well.

Before you configure authentication, make sure that you study what method is the best for your environment, because SONAS does not support authentication migration from one type to another. The only way to change the authentication is to clean it up and reconfigure.

Here, LDAP works as an authentication server in addition to working as an external ID mapping server. External ID mapping enables the ID mapping to remain out of the SONAS system. This method allows multiple SONAS systems to access the same UID and GID information of users and groups. So, for asynchronous replication, both the source and target are able to see the same user and group information, and therefore apply ACLs correctly.

2.6.1 Tasks before you get started

Before SONAS can be configured with an external LDAP server, you must verify the validity of the LDAP server. For this purpose, the following tasks must be done:

1. Validate that Management node can create SSH connections to the Interface nodes.
2. Check whether at least one of the LDAP servers is pingable. Verify that the LDAP server is up and responds by sending an LDAP query with the filter `objectClass=*`.
3. Verify that LDAP server is configured with at least a few users by sending an LDAP query with the filter `ou=People`.
4. 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.
5. The validation is run on each of the Interface nodes. If SSL/TLS is being configured, the certificate is also checked at this stage itself. Also, if Kerberos is enabled, the keytab file is checked for entries of the following types:
   - `cifs/`*`<cluster-name>`*`.`*`<domain>`*
   - `cifs/`*`<hostname>`*

> **Important:** The verification procedure does not change any configuration for an existing authentication (if any).
>
> For LDAP-KRB-CIFS configuration, plain LDAP users are still accepted. If the requirement is to allow access only for Kerberos users, set some invalid passwords in the LDAP database for the users.

2.6.2 Prerequisites for configuring the IBM SONAS system with LDAP

Before you configure SONAS, the LDAP schema must be extended to enable storing of additional attributes, such as SID, Windows password hash, and so on, to the LDAP user object.

The prerequisites are discussed in the Authentication chapter of *IBM SONAS Implementation Guide*, SG24-7962.

2.6.3 Preferred practices for configuring LDAP

This section describes the preferred practices to follow when you configure LDAP.

Configuring LDAP correctly before data is created on SONAS

When you configure authentication for the first time, it is essential you do it correctly before creating or migrating data on SONAS.

There is no way to correct any misconfiguration, except by cleaning up the authentication configuration.

User and group ID values must be greater than 1024

This is both a limitation and a preferred practice. Ensure that the UID for users and the GID for groups are greater than 1024. If the UID or GID is lower, access is denied.

Verify that all users and groups have valid UID and GID

Before migrating or copying data, make sure all the users and groups that will be accessing data have valid UID and GID. Do not change these values in the future, especially if data exists on SONAS.

Users with the same user name from different organizational units in the LDAP server

Users with the same user name from different organizational units within the same LDAP server trust are denied access to CIFS shares without regard to the LDAP User Suffix and LDAP Group suffix values configured on the system.

Limitation with the chkauth command and LDAP configuration

For LDAP-based authentication, the **`chkauth`** command validates access by using the userPassword attribute of the user in the LDAP directory server. If the userPassword attribute is inconsistent with the sambaNTPassword, CIFS access might be denied even if **`chkauth`** authentication is successful.

> **Attention:** The **`cfgldap`** command restarts the file services (FTP, SCP, HTTP, NFS, CIFS), which is disruptive for connected clients. The connected clients lose their connection and the file operations are interrupted. File services will resume few seconds after the command finishes successfully.
>
> When LDAP authentication is configured against the SONAS system, rerunning the **`cfgldap`** command overwrites the earlier configuration with the new one.

2.6.4 Common issues for LDAP setup

This section lists common issues that occur during LDAP setup.

Issue: LDAP is successfully configured, but some users cannot access data because the UID or GID is less than 1024

LDAP is set up correctly, but access is denied for some users.

How to debug this issue

Complete the following steps to debug user access:

1. If access is denied, check whether the ACLs are sufficient for user access. If the ACLs are set correctly, check for the UID or GID for the user or group that requires access.
2. Run the **`chkauth`** command to check for the UID or GID.

   ```
   $chkauth -i -u "<DOMAIN\\username>"
   ```

   In Example 2-13, it can be seen that UID is less than 1024. This is the cause of the issue.

   Example 2-13 Failed to get IDMapping for user

   ```
   $ chkauth -i -u "SONAS\user1"
   EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
   ```

Conclusion

The UID for all users or GID for all groups should be greater than 1024. If it is less than 1024, SONAS denies access for the user or group.

How to correct this issue

To correct the issue, you need to update the UID for the users whose UID is lower than 1024.

> **Attention:** Users lose access to data that has ACLs with the older UID or GID.

Preferred practice to be followed

It is essential that the UID and GID for users and groups is greater than 1024 when using LDAP authentication. Access is denied if any UID or GID is lower than 1024.

Issue: LDAP is successfully configured, but some users cannot access data because the user name is less than three characters

LDAP is set up correctly, but access is denied for some users.

How to debug this issue

Complete the following steps to debug access:

1. If access is denied, check whether the ACLs are sufficient for the user to access. If ACLs are set correctly, check for the UID or GID for the user or group that requires access.
2. Run the **`chkauth`** commands to check for the UID or GID:

   ```
   $chkauth -i -u "<DOMAIN\\username>"
   ```

In Example 2-14, it can be seen that UID is correctly set. However, it can be seen that the name of the user who is trying to access is fewer than three characters.

Example 2-14 Failed to get IDMapping for user

```
$ chkauth -i -u "SONAS\ab"
Command_Output_Data     UID      GID      Home_Directory
Template_Shell
FETCH USER INFO SUCCEED 10061795 10000513 /var/opt/IBM/sofs/scproot
/usr/bin/rssh
```

This is the root cause. This example is shown in Red Hat Enterprise Linux (RHEL). This is a known issue in RHEL 6.0 releases and has been fixed in RHEL 6.2 and later release. SONAS 1.4 uses RHEL 6.1. This is not an issue for SONAS 1.5.1 and later.

Conclusion

LDAP user and group names must have more than three characters to work correctly.

How to correct this issue

There is a workaround for this issue on older SONAS releases. This section provides tasks for applying manual configurations and for removing them.

Do the following steps to apply manual configurations. These steps require root login to one of the nodes:

1. Copy the stock `/etc/nslcd.conf` file that is generated to a temporary directory:

   ```
   cp /etc/nslcd.conf /tmp/tmp_nslcd.conf
   ```

2. Add the following lines to the end of the temporary file:

   ```
   # Workaround to allow 2 characters user names
   validnames /^[a-z0-9._@$()][a-z0-9._@$() \~-]*[a-z0-9._@$()~-]$/i
   ```

3. Update the registry with this new file:

   ```
   fileContent="$(cat /tmp/tmp_nslcd.conf; printf x)"
   # printf jig takes care of new line character at EOF
   net registry setvalue HKLM/SOFTWARE/IBM/SOFS/AUTH NSLCD_CONF sz
   "${fileContent%x}"
   ```

4. Reload cnscmc on all nodes:

   ```
   onnode all cnscmc reload auth
   onnode all service nslcd stop
   ```

5. Remove the temporary file:

   ```
   rm -f /tmp/tmp_nslcd.conf
   ```

Complete the following steps to remove the manual configurations:

1. Copy the stock `/etc/nslcd.conf` file that is generated to a temporary directory:

   ```
   cp /etc/nslcd.conf /tmp/tmp_nslcd.conf
   ```

2. Remove the following lines from the end of the temporary file:

   ```
   # Workaround to allow 2 characters user names
   validnames /^[a-z0-9._@$()][a-z0-9._@$() \~-]*[a-z0-9._@$()~-]$/i
   ```

3. Update the registry with this new file:

   ```
   fileContent="$(cat /tmp/tmp_nslcd.conf; printf x)"
   net registry setvalue HKLM/SOFTWARE/IBM/SOFS/AUTH NSLCD_CONF sz
   "${fileContent%x}"
   ```

4. Reload cnscmc on all nodes:

   ```
   onnode all cnscmc reload auth
   onnode all service nslcd stop
   ```

5. Remove the temporary file:

   ```
   rm -f /tmp/tmp_nslcd.conf
   ```

Important: If breakage is observed, you might need to remove the manual configurations.

Upgrade considerations: The upgrades for these manual configurations do not require any changes before or after upgrade. However, if the manual changes are lost after upgrade, repeat the steps for manual configuration.

Important: Users do not lose access to data because ACLs are stored for UID and GID, which remain intact.

Preferred practice to be followed

It is essential that user and group names are more than three characters long. Access is denied if any user or group has a name with fewer characters than that.

2.7 Preferred practices for Network Information Service

NIS is used in UNIX-based environments for centralized user and other services management. NIS is used for keeping user, domain, and netgroup information. Customers with UNIX-based environments use NIS for user management and host name management so all systems have the same user information. In SONAS, NIS is used for netgroup support and ID mapping.

This section explains some of the important points for configuring SONAS with NIS.

2.7.1 Prerequisites for configuring the SONAS system with AD and NIS

The following prerequisites apply to configuring SONAS with Active Directory and NIS:

- No data files have been stored on the SONAS system.
- AD authentication is configured on the SONAS system by using the **`cfgad`** SONAS CLI command.
- NIS is used for all User ID Mapping.
- All NIS user and group names must be in lowercase, with white space replaced by the underscore character (_). For example, an Active Directory user name `CAPITAL Name` should have a corresponding name on NIS as `capital_name`.

2.7.2 Limitations of configuring the SONAS system with AD and NIS

The following limitations apply to configuring SONAS with Active Directory and NIS:

- The low value of `idmapUserRange` and `idmapGroupRange` cannot be less than 1024.
- UNIX-style names do not allow spaces in the name. For mapping Active Directory users or group to NIS users, consider the following conversion on the NIS server:
  - Convert all uppercase characters to lowercase.
  - Replace all blank spaces with underscores.

    For example, an Active Directory user or group name `CAPITAL Name` should have a corresponding name on NIS as `capital_name`.

Important: If the preceding naming convention is not used for users, NIS mapping does not happen, and ID mapping for such users follows the user map rules that are defined in the **`cfgnis --userMap`** option for that AD domain.

2.7.3 Revisiting key parameters for configuring NIS

This section reviews key parameters for configuring NIS.

- **`--extend { extend }`**

  In extended mode, NIS can also be used as an ID mapping mechanism. ID mapping functionality can be activated using the **`--useAsIdmap`** option.

- **`--idmapUserRange { idmapUserRange }`**

  This parameter sets the user ID range. The UIDs of the users from the Active Directory domain that have a user map rule set to AUTO are assigned from this range. Samba will also use this range to map some of the well-known SIDs to UIDs. This option is mandatory when used with the **`--extend`** and **`--useAsIdmap`** options.

  Use the format *`<lowerID of the range>`* - *`<higherID of the range>`*. *`LowerID`* cannot be less than 1024:

  ```
  --idmapUserRange 100000-200000
  ```

- **`--idmapGroupRange { idmapGroupRange }`**

  This parameter sets the group ID range. The GIDs of all the Active Directory 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 used with the **`--extend`** and **`--useAsIdmap`** options.

  Use the format *`<lowerID of the range> - <higherID of the range>`*. *`LowerID`* cannot be less than 1024:

  ```
  --idmapGroupRange 100000-200000
  ```

Note: Absence of the **`--extend`** option indicates that NIS is to be configured in basic mode. Basic configuration supports NIS authentication for the NFS netgroup. It does not include any other protocol configuration, and does not support the ID mapping mechanism.

There should be a single mapping entry for a specified server. Having multiple entries for the same NIS server triggers an error.

`LowerID` cannot be less than 1024.

2.7.4 Preferred practices when you configure NIS

This section provides several preferred practices for configuring NIS.

Configuring NIS correctly before data is created on SONAS

When you configure authentication for the first time, it is essential that you do it immediately before creating or migrating data on SONAS.

There is no way to correct any misconfiguration, except by cleaning up the authentication that is configured.

User and group range values must be greater than 1024

This is both a preferred practice and a limitation. Users and groups range values must be greater than 1024.

User and group names must be in lowercase

This is both a preferred practice and a limitation. Users and groups must have names in lowercase letter.

Replace space or blank space in user and group names with an underscore (_)

This is both a preferred practice and a limitation. If there is a space or a blank space in a user or group name, it must be replaced with an underscore (_).

Establishing correct user mapping for Active Directory users who do not already have a user ID mapping in NIS

Use one of the following three options for the **`--userMap`** option in the **`cfgnis`** command to select the correct mapping for users who do not already have a user ID mapping in NIS:

- `AD_domain1:DENY_ACCESS`

  If a user from this domain does not have a mapping entry in NIS, the user is denied access. The Active Directory domain that has no user rule defined defaults to `DENY_ACCESS`.

- `AD_domain1:AUTO`

  If a user from this domain does not have a mapping entry in the NIS, a new ID is generated for the user. This ID is generated from the **`--idmapUserRange`** option and is auto-incremented. The administrator must ensure that existing NIS IDs do not fall in this ID range. This mapping is maintained in the system. NIS is not aware of this ID mapping.

- `AD_domain1:DEFAULT:ad_domain\\guest_user`

  If a user cannot be mapped in NIS, the user is mapped to a single guest user in Active Directory or Samba Primary Directory Controller (PDC), who must have a mapping entry in NIS.

  This requires that the `guest user` as specified in the **`--map`** instruction is mapped in NIS.

There can be only one rule out of the three listed options for the **`cfgnis`** command per Active Directory domain. Making multiple rules for the same Active Directory domain causes an error.

This is an optional parameter and the default setting is to deny all the users who do not have a NIS mapping.

This option can be used only when both the **--extend** and **--useAsIdmap** options are specified.

> **Important:** The **cfgnis** command restarts the file services (FTP, SCP, HTTP, NFS, and CIFS), which is disruptive for connected clients. The connected clients lose their connection and the file operations are interrupted. File services will resume few seconds after the command finishes successfully.
>
> When NIS service is configured against the SONAS system, rerunning the **cfgnis** command overwrites the earlier configuration with the new one.

Preferred practice to be followed

The preferred practice is to decide what options to use before installation. Use the default setting and repair failed user attempts with the specified NIS ID mapping.

Issue: AD+NIS successfully configured, yet some users are denied access - AD user does not have a mapping entry in NIS

Even after AD+NIS is successfully configured, some AD users are denied access to the IBM SONAS share.

How to debug this issue

Complete the following steps to debug user access:

1. If after sufficient ACLs are applied, data is inaccessible, check whether the UID for that user is set. Use the following command to check whether the user or group has a UID or GID assigned:

   ```
   #chkauth -i -u "<DOMAIN\\username>"
   ```

2. Run the **chkauth** command. Example 2-15 shows that the **chkauth** command does not show the UID or GID for autouser3. If the **chkauth** command does not show a UID or GID (even though SONAS cluster can resolve the users on the domain controller), even though the user is present in the AD server, a corresponding user might not be present in the NIS server.

   Example 2-15 Failed to get IDMapping for user

   ```
   # chkauth -i -u "SONAS\\autouser3"
   EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
   ```

Conclusion

Access is denied for users present only in the AD server but not present in NIS server.

How to correct this issue

There are multiple ways to resolve this issue:

- Define a corresponding user in the NIS server. See your NIS server documentation for creating new users.
- As per requirements, select the wanted option for the **--userMap** option while configuring authentication by using the **cfgnis** command. The default action for such users is to deny access. You can select the `AD_domain:AUTO` option in **--userMap** to let SONAS automatically generate a UID. Alternatively, use the `AD_domain:DEFAULT` option in **--userMap** to map such users to a predefined guest user in AD server. For the exact syntax and examples for **--userMap**, see the man page for the **cfgnis** command.

Preferred practice to follow

The preferred practice is to select options in advance. Configure the correct options for users who are present in AD without a corresponding user in NIS server while configuring authentication and before storing any data or accessing the cluster.

Issue: AD+NIS successfully configured, yet some users are denied access and the AD user has a mapping entry in NIS

Even after AD+NIS is successfully configured some windows users are denied access to the IBM SONAS share even though there is a corresponding user present in the NIS server.

How to debug this issue

Complete the following steps to debug user access:

1. If after sufficient ACLs, data is inaccessible, check if the UID for that user is set. Use the following command to check whether the user or group has a UID or GID assigned:

   ```
   #chkauth -i -u "<DOMAIN\\username>"
   ```

2. Example 2-16 shows that the **`chkauth`** command does not show the UID or GID for `AUTO user4`. If the **`chkauth`** command does not show UID or GID even though the SONAS cluster can resolve the users on the domain controller, the user might be present in the AD server without a corresponding user present in the NIS server.

   Example 2-16 Failed to get ID mapping for user

   ```
   chkauth -i -u "SONAS\\AUTO user4"
   EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
   ```

Conclusion

UNIX-style names do not allow spaces in the name. For mapping Active Directory users or groups to NIS users, consider the following conversion on the NIS server:

- Convert all uppercase alphabets to lowercase.
- Replace all blank spaces with underscores.

How to correct this issue

Complete the following steps to correct this issue:

1. Define the corresponding user in NIS server as for the UNIX-style names.
2. After the corresponding user is defined in the NIS server as for the UNIX-style names (in this case it should be `auto_user4`) you can use the **`chkauth`** command as shown in Example 2-17 to verify it.

   Example 2-17 Successfully displays ID mapping for a user

   ```
   # chkauth -i -u "SONAS\\AUTO user4"
   Command_Output_Data       UID   GID  Home_Directory           Template_Shell
   FETCH USER INFO SUCCEEDED 21015 2100 /var/opt/IBM/sofs/scproot /usr/bin/rssh
   EFSSG1000I The command completed successfully.
   ```

   If the default setting, which is to deny access, was used in the **`--userMap`** parameter of the **`cfgnis`** command, access is denied for such users. However, if the --userMap option chosen was either `AD_domain:AUTO` or `AD_domain:DEFAULT`, such unmapped users are either provided auto incremented UIDs by SONAS or are mapped to predefined guest users that are defined in AD. But then in such cases, even though the user might be able to access the shares, it might not have access to the data as expected.

Preferred practice to be followed

The preferred practice is to define all the users in NIS server as for the UNIX style names for all the AD users with uppercase letters, space in their names, or both. This renaming must be done before storing any data or accessing the cluster.

2.8 Preferred practices for local authentication

For version 1.5.1 and later, SONAS supports configuring authentication by using a local authentication server that is hosted internally by SONAS itself. For a detailed description of local authentication, see the Authentication chapter in *IBM SONAS Implementation Guide*, SG24-7962.

Local authentication supports both Windows and UNIX clients with some limitations that are described later in this section. Therefore, for mixed environments with both Windows and UNIX clients, this method 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 UIDs and GIDs on a first-come first-served incremental basis as the default setting.

2.8.1 Limitations for configuring SONAS with local authentication

The following limitations apply to configuring SONAS with local authentication:

- There is no support for migration of existing authentication server user and group data to local authentication server.
- There is no support for migration of user and group data from local authentication server to external authentication server.
- NAS user and group names are case-sensitive.
- NAS user and group names cannot collide with the CLI and system users.
- 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.
- NFS netgroups are not supported.
- There are no local data access user password policies (except 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.
- A maximum of 1000 users and 100 groups are supported. A user can belong to only 16 groups. A group can consist of 1000 users.
- Asynchronous replication is not supported.

2.8.2 Preferred practices for configuring SONAS with local authentication

This section describes preferred practices for configuring SONAS with local authentication.

Consistent user and group identity across multiple SONAS systems

It is preferred to make user and group names consistent across multiple SONAS systems in a customer environment. SONAS administrators must ensure this consistency of user and group identities across multiple systems. This configuration simplifies migrating authentication to an external LDAP server.

Unique user and groups IDs across systems

Do not reuse UIDs and GIDs across different SONAS systems. When they manage multiple SONAS systems, administrators should designate a primary SONAS system that will have all the users and groups. Create all new users and groups first on the primary SONAS system and then on the other SONAS system by reusing the same UID and GID as on the primary system. This method ensures that UIDs and GIDs are consistent across SONAS systems even if not all users connect to each SONAS system.

Always specify the UID and GID to create local users and groups

Always use a UID and GID you create users and groups with the CLI commands, rather than letting SONAS auto generate the UID and GID. This method provides increased control and administrative awareness to ensure that no security authorization issues are introduced over many years of use of the system. This method also helps in maintaining consistent user and group names and corresponding UIDs and GIDs across multiple SONAS systems.

Setting up Linux and UNIX systems for NFS access

To avoid UID or GID conflicts, ensure that user and group identities on the host system are consistent with those on IBM SONAS systems. UID/GID can be displayed on the host by using the UNIX **id** command.

Issue: Local authentication successfully configured but some NFS users are unable to access data.

Even after local authentication was successfully configured, some NFS users are not able to access data.

An NFS export that is mounted on a Linux host by the user `autouser1` is not able create files and gets a permission denied error message, as shown in Example 2-18.

Example 2-18 Permission denied message

```
[autouser1@gssvm01 testex]$ touch testfile
touch: cannot touch `testfile': Permission denied
```

How to debug this issue

Complete the following steps to debug user access:

1. If data is inaccessible, the first thing you need to check is whether the user that is trying to access the data is successfully resolved by the system and has sufficient ACLs.
2. If ACLs are sufficient, and data is still not accessible, check if the UID and GID for that user is set. Use the following command to check if the user or group has a UID or GID assigned as expected:

   ```
   #chkauth -i -u autouser1
   ```

As shown in Example 2-19, the UID and GID of the user autouser1 are set as expected by the SONAS administrator.

Example 2-19 The UID and GID are set correctly

```
[st001.virtual1.com]$ chkauth -i -u autouser1
Command_Output_Data       UID  GID  Home_Directory           Template_Shell
FETCH USER INFO SUCCEEDED 2000 2000 /var/opt/IBM/sofs/scproot /usr/bin/rssh
EFSSG1000I The command completed successfully.
```

3. Now on the NFS host, check whether the UID and GID of autouser1 set to what they are on the SONAS cluster by running the **id** command, as shown in Example 2-20.

Example 2-20 Verifying the UID and GUID of the autouser1 user

```
[autouser1@gssvm01 testex]$ id autouser1
uid=2001(autouser1) gid=2001(testgrp) groups=2001(testgrp)
```

As shown in Example 2-20, the UID of user autouser1 on the Linux host does not match that on the SONAS cluster.

Conclusion

Access is denied to user `autouser1` because the UID does not match to that of the UID on the SONAS cluster.

How to correct this issue

Re-create the local data access user or group on the SONAS cluster to match that of the users and groups on the NFS clients.

Preferred practice to follow

Ensure that UIDs and GIDs for the NFS client the UIDs and GIDs for the local data access users and groups on the SONAS cluster.

2.9 Common authentication issues

This section lists general authentication issues that can occur in SONAS.

2.9.1 Issue: No services are running after successful configuration of SONAS

In this case, SONAS is installed and configured successfully. However, none of the services, such as CIFS, NFS, or FTP, are running. Clients are not able to access data.

How to debug this issue

Complete the following steps to debug user access:

1. If none of the services are running, check whether authentication has been configured. Run the **lsauth** CLI command to check, as shown in Example 2-21.

Example 2-21 Checking whether authentication has been configured

```
$lsauth -c st001.virtual1.com
EFSSG0571I Cluster st001.virtual1.com is not configured with any type of
authentication server(ldap/ldap_krb/nt4/ad).
```

The result of running the **lsauth** command, as shown in Example 2-21 on page 66, shows that no authentication is configured.

2. In SONAS, the authentication commands start the services. Because no authentication is configured, none of the services started.

Conclusion

For services to be started and configured on SONAS, authentication must be configured after successfully installing and configuring SONAS. SONAS is ready to use only after that.

2.10 Troubleshooting authentication issues

This section shows some methods to debug issues. If your issue is not similar to any of the issues discussed in this chapter, use one of these commands to check what might be going wrong.

2.10.1 CLI commands to check

You can run the following commands to check whether output is as expected or something is wrong.

List the authentication configured

Check the authentication that is configured by using the **lsauth** command. This command provides information about the configuration. You can check for the different parameters if they are set correctly, depending on the authentication that is configured. Example 2-22 shows an example for AD authentication. You can check parameters similarly for all other authentication methods.

Example 2-22 Checking the authentication that is configured on the cluster

```
# lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
domain = SONAS
idMappingMethod = auto
clusterName = bhandar
userName = Administrator
adHost = SONAS-PUNE.SONAS.COM
passwordServer = *
realm = SONAS.COM
EFSSG1000I The command completed successfully.
```

Check the ID Mapping for users and groups

Run the **chkauth** command to check the user details such as the UID and GID for a user or group, as shown in Example 2-23.

Example 2-23 Check user information by using the chkauth command

```
# chkauth -c st002.vsofs1.com -i -u VSOFS1\\testsfuuser2 -p Dcw2k3dom01
Command_Output_Data     UID GID      Home_Directory           Template_Shell
FETCH USER INFO SUCCEED 250 10000011 /var/opt/IBM/sofs/scproot /usr/bin/rssh
```

Check for node synchronization

Run the **chkauth** command to check whether the nodes are in synchronization, as shown in Example 2-24.

Example 2-24 Check node synchronization by using the chkauth command

```
# chkauth
ALL NODES IN CLUSTER ARE IN SYNC WITH EACH OTHER
EFSSG1000I The command completed successfully.
```

Check whether the user can authenticate successfully

Run the **chkauth** command to check whether the user is able to authenticate with the authentication server, as shown in Example 2-25.

Example 2-25 Check if user can authenticate with server using ckhauth

```
# chkauth -c st002.vsofs1.com -a -u VSOFS1\\testsfuuser2 -p Dcw2k3dom01
Command_Output_Data       UID GID Home_Directory Template_Shell
       AUTHENTICATE USER SUCCEED
```

Check whether the authentication server is reachable

Run the **chkauth** command to check whether the authentication server is reachable, as shown in Example 2-26.

Example 2-26 Check whether authenticate server is reachable by using the ckhauth command

```
# chkauth -c st002.vsofs1.com -p
Command_Output_Data                UID GID Home_Directory Template_Shell
       PING AUTHENTICATION SERVER SUCCEED
```

2.10.2 Logs to check

The logs contain useful information to help resolve errors and determine what has happened in the SONAS.

System logs

Check the system logs to see whether there are any errors. The **lslog** CLI command displays the system log.

Audit logs

To check what commands recently ran and the command parameters, you can run the **lsaudit** CLI command. This shows all of the commands that were run. You might want to see the sequence of commands run, see if any of them were incorrect, and so on.

2.10.3 More logs to collect and check

If the logs in 2.10.2, “Logs to check” do not help, you can contact IBM Support. It is advisable to collect the following logs, which can help for further analysis or debugging:

- Samba Debug 10 Logs

  > **Important:** Read the man pages for the **starttrace** and **stoptrace** commands before you use them.

  Run the following commands to collect the Samba Logs:

  ```
  starttrace –cifs –client <client ip address>
  Recreate issue on a Windows Client.
  stoptrace #traceid
  ```

- UID and GID information for the user

  Along with the preceding logs, also collect UID and GID information for the users you see problems for. You can run the **chkauth** command to get the information:

  ```
  # chkauth -i -u <Username>
  ```

- Run the **cndump** command.

3

Networking

This chapter provides preferred practices for IBM Scale Out Network Attached Storage (SONAS) network configuration:

- Terminology
- SONAS networking
- Network bonding overview
- SONAS networking preferred practices
- SONAS routing

3.1 Terminology

Many terms are used by networking architects and specialists to describe networking. Linux and networking switch manufacturers often use different terms for the same concepts. The next sections list some SONAS-related networking terms and explain bonding modes.

3.1.1 Networking terms

This section provides terms and abbreviations that are common and relevant to network configurations in general. The purpose is to make you familiar with the terms and abbreviations that are used throughout this chapter:

Address Resolution Protocol (ARP)

ARP is a protocol that is used for resolution of network layer addresses (example Internet Protocol (IP) addresses) into link layer addresses (message authentication code (MAC) addresses). It is also the name of the program for manipulating these addresses in most operating systems.

Bonding A method for grouping several physical network ports into a single virtual adapter for one or more of the following reasons: Load sharing, throughput enhancement, redundancy, and high availability (HA). The term *link aggregation* is typically used for the same concept on network equipment.

Default route A routing table entry that is used to direct frames for which a next hop is not explicitly listed in the table.

Hop/hop count A hop in routed networks is the passage of a network packet through a router (or similar) on the way to its destination. Hop count refers to the number of hops of routers that a packet must pass through on its way to its destination, and is an estimate of distance in a network.

Jumbo frames Jumbo frames are Ethernet frames with more than 1500 bytes of payload. Conventionally, jumbo frames can carry up to 9000 bytes of payload as the maximum transmission unit (MTU), but variations exist and care must be taken when you are using the term.

Many Gigabit Ethernet (GbE) switches and Gigabit Ethernet network interface cards (NICs) support jumbo frames. Ethernet jumbo frames improve performance in some environments by enabling by moving more data with less required resources.

Management network

The SONAS network carries the configuration, health monitoring, and other low-bandwidth messaging among the nodes. Both management-network Ethernet adapters in each node are bonded to share a single IP address.

Netgroup A network-wide group of hosts and users. A netgroup can be used to restrict access to shared information on Network File Systems (NFS). SONAS supports netgroups only for grouping hosts to restrict access to NFS.

Network address translation (NAT)

NAT is a technique that is typically used with network routers. The benefit of the NAT gateway within SONAS is that it provides an internal network failover path by which local SONAS functions on a given management node or interface node can still access public services such as domain name server (DNS) and Lightweight Directory Access Protocol (LDAP).

Ping and round trip time (RTT)

Ping is a network diagnostic tool that sends an Internet Control Message Protocol (ICMP) Echo Request to a distant node. The remote node must immediately return an ICMP Echo Reply packet back to the originating node. The RTT reflects time between the transmission of a packet and the receipt of its acknowledgment, and can be used to determine the latency between two systems.

Virtual local area network (VLAN)

VLAN is a software-defined LAN that groups network elements in the same broadcast domain.

xmit_hash_policy With certain bonding protocols, when transmitting using multiple interfaces, a policy or algorithm must exist to determine which interface to use to attempt balance the network load.

3.1.2 Bonding modes

SONAS supports the following bonding modes. The simple numbers 0 - 6 are how Linux and SONAS refer to the bonding modes in the configuration file. Network equipment manufacturers and network support personnel generally use the term that is listed in parenthesis:

0 (balance-rr) Transmits packets in a sequential order from the first available subordinate through the last. Suppose that two real interfaces are subordinates in the bond, and two packets arrive destined out of the bonded interface.

The first is transmitted on the first subordinate and the second frame is transmitted on the second subordinate. The third packet is sent on the first, and so on. This technique provides load balancing and fault tolerance.

Note: Mode 0 is not recommended for general NAS protocols that heavily use Transmission Control Protocol/Internet Protocol (TCP/IP) versus User Datagram Protocol (UDP) because it can trigger congestion control actions (retransmission due to out of order packets) resulting in loss of intended performance gains. Generally bonding mode 2 or mode 4 are preferred over this mode when link aggregation performance and high availability are needed.

1 (active-backup) An active backup configuration. Only one subordinate in the bond configuration is active at a time. Other subordinates become inactive until the active, primary subordinate fails. To avoid switch confusion, the MAC address is externally visible only on one port.

This mode provides fault tolerance. Currently, 10 Gigabit converged network adapters (CNAs) in Interface nodes for external data connectivity are configured to handle IP over InfiniBand in this mode.

Moreover, all internal management NICs and internal data InfiniBand host channel adapters (HCAs) are configured in SONAS by default in this active backup configuration. Therefore, all internal SONAS networks share a single IP address, and work in hot standby configuration.

2 (balance-xor) Transmits based on an *exclusive or* (XOR) formula. The Source MAC address is XOR'd with destination MAC address Modulo subordinate count. It selects the same subordinate for each destination MAC address, and provides load balancing and fault tolerance.

4 (802.3ad) Also known as dynamic link aggregation based on the Link Aggregation Control Protocol (LACP). It creates aggregation groups that share speed and duplex settings. This mode requires a switch that supports Institute of Electrical and Electronics Engineers (IEEE) 802.3ad.

Mode 4: IEEE 802.3ad Dynamic link aggregation:

- Create aggregation groups that share speed and duplex settings.
- Here are the requirements for creating an 802.3ad link aggregate (bond):
  - All physical ports in the aggregate must have the same duplex setting and speed.
  - All physical ports in the aggregate must connect to the same Ethernet switch.
  - The Ethernet switch to which all ports in the aggregate connect must support 802.3ad LACP.

5 (balance-tlb) Adaptive transmit load balancing. The outgoing traffic is distributed according to the current load and queue on each subordinate interface. Incoming traffic is received by the current subordinate. Does not require any special switch support.

6 (balance-alb) Adaptive load balancing. The outgoing traffic is redistributed between all subordinates that are working in bond configuration according to the current load on each subordinate. The receive load balancing is achieved through ARP negotiation.

The receive load is redistributed by using a round robin algorithm among the group of subordinates in the bond. Effectively, this configuration combines bandwidth into a single connection, so it provides fault tolerance and load balancing. Adaptive low balancing does not require any special switch support.

3.2 SONAS networking

The SONAS hardware has several physical network ports for client communication, and for management tasks. The number of ports varies based on the SONAS configuration. Ports are used for management, backup, replication, and for data transfer from the clients. Each physical port is assigned to a bond. The default bonding mode of SONAS is mode 1 or active/backup. However, this setup can be changed in specific cases as needed to fulfill the wanted behavior in terms of performance and redundancy.

3.2.1 Interface node networking

Public interfaces are bonded redundantly into network groups of interface nodes that work together to share the load. Bonds for each node within a network group must have identical interface names (ethX0, ethX1, and so on). This configuration is required for proper node failover, where network IP addresses and supporting interfaces are moved from node to node within the network group due to a failover event. For example, if a network group is attached to ethX0, all nodes within that network group must support the ethX0 bond interface for managing the IP addresses that are assigned to the shared network.

When creating network groups, configuring uniform node characteristics for all nodes within a network group maximizes node behavior consistency on node failover. If node interfaces are not configured consistently within a network group, traffic routing becomes unpredictable. This can occur when a mixture of 1 GbE and 10 GbE speeds exist on assigned nodes within a network group.

If a node with only 1 GbE of network bandwidth capability for an interface assumes responsibility during failover from a node that had a 10 GbE network speed for the identical interface, reduced client performance and unwanted behavior can occur as 10 GbE incoming traffic is replied to out of 1 GbE port channels. Other node characteristic consistencies that maximize node failover behavior include bonding mode, MTU size, virtual local area network (VLAN) assignment, and `xmit_hash_policy`.

3.2.2 Understanding your network topology

Scale out NAS solutions are designed to simplify data access for a large contingency of your business, be it analytical or scientific computing, or simple home directories. A scale out solution by nature suggests that you plan to grow.

NAS is not a simplified block solution with isolated paths from application to storage or limited protocol and congestion to consider. NAS solutions are end-to-end network intensive protocol rich environments with many communication points of consideration in the path between the client demand and the service delivery.

Under-estimating the importance of the enterprise solution networked component interaction, bandwidth, latencies, and redundancy causes problems for the long-term strategy. It is important to pay close attention to these aspects regularly and in all troubleshooting events:

- First, consider flattening the network (mapping everything important and taking out the complexity).
- For any SONAS solution, ensure that an up-to-date network topology diagram is available to SONAS administration staff.
- Network switches should be well-managed and kept up to date. Network and client communication devices and firmware should be compatible with current switch manufacturer recommendations.
- Ensure that RTT distance and network hops between SONAS and client and authentication servers, antivirus servers, DNS, Network Time Protocol (NTP), backup and Hierarchical Storage Manager (HSM) primaries and secondaries, is understood at installation time. These should be regularly checked and validated. In some cases, high RTT times to such things as authentication servers can cause extraordinary latencies in client service response times.
- In some cases, it might be advantageous to build new service secondaries closer to your SONAS clusters (from a network perspective) to improve response and overall performance.

- Measure and record your RTT times to all your key solutions and key clients in different network segments of your environment. Ping can be a simple tool for testing and capturing those key points in your environment.
- When a disaster recovery (DR) site is assembled, it is equally important to ensure that RTT times to secondary servers are also quick to ensure the expectation of equal service response if there is a disaster. This is often overlooked in a complete solution, and never appreciated until disaster strikes (when it might be too late to fix the network quickly).
- Establish a troubleshooting protocol with your network team (in advance) to understand how to quickly collect, capture, record, share, and evaluate network diagnostic information if there is a service complaint. By establishing this protocol early in your deployment, you simplify time to resolution and greatly decrease administrative volley and service-related anxiety.

3.3 Network bonding overview

This section provides an overview of network bonding. It describes configuring the network for maximum throughput and availability, and when maximum throughput and availability can be obtained.

3.3.1 Link Aggregation Control Protocol

The preferred practice for networking for performance and reliability is to bond 2 or 4 ports by using Mode 4 (LACP or 802.3ad) bonds with the `xmit_hash_policy` policy set to Layer3+4. The switch ports to which the bond is connected to should be grouped into an 802.3ad compatible channel group. This grouping might be referred to as *dynamic link aggregation* or LACP, which is the control protocol that is used by 802.3ad in switch configuration options.

The precise method for these configurations varies by switch manufacturer and model. A balance algorithm should also be chosen for outgoing traffic on both the bond and the switch. The algorithms that are available on the switch and the method to select them vary by switch.

Static versus dynamic: Static link aggregation is the generic term that was originally referred to by Cisco as Etherchannel. As the name implies, *static* link aggregation is different from 802.3ad *dynamic* link aggregation.

3.3.2 Transmission policies

When network interfaces are bonded together, the switch and the server need to have algorithms to balance the load across the adapters. The balancing algorithm options are described in the `xmit_hash_policy` option of the bonding kernel documentation. The `xmit_hash_policy` option selects the transmit hash policy to use for output interface selection in balance-xor and 802.3ad modes. The default value of the `xmit_hash_policy` option is layer2. The other two option values are layer2+3 and layer3+4.

For LACP bonding, the preferred practice is to use `xmit_hash_policy` with layer3+4 option.

Important: The description of layer3+4 in the kernel documentation contains the following note: *This algorithm is not fully 802.3ad compliant. A single TCP or UDP conversation that contains both fragmented and unfragmented packets sees packets striped across two interfaces. This configuration might result in out-of-order delivery. Most traffic types do not meet this criteria, as TCP rarely fragments traffic, and most UDP traffic is not involved in extended conversations. Other implementations of 802.3ad might or might not tolerate this noncompliance.*

The layer3+4 note about 802.3ad noncompliance has not been a problem in practice. Strongly consider using the layer3+4 value for ports that use bonding modes 2 and 4, because it provides the best distribution of traffic for most use cases. The layer2+3 value is preferred only if there is a significant amount of fragmented IP traffic, such as with NFS over UDP as one example, and the dispersion of fragments causes performance degradation or reassembly failures, or the out-of-order delivery causes 802.3ad noncompliance or other issues.

3.3.3 Jumbo Frames

For improved performance, set up jumbo frames.

Jumbo frames: SONAS supports jumbo frames. However, use jumbo frames only if the end-to-end infrastructure can also support it. Setting jumbo frames without end-to-end support might lead to network instability and inconsistent performance.

If you do not specify the proper MTU setting on your network infrastructure, you will produce performance irregularity. Those problems present themselves in the following scenario:

- A client transmits a 9000-byte payload frame. The frame size that is transmitted to your network in a non-VLAN trunked transmission is 9018 bytes.
- If your network is not configured to accept a 9018-byte frame, it discards the transmission.
- If MTU discovery is enabled, the systems negotiate the payload size of each future frame transmission to a smaller value.
- If MTU discovery is not possible, the transmission fails.
- The process of negotiating a smaller payload increases latency and system processor use as the system is forced to continually negotiate smaller payload sizes.

Figure 3-1 illustrates the configuration requirements for a network infrastructure that has enabled jumbo frames for the clients on Bond1. The clients on Bond0 are using a smaller MTU. The routing on the SONAS server must be configured correctly, and the clients on the network must be segregated by MTU size.

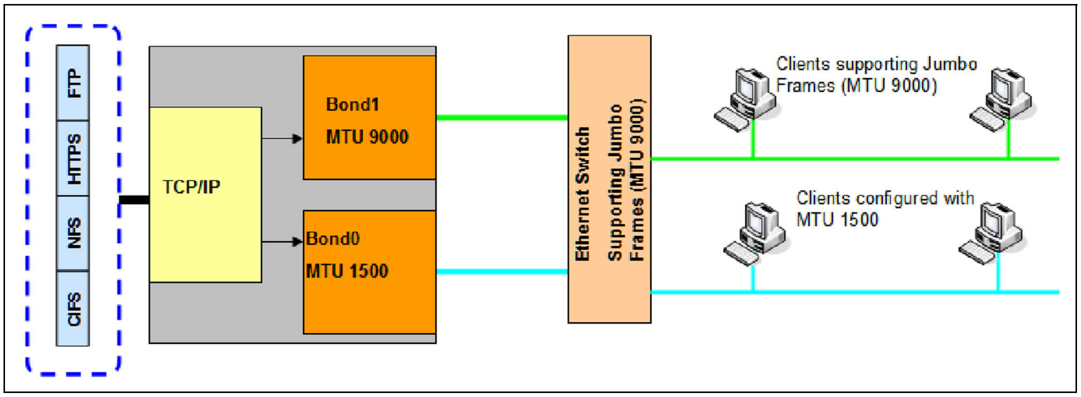


Figure 3-1 Jumbo frames configuration example

.

Remember: All VLAN networks that are created and attached on a jumbo frame MTU bond interface inherit the MTU size of the base bond. If you have a base bond, for example ethX0, a VLAN that you attach on that interface inherits the jumbo frame MTU setting. This setting requires VLANs to have the same MTU sizes as their parent bond interface. You cannot define separate VLAN MTU sizes.

Important: SONAS supports only MTU 1500 and MTU 9000 for 10 Gbit cards.

3.3.4 Configuring a system for high availability and throughput

Depending on your network switches, SONAS systems can be configured for high availability or throughput. Ideally both can be obtained together.

By default, the `mknwbond` command uses mode 1 (active-backup). This setting delivers high availability, but no increase in throughput. If your network switch supports LACP, use bonding mode 4. Otherwise, use mode 6.

The following sections describe bonding in more detail. The examples are similar for 1 GbE and 10 GbE, so in most cases the speed of the adapters is not described.

Single and multi-switch topology that uses bond mode 4

The 802.3ad mode requires that the switch have the appropriate ports configured as an 802.3ad aggregation. The precise method that is used to configure this setting varies from switch to switch, but, for example, a Cisco 3550 series switch requires that the appropriate ports first be grouped in a single Etherchannel instance, then that Etherchannel is set to mode LACP to enable 802.3ad (rather than standard Etherchannel).

In this example, these two ports are grouped to one single bond. In a single-switch topology, both ports are connected to the same switch, as shown in Figure 3-2. The switch ports to which the bond is connected to should be grouped into an 802.3ad compatible channel group.

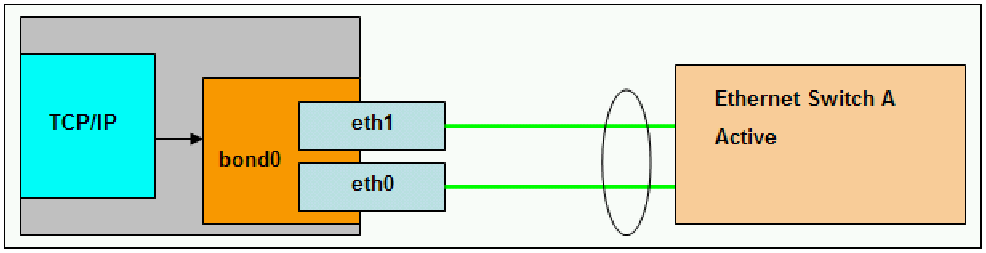


Figure 3-2 Single-switch layout example

Bond mode 6 (balance-tlb) with two ports that are connected to a single switch also looks like Figure 3-2.

Figure 3-3 shows a configuration with four network ports in a system that are all bonded together and attached to a single switch. This configuration does not make sense unless your network switch supports a bonding mode where you can have multiple ports active at the same time.

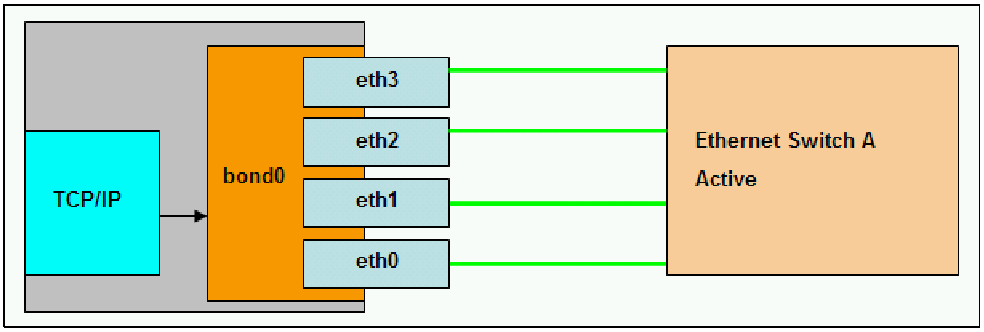


Figure 3-3 Example single-switch configuration layout

Multi-switch topology

Figure 3-4 shows two ports that are grouped to one single bond. In a multi-switch topology, ports are connected to different switches. The network switches need to be configured to support this configuration.

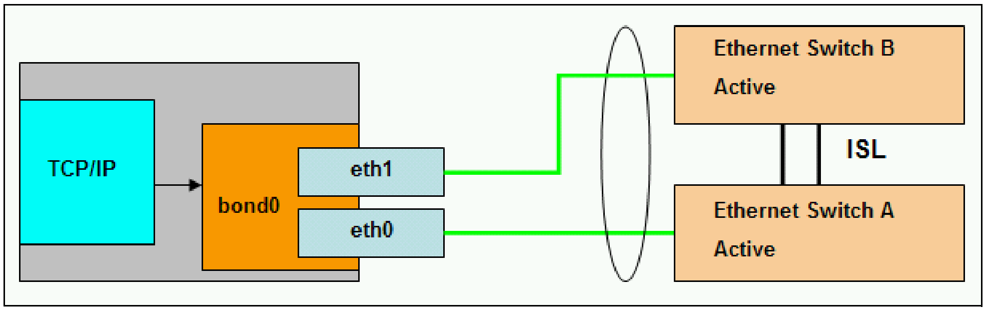


Figure 3-4 Multi-switch layout example

High availability considerations: When client I/O paths are moved from one SONAS node to another SONAS node, the failover operation issues gratuitous ARPs to inform the network of the MAC address changes for the IP addresses that the client is using on the SONAS interface nodes. Gratuitous ARPs to the network generally reach all of the equipment that directs the client traffic to the SONAS interface node. If this does not occur, you might consider changing timeout values for the ARP caches to values that are more appropriate for the particular network's traffic and the delay latency that SONAS I/O clients can accept. A timeout value of 5 to 10 minutes can accommodate most SONAS node failover operations.

Figure 3-5 shows an example multi-switch topology layout with four ports on the server.

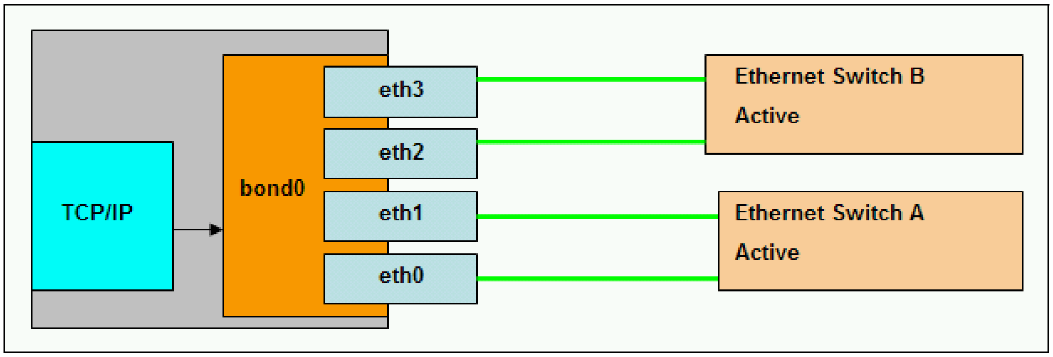


Figure 3-5 Example multi-switch topology layout

Mode 4 bonding example with two ports

The following section shows the commands that are used to create a mode 4 bond with the layer3+4 transmission hash.

Configuration requirements

The following prerequisites must be fulfilled:

- No network is assigned to the subordinates.
- No IP address is active on any of the subordinates.

Configuration Steps

Complete the following steps to create a multi-switch layout:

1. Detach the network from the interface and check the configuration. The commands are shown in Example 3-1.

Example 3-1 Detach the network from the interface and check the configuration

```
$ detachnw ethX0 10.0.0.0/24 -g int
Removing network NAT gateway...
EFSSG0087I NAT gateway successfully removed.
EFSSG0015I Refreshing data..

$lsnwinterface
Node          Interface MAC               Master/Subordinate Bonding mode
Transmit hash policy Up/Down Speed IP-Addresses MTU
int001st001  ethX0     02:1c:d1:02:03:00 MASTER active-backup (1) UP 10000 1500
mgmt001st001 ethX0     02:1c:d1:00:03:00 MASTER active-backup (1) UP 10000 1500
mgmt002st001 ethX0     02:1c:d1:01:03:00 MASTER active-backup (1) UP 10000 1500
```

2. Remove the existing bonds on all the servers, as shown in Example 3-2.

Example 3-2 removing existing bonds

```
$ rmnwbond int001st001 ethX0
EFSSG1000I The command completed successfully.

$ rmnwbond mgmt001st001 ethX0
EFSSG1000I The command completed successfully.

$ rmnwbond mgmt002st001ethX0
EFSSG1000I The command completed successfully.
```

3. Check the status of the network with the **lswinterface** command, as shown in Example 3-3.

> **Note:** The output in this section is abbreviated for clarity. Also note that typically, when bonded, the MAC address of all the subordinates and the bond is the same.

Example 3-3 Output of the lswinterface command

```
$ lsnwinterface -x
Node Interface MAC Master/Subordinate Bonding mode Transmit hash policy Up/Down
Speed IP-Addresses MTU
int001st001  eth0      02:1c:d1:02:03:01 UP      10000               1500
int001st001  eth1      02:1c:d1:02:03:02 UP      10000               1500
mgmt001st001 eth0      02:1c:d1:00:03:01 UP      10000               1500
mgmt001st001 eth1      02:1c:d1:00:03:02 UP      10000               1500
mgmt002st001 eth0      02:1c:d1:01:03:01 UP      10000               1500
mgmt002st001 eth1      02:1c:d1:01:03:02 UP      10000               1500
```

4. Create new bonds on all nodes by using the **mknwbond** command with the **mode** parameter, as shown in Example 3-4. Note the usage of bond mode 4 and the layer3+4 transmission hash. Both management and data traffic can be used on this bond.

   You can use the **mknwbond** command-line interface (CLI) command to create a bond interface by specifying the subordinate network interfaces on a specified node that are grouped to act as one logical network interface.

Example 3-4 Creating new bond interfaces

```
$ mknwbond int001st001 eth0,eth1 --mode 4 --mtu 9000 -xmit layer3+4
EFSSG0577W Warning: creating a bond with mode 802.3ad (4) might require
additional switch configuration work to access the network.
EFSSG0089I Network bond ethX0 successfully created.

$ mknwbond mgmt001st001 eth0,eth1 --mode 4 --mtu 9000 -xmit layer3+4
EFSSG0577W Warning: creating a bond with mode 802.3ad (4) might require
additional switch configuration work to access the network.
EFSSG0089I Network bond ethX0 successfully created.

$ mknwbond mgmt002st001 eth0,eth1 --mode 4 --mtu 9000 -xmit layer3+4
EFSSG0577W Warning: creating a bond with mode 802.3ad (4) might require
additional switch configuration work to access the network.
EFSSG0089I Network bond ethX0 successfully created.
```

5. Check the network and bonding configuration, as shown in Example 3-5.

Example 3-5 Check the network and bonding configuration

```
$ lsnwinterface -x
Node Interface MAC Master/Subordinate Bonding mode Transmit hash policy Up/Down
Speed IP-Addresses MTU
int001st001  ethX0     02:1c:d1:02:03:00 MASTER 802.3ad (4) layer3+4 UP 20000
9000
int001st001  ethXsl0_0 02:1c:d1:02:03:01 SUBORDINATE UP 10000 9000
int001st001  ethXsl0_1 02:1c:d1:02:03:02 SUBORDINATE UP 10000 9000
mgmt001st001 ethX0     02:1c:d1:00:03:00 MASTER 802.3ad (4) layer3+4 UP 20000
9000
mgmt001st001 ethXsl0_0 02:1c:d1:00:03:01 SUBORDINATE UP 10000 9000
mgmt001st001 ethXsl0_1 02:1c:d1:00:03:02 SUBORDINATE UP 10000 9000
mgmt002st001 ethX0     02:1c:d1:01:03:00 MASTER 802.3ad (4) layer3+4 UP 20000
9000
mgmt002st001 ethXsl0_0 02:1c:d1:01:03:01 SUBORDINATE UP 10000 9000
mgmt002st001 ethXsl0_1 02:1c:d1:01:03:02 SUBORDINATE UP 10000 9000
```

6. Attach the network or networks to the interface, as shown in Example 3-6.

Example 3-6 Attach the network or networks to the interface

```
$ attachnw 10.0.0.0/24 ethX0 -g int
EFSSG0015I Refreshing data.
```

3.3.5 Bonding examples

In this section, several examples are shown that use CLI commands to creates bonds and display the configurations.

Mode 6 bonding example with four ports

In the following example, all interface nodes and interface/management nodes have one 4-port 1 GbE Ethernet adapter installed. One bond with all four ports using active/active settings is created. On all nodes, eth0, eth1, eth2, and eth3 adapters are bonded to one single bond.

Configuration requirements

The following prerequisites must be fulfilled:

- No network is assigned to the subordinates.
- No IP address is active on any of the subordinates.

Configuration steps

Complete the following configuration steps:

1. Detach the network. The commands to detach the network from the interface and check the configuration are the same, as shown previously in Example 3-1 on page 80.
2. Remove the existing bonds on all the servers, as shown in Example 3-2 on page 81.

3. Create the new bond on all nodes using the **mknwbond** command with the **mode** parameter (Example 3-7).

Example 3-7 Configuring network bonding by using the CLI

```
$ mknwbond int001st001 eth0,eth1,eth2,eth3 --mode 6
EFSSG0089I Network bond ethX0 successfully created.

$ mknwbond mgmt001st001 eth0,eth1,eth2,eth3 --mode 6
EFSSG0089I Network bond ethX0 successfully created.

$ mknwbond mgmt002st001 eth0,eth1,eth2,eth3 --mode 6
EFSSG0089I Network bond ethX0 successfully created.
```

4. Use the **lsnwinterface** command to check the settings (Example 3-8).

Example 3-8 Using the lswinterface command to check the settings

```
$ lsnwinterface -x
Node         Interface MAC               Master/Subordinate Bonding mode
Transmit hash policy Up/Down Speed IP-Addresses MTU
int001st001  ethX0     02:1c:d1:02:03:00 MASTER balance-alb (6) UP 1000 1500
int001st001  ethXsl0_0 02:1c:d1:02:03:01 SUBORDINATE UP      1000 1500
int001st001  ethXsl0_1 02:1c:d1:02:03:02 SUBORDINATE UP      1000 1500
int001st001  ethXsl0_2 02:1c:d1:02:03:03 SUBORDINATE UP      1000 1500
int001st001  ethXsl0_3 02:1c:d1:02:03:04 SUBORDINATE UP      1000 1500
mgmt001st001 ethX0     02:1c:d1:00:03:00 MASTER balance-alb (6) UP 1000 1500
mgmt001st001 ethXsl0_0 02:1c:d1:00:03:01 SUBORDINATE UP      1000 1500
mgmt001st001 ethXsl0_1 02:1c:d1:00:03:02 SUBORDINATE UP      1000 1500
...................
```

5. Attach the network or networks to the interface, as shown in Example 3-9.

Example 3-9 Attach the network or networks to the interface

```
$ attachnw 10.0.0.0/24 ethX0 -g int
EFSSG0015I Refreshing data.
```

Configuring 1 GbE and 10 GbE adapters for high availability

Bonding two 1 GbE ports with mode 1 is typically a safe setting when less than 1 Gb per Interface node is required from the SONAS solution. Bonding mode 1 is the default configuration when creating a bond with 1 GbE and 10 GbE adapters. Mode 1 delivers failover, but no increase in throughput, as shown in Figure 3-6.

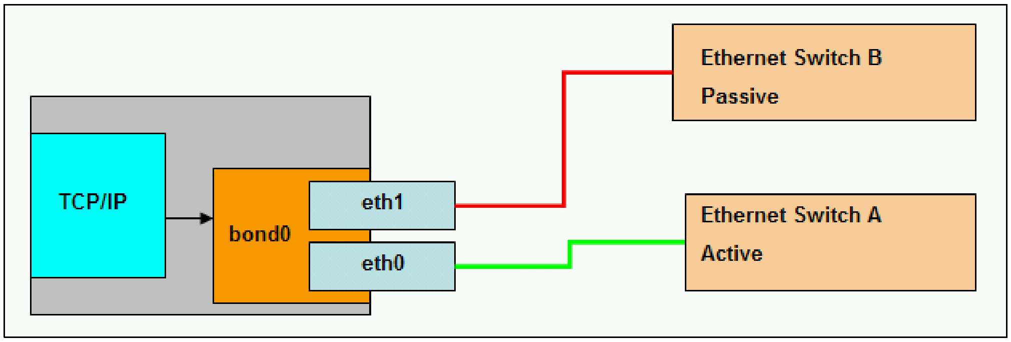


Figure 3-6 High availability example layout

Configuration requirements

The following prerequisites must be fulfilled:

- No network is assigned to the subordinates.
- No IP address is active on any of the subordinates.

Configuration steps

Complete the following configuration steps:

1. Detach the network. The commands to detach the network from the interface and check the configuration are the same, as shown previously in Example 3-1 on page 80.
2. Remove the existing bonds on all the servers, as shown in Example 3-2 on page 81.
3. Create the new bond on all nodes using the `mknwbond` command. No parameters are required, the default mode is active/backup, and the default MTU is 1500. Example 3-10 shows the command for two systems.

Example 3-10 Creating a default bond

```
$ mknwbond int001st001 eth0,eth1
EFSSG0089I Network bond ethX0 successfully created.
$ mknwbond mgmt001st001 eth0,eth1
EFSSG0089I Network bond ethX0 successfully created.
$ mknwbond mgmt002st001 eth0,eth1
EFSSG0089I Network bond ethX0 successfully created.
```

4. Use `lsnwinterface` command to check the settings. The output is shown in Example 3-11.

Example 3-11 Two interface, mode 1 bond

```
$ lsnwinterface -x
Node Interface MAC Master/Subordinate Bonding mode Transmit hash policy Up/Down
Speed IP-Addresses MTU
int001st001  ethX0     02:1c:d1:02:03:00 MASTER active-backup (1) UP 1000 1500
int001st001  ethXsl0_0 02:1c:d1:02:03:01 SUBORDINATE UP 1000 1500
int001st001  ethXsl0_1 02:1c:d1:02:03:02 SUBORDINATE UP 1000 1500
mgmt001st001 ethX0     02:1c:d1:00:03:00 MASTER active-backup (1) UP 1000 1500
mgmt001st001 ethXsl0_0 02:1c:d1:00:03:01 SUBORDINATE UP 1000 1500
mgmt001st001 ethXsl0_1 02:1c:d1:00:03:02 SUBORDINATE UP 1000 1500
mgmt002st001 ethX0     02:1c:d1:01:03:00 MASTER active-backup (1) UP 1000 1500
mgmt002st001 ethXsl0_0 02:1c:d1:01:03:01 SUBORDINATE UP 1000 1500
mgmt002st001 ethXsl0_1 02:1c:d1:01:03:02 SUBORDINATE UP 1000 1500
```

3.4 SONAS networking preferred practices

This section describes preferred practices for configuring the SONAS network. It considers configuring management and data traffic on the same physical network, and then on different physical networks.

3.4.1 Network naming

To clearly distinguish subordinate and bonded interfaces, the term *subordinate* is used for a subordinate interface and *bond* for a bonded interface. The following conventions are used in this book:

- Public interfaces are usually bonded into redundant groups.
- The installation process enforces a required naming convention for network interfaces:
  - `ethX0...ethXn` are bonded interfaces for the public network.
  - `ethsl0...ethsln` are subordinate interfaces for the public network.
  - `mgmt0...mgmtn` are bonded interfaces for the management network.
  - `mgmtsl0...mgmtsln` are subordinate interfaces for the management network.
  - `data0...datan` are bonded interfaces of an InfiniBand network.
  - `ib0...ibn` are subordinate interfaces of an InfiniBand network.

3.4.2 Management and data traffic on the same physical network

For traffic on the same physical network, this section considers management and data traffic on the same subnet.

Management and data traffic on the same subnet

If you configure one subnet for both management traffic and data traffic, you must assign both management traffic and data traffic to the same network interface (VLAN or no VLAN). For example, when both management IPs and external SONAS traffic IPs are assigned from the same network, do not use the 1 GbE ports if 10 GbE is available. Instead, move or combine management responsibilities and traffic IP addresses together on the same ports. If the SONAS has access to 10 GbE, combining both on the 10 GbE bonded network interfaces is the preferred practice.

In this case, the 10 GbE interface is preferred because the single subnet is carrying both management and data traffic, and the highest bandwidth rates with the least resource use. There is no need for specialized routing tables.

Using 10 Gb bonded ports is the preferred practice for SONAS external traffic. Active/Active configurations with Layer3+4 `xmit_hash_policies` are best for high-speed performance, when possible. See Figure 3-7.

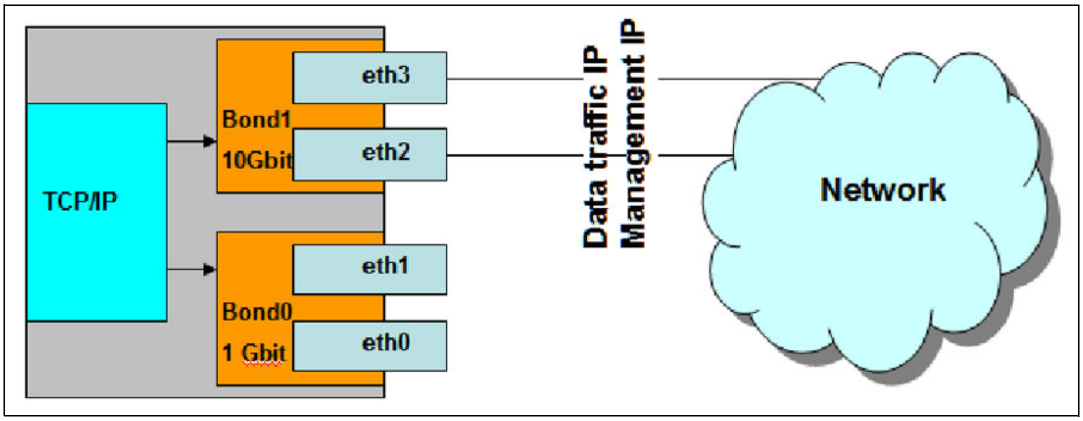


Figure 3-7 Example solution for management and data traffic on the same subnet

Configuration requirements

This scenario has the following configuration requirements:

- Configure the existing bond for optimal communications and bond mode.
- Network interface bonds are changed from the CLI with the **rmnwbond** and **mknwbond** commands on a node-by-node basis. However, the node must first be removed from a network before bonds can be changed at the CLI.
- Network groups must be set up to include nodes that are participating in IP managed external services. It is a preferred practice to create a netgroup for this purpose and not use the default netgroup (Default) because the default netgroup is not as flexible for special maintenance requirements.

Configuration steps

Complete the following steps to create this configuration:

1. Check, set, or validate management network settings.
2. Check, set, or validate network bonds, modes, and xmit_hash_policies.
3. Change the management network configuration.
4. Add a new network.
5. Attach the network to the interface.
6. Use the **lsnwinterface** command to check the settings.

Example 3-12 is an example of setting up the network by using the CLI.

Example 3-12 Network setup CLI example

```
$ lsnwmgt
Interface Service IP Node1 Service IP Node2 Management IP Network       Gateway
VLAN ID
ethX0     10.0.0.18        10.0.0.19        10.0.0.10     255.255.255.0 10.0.0.1

$ chnwmgt --interface ethX1
EFSSG0015I Refreshing data.

$ lsnwmgt
Interface Service IP Node1 Service IP Node2 Management IP Network       Gateway
VLAN ID
ethX1 10.0.0.18        10.0.0.19        10.0.0.10     255.255.255.0 10.0.0.1

$ mknw 10.0.0.0/24 0.0.0.0/0:10.0.0.1 --add 10.0.0.100,10.0.0.101

$ lsnw
Network     VLAN ID Network Groups IP-Addresses               Routes
10.0.0.0/24         int               10.0.0.121,10.0.0.122   0.0.0.0/0:10.0.0.1

$ attachnw 10.0.0.0/24 ethX1 -g int
EFSSG0015I Refreshing data.
```

3.4.3 Management and data traffic on different subnets

If you configure different subnets for both management traffic and data traffic, you can assign both management traffic and data traffic to the same network interface. In this case, the 10 GbE interface is preferred. For more information, see 3.5, “SONAS routing” on page 91 and 3.4.5, “Using and configuring VLANs” on page 89.

Figure 3-8 shows the management and data traffic on different subnets, but still physically using the same bond.

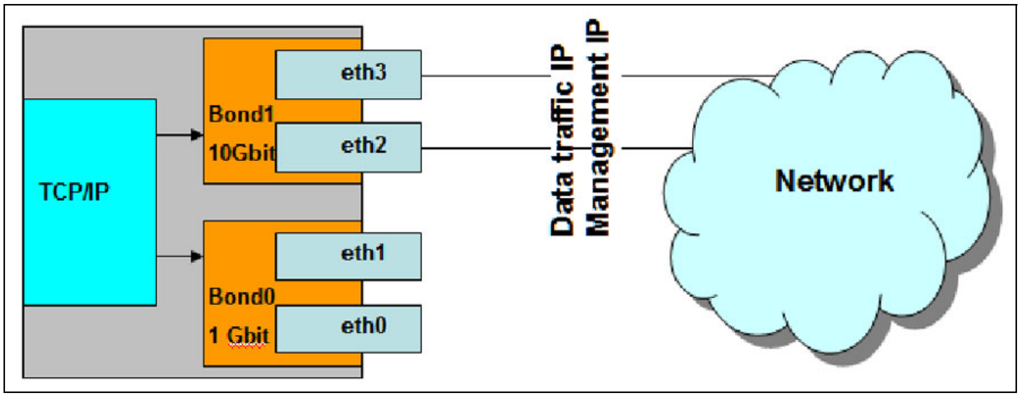


Figure 3-8 Management and data traffic on different subnets

Configuration requirements

This configuration has the following requirements:

- Configure the existing bond for optimal communications and bond mode.
- Network interface bonds are changed from the CLI with the **rmnwbond** and **mknwbond** commands on a node-by-node basis. However, the node must first be removed from a network before bonds can be changed at the CLI.
- Network groups must be set up to include nodes that are participating in IP managed external services. It is a preferred practice to create a netgroup for this purpose, and not to use the default netgroup (Default). The default netgroup is not as flexible for special maintenance requirements.

Configuration steps

Complete the following steps to create this configuration:

1. Check, set, or validate management network settings.
2. Check, set, or validate network bonds, modes, and `xmit_hash_policies`.
3. Change the management network configuration.
4. Add the new network.
5. Attach the network to the interface.
6. Use the **lsnwinterface** command to check the settings.

Example 3-13 shows setting up the management and data traffic on different subnets by using the CLI.

Example 3-13 Sample CLI setup

```
$ lsnwmgt
Interface Service IP Node1 Service IP Node2 Management IP Network Gateway  VLAN ID
ethX0     10.0.0.18        10.0.0.19        10.0.0.10     255.255.255.0 10.0.0.1

$ chnwmgt --interface ethX1
EFSSG0015I Refreshing data.

$ lsnwmgt
Interface Service IP Node1 Service IP Node2 Management IP Network Gateway  VLAN ID
ethX1 10.0.0.18        10.0.0.19        10.0.0.10     255.255.255.0 10.0.0.1
```

```
$ mknw 10.20.0.0/24 0.0.0.0/0:10.20.0.1 --add 10.20.0.100,10.20.0.101

$ lsnw
Network     VLAN ID Network Groups IP-Addresses               Routes
10.20.0.0/24        int                10.20.0.121,10.20.0.122
  0.0.0.0/0:10.20.0.1

$ attachnw 10.20.0.0/24 ethX1 -g int
EFSSG0015I Refreshing data.
```

3.4.4 Management and data traffic on different physical networks

This example shows an active management node that is also configured for data access. The system has both 1 GbE and 10 GbE network interfaces. If management traffic and data traffic are on two different subnets, the preferred practice is to configure management traffic on 1 GbE adapters (ethX0) and data traffic on the 10 GbE network interface, which is normally ethX1.

If there are two separate physical networks, complete the following tasks:

1. Create a bond to configure data traffic on the 10 GbE network interface.
2. Create a bond to configure the management network on the 1 GbE network interface, which is normally ethX0.

Figure 3-9 shows the management and data traffic on different physical networks.

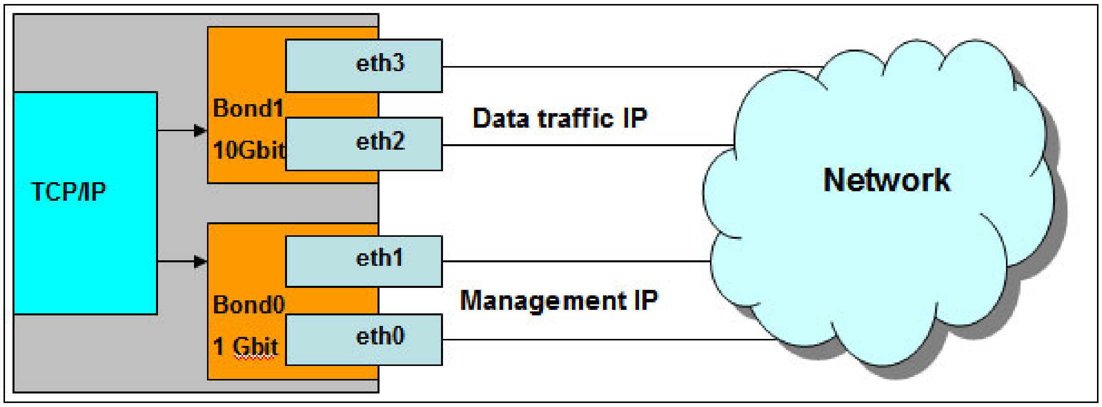


Figure 3-9 Management and data traffic on a different physical network sample solution

Configuration requirements

The following conditions must be met for this type of configuration:

- The existing bond should be configured for optimal communications and bond mode.
- Change network interface bonds from the CLI with the **rmnwbond** and **mknwbond** commands on a node-by-node basis. However, the node must first be removed from a network before bonds can be changed at the CLI.
- Network groups must be set up to include nodes that are participating in IP managed external services. It is a preferred practice to create a netgroup for this purpose, and not to use the default netgroup (Default), because the default netgroup is not as flexible for special maintenance requirements.

Configuration steps

Complete the following steps to configure the network:

1. Check, set, or validate management network settings.
2. Check, set, or validate network bonds, modes, and `xmit_hash_policies`.
3. Change the management network configuration.
4. Add a new network.
5. Attach the network to the interface.
6. Use the **`lsnwinterface`** command to check the settings.

Example 3-14 shows a sample setup of management and data traffic on different networks by using the CLI.

Example 3-14 CLI setup example

```
$ lsnwmgt
Interface Service IP Node1 Service IP Node2 Management IP Network Gateway  VLAN ID
ethX0     10.0.0.18        10.0.0.19        10.0.0.10     255.255.255.0 10.0.0.1

$ lsnwgroup
Network Group Nodes                    Interfaces
DEFAULT
ext           mgmt001st002
int           int001st002,mgmt002st002

$ mknw 10.0.0.0/24 0.0.0.0/0:10.0.0.1 --add 10.0.0.100,10.0.0.101

$ attachnw 10.0.0.0/24 ethX1 -g int
EFSSG0015I Refreshing data.

$ lsnwgroup
Network Group Nodes                    Interfaces
DEFAULT
ext           mgmt001st002
int           int001st002,mgmt002st002 ethX1
```

Requirement: Each bond must contain interfaces that all have the same speed, so that all of the 1 GbE interfaces are in one bond, and all of the 10 GbE interfaces are in a separate bond.

3.4.5 Using and configuring VLANs

VLAN-based subnets are supported with the public network for both management and file access functions. For management access, initially configure the system with a non-VLAN-based subnet and IP addresses. You can initially create the file access network and IP addresses with VLAN subnet support.

Restriction: VLAN 1 is not supported for SONAS client traffic. This restriction is intended to prevent security exposure and reduce the probability of network configuration errors.

VLAN 1 has been used within the industry as the default or native VLAN. Many vendors use VLAN ID value 1 for management traffic by default. Configuring VLAN 1 as available within the network can be a security exposure because VLAN 1 might span large parts of the switched network by default. Common practice in the industry strongly discourages the use of VLAN 1 for user client traffic. Setting VLAN 1 for user client traffic can require explicit steps that differ by vendor, and can be prone to configuration error.

VLAN Tagging is a method of providing deeper network segmentation on a single physical network. IEEE 802.1Q is the networking standard that supports VLANs on an Ethernet network. The 802.1Q standard defines a system of VLAN tagging for Ethernet frames, and the accompanying procedures to be used by bridges and switches in handling such frames.

Portions of the network that are VLAN-aware (IEEE 802.1Q conformant) can include VLAN tags. Traffic on a VLAN-unaware (IEEE 802.1D conformant) portion of the network does not contain VLAN tags.

When a frame enters the VLAN-aware portion of the network, a tag is added to represent the VLAN membership of the frame's port or the port/protocol combination, depending on whether port-based or port-and-protocol-based VLAN classification is being used. Each frame must be distinguishable as being within exactly one VLAN. A frame in the VLAN-aware portion of the network that does not contain a VLAN tag is assumed to be flowing on the native (or default) VLAN.

The data structure of VLAN tagging involves inserting VLAN tags into each data fragment.

Figure 3-10 depicts using different VLANs on the same adapter.

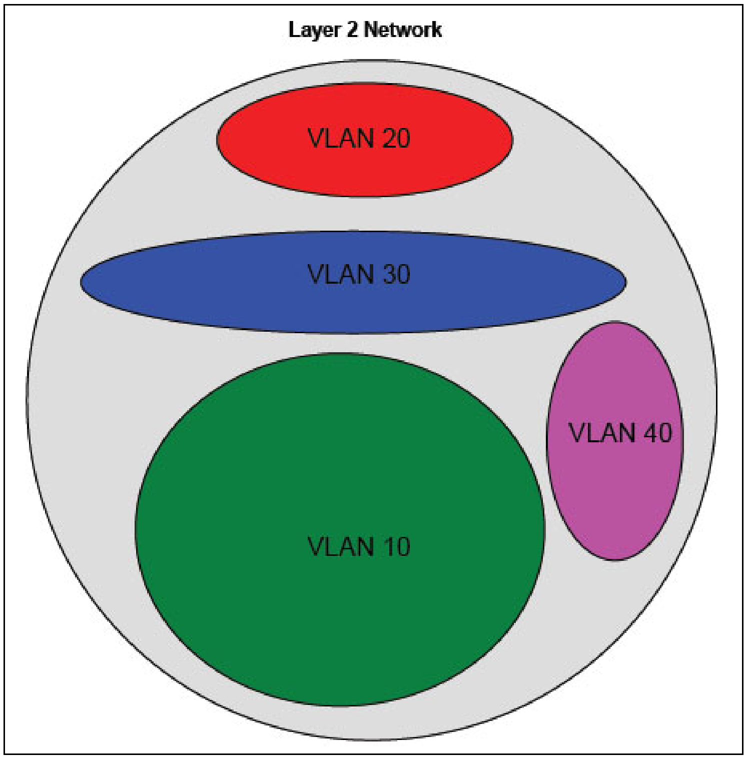


Figure 3-10 Using different VLANs on the same adapter

Example 3-15 shows a VLAN-based subnet. The example uses interface ethX1 as a 10 GB optional Ethernet interface. In the example, the Ethernet interface has already been created with the **mknwbond** command. Network groups are also created with the **mknwgroup** command. The **mknw** command is used with the **--vlan** parameter to create the network.

> **Important:** The IP ranges, routes, and VLANs used in this example are not likely to be used in a real configuration.

Example 3-15 Example CLI setup for VLANs on SONAS

```
$ mknw 10.0.0.0/24 0.0.0.0/0:10.0.0.1 --vlanid 10 --add 10.0.0.100,10.0.0.101
$ mknw 10.20.0.0/24 0.0.0.0/0:10.20.0.1 --vlanid 20 --add 10.20.0.100,10.20.0.101
$ mknw 10.30.0.0/24 0.0.0.0/0:10.30.0.1 --vlanid 30 --add 10.30.0.100,10.30.0.101
$ mknw 10.40.0.0/24 0.0.0.0/0:10.40.0.1 --vlanid 40 --add 10.40.0.100,10.40.0.101

$ lsnw
Network     VLAN ID Network Groups IP-Addresses Routes
10.0.0.0/24 10 int            10.0.0.100,10.0.0.101 0.0.0.0/0:10.0.0.1
10.20.0.0/24 20 int           10.20.0.100,10.10.20.101 0.0.0.0/0:10.20.0.1
10.30.0.0/24 30 int           10.30.0.100,10.30.0.101 0.0.0.0/0:10.30.0.1
10.40.0.0/24 40 int           10.40.0.100,10.40.0.101 0.0.0.0/0:10.40.0.1

$ lsnwinterface
Node         Interface MAC Master/Subordinate Bonding mode      Transmit hash
policy Up/Down Speed IP-Addresses MTU
int001st002  ethX0     02:1c:5b:02:03:00 MASTER active-backup (1) UP 10000 1500
int001st002  ethX0.10 02:1c:5b:02:03:00 MASTER UP 10000 10.0.0.100,10.0.0.101 1500
int001st002  ethX0.20 02:1c:5b:02:03:00 MASTER UP 10000 10.20.0.100,10.20.0.101
1500
int001st002  ethX0.30 02:1c:5b:02:03:00 MASTER UP 10000 10.30.0.100,10.30.0.101
1500
int001st002  ethX0.40 02:1c:5b:02:03:00 MASTER UP 10000 10.40.0.100,10.40.0.101
1500
```

When connecting the clients and the SONAS on multiple LAN switches, the connectivity between the switches must have the capability of transferring the data rate required for the data traffic. For better results, use 10 GbE connectivity between the switches. Another option is to define another link aggregation between the switches such that they can transfer the required bandwidth.

3.5 SONAS routing

The SONAS system supports three categories of routing definitions:

- Locally defined networks
- Statically routed networks
- System default gateway

These definitions provide the foundation for all routed traffic within the SONAS system. Each network is applied, or attached, to an interface, also termed a *bond*, which is built on top of a specific network adapter. Adapters support either 1 GbE or 10 GbE traffic speeds.

When the SONAS system is initially configured, the management network is created so that administrative tasks can be accomplished. The network is defined by a management subnet and an optional system default gateway. Static routes are not supported on this network. This network can be viewed by using the **lsnwmgt** CLI command. For detailed information, see *IBM SONAS Implementation Guide*, SG24-7962.

The data networks are configured to make data services available. Each network can define local networks, static routes, and a system default gateway. These network definitions can be viewed with the **lsnw** CLI command, as shown in Example 3-16.

Example 3-16 Sample VLAN configurations

```
$ lsnw
Network     VLAN ID Network Groups IP-Addresses               Routes
10.0.0.0/24          int        10.0.0.121,10.0.0.122       0.0.0.0/0:10.0.0.1
10.17.0.0/24 99      int        10.17.0.100,10.17.0.101     0.0.0.0/0:10.17.0.1
10.18.0.0/24 11      int        10.18.0.100,10.18.0.101     0.0.0.0/0:10.18.0.1
10.21.0.0/24         int        10.21.0.100,10.21.0.101     0.0.0.0/0:10.21.0.1
10.31.0.0/24                    10.31.0.200,10.31.0.201     0.0.0.0/0:10.31.0.1
10.35.0.0/24         int        10.35.0.100,10.35.0.101
```

Note: These configurations are only definitions. The networks might or might not be attached.

SONAS routing policies are designed to independently route traffic on each network that is configured to the SONAS system, including an optional system default gateway. For each network definition, this implementation enables the system default gateway to be configured to not conflict with other network system default gateways that are defined in the SONAS system. The **lsnwdg** command that is shown in Example 3-17 displays the default gateways and explicit declarations.

Example 3-17 Policy-based routing sample

```
$ lsnwsdg
Node         Active default route Interface Gateway
int001st001  10.11.136.1 (ethX0)   N/A       N/A
int002st001  10.11.136.1 (ethX0)   N/A       N/A
mgmt001st001 10.11.136.1 (ethX0)   N/A       N/A
mgmt002st001 10.11.136.1 (ethX0)   N/A       N/A
EFSSG1000I The command completed successfully.
```

The SONAS system includes many functions that require connections to remote services. Example connections include authentication servers, such as LDAP or Active Directory, NTP servers, and other SONAS servers for asynchronous replication. To maintain access to these services, it is a preferred practice to setup a NAT gateway. See the Configure the NAT gateway topic in the IBM Knowledge Center:

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

The following list summarizes preferred practices to remember:

- Flatten the network and reduce hop counts where possible for SONAS, clients, and NAS services. Keep round trip times (RTT) to a minimum between critical services:
  - SONAS and Client to DNS
  - SONAS and Client to Authentication Primary and Secondaries
  - SONAS to Antivirus (AV) server
  - SONAS to Backup or HSM server
  - SONAS and Client to SONAS Replication Site
- Use 10 GbE for external client to SONAS communications, backups, antivirus, and replication.
- When possible, use bonding mode 4 (LACP) over 10 GbE ports with layer3+4 `xmit_hash_policy` and MTU 9000 for high-speed active/active communications on networks. Bond ports before you establish your netgroups. Use a large MTU only if your network supports large packets end to end.
- Avoid aggressive or unnecessary use of VLAN tagging in simple network strategies.
- Avoid complex IP routing requirements in your networking strategy.
- Monitor network saturation levels on both clients and SONAS interface nodes for understanding when you are driving near rated line speeds or when you feel you might need to add nodes or network interfaces to improve bandwidth use.
- Monitor and manage TCP/IP traffic on SONAS interface nodes and clients to understand whether NAS share bandwidth is being compromised by competing network applications (such as backup or antivirus).
- Establish a tight relationship between network and the SONAS support teams for managing high performance from SONAS. Communicate and coordinate changes that might affect performance or reliability.
- Establish network segment based *test points* (trusted service savvy clients) in your infrastructure that can serve as a maintenance service point for validating network segment response assurance (when possible). This configuration simplifies troubleshooting and staff response to high-level network troubleshooting.
- Establish a maintenance and troubleshooting protocol with your network administration staff before going into production for establishing tools, communications, and requirements for the quickest capture and sharing of network diagnostic information for working together to troubleshoot issues that might arise.

4

Storage configuration

This chapter describes the preferred practice configuration and considerations for IBM Scale Out Network Attached Storage (SONAS) storage subsystems in general and covers the best configurations for different common workloads.

It is important to consider that at the heart of SONAS is the IBM General Parallel File System (IBM GPFS) engine that drives the basic value of the SONAS solution with its world-leading file system speed, reliability, and expandability (scale).

To take full advantage of the features of GPFS in SONAS, you also need to understand the key points that drive this value and how to implement your solution design to best use the highest standards of performance, redundancy, and scalability in your planned solution.

This chapter briefly introduces these concepts and outlines a *preferred practice* plan for achieving these high standards and provides design guidelines to meet that standard as closely as possible.

This chapter includes the following information:

- General Parallel File System
- SONAS supported hardware
- Failure groups and storage pools

4.1 General Parallel File System

It is important to understand the General Parallel File System (GPFS) before you consider how best to consider the hardware and software stacks in SONAS.

4.1.1 Overview

GPFS is a high-performance, shared-disk clustered file system that is developed by IBM. It is used by many of the world's largest commercial companies, and some of the largest supercomputers in the world's Top 500 List.

In common with typical cluster file systems, GPFS provides concurrent high-speed file access to applications that are running on multiple nodes of clusters. It can be used with IBM AIX 5L™ clusters, Linux clusters, Microsoft Windows Server, or a heterogeneous cluster of AIX, Linux, and Windows nodes. In addition to providing file system storage capabilities, GPFS provides tools for management and administration of the GPFS cluster, and enables shared access to file systems from remote GPFS clusters.

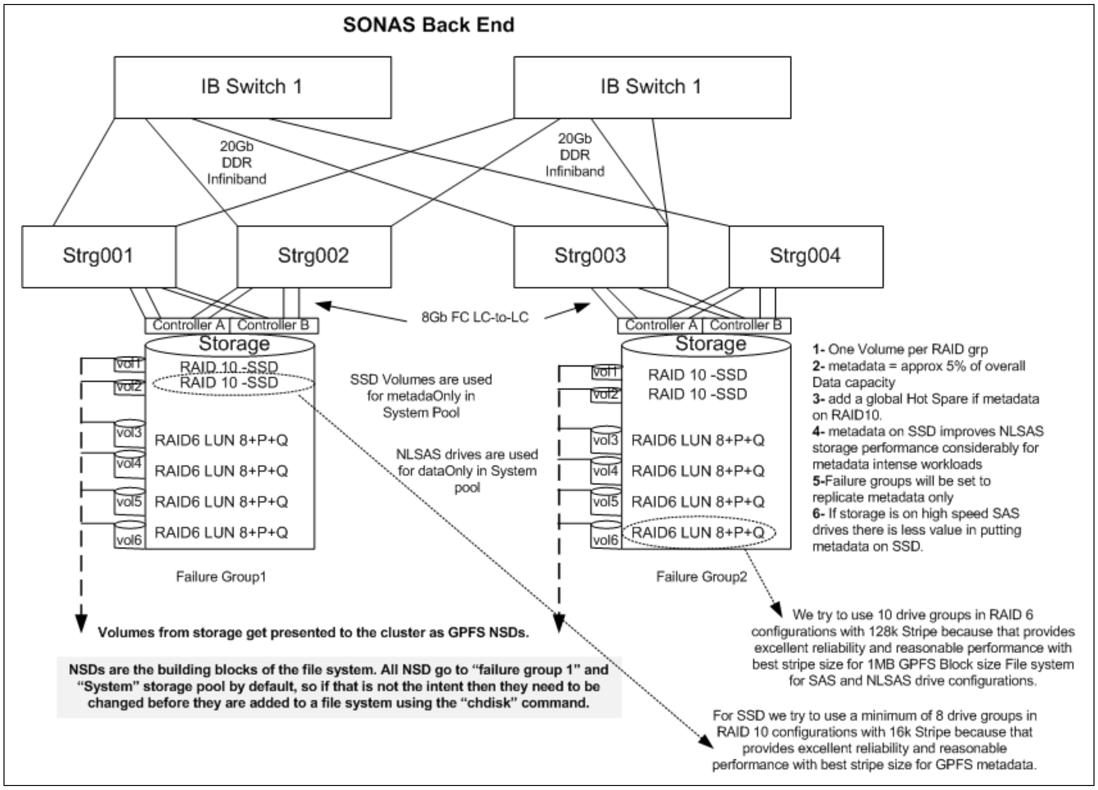
GPFS has been available on the AIX operating system since 1998, on Linux since 2001, and on Windows Server since 2008. GPFS is offered as part of the IBM System Cluster 1350. The most recent release of GPFS 3.5 offers Active File Management to enable asynchronous access and control of local and remote files, therefore enabling global file collaboration.

SONAS runs with GPFS release 3.4.0 - 15 (SONAS tests and sends current supported GPFS versions on a per-release basis. For example, SONAS 1.5.1 has GPFS version 3.5.0-21 and all the features and functions inherent with that build.) It is also important to note that the use of software in the SONAS stack is highly customized, and pre-tuned for optimal value and reliability within the product design goals. All software components are customized, and no tuning is required for optimal configuration. However, proper planning for system building, file systems, disk striping, and redundancy is required.

4.1.2 Back-end storage design for performance and redundancy

The SONAS hardware is defined to support a front-end set of service nodes (the GPFS Clients) that are highly available and highly reliable, and the back-end set of service nodes (the GPFS Servers) that are designed to provide the foundation of scale out storage device striping and scalable performance.

The back-end devices and their attached storage are the key subjects that are reviewed in this chapter. Figure 4-1 shows a sample SONAS back-end storage configuration.



1- One Volume per RAID grp
2- metadata = approx 5% of overall Data capacity
3- add a global Hot Spare if metadata on RAID10.
4- metadata on SSD improves NLSAS storage performance considerably for metadata intense workloads
5-Failure groups will be set to replicate metadata only
6- If storage is on high speed SAS drives there is less value in putting metadata on SSD.

Volumes from storage get presented to the cluster as GPFS NSDs.

NSDs are the building blocks of the file system. All NSD go to "failure group 1" and "System" storage pool by default, so if that is not the intent then they need to be changed before they are added to a file system using the "chdisk" command.

We try to use 10 drive groups in RAID 6 configurations with 128k Stripe because that provides excellent reliability and reasonable performance with best stripe size for 1MB GPFS Block size File system for SAS and NLSAS drive configurations.

For SSD we try to use a minimum of 8 drive groups in RAID 10 configurations with 16k Stripe because that provides excellent reliability and reasonable performance with best stripe size for GPFS metadata.

Figure 4-1 Back-end RAID and NSD considerations

Back-end designs are important for performance and redundancy. GPFS does a good job of striping data across all the devices in the file system, storage pool, layout, and so on. However, this functionality does not negate or diminish the value of good planning. Typically, for enterprise-class environments, you want to consider every opportunity for redundancy against its logical effect on performance and cost to determine the best possible reliability and performance at an affordable price. Price is relative to performance and reliability.

There is an old saying: "Cost, quality, speed, pick any two". If storage is fast and highly redundant, it typically costs more. If it is fast and cheap, quality is often lost. If it is high-quality and cheap, it is typically not fast. You need more spindles for higher redundancy, and the extra write activity with higher redundancy can decrease performance. The balance of performance and reliability is always a concern. However, depending on the configuration, more spindles can also give you more performance.

For non-IBM XIV based solutions, this chapter considers several types of drive technology behind SONAS:

- Solid-state drive (SSD)
- 15,000 (15 K) revolutions per minute (RPM) serial-attached SCSI (SAS)
- 10K RPM SAS
- Near Line SAS (NLSAS)

For now, tape is not described. Typically, the higher the speed, the smaller and more expensive (per terabyte (TB)) the drive technology.

The preferred practice for SONAS is to plan for performance with reliability.

4.1.3 Redundancy considerations

There is more than one level of redundancy. SONAS can have redundancy at both the drive or Redundant Array of Independent Disks (RAID) technology level, and the GPFS (file system) level.

For SONAS, with the expectation that clients are using it as an enterprise class storage solution, protect the storage at the RAID level. At a minimum, the file system metadata should also be protected at the file system level.

In this case, place appropriate RAID-protected storage into a minimum of two GPFS Network Shared Disk (NSD) Failure Groups with *metadata* replication. However, there are some exceptions to this practice that weight the back-end technology against the value of redundancy.

> **Note:** Failure group and RAID protection are described in depth later in this chapter. For now, the review is provided at a higher conceptual level.

4.1.4 GPFS metadata structures and preferred configurations

In the case where multiple tiers of drive types are invested, dedicate the fastest drive types to metadata in the default `system` pool, the next fastest be dedicated to data-only in the `system` pool, and the third fastest be dedicated to a tier 2 and 3 or *silver* storage pool within the file system.

This configuration provides isolation of metadata from the data drives, and reduces contention to normal data access workloads from the metadata scan activity that is used by lookups, backups, snapshots, replication, and virus scans. This configuration provides the best approach to high-performance solutions.

This type of solution requires thorough planning. It must be sized appropriately to ensure that the positive effect of this design is realized in actuality. Striping data across four SSD drives might not be better than striping data across 120 SAS drives. The more drives that are used, the more resources, or *read/write heads*, are available to support data search and the file system input/output (I/O) demands, and, typically, the better the overall performance.

There can be an advantage to separating metadata and data from a single drive type or drive set, because high-speed GPFS metadata scan time can be affected when the I/O shares or competes for access to read/write heads for normal data I/O, snapshots, backup, and restore, or replication scans. This is the case when NLSAS technology is used for metadata and file data I/O.

Metadata structures are small and use random I/O in contrast to many normal data I/O profiles. So, keep metadata on smaller, high-speed technology.

In this context, when a system is designed with predominantly large drive (NLSAS) technology, it is often advantageous to put the NLSAS drives into the default `system` storage pool as `dataOnly`, and add roughly 5% of the usable file system capacity as SSD or high-speed 15 K RPM SAS drives in the default `system` pool as `metadataOnly`.

This configuration enables all metadata scan transactions to happen independently from file data read/write I/O. This improves overall performance for systems that have high metadata scans for backup, replication, snapshots, directory listings, antivirus, and other metadata-intense operations.

Using 5% can be a reasonable and adequate standard for most file system data profiles. The 5% estimates the capacity when you are replicating metadata (which is a preferred practice). Now remember that the 5% estimate is for average file types and sizes.

If you anticipate having fewer large files, you can expect there to be a demand for less metadata, and that number can be reduced (in that case) to 3 or 4% safely. Alternatively, if you anticipate a huge number of small files or objects, you might choose to add a little extra capacity for metadata (say 6 or 7%).

It is important to consider both the metadata device type and capacity and the NSD count. Figure 4-2 is an example of a basic design of the back-end connectivity.

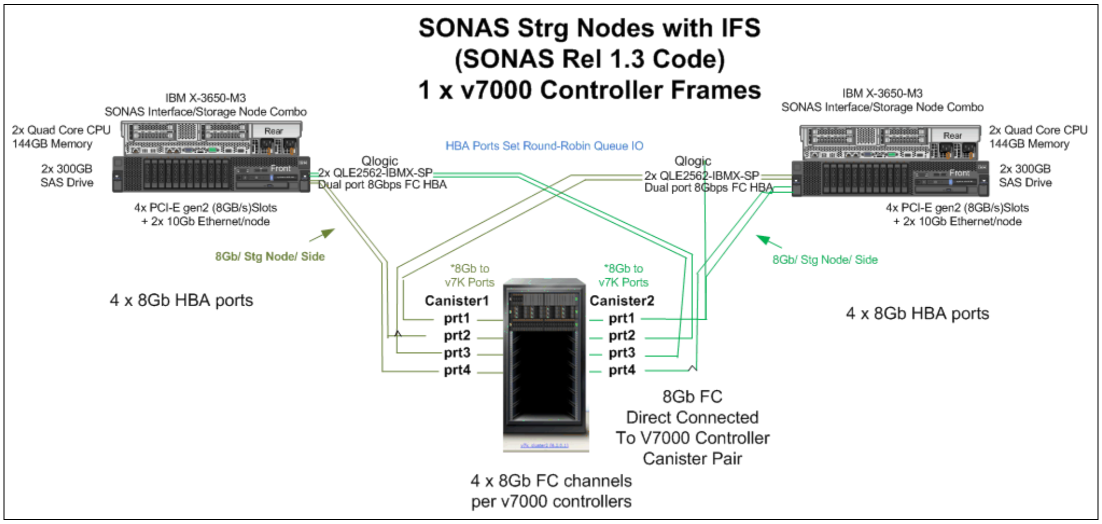


Figure 4-2   Back-end balance of virtual port connections (switches not included)

For the basic design of the back-end connectivity, there is an opportunity to optimize both the storage node host bus adapter (HBA) ports and the back-end subsystem storage controller channels by using NSDs (both data and metadata) in groups of four, per storage node, as you map your file system layout. So, it is a preferred practice to create volumes in groups of eight on the subsystems, and to map four active volumes (known to GPFS as NSDs) per storage node, per device type (or tier).

Note: The devices are mapped to the storage node pair (as a cluster) versus the individual nodes, and laid out in a balanced configuration of active/passive NSD devices per node. So, in groups of eight, you can map each device to an active port and channel to maximize access to the subsystem.

For a system that is designed with 240 x 15 K RPM drives, you can likely successfully serve metadata and data that is combined on the tier because the I/O bottleneck is not likely to track to the read/write heads of the spindles, when evenly distributed and even in heavy I/O situations. The higher the spindle counts, the broader the stripe, and the better the performance.

Purchase tip: It is common that clients buy as small a footprint as possible up front, with plans to grow. However, if substantial growth is expected within 18 months, it might be beneficial to purchase that capacity up front. This choice enables you to take advantage of drive spindle count performance and use new purchase discounts from the beginning of your solution implementation. More is always better, and early performance excitement helps the business community appreciate the purchase decision. To simplify planning, it is best to plan for annual growth in increments of no less than every six months.

When metadata is separated from data in the system pool, it is common to suggest putting metadata in RAID 1, RAID 5, or RAID 10 with GPFS replication. RAID 1 (mirroring), where data is written twice (a small write hit) mirrored to the secondary disk. A typical RAID 1 configuration consists of a minimum of four drives. Two drives are for primary data and two drives are for mirroring. A RAID 1 configuration takes the best concepts of read speed and redundancy, and combines them to provide better read performance along with reliability.

Some clients might argue the value of added redundancy of GPFS replication in this configuration. However, because the value and integrity of metadata is so high, the preferred practice is to use R1 RAID with metadata replication between disk failure groups in GPFS.

A second approach for metadata that is almost as good as the first approach is to use RAID 5 with GPFS replication between two failure groups. This configuration enables for more effective use of space, and a slight read performance compromise. In some cases, the advantage of RAID 1 (R1) is that it does not use the processor to calculate the parity, taking some of the burden off the storage node processor for these small, high-speed transactions.

RAID 1 overview

RAID 1 uses data mirroring. Two physical drives are combined into an array, and data is striped across the array. The first half of a stripe is the original data; the second half of a stripe is a mirror (that is, a copy) of the data, but it is written to the other drive in the RAID 1 array.

RAID 1 provides data redundancy and high levels of performance, but the storage capacity is diminished. Because the data is mirrored, the capacity of the logical drive when assigned RAID 1 is 50% of the array capacity. RAID 1 requires two physical drives. An array is created by using the two physical drives. Then, a logical drive is created within that array. The data is striped across the drives, creating blocks as shown in Figure 4-3.

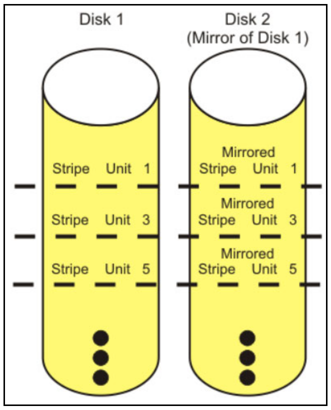


Figure 4-3 RAID 1 layout

Notice that the data on the drive on the right is a copy of the data on the drive on the left. With RAID level-1, if one of the physical drives fails, the controller switches read and write requests to the remaining functional drive in the RAID level-1 array.

RAID 5 overview

RAID 5 stripes data and parity across all drives in the array.

RAID 5 offers both data protection and increased throughput. When you assign RAID 5 to an array, the capacity of the array is reduced by the capacity of one drive (for data-parity storage). RAID 5 gives you higher capacity than RAID 1, but RAID 1 offers better performance.

RAID level-5 requires a minimum of three drives and, depending upon the level of firmware and the stripe-unit size, supports a maximum of 8 or 16 drives.

Figure 4-4 is an example of a RAID level-5 logical drive. The data is striped across the drives, creating blocks. If a physical drive fails in the array, the data from the failed physical drive is reconstructed onto the hot-spare drive.

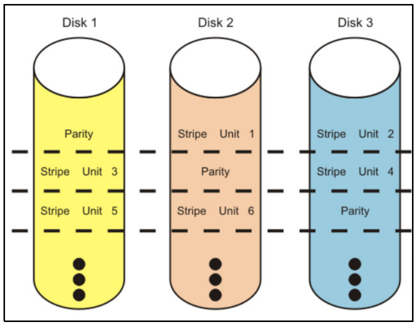


Figure 4-4 RAID5 layout

A parity block contains a representation of the data from the other blocks in the same stripe. For example, the parity block in the first stripe contains data representation of blocks 1 and 2.

RAID10 overview

RAID 10 uses mirrored pairs to redundantly store data. It is often referred to as RAID 1+0 (mirroring and striping). The array must contain an even number of disks. Two is the minimum number of disks that are needed to create a RAID 10 array. The data is striped across the mirrored pairs. RAID 10 tolerates multiple disk failures. If one disk in each mirrored pair fails, the array is still functional, operating in Degraded mode.

You can continue to use the array normally because for each failed disk, the data is stored redundantly on its mirrored pair. However, if both members of a mirrored pair fail, the array is placed in the failed state and is not accessible.

For example, a RAID 10 array of four disks would have data that is written to it in the pattern that is shown in Figure 4-5.

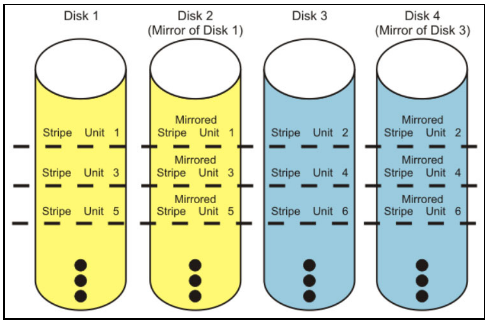


Figure 4-5 RAID10 layout

Metadata preferred practice configuration

To summarize the metadata preferred practice configuration, there are advantages in the following solutions.

Solution #1: When data is stored on NLSAS as the primary tier

When data is stored on NLSAS as the primary tier, perform the following actions:

- Plan for 5% of the file system capacity to be reserved for metadata placement.
- If you are using NLSAS drives for primary data storage, use either SSDs or 15 K RPM SAS drives for metadata isolation.
- Put the metadata on RAID 1, RAID 5, or RAID 10 protected devices (RAID 1 is preferred for performance because it offers the least burden on the processor for stripe management).

  > **Note:** RAID 1 is not supported on the IBM DCS3700 (1818C/2851-DR2).

- Set the GPFS attributes of the metadata disks as `metadataOnly` for Usage Type.
- Place the metadataOnly devices in two GPFS failure groups.
- Create the cluster-mapped storage volumes in increments of four per storage node to evenly balance the I/O ports and channels from storage nodes to subsystem controllers (groups of eight per storage node pair).

  Use **`-R meta`** when you are creating the file system for GPFS to replicate the metadata only between GPFS failure groups.

Solution #2: When data is stored on high-speed SAS as the primary tier

When data is stored on high-speed SAS as the primary tier, perform the following actions:

- Plan for 5% of the file system capacity to be reserved for metadata placement.
- Place the metadata with primary data storage in the GPFS `system` storage pool with the classification on usage type as dataAndMetadata.
- Place all the NSD devices in two GPFS Failure Groups, evenly distributed.
- Create the cluster-mapped storage volumes in increments of four per storage node to evenly balance the I/O ports and channels from storage nodes to subsystem controllers (groups of eight per storage node pair).
- Use **`-R meta`** when you create the file system for GPFS to replicate the metadata only between GPFS failure groups.

> **Note:** When XIV Gen2 or XIV Gen3 storage is used as the back-end device subsystem, due to the differences in RAID protection for XIV, do not isolate metadata for SONAS. Split the NSDs into two failure groups only when multiple XIV subsystems are in the solution. In this case, place the XIV systems into separate failure groups and replicate only file system metadata between them.

Again, the challenge is whether it is better to move data access contention to the drives used for metadata and use only a few spindles for it, or to share the data access types and use an abundance of spindles. Typically, the bigger and slower the drive type that is used for data, the greater the advantage for separating it.

Preferred practice configurations for data placement

Typically, use RAID 6 for data RAID protection.

> **Note:** XIV storage already provides a high-performance and highly reliable proprietary RAID protection scheme for all provisioned storage volumes. RAID 6 in this section refers to IBM Storwize V7000 and System Storage DCS3700.

RAID 6 is similar to RAID 5, but with two sets of parity information rather than one. RAID 6 stripes blocks of data and parity across all drives in the array like RAID 5, but adds a second set of parity information for each block of data.

When you assign RAID 6 to an array, the capacity of the array is reduced for data-parity storage (the exact amount depends on the size of the drives in the array). The second set of parity information is added to improve fault tolerance. RAID 6 can handle two simultaneous drive failures, where other single RAID levels can handle, at most, only one.

RAID 6 requires a minimum of four drives and supports a maximum of 16 drives. The maximum stripe-unit size depends on the number of drives in the array. RAID 6 provides two-drive fault tolerance and adequate performance for most NAS workloads, and it provides a consistent basis for preferred practice data solutions. In most cases, use global hot spares to help diminish risk that is associated with, for example, device failures on long weekends, and so on.

However, it is common that aggressive maintenance plans are chosen over provisioning hot spares with RAID 6 solutions today. The preferred practice is to ensure that hot spare capacity and devices are available in any event of device failure, and that hot sparing effects, functionality, and rebuild are well-understood by the client team that is managing the storage subsystem that is attached to the SONAS gateway devices.

4.2 SONAS supported hardware

SONAS supports various hardware (storage subsystems) and each have points to consider for implementing the storage in preferred practice for performance and reliability. These solutions include the following hardware:

- SONAS Appliance (SONAS with DataDirect Networks (DDN) storage)
- SONAS Gateway with XIV (XIV Gen2 or XIV Gen3)
- SONAS Gateway with Storwize V7000 Storage
- SONAS Gateway with IBM DS8000 Storage
- SONAS Gateway with System Storage DCS3700 Storage (1818 or 2851 feature codes)

As of SONAS 1.4.1, 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 and can be further expanded as an appliance by using DDN storage. The appliance information in this book is for existing installations.

Existing appliances can now also have capacity expansion with a different storage vendor to take advantage of SSD drive technology, using different storage tiers across the vendors. This configuration is offered using a request for price quotation (RPQ), and careful planning of logical disk layout is required.

You can choose to intermix storage across any of the SONAS supported gateway storage vendor types, to provide more flexibility for generating the wanted solution with cost, performance, and reliability characteristics.

This section describes the preferred practice configuration of all of the solutions in the preceding list.

Chapter 5, “File system configuration” on page 141 describes components of preferred practice configuration for all forms of back-end storage. This chapter is designed to cover specifics that are related to the storage itself. Read both chapters for the highest level of review and understanding.

4.2.1 SONAS appliance with DDN storage

For the SONAS appliance with DDN, the storage is automatically configured. Each DDN enclosure consists of a minimum of 60 drives. The solution consists of one or two controllers plus an add-on of 0, 1, or 2 expansion controllers. Figure 4-6 shows a SONAS with DDN storage.

Figure 4-6 SONAS with DDN storage

The DDN storage is automatically configured in 10-drive groupings of RAID 6 configurations that are presenting one volume or GPFS NSD per RAID group to the directly attached SONAS Storage nodes.

Each volume is managed by a preferred controller and subsequent storage node, with active Controller and Storage node failover if a preferred controller or storage node fails.

Each RAID 6 Group is striped with a stripe or segment size of 32 kilobytes (KB), which matches the 32 KB subblock size of a 1 megabyte (MB) file system block size. This configuration means that each subblock write from the SONAS file system is pushed to a single-spindle segment. Therefore, it is safe to conclude that SONAS is optimized for performance on a 1 MB file system block size (by design).

However, when the average file size is small, the preferred practice still warrants the best use of capacity and performance (with 4 x 8 KB writes per segment) when the file system is created with a block size of 256 KB (which is the SONAS default file system block size setting). Figure 4-7 illustrates this principle.

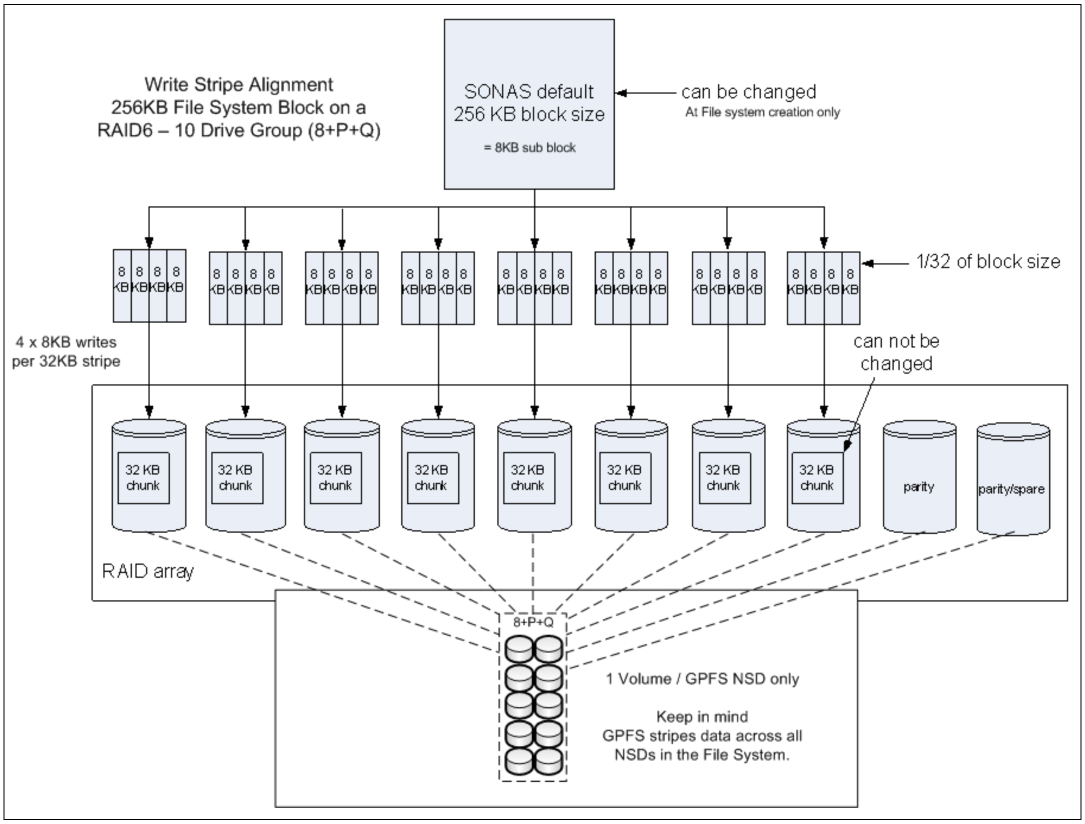


Figure 4-7 Illustration of write alignment with a 256 KB GPFS file system on a RAID 6 (8 + P + Q)

Figure 4-7 illustrates the pattern of data writes as the GPFS write for a 256 KB block size file system is broken into 8 KB writes and submitted to the NSD in alignment with four subblock writes per NSD Stripe of 32 KB.

Therefore, any write of 8 KB or less is submitted to a single stripe on an NSD in a single write. Any write of 16 KB or up to 32 KB is written to a single NSD in increments of 8 KB subblock writes. A write crosses over to another NSD only if the number of subblocks that are written exceed the stripe size and do so in 8 KB file chunk sizes.

The smallest capacity a single file uses in the file system is one subblock. Therefore, a 1 KB file uses a minimum of 8 KB, and a 34 KB file uses 40 KB (32 KB + 8 KB).

Small files are best-suited on a 256 KB block size, because less capacity is wasted on file size minimums, more files can be held in cache, and smaller chunks can be read back more quickly. However, larger files suffer read/write hits as the pieces are pulled together or assembled for a single read or write operation. For file systems with predominantly larger file sizes, it might be best to use a 1 MB block size (as shown in Figure 4-8).

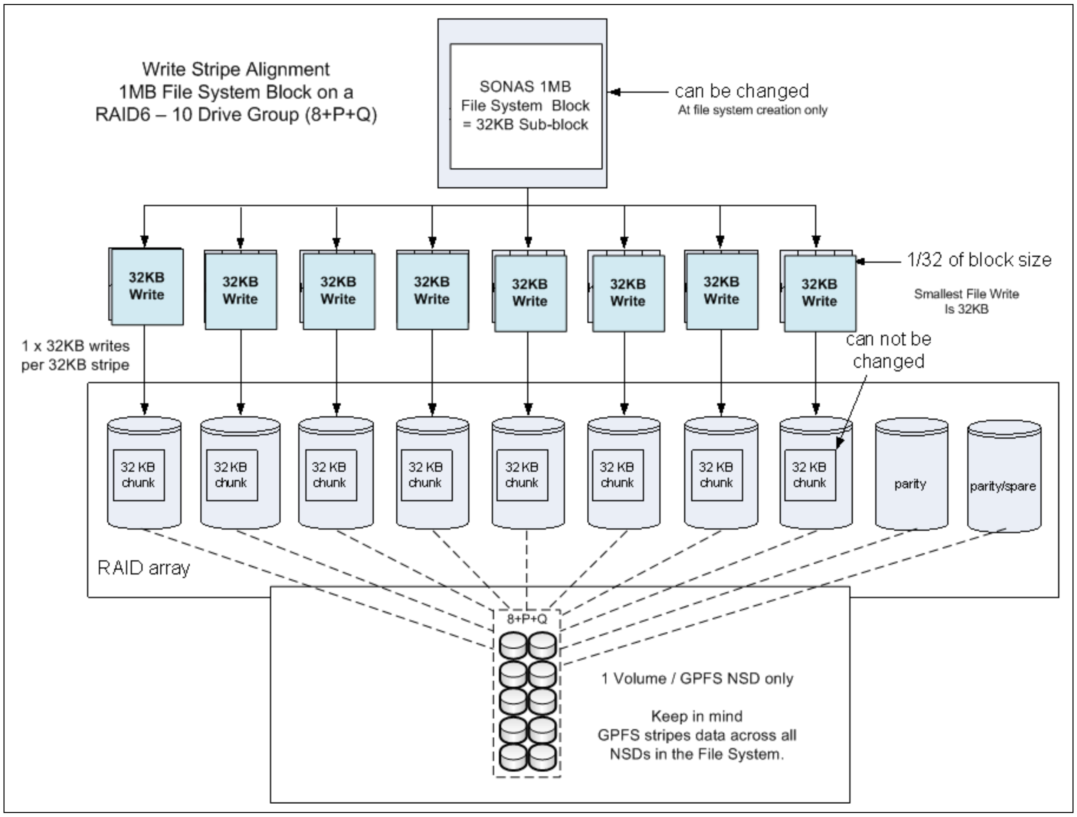


Figure 4-8 Write alignment with a 1 MB GPFS file system on a RAID 6 (8 + P + Q)

Figure 4-8 illustrates the pattern of data writes as the GPFS write for a 1 MB block size file system is broken into 32 KB writes and submitted to the NSD in alignment with one subblock write per NSD stripe of 32 KB.

Therefore, any write of 32 KB or less is submitted to a single stripe on an NSD in a single write. Any write of 32.1 KB or larger is written to more NSDs in increments of 32 KB subblock writes. A write crosses over to another NSD only if the number of subblocks that are written exceed the stripe size and it does so in 32 KB file chunk sizes.

The smallest capacity a single file uses in the file system is one subblock. Therefore, a 1 KB file uses a minimum of 32 KB, and a 34 KB file uses 64 KB (32 KB + 32 KB).

Small files are best suited on a 256 KB block size. Because less capacity is wasted on file size minimums, more files can be held in cache, and smaller chunks can be read back more quickly. However, larger files suffer read/write hits as the pieces are pulled together or assembled for a single read or write operation. For file systems with predominantly larger file sizes, it might be best to use a 1 MB block size (as shown in Figure 4-8).

SONAS and GPFS also support a 4 MB file system block size (shown in Figure 4-9). However, it is only efficient for managing file systems where most or all files are well over 128 KB. Capacity can be quickly lost with large block file systems if lots of small files are written to it.

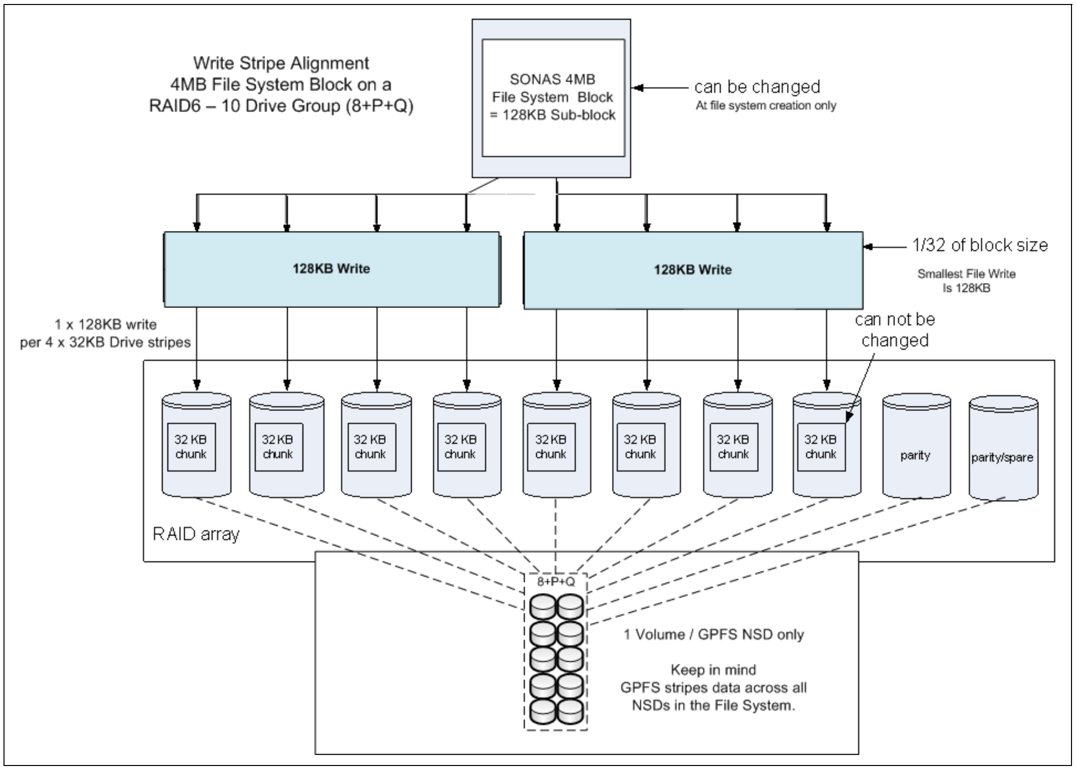


Figure 4-9 Illustration of write alignment with a 4 MB GPFS file system on a RAID 6 (8+P+Q)

Figure 4-9 illustrates the pattern of data writes as the GPFS write for a 4 MB block size file system is broken up into 128 KB writes and submitted to the NSD in alignment with 1 subblock write per NSD stripe of 128 KB.

Therefore, any write of 128 KB or less is submitted to four RAID 6 chunk stripes across four NSDs in a single write pattern. Any write of 128 KB or up is written to more NSDs in increments of 128 KB subblock writes.

The smallest capacity a single file uses in the file system is one subblock. A 1 KB file uses a minimum of 128 KB, and 130 KB file uses 256 KB (128 KB + 128 KB). There are few cases where setting a 4 MB file system block size is advised for general-purpose scale out NAS.

The DDN storage is managed using redundant Ethernet communication ports between the SONAS storage nodes and the DDN storage controller devices. However, modification of the DDN configuration is not authorized without explicit approval from SONAS support.

DDN service and firmware updates are managed directly through SONAS support engagements and code level upgrades (integrated in the SONAS software stack).

4.2.2 SONAS gateway with XIV (FC 9006)

For the SONAS gateway with XIV, the storage is manually configured.

In the *gateway* configurations, it is critical to understand that the storage subsystems are a client-managed solution. There are SONAS product guidelines, and key use and configuration restrictions, for each storage subsystem option. However, these storage subsystems are to be managed, monitored, and updated by the client support team independently from the SONAS solution support group.

It is important to coordinate and plan such activities between both the NAS solutions support team and the storage solution support team, to ensure that your actions do not conflict with either brand’s team advice or recommendations.

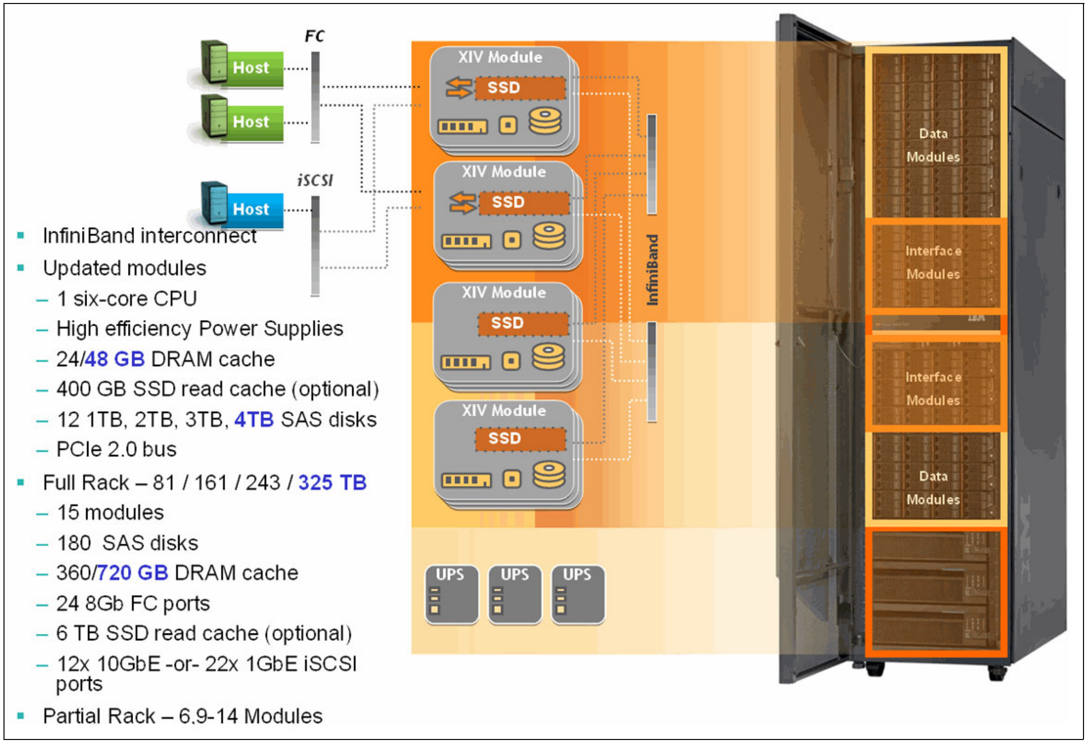
Each XIV subsystem consists of 9 - 15 modules.

Note: Six-module XIV configurations exist. However, do not run a SONAS production system with only a six-module XIV because the fabric zoning is complicated in expansion (to scale out from there) and performance is extremely limited. Use the six-module configurations only for demonstration centers and test labs (behind SONAS), or extremely small solutions with low performance requirements.

The XIV solution is not a typical dual-controller storage system, and works differently from all other supported storage backend solutions in SONAS. SONAS with XIV Gen3 is typically the highest-performance solution available. With XIV Gen2 or XIV Gen3, the XIV solution behind SONAS offers the highest performance persistence during component failure, and the fastest recovery to full redundancy in the industry.

SONAS with XIV is the preferred practice configuration for clients who need persistent high performance for largely sequential workloads and a simplified storage subsystem. It is a client favorite when reliability and performance persistence are most important.

XIV storage is not configurable from a RAID protection scheme perspective. It is a one size fits all solution that is optimized for ease of use, high performance, and high reliability (see Figure 4-10).



- 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 4-10 The SONAS XIV supports partial to full 15-module XIV configurations

XIV is a virtualized storage infrastructure where all data is striped across all the spindles in 1 MB chunks, and hot spare capacity is also spread across all devices. The XIV is connected to the SONAS with a Switch fabric to use access to storage through multiple advanced technology controllers called *interface modules*. This configuration provides extreme reliability, and provides up to 12 paths per volume device.

XIV Gen3 with SONAS is the best performing, most reliable of the SONAS solutions for most typical workloads. It offers high performance and reliability even during component failure, and works automatically at the quickest possible rate to return the solution to full redundancy.

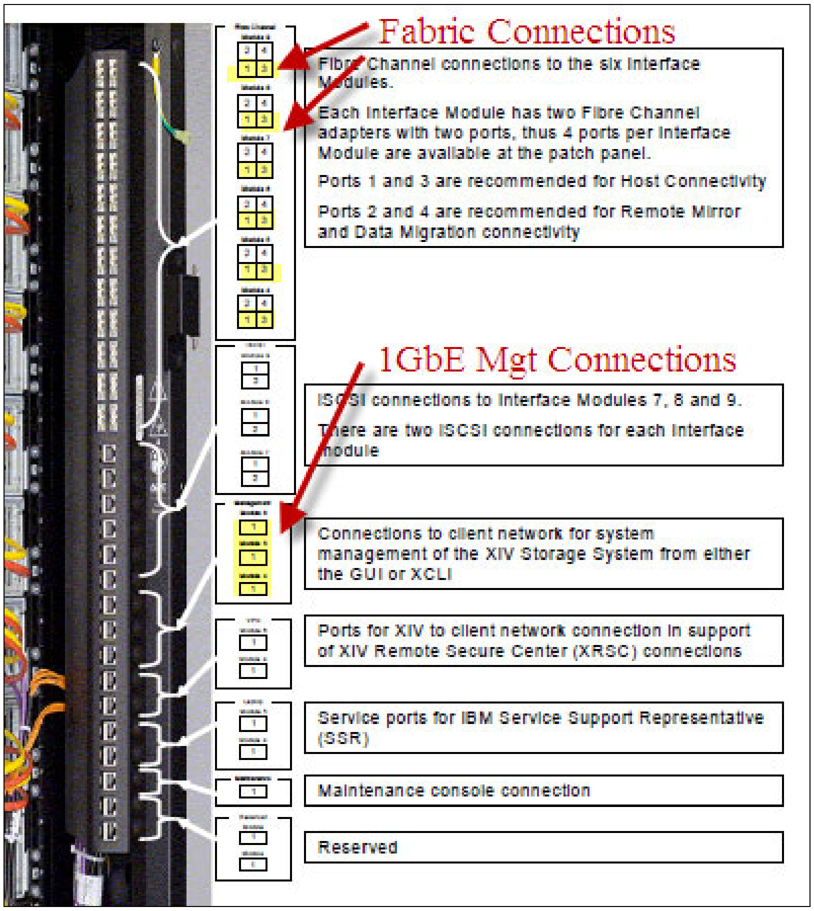
Because the XIV is so highly reliable, it does not break a SONAS cluster configuration into multiple failure groups where a single XIV subsystem is used. The consistency of the solution already provides high reliability. The entire frame works as a single proprietary type of RAID device, so much of the benefit of splitting synchronous replication across failure groups is lost with a single frame of XIV.

However, when multiple XIVs of the same type are stacked in your solution, it might be an added layer of protection to place each of the XIVs in a separate failure group and provide GPFS file system metadata replication in your preferred practice plan.

SONAS gateway with XIV fabric connection

The XIV must be fabric-connected to the SONAS, and redundant fabric is required for high availability (HA). It is a preferred practice to spread the XIV and Storage node connections across fabric switch gigabit interface converters (GBICs) and application-specific integrated circuits (ASICs). If possible, connect the XIV using port 1 and port 3 to ensure that the ports that are used on the XIV are not on the same HBA, Port, or ASIC.

This configuration means that port 1 of each module is connected to Fibre Channel (FC) Switch 1 and port 3 is connected to switch 2. Likewise, each HBA on each storage node is also connected across both switches (HBA 1 port 1 to switch 1, HBA 1 port 2 to switch 2, and so on). Figure 4-11 shows example XIV port connections.



Fabric Connections

Fibre Channel connections to the six Interface Modules.

Each Interface Module has two Fibre Channel adapters with two ports, thus 4 ports per Interface Module are available at the patch panel.

Ports 1 and 3 are recommended for Host Connectivity

Ports 2 and 4 are recommended for Remote Mirror and Data Migration connectivity

1GbE Mgt Connections

iSCSI connections to Interface Modules 7, 8 and 9.

There are two iSCSI connections for each Interface module

Connections to client network for system management of the XIV Storage System from either the GUI or XCLI

Ports for XIV to client network connection in support of XIV Remote Secure Center (XRSC) connections

Service ports for IBM Service Support Representative (SSR)

Maintenance console connection

Reserved

Figure 4-11 XIV example port connections

Zoning

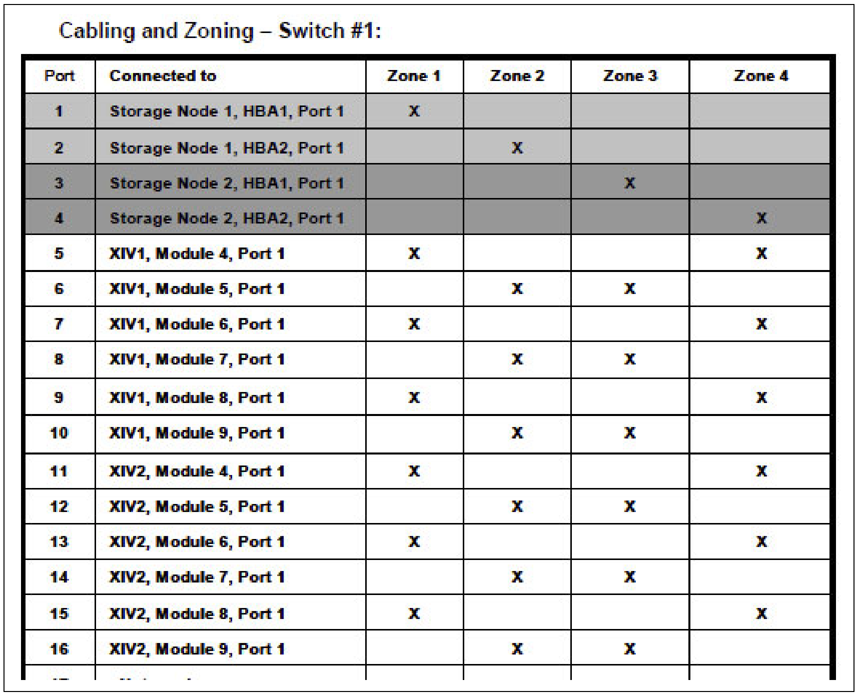
Zone the XIVs for 12 paths per volume. This zoning is done by using single-initiator zones and mapping only three XIV module target ports per zone.

The zoning of an XIV behind SONAS and the file system configurations are important for achieving preferred-practice performance. The size of the volume is less important. However, XIV performs much better with fewer, larger volumes than the traditional strategy of using many smaller volumes. That being said, the ideal volume size is typically 1 - 8 TB for best SONAS file system performance configurations. Tested results show that keeping volume quantities below 40 per storage node pair reduces resource use.

In addition, stacking volumes in multiples of four behind active file systems (per storage node) offers the best striping performance through the HBA ports of each node. The size 4 TB is a good target for most systems. Field standards typically include installers configuring the XIV volumes in 4 TB increments.

Figure 4-12 and Figure 4-13 on page 112 show XIV switch zone guides.

Tip: Too many volumes and too many paths add resource use to the already evenly distributed and heavily worked storage nodes.

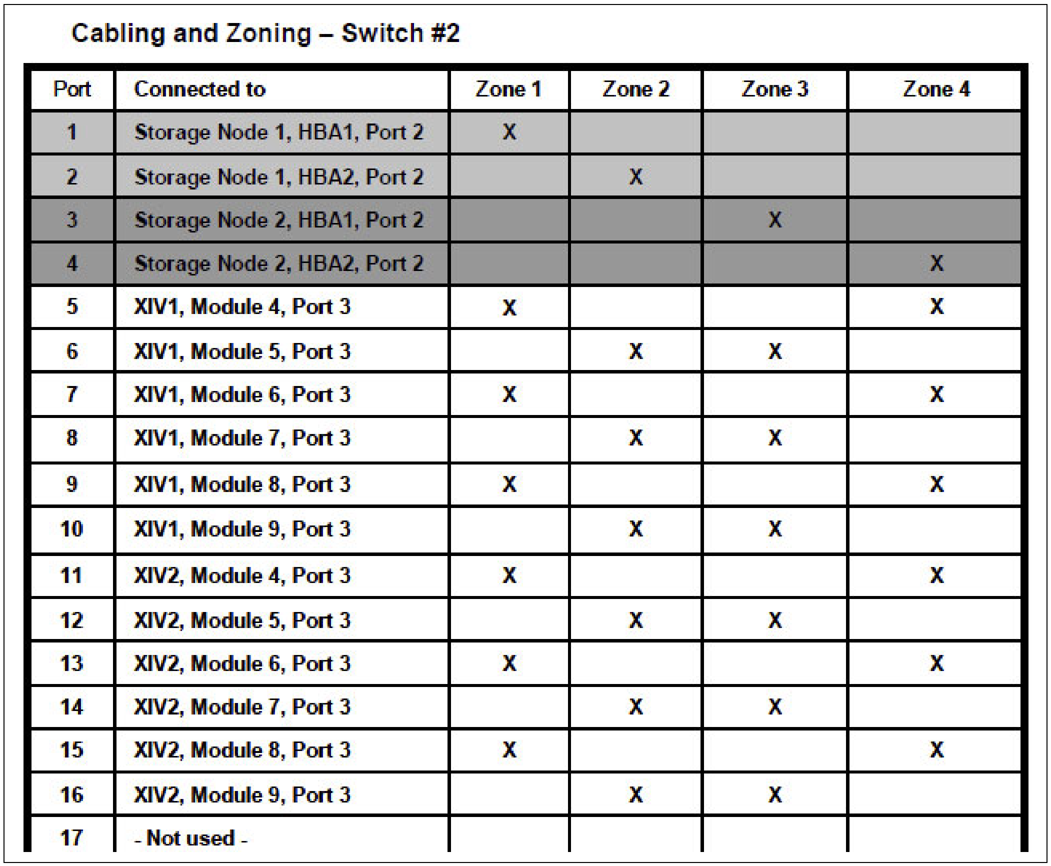


Cabling and Zoning – Switch #1:

Port	Connected to	Zone 1	Zone 2	Zone 3	Zone 4
1	Storage Node 1, HBA1, Port 1	X			
2	Storage Node 1, HBA2, Port 1		X		
3	Storage Node 2, HBA1, Port 1			X	
4	Storage Node 2, HBA2, Port 1				X
5	XIV1, Module 4, Port 1	X			X
6	XIV1, Module 5, Port 1		X	X	
7	XIV1, Module 6, Port 1	X			X
8	XIV1, Module 7, Port 1		X	X	
9	XIV1, Module 8, Port 1	X			X
10	XIV1, Module 9, Port 1		X	X	
11	XIV2, Module 4, Port 1	X			X
12	XIV2, Module 5, Port 1		X	X	
13	XIV2, Module 6, Port 1	X			X
14	XIV2, Module 7, Port 1		X	X	
15	XIV2, Module 8, Port 1	X			X
16	XIV2, Module 9, Port 1		X	X	

Figure 4-12 XIV Switch 1 preferred practice zones guide

Figure 4-13 shows the zone guide for Switch 2.



Cabling and Zoning – Switch #2

Port	Connected to	Zone 1	Zone 2	Zone 3	Zone 4
1	Storage Node 1, HBA1, Port 2	X			
2	Storage Node 1, HBA2, Port 2		X		
3	Storage Node 2, HBA1, Port 2			X	
4	Storage Node 2, HBA2, Port 2				X
5	XIV1, Module 4, Port 3	X			X
6	XIV1, Module 5, Port 3		X	X	
7	XIV1, Module 6, Port 3	X			X
8	XIV1, Module 7, Port 3		X	X	
9	XIV1, Module 8, Port 3	X			X
10	XIV1, Module 9, Port 3		X	X	
11	XIV2, Module 4, Port 3	X			X
12	XIV2, Module 5, Port 3		X	X	
13	XIV2, Module 6, Port 3	X			X
14	XIV2, Module 7, Port 3		X	X	
15	XIV2, Module 8, Port 3	X			X
16	XIV2, Module 9, Port 3		X	X	
17	- Not used -				

Figure 4-13 XIV Switch 2 preferred practice zones guide

XIV Gen2 versus XIV Gen3 comparison

There are some key differences to understand and acknowledge in the XIV Gen2 versus the XIV Gen3 subsystems. Key differences between the Gen2 and Gen3 subsystems are seen in the following areas:

- Performance

  In some cases, the XIV Gen3 storage can be three to four times faster than the Gen2 storage.

- Capacity

  Capacity is significantly increased for data and managed cache in Gen3 subsystems.

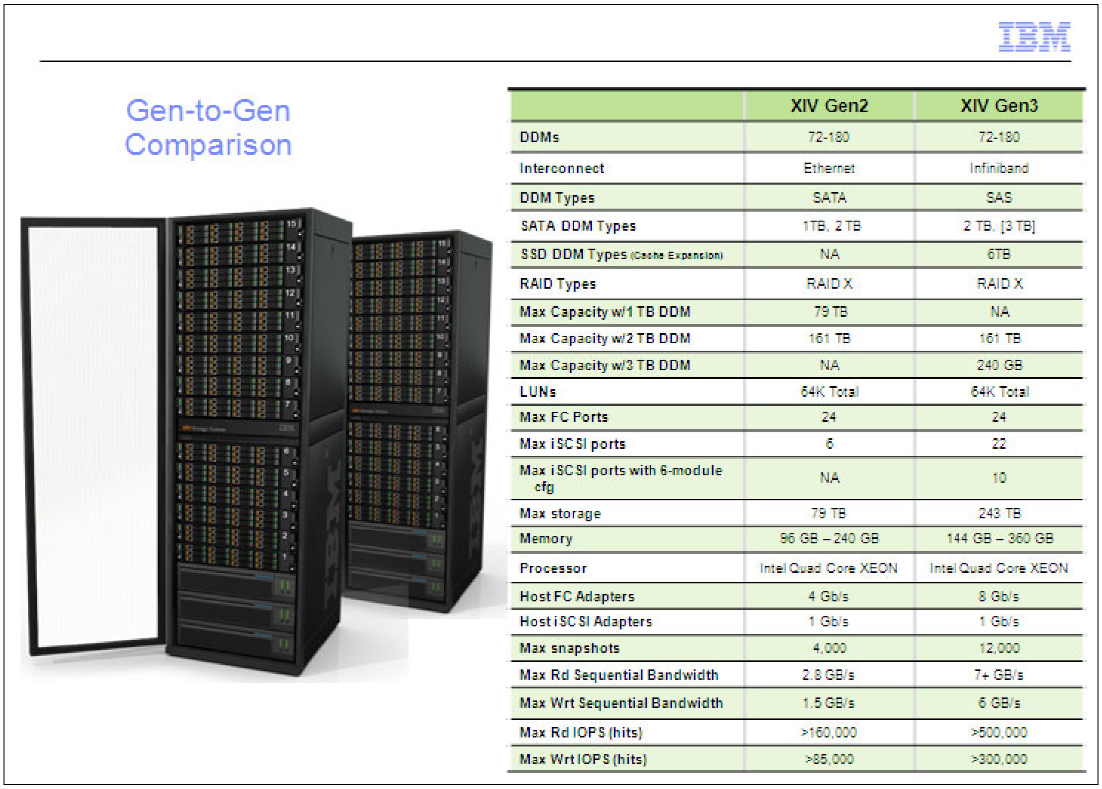
- Volume size

  The volume structure is slightly different in Gen3, which puts the volume sizing at a different offset. For this reason and others, it is only supported to mix Gen2 and Gen3 storage in SONAS when they are behind separate storage node pairs, and their volumes are in different storage pools (tiers).

- SSD-enabled performance enhancement

  SSD-enabled XIV (optional in Gen3 only) enables one SSD per XIV storage module, and it is used exclusively to improve storage cache read and data prefetch operations. SSD enablement can in many cases significantly improve performance for random access workloads.

Figure 4-14 shows differences between the XIV Gen3 and XIV Gen4 technologies.



Gen-to-Gen Comparison

	XIV Gen2	XIV Gen3
DDMs	72-180	72-180
Interconnect	Ethernet	Infiniband
DDM Types	SATA	SAS
SATA DDM Types	1TB, 2 TB	2 TB, [3 TB]
SSD DDM Types (Cache Expansion)	NA	6TB
RAID Types	RAID X	RAID X
Max Capacity w/1 TB DDM	79 TB	NA
Max Capacity w/2 TB DDM	161 TB	161 TB
Max Capacity w/3 TB DDM	NA	240 GB
LUNs	64K Total	64K Total
Max FC Ports	24	24
Max iSCSI ports	6	22
Max iSCSI ports with 6-module cfg	NA	10
Max storage	79 TB	243 TB
Memory	96 GB – 240 GB	144 GB – 360 GB
Processor	Intel Quad Core XEON	Intel Quad Core XEON
Host FC Adapters	4 Gb/s	8 Gb/s
Host iSCSI Adapters	1 Gb/s	1 Gb/s
Max snapshots	4,000	12,000
Max Rd Sequential Bandwidth	2.8 GB/s	7+ GB/s
Max Wrt Sequential Bandwidth	1.5 GB/s	6 GB/s
Max Rd IOPS (hits)	>160,000	>500,000
Max Wrt IOPS (hits)	>85,000	>300,000

Figure 4-14 Suggested differences between XIV generation technologies

Important: The performance information in Figure 4-14 is based on the block storage test results in a lab that was not specifically designed for benchmark performance anticipation. It is important to understand that the numbers do not relate to SONAS implementation. They are presented to show relative differences in XIV design only.

Note: The XIV Gen2 subsystem is no longer available. Support for the XIV Gen2 will continue to be provided until 2019.

Both generations support massive parallelism. The system architecture ensures full use of all system components. Any I/O activity that involves a specific logical volume in the system is always inherently handled by all spindles.

The system harnesses all storage capacity and all internal bandwidth, and it takes advantage of all available processing power. These benefits are for both host-initiated I/O activity and system-initiated activity, such as rebuild processes and snapshot generation. All disks, processors, switches, and other components of the system contribute to the performance of the system at all times.

In either case, because you cannot change the RAID or storage stripe width, there is less to consider with XIV. The two typical file system block sizes are 256 KB or 1 MB with XIV storage on the back-end.

When you choose the file system block size and allocation type on SONAS, follow the same guidelines as with DDN. However, with a single XIV, do not split the volumes (NSDs) into multiple failure groups. If there is a second XIV attached to the SONAS, use metadata-only replication.

As the XIV Gen3 subsystem has potential for high performance, it is understood that two SONAS storage nodes cannot drive the full potential performance from a single XIV Gen3 subsystem. In this case, clients might decide to span four SONAS storage nodes across one XIV Gen3 subsystem. In this case, the storage nodes are still added to SONAS as a storage node pair.

However, the XIV adds the storage nodes as two separate SONAS clusters of two Red Hat Enterprise Linux (RHEL) servers in each cluster, and half of the XIV volumes are mapped to the first storage node pair cluster, where the other half is mapped to the second storage node pair cluster. Use this configuration only for XIV Gen3 or DS8000 gateways where extreme performance is demanded from SONAS in smaller capacities.

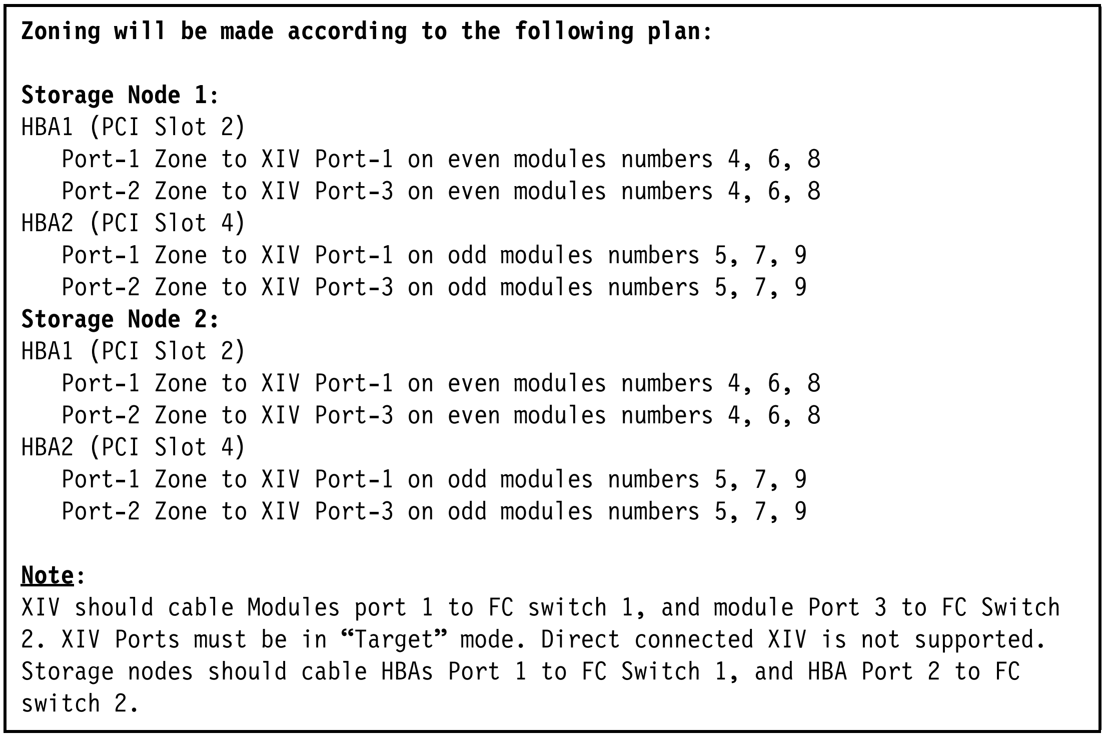
SONAS supports up to two XIV subsystems behind a single storage node pair and in some configurations where capacity is more important than performance, this configuration does not present a problem. However, the two XIVs can provide a huge performance improvement (along with capacity improvement) if they are placed behind separate storage node pairs.

When you put two XIVs behind the same storage node pair, do not place two different generation XIV subsystems (Gen2 and Gen3) behind the same storage node pair. XIV Gen2 operates with 4 Gb FC HBA port connections and XIV Gen3 operates with 8 Gb FC HBA port connections. Do not mix both on the same storage node ports through switch zoning. Volume sizes are also slightly different between XIV Gen2 and XIV Gen3. Therefore, because of capacity and performance difference between the two, do not mix the volumes into the same disk storage pools under SONAS.

In the preceding configuration, provision all volumes for SONAS evenly across both storage node pairs and provision all volumes up front and in even distribution for the highest possible performance and parallelism.

By design, the SONAS guide suggests mapping 4 TB volumes. It is more important to understand that volumes from XIV should be provisioned 1 - 8 TB, and that it is helpful to have an even distribution and like size, with a minimum of four volumes per storage node and a maximum of 20 volumes per storage node or 40 per storage node pair.

In a preferred practice configuration, these volumes have 12 paths per volume (NSD). Figure 4-15 shows the SONAS storage node to XIV module zoning.



```
Zoning will be made according to the following plan:

Storage Node 1:
HBA1 (PCI Slot 2)
   Port-1 Zone to XIV Port-1 on even modules numbers 4, 6, 8
   Port-2 Zone to XIV Port-3 on even modules numbers 4, 6, 8
HBA2 (PCI Slot 4)
   Port-1 Zone to XIV Port-1 on odd modules numbers 5, 7, 9
   Port-2 Zone to XIV Port-3 on odd modules numbers 5, 7, 9
Storage Node 2:
HBA1 (PCI Slot 2)
   Port-1 Zone to XIV Port-1 on even modules numbers 4, 6, 8
   Port-2 Zone to XIV Port-3 on even modules numbers 4, 6, 8
HBA2 (PCI Slot 4)
   Port-1 Zone to XIV Port-1 on odd modules numbers 5, 7, 9
   Port-2 Zone to XIV Port-3 on odd modules numbers 5, 7, 9

Note:
XIV should cable Modules port 1 to FC switch 1, and module Port 3 to FC Switch
2. XIV Ports must be in “Target” mode. Direct connected XIV is not supported.
Storage nodes should cable HBAs Port 1 to FC Switch 1, and HBA Port 2 to FC
switch 2.
```

Figure 4-15 SONAS Storage Node to XIV Module Zoning Guide

The following features of external storage systems are *not* supported when used with the IBM SONAS product:

- FlashCopy and volume copy of SONAS volumes
- XIV snapshots of SONAS volumes
- XIV based remote volume replication of SONAS volumes
- XIV based thin provisioning of SONAS volumes

The external storage system is configured, monitored, managed, serviced, and supported independent of the SONAS gateway system. There is no storage-subsystem-related integration into the SONAS system monitoring, reporting, alerting software stack that does not relate specifically to the port status of the connected storage nodes or the mapped logical unit number (LUN) devices.

Code and firmware levels

At the time of publication, the best supported XIV code level for use behind SONAS is 10.2.4e for XIV Gen2, and 11.3 for XIV Gen3. Advise your IBM service representative or service support representative (SSR) to ensure that this code is enabled on the XIV subsystems before you install SONAS. Check each SONAS release for the proper XIV firmware that was tested with the solution to get a suggested version for initial installation or to upgrade from a previous installation.

XIV storage preferred practice summary

The following items are preferred practices for XIV storage:

- Place one XIV subsystem behind each storage node pair for best performance to capacity scaling.
- For maximum performance, place each XIV Gen3 subsystem behind four storage nodes rather than two nodes. For maximum performance, place each XIV Gen2 subsystem behind two storage nodes. For maximum capacity, place two XIV subsystems behind each storage node set.
- To achieve the highest performance of small file random access workloads (such as metadata scans), add SSD card options to the XIV Gen3 system modules.
- Never mix XIV Gen2 and XIV Gen3 subsystems behind the same storage node pair.
- There is no benefit to separating data and metadata volumes on the XIV platform, unless both XIV Gen3 and XIV Gen2 subsystems are in the same cluster. In that case, make sure that the XIV Gen3 subsystem is in the system pool and metadata placement is only on XIV Gen3 volumes.
- Never mix XIV Gen2 and XIV Gen3 volumes in the same disk storage pool.
- Zone the XIV subsystem with 12 paths per volume by using single-initiator zoning on the switches (one initiator port per switch zone).
- Use port 1 and port 3 or port 2 and port 4, in target mode, on the XIV subsystems for Fibre Channel (FC) connectivity to the storage nodes.
- Zone two ports per XIV module with FC Switch redundancy between ports.
- Use large and consistent XIV volume sizes for the XIV subsystems. Four TB or 8 TB provide excellent performance and manageability.
- All LUNs attached to SONAS must not be usable or seen by any other host.
- Build the GPFS file system at 256 KB block size with the scatter allocation type for small file workloads, and 1 MB block size with scatter allocation type for large file sequential workloads.
- Review status on both SONAS and XIV storage events daily, and manage each as separate storage solutions that are working together to drive your requirements.
- Use nine or more XIV module configurations behind SONAS for production workloads with special exceptions only for demonstration, lab, or test environments.

4.2.3 SONAS gateway with Storwize V7000 (FC 9007)

The SONAS gateway with Storwize V7000 back-end storage is manually configured.

The SONAS gateway configuration with Storwize V7000, which is no longer an iRPQ-only solution (as of SONAS 1.3.2), is now listed as a feature code option for ordering the SONAS gateway. Storwize V7000 gateways are feature code 9007.

Much like the XIV (feature code 9006), you can stack one or two V7000 controllers behind each SONAS storage pod. The 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 using V7000 storage.

The SONAS 1.5.1 release supports V7000 Gen 2 hardware. The V7000 Gen 2 has nearly double the capacity and performance characteristics from the Gen 1 platform. The highest performance flexibility comes from SONAS V7000 gateway solutions.

This solution offers the greatest flexibility in disk speeds, types, and drive size options in the IBM storage catalog. It enables for flexible storage tiering with the storage pod between SSD, SAS, and Near Line SAS (NLSAS) storage options.

The V7000 storage solution does not support external storage devices for use with SONAS. Only local (internal) disks are to be used for SONAS shares. This restriction means that only devices in enclosures that are directly SAS-connected to the V7000 controllers can be used for SONAS disk (NSD) devices.

The Storwize V7000 can optionally be storage area network (SAN)-attached or direct-fiber-connected to the SONAS (as a gateway solution). Redundant fabric is required for high availability if it is SAN-attached. It is a preferred practice to spread the V7000 and SONAS Storage node connections across Fabric switch ASICs. If possible, connect the Storwize V7000 by using port 1 or port 2 and port 3 to ensure that the ports used on the Storwize V7000 are not on the same HBA, port, HCA, or ASIC.

Each Storwize V7000 must be dedicated to SONAS use (not Unified), and must not be clustered with other Storwize V7000 storage for virtualized storage sharing.

The following features of external storage systems are *not* supported when used with the IBM SONAS product:

- IBM FlashCopy and Volume Copy
- Remote Volume Mirroring (RVM)
- RAID 0, 1, 3, and 5
- Thin provisioning of SONAS volumes
- IBM Real-time Compression™

The external storage system is configured, monitored, managed, serviced, and supported independent of the SONAS gateway system. There are no storage subsystem-related integrations into the SONAS system monitoring, reporting, alerting software stack that does not relate specifically to the port status of the connected storage nodes or the mapped LUN devices.

Note: For the most current IBM Storwize V7000 configuration information, see the following information in the IBM Knowledge Center:

- IBM Scale Out Network Attached Storage (SONAS) 1.5.2 product documentation:

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

- IBM Storwize V7000 welcome page:

  http://www.ibm.com/support/knowledgecenter/ST3FR7/welcome

Figure 4-16 is the rear view of the Storwize V7000 controller.

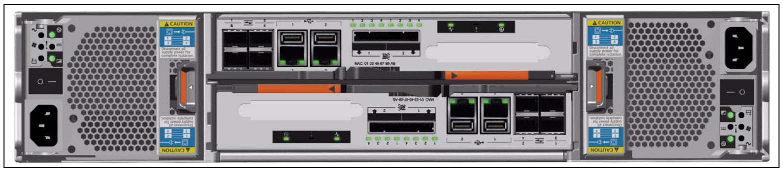


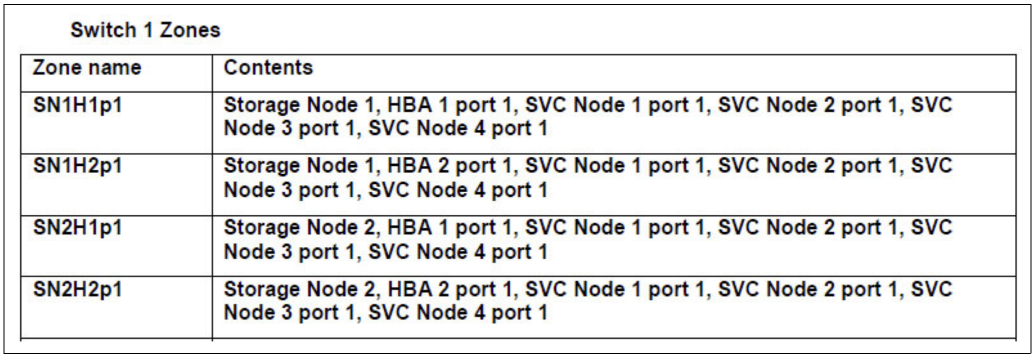
Figure 4-16 Storwize V7000 controller: rear view image

Zoning considerations

The Storwize V7000 can be FC switch-connected, and switches must be mounted in a customer-supplied rack. They cannot be mounted in the SONAS rack. Only single-initiator zoning is supported, and switch redundancy is required for SONAS configurations.

Zoning is made according to the following plan. In the following plan, SAN Volume Controller nodes can be substituted with V7000 Controller Canisters, such that SAN Volume Controller Node 1 = V7000 #1 Controller #1 (“upper”), SAN Volume Controller Node 2 = V7000 #1 Controller #2 (“lower”), SAN Volume Controller Node 3 = V7000 #2 Controller #1 (“upper”), and SAN Volume Controller Node 4 = V7000 #2 Controller #2 (“lower”).

Figure 4-17 shows Switch 1 zoning guidelines.

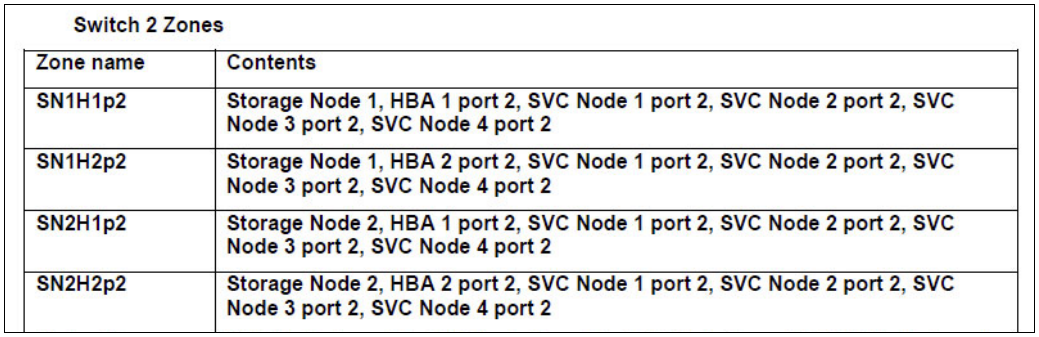


Switch 1 Zones

Zone name	Contents
SN1H1p1	Storage Node 1, HBA 1 port 1, SVC Node 1 port 1, SVC Node 2 port 1, SVC Node 3 port 1, SVC Node 4 port 1
SN1H2p1	Storage Node 1, HBA 2 port 1, SVC Node 1 port 1, SVC Node 2 port 1, SVC Node 3 port 1, SVC Node 4 port 1
SN2H1p1	Storage Node 2, HBA 1 port 1, SVC Node 1 port 1, SVC Node 2 port 1, SVC Node 3 port 1, SVC Node 4 port 1
SN2H2p1	Storage Node 2, HBA 2 port 1, SVC Node 1 port 1, SVC Node 2 port 1, SVC Node 3 port 1, SVC Node 4 port 1

Figure 4-17 V7000 gateway Preferred Practice Zoning Guide: Switch 1

Figure 4-18 shows Switch 2 zoning guidelines.



Switch 2 Zones

Zone name	Contents
SN1H1p2	Storage Node 1, HBA 1 port 2, SVC Node 1 port 2, SVC Node 2 port 2, SVC Node 3 port 2, SVC Node 4 port 2
SN1H2p2	Storage Node 1, HBA 2 port 2, SVC Node 1 port 2, SVC Node 2 port 2, SVC Node 3 port 2, SVC Node 4 port 2
SN2H1p2	Storage Node 2, HBA 1 port 2, SVC Node 1 port 2, SVC Node 2 port 2, SVC Node 3 port 2, SVC Node 4 port 2
SN2H2p2	Storage Node 2, HBA 2 port 2, SVC Node 1 port 2, SVC Node 2 port 2, SVC Node 3 port 2, SVC Node 4 port 2

Figure 4-18 V7000 gateway Preferred Practice Zoning Guide: Switch 2

Cabling preferred practices

The diagram in Figure 4-19 depicts the Fibre Channel and SAS cabling for the case when the first Storwize V7000 is direct-attached to a pair of SONAS storage nodes (2851-SSx). V7000 #1 is connected to port 2 of the HBAs, and V7000 #2 is connected to port 1 of the HBAs. These connections match the way DDN is cabled in SONAS appliances, which is reflected in parts of the RAS code. On each HBA, port 2 is on the left and port 1 is on the right.

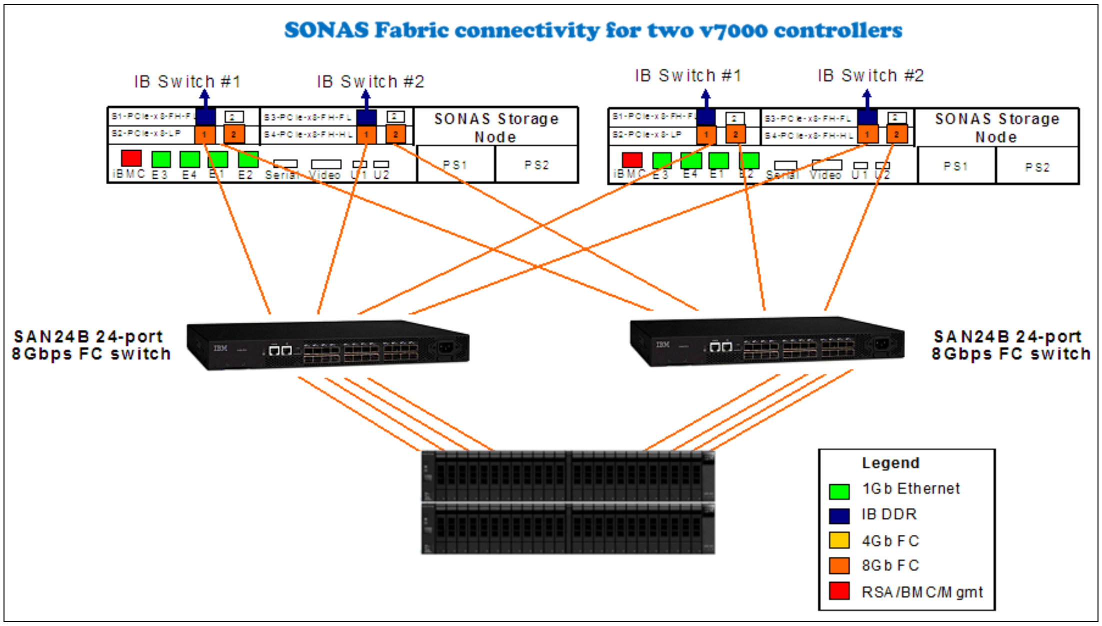

SONAS Fabric connectivity for two v7000 controllers
IB Switch #1
IB Switch #2
IB Switch #1
IB Switch #2
SONAS Storage Node
SONAS Storage Node
iBMC E3 E4 E1 E2 Serial Video U1 U2
PS1
PS2
SAN24B 24-port 8Gbps FC switch
SAN24B 24-port 8Gbps FC switch
Legend
1Gb Ethernet
IB DDR
4Gb FC
8Gb FC
RSA/BMC/Mgmt

Figure 4-19 Fabric overview (4 paths per volume)

Figure 4-20 shows cabling for a single Storwize V7000 and SONAS.

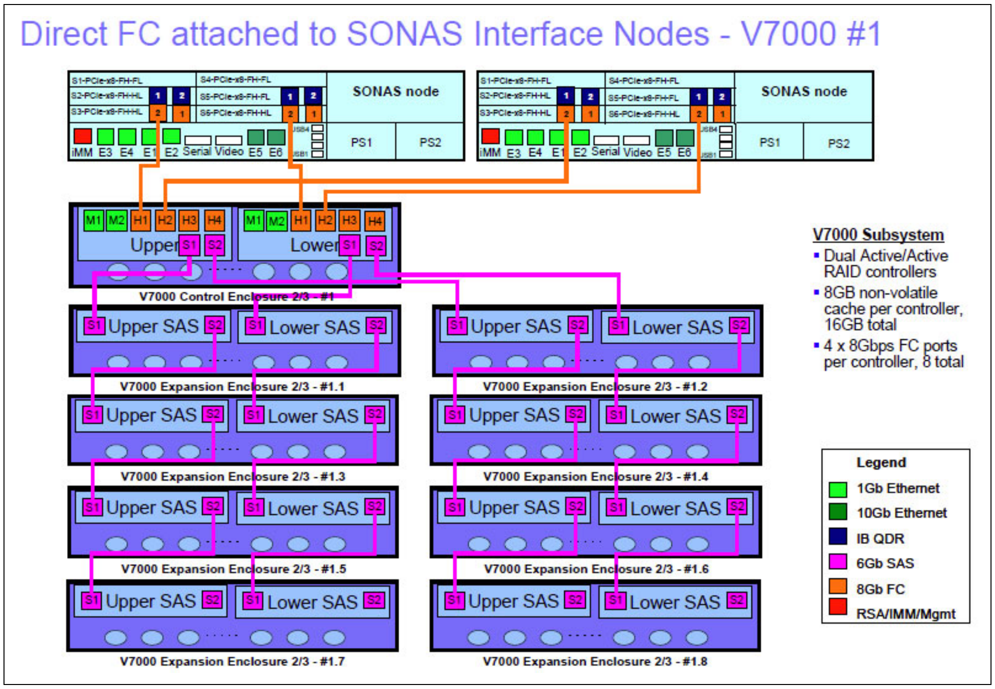
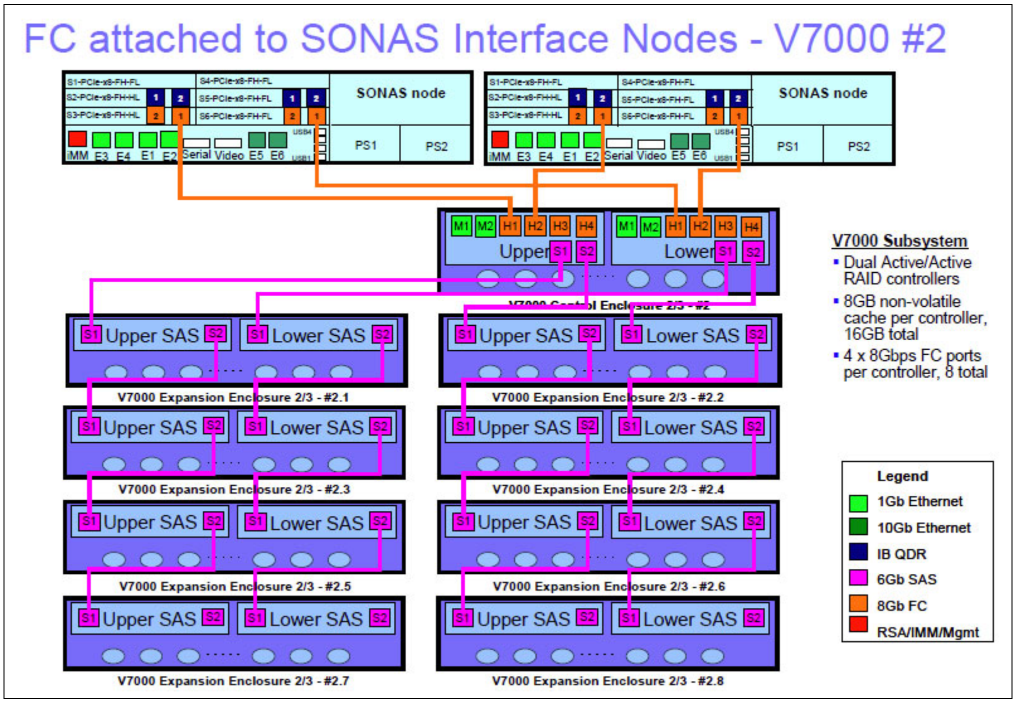

Direct FC attached to SONAS Interface Nodes - V7000 #1
SONAS node
SONAS node
iMM E3 E4 E1 E2 Serial Video E5 E6
PS1
PS2
M1 M2 H1 H2 H3 H4
Upper S1 S2
Lower S1 S2
V7000 Control Enclosure 2/3 - #1
S1 Upper SAS S2
S1 Lower SAS S2
V7000 Expansion Enclosure 2/3 - #1.1
V7000 Expansion Enclosure 2/3 - #1.2
V7000 Expansion Enclosure 2/3 - #1.3
V7000 Expansion Enclosure 2/3 - #1.4
V7000 Expansion Enclosure 2/3 - #1.5
V7000 Expansion Enclosure 2/3 - #1.6
V7000 Expansion Enclosure 2/3 - #1.7
V7000 Expansion Enclosure 2/3 - #1.8
V7000 Subsystem
Dual Active/Active RAID controllers
8GB non-volatile cache per controller, 16GB total
4 x 8Gbps FC ports per controller, 8 total
Legend
1Gb Ethernet
10Gb Ethernet
IB QDR
6Gb SAS
8Gb FC
RSA/IMM/Mgmt

Figure 4-20 SIngle V7000 cabling

Figure 4-21 shows a second Storwize V7000 attached to the SONAS.


FC attached to SONAS Interface Nodes - V7000 #2

S1-PCIe-x8-FH-FL
S2-PCIe-x8-FH-HL
S3-PCIe-x8-FH-HL
S4-PCIe-x8-FH-FL
S5-PCIe-x8-FH-FL
S6-PCIe-x8-FH-FL
1
2
2
1
SONAS node
iMM E3 E4 E1 E2 Serial Video E5 E6
USB4
USB1
PS1
PS2
M1 M2 H1 H2 H3 H4
Upper S1 S2
Lower S1 S2
V7000 Control Enclosure 2/3 - #2
S1 Upper SAS S2
S1 Lower SAS S2
V7000 Expansion Enclosure 2/3 - #2.1
V7000 Expansion Enclosure 2/3 - #2.2
V7000 Expansion Enclosure 2/3 - #2.3
V7000 Expansion Enclosure 2/3 - #2.4
V7000 Expansion Enclosure 2/3 - #2.5
V7000 Expansion Enclosure 2/3 - #2.6
V7000 Expansion Enclosure 2/3 - #2.7
V7000 Expansion Enclosure 2/3 - #2.8
Legend
🟩 1Gb Ethernet
🟩 10Gb Ethernet
🟦 IB QDR
🟪 6Gb SAS
🟧 8Gb FC
🟥 RSA/IMM/Mgmt

V7000 Subsystem

- Dual Active/Active RAID controllers
- 8GB non-volatile cache per controller, 16GB total
- 4 x 8Gbps FC ports per controller, 8 total

Figure 4-21 Second V7000 cabling

Storage configurations

Storage can be grouped in 10-drive or 12-drive groups to configure RAID arrays. Use RAID 6 for data storage for performance and reliability. However, it is supported to use RAID 5 with hot spares and RAID 1 or RAID 5 with hot spares for SSD metadata configurations.

The Storwize V7000 comes in either 12 x 3.5 in. drives or 24 x 2.5 in. drive capacities per enclosure as shown in Figure 4-22 and Figure 4-23.

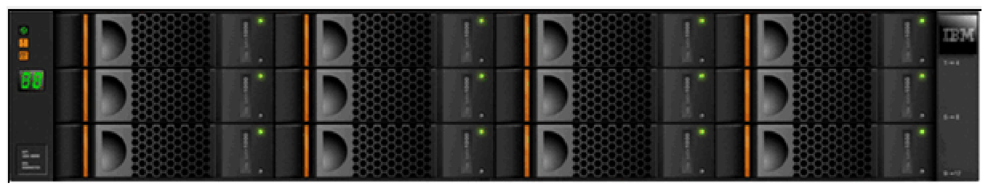


Figure 4-22 Image of the V7000 12 drive enclosure (3.5 in. drives)

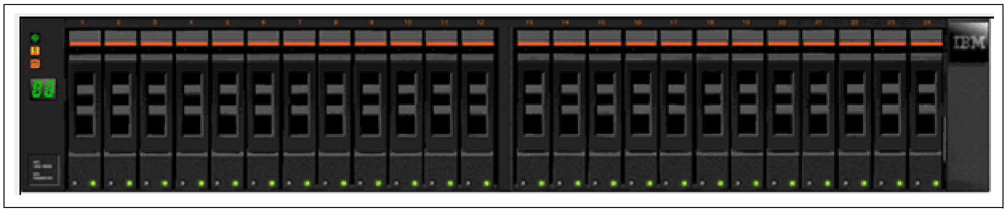


Figure 4-23 Image of the V7000 24 drive enclosure (2.5 in. drives)

SONAS supports one or two controllers with 0 - 9 expansion enclosures that are SAS connected to each storage node pair, as shown in Figure 4-24.

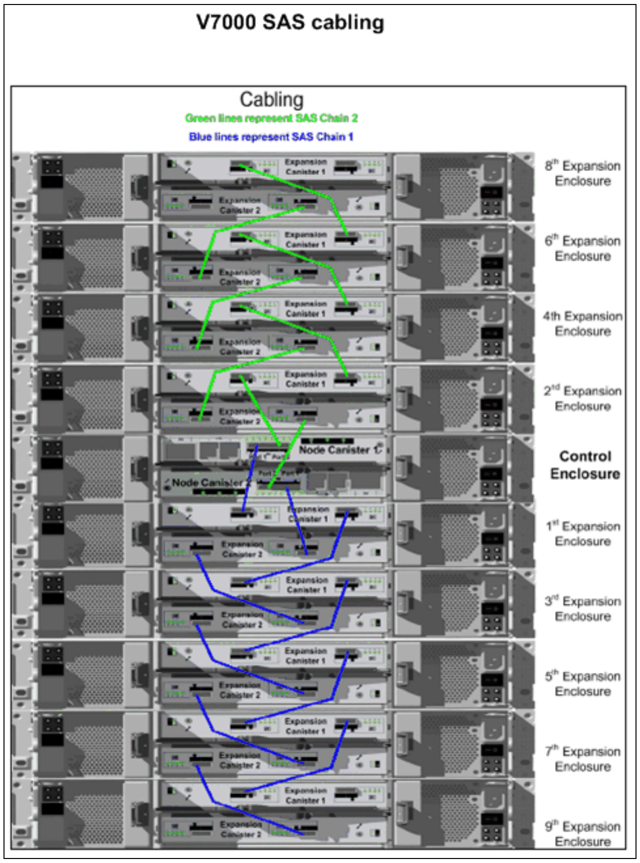


Figure 4-24 Image of the back of the V7000 controller with nine SAS connected expansions

It is a preferred practice to place the fastest drive technology closest to the controllers in the storage chain. It is also best to use drives from within the enclosure for RAID groups. In this case, using 10 drives in a RAID 6 group leaves an unbalanced number of drives remaining: Two out of 12 or four out of 24 drives in the v7000 enclosure (that can be used as hot spares). However, it is a common and good practice to use RAID 6 over 10 drives with no hot spares and span the 10 drives over multiple enclosures for the RAID group protection.

The current preferred practice is to build 1 managed disk (MDisk) group across each 10-drive RAID 6 group of (8 + P + Q) with a 32 KB segment size. Also, use a 256 KB scatter file system for small files, and span the MDisk groups across the enclosures. Further segment the MDisk and create multiple virtual disks (VDisks) from each MDisk group to create (populate) the file systems and limit the number of VDisks to four or less per MDisk group when possible.

This 256 KB block size considers the most effective capacity use for large groups of small files. When large file sequential workloads are expected and few small files exist, it would be better to use a 1 MB block size with scatter allocation pattern, and build your MDisk RAID 6 groups with a 128 KB segment size.

Note: Configure the V7000 logical storage configuration with the command-line interface (CLI) rather than the graphical user interface (GUI). The GUI selects many default values that cannot be overridden. Using the CLI enables you to choose parameters that are better for a SONAS file system (sequential versus striped VDisk, 128 KB strip size, IBM Easy Tier® off, and so on).

The following rules must be applied to that part of the Storwize V7000 configuration that is serving the SONAS subsystem:

- MDisks must be RAID arrays (“just a bunch of disks” (JBODs) are not supported).
- All RAID arrays must have redundancy (RAID 0 is not supported).
- All MDisks that are allocated to a single MDisk group must be of the same RAID type and the same drive type with the same performance characteristics. Each unique set of MDisks needs to go into a separate information lifecycle management (ILM) Pool in SONAS. You *cannot* mix different MDisks of different RAID or drive types in one pool.
- An MDisk group that consists of SSDs is supported.
- The MDisks in each MDisk group must come from the same physical back-end RAID controller.
- Single-port attached RAID systems are supported by the Storwize V7000. However, they are not supported for use with SONAS because of the lack of redundant connections between the Storwize V7000 and the RAID controller.
- The V7000 extent size must be chosen based on the maximum size of the cluster. Use an extent size of 256 MB for the MDisk groups that are used for SONAS storage. Other extent sizes can be used in non-SONAS MDisk groups.
- MDisk sizing and naming are chosen by the user. Follow the guidelines in the Storwize V7000 documentation.
- Storwize V7000 must configure a single host object that contains the worldwide port names (WWPNs) of all of the FC Ports in each of the SONAS storage nodes. This V7000 host object is called `SONAS_<serialnumber>` where `<serialnumber>` is the serial number of the SONAS node. This host object is used to map the VDisks that are in use by SONAS to the SONAS nodes.

Using 10-drive groups for RAID 6 (8 + P + Q) helps with drive stripe and write alignment and, as such, follows the preferred practice for data writes. The latency of spanning enclosures does not prove a significant latency over write stripe to storage RAID segment alignment. Therefore, in most cases, it is a preferred practice to use RAID 6 (8 + P + Q).

Design the RAID group to use a segment size that best aligns with the intended file system block and subblock size. This choice is described in Chapter 5, “File system configuration” on page 141.

Separating drives for metadata and data

It is often helpful to separate drives for metadata and data use to help reduce I/O contention to drives during mixed high performance data I/O and heavy metadata scan operations. Heavy normal read and write data access can compete with metadata scans for backup worklists, ILM scans, antivirus, and so on. However, as described in Chapter 5, “File system configuration” on page 141, it is important to recognize that many things are important to consider when you make this decision.

Having enough devices for metadata, along with having enough devices for data, is a delicate balance for many cost-conservative customers. Where extreme performance is not the requirement, it is often advised that sharing devices in the system pool for data and metadata is the best solution, especially when many spindles are used to support the file system.

For clients that require extreme performance where capacity is not the main requirement, it is often helpful to reserve a set of NSDs for metadata only in the system pool. However, when NSDs are set aside for metadata only I/O, and scan time performance is critical to success, it is also important to use multiples of four NSDs per storage node (16) for optimal port and channel saturation.

The HBA port I/O and striping capabilities, and the cache and channel capability of the storage controllers, play a large part in tuned I/O workload balance for data and metadata. So typically, for isolated metadata-only NSDs, the size can be smaller and the quantities should be presented to the file systems in multiples of four per storage node, or optimally 16.

Size is a major consideration for isolated metadata-only NSD devices as well. For example, you can set aside 5% of the file system capacity in drives, and isolate the number of spindles for metadata only, in the system pool. In this configuration, only metadata is hosted by those drives (NSDs) where data-only NSDs are used for data in the system pool.

Using 5% for metadata is a typical sizing, and it includes average file size and typical file count patterns, with GPFS replication of metadata only. If you anticipate a huge number of small files, make that number higher. If you expect a smaller number of larger files, it can be lower. If you choose not to replicate metadata, the size can also be smaller. Plan for more, rather than less, and always replicate metadata across two or more failure groups in GPFS.

In every case where metadata-only devices are to be used for a file system, the preferred practice is to consult a GPFS expert or the SONAS performance team to help make the best choice about what to use and how much to buy.

When SSDs are used for metadata, use RAID 1 or RAID 5 with hot spares, as previously mentioned in this chapter.

Defining volumes

Defining volumes on the V7000 can be done in the GUI or using the CLI. However, defining RAID groups in the GUI always creates the Stripe or Segment size at 256 KB. This does not configure optimal striping for file system subblock-to-storage segment size alignment. See Chapter 5, “File system configuration” on page 141 for block size considerations, and always try to align your storage stripe or chunk size to file system stripe.

It is a preferred practice to configure storage on the Storwize V7000 with the CLI. This practice follows the Storwize V7000 storage guides and Storwize V7000 preferred practice for SONAS gateway solutions, and it offers the best use of that storage platform in most cases.

Considerations for Storwize V7000

The IBM Storwize V7000 Storage configuration behind the SONAS gateway platform is a solution where the storage and the SONAS are independently managed. Therefore, your staff must understand both platforms for overall management of an effective SONAS gateway solution.

The SONAS solution supports only V7000 in the gateway by using internal storage enclosures, and not external V7000 mapped storage for SONAS shares. The Storwize V7000 behind SONAS must be dedicated and not in clustered storage solutions.

The SAN FC switches, and the SAN Volume Controller and Storwize V7000 storage system, are externally managed by using their own native management GUI. As such, the SAN FC switches and the SAN Volume Controller and Storwize V7000 Interface modules are not attached to the internal management network within the SONAS system.

The SAN FC switches and the SAN Volume Controller and Storwize V7000 storage system are externally serviced by using their own native service management interface and any guided maintenance procedures that are provided by these products. As such, the SAN FC switches and the SAN Volume Controller and Storwize V7000 system are *not* serviced and supported by the SONAS RAS package.

Code and firmware levels

At the time of the writing of this book, the best supported code level on V7000 firmware, for use behind the SONAS, is 7.3.0.7. Before installation, consult your IBM technical advisor (TA) and SSR to ensure that this code is enabled on the Storwize V7000, or to get guidance on currently supported V7000 firmware versions for use behind a SONAS.

Storwize V7000 storage preferred practice summary

This section summarizes the following preferred practices:

- For maximum performance, place only one Storwize V7000 behind each pair of SONAS storage nodes. For maximum capacity, place two Storwize V7000 subsystems behind each storage node pair.
- Never mix Storwize V7000 disk types in the same SONAS disk storage pool.
- Zone the Storwize V7000 with four paths per volume by using single-initiator zoning on the switches (one initiator port per switch zone).
- Zone 4 ports per Storwize V7000 controller with FC Switch redundancy between ports.
- Use large, and consistent, Storwize V7000 volume sizes. 1 TB, 2 TB, or 4 TB volume sizes provide excellent performance and manageability.
- All LUNs that are attached to SONAS must not be usable or seen by any other host.
- Provision groups of four NSDs per storage node to each file system for data or metadata to ensure the highest bandwidth by maximizing use of all ports, buses, channels, and controllers.
- Build the GPFS file system at a 256 KB block size with scatter allocation type for small file workloads, and 1 MB block size with scatter allocation type for large file sequential workloads.
- Create your RAID array segment size to match the file system block stripe. For a file system block size of 256 KB, use a stripe segment size of 32 KB for RAID 6 (8 + P + Q). For file system block size of 1 MB, use a stripe segment size of 128 KB for RAID 6 (8 + P + Q).
- Provision your storage with global hot spares.
- Review status on both SONAS and Storwize V7000 storage events daily, and manage each as separate storage solutions that are working together to drive your requirements.
- When your primary data storage tier is large-disk (3 TB) NLSAS technology, separate placement of metadata only on SSDs or high-speed SAS disk (in the system pool) to dramatically improve performance when heavy metadata scans are anticipated.

4.2.4 SONAS gateway with DS8000 storage

For the SONAS gateway with DS8000 (RPQ #631-21686), the storage is manually configured and separately managed.

The SONAS gateway solution with DS8000 is especially attractive to customers who decide to use DS8000 storage as their platform of choice. This configuration adds yet another storage tool from the expansive storage platform, and enables the SONAS to be ordered as an independent addition and fabric-connected or direct-connected to the DS8000 for use in a Storwize V7000 Unified platform.

Because of the high-performance capabilities of the IBM DS8000 storage platform, the SONAS can span two or four storage nodes in either single-node or double-node pair solutions to ensure maximum bandwidth and alleviate the storage node pair as a bottleneck (as with XIV Gen3).

However, volumes from the DS8000 can be provisioned only to a single storage node pair. In this scenario, an equal number of volumes are provisioned and distributed across to separate SONAS storage node pairs for preferred-practice, high-performance considerations. Figure 4-25 shows a DS8000 frame.

Tip: This configuration is a specialized storage solution for SONAS, and is strictly controlled by RPQ for ensuring the highest success with SONAS subject matter expert (SME) specialist assistance. It requires special resource attention for solution preparation and installation.

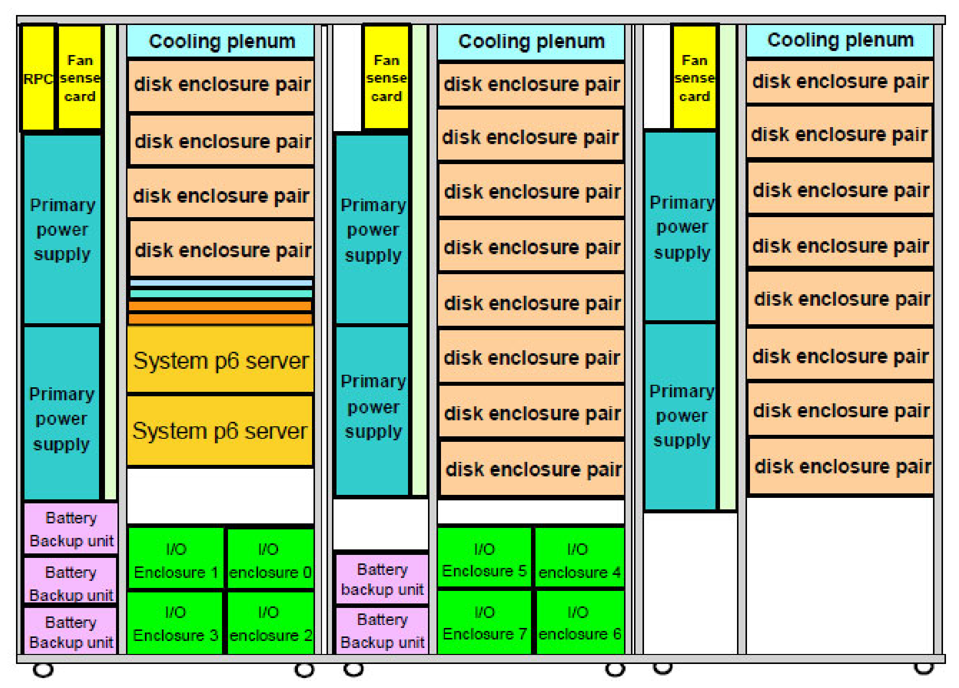


Figure 4-25 IBM DS8700 frame

Figure 4-26 shows a DS8800 frame.

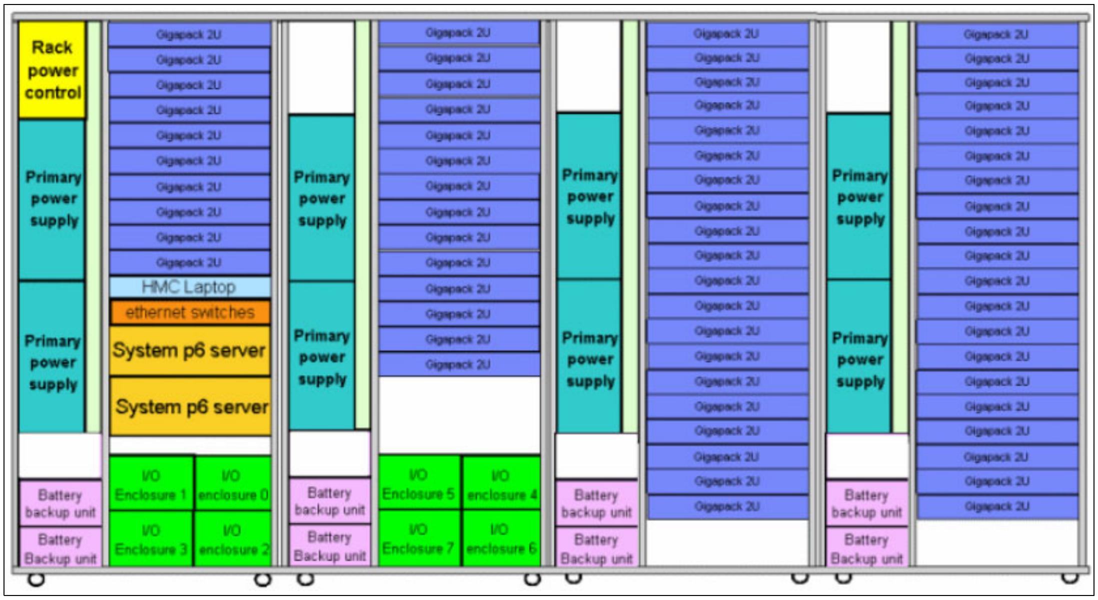
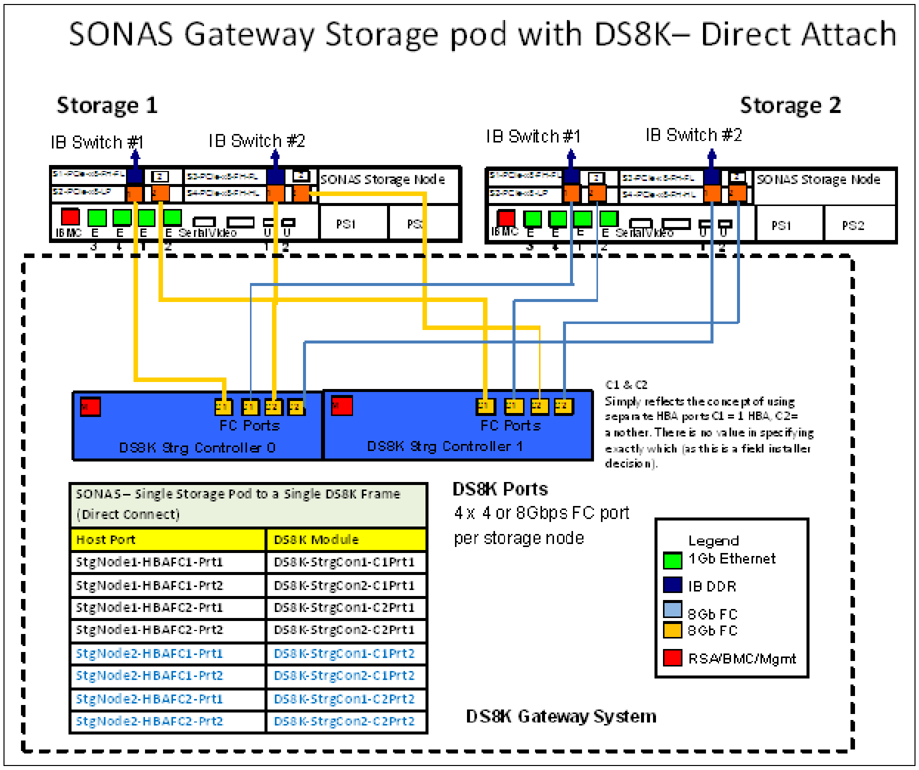


Figure 4-26 DS8800 frames image

Connectivity to the Storage node pairs is managed using LC-to-LC Fibre Channel cable (see Figure 4-27).



C1 & C2
Simply reflects the concept of using separate HBA ports C1 = 1 HBA, C2= another. There is no value in specifying exactly which (as this is a field installer decision).

SONAS – Single Storage Pod to a Single DS8K Frame (Direct Connect)

Host Port	DS8K Module
StgNode1-HBAFC1-Prt1	DS8K-StrgCon1-C1Prt1
StgNode1-HBAFC1-Prt2	DS8K-StrgCon2-C1Prt1
StgNode1-HBAFC2-Prt1	DS8K-StrgCon1-C2Prt1
StgNode1-HBAFC2-Prt2	DS8K-StrgCon2-C2Prt1
StgNode2-HBAFC1-Prt1	DS8K-StrgCon1-C1Prt2
StgNode2-HBAFC1-Prt2	DS8K-StrgCon2-C1Prt2
StgNode2-HBAFC2-Prt1	DS8K-StrgCon1-C2Prt2
StgNode2-HBAFC2-Prt2	DS8K-StrgCon2-C2Prt2

Figure 4-27 DS8000 connectivity chart

Figure 4-28 shows a DS8000-to-SONAS configuration.

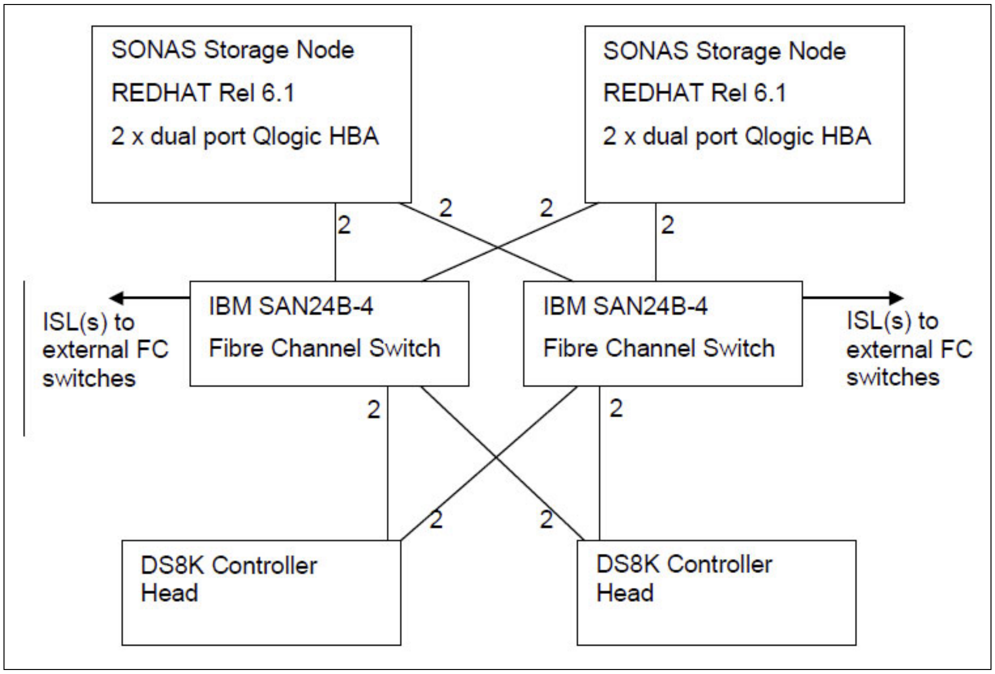
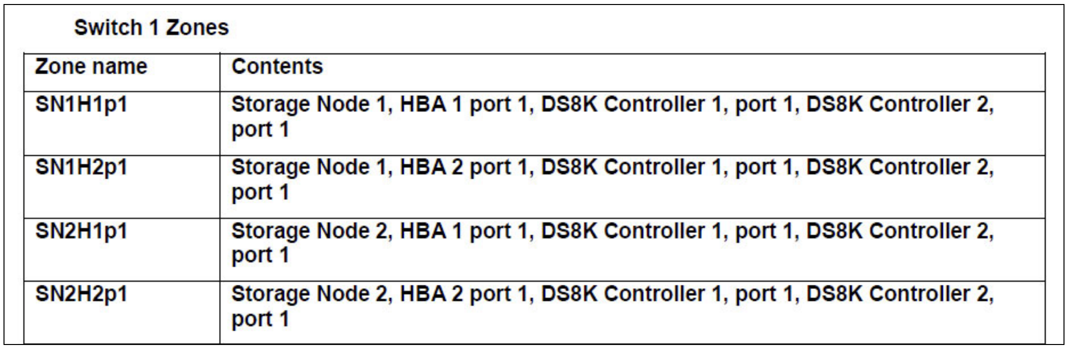


Figure 4-28 SONAS to DS8000 connection overview

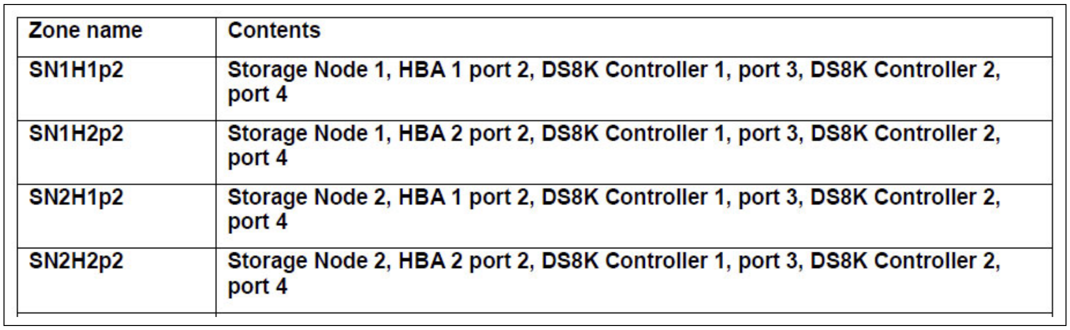
Figure 4-29 and Figure 4-30 show a sample zoning configuration for this solution.



Switch 1 Zones

Zone name	Contents
SN1H1p1	Storage Node 1, HBA 1 port 1, DS8K Controller 1, port 1, DS8K Controller 2, port 1
SN1H2p1	Storage Node 1, HBA 2 port 1, DS8K Controller 1, port 1, DS8K Controller 2, port 1
SN2H1p1	Storage Node 2, HBA 1 port 1, DS8K Controller 1, port 1, DS8K Controller 2, port 1
SN2H2p1	Storage Node 2, HBA 2 port 1, DS8K Controller 1, port 1, DS8K Controller 2, port 1

Figure 4-29 Switch 1 zone list



Zone name	Contents
SN1H1p2	Storage Node 1, HBA 1 port 2, DS8K Controller 1, port 3, DS8K Controller 2, port 4
SN1H2p2	Storage Node 1, HBA 2 port 2, DS8K Controller 1, port 3, DS8K Controller 2, port 4
SN2H1p2	Storage Node 2, HBA 1 port 2, DS8K Controller 1, port 3, DS8K Controller 2, port 4
SN2H2p2	Storage Node 2, HBA 2 port 2, DS8K Controller 1, port 3, DS8K Controller 2, port 4

Figure 4-30 Switch 2 zone list

Because the DS8000 with SONAS is a highly specialized solution, and has limited engagements in the field today, only a brief description is provided. Work with IBM to engage subject matter experts for this solution.

DS8000 storage preferred practice summary

This section summarizes the preferred practices for DS8000 storage:

- For maximum performance, place only one DS8000 behind each set of four SONAS storage nodes. For maximum capacity, place two DS8000s behind each set of storage nodes. The DS8000 can be placed behind a single pair of storage nodes for maximum capacity at lowest cost, at reduced performance ceilings.
- Never mix DS8000 disk types in the same SONAS disk storage pool.
- Zone the DS8x00 with four paths per volume by using single-initiator zoning on the switches (one initiator port per switch zone).
- Use large and consistent DS8000 volume sizes. 1 TB, 2 TB, or 4 TB provide excellent performance and manageability.
- All LUNs that are attached to SONAS must not be usable or seen by any other host.
- Provision groups of four NSDs per storage node to each file system for data or metadata to ensure the highest bandwidth by maximizing use of all ports, buses, channels, and controllers.
- Build the GPFS file system at a 256 KB block size with scatter allocation type for small file workloads, and a 1 MB block size with scatter allocation type for large file sequential workloads.
- Create your RAID array segment size to match the file system block stripe. For a file system block size of 256 KB, use a stripe segment size of 32 KB for RAID 6 (8 + P + Q). For a file system block size of 1 MB, use a stripe segment size of 128 KB for RAID 6 (8 + P + Q).
- Provision your storage with global hot spares.
- Review status on both SONAS and DS8x00 storage events daily, and manage each as separate storage solutions that are working together to drive your requirements.
- When your primary data storage tier is large-disk (3 TB) NLSAS technology, separate placement of metadata only on SSDs or high-speed SAS disk (in the system pool) to dramatically improve performance when heavy metadata scans are anticipated.

4.2.5 SONAS gateway with IBM DCS3700

For the SONAS gateway with IBM DCS3700 (FC 9008), the storage is manually configured, and independently managed.

The IBM DCS3700 storage with a SONAS gateway offers a highly flexible storage configuration, plus high-density storage in a small footprint. The IBM DCS3700 offers a wide variety of drive options with the inclusion of SSD for high performance. This selection can bring you great flexibility and performance with a small footprint, and at the lowest potential price per terabyte. The IBM DCS3700 storage solution enables for capacity to be added in 10-drive increments with a minimum of 20 drives per enclosure.

Note: The IBM DCS3700 must be installed in customer-provided frames and managed independently of the SONAS storage frame solution. It requires a DS Host Manager for initial configuration, management, and monitoring. The addition of an IBM DCS3700 Remote Storage Manager server is also required for call home support.

The following performance and flexibility benefits are provide by SONAS DCS3700 gateway solutions:

- This solution offers great flexibility in disk speeds, drive technology types, and drive size options with a highly flexible management of incremental growth. It enables for flexible storage tiering with the storage pod between SSD, SAS, and Near Line SAS storage options, and incremental growth in 10-drive increments from a 20-drive minimum.
- The IBM DCS3700 solution is also currently one of two possible solutions that accommodate support for an SSD tier of SONAS storage. The other solution is Storwize V7000. It is a preferred-practice solution for supporting a cheap and deep NLSAS drive storage tier for data storage with SSD dedicated for metadata.

  With one or two controllers that support 0, 1, or 2 expansion enclosures behind it, it offers the highest overall capacity at the smallest footprint in the SONAS gateway solution offerings. It uses 3 TB or 4 TB NLSAS drives, and 60 drives per enclosure for file system capacity depth.

The SONAS gateway configurations are sent in pre-wired racks that are made up of Internal switching components for the SONAS Interface nodes (with Integrated Management services) and SONAS Storage nodes. The storage is sold and framed separately. The solutions do not provide for cross-frame connectivity between SONAS, fabric, and storage. It is something that you must plan for independently. This section can help you accomplish that goal with the highest level of success. Figure 4-31 shows the SONAS gateway base frame.

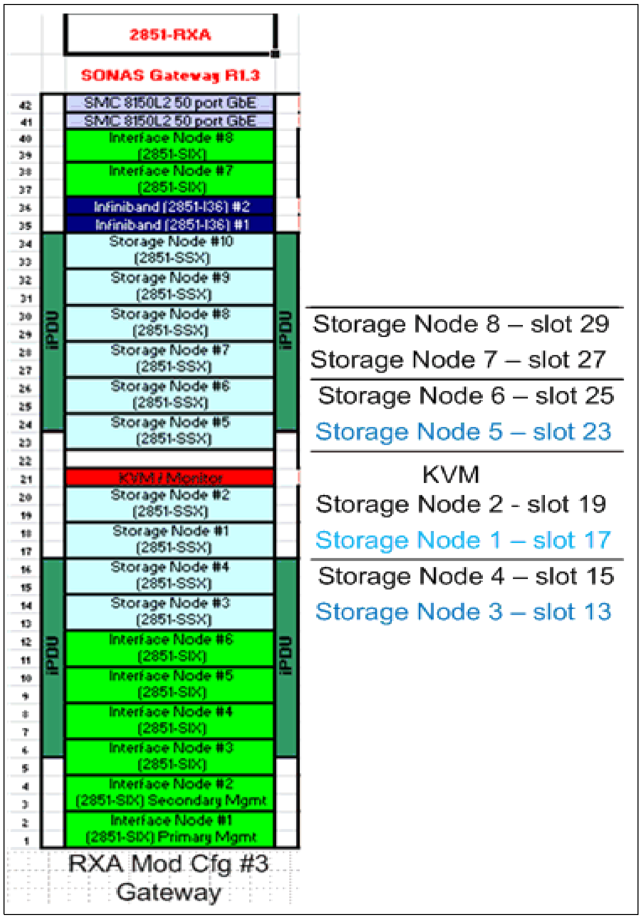

2851-RXA
SONAS Gateway R1.3
SMC 8150L2 50 port GbE
SMC 8150L2 50 port GbE
Interface Node #8 (2851-SIX)
Interface Node #7 (2851-SIX)
Infiniband (2851-I36) #2
Infiniband (2851-I36) #1
Storage Node #10 (2851-SSX)
Storage Node #9 (2851-SSX)
Storage Node #8 (2851-SSX)
Storage Node #7 (2851-SSX)
Storage Node #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 4-31 The SONAS gateway base frame lines up interface nodes from the frame bottom

IBM DCS3700 enclosures support up to 60 drives in each enclosure. All storage must be assembled and installed in a customer-provided frame and not the SONAS frames (see Figure 4-32).

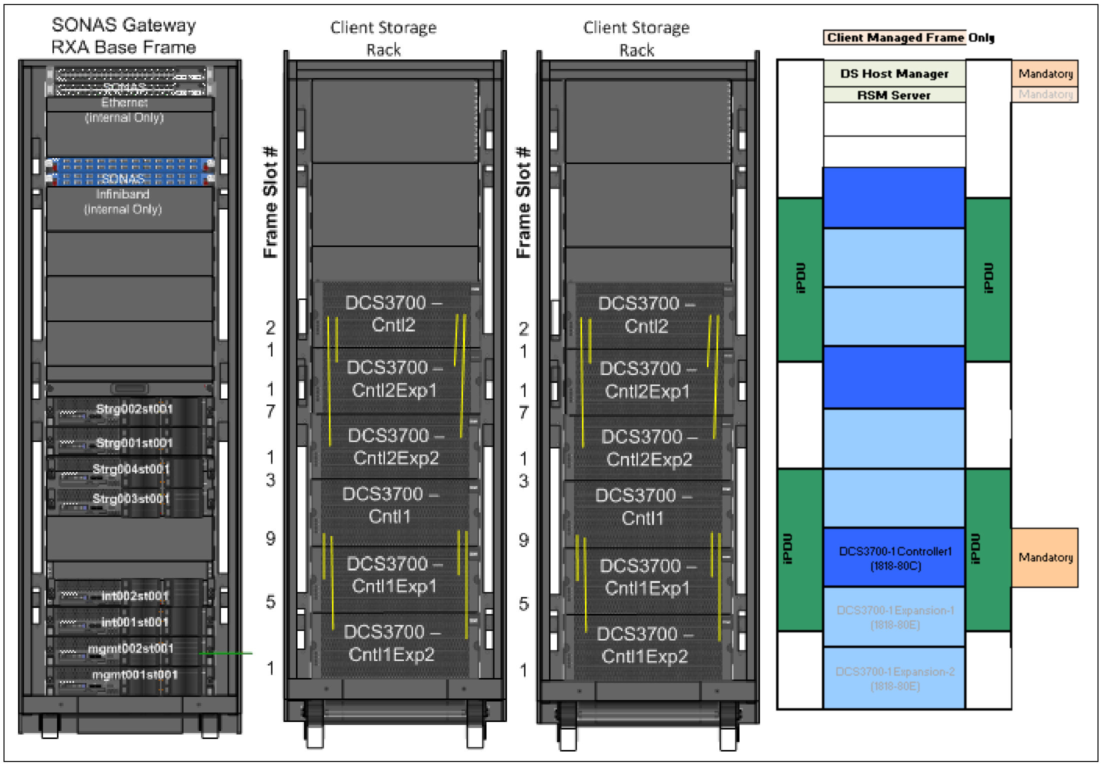


Figure 4-32 SONAS base frame plus client DCS3700 frames

Types of enclosures

There are effectively four types of IBM DCS3700 enclosures supported by the SONAS gateway. These types are 1818-80C, 1818-80E, 2851-DR2, and 2851-DE2. Each enclosure can hold up to 60 drives. The following list provides an explanation of each type:

- 1818-80C. This type is an IBM 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 when the minimum configuration and code levels (nonvolatile random access memory (NVRAM), and firmware) are met.
- 1818-80E. This type is an IBM 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 when the minimum configuration and code levels (environmental service modules (ESMs)) are met.
- 2851-DR2. This type is an IBM DCS3700 controller enclosure that is ordered with or specifically for a SONAS gateway. Though ordered as a separate line item, the 2851-DR2 is directly linked to the SONAS gateway and includes the correct code levels. It is ready to be integrated into a SONAS gateway solution.
- 2851-DE2. This type is an IBM 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 and includes the correct code levels. It is ready to be integrated into a SONAS gateway solution.

Note: Though the 1818 and 2851 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 an 1818-80C enclosure. Likewise, an 1818-80E enclosure cannot be used as an expansion enclosure for a 2851-DR2.

Expect to start with a minimum of 60 drives behind SONAS, and grow in increments of 10 drives. However, plan for growth wisely. rather than adding storage on a quarterly basis, it is wise to grow in targets of semi-annual or annual growth. This approach not only simplifies and decreases the number of maintenance operations that are required for upgrade and expansion, but it also reduces the cumulative burden of normal I/O competing with each data redistribution or restripe operation.

IBM DCS3700 is only supported in direct connection configurations (not fabric switch connections), so zoning is not a concern with this solution.

Gateway storage cannot be placed in SONAS frames

The storage must be configured independent of the SONAS and before SONAS initial configuration. All IBM DCS3700 product preferred practices for use on Red Hat Enterprise Linux (RHEL) clustered servers should be adhered to as a component of preferred practice in use with SONAS attachment.

The SONAS with IBM DCS3700 gateway configuration is not a unified storage platform and it is not supported to use part of the storage for other block or NAS clients. This restriction means that you can connect only one or two IBM DCS3700 controllers to any SONAS storage node pair, although you can use multiple storage node pairs to extend IBM DCS3700 storage in your SONAS solution. Figure 4-33 provides an orientation of the IBM DCS3700.

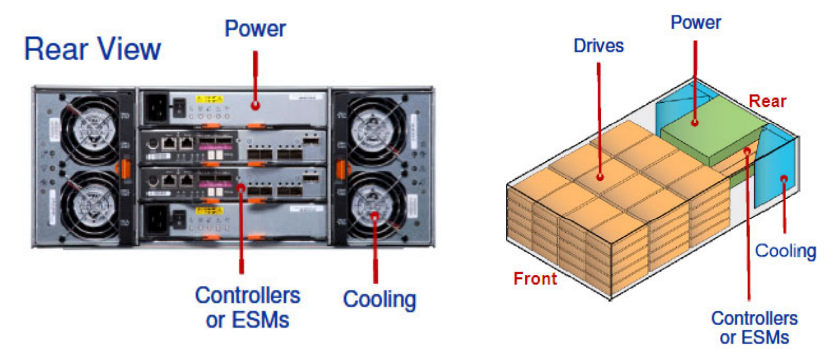


Figure 4-33 IBM DCS3700 component orientation

The IBM DCS3700 controllers and expansion enclosures are 60-drive enclosures that consist of five 12-drive trays in each enclosure (see Figure 4-34). In partial population configurations, it is mandatory to fill the front four drive slots in each tray to ensure correct system cooling and enable correct air flow. Therefore, the minimum drive capacity that is allowed in an IBM DCS3700 enclosure is 20 drives.

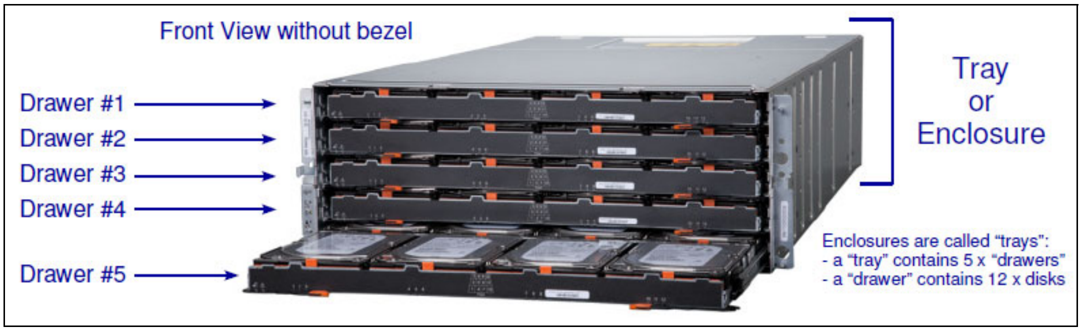
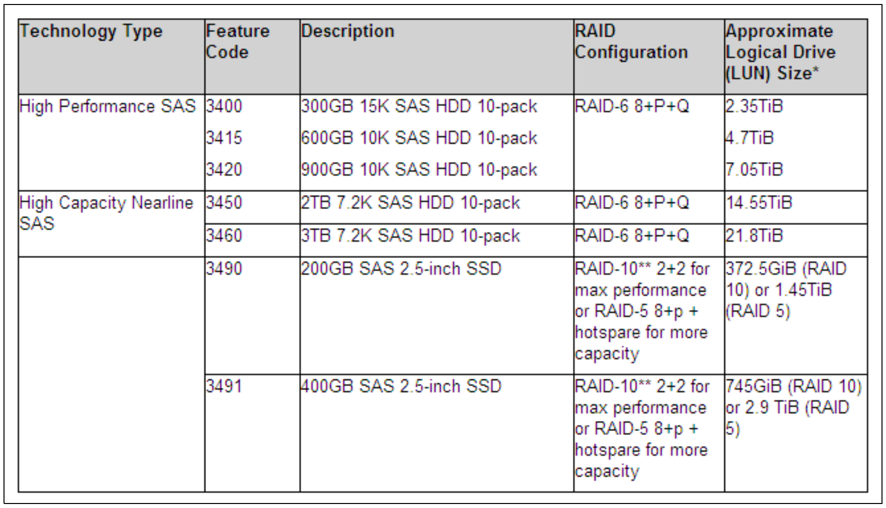

Front View without bezel
Drawer #1
Drawer #2
Drawer #3
Drawer #4
Drawer #5
Tray or Enclosure
Enclosures are called "trays":
- a "tray" contains 5 x "drawers"
- a "drawer" contains 12 x disks

Figure 4-34 IBM DCS3700 4U chassis and drive trays

IBM DCS3700 supports two types of controllers

IBM DCS3700 storage supports the following types of controllers:

- The controller that is based on the IBM DS3500 controller that supports 4 GB or 8 GB of Cache, and the drive types, are shown in Figure 4-35. This controller supports up to two expansion enclosures for a total of three enclosures and up to 180 drives. This configuration is supported in SONAS 1.3.2 and later.
- The Performance Module option Feature code 3100 is based on the IBM DS5000™ controller and supports cache sizes of 12 GB, 24 GB, or 48 GB. This controller supports up to five expansion enclosures for a total of six enclosures and up to 360 drives. This configuration is supported in SONAS 1.4.1 and later.



Technology Type	Feature Code	Description	RAID Configuration	Approximate Logical Drive (LUN) Size*
High Performance SAS	3400 3415 3420	300GB 15K SAS HDD 10-pack 600GB 10K SAS HDD 10-pack 900GB 10K SAS HDD 10-pack	RAID-6 8+P+Q	2.35TiB 4.7TiB 7.05TiB
High Capacity Nearline SAS	3450	2TB 7.2K SAS HDD 10-pack	RAID-6 8+P+Q	14.55TiB
	3460	3TB 7.2K SAS HDD 10-pack	RAID-6 8+P+Q	21.8TiB
	3490	200GB SAS 2.5-inch SSD	RAID-10** 2+2 for max performance or RAID-5 8+p + hotspare for more capacity	372.5GiB (RAID 10) or 1.45TiB (RAID 5)
	3491	400GB SAS 2.5-inch SSD	RAID-10** 2+2 for max performance or RAID-5 8+p + hotspare for more capacity	745GiB (RAID 10) or 2.9 TiB (RAID 5)

Figure 4-35 Drive types that are supported by the SONAS gateway IBM DCS3700 controller

Connectivity is only supported using direct connect between the SONAS storage nodes and the IBM DCS3700 controllers. Currently, one or two IBM DCS3700 controllers are supported behind each SONAS storage node pair. The IBM DCS3700 base controller supports up to two expansion enclosures, where the IBM DCS3700 Performance Module supports up to five expansion enclosures.

Each storage node has two connections to each controller enclosure, providing two paths to every volume, as shown in Figure 4-36 and Figure 4-37.

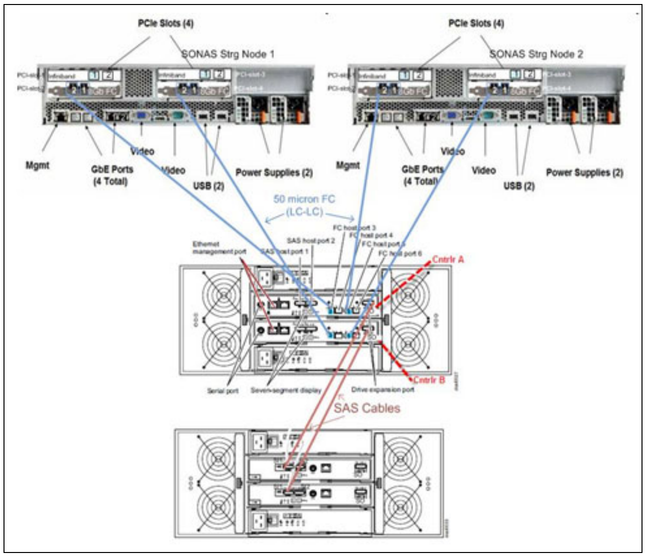


Figure 4-36 SONAS gateway with a single IBM DCS3700 and one expansion cable connectivity map

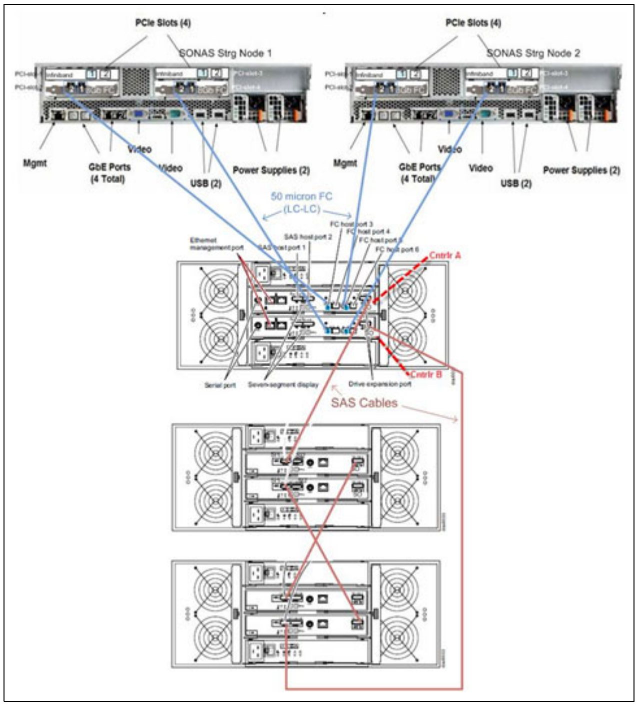


Figure 4-37 SONAS gateway with a single IBM DCS3700 and two expansion cables connectivity map

Figure 4-38 shows the preferred-practice configuration with SONAS storage nodes. The cables in blue are LC-to-LC Fibre Channel cables, where the brown lines are showing SAS cables from the controller to the expansion unit, and from the first expansion unit to the second expansion unit.

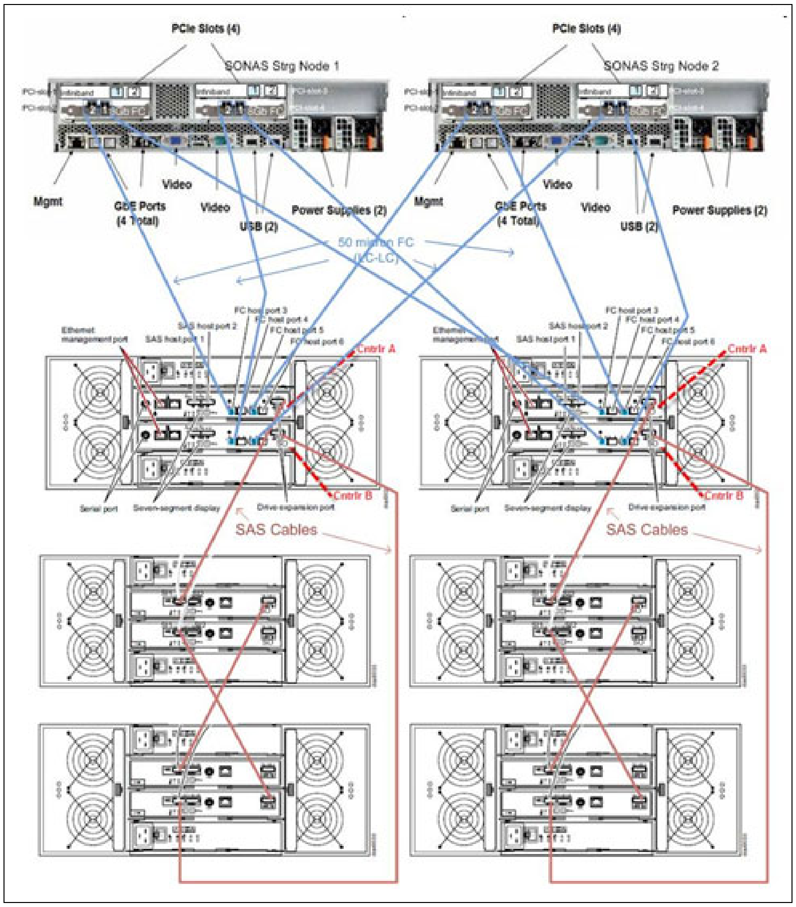


Figure 4-38 SONAS storage node cable connections with two separate controllers

IBM DCS3700 configuration

The IBM DCS3700 is initially configured before it is added to the SONAS storage nodes. Follow the *IBM System Storage DCS3700 Installation Guide* for configuring the IBM DCS3700, by using the IBM DS Storage Manager 10.84 software (or the latest supported). On a customer-managed server, configure the IBM DCS3700 controller IP addresses, storage pool, arrays, and volumes before you connect and provision the SONAS storage nodes. The storage installation follows the process in the *IBM System Storage DCS3700 Storage Subsystem and DCS3700 Storage Subsystem with Performance Module Controllers Installation, User's, and Maintenance Guide*, GA32-0959.

The SONAS storage node is added to the IBM DCS3700 as a cluster of two nodes (the SONAS storage node pair). The storage nodes are added by their connected Fibre Channel HBA WWPNs. They can be found on a sticker in the back of each SONAS Storage or by running **`cn_get_wwpns`** on the SONAS management node.

They can also be detected from the IBM DCS3700 after the storage node FC cable is connected and the SONAS storage nodes are powered on. The connection is made between the SONAS and the previously configured IBM DCS3700 by using the *SONAS Gateway Installation Guide*. This process is typically done by the IBM solution installation team.

The storage nodes are added to the defined cluster, and the volumes are mapped to the cluster rather than the individual nodes (this mapping is the preferred practice).

When mapping volumes of one IBM DCS3700, a `Lun0` device ID can be mapped to the SONAS cluster. However, it is a preferred practice to avoid mapping the Lun0 device ID for the second IBM DCS3700 behind any SONAS storage node pair to avoid issues in LUN discovery on the SONAS side.

Creating file systems with SONAS gateway IBM DCS3700 devices

For best performance with highly random I/O workload profiles, use SAS drive technology for the file system that you are planning to serve. The more, the better, because GPFS distributes I/O evenly between them all.

Generally, you also spread all data and metadata across all the SAS drives to ensure adequate capacity for both data and metadata in the system pool and enable the full distribution to support sharing high-performance through even striping.

When extremely high file system scanning operations are expected, such as extremely high-frequency operations of replication, backup, snapshots, and snapshot deletions, there can be an advantage to isolating the metadata devices from the data devices to remove contention between operation types. In this case, it is important to make sure that you have enough devices for both types of data.

Assuming that file types and sizes are a mixed average of most common scenarios, you can assume that you need approximately 5% of the overall data set size for metadata. This configuration assumes that metadata replication is wanted. In this case, a 100 TB file system capacity requires a 5 TB metadata capacity.

This allocation can be an effective way to improve your scan time on otherwise busy file systems. However, if that 5 TB comes from only a few spindles, there might not be enough read and write heads or I/O channels to handle the heavy random reads and writes of a common metadata workload pattern. In this case, you still need to make sure that you have enough spindles and volumes to manage both capacity and performance for the smaller metadata component of the system pool:

- Create at least 16 equally-sized LUNs in the pool that houses the GPFS metadata, even if this process requires creating smaller LUNs than specified in Figure 4-35 on page 132.
- High-performance SAS disk drives can be helpful for building high-performance on 10,000 and 15,000 RPM SAS disks drives in RAID 6 (8 + P + Q) arrays.
- For system availability reasons, consider having some number (a 10-pack) of high-performance SAS disk drives in the IBM DCS3700 system available as hot spares for RAID rebuilds.
- One or more 10-packs of SAS disks drives can be ordered on the IBM DCS3700 by ordering FC 3400 (for example, a 300 GB 15 K SAS HDD 10 pack).

High capacity Nearline SAS disk drives

Configure high-capacity 7.2K RPM Nearline SAS disks drives in RAID 6 8 + P + Q arrays.

A single logical drive (LUN) should be created out of the available space on each RAID array using the IBM DS Storage Manager. For system availability reasons, consider having some number (a 10-pack) of high capacity 7.2 K RPM Nearline SAS disk drives in the IBM DCS3700 system available as hot spares for RAID rebuilds.

One or more 10-packs of high capacity Nearline SAS disks drives can be ordered on the IBM DCS3700 by ordering FC 3450.

Disk pools

A new feature of IBM DCS3700 called Dynamic Disk Pools (DDP) or just disk pools, is now supported by SONAS release 1.4.1. DDP dynamically distributes data, spare capacity, and protection information across a pool of disk drives. DDP is designed to deliver and maintain predictable performance under all conditions, including recovering from drive failures.

These pools can range in size from a minimum of 11 drives to potentially as large as all of the drives in the IBM DCS3700 storage systems. There are no spare drives, only spare capacity. In other words, there are no idle drives; every drive participates.

The four key tenets of DDP technology are:

- Elimination of complex RAID management
- No idle spares to manage
- No reconfiguring of RAID when expanding
- Significant reduction of performance effect after a drive failure (or multiple drive failures) when compared to traditional RAID schemas

LUNs that are created from a disk pool are configured as RAID 6 LUNs and a segment size of 128 KB (the segment size cannot be changed). However, unlike RAID 6 arrays, 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 IBM DCS3700 storage for SONAS. The drives within each disk pool must be of the same type, size, and speed. SSDs cannot be in a disk pool.

Configure disk pools in groups of 10 (the minimum is 11). This setting is because LUNs are created as RAID 6 devices (8 + P + Q). A larger disk pool has a smaller percentage of capacity allocated to reserved capacity, and is affected less by a rebuild. In a typical configuration, the minimum disk pool is 40 disks. Make the number of equally sized LUNs the number of disks in the disk pool divided by 10 (for example, 40 / 10 = 4).

Also, remember to allocate LUNs in fours (4, 8, 12, and so on) to evenly distribute the LUNs across storage nodes and storage controllers. It is also important to note that LUNs are created in 4 GB chunks. Therefore, make the LUN size evenly divisible by 4 GB to maximize the used capacity of the disk pool.

Determining the most space-efficient DDP LUN size

This section provides an example of determining the most space-efficient LUN size:

1. Look at the Free Space (in the GUI) after the pool is created (for example, 53.246 TB).
2. Convert to GB (53.246 * 1024 = 54,523.904).
3. Divide by 4 (4 GB chunk) and drop the remainder (54,523.904 / 4 = 13,630).
4. Divide by the number of LUNs and drop the remainder (13,630 / 4 = 3,407).
5. Multiply by 4 (4 GB chunk) (3,407 * 4 = 13,628 GB).
6. Create 4 LUNs that are 13,628 GB or 13.308 TB.

Solid-state disks

It is becoming popular to use NLSAS for data stores, and SSD for improving scan rates of the metadata.

For the maximum performance from metadata scans, configure solid-state disks in RAID 1 2 + 2 hot spare, RAID 5, or RAID 10 2 + 2 + HS arrays. A single logical drive (LUN) is optimal for available space on each RAID array. However, for best performance, use 8 or 16 NSDs on the SONAS for maximum port and channel saturation. It is common to split the difference with two mapped LUNs from each of four or eight array groups for metadata, and use the IBM DS Storage Manager to create and map them.

For maximum protection, evenly divide the NSDs across two controller-based failure groups and replicate metadata with your GPFS file systems.

Order SSDs in groups of five, so that with each group of five SSDs, you have one RAID 1 (2 + 2 array, plus a hot spare SSD). Use the 400 GB SAS 2.5-inch SSD model for improved capacity.

Solid-state disk preferred practices

Because the most common use case for solid-state disks is for GPFS file system metadata, order a minimum of two groups of five (a total of 10 SSDs).

These two groups of five SSDs support the following configuration:

- One group of five SSDs can be used to create a RAID-10 2 + 2 array (with a spare SSD set aside in a global spare pool). One logical drive is created out of the RAID array, and the GPFS NSD that corresponds to this logical drive are assigned to one GPFS failure group.
- Another group of five SSDs can be used to create another RAID-10 2 + 2 array (with a spare SSD set aside in a global spare pool). One logical drive is created out of the RAID array and the GPFS NSD that corresponds to this logical drive assigned to a different GPFS failure group.

If you are going to use solid-state disks specifically for GPFS file system metadata, you need to estimate the amount of space that is needed for GPFS file system metadata.

Code and firmware levels

At the current date (date of this release), the best-supported IBM DCS3700 code level for use behind SONAS is IBM DS Storage Manager = 10.86.G5.x and FW= 7.86.46.00. Ask your IBM SSR to ensure that this code is enabled on the IBM DCS3700s before the SONAS installation, and inquire what the current SONAS release suggests for the IBM DCS3700 firmware.

IBM DCS3700 storage preferred practice summary

This section summarizes the preferred practices for IBM DCS3700 storage:

- Place one IBM DCS3700 behind each storage node pair for best performance to capacity scaling.
- For maximum performance, place only one IBM DCS3700 behind each pair of SONAS storage nodes. For maximum capacity, place two IBM DCS3700 subsystems behind each storage node pair.
- Never mix IBM DCS3700 disk types in the same SONAS disk storage pool.
- Use large, and consistent, IBM DCS3700 volume sizes of 1 TB, 2 TB, or 4 TB (the preferred practice is to have one logical drive per RAID 6 array) to provide performance and manageability. Smaller volume sizes are preferred for metadata-only solutions.
- Use Dynamic Disk Pools (disk pools) where possible. Ensure that disk pools are a minimum of 40 disks and must be the same type, size, and speed.
- If only the control enclosure is present in a controller string, the following guidelines apply:
  - Configure RAID 6 groups to contain no more than two drives per tray.
  - Configure RAID 5 SSD groups as 4 + P arrays, with one drive per tray, rather than an 8 + P + S array.
  - Configure RAID 1 and RAID 10 groups such that no mirror drives are contained in the same tray and hot spares are available.

- Higher-availability characteristics are achieved when one or more expansion enclosures are present in a controller string. If one or more expansion units are present in a controller string, apply the following guidelines:
  - Configure RAID 5 and RAID 6 groups to contain only one drive per tray.
  - Configure RAID 1 and RAID 10 groups such that no mirror drives are contained in the same tray.

  The two preceding guidelines are generally the default choices that are made by the Create Array and Disk Pool wizard in the IBM DCS3700 Storage Manager GUI when automatic configuration is used. However, depending on the number and type of drives present, their physical locations, and previously created arrays, this selection might not always occur. If you are concerned about availability, verify the drive selections in each array manually.
- All LUNs that are attached to SONAS must not be allocated or seen by any other host.
- Provision groups of four NSDs per storage node to each file system for data or metadata to ensure the highest bandwidth by maximizing use of all ports, buses, channels, and controllers.
- Build the GPFS file system at 256 KB block size with the scatter allocation type for small file workloads, and the 1 MB block size with the scatter allocation type for large file sequential workloads.
- Create your RAID array segment size to match the file system block stripe. For a file system block size of 256 KB, use a stripe segment size of 32 KB for RAID 6 (8 + P + Q). For a file system block size of 1 MB, use a stripe segment size of 128 KB for RAID 6 (8 + P + Q).
- Provision your storage with global hot spares.
- Review the status on both SONAS and IBM DCS3700 storage events daily, and manage each as separate storage solutions that are working together to drive your requirements.
- When your primary data store tier is large disk (3 TB) NLSAS technology, separate placement of metadata only on SSDs or high-speed SAS disk (in the system pool) to dramatically improve performance when heavy metadata scans are anticipated.

4.3 Failure groups and storage pools

This section provides an overview of failure groups and storage pools, and a few preferred practices.

4.3.1 Failure groups

GPFS disk failure groups are means of enabling GPFS to manage groups of NSDs with special care (such as synchronous replication of data, or metadata, or both). It is yet another layer of flexibility, power, and protection that is offered by GPFS to the SONAS and Storwize V7000 Unified storage platforms.

The preferred practice is that, whenever possible, you provision all metadata storage evenly into two separate failure groups to support the use of GPFS replication, when and if that level of extra protection is wanted or deemed necessary.

Also understand that synchronous replication adds mirroring of data that is often already adequately protected in an intelligent storage RAID set protection. Using it often adds a significant burden to write performance. Carefully consider this option and expert advice before making these decisions and creating the initial file system.

If replication is created to support the file system write activity at onset or creation time, it can be turned off subsequent to activating file shares non-disruptively.

4.3.2 Storage pools

Storage pools are a means of keeping separation in the types (tiers) of storage that are used by GPFS in SONAS or Storwize V7000 Unified storage platforms.

> **Tip:** If all storage is placed in the default failure group 1 only, no GPFS based synchronous replication can exist. However, if all devices are evenly distributed across 2 or more failure groups, the choice about whether to replicate that data or metadata can be made at any time.

Typically, different tiers of storage or storage types are added into different storage pools (or tiers) except for the system pool. The system pool is the default primary data placement pool and the default placement pool for metadata. It is possible to have SSD in the system pool with a data usage type tag of `metadataOnly`, while SAS or NLSAS devices are placed in the system storage pool with a data usage tag `dataOnly`.

Changes to disk pools and types must be made before you allocate devices to the file system.

If you are intermixing different storage vendors behind a SONAS (that is, DDN and IBM DCS3700), use different storage pools within the same file system to separate the different drive vendors and technologies. Placement and migration policies are a powerful way to use multiple pools and direct your data to the storage pool that you want.

For example, “hot” data can remain on a 15 K SAS disk pool, while “cold” data can move to a high capacity Nearline pool. There are many possibilities that you can consider when you intermix storage types and careful thought is critical for a successful deployment.

> **Tip:** Metadata can be placed only in the system pool. Therefore, all other storage pools can contain dataOnly usage types.

5

File system configuration

Managing the IBM General Parallel File System (IBM GPFS) file systems involves several tasks, such as creating, expanding, removing, mounting, unmounting, restriping, listing, and setting the attributes of the file systems. This chapter describes some of the different ways to create and modify the file system for preferred practice applications in your environment, and explains the key points of logic behind these guidelines.

It covers the preferred practice considerations for file systems in general, and reviews some conceptual configurations for different common workloads. This chapter cannot cover every specific requirement. However, the intent is to help you make better decisions in your cluster planning efforts.

This chapter describes creating, modifying, restriping, and removing a file system. It also includes use cases and typical scenarios.

Specifically, this chapter describes the following topics in terms of preferred practices:

- Getting access to file system content
- Creating a file system
- Removing a file system
- File system performance and scaling

5.1 Getting access to file system content

You can access the file system content in a SONAS solution by using file services like Common Internet File System (CIFS) or Network File System (NFS). However, when you are using the administration graphical user interface (GUI) or command-line interface (CLI), data access is not possible. This configuration is intentional.

It is not intended that the administrators have access to file or directory content, because file or directory names have the potential to disclose guarded business information. Although access rights are covered in Chapter 2, “Authentication” on page 25, some authentication topics that pertain to file systems are also considered.

5.1.1 Access rights

CLI and GUI administrators cannot change the access control lists (ACLs) of existing files or directories. However, they can create a new directory when they define a new share or export, and assign an owner to this still-empty directory. They can also create a share or export from any path they know. However, exporting a directory does not grant any additional access. The existing ACLs remain in place. Therefore, an administrator cannot get or create access for himself or others to files solely by creating a share or export.

During share or export creation, when a new directory is created, a user can be specified as the owner of the new directory. This user owns the underlying directory and therefore acts as the security administrator of this directory. Therefore, it is important that the owner is set specifically and thoughtfully on creation of the directory (or file set) to be shared.

Access rights must be managed by the owner and the authorized users, as specified in the ACL by using a Microsoft Windows workstation that is connected to the share or export by using the CIFS protocol. An NFS user can set access rights by using Portable Operating System Interface (POSIX)-style access controls, not the ACLs that the Windows operating system uses.

In addition, during the initial setup of the file system, a default ACL can be set on the file system root directory, implementing inheritance, where any newly-created directory inherits the ACLs of the default set. This configuration is important to remember, and is described further in Chapter 6, “Shares, exports, and protocol configuration” on page 171.

Restriction: Access rights and ACL management can be run only on Windows clients. The command-line utilities of the UNIX stations that deal with the POSIX bits must not be used on files that are shared with Windows or CIFS clients (multiprotocol shares), because they destroy the ACLs. This restriction applies to the UNIX commands, such as `chmod`.

Tip: SONAS 1.5.1 provides functions for managing ACLs in the file sets, file systems, and Shares section of the GUI. The `chacl` and `lsacl` commands support this function in the CLI. You can use this function to manage ACL for all clients. For more information, see *IBM SONAS Implementation Guide*, SG24-7962.

5.2 Creating a file system

When you are creating a file system, there are two types of parameters:

- Parameters that can be changed dynamically
- Parameters that cannot be changed dynamically

The parameters that cannot be changed dynamically are likely to require downtime to unmount the file system. Therefore, it is important to plan the file system well before it goes into production.

5.2.1 File system block size

One key parameter that must be determined at file system creation is the file system block size. After the block size is set, the only way to change it is to create a new file system with the changed parameter and migrate the data to it. In most cases, test the block size and usage type on a test system before you apply your configuration to your production deployment.

GPFS in SONAS supports a few basic block sizes: 256 kilobytes (KB), 1 megabyte (MB), and 4 MB, with a default size of 256 KB. The block size of a file system is determined at creation by using the **-b** parameter to the **mkfs** command.

This chapter presents some guidelines for how to choose an appropriate block size for your file system. Whenever possible, it is best to test the effect of various block size settings with your specific application before you define your file system.

5.2.2 Methods to create the file system

You can use IBM SONAS CLI commands or the IBM SONAS GUI to create a file system. A maximum of 256 file systems can be mounted from an IBM SONAS system at one time. However, the SONAS system reduces the value of managing so many independent file systems. This concept is explained later in this section.

Typically, creating file systems from the CLI offers more specific options and better control of the task, and more technical clients prefer the CLI (as a leading practice). However, if devices are properly provisioned, well-balanced, and easily identified by type, the GUI offers the simplest way of creating file systems in SONAS. Choose the device type and size and create the file system.

Tip: The CLI offers the preferred-practice flexibility and control for file system creation.

5.2.3 File system options

Every file system contains an independent set of Network Shared Disks (NSDs). So, reducing the number of file systems enables you to load more NSDs (logical unit numbers (LUNs) and Redundant Array of Independent Disks (RAID)-protected drive resources) behind the file systems that you create. Loading more NSDs improves your potential back-end performance by striping data across more spindle surfaces and read/write (RW) access controller arms.

Therefore, the best performance that is achievable in a file system is typically the one that contains the greatest number of spindles. Further performance assurance can be obtained when the file system structure is configured for highest suitability for the file I/O type and workload access patterns that are used by that file system.

By increasing the number of spindles in the file system, you achieve the following benefits:

- Maximize the RW head devices that work together to share the input/output (I/O) demands
- Reduce the seek times
- Improve latency
- Improve capacity in scale with performance from the back-end storage

Consider the following file system NSD and GPFS disk options before file system creation:

- Failure group assignment
- NSD allocation and storage node preference
- Storage pool use
- Usage type

Consider the following file system options before you create file systems:

- File system naming
- NSD allocations
- Usage types
- GPFS data or metadata placement and replication
- Block size
- Segment size
- Allocation types
- Quota management
- Data management application program interface (DMAPI) enablement
- File system logging
- Snapshot directory enablement
- iNode allocations and metadata

This section examines these options as they apply to preferred practices for several common use case scenarios to help illustrate the value of these considerations.

5.2.4 File system naming option

The simplest of these options is file system naming. From the file system, there can be many independent or dependent file sets. File system naming serves different clients differently. Some clients choose to name file systems by simply applying the IBM generic standard that typically appears in documentation with a numeric value, such as `gpfs0`, `gpfs1`, and `gpfs2`. Some clients use descriptive titles to help them identify the content type such as `PACS1`, `Homes1`, `FinanceNY`.

In either case, it is left to the customer to decide. The preferred practice is to keep it short, simple, and descriptive for your own purposes. Do not include spaces or special characters that can complicate access or expand in the shell.

Every file system also requires a specified path (such as file system *device* `gpfs0` on the specified path `/ibm/gpfs0` or `/client1/gpfs0`). It is required that the path is consistent in the naming convention for simplicity of overall management (such as: `/ibm/gpfs0`, `/ibm/gpfs1`, `/ibm/gpfs2`, and so on).

The name must be unique and is limited to 255 characters. 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 path entry such as `/dev`.

For IBM Storwize V7000 Unified systems, the GPFS file name is limited to 63 characters.

5.2.5 NSD allocations for a balanced file system

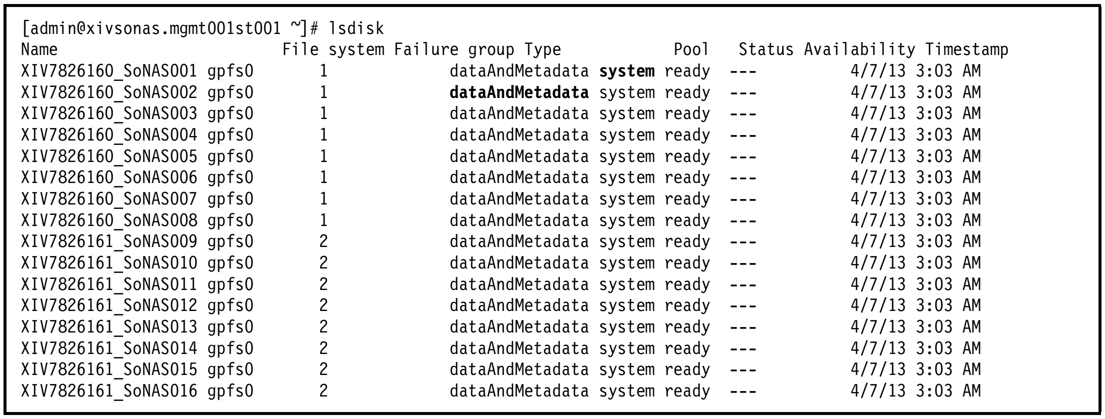
GPFS NSDs are RAID-protected disk volumes that are provisioned from back-end storage solutions to the SONAS GPFS Server (which are the SONAS Storage nodes in a SONAS cluster), for use in building SONAS file systems. Back-end storage systems, such as DataDirect Networks (DDN) for appliance storage, Storwize V7000, IBM XIV, IBM DS8000 and the DS8x00 series, and IBM DCS3700 are attached to NSD Server sets (and are typically configured as SONAS Storage node pairs).

An exception to this rule is in the case of DS8x00 or XIV Gen3, where performance benefits can be extended beyond storage node pair limitations by attaching to a four-node storage node set (a special solution that requires special development team consultation). When the back-end devices can support a great deal more bandwidth than can be managed by the storage node pair that they are attached to, special allowance is provided for spreading the volumes of those back-end devices across two storage node pairs rather than just one.

At installation, after adding new storage, or when new storage is provisioned to the SONAS cluster, the GPFS servers (storage nodes) discover the newly attached storage. By default the provisioned volumes are mapped as multipath devices by the Linux multipath daemon. Then, GPFS converts the volumes into NSDs. The NSD headers are placed (written) on the volume (LUN), and the devices are added to the general disk availability list in the cluster as the `system` storage pool devices for `dataAndMetadata` type storage, in failure group 1 (default).

Regardless of the disk type or disk size, the NSDs are automatically added to the `system` storage pool, with all the defaults mentioned previously. It is specifically for this reason that the storage use must be planned and defined correctly before it is assigned to a file system.

Ideally, the devices should be identified by controller serial number and separated into two separate failure groups (failure group 1 and failure group 2). Then, depending on the intended use of the devices, you put the devices into the intended storage pool and specify the *type* of data that is allowed to be placed on that device (`dataOnly`, `metadataOnly`, or `dataAndMetadata`). Figure 5-1 shows two XIVs behind a SONAS pair.

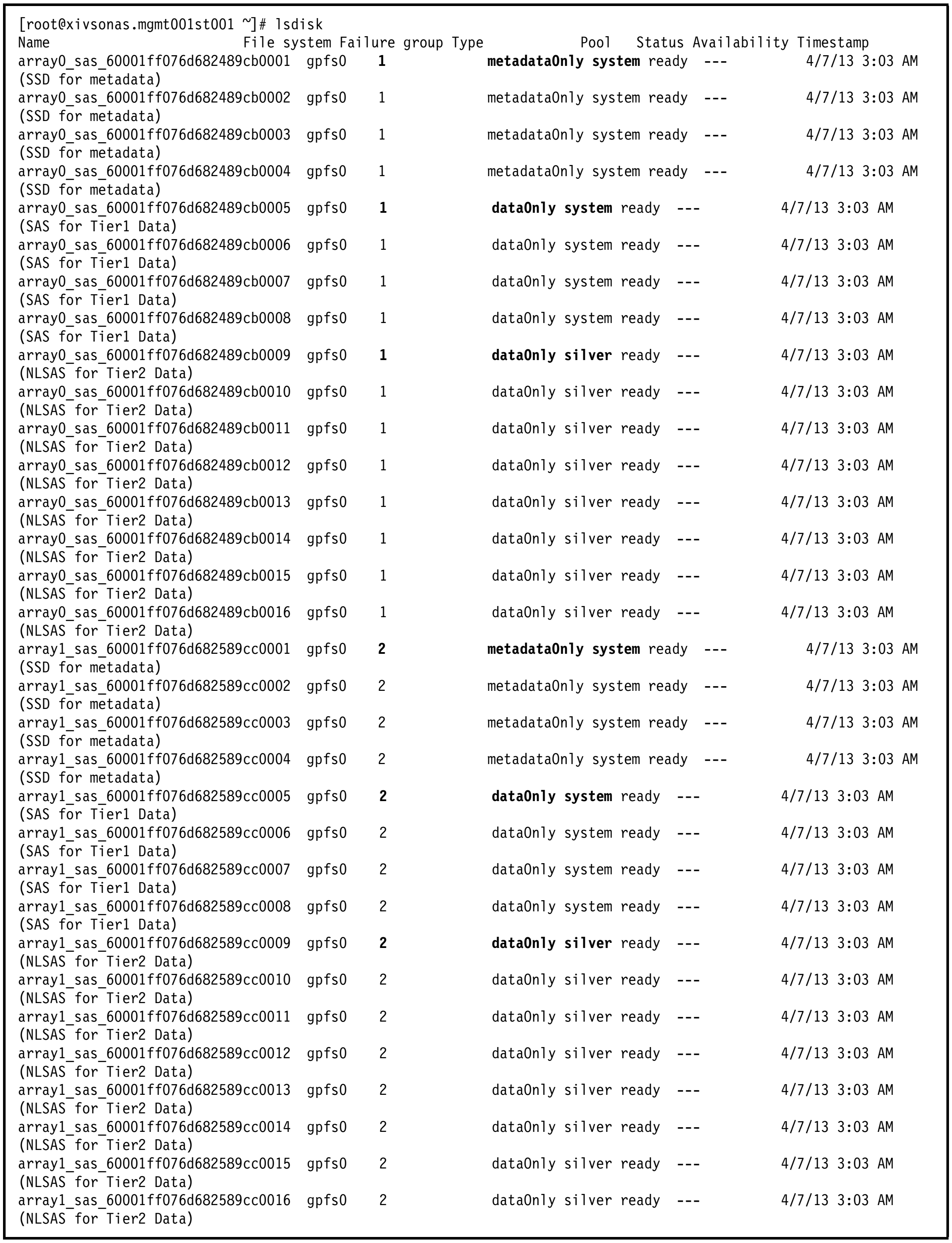


```
[admin@xivsonas.mgmt001st001 ~]# lsdisk
Name                       File system Failure group Type            Pool   Status Availability Timestamp
XIV7826160_SoNAS001 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS002 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS003 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS004 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS005 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS006 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS007 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826160_SoNAS008 gpfs0      1             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS009 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS010 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS011 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS012 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS013 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS014 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS015 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
XIV7826161_SoNAS016 gpfs0      2             dataAndMetadata system ready  ---          4/7/13 3:03 AM
```

Figure 5-1 Example lsdisk command output from the CLI

Figure 5-1 shows two XIVs behind SONAS, each providing 8 volumes. The NSDs are split into two failure groups to enable GPFS replication of metadata across the two XIVs. Note that all the devices remain in the `system` pool to enable these devices to be used for primary data placement and set to enable both data and metadata to be spread across all listed devices. If you had set aside a few devices with a specific usage type of `metadataOnly`, they can be added to the file system with all the other devices. However, only metadata would be written to these devices. The same rule applies if you set aside devices with usage type `dataOnly`.

As shown in Figure 5-2, you can see that, by using the **chdsk** command, you can set the designated pool for solid-state drives (SSDs), serial-attached SCSI (SAS), and Near Line SAS (NLSAS) drives, along with usage type and failure group. This enables GPFS to manage specific data to specific targets.



```
[root@xivsonas.mgmt001st001 ~]# lsdisk
Name                        File system Failure group Type           Pool   Status Availability Timestamp
array0_sas_60001ff076d682489cb0001  gpfs0    1           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array0_sas_60001ff076d682489cb0002  gpfs0    1           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array0_sas_60001ff076d682489cb0003  gpfs0    1           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array0_sas_60001ff076d682489cb0004  gpfs0    1           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array0_sas_60001ff076d682489cb0005  gpfs0    1            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array0_sas_60001ff076d682489cb0006  gpfs0    1            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array0_sas_60001ff076d682489cb0007  gpfs0    1            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array0_sas_60001ff076d682489cb0008  gpfs0    1            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array0_sas_60001ff076d682489cb0009  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0010  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0011  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0012  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0013  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0014  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0015  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array0_sas_60001ff076d682489cb0016  gpfs0    1            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0001  gpfs0    2           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array1_sas_60001ff076d682589cc0002  gpfs0    2           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array1_sas_60001ff076d682589cc0003  gpfs0    2           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array1_sas_60001ff076d682589cc0004  gpfs0    2           metadataOnly system ready  ---         4/7/13 3:03 AM
(SSD for metadata)
array1_sas_60001ff076d682589cc0005  gpfs0    2            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array1_sas_60001ff076d682589cc0006  gpfs0    2            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array1_sas_60001ff076d682589cc0007  gpfs0    2            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array1_sas_60001ff076d682589cc0008  gpfs0    2            dataOnly system ready  ---         4/7/13 3:03 AM
(SAS for Tier1 Data)
array1_sas_60001ff076d682589cc0009  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0010  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0011  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0012  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0013  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0014  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0015  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
array1_sas_60001ff076d682589cc0016  gpfs0    2            dataOnly silver ready  ---         4/7/13 3:03 AM
(NLSAS for Tier2 Data)
```

Figure 5-2 Example of three tiers of disk in one gpfs0 file system

Note: The data type and target definition must be set before file system creation.

The CLI command **`lsdisk -r -v`** displays the status of all NSDs (storage volumes) from the back-end storage. This command lists the device name, with its assigned failure group, storage pool, data allocation type, size and available capacity, and storage node preference. This data is critically important to building a well-balanced, high-performance SONAS file system. Consider a few of these properties and why they might be important.

If the disk is already associated with a file system, the file system device name is included in the **`lsdisk`** output.

The CLI command **`lsdisk -v`** reports the device name, file system association, failure group association, disk pool association, data allocation type, and storage node preference.

The **`chdisk`** command enables you to change the settings of disks before they are added to a file system. However, if the disk is already in a file system, it must be removed, changed, and re-added to the file system (in normal circumstances, this task can also be done non-disruptively with GPFS).

NSD allocation and storage node preference

NSD allocation and storage node preference is an important component of workload balance. The **`lsdisk -v`** command can be used to detect the NSD's storage node preference (such as `strg001st001`, `strg002st001`). What this list defines is the storage node that owns the primary workloads of the defined NSD.

The first node in the storage node preference list always does the work unless the primary preferred node is down or otherwise unavailable. In that case, the second node that is listed in the storage node preference takes over control of access and use of the defined NSD until the primary node comes back online and GPFS regains control of those NSDs.

Number of NSDs used for the file system

The number of NSDs that are used for the file system data targets are in a preferred practice configuration when there are enough NSDs provisioned to fill all of the busses and channels, and to maximize both the storage node host bus adapter (HBA) port channels and storage device controller port channels available for best performance. In this regard, typically NSDs in multiples of four, per storage node, per NSD pool, and usage type, is best.

The performance team has tested that a minimum of eight metadata NSDs per storage node has shown excellent performance for metadata scan rates in SONAS performance testing (or 16 NSDs per storage node pair). Conversely, if you have only two NSDs in a file system, you greatly limit the parallelism of I/O at the storage node level, because one NSD per storage node is written, and the file system data is striped across only two NSDs.

Usage type that is assigned to disks

The usage type that is assigned to disks defines the type of data that will be targeted on that device. The `system` pool is the default pool and, by default, it sends all data and metadata to devices in the system pool that are added to a file system. If you want to isolate metadata to go only to designated disk targets in the system, the `usageType` definition must be `metadataOnly`. This can be set with the **`chdisk`** command from the CLI before assigning the disk (or adding the disk) to the file system. See Example 5-1 for the syntax of the **`--usagetype`** option of the **`chdisk`** command.

Example 5-1 Example chdisk option for setting usageType

```
--usagetype usageType
             Specifies  the  usage type for the specified disks. Valid usage
types are dataAndMetadata, dataOnly, metadataOnly, and descOnly.
```

The system pool requires that some devices be added to the system pool for data usage type. Whether the devices are added as dataAndMetadata or dataOnly, you need to have some devices in the system pool that support metadata, and some devices that support data.

Any devices that are added to a pool other than the system pool are added as dataOnly, because GPFS puts all metadata on devices in the system pool.

Important: Setting disk failure groups, storage pools, and allocation types before they go into a file system is critically important. They should always be balanced across storage node preference to avoid having a single storage node carrying the workload of a defined file system and all its associated file sets.

5.2.6 GPFS synchronous replication and failure group definitions

When the pools, usage type, and failure groups are established, it is time to create the file system. An even distribution of NSDs suggests that you have an even number of devices for each storage node preference. Also, each storage pool that is included, and each data allocation type that is defined, are also evenly balanced across the two failure groups and storage node preferences.

The primary reason for using failure groups is to replicate metadata only. As mentioned in 5.2.5, “NSD allocations for a balanced file system” on page 145, your metadata is placed in failure group 1 on any device in the system pool that is defined for metadata allocation type and replicated to devices in failure group 2.

If the device is in the file system and GPFS file system replication is not enabled, the file system can use all failure groups for data or metadata. Replication is synchronous, and requires that both writes be committed before it gets acknowledged.

For example, assume that you have eight NSDs (volumes) with an assigned allocation type of data and metadata. Further assume that all eight devices are in the system (default) storage pool, but half the devices are in failure group 1 and the other half are in failure group 2. At file system creation, all devices are added.

Without further enablement of replication at the time of file system creation, the metadata and data are striped across all the NSDs in both failure groups, and no data or metadata is replicated. For GPFS to replicate metadata between the file system failure groups, it must be defined or set using **mkfs** or **chfs** commands. This option can be applied or disabled without unmounting the file system. See Figure 5-3.

```
mkfs fs-device-name [-R { none | meta | all }]
chfs fs-device-name [-R { none | meta | all }]
```

Figure 5-3 example commands for changing GPFS replication level activity

Performance of data and metadata is optimal for the devices of the file system. Performance scales as new, similarly-defined devices are added and restriped. Data is redistributed in balance across more devices, because more spindles and read/write heads are available to manage distributed data access without the extra write latency.

The chfs command

Replication can be set or modified with the **chfs** CLI command.

The **chfs** command has the following options for setting or changing replication:

```
-R { none | meta | all }
```

-R Sets the level of replication that is used in this file system. Optional. The variables are as follows:

- The `none` option indicates no replication at all.
- The `meta` option indicates that the file system metadata is synchronously mirrored across two failure groups.
- The `all` option indicates that the file system data and metadata is synchronously mirrored across two failure groups.

If this property is changed on an existing file system, it takes effect on new files only. To apply the new replication policy to existing files, run the **restripefs** command.

Using synchronous file system replication

Synchronous file system replication is a feature that is provided by the GPFS file system. The file system can use synchronous replication only if the NSDs that make up the file system exist in two or more predefined failure groups. If two or more failure groups do exist, the standard preferred practice is to replicate metadata across the failure groups.

Typically, metadata is equal to roughly 2.5 - 3% of the file system capacity, and replicating metadata adds a layer of file identification protection by enabling GPFS to have two separate copies to read from if GPFS determines that one copy is faulty or corrupted. In this case, the metadata capacity can take up to 5 or 6% of the file system capacity.

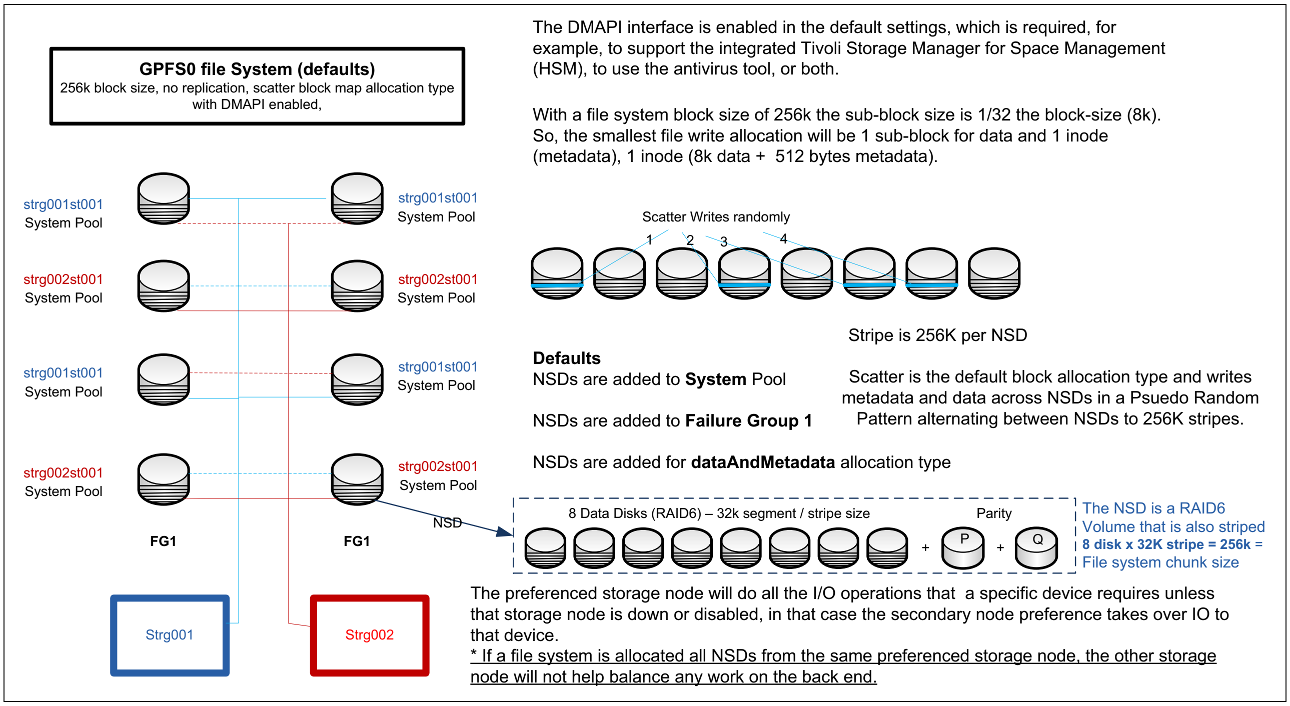
The preferred practice for most configurations is to use two failure groups for all back-end storage, based on a dual-controller technology. However, do not use two failure groups for XIV storage frames when a single XIV is defined as the back-end storage to SONAS. When two or more XIVs are configured in the SONAS solution, place the volumes from each serial number in two separate failure groups for metadata replication.

The preferred practice for highest reliability is to implement replication of both data and metadata. This practice creates multiple synchronous copies of all data and metadata on redundant RAID-protected volumes, and provides the highest level of redundancy that is practical within the immediate cluster. However, this protection also comes at the highest cost penalty in capacity and performance.

For highest performance purposes, do not implement replication. All NSDs would exist in a Single (default) failure group, and the only protection of the data or metadata would be the RAID-protected volumes (that make up the NSDs).

Select the configuration that is appropriate to your environment.

Figure 5-4 shows a SONAS file system that is built on default settings.



The DMAPI interface is enabled in the default settings, which is required, for example, to support the integrated Tivoli Storage Manager for Space Management (HSM), to use the antivirus tool, or both.

With a file system block size of 256k the sub-block size is 1/32 the block-size (8k). So, the smallest file write allocation will be 1 sub-block for data and 1 inode (metadata), 1 inode (8k data + 512 bytes metadata).

Defaults
NSDs are added to **System** Pool

NSDs are added to **Failure Group 1**

NSDs are added for **dataAndMetadata** allocation type

Scatter is the default block allocation type and writes metadata and data across NSDs in a Psuedo Random Pattern alternating between NSDs to 256K stripes.

The NSD is a RAID6 Volume that is also striped **8 disk x 32K stripe = 256k =** File system chunk size

The preferenced storage node will do all the I/O operations that a specific device requires unless that storage node is down or disabled, in that case the secondary node preference takes over IO to that device.
* If a file system is allocated all NSDs from the same preferenced storage node, the other storage node will not help balance any work on the back end.

Figure 5-4 A SONAS file system that is built on default settings

5.2.7 Placing metadata on SSD

Put metadata on the fastest tier (for most implementations and services). It is common to put both data and metadata on the same storage. Consider a few scenarios to drive the point on preferred practice.

Scenario 1

The client has a cluster with three interface nodes, two storage nodes, and one storage controller with 120 SAS drives.

Here, there are not many choices:

- Option 1. Put all devices in two failure groups. Put both metadata and data on all devices. If you need to speed up GPFS metadata scan rate, you must add more spindles and restripe the file system. Alternatively, add a new tier and move some performance-contentious data off the primary tier of storage to reduce the competition for metadata scans on read/write heads.
- Option 2. Use several NSDs from the mix for metadata only. This means that the metadata-only devices do not get busy with data read/write I/O, and metadata scans use only those designated devices (free from normal data access contention).

Be sure that you have enough read and write heads in this approach to satisfy both data types. In solutions with a huge amount of spindles, option one might be faster. In a solution with high user contention, option two might work better.

Either scenario can be improved by adding spindles to that pool and data allocation usage type. For any pool and usage type, there should be four or more devices per storage node.

Scenario 2

The client has a cluster with three interface nodes, two storage nodes, and one storage controller with 60 SAS drives and 60 NLSAS drives.

Here you have the same choice as Scenario 1, but now you can use SAS or NLSAS for metadata:

- Option 1. Put all devices in two failure groups. Put both metadata and data in an all-SAS device storage pool. If you need to speed up the GPFS metadata scan rate, you need to add more spindles to that pool and restripe the file system. Put the NLSAS devices into a silver pool for tier 2 data only placement.
- Option 2. Use a few SAS NSDs from the mix for metadata only. This means that the devices do not get busy with data read/write I/O and metadata scans, but instead use only those designated devices for each specific workload.

When you have enough read/write heads in this approach to satisfy both data types, the storage is optimally configured. In solutions with a huge number of spindles, option 1 might be faster. In a solution with high user contention, option 2 might work better.

Again, either scenario can be improved by adding spindles to that pool and data allocation type.

Scenario 3

You have a cluster with three interface nodes, two storage nodes, and one storage controller with 20 SSD drives and 100 NLSAS drives.

Here, you might use the 20 SSD drives for metadataOnly:

- Option 1. Put all devices in two failure groups. Assign `metadataOnly` on all SSD devices in the system storage pool. If you need to speed up GPFS metadata scan rate, you need to add more devices to that pool and restripe the file system. Then, put all NLSAS NSDs in the system pool as dataOnly usage type. This placement can greatly improve the performance of GPFS scan rates for tasks that are metadata-intensive, like backup and restore, replication, snapshot deletion, or antivirus.

> **Consider:** If the client tier 1 is on SAS, SSDs might not hugely improve scan rates in this case. It might be better to have more SAS spindles that share both data and metadata workloads.
>
> **Preferred practice:** In all cases, when you assign a designated `metadataOnly` set of devices in the `system` storage pool, you want to make sure that each storage node sees at least eight devices of that type to optimize the node data path and channels, and the back-end storage controller cache. In some cases, 16 devices per storage node pair (for metadata) has tested as an optimal performance minimum. Therefore, sizing your storage volumes for achieving this minimum is a good exercise before putting your file system into production.

Summary

This section describes several options. When you plan to use a large number of SAS drives for your tier 1 storage, a preferred practice is to put your metadata and data on as many of the fastest tier of drive types possible, and to replicate all metadata. If you plan to use NLSAS drives for your tier 1 data storage, and you have processes planned that require fast metadata scan rates, it is best to allocate a subset of SSD drives for `metadataOnly` (approximately 5%) in your `system` pool and NLSAS for `dataOnly` (also in your `system` pool).

XIV Gen2: The XIV Gen2 does not support an SSD option. Therefore, placing data and metadata in the system pool is common, and a preferred practice for high performance and high reliability on that platform.

XIV Gen3: The XIV Gen3 is considerably faster than the XIV Gen2, and, although the SSD option to XIV Gen3 is not directly assigned as volumes provisioned to SONAS, using the SSD option with SONAS XIV Gen3 can significantly improve data and metadata random read performance. Allocate XIV Gen3 storage to the system pool for data and metadata when the Gen3 is used with the SSD enabled volume option. Considerable improvement can be achieved with metadata and small-file random I/O workloads.

To yield the fullest value from SONAS with XIV Gen3 or DS8x00, manage only one XIV Gen3 behind each SONAS storage node pair. Highest performance from the storage is yielded when placing the XIV behind four storage nodes (this might require special consultation with SONAS implementation experts on the XIV or DS8x00 platforms). However, in this case, half of the volumes from that XIV are mapped to one storage node pair, and the other half are mapped to the second storage node pair.

5.2.8 Block size considerations

For a preferred-practice SONAS file system configuration for small file, mixed small, and large file random or sequential work loads, you can consider a few changes from default or typical installation configurations.

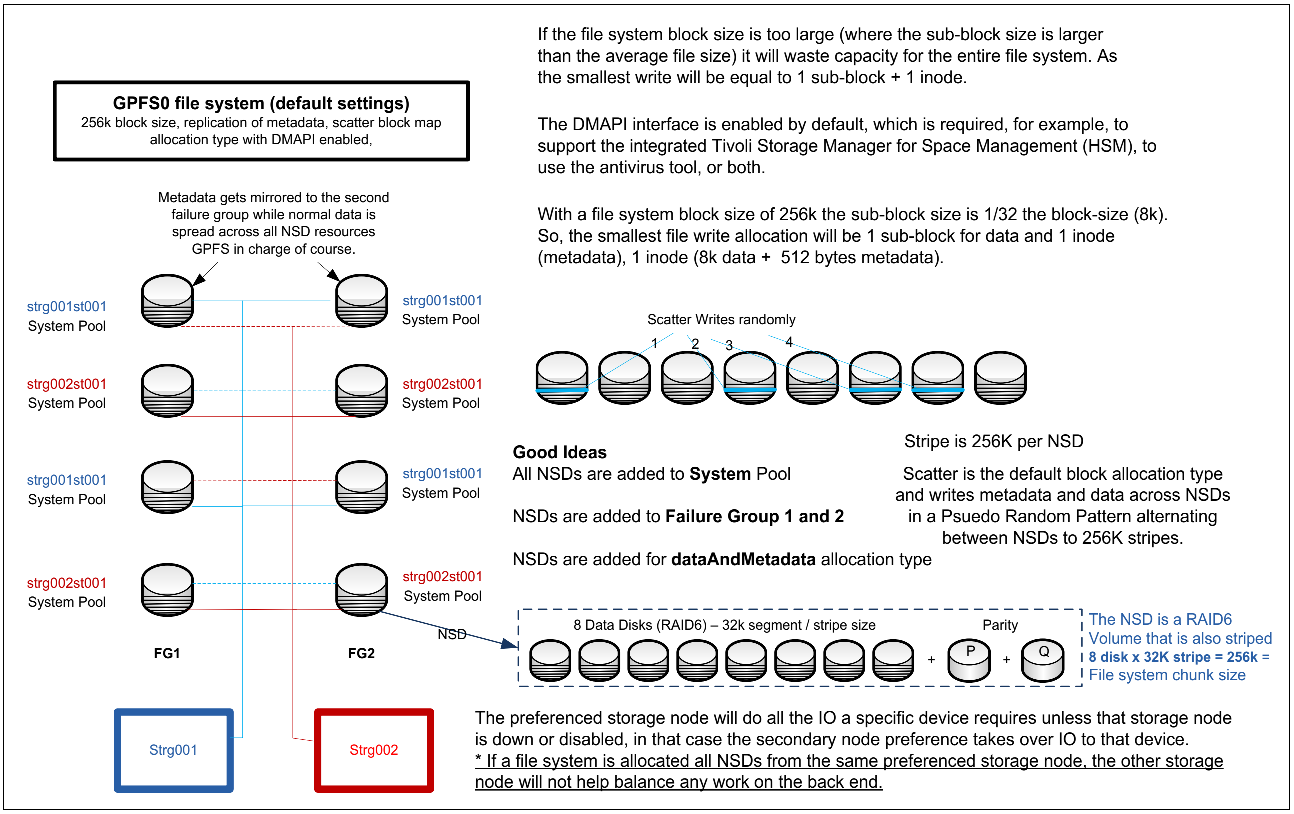
If many or most of the files are small, you also need to keep the GPFS subblock size small. Otherwise, you risk wasting significant capacity for files that are smaller than the file system subblock size. This is because all files use at least one subblock for each write. So when writing a new 2 KB file to a file system that is set to 1 MB file system block size, it uses 32 KB of disk space because that is the GPFS file system subblock size (1/32 of the assigned block size) and you cannot write to a smaller space.

You can modify that file to growth up to 32 KB without using more capacity, but consider the potential waste. However, files that are larger than the block size will waste performance from the extra read/write operations in patching together the larger files. These points are not uncommon to any file system layout and performance or capacity consideration.

Note: For SONAS gateway solutions with an XIV back end, it is more of a one size fits all process regarding the back-end segment size of the storage. However, the file system block size can still provide an effect on performance and capacity consumption. Gen3 XIV with SSD does an excellent job with small block I/O, where XIV Gen2 does a good job for larger blocks and more sequential workloads.

XIV overall provides more persistent performance, even if there are drive failures, and the fastest recovery time to full redundancy in RAID protection. XIV Gen3 shows the highest persistent performance and client satisfaction in the field today, and it is the easiest storage to configure and maintain.

For a preferred-practice SONAS file system configuration for large-file and more sequential work loads, you can consider a different change from the default settings (see Figure 5-5).



GPFS0 file system (default settings)
256k block size, replication of metadata, scatter block map allocation type with DMAPI enabled,
strg001st001
System Pool
strg002st001
System Pool
strg001st001
System Pool
strg002st001
System Pool
FG1
FG2
NSD
Strg001
Strg002
Scatter Writes randomly
1
2
3
4
8 Data Disks (RAID6) – 32k segment / stripe size
Parity
P
Q
+
+

Metadata gets mirrored to the second failure group while normal data is spread across all NSD resources GPFS in charge of course.

If the file system block size is too large (where the sub-block size is larger than the average file size) it will waste capacity for the entire file system. As the smallest write will be equal to 1 sub-block + 1 inode.

The DMAPI interface is enabled by default, which is required, for example, to support the integrated Tivoli Storage Manager for Space Management (HSM), to use the antivirus tool, or both.

With a file system block size of 256k the sub-block size is 1/32 the block-size (8k). So, the smallest file write allocation will be 1 sub-block for data and 1 inode (metadata), 1 inode (8k data + 512 bytes metadata).

Stripe is 256K per NSD

Scatter is the default block allocation type and writes metadata and data across NSDs in a Psuedo Random Pattern alternating between NSDs to 256K stripes.

Good Ideas
All NSDs are added to **System** Pool

NSDs are added to **Failure Group 1 and 2**

NSDs are added for **dataAndMetadata** allocation type

The NSD is a RAID6 Volume that is also striped **8 disk x 32K stripe = 256k =** File system chunk size

The preferenced storage node will do all the IO a specific device requires unless that storage node is down or disabled, in that case the secondary node preference takes over IO to that device.
* If a file system is allocated all NSDs from the same preferenced storage node, the other storage node will not help balance any work on the back end.

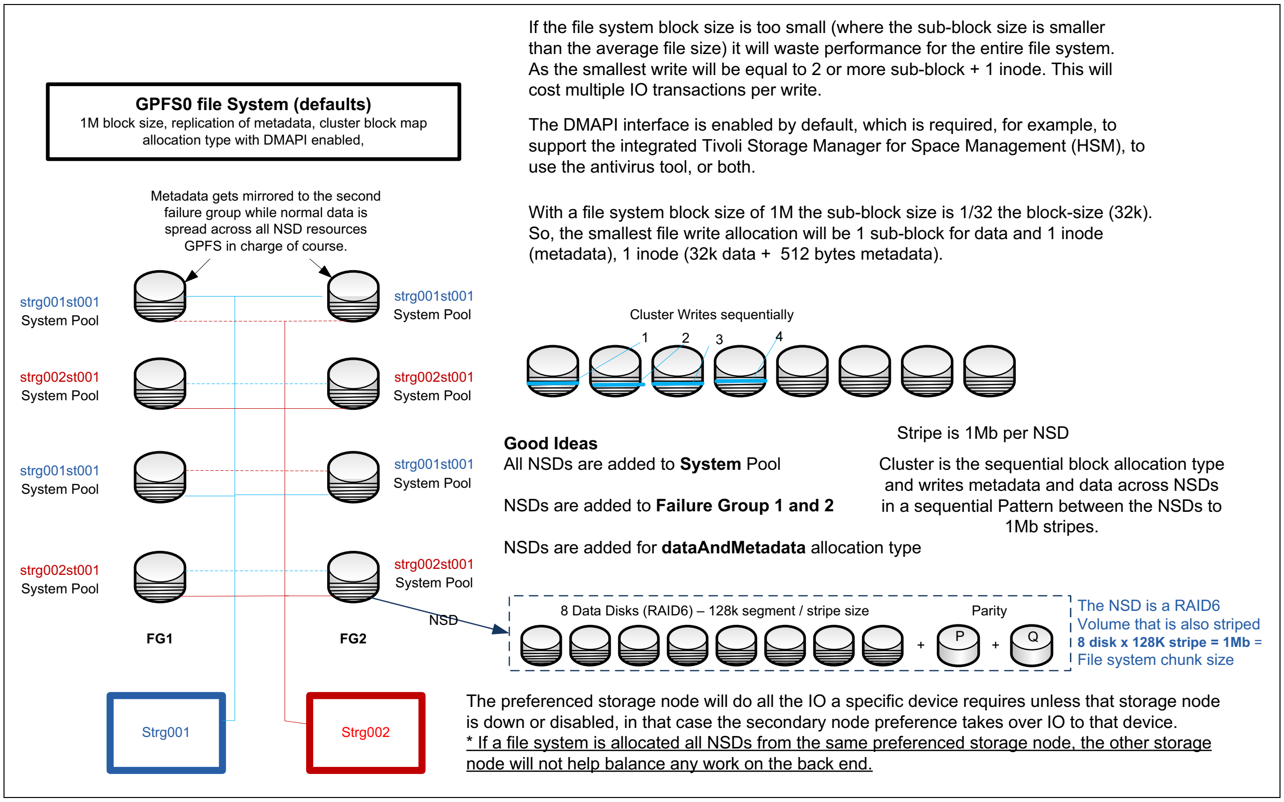
Figure 5-5 Preferred practice for small or mixed file sizes

If many or most of the files are large (> 256 KB), keep the GPFS subblock small enough to make the best use of space, but large enough to avoid wasting significant performance by breaking up the writes for I/O files. So, when writing a file, always take multiple subblocks.

Of course, if balanced well, you push a better performance. However, if files take a little more space than the subblock size, you can still waste a little capacity on the carry-over bits. These larger file writes and reads tend to be more sequential in nature. Sequential workloads tend to read and write more efficiently to a GPFS *cluster* block allocation map type.

Because XIV acknowledges writes from its large distributed cache, it is the fastest performance for back-end storage options with large file, sequential workloads of all the storage solutions that are available to SONAS today.

Now consider a file system that is designed for service where the average file size is large (> 10 MB), as shown in Figure 5-6.



GPFS0 file System (defaults)
1M block size, replication of metadata, cluster block map allocation type with DMAPI enabled,

Metadata gets mirrored to the second failure group while normal data is spread across all NSD resources GPFS in charge of course.

If the file system block size is too small (where the sub-block size is smaller than the average file size) it will waste performance for the entire file system. As the smallest write will be equal to 2 or more sub-block + 1 inode. This will cost multiple IO transactions per write.

The DMAPI interface is enabled by default, which is required, for example, to support the integrated Tivoli Storage Manager for Space Management (HSM), to use the antivirus tool, or both.

With a file system block size of 1M the sub-block size is 1/32 the block-size (32k). So, the smallest file write allocation will be 1 sub-block for data and 1 inode (metadata), 1 inode (32k data + 512 bytes metadata).

Good Ideas
All NSDs are added to **System** Pool

NSDs are added to **Failure Group 1 and 2**

NSDs are added for **dataAndMetadata** allocation type

Stripe is 1Mb per NSD

Cluster is the sequential block allocation type and writes metadata and data across NSDs in a sequential Pattern between the NSDs to 1Mb stripes.

The NSD is a RAID6 Volume that is also striped
8 disk x 128K stripe = 1Mb =
File system chunk size

The preferenced storage node will do all the IO a specific device requires unless that storage node is down or disabled, in that case the secondary node preference takes over IO to that device.
* If a file system is allocated all NSDs from the same preferenced storage node, the other storage node will not help balance any work on the back end.

Figure 5-6 Preferred practice large file system layout

In a typical configuration, you can use a 4 MB file system block size for clients that use predominantly large files. These workloads are typically sequential, and you can minimize performance reduction by writing to larger chunks in fewer transactions. Sequential workloads tend to write more efficiently to a GPFS cluster block allocation map type.

When you are deciding on the block size for a file system, consider these points:

- Supported SONAS block sizes are 256 KB, 1 MB, or 4 MB. Specify this value with the character K (Kilobytes) or M (megabytes); for example, 4M. The default is 256K.
- The block size determines the following values:
  - The minimum disk space allocation unit. The minimum amount of space that file data can occupy is a subblock. A subblock is 1/32 of the block size.
  - The block size is 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 degraded, especially for write operations.
  - In file systems with a high degree of variance in the sizes of the files within the file system, using 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 and an 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 usually insignificant.

Looking at the preceding images, you can see that when you are setting up a file system optimally you need to share the RAID configuration below it to optimize performance to align the data striping of the file system to match the design of the RAID configuration below the volumes.

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-file, sequential read and write workloads.
- A smaller block size is likely to offer better performance for small file, 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 of a given size, 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 of 256 KB, you might also want to increase the page pool size in proportion to the block size.

  Remember that this increase requires special approval through a request for price quotation (RPQ), and clearly warrants value validation through testing.
- The file system block size must not exceed the value of the maxblocksize configuration parameter, which is 4096 KB (4 MB).

GPFS uses a typical file system logical structure, common in the UNIX space. A file system is essentially composed of a storage structure (data and metadata) and a file system device driver. It is in fact a primitive database that is used to store blocks of data and keep track of their location and the available space.

The file system can be journaled (all metadata transactions are logged in to a file system journal) for providing data consistency in case of system failure. GPFS is a journaled file system with in-line logs (explained later in this section).

Note: Network Data Management Protocol (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. Therefore, *do not* use a larger block size if you plan to use NDMP backups.

5.2.9 Segment size considerations

Of the storage options available as back-end devices for the SONAS solution, the XIV platform is the only one that does not allow choosing the segment size or RAID type that is used for your file system storage solution.

For DDN storage, these parameters are chosen for you. Therefore, the platforms that do allow setting these parameters are the Storwize V7000, the IBM DCS3700, and the DS8*x*00.

The segment size of the back-end storage should clearly be aligned with the RAID type, spindle count, and file system block size for preferred practice performance.

In this explanation, RAID 6 (8 + P + Q) is used because this is a preferred practice data storage configuration for data on non-XIV storage solutions. It is a preferred practice because it offers the best overall performance and capacity preservation with two-drive fault tolerance.

Segment size effect

Now consider the science of the segment size effect.

If the file system has a 1 MB block size, SONAS tries to write data in prescribed subblock sizes (which in this case would be 32 KB). However, a full write is considered 1 MB.

If the file system has a 256 KB block size, SONAS tries to write data in prescribed subblock sizes (which in this case would be 8 KB). However, a full write is considered 256 KB.

So, if you write data in a 256 KB write from the host to a RAID 6 volume that is made up of 8 data drives plus 2 Parity (P and Q), the write is striping across the 8 data drives. The host write should ideally be divided into 8 drive writes. A 256 KB write divided by 8 drives has 32 KB segments. In this case, splitting the 256 KB stripe into 32 KB segments is ideal.

For a 1 MB file system block, that same division equates to a 128 KB segment size because the full 1 MB write would be divided into 128 KB segments that align evenly across the 8 data drives.

If the segment size is 32 KB and the data write from the host is 1 MB, it takes 4 writes across all 8 drives (plus parity) to complete the write, because 8 drives x 32 KB segments x 4 = 1 MB.

Therefore, by aligning the segment size in the RAID group to the file system block size and data stripe, you can avoid more writes in I/O transactions and make your storage performance more efficient.

5.2.10 Allocation type

Consider the allocation type before file system creation because it cannot be changed after a file system is created. There are basically two options to choose from with SONAS GPFS file systems:

- `-j {cluster | scatter}`
- `-j` Specifies the block allocation map type. The block allocation map type cannot be changed after the file system is created. The variables are as follows:
  - The `cluster` option. Using the cluster allocation method might provide better disk performance for some disk subsystems in relatively small installations. The cluster allocation stores data across NSDs in sequential striping patterns.
  - The `scatter` option. Using the scatter allocation method provides more consistent file system performance by averaging out performance variations because of block location. (For many disk subsystems, the location of the data relative to the disk edge has a substantial effect on performance.) The scatter allocation pattern attempts to balance a scattered pattern of writes across NSDs to provide the maximum spread of data to all channels in a pseudo-random pattern.

In most cases, using a scatter allocation type can yield better performance across large distributions of NSDs with mixed workloads. Therefore, unless the data profiles are large files and sequential workloads, it is usually better to use a scatter allocation type, which is the GPFS default assignment.

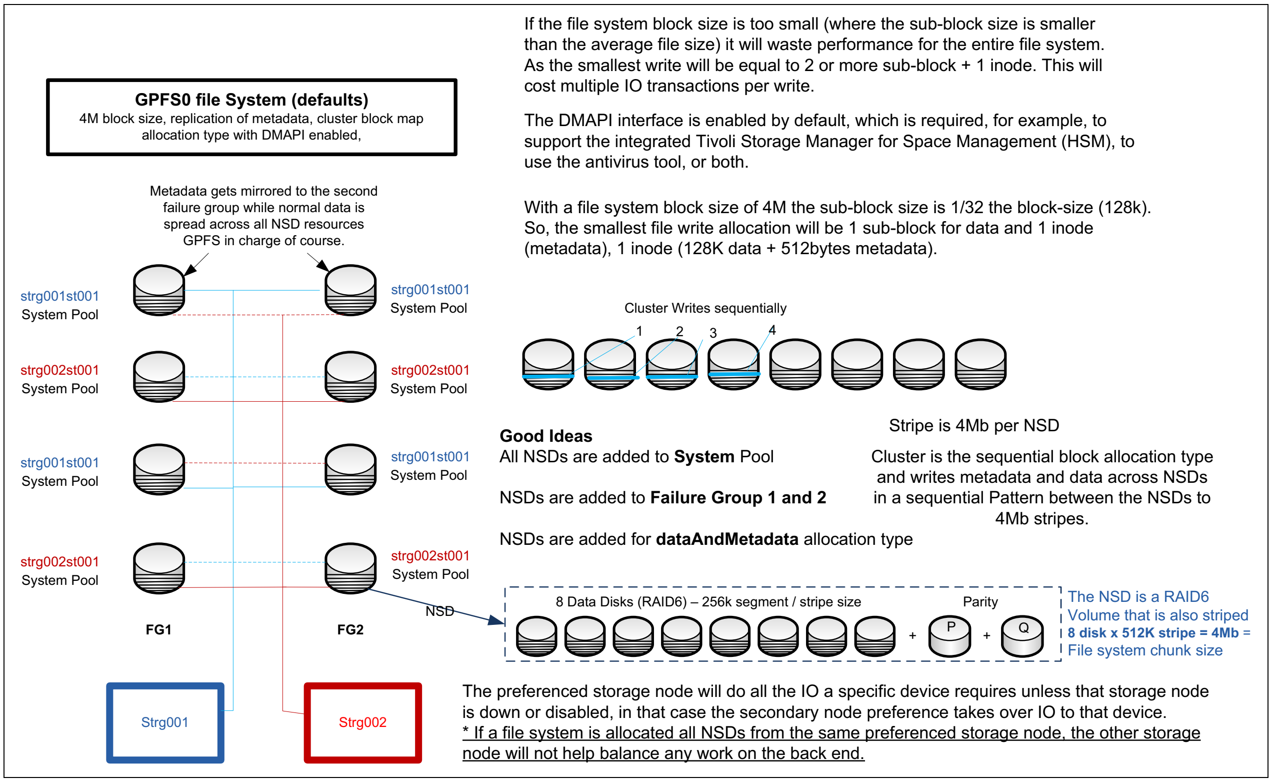
Summary of file system block size and allocation type preferred practice

Clients with many small files waste a considerable amount of capacity when they use a large file system block size. Clients with heavy read and write patterns of small files often perform best with a 256 KB block and scatter allocation type.

Clients with many large files degrade performance if they use a small file system block size. Clients with large files tend to use sequential data access patterns, and therefore often benefit from 1 MB block size and the scatter allocation type. However, if the large file manager also has many small files and they are space-sensitive, it might benefit them (from a capacity usage perspective) to use 256 KB file system block sizes with scatter access patterns:

- Workloads that are sequential with file systems smaller than 60 TB might benefit from the cluster allocation type.
- Only file patterns that are confirmed large and mostly sequential should use a 4 MB file system block size.
- The 4 MB block size should not be used when NDMP is planned for backup.

Figure 5-7 shows considerations for a huge SONAS file system.



GPFS0 file System (defaults)
4M block size, replication of metadata, cluster block map allocation type with DMAPI enabled,

Metadata gets mirrored to the second failure group while normal data is spread across all NSD resources GPFS in charge of course.

If the file system block size is too small (where the sub-block size is smaller than the average file size) it will waste performance for the entire file system. As the smallest write will be equal to 2 or more sub-block + 1 inode. This will cost multiple IO transactions per write.

The DMAPI interface is enabled by default, which is required, for example, to support the integrated Tivoli Storage Manager for Space Management (HSM), to use the antivirus tool, or both.

With a file system block size of 4M the sub-block size is 1/32 the block-size (128k). So, the smallest file write allocation will be 1 sub-block for data and 1 inode (metadata), 1 inode (128K data + 512bytes metadata).

Good Ideas
All NSDs are added to **System** Pool

NSDs are added to **Failure Group 1 and 2**

NSDs are added for **dataAndMetadata** allocation type

Stripe is 4Mb per NSD

Cluster is the sequential block allocation type and writes metadata and data across NSDs in a sequential Pattern between the NSDs to 4Mb stripes.

The NSD is a RAID6 Volume that is also striped **8 disk x 512K stripe = 4Mb** = File system chunk size

The preferenced storage node will do all the IO a specific device requires unless that storage node is down or disabled, in that case the secondary node preference takes over IO to that device.
* If a file system is allocated all NSDs from the same preferenced storage node, the other storage node will not help balance any work on the back end.

Figure 5-7 Preferred practice huge file system layout

5.2.11 Quota enablement set with file system creation

This section describes preferred practices for enabling quotas when the file system is created with the **mkfs** command. The following options are available:

```
-q, --quota { filesystem | fileset | disabled }
```

Enables or disables the ability to set a quota for the file system.

File systems must be in an unmounted state to enable or disable quotas with the `-q` option.

The file system and file set quota system helps you to control the allocation of files and data blocks in a file system or file set.

File system and file set quotas can be defined for:

- Authenticated individual users of an IBM SONAS file system.
- Authenticated groups of users of an IBM SONAS file system.
- Individual file sets.
- Authenticated individual users within a file set.
- Authenticated groups of users within a file set.
- The `filesystem` option, the default value, enables quotas to be set for user, groups, and each file set. A user and group quota applies to all data within the file system regardless of the file set.
- The `fileset` option enables user and group quotas to be set on a per file set basis. This option also enables file set level quotas, therefore restricting total space or files that are used by a file set.
- The `disabled` option indicates that a change in this setting requires the file system to be unmounted.

Because quota information in the database is not updated in real time, you must submit the **chkquota** CLI command as needed to ensure that the database is updated with accurate quota information. This update is critical after you enable or disable quotas on a file system, or after you change the basis of quotas in a file system to enable or disable quotas per file set within the file system by using the **chfs** CLI command with the `-q` option. The file system must be in an unmounted state when the **chfs** CLI command is submitted with the `-q` option.

The default setting enables quotas on a file system basis.

Reaching quota limits

When you reach the soft quota limit, a grace period starts. You can write until the grace period expires, or until you reach the hard quota limit. When you reach a hard quota limit, you cannot store any additional data until you remove enough files to take you below your quota limits or your quota is raised. However, there are some exceptions:

- By default, file set quota limits are enforced for the root user.
- If you start a file create operation before a grace period ends, and you have not met the hard quota limit, this file can take you over the hard quota limit. The IBM SONAS system does not prevent this behavior.
- If quota limits are changed, you must recheck the quotas to enforce the update immediately (use the **chkquota** command).

It is not advised to run **chkquota** command during periods of high workloads. However, it is advised that quotas are monitored along with file system, file set, and storage pool capacities by SONAS or v7000 Unified administrators on a daily and weekly basis as a standard component of monitoring cluster health.

5.2.12 DMAPI enablement

This section describes preferred practices for enabling quotas when DMAPI is enabled with the **mkfs** command. The following options are available:

```
--dmapi | --nodmapi
```

If you use the `-b` option to specify a block size that is different from the default size, be aware that the 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 greater than 1 MB.

`<`*`list of disks`*`>` is a set of the disk names that are displayed by the **lsdisk** command in the previous section. The names are separated by commas.

The `--dmapi` and `--nodmapi` are keywords that are used to designate that support for external Storage Pool Management software (for example, IBM Tivoli Storage Manager) is enabled (`--dmapi`) or disabled (`--nodmapi`).

Specify either `--dmapi` or `--nodmapi` on file system creation. If you do not specify it on file system creation, the default is `--dmapi enabled` (which is required for supporting backup, external space management, and NDMP).

> **Important:** If DMAPI is disabled, the client cannot recall data in space management with Tivoli Storage Manager and Hierarchical Storage Manager (HSM). Enabling and disabling DMAPI requires that the file system is unmounted.

5.2.13 File system logging

This section describes preferred practices for enabling quotas when DMAPI is enabled with the **mkfs** command. The following options are available:

```
--logplacement { pinned | striped }
```

These options set the data blocks of a log file, which is to be striped, across all metadata disks, like all other metadata files:

- The `pinned` option indicates that the data blocks are not striped. Use pinned logs only for small file systems that are spread across few NSDs.
- The `striped` option, the default, indicates that the data blocks of a log file are striped across all metadata disks. If not specified, the default is `striped` and, if this property is set to `striped`, it is *not* possible to turn it back to `pinned`. Striped is the preferred practice high-performance solution for log placement for most file system work loads with eight or more NSDs.

The IBM SONAS system by default stripes data blocks for file system log files across all metadata disks, like all other metadata files, for increased performance. When you create a file system, you can specify that the file system log file data blocks instead are pinned, and later change to striped, *but when striped, you cannot change to pinned.*

When you are creating the file system with the **mkfs** command, the `--logplacement { pinned | striped }` option decides how you want your logging done. Typically, striped logging runs faster than pinned, and striped is the default.

5.2.14 Snapshot enablement

This section describes preferred practices for enabling quotas when snapshots are enabled with the **mkfs** command. The following options are available:

--snapdir | --nosnapdir

This selection enables or disables the access to the snapshots directories. The --snapdir option to the **mkfs** command is the default value. This setting creates a special directory entry named .snapshots in every directory of the file system and in independent file sets. The .snapshots directory entry enables the user to access snapshots that include that directory.

If the --nosnapdir option is used, access to all snapshots can be done only by using the .snapshots directory entry at the root of the file system or file set.

For example, if a file system is mounted at /ibm/gpfs0, the user needs to use the /ibm/gpfs0/.snapshots directory entry.

If, for any reason, the default .snapshot directory is changed, remember that the default backup and replication exclude file might also require changes to reflect the new .snapshot directory location. Otherwise, backups might *not* get excluded. The amount of data that you back up to tape reflects the number of snapshots, and can be 10x or 100x the size of the data footprint, because each snapshot file gets backed up without space efficiency. Therefore, the preferred practice is to use the default location for snapshot directories.

Automatically generated snapshots are managed by rules for creation and retention-based deletion. It is best to avoid using the same snapshot rules for all file system and independent file set snapshots. If the same rules are used, it increases the amount of metadata that must be scanned.

It also increases the work that is associated in snapshot deletion operations in that specific time frame. Therefore, during the snapshot deletion operation, metadata transactions and data I/O can be so intense that they affect normal user data access or other operations, such as backup, replication, or antivirus scanning.

When you are creating file set snapshot rules, try to break up the number of file sets that share the same rules at any given time, and stagger the snapshot times to avoid or limit contention.

Rule example

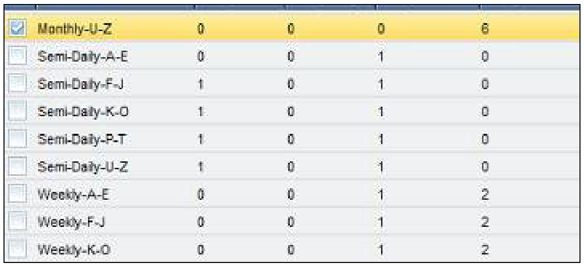
Here is an example of a rule that is defined at one customer site to simplify and stagger the snapshot management policies, and yet accomplish enough user-recoverable file protection.

All independent file sets are on an automatic snapshot schedule for semi-daily snapshots, weekly snapshots, and monthly snapshots. Snapshots are not run on the underlying file system directly, but instead the root file set of the underlying file system is snapshot.

The root file set snapshots of the underlying file system capture any data that is not managed in independent file sets (so that it captures dependent file set data).

In this example, the *snapshot rule* that is applied to each file set is assigned alphabetically in groups to ensure that not all file sets get applied snapshots or snapshot deletions occurring at the same time of day. This enables the schedules to stagger, and distributes the heavy metadata scan requirements that are associated with snapshots at varying times in the daily schedule.

A file set, whose name begins with the letters A - E, is assigned a semi-daily, weekly, and a monthly snapshot rule that ends in A - E (see Figure 5-8).



☑	Monthly-U-Z	0	0	0	6
□	Semi-Daily-A-E	0	0	1	0
□	Semi-Daily-F-J	1	0	1	0
□	Semi-Daily-K-O	1	0	1	0
□	Semi-Daily-P-T	1	0	1	0
□	Semi-Daily-U-Z	1	0	1	0
□	Weekly-A-E	0	0	1	2
□	Weekly-F-J	0	0	1	2
□	Weekly-K-O	0	0	1	2

Figure 5-8 GUI for example snapshot rule distribution

The *semi-daily* snapshot takes a snapshot at roughly 5 am or 6 am and 5 pm or 6 pm daily, and these snapshots are automatically deleted after one week. The *weekly* snapshot takes a snapshot at roughly 9 pm on a weekend and these snapshots are automatically deleted after five weeks of age. The *monthly* snapshot takes a snapshot at roughly 1 am on a weekend and these snapshots are automatically deleted after six months of age.

This process ensures adequate protection of file set data (restorable by clients), and yet limits the number of active snapshots to roughly 25 for any file set, while it mixes the activation time and distributes the metadata scan burden of deletion schedules.

Do *not* carry the number and frequency beyond practical value.

Most clients keep daily snapshots for 10 days, weekly snapshots for 5 or 6 weeks, and monthly snapshots for 10 months. Otherwise, the number of snapshots become overwhelming and offers more pain than value.

You must follow a naming convention if you want to integrate snapshots into a Microsoft Windows environment. Therefore, do not change the naming prefix on snapshots if you want them to be viewable by Microsoft Windows clients in *previous versions* listings.

Example details

The following example shows the correct name format for a snapshot that can be viewed on the Microsoft Windows operating system under the previous version:

```
@GMT-2008.08.05-23.30.00
```

Use the **chsnapassoc** CLI command or the GUI to change a snapshot rule association between a single snapshot rule and a single file system or independent file set.

To optionally change the rule that is associated with the file system or independent file set, use the `-o` or `--oldRuleName` option to specify the rule that should no longer be associated, and use the `-n` or `--newRuleName` option to specify the rule that should become associated with the file system or independent file set.

To optionally limit the association to an independent file set within the specified device, use the -j or --filesetName option and specify the name of an existing independent file set within the specified file system. The following CLI command example removes the association of a rule named `oldRuleName` with an independent file set named `filsetName` within a file system that is named `deviceName` and replaces the old rule in the association with a new rule named `newRuleName`:

```
# chsnapassoc deviceName -o oldRuleName -n newRuleName -j filsetName
```

Important: When a snapshot rule association is changed from an old rule to a new rule, all of the corresponding snapshot instances are immediately altered to reflect the changed snapshot rule association so that the existing snapshots continue to be managed, but according to the new rule.

If the retention attributes of the new rule are different from the previous rule, these alterations might cause more than the usual number of existing snapshots to be deleted.

If, instead, a snapshot rule association is removed and a new association is created, all of the snapshots corresponding to the old rule in the association are either deleted or marked as unmanaged.

5.2.15 inodes and metadata

The inodes and indirect blocks represent pointers to the actual blocks of data as shown in Figure 5-9.

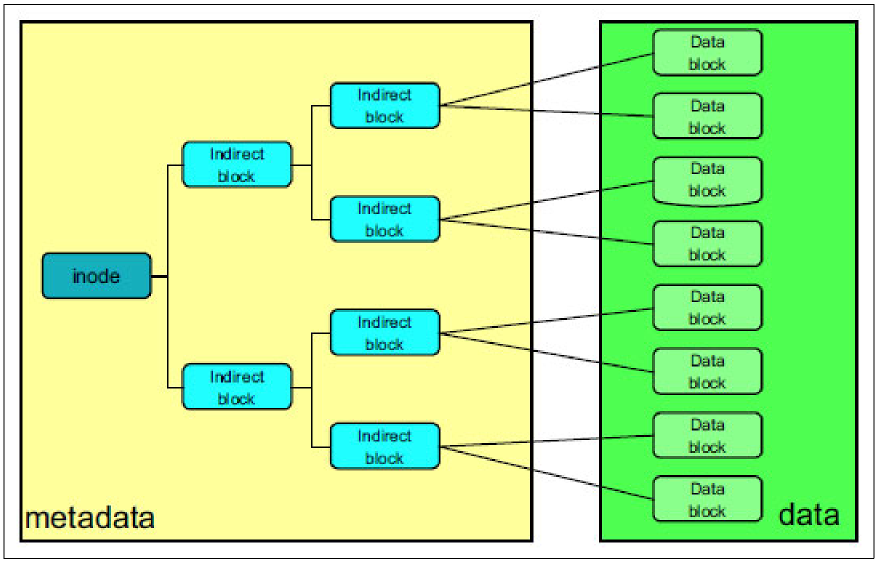


Figure 5-9 An inode to data relationship diagram

For a journaled file system, all inodes and indirect blocks that are subject to a modification are stored into the file system log before they are modified.

In addition to the inodes and indirect blocks, which keep track of where the data is stored, the file system also keeps information about the available space by using a block allocation map.

The block allocation map is a collection of bits that represent the availability of disk space within the disks of the file system, which is maintained by the FSMgr. One unit in the allocation map represents a subblock, or 1/32 of the block size of the GPFS file system. The allocation map is broken into regions, which are on disk sector boundaries. However, this is deeper than you need to go to help define the point and purpose of metadata.

The key is that metadata is small in comparison to file system data and it gets read every time a job list must be created to support a task that involves all files.

The typical preferred practice is to enable capacity for metadata at approximately 5% of the file system usable capacity (with `metadataOnly` replication enabled). Then, if the metadata is on an independent disk type from where you put data (within the system pool) (such as SSD), you need to monitor that capacity regularly and validate that your file systems and independent file sets do not run out of space or pre-allocated inodes. When a file system or independent file set runs out of inodes, you stop writing data to that store.

If you choose not to do metadata replication, metadata typically does not use more than 2.5 - 3% of the file system usable data capacity. This usage depends on how many files and objects exist in the file system. One thousand large files take much less metadata capacity than 100 million small files.

5.3 Removing a file system

Removing a file system from the active management node deletes all of the data on that file system. This is not a common task, and there are a few considerations. Use caution when you are performing this task. There is a sequence of events that should be followed before you remove a file system.

Do the following steps to remove a file system:

1. Ensure that the clients are not using shares from the file system.
2. Stop any replication of the file system.
3. Stop any backup of the file system.
4. Stop any antivirus scan scope that includes the file system.
5. Unmount shares to the file system from all clients.
6. Disable and remove all shares.
7. Remove any existing links to the file system.
8. Remove the file system.
9. Validate that the disks are no longer associated with the file system.

5.4 File system performance and scaling

When performance requirements increase, you need to see whether the back-end devices are heavily used, or if the front-end devices have maximized their throughput capacity.

If the front-end devices are at their capacity ceiling (system resource use is nearing maximum potential), adding interface nodes is a good first step. However, if the NSD performance use is high, it is important to consider growing the back end.

The sizing section of this book (1.4, "Sizing" on page 20) provides more information about how to increase back-end performance. However, there are three ways to grow performance on the back end:

1. Increase the number of spindles behind each storage node pair.
2. Increase the tier pool options to add a faster drive technology, or to offload contention of some user data to a lower-performance tier of storage.
3. Add more storage node pairs with more spindles and restripe the file system to better disperse the I/O across more devices all together.

When the storage node pair processing power is at its maximum, you need to add a storage node pair with additional spindles behind it, then restripe the file system with the **`restripefs`** command. Doubling capacity in this way can typically double the back-end performance.

Modifying file system attributes

You can use the **chfs** CLI command to modify the file system attributes. This command enables you to add NSDs, remove NSDs, add inode allocations, and set file system replication rules for data or metadata when the NSDs are in two or more predefined failure groups.

Listing the file system mount status

Use the **lsmount** CLI command to determine whether a file system mount status is mounted, not mounted, partially mounted, or internally mounted. Validating the file system mount status in the cluster is an important troubleshooting step.

Running **lsmount -v** shows the status of all nodes mounting all file systems. Occasionally, make sure that there are no inconsistencies with file system mounts. When there are issues, this should be one of the first things that is checked because it is often easily corrected.

Mounting a file system

Before you mount the file system on one interface node or all interface nodes, ensure that all of the disks of the file system are available and functioning properly. When replication is configured, all of the disks in at least one failure group must be fully available.

Unmounting a file system

You can use the GUI or the **unmountfs** CLI command to unmount a file system on one interface node or on all interface nodes.

> **Tip:** There must be nothing accessing the file system before it can be unmounted. Therefore, make certain that clients discontinue use of and unmount the file shares, then disable the shares. Then, you can unmount the file system with the **unmountfs** command from the management node. This is typically done only when removing a file system, or if support asks you to unmount it for maintenance reasons.

Restriping file system space

You can rebalance or restore the replication of all files in a file system by using the **restripefs** CLI command:

```
restriprefs gpfs0 --balance
```

It is important that data that is added to the file system is correctly striped. Restriping a large file system requires many insert and delete operations, and might negatively affect system performance temporarily. Plan to perform this task when system demand is low (if possible).

When a disk is added to an existing file system, data is moved to rebalance data across all of the disks in the file system slowly over time (running at a low priority), as files are accessed, in the background. In the case where you want to rebalance data across all of the disks in a file system immediately on demand, you can use the **restripefs** CLI command after considering the potential effect to overall system performance because I/O is performed to read and write data during the restriping operation.

By default, the restriping operation is performed using all of the nodes in the IBM SONAS system. You can lessen the overall effect to the system by using the **-n** option, specifying a subset of nodes on which the restriping operation is performed.

If you are increasing the available storage of an existing file system because the current disks are nearing capacity, consider the following guidelines when you determine whether an immediate on-demand restriping operation is appropriate and needed.

Rebalancing applies only to user data.

The file system automatically rebalances data across all disks as files are added to, and removed from, the file system.

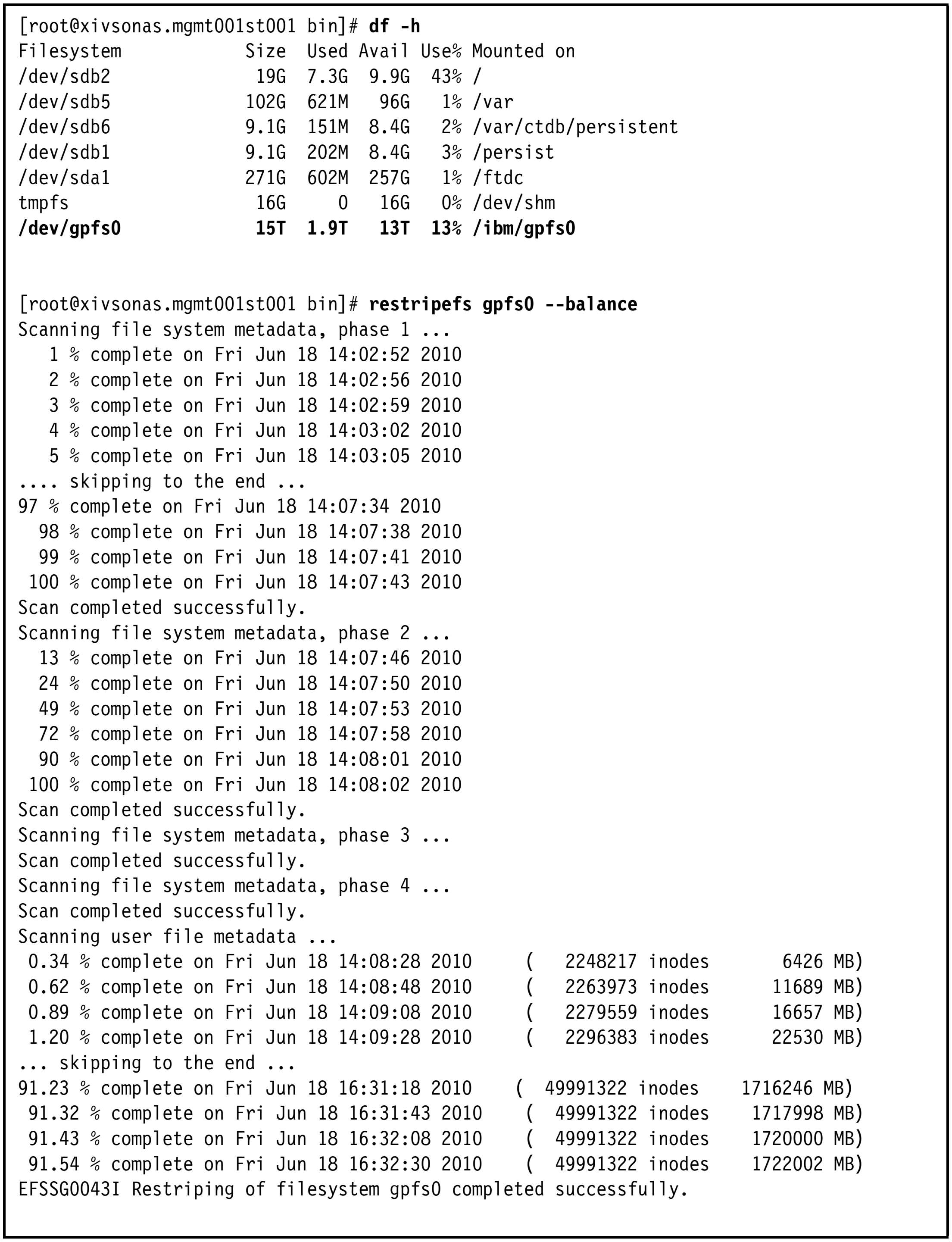
Important: If the number of disks that you are adding is small relative to the number of disks that already exist in the file system, rebalancing is not likely to provide significant performance improvements.

Forcing restriping might be necessary if all of the existing disks in a file system are already near capacity. However, if, for example, only 1 of 10 disks in a file system is full, that might be considered 90% performance. If the current level of performance is sufficient, an immediate, on-demand restriping operation might not be required.

If you are increasing the available storage of an existing file system because the current disks are nearing capacity, consider the following guidelines when you are determining whether an immediate on-demand restriping operation is appropriate and wanted:

- Rebalancing applies only to user data.
- The file system automatically rebalances data across all disks as files are added to, and removed from, the file system.
- If the number of disks that you are adding is small relative to the number of disks that already exist in the file system, rebalancing is not likely to provide significant performance improvements.
- Restriping might be necessary if all of the existing disks in a file system are already near capacity. However, if, for example, only 1 of 10 disks in a file system is full, that might be considered 90% performance. If the current level of performance is sufficient, an immediate, on-demand restriping operation might not be required.

Figure 5-10 shows output from a **restripefs** operation.



```
[root@xivsonas.mgmt001st001 bin]# df -h
Filesystem           Size  Used Avail Use% Mounted on
/dev/sdb2             19G  7.3G  9.9G  43% /
/dev/sdb5            102G  621M   96G   1% /var
/dev/sdb6            9.1G  151M  8.4G   2% /var/ctdb/persistent
/dev/sdb1            9.1G  202M  8.4G   3% /persist
/dev/sda1            271G  602M  257G   1% /ftdc
tmpfs                 16G     0   16G   0% /dev/shm
/dev/gpfs0            15T  1.9T   13T  13% /ibm/gpfs0


[root@xivsonas.mgmt001st001 bin]# restripefs gpfs0 --balance
Scanning file system metadata, phase 1 ...
   1 % complete on Fri Jun 18 14:02:52 2010
   2 % complete on Fri Jun 18 14:02:56 2010
   3 % complete on Fri Jun 18 14:02:59 2010
   4 % complete on Fri Jun 18 14:03:02 2010
   5 % complete on Fri Jun 18 14:03:05 2010
.... skipping to the end ...
97 % complete on Fri Jun 18 14:07:34 2010
  98 % complete on Fri Jun 18 14:07:38 2010
  99 % complete on Fri Jun 18 14:07:41 2010
 100 % complete on Fri Jun 18 14:07:43 2010
Scan completed successfully.
Scanning file system metadata, phase 2 ...
  13 % complete on Fri Jun 18 14:07:46 2010
  24 % complete on Fri Jun 18 14:07:50 2010
  49 % complete on Fri Jun 18 14:07:53 2010
  72 % complete on Fri Jun 18 14:07:58 2010
  90 % complete on Fri Jun 18 14:08:01 2010
 100 % complete on Fri Jun 18 14:08:02 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 ...
 0.34 % complete on Fri Jun 18 14:08:28 2010     (   2248217 inodes       6426 MB)
 0.62 % complete on Fri Jun 18 14:08:48 2010     (   2263973 inodes      11689 MB)
 0.89 % complete on Fri Jun 18 14:09:08 2010     (   2279559 inodes      16657 MB)
 1.20 % complete on Fri Jun 18 14:09:28 2010     (   2296383 inodes      22530 MB)
... skipping to the end ...
91.23 % complete on Fri Jun 18 16:31:18 2010    (  49991322 inodes    1716246 MB)
 91.32 % complete on Fri Jun 18 16:31:43 2010    (  49991322 inodes    1717998 MB)
 91.43 % complete on Fri Jun 18 16:32:08 2010    (  49991322 inodes    1720000 MB)
 91.54 % complete on Fri Jun 18 16:32:30 2010    (  49991322 inodes    1722002 MB)
EFSSG0043I Restriping of filesystem gpfs0 completed successfully.
```

Figure 5-10 Sample output from a restripefs operation

When NSD devices are added to the SONAS file system, the capacity is available immediately, but, to take full advantage of a balanced capacity and workload configuration (which you should) across all the devices in the set, you must run a **restripefs --balance** command against the file system.

This command takes some performance away from the file system to initiate (up to about 12%) during redistribution. For huge file systems, it can take a long time to complete. In most cases, it must be done to maintain balance and linearly improve performance.

Most clients choose to start this process on Friday evening and allow it to run all weekend to complete. However, the operation does not disrupt services or create a need to bring any services down. It is a nondisruptive service. Again, adding NSDs to the file system provides immediate relief to capacity issues, and restriping later redistributes for optimal load balancing. This is the preferred practice for adding capacity.

Figure 5-10 on page 166 shows a small lab system taking 2 hours and 30 minutes to redistribute data balance for 2 terabyte (TB) of data across the old and the added devices.

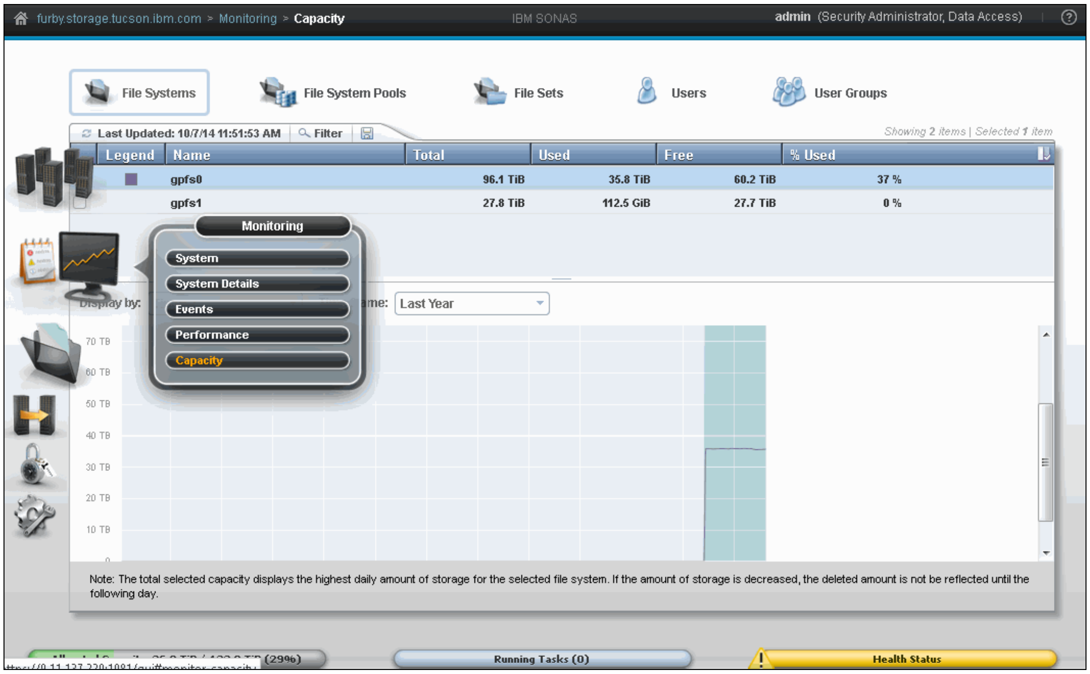
Viewing file system capacity

You can use the GUI to determine the available capacity of a file system. This task cannot be performed with a CLI command.

GUI usage

1. Log in to the GUI.
2. Click **Monitoring** → **Capacity**.
3. Ensure that the File System tab is selected.
4. Select the file system for which you want to view the system utilization. A chart that shows the system capacity in percentage displays.

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 does not display until the following day. See Figure 5-11.



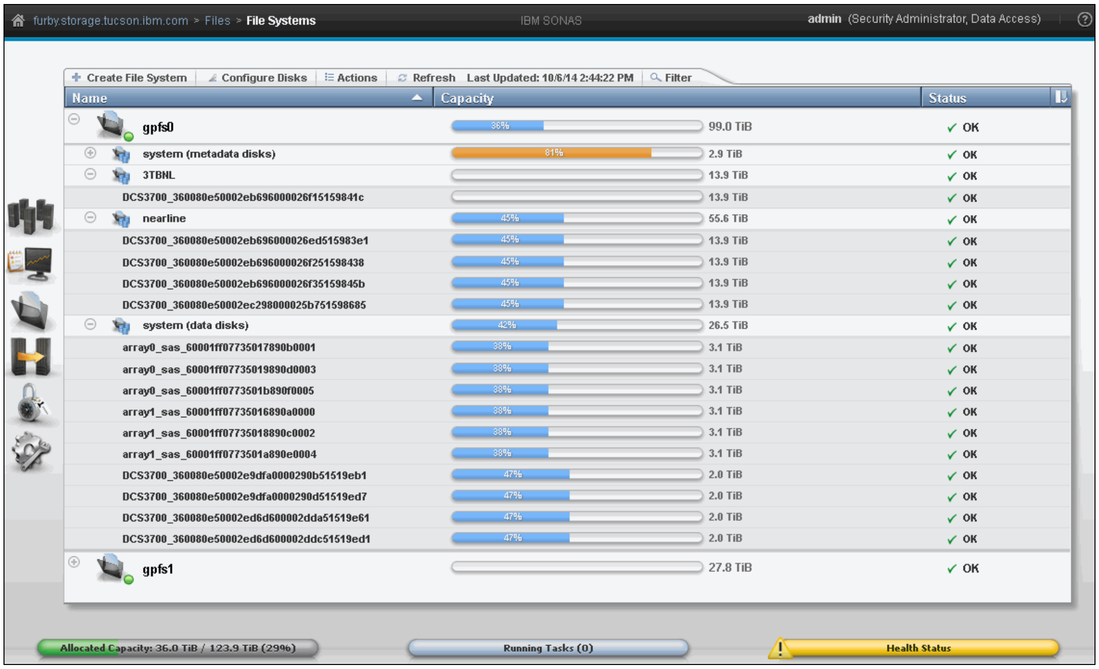
Legend	Name	Total	Used	Free	% Used
■	gpfs0	96.1 TiB	35.8 TiB	60.2 TiB	37 %
	gpfs1	27.8 TiB	112.5 GiB	27.7 TiB	0 %

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 5-11 Viewing capacity through the GUI

Perhaps a better way to look at capacity is to look at the file system statistics.

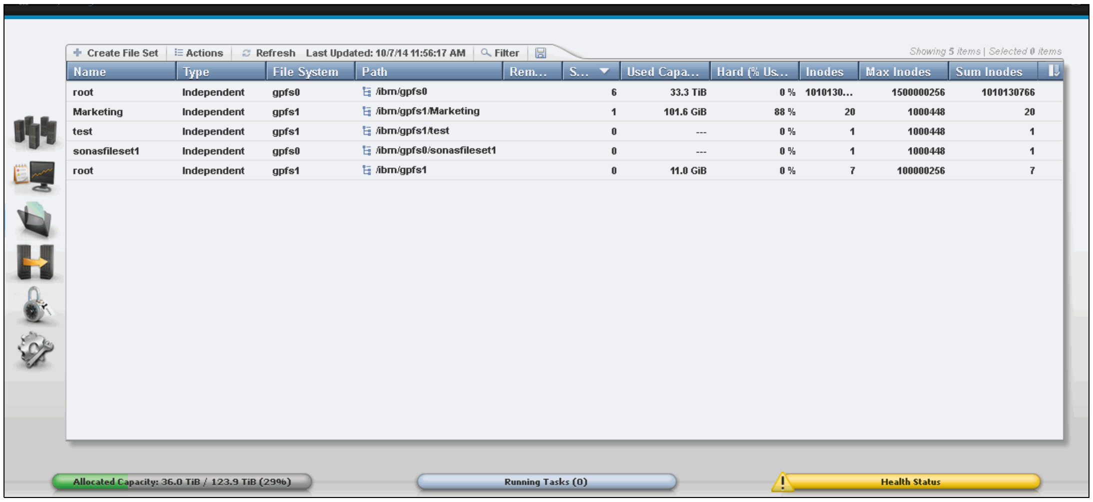
By expanding the tiers under each file system, you can see capacity that is related to the storage pools behind the file system. This solution is much more detailed (see Figure 5-12).


furby.storage.tucson.ibm.com > Files > File Systems
IBM SONAS
admin (Security Administrator, Data Access)
Create File System
Configure Disks
Actions
Refresh
Last Updated: 10/6/14 2:44:22 PM
Filter

Name	Capacity		Status
gpfs0	36%	99.0 TiB	✓ OK
system (metadata disks)	81%	2.9 TiB	✓ OK
3TBNL		13.9 TiB	✓ OK
DCS3700_360080e50002eb696000026f15159841c		13.9 TiB	✓ OK
nearline	45%	55.6 TiB	✓ OK
DCS3700_360080e50002eb696000026ed515983e1	45%	13.9 TiB	✓ OK
DCS3700_360080e50002eb696000026f251598438	45%	13.9 TiB	✓ OK
DCS3700_360080e50002eb696000026f35159845b	45%	13.9 TiB	✓ OK
DCS3700_360080e50002ec298000025b751598685	45%	13.9 TiB	✓ OK
system (data disks)	42%	26.5 TiB	✓ OK
array0_sas_60001ff07735017890b0001	38%	3.1 TiB	✓ OK
array0_sas_60001ff07735019890d0003	38%	3.1 TiB	✓ OK
array0_sas_60001ff0773501b890f0005	38%	3.1 TiB	✓ OK
array1_sas_60001ff07735016890a0000	38%	3.1 TiB	✓ OK
array1_sas_60001ff07735018890c0002	38%	3.1 TiB	✓ OK
array1_sas_60001ff0773501a890e0004	38%	3.1 TiB	✓ OK
DCS3700_360080e50002e9dfa0000290b51519eb1	47%	2.0 TiB	✓ OK
DCS3700_360080e50002e9dfa0000290d51519ed7	47%	2.0 TiB	✓ OK
DCS3700_360080e50002ed6d600002dda51519e61	47%	2.0 TiB	✓ OK
DCS3700_360080e50002ed6d600002ddc51519ed1	47%	2.0 TiB	✓ OK
gpfs1		27.8 TiB	✓ OK

Figure 5-12   File system capacity view in the SONAS GUI

By using the GUI to review file set capacity, you also get to see the inode consumption and availability. Remember every file system (root file set) and independent file set displays capacity and inode consumption in this view (see Figure 5-13).


Create File Set
Actions
Refresh
Last Updated: 10/7/14 11:56:17 AM
Filter
Showing 5 items | Selected 0 items

Name	Type	File System	Path	Rem...	S...	Used Capa...	Hard (% Us...	Inodes	Max Inodes	Sum Inodes
root	Independent	gpfs0	/ibm/gpfs0		6	33.3 TiB	0 %	1010130...	1500000256	1010130766
Marketing	Independent	gpfs1	/ibm/gpfs1/Marketing		1	101.6 GiB	88 %	20	1000448	20
test	Independent	gpfs1	/ibm/gpfs1/test		0	---	0 %	1	1000448	1
sonasfileset1	Independent	gpfs0	/ibm/gpfs0/sonasfileset1		0	---	0 %	1	1000448	1
root	Independent	gpfs1	/ibm/gpfs1		0	11.0 GiB	0 %	7	100000256	7

Figure 5-13   File set Capacity View: GUI

CLI listing of file system attributes and capacity

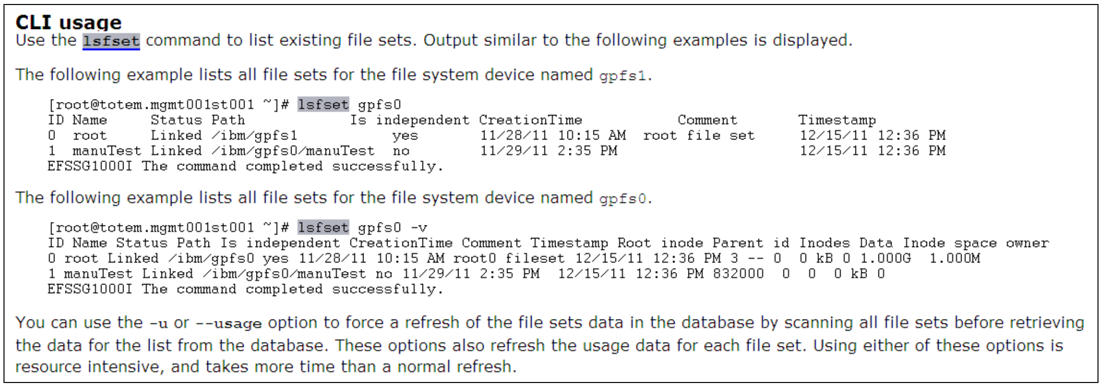
This section describes how to use the verbose option of the `lsfs` command:

```
lsfs -v
```

This is the same example as shown previously, with the verbose option added. The following columns are added:

- Min.frag.size
- Inode size
- Ind.blocksize
- Max.Inodes
- Locking type
- ACL type
- Version
- Logf. size
- Mountpoint
- Type
- Remote device
- atime
- mtime
- logplacement
- state
- snapdir

Information about the **lsfset -v** command is shown in Figure 5-14.



CLI usage

Use the lsfset command to list existing file sets. Output similar to the following examples is displayed.

The following example lists all file sets for the file system device named gpfs1.

```
[root@totem.mgmt001st001 ~]# lsfset gpfs0
ID Name     Status Path                Is independent CreationTime          Comment       Timestamp
0  root     Linked /ibm/gpfs1           yes           11/28/11 10:15 AM  root file set    12/15/11 12:36 PM
1  manuTest Linked /ibm/gpfs0/manuTest  no            11/29/11 2:35 PM                    12/15/11 12:36 PM
EFSSG1000I The command completed successfully.
```

The following example lists all file sets for the file system device 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
EFSSG1000I The command completed successfully.
```

You can use the -u or --usage option to force a refresh of the file sets data in the database by scanning all file sets before retrieving the data for the list from the database. These options also refresh the usage data for each file set. Using either of these options is resource intensive, and takes more time than a normal refresh.

Figure 5-14 Sample output from the lsfset command with -v

Summary of preferred practices for file system creation

Here are preferred practice guidelines for file system creation:

- When naming file systems, keep the names short (fewer than 63 characters), simple, and descriptive for your purposes. Do not include spaces or special characters that can complicate access or expand in the shell.
- Before provisioning storage, the RAID and RAID segment size and volume should be properly calculated to provide the best performance and capacity efficiency for the file system block size that is planned and the file data workload characteristics.
- Before assigning NSDs to the file system, make sure that they are in sufficient quantity to maximize I/O resources of the storage node HBA ports, busses, channels, and back-end storage controller channels. Also ensure that the NSDs are evenly spread across storage node preferences and two failure groups (except for single XIV applications).

  NSDs that are assigned to file systems (regardless of disk or pool type), should be assigned in groups of four per storage node (16 NSDs is the optimal minimum per storage node pair for most non-XIV type storage, and eight or 16 is optimal for XIV applications).
- NSDs must be of a consistent size and multipath structure when used in the same SONAS GPFS storage pool for the same usage type.

- The file system block size and allocation type are best planned against the average file size, the quantity or percentage of small files, and workload access patterns (random or sequential). This must be carefully considered before back-end volume provisioning and file system creation is applied.
- For small file random workloads, ensure that the file system is created with a 256 KB block size and scatter allocation type, and ensure that the underlying block device creation is assembled with appropriate configuration. Follow the guidelines in Chapter 4, “Storage configuration” on page 95.
- Using large block size definitions in file system creation when there are lots of small files wastes space. Consider this before creation.
- For large file sequential workloads, consider a 1 MB block size and general scatter allocation type unless the file system will be small and access will be sequential. In that case, choose the cluster allocation type.
- Using 4 MB limits backup capabilities, so only use the 4 MB block size when instructed by development.
- Use the striped file system logging (also described as log placement). Remember that stripe is the default.
- When you are in doubt, use the default `DMAPI enabled` setting for file system creation for best support with backup and special DMAPI enabled applications.
- When you set quota enablement, use **`-q fileset`** as the defined option, rather than the default **`file system`**. For capturing dependent file sets, run snapshots against the root file set of a file system along with the independent file sets, and *not* the file system (because that captures all dependent and independent file set data, which is not the preferred method of capturing true file set granularity.
- When you add storage for capacity and performance, it is best to add it when capacity is less than 75% used. Force a restripe when you apply a **`restripefs`** command if your capacity is over 80% when you decide to expand.
- Monitor front-end and back-end performance with a solid baseline, when your storage is set up and running well, to have something to compare to weekly or monthly snapshots as your services grow.
- Monitor file system, file set capacity, and inode consumption regularly and add allocated inodes or capacity when systems reach management action thresholds.
- If your file systems are built on gateway storage, it is important to monitor back-end storage with SONAS reports daily, and manage negative events aggressively.
- Work with your storage and support team to prepare a GPFS file system management and diagnostics readiness program.

6

Shares, exports, and protocol configuration

This chapter explains file exports or shares for the various client platforms, the different export options, and the different client mount options. It explains permissions and access control list (ACL) management for these shares and preferred practices for data access, performance, and security.

This chapter describes the following topics:

- Choices when you are creating a share or export
- Background information for creating shares and exports
- Managing authentication and ID mapping
- Establishing user and group mapping for client access
- Managing authorization and access control lists
- Changing shares and exports
- Working with Common Internet File System (CIFS) shares
- Working with Network File System (NFS) exports
- Connecting with other protocols

Shares or exports can be managed through the graphical user interface (GUI) or command-line interface (CLI) by IBM Scale Out Network Attached Storage (SONAS) administrator users or users that have a user role of `Administrator` defined in the cluster. SONAS users are not authorized to manage SONAS system shares or exports.

For example, a user, as an external user, is not able to change the name or the SONAS system configuration settings of a share or export, even if that user is the owner of the shared or exported directory. They control the data, but not the share. For that reason SONAS administrators must understand and protect the needs and requirements of the data owners when they create and manage the shares.

A share or export results from making a disk space accessible through the protocols that are specified during its creation. The following shares and exports can be created if the corresponding protocol is enabled for the system:

- Hypertext Transfer Protocol (HTTP)
- Secure Copy Protocol (SCP)
- File Transfer Protocol (FTP)
- CIFS
- NFS

Shares and exports can be created only for data that is stored in the IBM General Parallel File System (IBM GPFS) file system of the SONAS system. Non-GPFS file system content cannot be shared or exported.

6.1 Choices when you are creating a share or export

You have the following choices when you are creating a share or export:

- Sharing or exporting an existing directory that contains user files or data other than.snapshots and quota metadata, where no changes to ACLs or ownership are possible.

  > **Tip:** Because ownership of an existing file set path cannot be changed when you are creating a share or export that contains user files or data, ensure that the ownership and corresponding ACLs of all of the directories in the entire path to which the file set is linked are correct. Do this task before you create a share or export on an existing path that contains user files or data.

- Sharing or exporting a new directory, which is created during execution, and optionally specifying an owner for this newly created directory.

  > **Tip:** When a file system or file set is created, and the owner is not specified, the default owner/group/permissions of the new path is root/root/700.

  Creating a share or export on a path with owner `root` might render the share or export inaccessible to users. When you use the GUI to create a share or export, specifying an owner is not mandatory if the directory exists and the owner is not `root`. Otherwise, when you use the GUI to share or export a new directory, or an existing directory that is owned by `root`, specifying an owner for the share or export directory is mandatory.

  > **Tip:** This requirement is true even when ACLs are configured with inheritance. Owner specification does not affect ACL settings, including group ACL definitions.

Owner specification is still required when you are using the GUI to create a share or export, even though the share or export definition enables ACL manipulation on subdirectories. For example, an NFS export that is configured with the `no_root_squash` option.

6.2 Background information for creating shares and exports

To access the share or export, a user must have appropriate permissions for accessing the path, and ACL authorization to read (r) and execute (x) each directory in the full path.

If a SONAS system administrator disables the `--bypassTraversalCheck` option of the **`chcfg`** SONAS CLI command, *which is enabled by default,* retains the traversal rights entry to ensure that users are able to access the share or export and its subdirectories.

For example, to access a share or export that is mounted at `/ibm/mydir/mysubdir/myexport`, the preceding ACLs, at a minimum, must be applied to `/ibm`, `/ibm/mydir`, `/ibm/mydir/mysubdir`, and `/ibm/mydir/mysubdir/myexport`.

6.3 Managing authentication and ID mapping

To enable user read and write access to directories and files on the SONAS system, you must configure the SONAS environment for user authentication. You can configure an external server for authentication and authorization of users and groups. Only one user authentication method, and only one instance of that method, can be supported at any time. If Active Directory (AD) is configured, use the principle AD domain.

If Lightweight Directory Access Protocol (LDAP) is used and multiple LDAP servers are configured, they must be replicas of the same master LDAP server, or they can be any LDAP hosts with the same schema, which contain data that is imported from the same LDAP Data Interchange Format (LDIF) file.

6.3.1 Authentication server conditions

The following conditions must be met:

- Ensure that an authentication server external to the SONAS system is installed with the correct connectivity to and from the SONAS environment.

> **Note:** The CLI and GUI do *not* provide any means to configure or manage the external authentication server.

- You can configure *only one type of authentication server,* at any time. The following authentication servers are supported:
  - LDAP
  - LDAP with MIT Kerberos
  - SAMBA Primary Domain Controller (PDC) on Microsoft Windows NT version 4 (NT4)
  - Active Directory Server (ADS), which works as Kerberos, and AD with Microsoft Windows Services for UNIX (SFU)
  - Network Information Service (NIS) as an extension to AD/Samba PDC
  - Added support in SONAS 1.5.1.0 for local authentication server, and hosting an authentication server within SONAS
- Obtain, in advance, the administrative information as required by the implementation steps for the authentication server, such as the administration account, password, Secure Sockets layer (SSL) certificate, and Kerberos keytab file.

Important: Obtaining this administrative information is an important requirement for SONAS or IBM Storwize V7000 Unified configuration.

6.3.2 Server-side authentication configuration

The SONAS system provides server-side authentication configuration for these services:

- CIFS
- FTP
- SCP
- NFS version 3 (NFSv3)
- NFS version 4 (SONAS 1.5.1 and later)
- HTTP Secure (HTTPS)

For the configuration of NFSv3, only the protocol configuration is done. Because *authentication occurs on the NFSv3 client side*, configure the authentication on the clients that are attaching to an SONAS system. This process is a manual procedure, which is described in detail in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/usgr_cnnctng_2_clstr.html

- Only the interface nodes are configured for authentication. Storage nodes are not part of this configuration.
- The SONAS system must be cluster and network configured (as a part of initial installation) before authentication server configuration is started.
- Ensure that all SONAS nodes are in time synchronization (by using an Network Time Protocol (NTP) Server) with the authentication server. Authentication does not work correctly if the time is not synchronized across the cluster nodes.
- For Kerberos, create service principals for all of the services that are managed by the Kerberos domain controller; that is, the host, CIFS, and NFS.

6.3.3 Other authentication elements

The following other authentication elements are supported by the SONAS system:

- Netgroups. Groups of hosts are used to restrict access for mounting NFS exports on a set of hosts, and deny mounting on the remainder of the hosts. The SONAS system supports netgroups that are stored in NIS, in LDAP, in SFU as Server for NIS, or in local files.
- Kerberos. The SONAS system supports Kerberos only with AD. It also supports Kerberos optionally with LDAP, but not with the internal OpenLDAP server.
- SSL or Transport Layer Security (TLS). These protocols are primarily used to increase the security and integrity of data that is sent over the network. These protocols are based on public key cryptography and use digital certificates that are based on X509 for identification.

6.4 Establishing user and group mapping for client access

Map the Microsoft Windows security identifiers (SIDs) to the UNIX 32-bit user identifiers and group identifiers (UID and GID).

6.4.1 Central ID mapping for users and groups

A Microsoft Windows system uses SIDs internally to identify users and groups.

A UNIX system like the SONAS system requires a 32-bit UID/GID. It is necessary to map the Windows SID to the UNIX UID/GID. The mapping of Windows security identifiers is done by using the winbind system, which supports several pluggable mapping systems, depending on the customer environment. Common SID mapping schemes are the RFC2307 SFU scheme or a reserved ID range (RID) scheme. Both mapping schemes are coherent across the system.

The SONAS system supports the following authentication server integrations:

- LDAP
- LDAP with MIT Kerberos
- Samba Primary Domain Controller (PDC) on Microsoft Windows NT version 4 (NT4)
- Active Directory Server (ADS), which works as Kerberos, and Active Directory (AD) with Microsoft Windows Services for UNIX (SFU)
- Network Information Service (NIS) as an extension to AD/Samba PDC
- Local authentication server hosted within SONAS

The following types of server integration support central ID mapping**:**

- LDAP, where the UID or GID is stored in a dedicated field in the user or group object on the LDAP server

  > **Tip:** Netgroups, which are managed by LDAP, are supported.

- Active Directory with SFU, where the UID or GID is stored in a dedicated field in the user or group object on the ADS

  > **Tip:** The SFU schema extension or W2003R2 is required. NIS in extended mode can be used for netgroups. This is often considered the preferred practice method for SONAS authentication, and the easiest service to maintain overall.

- SONAS 1.5.1 supporting deterministic ID mapping and tools to maintain consistent ID maps of 2 or more SONAS clusters
- NIS as an extension to AD/Samba PDC

  > **Tip:** The SONAS system supports NIS as an extension over AD/Samba PDC in two modes, either netgroups only or netgroups and ID mapping. In the latter mode, ID mappings are stored on, and read from, the NIS server.

The UID/GID can be used by the SONAS system and also by other systems, such as NFS clients and multiple SONAS systems (including replication sites).

6.4.2 Manual ID mapping for users and groups

When central ID mapping is not established or cannot be used, such as when pure AD without SFU or a Samba PDC is used, the mapping must be manually established on every NFS client that must to attach to, and work with, the files of the SONAS system. This ensures that the ID mapping on the client is identical to the mapping from the SONAS system.

Manual mapping

To establish the manual mapping, complete the following steps:

1. Determine the UID/GID of a user that exists in the SONAS system by using the **`chkauth`** command. See the man page of the **`chkauth`** CLI command for complete usage information. For example:

   `# `**`chkauth --userName w2k3dom01\\laura -iOutput`** is displayed in the following format:

   ```
   Gid = 10000017, Uid = 10000000, Home Directory = /var/opt/IBM/sofs/scproot,
   Template Shell = /usr/bin/rssh2
   ```

2. Define users manually on the client with the same UID/GID. Edit the `/etc/passwd` file on all clients that require access. Using the example from step 1 on page 176, the following line must be added:

   ```
   laura::10000000:10000017::/:
   ```

   > **Important:** The information on the SONAS system for the home directory and template shell are more restrictive than what must be applied on the client. The previously shown example does *not* provide or apply such settings, and only describes the information that is required for UID/GID to be consistent on all the clients and the SONAS system.

6.5 Managing authorization and access control lists

Authorization grants or denies an already authenticated identity. Access control must prevent unauthorized access to SONAS system resources.

Detailed information about SONAS or Storwize V7000 Unified authentication is discussed in Chapter 2, "Authentication" on page 25. It is important to fully understand data access and authentication to understand data shares or exports from SONAS. Managing authentication and ID mapping for SONAS authentication information is described in 6.9, "Connecting with other protocols" on page 202.

6.5.1 Access Control List and Access Control Entry

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 the UNIX and Windows operating systems.

The SONAS system uses the Linux built-in ACL mechanism for access control to files that are stored on the SONAS system. Types of access include read, write, and execute. A file that can be accessed by a user is in a SONAS file system that is created using the General Parallel File System (GPFS). Example 6-1 on page 177 shows sample output from running the **`mmgetacl`** command on a file in SONAS.

Example 6-1 Sample output from running “mmgetacl” on a file in SONAS

```
[root@xivsonas.mgmt001st001 testdir]# mmgetacl file418
#NFSv4 ACL
#owner:root
#group:root
special:owner@:rw-c:allow
 (X)READ/LIST (X)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (X)CHOWN (-)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR (-)WRITE_NAMED

special:group@:----:allow
 (-)READ/LIST (-)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (-)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR (-)WRITE_NAMED

special:everyone@:----:allow
 (-)READ/LIST (-)WRITE/CREATE (-)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (-)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR (-)WRITE_NAMED
```

Note: (-) means not selected and (X) means selected in the output in Example 6-1.

Starting with SONAS 1.5.1, the on-disk format of the ACL changes, and is as shown in Example 6-2.

Example 6-2 New ACL format with SONAS 1.5.1

```
mmgetacl redbookexport2
#NFSv4 ACL
#owner:STORAGE4TEST\redbook2
#group:STORAGE4TEST\domain users
#ACL flags:
#  OWNER_DEFAULTED
#  GROUP_DEFAULTED
#  DACL_PRESENT
#  DACL_DEFAULTED
#  SACL_PRESENT
#  NULL_SACL
special:owner@:rwxc:allow:FileInherit:DirInherit
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR
(X)READ_NAMED
 (X)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR
(X)WRITE_NAMED

special:group@:rwxc:allow
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR
(X)READ_NAMED
 (X)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR
(X)WRITE_NAMED

special:everyone@:--x-:allow:DirInherit
 (-)READ/LIST (-)WRITE/CREATE (-)MKDIR (-)SYNCHRONIZE (-)READ_ACL  (-)READ_ATTR
(-)READ_NAMED
 (-)DELETE    (-)DELETE_CHILD (-)CHOWN (X)EXEC/SEARCH (-)WRITE_ACL (-)WRITE_ATTR
(-)WRITE_NAMED

user:STORAGE4TEST\redbook2:rwxc:allow:FileInherit:DirInherit:InheritOnly
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR
(X)READ_NAMED
```

```
 (X)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR 
(X)WRITE_NAMED

group:STORAGE4TEST\domain users:rwxc:allow:FileInherit:DirInherit:InheritOnly
 (X)READ/LIST (X)WRITE/CREATE (X)MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR 
(X)READ_NAMED
 (X)DELETE    (X)DELETE_CHILD (X)CHOWN (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR 
(X)WRITE_NAMED
```

Note that all users and groups must be part of the domain and known to the domain server that is used by the SONAS system. If the domain server does not recognize the user and group used in the ACL, the SONAS system does not recognize them, and does not add the correct entries to the internal database.

There are a broad range of ACL formats, which differ in syntax and semantics. The ACL format that is defined by Network File System version 4 (NFSv4) is called NFSv4 ACL.

GPFS supports the NFSv4 ACL format; this implementation is sometimes referred to as GPFS NFSv4 ACL. Therefore, this is also what is supported with SONAS and Storwize V7000 Unified.

The SONAS system stores all user files in GPFS. Access protection in the SONAS system is implemented in GPFS using NFSv4 ACLs, and is enforced for all of the protocols that are supported by the SONAS system: CIFS, NFS, FTP, HTTPS, and SCP.

The implementation of NFSv4 ACLs in GPFS does not imply that GPFS or the SONAS system supports NFSv4. The SONAS system supports NFS version 2 (NFSv2) and NFS version 3 (NFSv3), in the current release**.**

NFSv4.0 on SONAS 1.5.1: For SONAS 1.5.1 and later, support for NFSv4.0 is added. This version is the IBM implementation of the NFSv4 protocol. This implementation also supports the NFSv3 protocol.

Changes to on-disk ACL format and support for Creator Owner and Creator Group in SONAS 1.5.1

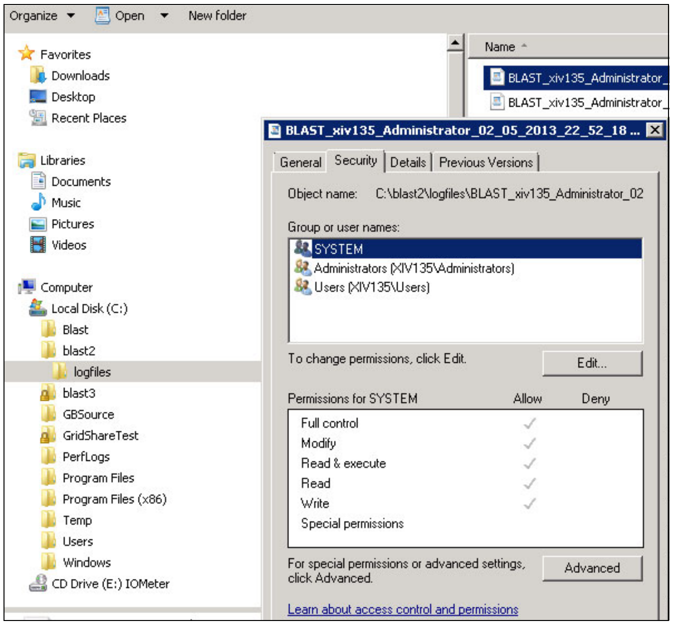
Starting with SONAS 1.5.1, the on-disk format of the ACLs is changed to improve the inheritance behavior of ACLs to match the Windows behavior, and to support Creator Owner and Creator Group entries. All newly created file systems benefit immediately from this improvement. Existing file systems must be updated as described in the *Upgrading ACL from V 1.4.x to V 1.5.1* section of the SONAS information in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonasWelcome.htm

Without an update, the inheritance behavior matches the Windows behavior only for folders with newly set ACLs, but not for all new folders. Also, the ACL update is required to enable Creator Owner ACLs.

The only method to administer ACLs in the SONAS system is with a Windows client and CIFS protocol, by using the Windows Security dialog in Windows Explorer (see Figure 6-1 on page 179) to adjust the ACLs.

Tip: SONAS v1.5.1 and later provides support to view and modify ACLs by using either the CLI or the GUI. For more information, see the Authentication chapter in *IBM SONAS Implementation Guide*, SG24-7962.



Object name: C:\blast2\logfiles\BLAST_xiv135_Administrator_02

Group or user names:

- SYSTEM
- Administrators (XIV135\Administrators)
- Users (XIV135\Users)

To change permissions, click Edit.

Permissions for SYSTEM	Allow	Deny
Full control	✓	
Modify	✓	
Read & execute	✓	
Read	✓	
Write	✓	
Special permissions		

For special permissions or advanced settings, click Advanced.

Figure 6-1   Sample session of the Windows Security Dialog pane

The owner can be changed, and has permission to read and write the ACL of the owned file regardless of the settings in the ACE. This implementation prevents a user from locking themselves out of setting the access rights for files that they own.

A correctly authorized SONAS system administrative user can change the owner, group, or both, of a file system, file set, share, or export by using the **chowner** CLI command, if the object does not contain any directories or files other than the.snapshots subdirectory and quota files. An administrative user cannot modify ACLs.

6.5.2 ACL permissions

The ACL permissions Read Permissions and Read Attributes are required to list a file.

A file owner requires only the Read Attributes permission to list a file, because the permission Read Permissions is implied. A different user must have both the Read Permissions and Read Attributes permissions enabled to reliably list the file. These permissions are both automatically granted together when read access is granted.

The inheritance of ACEs in the ACL from the owner of a directory to subdirectories and files works only for subdirectories and files that have the same owner as the parent directory.

A subdirectory or file that is created by a different owner does not inherit the ACE of a parent directory that is owned by another user. In this case, the owner of the newly created subdirectory or file must explicitly set an ACE for the newly created subdirectory or file that is appropriate for the owner of the parent directory. All of the other ACEs are inherited, whether the owner of the new file or directory is changed or retained.

Remember: The preceding restriction does not apply to SONAS version 1.5.1 and later. The inheritance behavior of ACLs matches the Windows behavior and also now supports Creator Owner and Creator Group entries.

For Microsoft Windows clients, the SONAS system maps NFSv4 ACLs to Windows NT ACLs, and does a *reverse mapping* for Windows NT ACLs that are set by Windows workstations. GPFS NFSv4 ACLs and CIFS ACLs are not equivalent.

For example, CIFS supports unlimited nested groups, which are not fully supported by GPFS NFSv4 ACLs. The SONAS system maps most of the CIFS ACL features to GPFS NFSv4 ACLs, with some limitations.

The ownership of a file cannot be migrated by using a user ID; you must configure and use an SONAS system administrative user to do data migration. When migrating existing files and directories from other systems to the SONAS system, the ACL might not contain explicit traversal rights for the users because the source system can grant this right implicitly. After migrating the files with ACLs, ensure that traversal rights are granted to the parent directory of each exported path.

6.5.3 BypassTraversalCheck privilege

One such example is the Windows BypassTraversalCheck privilege. You can enable the optional `--bypassTraversalCheck` option of the **`chcfg`** SONAS CLI command to allow CIFS users to traverse through directories without explicit ACL permissions.

Listing of files in a directory must still be permitted by the ACL read permission.

For example, in the directory structure `/A/B/C`, assume that a CIFS user has read permission on C but no permissions on A and B. When the `--bypassTraversalCheck` option is set to its default value of `yes`, this CIFS user can access C without having Traverse Folder and Execute File permissions set to allow on A and B, but it is still not allowed to browse the contents of A and B.

Tip: Setting the `--bypassTraversalCheck` option enables a user to directly access files and folders that the user owns, and that are contained under parent folders for which the user does not have Read or Write permissions. Users without Read and Execute access to the share or export in which the user-owned files and folders are located can Read and Modify the files inside the export for which the user has permissions granted by the `--bypassTraversalCheck` option.

However, in this case, operations like Rename file and Delete file are not granted by default. This is normal CIFS behavior. Modify ACLs as required to enable these operations.

6.5.4 POSIX bits

The Portable Operating System Interface (POSIX) bits of a file are another authorization method, which is different from ACLs. POSIX bits can also be used to specify access permissions for a file. UNIX file systems enable you to specify the owner and the group of a file. You can use the POSIX bits of a file to configure access control for an owner, a group and for all users to read, update, or execute the file. POSIX bits are less flexible than ACLs.

Tip: Changing the POSIX bits of a GPFS file system triggers a modification of the file system’s GPFS NFSv4 ACL, including the deletion of some ACEs. Because the SONAS system uses GPFS NFSv4 ACLs for access control, SONAS administrators should avoid changing the POSIX bits of files that are stored on the SONAS system, unless these specific GPFS NFSv4 ACL modifications and ACE deletions are clearly intended.

NFSv3 clients can set and read the traditional UNIX permissions. NFSv3 clients setting UNIX permissions reduce the ACL to match the UNIX permissions. In most NFS-only cases, the POSIX permissions are used directly. For NFSv3 clients, file sharing with CIFS access protection is done by using NFSv4 ACLs in GPFS, but NFSv3 clients only see the mapping of ACLs to traditional UNIX access permissions. The full NFSv4 ACLs are enforced on the server, and not necessarily the client.

6.5.5 Displaying and changing a file system directory's owner and group

Use the **`lsowner`** CLI command to display the owner and group of a directory in a file system. (see Example 6-3). An authorized SONAS administrative user can use the **`chowner`** CLI command to change the owner, the group, or both the owner and group, of an empty directory in a file system.

Example 6-3 Sample output of checking the owner of a file when the fully qualified path is used

```
$lsowner /ibm/gpfs0/testdir/file418
Owner Uid Group Gid
root  0   root  0
EFSSG1000I The command completed successfully.
```

The **`chowner`** CLI command changes only the owner and group that is assigned to the specified directory. It does not change the authorization of the owner or group. The authorization is derived from the system umask, or from inherited ACLs from higher-level directories. The **`chowner`** CLI command can only be used for a file system directory path that does not contain user files or subdirectories. You must specify the directory path enclosed in quotes, and you can optionally specify an owner, a group, or both using the syntax `owner:group`, as described in the following scenarios:

- To change only the group and not the owner, you must precede the group name with a colon, using the syntax `:group`.
- To change only the owner and not the group, use the syntax `owner`, without a colon.
- To change the owner and also change the group to the new owner's primary group, suffix the owner with a colon, using the syntax `owner:`.

The following paragraphs provide some specific examples:

- To change only the owner of directory `/nas/gpfs0/myexport` to `mydomain\\testuser`, and change the specified directory's group to `mydomain\\testgroup`, enter the following command:

  ```
  $ chowner 'mydomain\\testuser:mydomain\\testgroup' /nas/gpfs0/myexport/
  ```

- To change only the owner of directory `/nas/gpfs0/myexport` to `mydomain\\testuser` without changing the specified directory's group, enter the following command:

  ```
  $ chowner 'mydomain\\testuser' /nas/gpfs0/myexport/
  ```

- To change the owner of directory `/nas/gpfs0/myexport` to `mydomain\\testuser`, and also change the specified directory's group to be the primary group of the new owner, enter the following command:

  ```
  $ chowner 'mydomain\\testuser:' /nas/gpfs0/myexport/
  ```

- To change the group of directory `/nas/gpfs0/myexport` to `mydomain\\testgroup`, without changing the specified directory's owner, enter the following command:

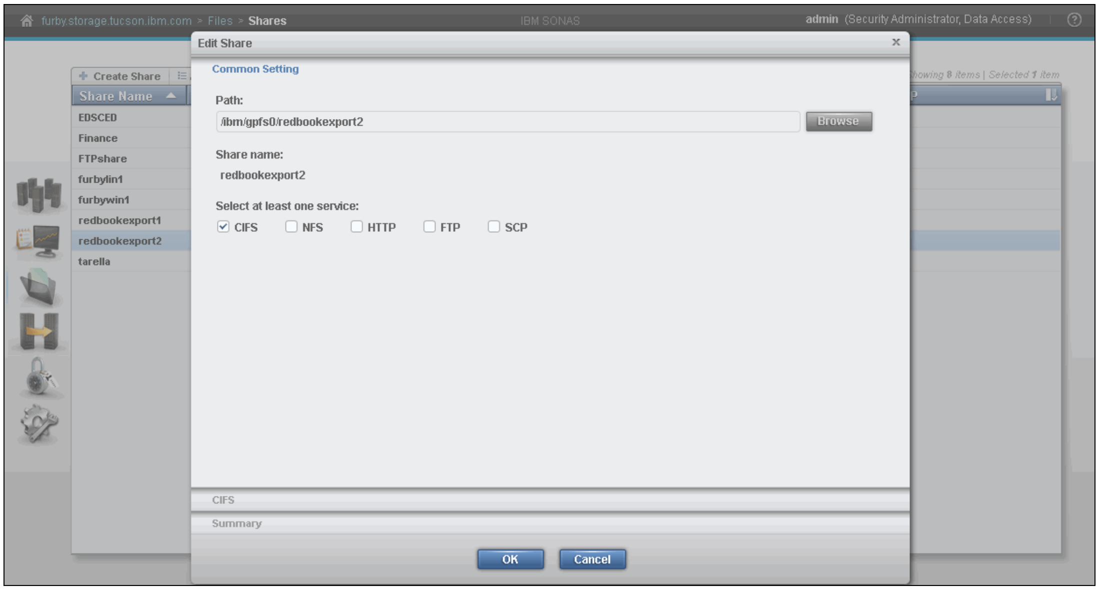
  ```
  $ chowner ':mydomain\\testgroup' /nas/gpfs0/myexport/
  ```

6.6 Changing shares and exports

You can add or remove a service to a share or export by using the GUI or by using the **chexport** CLI command.

6.6.1 GUI navigation

To work with this function in the management GUI, log on to the GUI and select **Files** → **Shares** (see Figure 6-2).



Path: /ibm/gpfs0/redbookexport2

Share name: redbookexport2

Select at least one service:
☑ CIFS ☐ NFS ☐ HTTP ☐ FTP ☐ SCP

Figure 6-2 Image of edit panel on GUI managed Share protocols

6.6.2 CLI usage

The **chexport** CLI command can be used to add a service to a share or export.

To modify an export named `testexport` by adding the FTP service and removing the NFS service, enter the following command:

```
# chexport testexport --ftp --nfsoff
```

6.7 Working with Common Internet File System (CIFS) shares

CIFS clients are widely used with Windows clients. The term CIFS is generally used interchangeably with the older name of Server Message Block (SMB).

6.7.1 Creating a CIFS share

The following steps outline an example scenario for creating a new CIFS share. When you use the CLI to create a share or export, specifying the owner is optional.

Details can also be found in *IBM SONAS Implementation Guide*, SG24-7962, or by using the Help button in the GUI and the SONAS information in the IBM Knowledge Center (*Creating shares or exports*):

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_t_export_create_GUI.html

To create a share, complete the following steps:

1. Create a new independent file set within an existing file system. For more detailed information, see the Creating a file set section in the SONAS section of the IBM Knowledge Center:

   http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_t_filesys_fileset_create.html

2. Link the newly created file set to a directory (junction path) on the underlying file system or file set. For more detailed information, see Linking a file set in the SONAS section of the IBM Knowledge Center:

   http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_t_filesys_fileset_link.html

3. Change the owner and group of the file set junction path. For more detailed information, see Displaying and changing a file system directory's owner and group in the SONAS section of the IBM Knowledge Center.

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

4. Create a quota definition for the newly created file set. For more information, see Managing quotas in the SONAS section of the IBM Knowledge Center:

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

5. Create a CIFS share or export by using the newly created file set. For more information, see Creating shares and exports in the SONAS section of the IBM Knowledge Center:

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

6. Modify ACLs and define inheritance. For more detailed information, see the following topics in the SONAS section of the IBM Knowledge Center:

   - Managing authorization and access control lists:

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

   - Authorization limitations:

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

   > **Tip:** Depending on the default ACLs and the usage of inheritance, the directory might become inaccessible. This might be the case when the ACLs are not inherited from the owning directory, and no owner has been specified. In this case, the directory is owned by the default owner `root`, and therefore inaccessible to users.

7. Modify the newly created share or export. For more detailed information, see the following SONAS topics in the IBM Knowledge Center:
   - Changing shares and exports:

     http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_t_export_chang.html?lang=en
   - Adding a service to a share or export:

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

As previously mentioned, the SONAS Knowledge and *IBM SONAS Implementation Guide*, SG24-7962 provide information to help with these tasks, whether you choose to use the GUI or the CLI. See these resources for detailed information about each specific requirement.

SONAS 1.5.1 includes support for the Server Message Block (SMB) 2.1 protocol. SMB 2.1 was introduced in Microsoft Windows Server 2008 R2 and Windows 7, and is supported by the Samba server in SONAS.

SMB 2.1 support brings important performance and usability enhancements including large maximum transmission unit (MTU) support (increased from 64 kilobytes to 1 megabyte), support for an interim response, and asynchronous I/O (read, write) support. An interim response from the server enables the server to tell the client that more information will follow, do not time out.

If a file is recalled from tape, the client waits longer for the SONAS to deliver the actual file. SMB 2.1 support delivers more efficient use of 10 gigabit Ethernet (GbE) network bandwidth, improving the performance of various tasks, such as copying large files.

SONAS also now supports SMB 2 signing. SMB signing adds a cryptographic checksum to digitally sign a packet. This enables the communication between a client and server to be validated, and prevents data tampering or man-in-the-middle attacks.

For more information about CIFS, see the CIFS limitations section in the SONAS area of the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/adm_smb_limitations.html

6.7.2 CIFS data integrity options

When you submit the **mkexport** or **chexport** SONAS CLI commands and use the `--cifs` option, you can optionally specify parameter values in a comma-separated, key-value pair list. Some of these CIFS options regulate SONAS CIFS server behavior that is related to data integrity.

oplocks=yes/no

The default value is yes. You can disable opportunistic locks (oplocks) to avoid the loss of data if a CIFS connection breaks or times out.

An oplock is an SMB/CIFS mechanism that promises a client that a block region of the requested file will not be written or modified by the server without correct notification until the oplock is cleared by the client. It does not apply to NFS, but only SMB/CIFS shared files and Windows-based clients. Oplocks are enabled on all Windows clients and servers by default.

Oplocks are a caching mechanism. If the oplocks option is set to yes, the client can request an oplock from an SMB server when it opens a file. If the server grants the request, then the client can cache large chunks of the file and not notify the server what it is doing with those cached chunks until it is finished. That reduces network I/O significantly, and is a big boost to performance.

Although oplocks give applications a performance boost, they can also contribute to data loss when a CIFS connection breaks or times out. Some examples of events that can cause connection interruptions are an interface node failover and GPFS or storage hang.

When an oplock is granted for a file, when Windows Redirector has stored the data into its cache, it acknowledges to the application that the data was written. From the application's perspective, the data was written successfully, and the application proceeds. Windows later destages the data to the SONAS system in the background. If any of the preceding events occurs, and the network connection is interrupted while there is still data in Windows Redirector to write to the SONAS system, the Redirector cache is erased.

When oplocks are in use and an out-of-space error occurs, the same data loss can happen. The write has already been acknowledged by the application, but the Windows Redirector cannot write the data to the SONAS system. In these cases, the Windows operating system usually logs a `Delayed Write Failed` event. Monitor the Windows event log for these events.

When files are in high contention, oplocks begin to lose their value quickly as oplock revocation communication creates resource demands. In many cases where CIFS performance becomes a bottleneck due to high contention, disabling oplocks on CIFS shares can provide significant performance improvement for the client and server stacks.

syncio=yes/no

The syncio parameter is disabled (set to a value of no) by default. A value of yes specifies that files in a share for which the setting is enabled are opened with the `O_SYNC` flag. The file is opened for synchronous I/O, and writes on the resulting file descriptor block the calling process until the data has been physically written to the underlying hardware.

This minimizes data loss in the case of a node failover because otherwise, a value of no might result in data being acknowledged as written to the client while the interface node has the data only in its cache and has not yet written that data to disk. That data is lost if a node failover occurs.

However, a value of yes can result in a significant performance decrease for most applications.

The preferred practice is to avoid operating data-critical workloads over CIFS, unless oplocks are disabled and synchronous I/O is enabled.

Workloads, such as database applications that update only small blocks in a large file, might have improved performance when syncio is enabled, because, in that case, GPFS does not read a complete block if there is only a small update to it.

leases=yes/no

This option is enabled (set to a value of yes) by default. A value of yes specifies that clients that are accessing the file over other protocols (such as NFS) can break the opportunistic lock of a CIFS client, so the CIFS client is informed when another client is now accessing the same file at the same time.

Disabling this feature provides a slight performance increase each time that a file is opened, but it increases the risk of data corruption when files are accessed over multiple protocols without correct synchronization. This option (leases=`yes`) is less important if oplocks are disabled).

locking=yes/no

This option is enabled (set to a value of `yes`) by default. A value of `yes` specifies that, before granting a byte range lock to a CIFS client, a determination is made whether a byte range file control lock is already present on the requested portion of the file. Clients that access the same file by using another protocol, such as NFS, are able to determine whether a CIFS client has set a lock on that byte range of that file.

sharemodes=yes/no

The CIFS protocol enables an application to permit simultaneous access to files by defining share modes, which can be any combination of `SHARE_READ`, `SHARE_WRITE`, and `SHARE_DELETE`.

If no share mode is specified, all simultaneous access attempts by another application or client to open a file in a manner that conflicts with the existing open mode are denied, even if the user has the appropriate permissions granted by share and file system access control lists.

The sharemodes option is enabled (set to a value of `yes`) by default. A value of `yes` specifies that the share modes that are specified by CIFS clients are respected by other protocols.

A value of `no` specifies that the share modes apply only to access by CIFS clients, and clients using all other protocols are granted or denied access to a file without regard to any share mode that is defined by a CIFS client.

synconclose=yes/no

This option is enabled (set to a value of `yes`) by default. A value of `yes` specifies that the file system synchronizes data to disk each time that a file is closed after a write to ensure that the written data is flushed to disk.

> **Important**: Disabling this option (setting it to a value of `no`) increases the risk of possible data loss if there is a node failure.
>
> **Tip:** This option applies only to file system data, not metadata.

coherency={yes|no|nodirs|norootdir}

In the CIFS specification, lock enforcement is mandatory if a client application has requested a lock. By default (coherency=`yes`), which is the preferred practice for data integrity, the SONAS system enforces this specification by using system-wide locking, which can decrease performance for some special workloads. A warning message displays when a SONAS system administrative user sets the `--cifs` option coherency to a value other than `yes`, and prompts for confirmation.

The `--cifs` coherency option values `yes` (default), `no`, `nodirs`, and `norootdir` are mutually exclusive. You might consider using a non-default coherency value if performance is a consideration, and if data integrity and consistency is ensured at the application level. These are ensured because the application enforces a strict access control model, and requires that files and directories are not modified by multiple processes at the same time.

Also, it requires that reading a file's contents does not occur while another process is writing to that file. The application must coordinate all file accesses to avoid conflicts, because conflicts are no longer managed at the CIFS protocol level.

If coherency is set to `no`, the SONAS system does not enforce any CIFS protocol system-wide consistent locking, which therefore must be ensured by the application to avoid data corruption and data loss. Avoid this configuration.

If coherency is set to `nodirs`, the SONAS system does provide consistent file locking, but not consistent directory locking.

If coherency is set to `norootdir`, the SONAS system disables the synchronization of directory locks for the root directory of the specified share, keeping the lock coherency for all the files and directories within the share root.

Attention: This option should only be used if data integrity is ensured on the application level (rather than the protocol level). The applications must ensure that files/directories are not modified by multiple processes at the same time, and that reading of file content does not happen while another process is still writing the file.

Alternatively, if the application is coordinating all file accesses to avoid conflicts. Without locking, the consistency of files is no longer guaranteed on protocol level. If data integrity is not ensured on application level this can lead to data corruption.

Tip: When the `--cifs` option coherency has a value other than `yes`, the **`lsexport`** SONAS CLI command displays the coherency value of a share or export.

6.7.3 Connecting by using a CIFS client

You can connect to a CIFS share by using a CIFS client from many different client operating systems.

Attention: The CIFS protocol does not include a built-in failover capability to transparently reestablish a lost CIFS connection. If a CIFS connection to the SONAS system is lost due to a failover or a network event, the CIFS client must establish a new CIFS connection and reopen all of the file handles that were previously open when the connection was lost. This is a limitation of the CIFS protocol. The SONAS system synchronizes file data to disk only when it receives a close request for a file.

If an SONAS interface node failover occurs during a write request, the most recent changes might not be written to disk. Therefore, only file changes that were both written and closed are always saved, and CIFS clients might experience corrupted files for any files that were open for write at the time of failover.

Also note that used space and free space that is reported to CIFS clients depends on the quotas that are applicable to the current user.

Scenarios for connecting to SONAS with a Windows client

The following examples are not meant to be in-depth. The examples use the following definitions:

System name `SONAS03`
Share name `gpfs0all`

Several methods are available for connecting to the SONAS system with a Windows client:

- Using the Universal Naming Convention (UNC) syntax
- Mapping a network drive using Windows Explorer
- Mapping a network drive using **`NET.EXE`** from the Windows command line

Example 1: UNC mapping

Using Windows Explorer, enter the path to the network share in the address bar using UNC syntax:

```
\\sonas03\gpfs0all
```

> **Note:** The preferred practice is to use a fully qualified domain name when connecting to shares. If your organization has a flat domain name server (DNS) namespace, this is less important, but in a large organization with an internal DNS hierarchy, using the full name means the server will be found even if the client moves around the organization.

If you are not currently authenticated to the domain on which the share resides, you are prompted for your credentials, as shown in Figure 6-3.

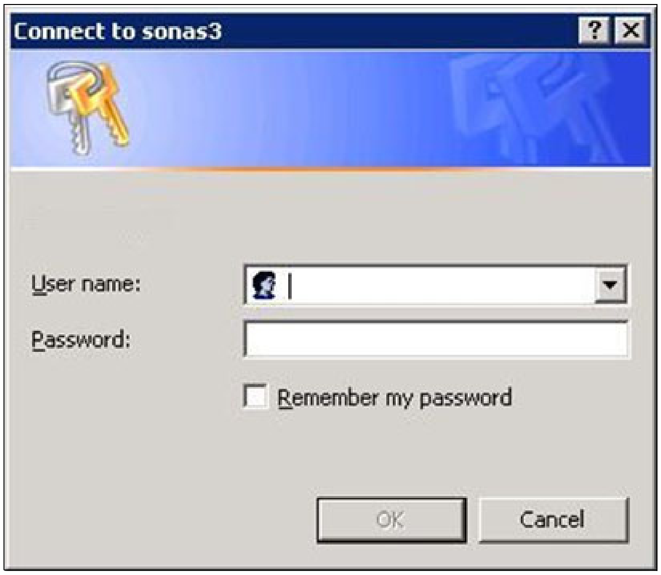
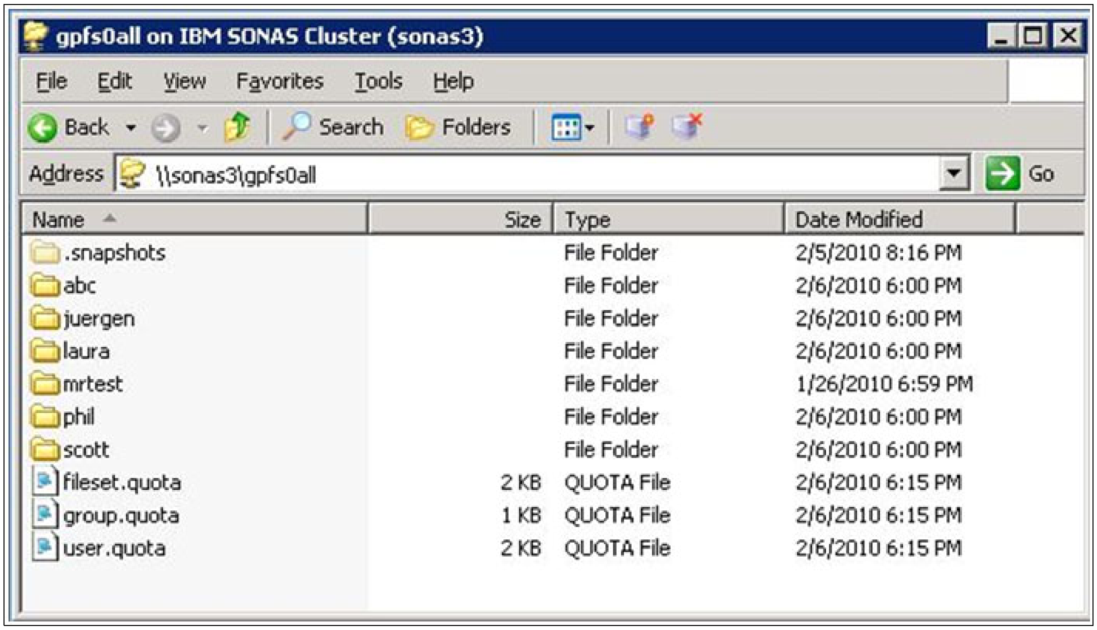


Figure 6-3 Windows Authentication Prompt

After authentication, a new window opens showing the contents of your share, as shown in Figure 6-4.



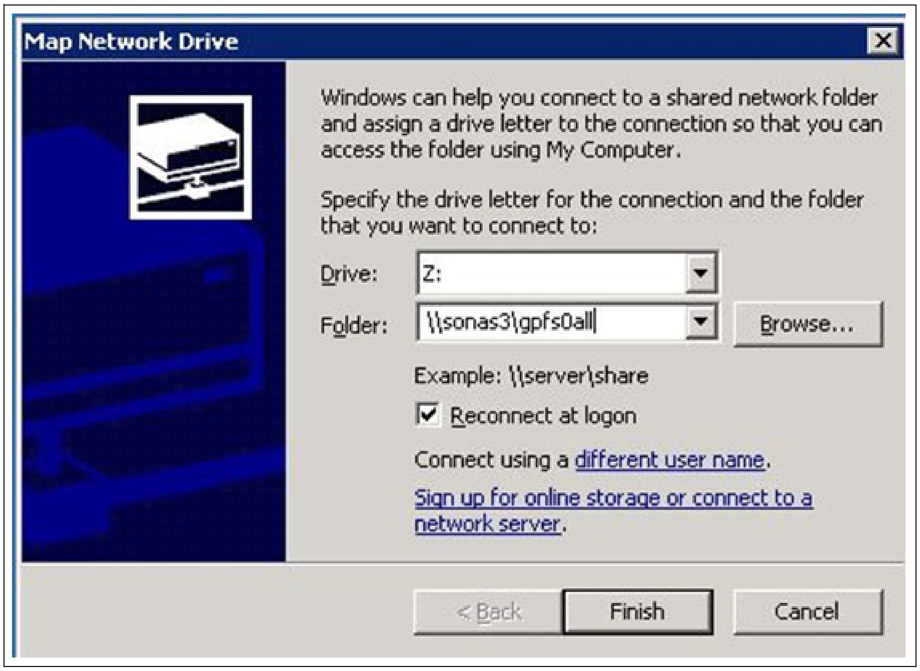
Address: \\sonas3\gpfs0all

Name	Size	Type	Date Modified
.snapshots		File Folder	2/5/2010 8:16 PM
abc		File Folder	2/6/2010 6:00 PM
juergen		File Folder	2/6/2010 6:00 PM
laura		File Folder	2/6/2010 6:00 PM
mrtest		File Folder	1/26/2010 6:59 PM
phil		File Folder	2/6/2010 6:00 PM
scott		File Folder	2/6/2010 6:00 PM
fileset.quota	2 KB	QUOTA File	2/6/2010 6:15 PM
group.quota	1 KB	QUOTA File	2/6/2010 6:15 PM
user.quota	2 KB	QUOTA File	2/6/2010 6:15 PM

Figure 6-4 Explorer view using UNC paths to connect to the SONAS system

Example 2: Mapping a network drive by using Windows Explorer

To map a network drive to a drive letter from Windows Explorer, click **Tools** → **Map Network Drive** or use the icon. Figure 6-5 shows the Explorer pane using drive letters to connect to the SONAS system.



Windows can help you connect to a shared network folder and assign a drive letter to the connection so that you can access the folder using My Computer.

Specify the drive letter for the connection and the folder that you want to connect to:

Drive: Z:

Folder: \\sonas3\gpfs0all

Example: \\server\share

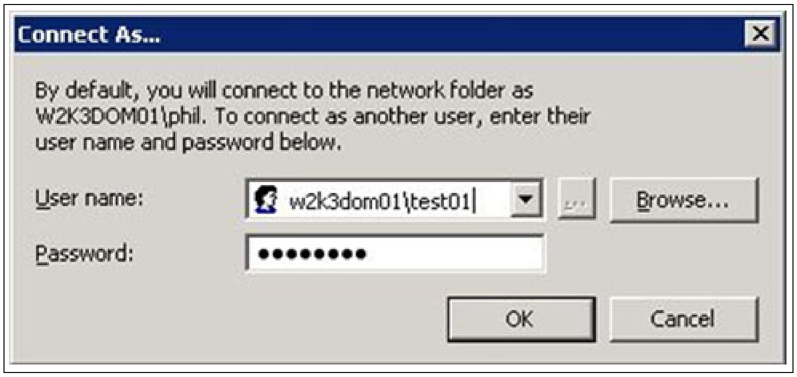
☑ Reconnect at logon

Connect using a different user name.

Sign up for online storage or connect to a network server.

Figure 6-5 Windows Authentication and Drive Letter Attachment to Map a Network Drive

Optionally, you can specify a user name and password to access the share using the connection with a different user name option. Figure 6-6 shows the window prompting for authentication information.



Connect As...

By default, you will connect to the network folder as W2K3DOM01\phil. To connect as another user, enter their user name and password below.

User name: w2k3dom01\test01

Password: ●●●●●●●●

Figure 6-6 Different user auth panel pop-up

The domain separator on Windows clients is a backslash (\). That means you must enter `w2k3dom01\test1` for the user `test01` from the domain `w2k3dom01`.

Example 3: Mapping a network drive by using NET.EXE

The following example maps the share `gpfs0all` from `sonas03` to the drive letter `X:` and prompts for the password without echoing the password to the screen:

```
C:> net use x: \\SONAS03\gpfs0all/user:DOMAIN\\username
```

The command waits for your password input. If your authentication information is successful, you are notified that the command completed successfully.

Connecting with CIFS from the Linux operating system

To attach to a CIFS share or export by using a Linux operating system, use a CIFS client. To access the share or export, a user must have appropriate permissions for accessing the path, and ACL authorization to at least allow read (r) and execute (x) each directory in the full path of the directory on which the share or export is mounted.

Tip: Connecting by using CIFS from IBM AIX is not supported.

You can use the CIFS client in the Linux kernel to mount a share, similar in concept to mounting an NFS export.

Tip: Linux clients that are running the CIFS protocol with a version of CIFS before version 1.69 might experience issues that are related to I/O transactions. Use CIFS version 1.69 or higher installed on the client to improve reliability.

You can also use interactive CIFS clients, such as **`smbclient`** and **`smbget`**, from the Samba Software Suite that is available for any major Linux distribution. Some desktop environments have their own CIFS clients, such as Gnome, where you can use the menu item **Places** → **Connect to server**, and select **CIFS**, after which the mounted share displays in the desktop environment’s file browser.

Example 1: Attaching to a CIFS share or export by using a Linux host

The following example describes how to connect to a Microsoft Windows network by using the CIFS protocol from the Linux operating system. For more information, administrators of other UNIX operating systems should consult the documentation for their CIFS client. It is assumed that you have root access to the host.

Example 6-4 uses these definitions:

System name `SONAS03`
Share or export name `gpfs0all`

If you do not already know the name of the share or export that you want to access, you can use the **`smbclient`** command to obtain a list of available resources.

Example 6-4 Sample smbclient command to list resources

```
(root@linuxhost)~ # smbclient -L sonas03 -U DOMAIN\\username
Password: <password not displayed>
Domain=[DOMAIN] OS=[UNIX] Server=[SONAS Cluster]
Exportname       Type      Comment
   ---------       ----      -------
   IPC$            IPC       IPC Service ("SONAS Cluster")
   phil            Disk      CIFS share or export of SONAS3 cluster for user Phil
   gpfs0all        Disk      CIFS share or export of SONAS3 to share or export FS
gpfs0
Domain=[DOMAIN] OS=[UNIX] Server=[SONAS Cluster]
Server              Comment
   ---------           -------
Workgroup           Master
   ---------           -------
```

You can access the share or export with an FTP-like utility by using the following command from an interactive system shell, as shown in Example 6-5.

Example 6-5 Sample access by using smbclient

```
(user@linuxhost)~ # smbclient //SONAS03/gpfs0all -U DOMAIN\\username
Password: <password not displayed>
Domain=[DOMAIN] OS=[UNIX] Server=[SONAS Cluster]
smb: \>
```

Furthermore, you can mount the CIFS share or export to a local directory, `/tmp/mount` in this example, by submitting the command in Example 6-6 as `root` from a system shell.

Example 6-6 Mount a CIFS share

```
(root@linuxhost)~ # mount -t cifs -o user=DOMAIN\\username //server/export
/tmp/mount
```

Verify that the mount occurred by running the **mount** command without any arguments, as shown in Example 6-7. The output will display the file systems that are mounted on the Linux system.

Example 6-7 Sample output from the mount command

```
(root@linuxhost)~ # mount
/dev/hda1 on / type ext3 (rw)
proc on /proc type proc (rw)
sysfs on /sys type sysfs (rw)
debugfs on /sys/kernel/debug type debugfs (rw)
udev on /dev type tmpfs (rw)
devpts on /dev/pts type devpts (rw,mode=0620,gid=5)
/dev/hda2 on /boot type ext3 (rw)
proc on /var/lib/ntp/proc type proc (rw)
//sonas03/gpfs0all on /tmp/mount type cifs (rw,mand)
```

Conversely, to unmount the share or export, run the following command at the system shell:

```
# umount /tmp/mount
```

6.7.4 Using substitution variables for CIFS shares

Although it is possible to use substitution variables for creating a CIFS share, in most cases you should *not* use this configuration for reasons of simplicity and consistency.

However, having a large number of Windows users all concurrently accessing the same CIFS share can lead to performance bottlenecks, because Windows clients automatically open the root folder of a share when connecting. In a home directory environment, it is helpful to use substitution variables when you create CIFS exports for home directories. For example, home directory exports can be created by using the `%U` substitution variable to represent the user name on the **mkexport** command:

```
mkexport home /ibm/gpfs0/.../%U --cifs
```

For more information about substitution variables, see the Using substitution variables topic in the SONAS area of the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/use_substitute_var.html

6.8 Working with Network File System (NFS) exports

When you submit the **mkexport** or **chexport** SONAS CLI commands and use the `--nfs` option, you can optionally specify NFS options and corresponding parameter values in a semicolon-separated list that is enclosed by double quotation characters. Each option is followed by its corresponding parameter values, if any exist, enclosed in parentheses (see Example 6-8).

Example 6-8 NFS options example

```
--nfs "master(rw);trusty(rw,no_root_squash);sync"
```

6.8.1 NFS options

Some of these NFS options regulate SONAS CIFS server behavior that is related to data integrity.

async/sync

Using the `async` option clears the default sync option. The `async` option enables the SONAS NFS server to respond to NFS write and commit requests without committing data to stable storage, such as permanent disk media, violating the NFS protocol specification.

When configured with the sync option, the NFS server behavior strictly conforms to the NFS protocol specification. Configuring the `async` option enables for the possibility that data that has been written by an NFS client, but not yet committed to stable storage by the NFS server, can be lost when a failure condition, such as the loss of an SONAS interface node, occurs. The result is data loss. This data loss can occur without notice to the application, even after successful returns from the fsync or close system calls.

Note that the `async` option *might not* significantly improve the performance of NFS write requests, because asynchronous optimization is now embedded in the protocol and a feature of the NFS protocol. When the async option is set, the `no_wdelay` option has *no effect*.

The `async` option, should only be used in environments where the potential for data loss that occurs without notification to the application can be tolerated, and where the performance benefits of setting the option have been verified.

wdelay/no_wdelay

NFS has an optimization algorithm *enabled by default,* by the NFS `wdelay` option that delays disk writes if NFS calculates that a sufficient probability exists that a related write request might arrive soon.

The write delay reduces the number of disk writes, and might increase performance; however, when the actual related write request pattern does not reach the predicted probability, performance might decrease because this behavior causes delay in every request.

The mutually exclusive NFS option `no_wdelay` disables the `wdelay` behavior.

In general, the `no_wdelay` NFS option is considered preferred practice when most of the NFS requests are small and unrelated. The `no_wdelay` NFS option is enabled by the SONAS system by default, and directs NFS to write to disk as soon as possible.

6.8.2 NFSv4 support in SONAS 1.5.1

SONAS 1.5.1 supports two types of NFS:

- The kernel-based NFSv3 server that is implemented in the operating system that is the Red Hat Enterprise Linux (RHEL) implementation
- The user space-based NFSv4 server that is the IBM implementation, which also supports the NFSv3 protocol

By default, the SONAS cluster is configured to use the kernel-based NFSv3 server. You can migrate to the IBM NFSv4 stack to use the NFSv4 protocol using the **`chnfsserver`** command. You can also switch back to the kernel-based Red Hat NFSv3 stack.

For complete details about migration to and fro between IBM NFSv4 and RHEL NFSv3 stack, see the Managing the NFS stack topic in the SONAS section of the IBM Knowledge Center:

`http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_nfs_stack_welcome.html`

The following NFS options are supported on both RHEL NFSv3 and IBM NFSv4:

- rw
- no_root_squash
- all_squash
- anonuid=<uid> and anongid=<gid>
- insecure
- sec=<seclist>

The following NFS options are supported only on RHEL NFSv3:

- async
- no_wdelay
- nohide
- crossmnt
- no_subtree_check
- insecure_locks, no_auth_nlm
- no_acl
- mountpoint=<path>

The following NFS options are supported only on IBM NFSv4:

- transportprotocol
- protocol

For a detailed description of these NFS options, see the man page for the **`mkexport`** command.

Preferred practices for using NFS

The following list describes some preferred practices for using NFS:

- If you want to have the exact NFSv4 protocol semantics, switch all the NFSv3 clients to NFSv4 because some features of NFSv3 and NFSv4 protocols are incompatible. For example, the NFSv3 protocol does not support Share Reservations or open operations that the NFSv4 protocol supports.
- Before migrating from kernel-based RHEL NFSv3 stack to the IBM NFSv4 stack, save the export configuration by using the **`lsexport -v`** command. You might need this information if you migrate back to the kernel-based RHEL NFSv3 stack.
- Frequent migration between the kernel-based RHEL NFSv3 stack and IBM NFSv4 stack is not recommended.
- If you have nested exports, such as /a and /a/b/c on the same file system for the same NFS client, the NFS protocol cannot determine that the client is accessing the same file through different exports and data corruption might occur. Each export is assigned a new file system ID even if the exports are from the same file system. Therefore, do not create nested exports on the same file system for the same NFS client.
- The NFSv4 protocol does not support nested exports. If you want to use the NFSv4 protocol after you migrate to the IBM NFSv4.0 stack, ensure that the export with the common path, called as the top-level export, has all the permissions.

- The NFSv4 protocol uses string names for user and group identification, but the NFSv3 protocol uses UIDs and GIDs. After you migrate to the IBM NFSv4.0 stack, if all clients are to use the NFSv4 protocol or a mix of NFSv3 and NFSv4 protocols, install a common ID mapping server so that the clients and the storage server obtain the same names or UIDs and GIDs.
- A NFS client must mount an NFS export by using the IP address of the Storwize V7000 Unified system. If a host name is used, ensure that the name is unique and remains unique.
- The default NFS protocol that is used on client systems might differ. For example, the AIX 6.3 client system uses NFSv3, where the RHEL 6.2 client system uses NFSv4 by default. If you are using a client that uses NFSv3 by default, you must explicitly specify NFSv4 in the mount command.
- For more information about important things to be considered for different types of NFS clients, see the NFS client considerations topic in the SONAS section of the IBM Knowledge Center:

  http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonasWelcome.htm

Limitations of NFS

For detailed information about NFS limitations, see the NFS limitations section in the SONAS section of the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonasWelcome.htm

6.8.3 Netgroups

Netgroups are used in UNIX environments to control access for NFS exports, remote logins, and remote shells. Netgroups cannot be used to control access for any NAS protocols other than NFS. Each netgroup has a unique name and defines a set of hosts, users, groups, and other netgroups. Netgroups were introduced as part of NIS, and can also be stored in LDAP, in SFU as Server for NIS, or in local files on the client.

The SONAS system versions 1.1.1 and higher support netgroups by using an NIS database, and lookup of netgroups in NIS can be configured in either of the first two configurations that are listed in the table in Table 6-1. SONAS system versions 1.3.1 and higher also support netgroups by using the NIS and LDAP database configurations that are listed in rows 3 and 4 in Table 6-1.

Table 6-1 SONAS netgroup support

Netgroup database	Authentication	ID mapping	SONAS release	CLI commands to configure
NIS	None	None	1.1.1	`cfgnis`
NIS	Active Directory	Auto	1.1.1	`cfgad, cfgnis`
NIS	Active Directory	SFU	1.3.1	`cfgad, cfgsfu, cfgnis`
LDAP	LDAP	LDAP	1.3.1	`cfgldap`

Note: Local authentication scheme available in SONAS 1.5.1 does not support netgroups.

Netgroup host definition format

The SONAS system supports netgroups only for grouping hosts to restrict access to NFS file systems that are exported by the SONAS system. The /etc/netgroup file of a server defines network-wide groups.

Netgroups are used in UNIX environments to control access for NFS exports, remote logins, and remote shells. Netgroups cannot be used to control access for any NAS protocols other than NFS. Each netgroup has a unique name and defines a set of hosts, users, groups, and other netgroups. Netgroups were introduced as part of NIS, and can also be stored in LDAP, in SFU as Server for NIS, or in local files on the client.

You can define a netgroup host in one of the following formats:

- By name. For example, myhost.
- By fully qualified domain name. For example, myhost.mydom.com.

Because NFS inherently works with host names for netgroups, avoid using Internet Protocol (IP) addresses in netgroup definitions.

> **Tip:** The host name in the netgroup definition must have both forward and reverse DNS lookup configured, so that the SONAS system can resolve both the host name and the host IP address, with which the mount service is requested on the SONAS system. Otherwise, a mount request fails with an access denied error.

Displaying a netgroup configuration

There is no SONAS CLI command to verify how the SONAS system resolves a netgroup reference.

You can, however, submit the **lsauth** SONAS CLI command to display the SONAS system authentication method configuration. Use the methods that are appropriate to the authentication method that you have configured to verify the configuration of external authentication servers.

6.8.4 NFS and ID mapping

The NFS authorization is based on UIDs and GIDs. The NFS client sends the UID of the current user to the NFS server. In this case, it is the SONAS system. This UID is only correct if the client has the same ID mapping as the server.

For example:

- Client has user johndoe with UID 2323
- Server has user janedoe with UID 2323

When the client (johndoe) accesses the server, the server thinks that janedoe is connecting and grants access to janedoe's files.

One solution is to manually define users on the client with the correct IDs. Use the **chkauth** CLI command to assist with this task (see Example 6-9).

Example 6-9 Sample chkauth command and output

```
# chkauth  -c sonas3.strg001st001 -u w2k3dom01\\laura -i
Gid = 10000017, Uid = 10000000, Home Directory = /var/opt/IBM/sonas/scproot,
Template Shell = /usr/bin/rsshAuthorization
```

The SONAS system uses the group IDs (GIDs) that are supplied by the NFS client to grant or deny access to file system objects, as defined in RFC 5531. When a user is defined in more than 16 groups, to get the wanted access control, you must define the groups that are transmitted from the client, and define mode bits or ACLs on the SONAS system.

Changes to NFS and ID mapping in 1.5.1

The NFSv4 clients cannot get UIDs and GIDs from NIS because the NFSv4 protocol uses string names (user@host.domain) for user and group identification and not UIDs and GIDs.The client and server must then map these strings to traditional POSIX user and group IDs because GPFS is a POSIX-based file system.

NFSv4 mandates support for a strong security model, where client/server interactions are done using the GSS-API framework. The required security type is Kerberos. The quality of protection, such as which crypto techniques are used, and service (authentication only, integrity, or privacy) are negotiated between the client and server.

As with the previous protocol versions, NFSv4 defers to the authentication provisions of the supporting RPC protocol. Because the RPC mechanisms are the same for versions 3 & 4, the NFS implementation supports Kerberos for both Version 3 & 4.

The NFSv4 clients and the Storwize V7000 Unified system must be configured with the same authentication and ID mapping server. The Storwize V7000 Unified system does not support NFSv4 clients to be configured with different authentication and ID mapping servers. The NFSv4 implementation still supports traditional UNIX style authentication and security though it is not recommended.

6.8.5 Connecting with an NFS client

This section describes how to connect to SONAS with an NFS topic:

- “Connecting from the UNIX operating systems”
- “Connecting to IBM NFSv4/NFSv3 (Kerberized) Server from UNIX clients” on page 199
- “Connecting with Apple Mac OS X clients” on page 202
- “Connecting by using NFS from a Windows client” on page 202

Connecting from the UNIX operating systems

As a prerequisite, NFS client packages must be installed, and the administrator give the user the right to mount remote file systems. Otherwise, the administrator must create the mount.

> **Tip:** NFS clients are responsible for maintaining data integrity when a server restarts, crashes, or fails over. In the NFS protocol, the NFS client is responsible for tracking which data is destaged, for detecting that a server has crashed before destaging all data, and for tracking which data must be rewritten to disk.

Failover is not visible to most applications in NFS, with the following exceptions:

- Client applications might experience `-EEXIST` or `-ENOENT` errors when they create or delete file system objects.
- *NFS clients should always use IP addresses when they are mounting NFS exports* because reliable file locking and lock recovery requires that the NFS client reconnect all of the NFS locked and `statd` services to the original server node after a client restart or crash. In most cases, the SONAS IP addresses are floating IPs controlled by clustered trivial database (CTDB) that move between SONAS interfaces. If a node crashes or restarts, there is no default guarantee that the same IP will be put back on the same server after restart.

The following example shows connecting to a SONAS using NFS.

To check which exports are available, run the following command on the command line:

```
showmount -e SONAS <public name or IP address>
```

The output, showing the export list, is shown in Example 6-10.

Example 6-10 Sample showmount export list

```
$ showmount -e furby.tuc.stglabs.ibm.com
Export list for furby.tuc.stglabs.ibm.com:
/ibm/gpfs1               *
/ibm/gpfs0/redbookexport1 *
/ibm/gpfs0/furbywin1      *
/ibm/gpfs0/furbylin1      *
```

Tip: By default, an AIX client (much like a MAC OS client) uses non-reserved ports. However, the SONAS system rejects requests from non-reserved ports. To enable the use of reserved ports on an AIX client, submit the AIX command **`nfso -p -o nfs_use_reserved_ports=1`**.

The AIX NFS client also has a known issue in datagram retransmission behavior. There are two actions to address this issue, both of which must be used.

The NFS client is responsible for detecting whether it is necessary to retransmit data, and for retransmitting all uncommitted cached data to the NFS server if retransmission is required. SONAS system failover relies on this expected client behavior.

Use the specific NFS mount options `hard,intr,timeo=1000,dio,noac` on the AIX client to extend the retransmit time to 100 seconds on failure. To mount the remote NFS export to a local mount point, submit the following command:

```
mount -t nfs -o hard,intr,timeo=1000 <server>:/<path>
```

The following things are important to note:

- If you use NFS file locking, use the server IP address rather than the host name when you create the mount. This step is required to ensure that file locking is reliable and to prevent orphaned locks.
- For Linux users: If you want file changes to be immediately visible to applications on NFS clients, it might be necessary to add the `lookupcache=none` option when you mount the export. This setting can adversely affect performance.

Check the NFS man page on your client system for recommendations on mount parameters, and more information about soft and hard mounts. Use the **`-o hard,intr`** parameter when you create a mount. On AIX and older Linux systems, the **`intr`** parameter enables NFS requests to be interrupted if the server goes down or cannot be reached.

Set the default timeout to `1000` (which equals 100 seconds).

For example, issue the following command:

```
mount -t nfs -o hard,intr,timeo=1000 sonas3:/ibm/gpfs0 /mnt
```

Tip: The **`mount`** command is an administrative command unless the administrator gives access privileges to users that include the ability to mount devices. The NFS mount point might already be configured by the administrator.

Connecting to IBM NFSv4/NFSv3 (Kerberized) Server from UNIX clients

A detailed description of setting up Kerberos and UNIX clients to support NFSv4 exports is beyond the scope of this book. The process typically includes the following steps:

1. Enable IBM NFSv4 stack using the **`chnfsserver`** CLI command.
2. Ensure all the domains are set up correctly by using the **`setnwdns`** CLI command.
3. Ensure that the fully qualified domain name (FQDN) of SONAS and its corresponding public IPs are registered in the DNS.
4. Create a proxy AD NFS user account, set the NFS service principal attribute for the SONAS FQDN, and register the proxy AD NFS user account with it.
5. Create a keytab file for SONAS on the AD server by using this principal and map it to the proxy AD user.
6. Using the keytab file, configure AD with Kerberos support on SONAS.
7. Adapt the general NFS server settings to the Kerberos domain (which would also be the NFSv4 domain) and enable NFSv4 protocol by using the **`chservice`** CLI command and verify the settings by using the **`lsservice`** CLI command, as shown in Example 6-11.

Example 6-11 General NFS server settings

```
$chservice nfs --options
'squash=root_squash,domain=sonas,realm=SONAS.com,NFS4_service=enable'
Do you really want to perform the operation (yes/no - default no):yes
EFSSG1000I The command completed successfully.

lsservice --protocoloptions
CIFS
=====
serverDescription : "IBM NAS"
diskFreeQuota : yes
NFS
=====
Lease_Lifetime : 90
NFS4_service : enable
domain : sonas
realm : SONAS.com
anongid : -2
anonuid : -2
access : ro
squash : root_squash
sec : sys
secure : true
transportprotocol : tcp
protocol : 3;4
EFSSG1000I The command completed successfully.
```

8. Create a share and view the share by using the **lsexport** command, as shown in Example 6-12.

Example 6-12 View the share using the lsexport command

```
$ mkexport nfsv4test '/ibm/gpfs0/nfsv4test' --nfs
'*(rw,root_squash,sec=krb5,protocol=3;4)' --owner 'SONAS\autouser1:SONAS\domain
users'
EFSSG0019I The export nfsv4test has been successfully created.
EFSSG1000I The command completed successfully.

$lsexport -v | grep nfsv4test
nfsv4test /ibm/gpfs0/nfsv4test NFS      true   10/23/14 8:24 PM
*(root_squash,rw,anonuid=-2,anongid=-2,secure,protocol=3;4,transportprotocol=tc
p,sec=krb5)
```

9. Configure the RHEL NFSv4 client:

   a. Ensure that the Linux client details are configured in the DNS with IP, FQDN, and that the forward and reverse name IP lookups work.

   b. On the AD server, create a client computer account that will be used for the Kerberized NFS service.

   c. On the AD server, create a proxy AD NFS user account that will the NFS service principal for the client computer account.

   d. On the AD server map, the NFS service principal for this client computer account to the proxy AD NFS user.

   e. Create a keytab file for the Linux client with the previous principal, and map to the dummy AD user.

   f. Copy the created keytab file to the Linux client under `/etc/krb5.keytab`.

   g. Configure the Linux client to authenticate with the AD server and ensure that Kerberos is configured properly. AD users should be able to log in to this Linux client. Also ensure that `/etc/idmapd.conf` is configured with the correct domain and realm.

   h. Ensure that the Linux client can get the granting ticket from AD with **kinit** by using a valid principal.

   i. Configure the client for Kerberized services (mount, I/O) which require the `gssd` daemon and the secure NFS setting, as shown in Example 6-13.

   Example 6-13 Configuring the client for Kerberized services

```
#cat /etc/sysconfig/nfs | grep SECURE_NFS
SECURE_NFS="yes"

#service rpcgssd restart
Stopping RPC gssd:                                          [FAILED]
Starting RPC gssd:                                          [  OK  ]
[root@client001 nfsv4test]# service rpcgssd status
rpc.gssd (pid 2272) is running...
[root@client001 nfsv4test]#
```

10. Mount the Kerberized export with the `sec=krb5` option and verify the configuration, as shown in Example 6-14.

Example 6-14 Mounting an NFSv4 share

```
#showmount -e newinstsonas.sonas.com
Export list for newinstsonas.sonas.com:
/ibm/gpfs0/nfsv4test *
/ibm/gpfs0/abc       *

#mount -vv -t nfs4 -o tcp,soft,intr,sec=krb5
newinstsonas.sonas.com:/ibm/gpfs0/nfsv4test /mnt/nfsv4test
mount.nfs4: timeout set for Sat Oct 25 16:07:32 2014
mount.nfs4: trying text-based options
'tcp,soft,intr,sec=krb5,addr=10.0.100.141,clientaddr=10.0.100.240'
newinstsonas.sonas.com:/ibm/gpfs0/nfsv4test on /mnt/nfsv4test type nfs4
(rw,tcp,soft,intr,sec=krb5)

# cat /proc/mounts | grep nfsv4test
newinstsonas.sonas.com:/ibm/gpfs0/nfsv4test/ /mnt/nfsv4test nfs4
rw,relatime,vers=4,rsize=1048576,wsize=1048576,namlen=255,soft,proto=tcp,port=0
,timeo=600,retrans=2,sec=krb5,clientaddr=10.0.100.240,minorversion=0,local_lock
=none,addr=10.0.100.141 0 0
```

11. Log in as an AD user on the Linux client. Run the **`kinit`** command to get the Kerberos credentials, access the mount, and do file operations. See Example 6-15.

Example 6-15 Accessing NFSv4 mount

```
# su - sonas\\autouser1
bash-4.1$ id
uid=16777216(SONAS\autouser1) gid=16777218(SONAS\autouser)
groups=16777218(SONAS\autouser),16777217(BUILTIN\users),16777219(SONAS\domain
users) context=unconfined_u:unconfined_r:unconfined_t:s0-s0:c0.c1023

bash-4.1$ id
uid=16777216(SONAS\autouser1) gid=16777218(SONAS\autouser)
groups=16777218(SONAS\autouser),16777217(BUILTIN\users),16777219(SONAS\domain
users) context=unconfined_u:unconfined_r:unconfined_t:s0-s0:c0.c1023

bash-4.1$ kinit
Password for autouser1@SONAS.COM:

bash-4.1$ klist
Ticket cache: FILE:/tmp/krb5cc_16777216
Default principal: autouser1@SONAS.COM

Valid starting     Expires            Service principal
10/25/14 16:14:15  10/26/14 02:14:18  krbtgt/SONAS.COM@SONAS.COM
        renew until 11/01/14 16:14:15

bash-4.1$ cd /mnt/nfsv4test/

bash-4.1$ echo "This is write to file from nfsv4 client" > testfile1.txt
-bash-4.1$ ls -l
total 0
-rwx------. 1 SONAS\autouser1 SONAS\autouser 40 Oct 25  2014 testfile1.txt
```

```
bash-4.1$ cat testfile1.txt
This is write to file from nfsv4 client

bash-4.1$ chown :sonas\\"domain users" testfile1.txt

bash-4.1$ ls -l
total 0
-rwx------. 1 SONAS\autouser1 SONAS\domain users 40 Oct 25  2014 testfile1.txt

bash-4.1$ chmod 750 testfile1.txt
-bash-4.1$ ls -l
total 0
-rwxr-x---. 1 SONAS\autouser1 SONAS\domain users 40 Oct 25 16:18 testfile1.txt
```

Connecting with Apple Mac OS X clients

Consider the following Mac OS X requirements when you configure and manage the SONAS or Storwize V7000 Unified shared file system:

- Mac OS X Version 10.6.8 does not support display of backup files in a `/.snapshots` subdirectory. To display backup files in a `/.snapshots` subdirectory with Mac OS X Version 10.7.4 or later, you can use **Finder** → **Search** to find the file, click **Get Info** on the file, and click **General**.
- By default, a Mac OS client uses non-reserved ports. The SONAS and Storwize V7000 Unified system reject requests from *non-reserved ports*. To enable the use of *reserved ports* on a MAC OS client, run the **`mount`** command with the **`-o resvport`** option.

  For example:

  ```
  # sudo mount -t nfs -o resvport 192.168.101.100:/ibm/gpfs0/abc /mnt
  ```

Connecting by using NFS from a Windows client

Although you can connect to an NFS export by using the NFS client that is included in some versions of the Microsoft Windows operating system, the preferred practice is to use CIFS to connect to SONAS.

6.9 Connecting with other protocols

Although most clients likely connect to CIFS or NFS, interactive access with a web browser is an easy interface for users to download files. FTP and SCP protocols are often used when writing automated scripts.

6.9.1 Using FTP

Users can connect to a SONAS environment by using the FTP protocol. As a prerequisite, an FTP client must be installed and working correctly. Microsoft Windows clients have an FTP client that is installed by default, but this FTP client does not support automatic resume or reconnect in the case of an error.

> **Important:** The FTP protocol does *not* use encryption. The user name, password, and the file contents are sent across the network in plain text. When possible, use SCP rather than FTP.

The SONAS environment provides failover capabilities for FTP. If a failure occurs, the FTP client must reconnect and resume the transfer. Availability of the resume and the automatic reconnection features depends on the capabilities of the FTP client. The FTP server of the SONAS environment does support resume and reconnecting FTP clients. The FTP server supports the Representational State Transfer (REST) command to assist with resuming abnormally stopped operations on the SONAS system.

When you are using FileZilla to view a directory listing on an SONAS system, all file time stamps have a constant time offset. The time offset is caused by FileZilla automatically converting the time stamps from Coordinated Universal Time to the local time zone. This conversion can be customized by adding the SONAS system to the site manager and adjusting the server time offset in the Advanced tab.

For the system named `sonas3.w2k3dom01.com`, go to the command-line interface and start the FTP client:

```
ftp sonas3.w2k3dom01.com
```

The system asks for a user name and password and, if successful, the panel message displays in the format shown in Example 6-16.

Example 6-16 Sample FTP

```
# ftp sonas3.w2k3dom01.com
Connected to sonas3.w2k3dom01.com.
220 -FTP Server (user 'testuser@w2k3dom01.com')
220
User (sonas3.w2k3dom01.com:(none)): sonasdm\testuser
331 -Password:
331
Password:
230-230 Login successful.
```

6.9.2 Using Secure Copy Protocol

Secure Copy Protocol (SCP) is a protocol for securely transferring files between a local and a remote host or between two remote hosts. The protocol has certain options that can be displayed on a Linux or UNIX system by using the **`man scp`** command. One option (`-o`) can be used to pass Secure Shell (SSH) options in the format that is used by the system-wide configuration file, `/etc/ssh/ssh_config`.

Search the client operating system documentation for information about the `ssh_config` options that are available by default to understand what options are available for your use in the SONAS environment. Describing those options is beyond the scope of this section.

To connect to an SONAS environment by using SCP, a prerequisite is that an SCP client is installed, available, and functioning correctly. Windows clients do not have an SCP client installed by default, so one must be installed before this protocol can be used.

To copy data from the local system to the SONAS system by using the SCP protocol, submit the following command:

```
scp local_path/files user:sharename
```

An example of the **scp** command and output is shown in Example 6-17.

Example 6-17 Sample scp command

```
# scp /tmp/*
"w2k3dom01\test1"@sonas01.w2k3dom01.com:/sonas01.w2k3dom01\test1@sonas01.w2k3dom01
.com's password:
```

Provide the domain and user name in the syntax `domain\user` to authenticate against the HTTPS server, which is `w2k3dom01\test1` in this example. Otherwise, the login fails.

When the password of the corresponding user is entered, the files are copied.

6.9.3 Using HTTPS

You can connect to a SONAS share or export by using the HTTP/HTTPS service with the following procedure.

The HTTPS protocol does not include an automated failover capability to transparently reestablish a lost connection. If an HTTPS connection to the SONAS system is lost due to a node IP failover event, the client must establish a new connection and restart the abnormally stopped operation. You can restart only a **GET** operation; resuming a **GET** operation with the `Range:` option in the HTTPS header is not supported.

Complete the following steps:

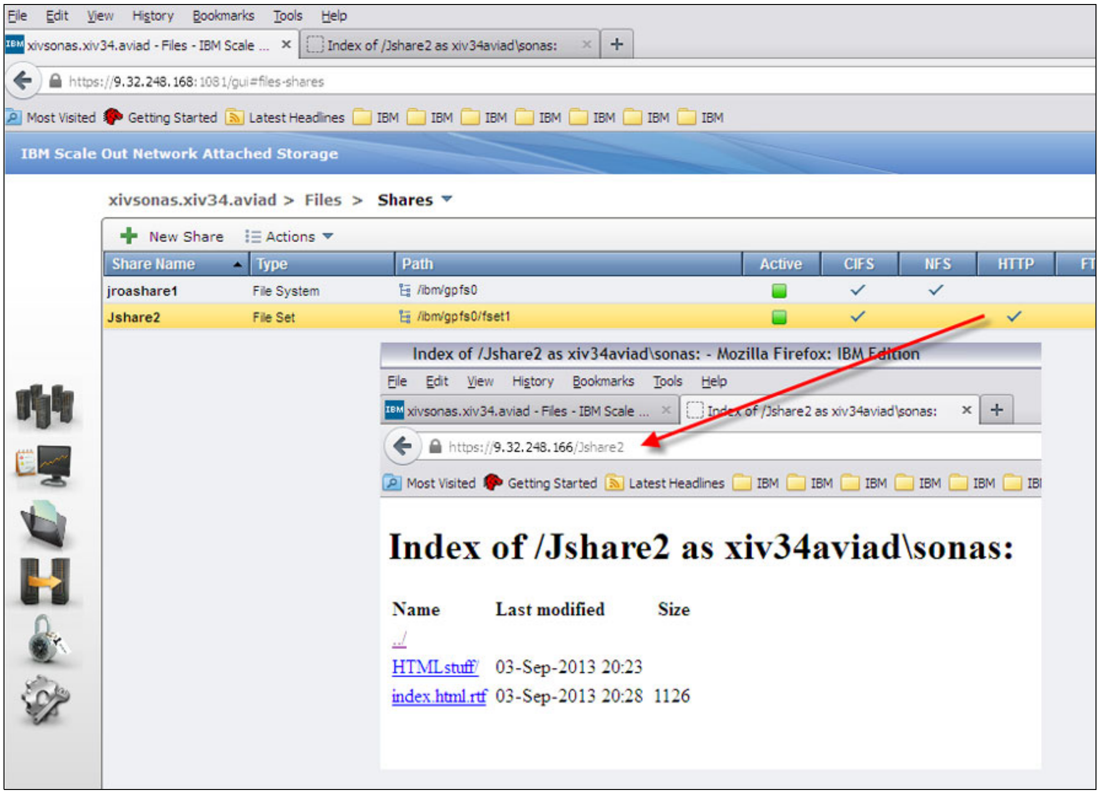
1. Open a browser, and enter the following address:

   `http://<sonas_cluster_name>/<share_or_export_name>`

   > **Note:** The share or export name is optional. If you do not provide the share name, a listing of all available shares will be presented.

   The system redirects all HTTP access requests to HTTPS.

For example, enter: `http://9.32.248.166/Jshare2/`, as shown in Figure 6-7.


File Edit View History Bookmarks Tools Help
xivsonas.xiv34.aviad - Files - IBM Scale ...
Index of /Jshare2 as xiv34aviad\sonas:
https://9.32.248.168:1081/gui#files-shares
Most Visited
Getting Started
Latest Headlines
IBM
IBM Scale Out Network Attached Storage
xivsonas.xiv34.aviad > Files > Shares
New Share
Actions
Index of /Jshare2 as xiv34aviad\sonas: - Mozilla Firefox: IBM Edition
https://9.32.248.166/Jshare2

Share Name	Type	Path	Active	CIFS	NFS	HTTP	FT
jroashare1	File System	/ibm/gpfs0	🟩	✓	✓		
Jshare2	File Set	/ibm/gpfs0/fset1	🟩	✓		✓	

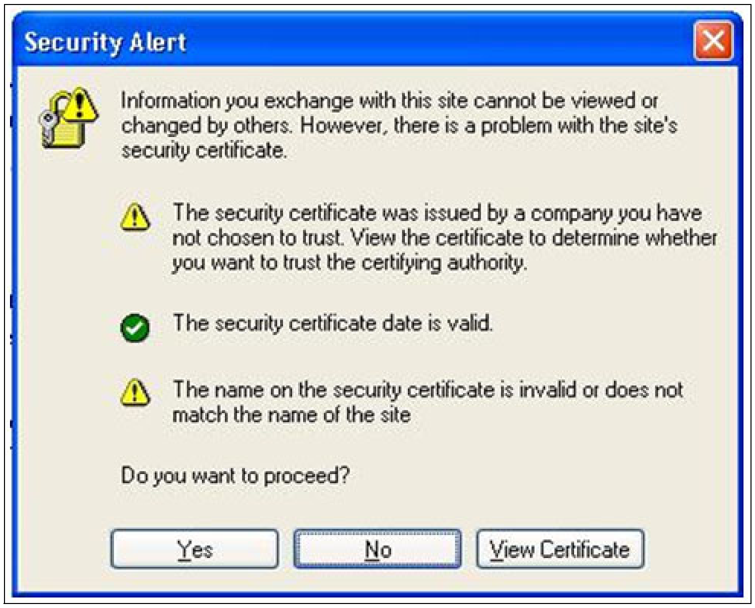
Index of /Jshare2 as xiv34aviad\sonas:

Name	Last modified	Size
../		
HTMLstuff/	03-Sep-2013 20:23	
index.html.rtf	03-Sep-2013 20:28	1126

Figure 6-7 Internet Explorer view: Connecting to export Jshare2

When only a self-signed certificate is used, the browser returns an error message that the certificate is not trusted. You can proceed to get access, or you can select to install the certificate to avoid this message in the future.

2. When you connect for the first time, you might see a security alert, as shown in Figure 6-8.



Information you exchange with this site cannot be viewed or changed by others. However, there is a problem with the site's security certificate.

⚠ The security certificate was issued by a company you have not chosen to trust. View the certificate to determine whether you want to trust the certifying authority.

✓ The security certificate date is valid.

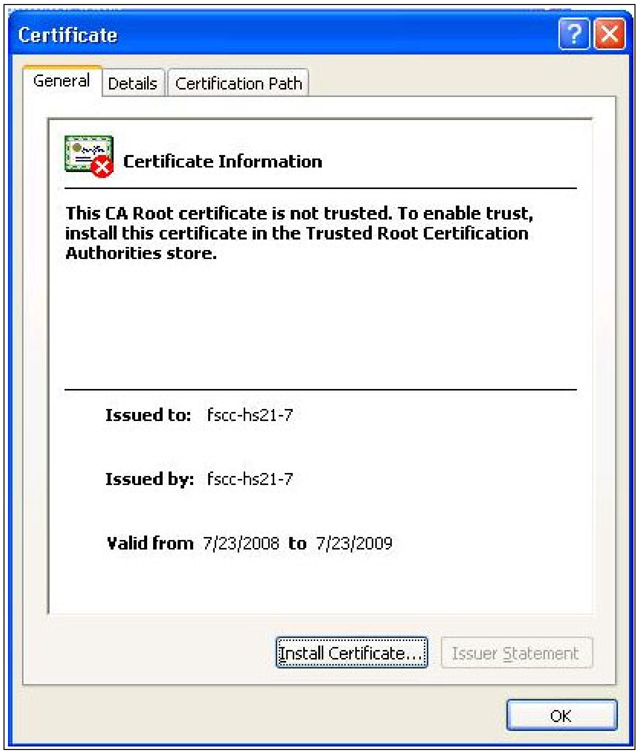
⚠ The name on the security certificate is invalid or does not match the name of the site

Do you want to proceed?

Figure 6-8 SONAS Security alert on a self-signed certificate

You can click **Yes** to proceed, but then you must repeat this procedure every time you want to connect using HTTPS to the SONAS software system.

3. Alternatively, you can click the **View Certificate** button and install the certificate, as shown in Figure 6-9.



Certificate Information

This CA Root certificate is not trusted. To enable trust, install this certificate in the Trusted Root Certification Authorities store.

Issued to: fscc-hs21-7

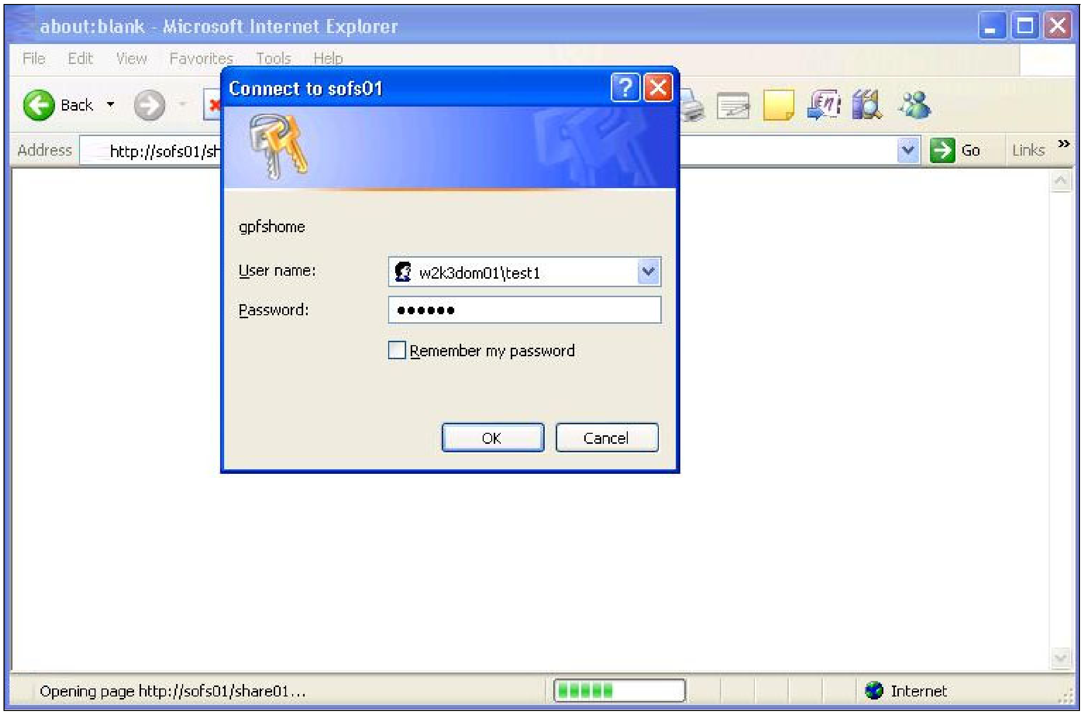
Issued by: fscc-hs21-7

Valid from 7/23/2008 to 7/23/2009

Figure 6-9 Self Installed Certificate

If you want to install this certificate on your client, proceed with installing this certificate and answer all of the questions in this wizard.

4. Next, you must authenticate (see Figure 6-10).


about:blank - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://sofs01/sh
Go
Links

Connect to sofs01

gpfshome

User name: w2k3dom01\test1

Password: ••••••

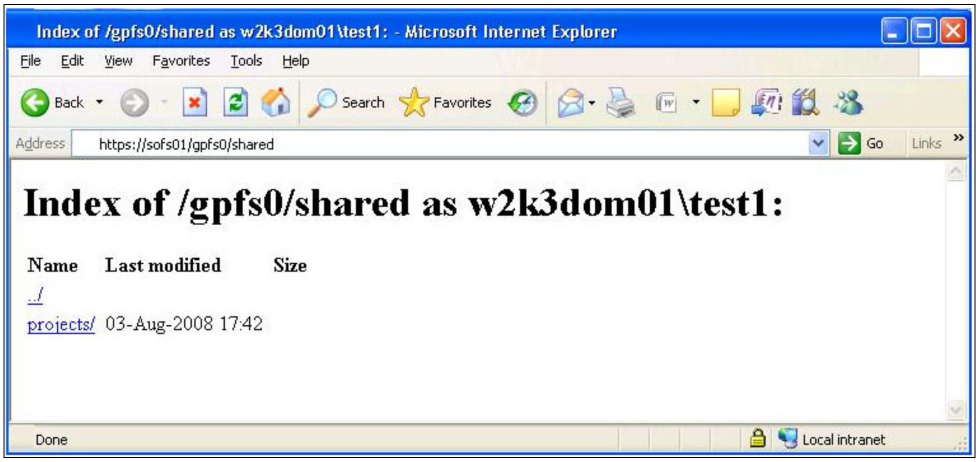
□ Remember my password

Opening page http://sofs01/share01…

Figure 6-10 HTTP connection authentication panel

5. Enter the domain and user name by using the syntax `domain\user` to authenticate against the HTTPS server, which is `w2k3dom01\test1`, as shown in Figure 6-11. Otherwise, the login fails.

   When the user is authenticated and has access to the system, the user can browse all files and folders to which the user has access.


Index of /gpfs0/shared as w2k3dom01\test1: - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address https://sofs01/gpfs0/shared
Go
Links

Index of /gpfs0/shared as w2k3dom01\test1:

Name	Last modified	Size
../		
projects/	03-Aug-2008 17:42	

Done

Figure 6-11 Explorer View of HTTP connection

7

Data protection

Data protection is a high priority for most enterprise-class data services. It is not only critical to business continuance, but critical to the protection of corporate intellectual property and general business value to its consumers. Clearly data protection is equal to business protection and IBM Scale Out Network Attached Storage (SONAS) is not exempt from that requirement.

Also, consider the fact that data protection gets more complex on a larger scale, so the bigger the data, the more expensive and complicated it is to protect efficiently. This is true with all data, large files, and small files. The more you have, the better you need to plan for its protection.

This chapter describes the following topics:

- Cluster security
- Asynchronous replication
- Backup and restore solutions
- Snapshots
- File clones
- Management node configuration backup tasks
- External log service
- Antivirus
- Failover, failback, and disaster recovery

A well-planned (preferred practice) replication requires that all services and clients are ready to go from the second site when disaster strikes the primary site, and that all the backup site or secondary site resources are all sized appropriately for managing failover workloads, bandwidth, and capacity requirements.

Services, such as domain name server (DNS) and domain controllers, must be robust and reliable enough to handle all failover traffic. Network and staffing tables should be ready to take over new demands.

Finally, the more protocols you offer for data access, the more complex it becomes to protect.

The IBM Redbooks publication *IBM SONAS Implementation Guide*, SG24-7962 discusses the various concepts of data protection and how to install, configure, and manage this data protection service in accordance with known preferred practices. This publication describes the preferred practices considerations of these requirements, how they should and should not be implemented, and topics that you might consider obtaining expertise in to serve a preferred practice approach.

Remember, as you read this chapter, that a preferred practice for SONAS service has some form of data protection scenario locally (snapshots and backup), plus remote data replication (such as asynchronous replication) for disaster recovery (DR).

7.1 Cluster security

Cluster security is a critical point of concern for most clients who are preparing to engage in a Scale out NAS solution. This chapter introduces common topics that are related to cluster security and how you can best position your security efforts for a SONAS or IBM Storwize V7000 Unified cluster solution. For information about the preferred practice network and network port security topics that are related to SONAS, see Chapter 3, “Networking” on page 71.

These are several things to consider for protection of SONAS or Storwize V7000 Unified. This chapter explains preferred practices for these protection considerations.

7.1.1 Protection of the cluster service

This section describes several aspects of protecting the cluster service:

- Hardware security:
  - The SONAS hardware can be a point of vulnerability if access to the SONAS cluster hardware is not controlled. Typically, this concern is answered by having SONAS in a controlled access facility (such as a secure data center) where conditioned, redundant power and cooling can be provided and monitored to the SONAS hardware, and power, temperature, and access is controlled.
  - Guests and service agents are monitored and tracked, and ideally physical access to service hardware is strictly controlled.
- Code Security:
  - Code must not be modified by the administrators unless instructed by IBM development through support escalation, and access to code level changes should only be administered by requirement if SONAS development or support agents mandate and supervise the changes. It can also be modified at the direction and guidance of a service engagement professional, where all changes are recorded in a problem management record (PMR) or Code Defect for explicit tracking.
  - Administrative accounts should only be given by and to qualified administrators and all levels of service administrators should be properly trained on SONAS use of command-line interface (CLI) or graphical user interface (GUI) and be officially aware of cluster management policies in accordance with their user account privilege level.

    Administrators should avoid using a general `Admin` account to ensure that changes are tracked by the user, and the general `Admin` account should be protected by a staff integrity managed process.

  - When root-level access is given to an account, it is important that access is controlled (this topic is further described later in the Privileged User Access Control bullet). For example, the root password can be changed with the **chrootpwd** command for a specific maintenance requirement. Then it can be changed back and the password can be protected by a *management-sealed process*. This process ensures that privileged anonymous access is only used for a specific or planned purpose.
  - Remote access should be controlled by firewall, controlling port, and Internet Protocol (IP) access to the SONAS management IP and Service IP addresses to the clusters.
- Node configuration security:
  - Node configuration, logs, and events are managed and protected by configuration files and service logs. A node can be reimaged from the Trivial File Transfer Protocol (TFTP) boot server image in the management node, but the vital product data and configuration files on the node are protected by a node-backup process.
  - The command **cndumps** (service dumps) should be run occasionally or regularly (such as monthly) to capture cluster configuration and log changes, and those **cndumps** should be stored off cluster to a central protected log repository.
  - It is important that the management node runs a **backupmanagmentnode** task from cron regularly, and stores the backup information on the secondary management node, and possibly a third node for safekeeping. This task is thoroughly explained in the *IBM SONAS Implementation Guide*, SG24-7962 IBM Redbooks publication. As a part of preferred practice, it is important to validate that the task is active and successfully completing (note that this task is automatically run in SONAS clusters today).
  - The **backupmanagementnode** command makes a backup from the local management node, where the command is running, and stores it on another host or server. *This command should only be run manually in configurations where there is a single dedicated management node (as with SONAS v1.1.1 and SONAS V1.2).* For Storwize V7000 Unified and for dual management node clusters, the backup (or sync) is automatic, and the command should never be run from the command line.

> **Note:** Run the following command to ensure that the tasks are scheduled and successful:
>
> ```
> # lstask -t CRON -s -v
> Name                      Status Last run        Runs on
> Type Scheduled
> BackupMgmtNode            OK     8/21/13 2:22 AM Management node
> CRON yes
> ```

- Privileged user access control:
  - Code security at the SONAS or Storwize V7000 Unified cluster is protected by password enabled access and limited service access accounts. Customers are typically not given access to privileged users, such as root, in a SONAS environment. In some cases, for special reasons, a customer might have root-level access for special maintenance reasons. In this case, the root level access must be protected by policy and practice.
  - Root-level access should not be used for day-to-day activity and should only be used when administrative-level user privilege is not sufficient to complete a task. In this case, the root password should be temporarily changed to support the required activity for the time frame of the required maintenance, and the usage of that access should be planned and collaborated to avoid mishap. A root user has enough privilege to harm the cluster, and much of what is started by the root user can be only loosely tracked from a history perspective.

7.1.2 Compliance with security policies

Many companies and government institutions require compliance with formal security policies. SONAS can configure the underlaying Linux operating system and hardware to meet many common security policy requirements.

The SONAS CLI, using the **chsettings** and the **chpasswordpolicy** commands, enables an administrator to make changes to the security policy in the following areas:

- Boot loader settings
- Secure Shell (SSH) server parameters
- Password requirements

In the next sections we'll describe these options in more detail.

Boot loader settings

Security policies can have requirements that reduce the exposure from physical access to the hardware. SONAS 1.5 and later provide a way to meet the following requirements:

- Configure password for the Grand Unified Bootloader (GRUB) using an MD5 or stronger cryptographic hash to protect the password
- Disable Boot From Removable Media

The **chsettings** command is used with the following parameters:

bootLoaderPassword *password* Sets the new password for the GRUB for all the nodes within the cluster. Optional.

disableBootLoaderPassword Disables password for the GRUB for all the nodes within the cluster. Optional.

bootFromRemovableMedia `'yes|no'` Specifies if default boot from removable media is enabled for all the nodes or not. Optional.

> **Note:** Setting the value of **bootFromRemovableMedia to** `no` disables the default boot from removable media, such as CD, DVD, USB, and so on for all of the SONAS nodes. However, the system can still boot from removable media by changing the boot option in the system's basic input/output system (BIOS) or system controller.
>
> To protect against unauthorized access to the BIOS or system controller and completely prevent booting from removable media, the SONAS administrator needs to set a BIOS supervisor/administrator password if one has not been set and disable a user-level password if one has been set.

SSH server and client settings

The following list describes SSH server hardening requirements and the respective configuration file values. These requirements have already been satisfied on all SONAS systems since the 1.3 release:

- The SSH daemon must not permit tunnels:

  ```
  PermitTunnel no
  ```

- The SSH daemon must do strict mode checking of home directory configuration files:

  ```
  StrictModes yes
  ```

- The SSH daemon must use privilege separation:

  ```
  UsePrivilegeSeparation yes
  ```

- The SSH daemon must not allow Rhosts Rivest-Shamir-Adleman algorithm (RSA) authentication:

  ```
  RhostsRSAAuthentication no
  ```

The new **chsettings** CLI command changes the SSH configuration on all of the systems in the SONAS cluster.

The chsettings security --sshHardening yes option

The following list describes various aspects of the `--sshHardening yes` option.

- The SSH daemon is configured to use only Federal Information Processing Standard (FIPS) 140-2 approved ciphers by adding the following line to the SSH server configuration files on all SONAS nodes:

  ```
  Ciphers aes128-ctr,aes192-ctr,aes256-ctr
  ```

- The SSH daemon is configured to not use cipher block chaining (CBC) ciphers.
- The SSH daemons are configured to only use message authentication codes (MACs) that employ FIPS 140-2 approved cryptography hash algorithms by adding the following line to all SSH server configuration files on all SONAS nodes:

  ```
  MACs hmac-sha1
  ```

- The SSH client is configured to only use FIPS 140-2 approved ciphers by adding the following line to SSH client configuration file on all SONAS nodes:

  ```
  Ciphers aes128-ctr,aes192-ctr,aes256-ctr
  ```

- The SSH client is configured to not use CBC-based ciphers.
- The SSH client is configured to only use message authentication codes (MACs) that employ FIPS 140-2 approved cryptographic hash algorithms by adding the following line to SSH client configuration file on all SONAS nodes:

  ```
  MACs hmac-sha1
  ```

- SSH daemon does not allow compression, or must only allow compression after successful authentication by adding the following line to all SSH server configuration files on all SONAS nodes:

  ```
  Compression delayed
  ```

- The SSH daemon uses a FIPS 140-2 validated cryptographic module (operating in FIPS mode). This means, use only cryptographic algorithms available in the validated Crypto++ module, which means the encryption algorithm list is limited to Advanced Encryption Standard (AES) and Triple-Data Encryption Standard (DES).
- The SSH client uses a FIPS 140-2 validated cryptographic module (operating in FIPS mode).

No additional hardening is done by default on the SONAS cluster. If you want to enable these settings, the following command can be used:

```
chsettings security --sshHardening yes
```

To reverse these settings, use the following command:

```
chsettings security --sshHardening no
```

Note: In SONAS 1.4 and later, async replication has the `--encryption [strong|fast]` option. The default strong authentication is compliant with the SSH hardening rules. However, async replication with the `fast` option does not work, because a weak cipher is used.

For more information, see the *Configuring asynchronous replication* topic in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/mng_t_arepl_config.html

Displaying the current security settings

The current security settings can be displayed with the **`lssettings`** command as shown in Example 7-1.

Example 7-1 Displaying the current security settings

```
$ lssettings security
SSH hardening  =  no
Boot from removable media  =  yes
Boot loader password  =  no
EFSSG1000I The command completed successfully.
```

Linux password settings

Security policies generally require certain characters be included in passwords to make the passwords stronger. The **`chpasswordpolicy`** command in SONAS 1.3 and later can set the following password requirements:

minLength Minimum length of the password. Defaults to 6 positions. Optional.

minAge Minimum age of the password in days. Defaults to zero indicating that password can be changed multiple times in a day. Minimum age should be less than the maximum age. Optional.

maxAge Maximum age of the password in days. Defaults to 90 days. Maximum age should be greater than the minimum age. Optional.

resetOnLogin Reset password on next login. Defaults to do not reset password on login if not specified. Optional. This applies to existing CLI users and not to new users.

dontResetOnLogin Do not reset password on next login. Defaults to do not reset password on login if not specified. Optional. This applies to existing CLI users and not to new users.

SONAS 1.4 added the ability to set the following password requirement:

remember The number of old passwords for each user to be remembered. Sets how frequently old passwords can be reused. The system remembers this many passwords, and the user cannot reuse any of those passwords. To disable this feature, set it to `0`. The initial default value is `3`. Optional.

The following options are added in SONAS 1.5.1:

minUpperChars Minimum number of uppercase characters that must be specified in the password. Optional.

minLowerChars Minimum number of lowercase characters that must be specified in the password. Optional.

minSpecialChars Minimum number of special case characters that must be specified in the password. Optional.

minDigits Minimum number of digits that must be specified in the password. Optional.

maxRepeat Maximum number of repeat characters that can be specified in the password. Optional.

minDiff Minimum number of characters that must be different in new password as compared to the old password. Optional.

rejectUserName User-name is not allowed as a part of password. Optional.

allowUserName User-name can be used as a part of password. Optional.

Note: These parameters are mixed-case, and must be entered as shown.

The **lspasswordpolicy** command is used to display the current password settings, as seen in Example 7-2.

Example 7-2 Output of the lspasswordpolicy command

```
$ lspasswordpolicy
Minimum special case characters  =  0
Minimum different characters with respect to old password  =  0
Minimum digits  =  0
Minimum lower case characters  =  0
Reset password on next login  =  false
Minimum password length  =  6
Minimum upper case characters  =  0
Maximum password age  =  90
Reject user-name  =  false
Remember old passwords  =  3
Minimum password age  =  0
Maximum number of repeat characters  =  0
EFSSG1000I The command completed successfully.
```

Session Policy settings

The session policy settings deal with bad password attempts, and how long an account will remain locked after too many bad password attempts. The **chsessionpolicy** command can be used with the following parameters:

maxLoginAttempts Specifies the maximum number of login attempts allowed before the user gets locked out. The default value is 0, indicating unlimited login attempts. A valid value would be 0 - 9. Optional.

timeout Specifies the time in hours for which the user gets locked out before the account is enabled. The default is 1 hour, indicating that the user gets locked out for an hour. A valid value would be 1 - 24. Optional.

loginPromptDelay Specifies the time in seconds between login prompts following a failed login attempt. The default is 4 seconds. A valid value would be 0 - 9. Optional.

The **lssessionpolicy** command is used to display the current settings as shown in Example 7-3.

Example 7-3 Using the lssessionpolicy command to display the current settings

```
$ lssessionpolicy
Locked user timeout  =  1
Maximum login failures allowed  =  0
Login prompt delay  =  4
EFSSG1000I The command completed successfully.
```

If you want to implement a password policy that requires a 10-letter password with an uppercase letter, a lowercase letter, a digit, and that does not allow the user name to be in the password, you can use the **chpasswordpolicy** command with the options shown in Example 7-4.

Example 7-4 Implementing a password policy

```
$ chpasswordpolicy --minLength 10 --minUpperChars 1 --minLowerChars 1 --minDigits
1 --rejectUserName
EFSSG0465I The password policy was set successfully on the cluster
4483305904013673286
EFSSG1000I The command completed successfully.
```

7.1.3 Protection of the data

This section describes protecting data:

- Share access control:
  - Data shares should be created with an explicit user owner definition. Access control lists (ACLs) must be immediately set by that owner for files within the file set and share before data is written to it. The ACLs exist to control the access to files and directories. Therefore, a full understanding of these ACLs, owners, and data access is necessary for the administrators of the NAS service.
- ACL management:
  - ACL management is described in the *IBM SONAS Implementation Guide*, SG24-7962. For preferred practice understanding of ACL management and data authentication see Chapter 2, “Authentication” on page 25. Access and authorization knowledge should be a requirement for all SONAS or Storwize V7000 Unified solution administrators before administrative level privileges are assigned to administrators.
  - You can and should create a new file system with a specific owner and group. Do not use the default root owner. When a file system or file set is created, if the owner is not specified, the default `owner/group/permissions` of the new path are `root/root/771`. You can change the ownership of a file system or file set by using the **chowner** CLI command if the file system or file set does not contain any directories or files, or during share or export creation, by setting the name of the owner of the share or export.
  - If you use the root squash option for NFS users, which is the default and used for security reasons, you should set the owner of the file system or file set immediately after the file system or file set is created because you are unable to set it using the Network File System (NFS) client.

Note: When a share or export contains files or directories, you cannot change the owner.

- Snapshots, clones, and replication:
  - Snapshots, file clones, wide area network (WAN) caching, and replication are ways to protect data locally and remotely by creating copies of the data within the file system space. Copies of a file should also be managed carefully. Customers can choose to enable privileges for their clients to be permitted to restore from their own snapshots. In this case, a client can restore any file that they have access to from previous snapshots. If two clients have access to any file, either one of them can restore from a previous snapshot.

    This activity is not tracked or audited in the cluster, so carefully consider this possibility for use in your environment. It is important to take time to carefully consider your snapshot management practice.
  - Too many snapshots are typically a waste of space, and improperly managing retention and deletion schedules or proper distribution of rules can tie up much-needed metadata performance when it is otherwise needed for routine tasks. For more information, see Chapter 4, “Storage configuration” on page 95.
- Client access security:
  - Client environment protection is also a serious consideration. If a user’s client system and user account is compromised, the attacker potentially has access to controlled data in the SONAS environment. Client systems should be maintained and up-to-date for security, and virus scanning. When a customer supports heritage host platforms without mandates for up-to-date internet security and password management practices, they put corporate data protection in jeopardy.
  - If a client system is infected by a virus, that virus can be spread to other clients through the data that is stored in the SONAS file system. For this reason, use an antivirus service for SONAS and Storwize V7000 Unified accessed data (whether that solution is tied to the client or the data service).

    When an antivirus service is installed for the SONAS, carefully define the scopes. Scan the data on ingest (preferred practice) rather than bulk scan. Bulk scan can place a burden on the cluster and degrade client response time during the scheduled event. For detailed information about SONAS antivirus see 7.8, “Antivirus” on page 253.
  - For large data repositories, random bulk file scanning can take large amounts of metadata and input/output (I/O) cycles away from the cluster. For data that is stored on tape devices such as Hierarchical Storage Manager (HSM) migrated files, data must be recalled to disk to support antivirus scan operations. For this reason, it is best to have a defined scope that only reads files on specified ingest.
  - It is a preferred practice to consult with SONAS or Storwize V7000 Unified solution experts before defining antivirus sizing and scopes for antivirus protection.

Tip: Antivirus scanning is only supported for Common Internet File System (CIFS) shares.

7.1.4 Protection of the access to data

All protocols have benefits and drawbacks. SONAS supports NFS, CIFS, Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), and Secure Copy Protocol (SCP) protocols to satisfy client requirements. However, because these protocols are not proprietary, limitations of the underlying protocol are not necessarily mitigated by the SONAS implementation. It is the preferred practice to be mindful of the protocols you use in your SONAS environment, and that you keep up to date with client operating system (OS) patching and client protocol tuning for optimal use of NAS protocol services.

Note: When using FTP, the user ID and password are passed on the network in plain text. Access controls in NFSv3 are basically provided by the client, and not the server. If your client is compromised, your data could be at risk.

SONAS and Storwize V7000 Unified V1.4.1 include the capability to turn off unwanted or unneeded protocols. So, if you are a pure CIFS shop, you can turn off NFS. If you are a pure NFS shop, you can turn off CIFS as a cluster service.

7.2 Asynchronous replication

SONAS asynchronous replication provides for command-driven consistent remote copies of sets of files from a source SONAS system (or file tree) to a target SONAS system.

The basis of disaster recovery capabilities for a SONAS solution is built around the asynchronous replication. Asynchronous replication enables for one or more file tree structures within a SONAS file name space to be defined for replication to another SONAS system.

The asynchronous replication process looks for changed files in the source file tree since the last replication cycle completed and uses `rsync` to efficiently move only the changed portions (blocks) of those files from the source location to the target location. In addition to the file contents, all extended attribute information about the changed files is also replicated to the target system.

Asynchronous replication is defined in a single direction, such that a file system at one site is considered the source of the data, and the file system at the other site is the target. The replica of the file tree at the remote location is to be used Read-Only until it is needed to become usable.

Figure 7-1 shows a basic asynchronous replication process.

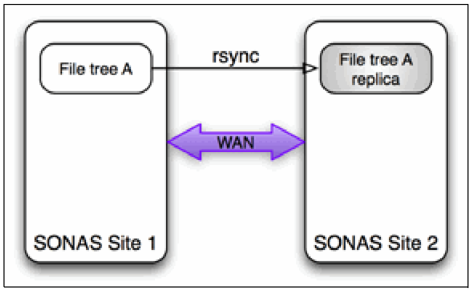


Figure 7-1 Basic uni-directional asynchronous replication

When using asynchronous replication, the SONAS system detects the modified files from the source system, and only moves the changed contents (changed blocks) from each modified file to the target system to create a replica. In most cases, by moving only the changed portions of each modified file, the network is used efficiently for replication.

Since SONAS V1.4.1, SONAS copies entire files when they are smaller than 1 megabyte (MB) and have changed blocks, or when the files are larger than 50 gigabyte (GB) and they have changed blocks. This process improves performance in replication because it reduces the complexity of tasks for reading the files to worklist the changed blocks in replication. In these cases, it is far more efficient to replicate the whole file instead.

The file-based movement enables the source and target file trees to be of differing sizes and configurations if the target file tree is large enough to hold the contents of the files from the source. For example, differing configurations support options like local synchronous copies of the file tree to be used at the source location but not at the target. This flexibility helps make the solution adapt to many different needs. Figure 7-2 shows a high-level asynchronous replication process flow.

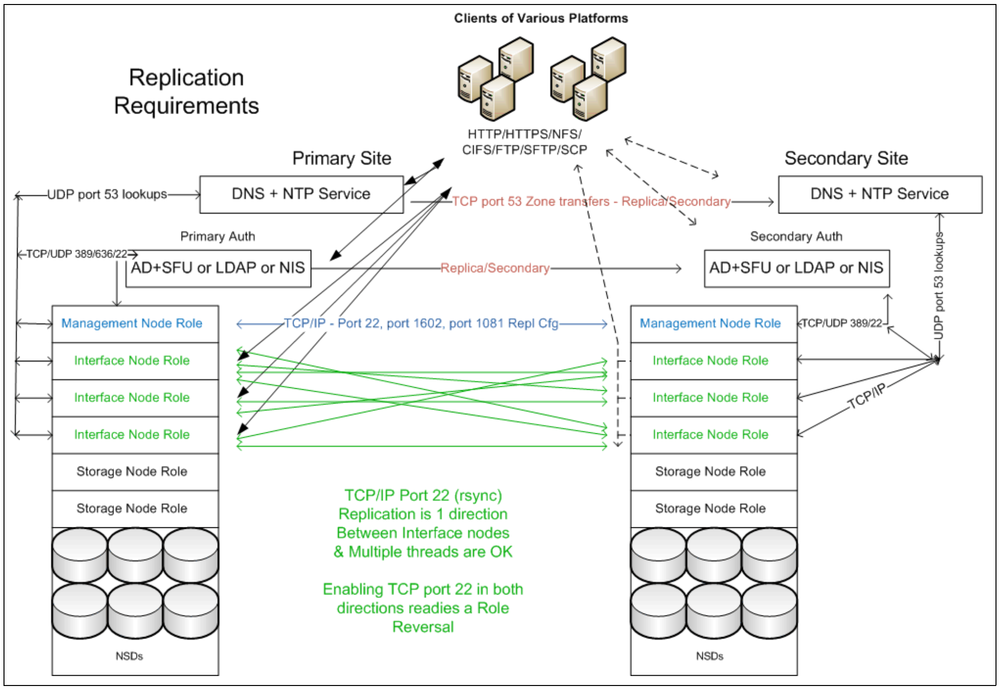


Figure 7-2 High-level image of the asynchronous replication service

The asynchronous replication function is intended to be run on a periodic basis to replicate the contents of a file system from a source SONAS system to a file system on a target SONAS. When it is run, the following major steps are performed during the replication:

- A snapshot of the source file system is created.
- The source file system snapshot is scanned for created, modified, and deleted files and directories since last asynchronous replication completed.
- A full read of each changed file for finding actual changed data is done.
- Changed data contents are replicated to the target system.
- A snapshot of the target file system is created.
- The source file system snapshot is removed.

Figure 7-3 shows SONAS replication process workflow.

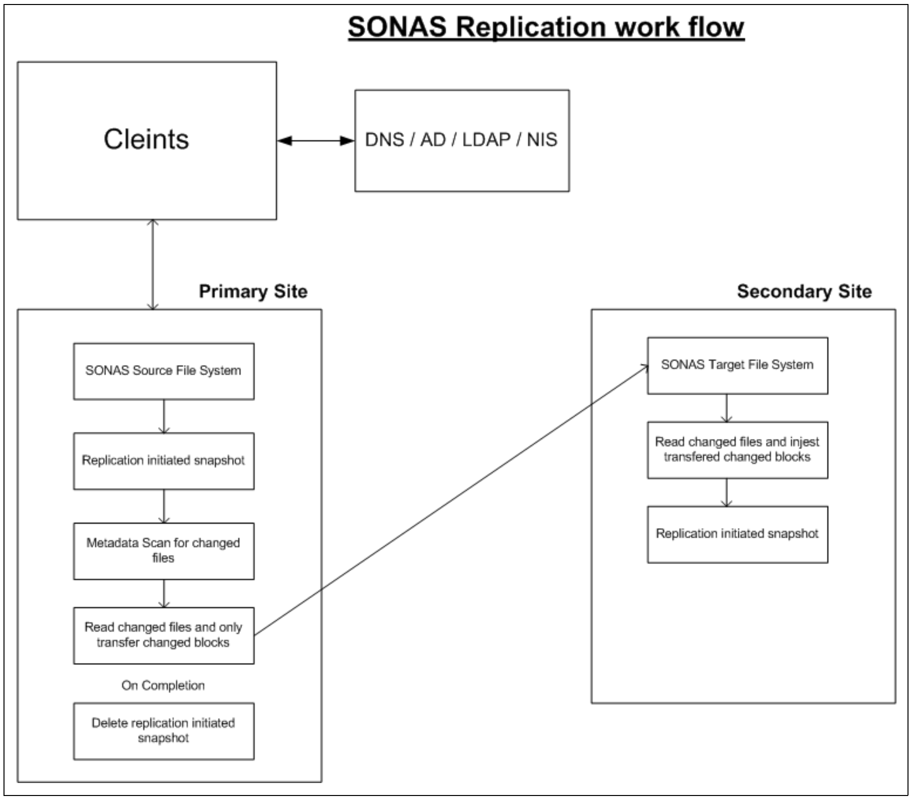

SONAS Replication work flow
Cleints
DNS / AD / LDAP / NIS
Primary Site
SONAS Source File System
Replication initiated snapshot
Metadata Scan for changed files
Read changed files and only transfer changed blocks
On Completion
Delete replication initiated snapshot
Secondary Site
SONAS Target File System
Read changed files and injest transfered changed blocks
Replication initiated snapshot

Figure 7-3 SONAS replication work flow diagram

The source and target snapshots *can* be configured to be omitted from the replication process. However, avoid this configuration. The source-side snapshot creates a point-in-time image of the source file system when the asynchronous replication process is started. The asynchronous replication process walks through this snapshot and looks for changes to build a changed file list.

This list is used as the basis for the replication to the target. The asynchronous process can use the file system for this function, but, if it does, it must scan and replicate the directory tree while it is actively being accessed and modified by the product applications.

Asynchronous replication is the only supported disaster recovery solution for SONAS or Storwize V7000 Unified. However, success depends on implementation and management practices, because it is a manual process that is initiated and managed by SONAS administrators.

It is critical that, for every file set created on the source, an equal file set must be created on the target and base file set permissions must be created on the target before data in that file set is replicated. Therefore, the preferred practice includes stopping replication to create a new file set on the source, then create the file set and set its base permissions on both source and target systems before putting data in the source. Then, restart replication on completion of the file set creation.

If data is pushed to the newly created source file set, and that data is replicated to the target file system before the target site file set is created and permissions are properly set on it, the data gets replicated to a directory in the target file system rather than a properly defined and linked target file set. This situation might not prove optimal in a disaster recovery solution.

7.2.1 Replication port requirements

The source management node and interface nodes must be able to communicate with the destination system’s management and interface nodes over the customer network.

The following Transmission Control Protocol (TCP) ports must be open between the source and destination nodes:

- TCP 22 (SSH)
- TCP 1081 (HTTP)

Open these ports in both directions.

7.2.2 Replication requirements

This section lists the requirements for SONAS replication:

- 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.
- The target file system must have enough free space to support replication of the source file system with the resources required 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) requirements during peak use.
- 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 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 management node and interface nodes to the target management node and interface nodes.
- For replication in both directions, or for potential failback after a recovery, open ports 1081 and 22 in both directions.
- Consistent authentication and ID mapping between all the sites is required:
  - If you are using a version before 1.5.1, SONAS must be configured for authentication and ID mapping with Active Directory Server (ADS) with and the Services for UNIX (SFU) extension, or with a Lightweight Directory Access Protocol (LDAP) server. Authentication with Active Directory (AD) without the SFU extension is not supported. Active Directory plus SFU is the preferred practice authentication for replication.
  - With version 1.5.1 and later, the auto ID mapping method is also supported.

> **Tip:** Adding more participating nodes to the replication, and more threads and processes per node, improves data replication performance for both scanning and data transfer, when the networks are not the bottleneck.

7.2.3 Multi-directional replication

Multi-directional replication is possible with different file systems, meaning file trees only. Because each file system can have only one target for replication, different file systems in the same cluster can replicate with different targets. Figure 7-4 shows multi-directional replication.

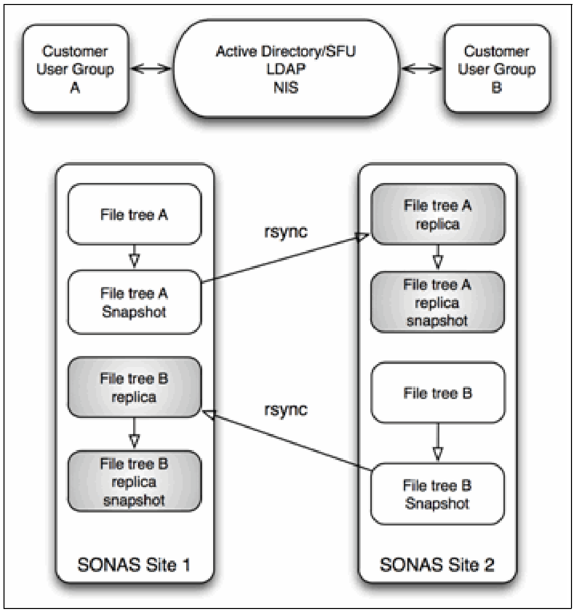


Figure 7-4 Example of multi-directional replication

7.2.4 Asynchronous replication preferred practice considerations

Remember the following considerations when you set up asynchronous replication:

- Time should be coordinated between the source and target systems. A recovery point objective (RPO) cannot be established unless replication can be consistently completed within that specified time frame.
- The first replication needs time to complete before an ongoing RPO can be established.
- Time and date stamps are carried forward from the source files or directories to the destination SONAS.
- If the destination system is used for business continuance, a time skew at the target can cause confusion. Therefore, coordinate both systems with Network Time Protocol (NTP) time services.
- The *Scan time* part of the replication time scales linearly with number of inodes in the source file tree (source file system).

Important: Snapshot capabilities use IBM General Parallel File System (IBM GPFS) scans. This GPFS capability provides the file system with a way to do high-speed scans of the GPFS inodes. As such, it is important to understand the different elements that affect GPFS scan performance.

The multiple interface nodes equally spread the policy engine rule evaluation, file scan identification, and subsequent data movement for increased performance. If greater scan speed is required, more SONAS nodes can be allocated to the scan, and each node scans only its equal portion of the total scan. The results of the parallel scan are aggregated, and returned as the usable list of candidate files.

File set information is not automatically carried forward to the destination SONAS. If a file set is created on the source file system, also configure it on the target file system. Base or root folder ACLs must be configured before data replication begins from source to target.

Source file systems that contain file sets have their directory structure and files replicated to the target, but file set information is not preserved. It is similar to what a `restore` or `cp -R` command enacts across file sets.

Restriction: Special characters in the source and target paths that are given to the asynchronous replication CLI commands must not contain the following characters:

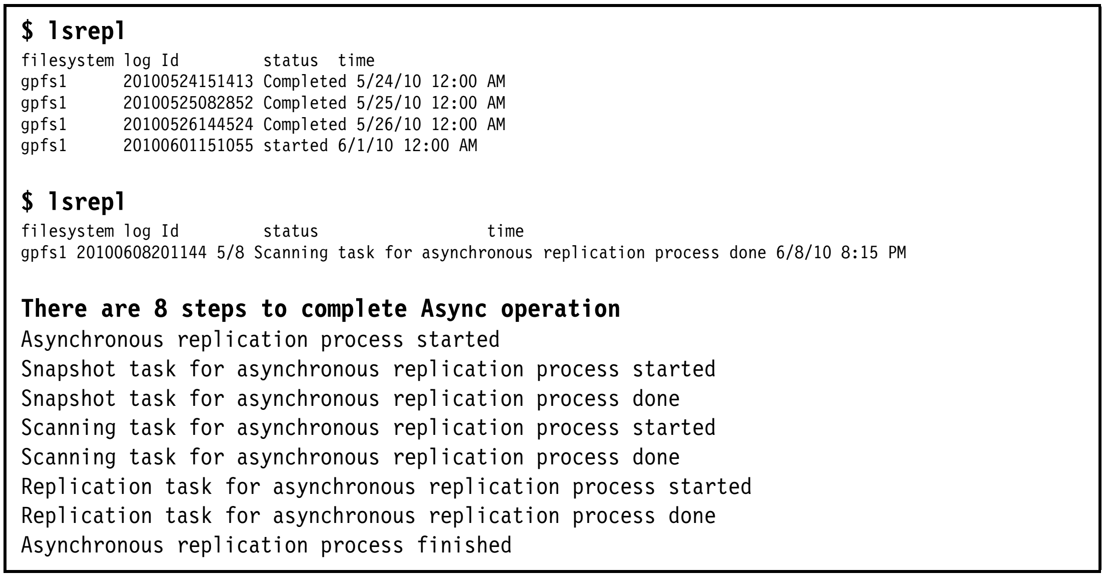
- Colon (:)
- Backslash (\)
- Backslash + lowercase n (\n)
- Backslash + lowercase r (\r)
- Backslash + lowercase t (\t)
- Asterisk (*)
- Question mark (?)
- Exclamation mark (!)
- Comma (,)
- Percentage sign (%)
- White space
- Open or close parenthesis
- Single or double quotation mark

The underlying paths in the directory tree that is being replicated are allowed to have the previously mentioned special characters.

7.2.5 Replication status

In SONAS V1.4.1 and later the **lsrepl** command displays the status of asynchronous replications the amount of data that is being replicated, the current rate of transfer, and estimated time of completion.

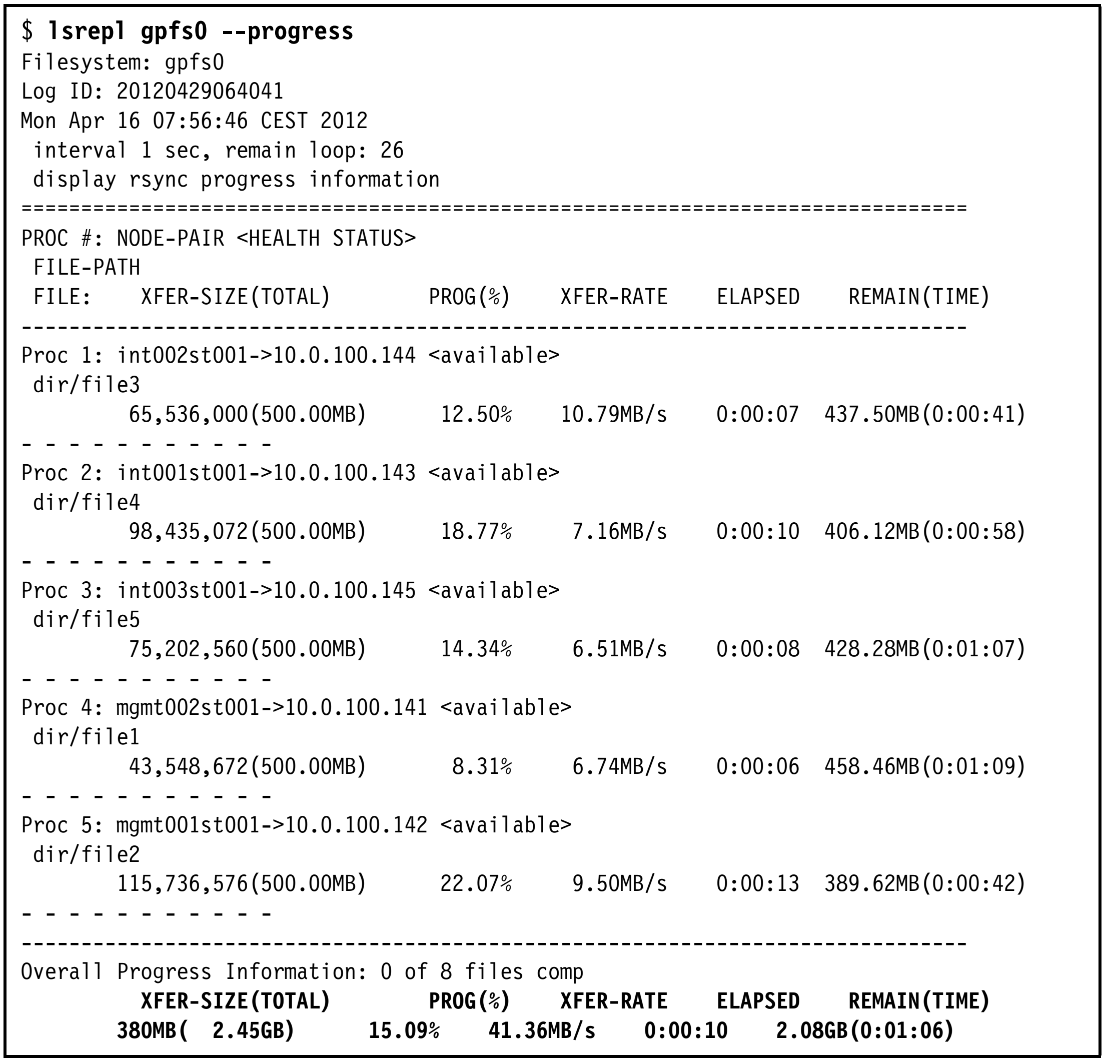
Figure 7-5 shows the **lsrepl** command with no parameters.



```
$ lsrepl
filesystem log Id         status  time
gpfs1      20100524151413 Completed 5/24/10 12:00 AM
gpfs1      20100525082852 Completed 5/25/10 12:00 AM
gpfs1      20100526144524 Completed 5/26/10 12:00 AM
gpfs1      20100601151055 started 6/1/10 12:00 AM

$ lsrepl
filesystem log Id         status                 time
gpfs1 20100608201144 5/8 Scanning task for asynchronous replication process done 6/8/10 8:15 PM

There are 8 steps to complete Async operation
Asynchronous replication process started
Snapshot task for asynchronous replication process started
Snapshot task for asynchronous replication process done
Scanning task for asynchronous replication process started
Scanning task for asynchronous replication process done
Replication task for asynchronous replication process started
Replication task for asynchronous replication process done
Asynchronous replication process finished
```

Figure 7-5 Example of lsrepl command with no options

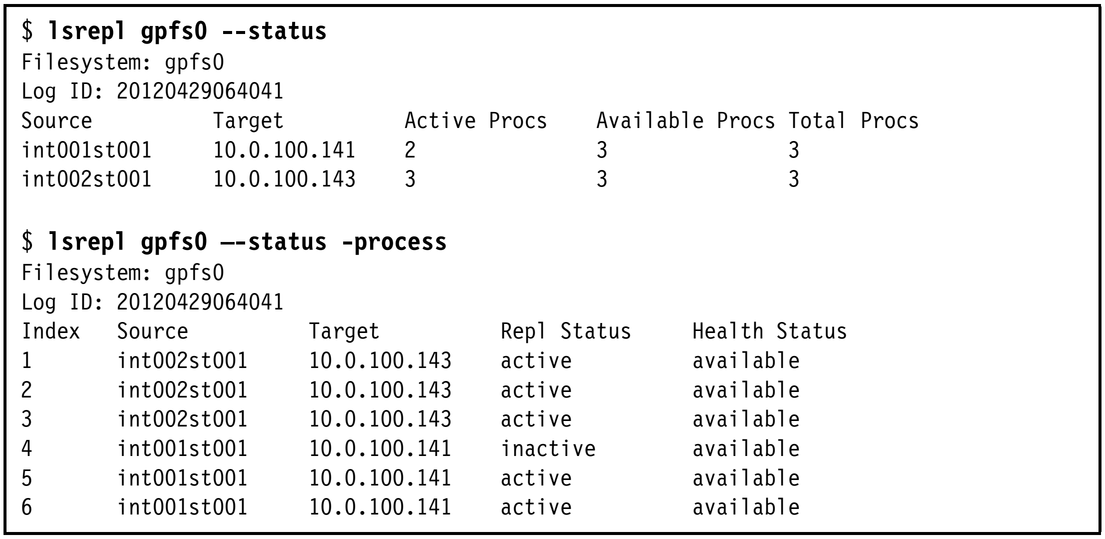
You can use the `--progress` option to display the status of active (running), available (not banned), and banned replication processes for each file module, as shown in Figure 7-6. Information about the transferred size, progress, transfer rate, elapsed time, remaining size, and remaining time for each process that is running in the asynchronous replication is displayed. The **`rsync`** progress and overall progress information are provided.



```
$ lsrepl gpfs0 --progress
Filesystem: gpfs0
Log ID: 20120429064041
Mon Apr 16 07:56:46 CEST 2012
 interval 1 sec, remain loop: 26
 display rsync progress information
===============================================================================
PROC #: NODE-PAIR <HEALTH STATUS>
 FILE-PATH
 FILE:    XFER-SIZE(TOTAL)        PROG(%)    XFER-RATE    ELAPSED    REMAIN(TIME)
-------------------------------------------------------------------------------
Proc 1: int002st001->10.0.100.144 <available>
 dir/file3
         65,536,000(500.00MB)      12.50%    10.79MB/s    0:00:07  437.50MB(0:00:41)
- - - - - - - - - - -
Proc 2: int001st001->10.0.100.143 <available>
 dir/file4
         98,435,072(500.00MB)      18.77%     7.16MB/s    0:00:10  406.12MB(0:00:58)
- - - - - - - - - - -
Proc 3: int003st001->10.0.100.145 <available>
 dir/file5
         75,202,560(500.00MB)      14.34%     6.51MB/s    0:00:08  428.28MB(0:01:07)
- - - - - - - - - - -
Proc 4: mgmt002st001->10.0.100.141 <available>
 dir/file1
         43,548,672(500.00MB)       8.31%     6.74MB/s    0:00:06  458.46MB(0:01:09)
- - - - - - - - - - -
Proc 5: mgmt001st001->10.0.100.142 <available>
 dir/file2
        115,736,576(500.00MB)      22.07%     9.50MB/s    0:00:13  389.62MB(0:00:42)
- - - - - - - - - - -
-------------------------------------------------------------------------------
Overall Progress Information: 0 of 8 files comp
          XFER-SIZE(TOTAL)        PROG(%)    XFER-RATE    ELAPSED    REMAIN(TIME)
        380MB(  2.45GB)      15.09%    41.36MB/s    0:00:10    2.08GB(0:01:06)
```

Figure 7-6 Example of lsrepl command with file system details

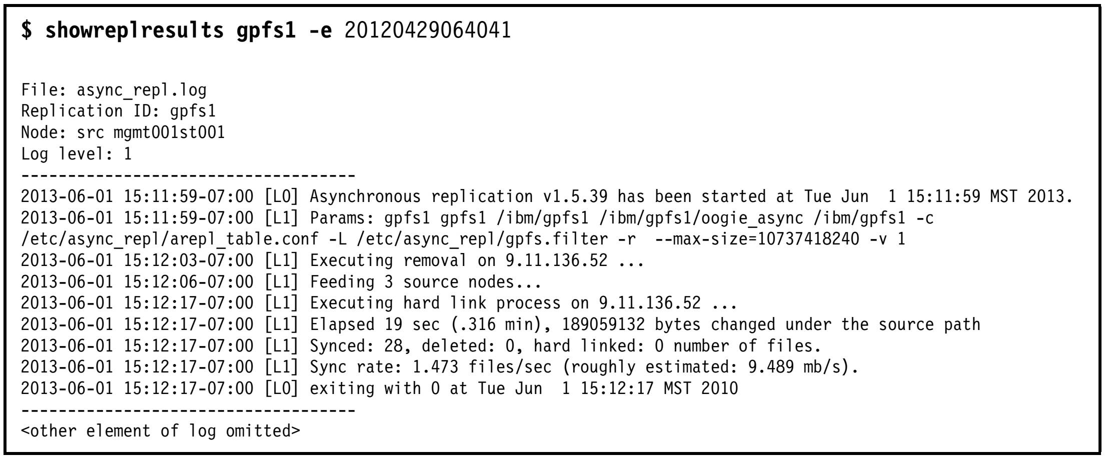
You can only use the `--process` option when you are also using the `--status` option in the same instance of an **`lsrepl`** CLI command submission. The Health Status column displays whether the replication process is available (not banned) or not available (banned due to an error). See Figure 7-7.



```
$ lsrepl gpfs0 --status
Filesystem: gpfs0
Log ID: 20120429064041
Source          Target          Active Procs    Available Procs Total Procs
int001st001     10.0.100.141    2               3               3
int002st001     10.0.100.143    3               3               3

$ lsrepl gpfs0 --status -process
Filesystem: gpfs0
Log ID: 20120429064041
Index   Source          Target          Repl Status     Health Status
1       int002st001     10.0.100.143    active          available
2       int002st001     10.0.100.143    active          available
3       int002st001     10.0.100.143    active          available
4       int001st001     10.0.100.141    inactive        available
5       int001st001     10.0.100.141    active          available
6       int001st001     10.0.100.141    active          available
```

Figure 7-7 Example of options for the lsrepl command

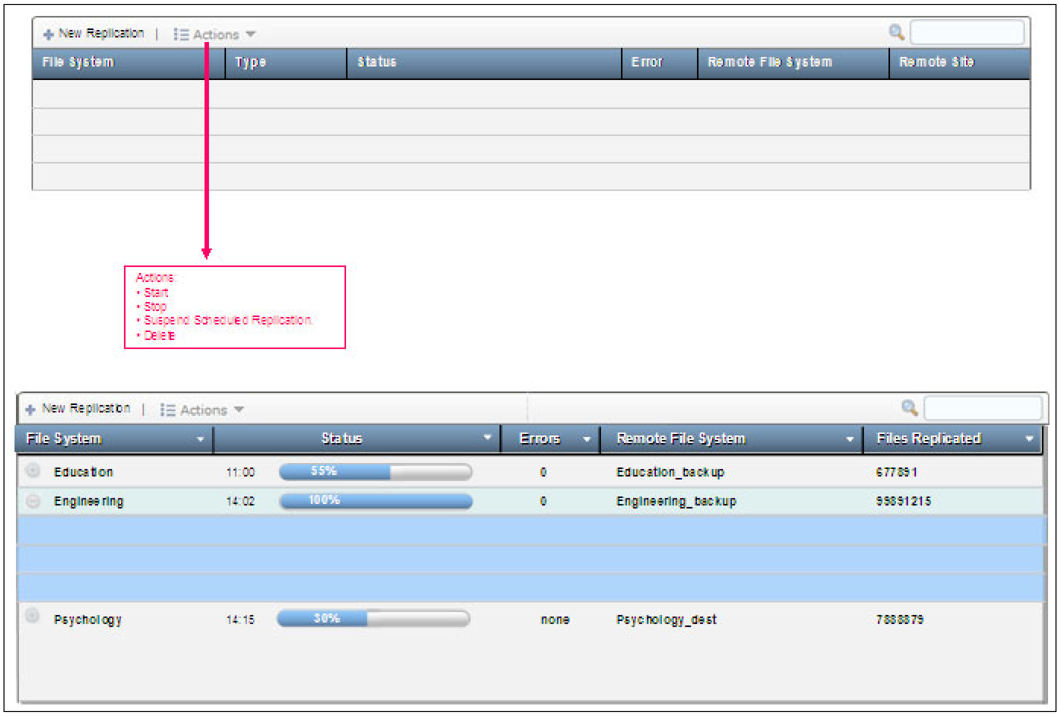
The **`showreplresults`** command displays the number of files that are changed and the number of delta changes in the data between two snapshots or a snapshot and the current file system state. Figure 7-8 shows sample output from the **`showreplresults`** command.



```
$ showreplresults gpfs1 -e 20120429064041

File: async_repl.log
Replication ID: gpfs1
Node: src mgmt001st001
Log level: 1
------------------------------------
2013-06-01 15:11:59-07:00 [L0] Asynchronous replication v1.5.39 has been started at Tue Jun  1 15:11:59 MST 2013.
2013-06-01 15:11:59-07:00 [L1] Params: gpfs1 gpfs1 /ibm/gpfs1 /ibm/gpfs1/oogie_async /ibm/gpfs1 -c
/etc/async_repl/arepl_table.conf -L /etc/async_repl/gpfs.filter -r  --max-size=10737418240 -v 1
2013-06-01 15:12:03-07:00 [L1] Executing removal on 9.11.136.52 ...
2013-06-01 15:12:06-07:00 [L1] Feeding 3 source nodes...
2013-06-01 15:12:17-07:00 [L1] Executing hard link process on 9.11.136.52 ...
2013-06-01 15:12:17-07:00 [L1] Elapsed 19 sec (.316 min), 189059132 bytes changed under the source path
2013-06-01 15:12:17-07:00 [L1] Synced: 28, deleted: 0, hard linked: 0 number of files.
2013-06-01 15:12:17-07:00 [L1] Sync rate: 1.473 files/sec (roughly estimated: 9.489 mb/s).
2013-06-01 15:12:17-07:00 [L0] exiting with 0 at Tue Jun  1 15:12:17 MST 2010
------------------------------------
<other element of log omitted>
```

Figure 7-8 Example output of the showreplresults command

Figure 7-9 shows the SONAS GUI for managing replication tasks.



File System	Type	Status	Error	Remote File System	Remote Site

Actions:
- Start
- Stop
- Suspend Scheduled Replication
- Delete

File System	Status	Errors	Remote File System	Files Replicated
Education	11:00 55%	0	Education_backup	677891
Engineering	14:02 100%	0	Engineering_backup	99891215
Psychology	14:15 30%	none	Psychology_dest	7888879

Figure 7-9 GUI for managing the replication tasks (start, stop, suspend, delete)

7.2.6 Recovery point objectives

The RPO is the amount of time between data being written at the production site and the same data being written at the DR site. If a disaster at the production site occurs during this time, any data that is written during that time frame might be lost if the production site is unrecoverable. The RPO for asynchronous replication is based on the frequency of the **`startrepl`** command, and is limited by the duration of the asynchronous replication operation process.

Only one asynchronous replication can be running at any time. Therefore, the minimum RPO is defined by the time of the completion from one incremental to the time of completion of the next incremental.

The first step in the asynchronous process is to create a point-in-time snapshot of the source file system that creates a data consistency point. This snapshot is used as a base to do the changed file (delta) scan and the source of the file replication.

New data that arrives to the source file system after the replication was initiated, when the point-in-time snapshot was created, is not replicated during this replication cycle. This new data is replicated as part of the next replication window.

Replication duration

The duration of the replication depends on several factors, which are described here:

- Number of files in the source file system

  The number of files that are contained in the source file system affects the duration of the scan time. Asynchronous replication uses a high-speed file system scan to quickly identify the files that changed since the last successful asynchronous replication. Although the scan is optimized, the number of files that is contained in the file system adds to the time it takes to run the scan.

  The scan process is distributed across the interface nodes that are configured to participate in the asynchronous process. Also, the files with changed data must be read to understand the data that changed. Therefore, what needs to be sent to the target system, the data size, and the number of files that are changed all affect replication performance.

- Amount of changed data that requires replication

  The time that it takes to transfer the contents from the source SONAS to the target is a direct result of the amount of data that was modified since the last asynchronous operation. The asynchronous operation moves only the changed contents of files between the source and target to minimize the amount of data that needs to be sent over the network.

- Bandwidth and capabilities of the network between SONAS systems

  The network capabilities play a large factor in the time that it takes the replication process to complete. Enough network bandwidth must be available to handle all of the updates that occurred since the start of the last increment before the start of the next scheduled increment. Otherwise, the next **`startrepl`** command fails, which effectively doubles the RPO for this time.

- Number of source and target nodes that are participating in asynchronous replication

  The number of interface nodes and the network capabilities between the systems is also a factor in the time that it takes to replicate the amount of changed data. The interface nodes work in parallel to transfer the changed content to the target.

- Other workloads that are running concurrently with replication
- Type and number of disks that form the source and target file systems

  The disk type of the file systems at the source and target help determine the I/O capabilities of the nodes that are participating in the replication process. Serial-attached SCSI (SAS) drives have a higher degree of I/O capacity than Serial Advanced Technology Attachment (SATA) drives, which enables them to move data faster, especially if they are doing replication alongside production workload.

- Using HSM managed file systems at source or target file systems

  HSM managed file systems can negatively affect the time that it takes for asynchronous replication to complete if the changed files in the source file system must be recalled from secondary media before asynchronous replication replicates the files to the target.

- Number of replication processes per node

  This configuration parameter enables for more internal **`rsync`** processes to be used on each node to replicate data to the other side. This increase provides more parallelism, and increases the potential replication bandwidth and rate.

- Encryption cipher

  The data that is transferred between SONAS systems is encrypted using SSH as part of the replication. The default cipher is strong, but limits the maximum per-process network transfer rate to approximately 35 - 40 megabytes per second (MBps). The fast cipher is not as strong, but increases the per-replication process network transfer rate to approximately 95 MBps. Consider requirements for security versus speed when you make the choice.

- Software compression

  This compresses the data to be transferred over the network before being encrypted. The compression is done with the processor of the node that is transferring the data.

  For data that is compressible, this compression can provide a means of reducing the bytes being sent over the network to increase the effective bandwidth. When large data transfers are expected, and overall cluster use shows that it can handle the extra processor workload on the interface nodes, compression is preferred. If the interface nodes are at high usage levels, and the size of the changes is small, compression might not be advantageous.

7.2.7 Replication validation

The **lsrepl** command displays the list of currently running and previous historical asynchronous operations from this SONAS cluster. This command can be run from the source SONAS management node.

This command displays all of the currently running asynchronous operations, what step of the replication process they are on, and the history of previous asynchronous operations. Asynchronous history is kept until it is cleared with the **cleanuprepl** CLI command. The point here is that it is important to put review of replication task success in your daily task list.

Each replication keeps an archive of the logs that are generated by the processes on both the source and target SONAS systems. These archives are preserved until they are removed using the **cleanuprepl** command that is described later in this section.

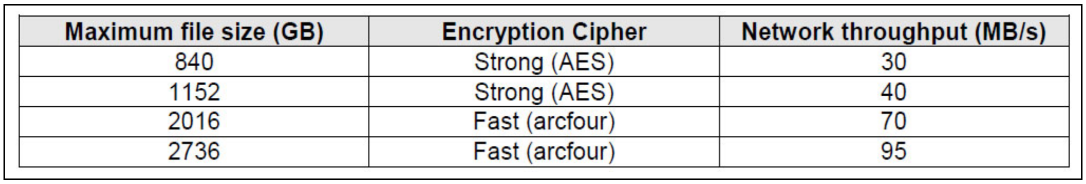
The contents of the logs are helpful for determining specifics of each replication cycle. The logs include detailed error information when a replication fails. This information in these logs is available using the **showreplresults** CLI command.

Cleaning up old asynchronous replication information is useful for freeing the space that is used to keep the information. Therefore, it is a good idea to add cleaning up the replication log once a week as a matter of routine weekly tasks for the cluster admin (when replication proves successful). For more information, see the documentation in the SONAS section of the IBM Knowledge Center:

http://pic.dhe.ibm.com/infocenter/sonasic/sonas1ic/index.jsp

For workloads that create larger files, there can be an issue with being able to replicate a large file within an 8-hour window, depending on the capabilities of the system (disk types, number of disks, network bandwidth, encryption cipher, and so on). The option of disabling the timer to allow these files to complete replication is required in these cases. Contact IBM support for assistance in temporarily disabling the timer.

Figure 7-10 shows some of the file sizes that can be replicated within the 8-hour window against the encryption cipher and varying average network throughputs.



Maximum file size (GB)	Encryption Cipher	Network throughput (MB/s)
840	Strong (AES)	30
1152	Strong (AES)	40
2016	Fast (arcfour)	70
2736	Fast (arcfour)	95

Figure 7-10 Cypher encryption replication modeling table

7.2.8 Asynchronous replication logs

There are several logs that are related to asynchronous replication:

- Current logs. Active logs that are being used by asynchronous replication in each of the participating source and destination nodes.
- Collected logs. Logs that were collected for the last asynchronous replication from all the nodes to one dedicated location.
- Archived logs. Historical logs from previous asynchronous replications.

Replication logs are segmented in two locations

Small files are kept under the node's local file system on the source management node (`/var/log/cnlog/async_repl/archive`).

Large files are in the GPFS file system that is participating in the asynchronous relationship (for example, `/ibm/gpfs0/.async_repl`, where `gpfs0` is the file system that is defined in this asynchronous relationship).

7.2.9 Issues and expectations

This section describes issues that are related to asynchronous replication:

- Source or target interface node failures during an asynchronous replication are automatically recovered by the asynchronous process:
  - Files that are replicated through these nodes are reassigned to surviving nodes.
  - Source interface node failures move another source interface node to the configured target interface node.
  - Target interface node failures cause the source node to resume the replication to the destination IP address of the failed node, which has been reassigned to another destination node.
- Source management node failures cause the currently running async operation to fail:
  - If the asynchronous process is still in the scan phase, the current async process fails without transferring any files.
  - If the asynchronous process is in the file transfer phase, the interface nodes continue to transfer their current file lists, but the overall asynchronous process fails.
- Network failures between a subset of nodes cause work to be distributed to a source interface node that can still reach the destination.
- Network failures that cause all communication to be lost fail the overall asynchronous operation.

Effects of an overall failure during asynchronous replication

An overall failure to the asynchronous process has the following effects:

- The current replication halts until the next replication is initiated, either manually or when the next scheduled starts.
- The source-side snapshot is retained at the point in time that the failed asynchronous operation created it.
- The next asynchronous operation continues to attempt the failed replication attempt.
- An asynchronous operation must complete successfully before another asynchronous operation can take place at a new RPO point in time.
- Upon successful completion of the replication, the recovery point is the one established at the start of the first failed replication.
- Depending on the type of failure condition, an asynchronous lock file is left to suspend future asynchronous operations from starting until corrective action is taken.
- The `startrepl` command has a parameter to clear the lock file, which tells the asynchronous replication service to attempt to do asynchronous operation again (presumably after corrective action is taken to clear the condition).

For information about asynchronous cluster replication, see the *IBM SONAS Implementation Guide*, SG24-7962 IBM Redbooks publication.

7.3 Backup and restore solutions

Backup and restore solutions are a common requirement for file data protection. SONAS currently supports two solutions for backup and restore. The first solution is full integration with IBM Tivoli Storage Manager. Alternatively, a Network Data Management Protocol (NDMP) backup and restore solution can be used.

Backup and restore solutions add value to most NAS solutions with little added complexity. As you approach the higher scale of NAS (such as multi-petabyte (PB) solutions), SONAS, like other Scale Out NAS solutions, begins to lose its value with backup and restore tape-based solutions. It soon becomes too much data to back up, track, and manage.

The cost, complexity, and data scan times soon become too much of a burden, and large-scale data restore from tape takes longer than is tolerable. When the scale reaches these boundaries, it becomes important to back up only data that requires backup, and consider replication as the better option for protecting large-scale data.

Replication is the fastest way to recover access to replicated data. A well-planned disaster recovery strategy for even a multi-petabyte SONAS solution can typically be recovered in less than 30 minutes by a disaster recovery team that is familiar with SONAS. For this reason, replication is the best-supported vehicle for disaster recovery for SONAS.

7.3.1 Tivoli Storage Manager

This section describes Tivoli Storage Manager and how it works with SONAS.

Key backup and archive concepts

The key concepts for consideration with any backup system are the amount of new, changed, or deleted file-based data that might be lost because it is not yet backed up, and the time it takes to recover an entire system if the production system suffers a failure or disaster.

Consider the following terminology for Tivoli Storage Manager backup:

Tivoli Storage Manager Server

This is the Tivoli Storage Manager Server that holds Tivoli Storage Manager license keys, and the backup database that is used for Tivoli Storage Manager backups of SONAS. The preferred practice is to size and tune it appropriately for the size of the SONAS environment that you want to back up.

Tivoli Storage Manager Clients

These clients are the SONAS interface nodes that participate in the workload that is associated with SONAS data backup and restore.

Tivoli Storage Manager Storage Class

This is a device class that represents a device type (such as disk or tape) that Tivoli Storage Manager can use to determine which types of devices and volumes are available to store client-node data in primary storage pools, copy storage pools, and active-data pools. Device classes are also important for storing database backups and for exporting and importing data.

Sequential-access device types include tape, optical, and sequential-access disk.

For random access storage, Tivoli Storage Manager supports only the *DISK* device class, which is defined by Tivoli Storage Manager.

Figure 7-11 shows a basic Tivoli Storage Manager backup solution for SONAS.

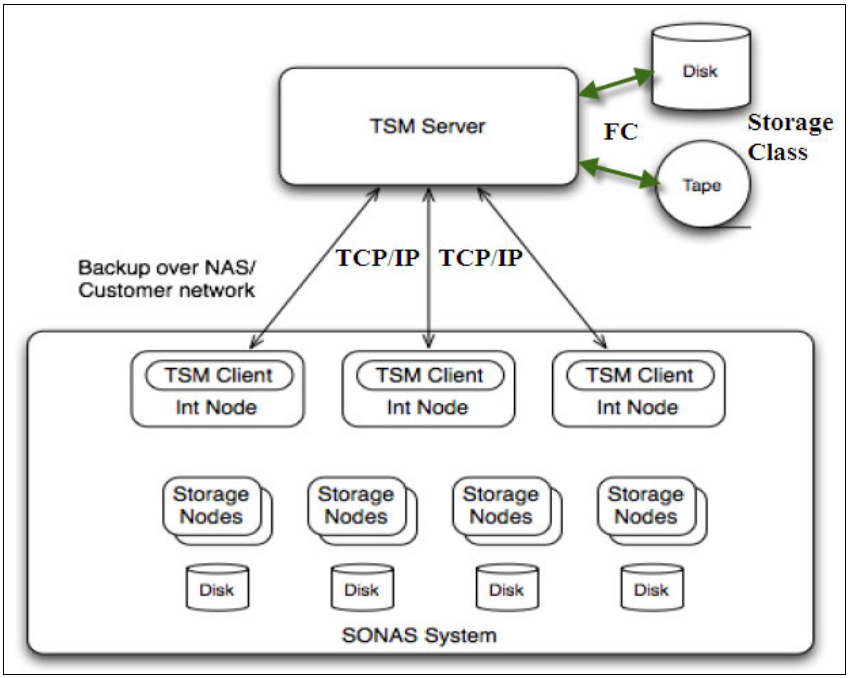


Figure 7-11 Basics in a Tivoli Storage Manager backup solution for SONAS

SONAS and Tivoli Storage Manager considerations

Although Tivoli Storage Manager is tightly integrated with SONAS and Storwize V7000 Unified, it is important to remember that the Tivoli Storage Manager Server side of that relationship is independently managed. It is out of the general scope of SONAS configuration and preferred practice.

However, this section provides a few preferred practices for the Tivoli Storage Manager Server and supporting SONAS side preferred practice considerations.

Configuration

Before configuring SONAS backups on a Tivoli Storage Manager server, complete the following tasks:

- Work with Tivoli Storage Manager subject matter experts (SMEs) to ensure that the server is configured optimally for reliability, sizing, and performance. IBM offers fee-based Tivoli Storage Manager Server sizing and installation services:
  - Networking bandwidth and redundancy
  - Processor and memory
  - Kernel and protocol tuning
  - Database, logs, and file system sizing and performance
  - Software updates and system patching
- Run annual or semi-annual Tivoli Storage Manager Server health checks to ensure that the server and all associated hardware, software, tuning parameters, licenses, and support are correctly sized, configured, and maintained for use in protecting SONAS or Storwize V7000 Unified data.

Recovery point objective, recovery time objective, and versioning

The industry terms for these concepts are *recovery point objective* (RPO) and *recovery time objective* (RTO). Another concept that is important for backup is *versioning*. Versioning is the ability to retain older versions of files for restoration in the case of some operational or accidental corruption.

Recovery point objective

The RPO is expressed as the amount of time for which data can be lost in a disaster. Specifically, it is the amount of time that it takes between the write of data to a storage device and the time it takes to back it up and move it to a location that will not participate in the same disaster that befalls the production site.

Usually, continuous copy replication methods have low RPOs, ranging from zero RPO for synchronous replication to tens of seconds for some of the asynchronous replication methodologies. Traditional backup to tape, such as Tivoli Storage Manager backup, can have daily or even weekly RPOs.

If you are using tape to make a backup copy, it is not just the time between backups and the time it takes for the backup to complete. The tape data must be in a location that is secure from a disaster before the RPO is met. This time includes boxing the tapes and transporting them to a safe facility, or running the backups from a remote location from the primary data. The use of “hot” sites and electronic vaulting are used to reduce and sometimes eliminate this aspect of the RPO that is associated with tape backups.

Recovery time objective

The recovery time objective is also expressed as the time that it takes to recover and get back into production after a disaster. This time includes the time that it takes to activate the recovery site and recover all of the data that is necessary to start production, including starting the applications on the servers to resume production.

There might be different RTOs for different applications or lines of business. Critical operations might have much smaller RTOs than non-critical applications. For example, a customer order system might be considered a critical system to protect order status and cash flow. Studies that have been done on RTO suggest that most business-critical applications must have an RTO of 48 hours or less.

This might or might not reflect the requirements of your company. However, the availability of tapes, tape drives, drive restore speeds, and bandwidth from the backup server and recovery network are extremely important to consider in the RTO objectives for extremely large NAS services. Petabytes from tape might take weeks to recover, where DNS updates for replication sites might be adequately updated within a few minutes, or at worst hours, with even 5 - 10 PB in shared data repositories.

Versioning

Versioning is the ability to retain multiple versions of changed files for later recovery. It is important if the detection of the corruption has not been identified before a file is saved, or if a history of updates to a file is needed. When a file is detected as being changed during the backup process, a customer might retain several versions even after the latest version of the file is deleted. This retention policy is applicable to the number of versions that are maintained and the duration of time that they are retained before the version is deleted.

GPFS scans

The Scan time part of the replication time scales linearly with the number of inodes in the source file tree (source file system).

> **Important:** Snapshot capabilities use GPFS scans. This GPFS capability provides the file system with a way to do a high-speed scan of the GPFS inodes. As such, it is important to understand the different elements that affect GPFS scan performance.
>
> The multiple interface nodes equally spread the policy engine rule evaluation, file scan identification, and subsequent data movement for increased performance. If greater scan speed is required, more SONAS nodes can be allocated to the scan, and each node scans only its equal portion of the total scan. The results of the parallel scan are aggregated, and returned as the usable list of candidate files.

The SONAS software high-performance scan engine is designed to use the multiple interface nodes of the SONAS in parallel to scan the internal file system metadata called inodes. SONAS asynchronous replication, the **`startbackup`** SONAS backup CLI command, and SONAS snapshot capabilities all use GPFS scans.

Given this description, it is easy to understand that scans use a large amount of I/O and processing resources. This adds to the overall stress of the system. Therefore, minimize scans and only run them when there are sufficient resources both in the interface nodes and storage controller that contain the GPFS metadata. This availability is generally not a problem because the backup window normally is done during a slack production period.

Scan factors

The time that it takes to complete a scan is influenced by several factors:

- System load. In particular if other scans are running at the same time.
- The number and type of system disks that hold the inodes (metadata).
- How the disk pools are set up. If the system disks are dedicated to inodes or are shared with user data.
- The number of interface nodes.
- The number of files and directories that are contained in the file system or independent file set that is being scanned.
- The complexity and depth of the directory structure that is being scanned.
- The size of the files.

Metadata scans

There are several processes that use metadata scans:

- The SONAS `startbackup command`
- SONAS asynchronous replication
- Policy scan to determine new, changed, and deleted files, or their base file placement
- Snapshot deletions, even as part of the asynchronous replication consistency mechanism
- Snapshot deletes that are initiated by the customer for customer-created scans, and snapshots that are initiated by asynchronous replications on both the source and target machines
- Antivirus scans

Policy scans

The GPFS scan has two phases:

1. A directory traversal to collect the full path to each file.
2. A sequential scan over each file’s inode to collect the remaining file attributes. This second scan also does the policy rule evaluation, which selects the files that match the policy rules.

Startbackup LAN-based backup through Tivoli Storage Manager

The SONAS `startbackup` command uses Tivoli Storage Manager backup processing to backup files to an external Tivoli Storage Manager server by using the embedded Tivoli Storage Manager backup client. This client runs the GPFS `mmbackup` command with special parameters and backs up file by file system.

This command is unaware of the nature of file sets. It backs up files by path association and directory, whether file sets are linked or not. For this reason, it is important to make sure that file sets are not unlinked during replication or backup processes.

Advantages of using Tivoli Storage Manager

The advantages to using Tivoli Storage Manager are described in this section.

GPFS scan

As the number of files continues to grow in a single file system, the time that is required for this scan using the traditional “walk the directory tree” method becomes a major obstacle to completing the backup within the backup window.

At the mid-end or high-end scale of SONAS, it becomes infeasible to use the traditional backup method of walk the directory tree to identify new changed and deleted files for backup processing. To make the process of identifying files that need to be processed by Tivoli Storage Manager backup, the SONAS `startbackup` command is specifically designed to use a high-performance, high-speed GPFS scan engine that is available through the GPFS `mmbackup` command.

The GPFS `mmbackup` command is an integrated part of the GPFS file system and serves as an interface to the embedded RPO and RTO client to assist the backup process. As a part of `mmbackup`, there is an internal database of file system metadata that is called the *shadow file*. This database is designed for the integrated scan engine. The goal of `mmbackup` is to replace the walk the directory tree method with a higher-performance GPFS scan.

HSM

The SONAS **startbackup** support is the only solution that enables for the mix of data protection and tape-based storage tiering using the IBM HSM components of the Tivoli Storage Manager product.

GUI and CLI

Setup, monitoring, and administering backup is supported through the SONAS CLI and GUI.

Additional considerations

It is important to understand the functions of the SONAS `startbackup` command, which uses the GPFS `mmbackup` command and an internal Tivoli Storage Manager client to an external Tivoli Storage Manager server for backup. Backup of a cluster, file system, and file set configuration data is not captured with file system backups. For this, it relies on secondary management nodes, site cluster replication, and internal tasks that replicate critical cluster management data between the management nodes.

Tivoli Storage Manager database object count considerations

There are limitations on how many objects the Tivoli Storage Manager IBM DB2® database has been tested with, and, therefore are officially supported. With Tivoli Storage Manager 6.3, the statement of support was a maximum of 4 billion objects per server. However, this number is unlikely to be achievable in reality due to operational inefficiencies or limitations in the usual disk, network, processor, or memory in the SONAS or Tivoli Storage Manager infrastructure.

An object is used for each version of a file. If the file is being managed by HSM, it uses an additional object (for a stub and migrated file). For example, if there are 200 million files (all backed up) and each file has three inactive versions retained, and the files are being HSM managed, there are about 1 billion objects that are used in the Tivoli Storage Manager DB2 database. Tivoli Storage Manager data deduplication also affects capacity. You can see how a single file becomes multiple objects in the Tivoli Storage Manager database.

Check the current statement of support for the version of Tivoli Storage Manager server that is being used to determine the maximum number of files that can be backed up and managed by a single Tivoli Storage Manager server. With extremely large clusters, the value of replication exceeds that of backup for some parts data protection. However, replication will not necessarily provide file versioning or a way to recover from data loss or corruption. Large data structures might require multiple file systems for improved backup (or data protection) strategies.

> **Tip:** When planning for backup, if you have a file system backup that will approach 1 billion objects, you should consider creating additional file systems. Contact IBM support for more guidance and assistance.

Single Tivoli Storage Manager server for a single GPFS file system

There is an **mmbackup** restriction that a GPFS file system can be backed up only by a single Tivoli Storage Manager server. The design of **mmbackup** does not support the coordination of backups across multiple Tivoli Storage Manager servers. It is possible to back up multiple file systems to a single Tivoli Storage Manager server. This restriction, combined with the Tivoli Storage Manager DB2 restriction mentioned previously, limits the number of files that can be backed up and managed in a single Tivoli Storage Manager server database.

GPFS scan times

The **mmbackup** command uses GPFS scans to significantly decrease the time when compared to the traditional walking the directory tree scans to determine new, changed, and deleted files for backup processing. Although faster, the GPFS scan does take time to complete. Generally, GPFS scans take about 40 minutes for each 100 million files. So, the preferred practice is to set RPOs that are appropriate to these expectations and adjust them based on performance.

Sequential mark for expiration

The **mmbackup** command manages deleted files with a step that is called *mark for expiration* or *expiration processing*. This expiration processing must be completed before the start of the process that copies new and changed files for backup. Generally, expiration processing can process about 5 million files per hour. If many files are deleted between backups, a significant amount of time in the backup window can be taken by expiration processing. Consider this time when you are setting RPO time expectations.

Shadow file rebuilds

The **mmbackup** process uses an internal file to track information that is needed to determine if a file is new, modified, or deleted. This file is stored in GPFS and is a shadow of the Tivoli Storage Manager DB2 database. This file is rebuilt if the **mmbackup** command detects a problem with the shadow files or if the **mmbackup** process did not complete successfully.

The shadow file is rebuilt using an interface with Tivoli Storage Manager and might take time, depending on the number of files that Tivoli Storage Manager is tracking (Tivoli Storage Manager DB2 objects). Extremely high numbers of objects affect the RPO expectations.

Considerations on file sets

The **mmbackup** command and Tivoli Storage Manager are unaware of the existence of file sets. When you are restoring a file system that was backed up by Tivoli Storage Manager, the files are restored to their original path names, regardless of which file sets they were part of. Tivoli Storage Manager has no mechanism to create or link file sets during restore.

Therefore, if a file system is migrated to Tivoli Storage Manager and file sets are unlinked or deleted, or linked files are restored to a different location, restore or recall of the file system does not restore the file sets. During a full restore from backup, all file set information is lost and all files are restored into the root file set. This is a limitation that affects all backup products, but it is important to understand.

Backups on entire file systems only

Tivoli Storage Manager backups can be taken only for a file system, but not for a specific file, or for a path (no selectable backup). The first backup is of the entire file system, then incremental are captured from there. Restore can be done on file or path level. However, individual files or directories can be restored.

Application-consistent backups are not supported

The **startbackup** command does not support an application-consistent backup. It uses the active file system view to back up data, and therefore changes from file to file occur as the backup proceeds.

Special Tivoli Storage Manager tuning considerations

RESOURCEUTILIZATION in the SONAS Tivoli Storage Manager (backup-archive client tuning parameters) is set to 10. This might cause issues, for example, when Tivoli Storage Manager has no disk storage pool and the tape library does not have enough tape drives attached to manage the work. This limitation can result in frequent errors, such as This node has exceeded its maximum number of mount points or something similar. To resolve the issue, reduce the number to improve work distribution to a specified number of targets.

This value regulates the number of concurrent sessions during processing.

Sessions

Each interface node can open up to eight sessions on the Tivoli Storage Manager server. Therefore, the Tivoli Storage Manager server should set its maximum connections to accommodate this limit. A typical value for this is 100 sessions. Resource utilization in SONAS is set to 10, which means that a node that is configured as a Tivoli Storage Manager client can use up to seven tape drives per IBM SONAS configured Tivoli Storage Manager node.

> **Note:** Although you can run backups from several or many interface nodes, restores come from a single node. Avoid restoring large numbers of files with the Tivoli Storage Manager restore operation. Restoring a large number of files might require a lot of time.

Table 7-1 shows the maximum number of mount points that are used, by type of operation, per IBM SONAS-configured Tivoli Storage Manager node.

Table 7-1 Mount restrictions for resource utilization settings

Type of operation	Maximum number of mount points
Backup	Six per interface node, when RESOURCEUTILIZATION is set to 10
Restore	Six, when RESOURCEUTILIZATION is set to 10
Migration	Three per interface node
Recall	20 (This value is the default setting for a Tivoli Storage Manager server)
Reconcile	0

Use the RESOURCEUTILIZATION client option to regulate the number of concurrent sessions that are handled by Tivoli Storage Manager server and client operations, such as multiple backup, restore, archive, or retrieve operations. Tivoli Storage Manager Hierarchical Storage Management operations, such as migrate or recall, that are initiated by the IBM SONAS interface nodes that are configured for Tivoli Storage Manager Hierarchical Storage Management, require one or more sessions to the Tivoli Storage Manager server.

Multiple concurrent sessions might be required to perform one operation, depending upon the volume of the operation and the value that is set for the RESOURCEUTILIZATION option. The number specifies the level of resources the Tivoli Storage Manager server and client can use during processing. The range of values that you can specify is 1 - 10.

When set at the default value 10, up to 7 tape drives can be used per SONAS node. When there is no `disk` storage pool that is defined for Tivoli Storage Manager and HSM, reduce this number to match the number of drives. For example, if your tape library has 4 LTO5 drives, reduce this number to 4.

Tracking backup status and monitoring failures

It is important to consider daily status and failures in the cycle of daily backups. Review the status of backups daily. Complete the following tasks:

- Validate error codes and the status of files that failed in backups.
- Search for trends and common behavior in order to catch when the result is different from normal.
- Track scratch tape quantities regularly, and determine how quickly scratch tapes are normally used.
- Watch both SONAS file system capacities and Tivoli Storage Manager server file system capacities.

For more information about these tasks, see Chapter 8, “Monitoring” on page 261 and Chapter 9, “Troubleshooting and support” on page 279.

For more information and methods to monitor your SONAS, see the *Scale Out Network Attached Storage Monitoring*, SG24-8207 IBM Redbooks publication:

http://www.redbooks.ibm.com/abstracts/sg248207.html?Open

Additional Information

Additional limitations and planning items are described in the SONAS section of the IBM Knowledge Center:

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

7.3.2 IBM HSM

HSM is Hierarchical Storage Manager or Hierarchical Space Manager. It is a separate set of licensed binary files in the Tivoli Storage Manager product that require Tivoli Storage Manager backup licenses. It enables you to extend your file system to tape through a managed archive.

The HSM solution enables you to define criteria for which, when met, files are migrated from their disk locations to a location on tape devices. A simple file stub (or small subset of a file and the metadata) is used for a client to access the file.

For example, a client navigates a directory that has some HSM migrated files. What they see is an icon in Windows Explorer that indicates that the file is a *stub* file. When the client opens the file, it calls for the HSM manager to read the whole file from tape.

If supported by the application, the client can read the subset that is stored in the stub while the file is recalled back to a disk location. In the background, Tivoli Storage Manager calls for the data, tracks it to the tape or tapes that it is stored on, and restores that data to disk locations on the file server, where it is completely readable by the client.

Recall storm

The previously mentioned process is known as a *file recall*. This process is slower than reading files from disk locations. In some cases, it can be much slower (taking many minutes to recall), and, if there is tape drive contention for other tasks (such as backup or network issues), it can take hours.

For this reason, users of data in *space-managed data* should be educated on the possibility of recall delays, and how to avoid a *recall storm*. A recall storm is a process where a client recalls an exceptional amount of data, such as a recall of a large directory of files. This can tie up file system management resources, tape drive resources, and network bandwidth.

One example of a recall-storm-inducing task is to force an antivirus scan of millions of files in a large SONAS-migrated file set. This data can include millions of files that have not been read for a long time, and have been archived to tape because of a space management policy. These files are later scanned for antivirus signatures. In this scenario, all the files that are scanned would be recalled from tape and placed back on file system disk pool resources sequentially.

It is the preferred practice to ensure that files are scanned before being placed into space managed pools, and that you avoid any action that tries to read a massive number of files from space managed capacity, unless that is the wanted result and enough time and resources are known to be available to support the task. It is the preferred practice to use HSM migration policies as a tool for true archiving, and to manage capacity that might be accessed or read on a cheaper tier of storage.

As Tivoli Storage Manager and HSM can require much knowledge, take time to read about these externally managed services in the product brand materials, IBM Knowledge Centers, and Tivoli Storage Manager Administrator Guides, along with reading the sections for Tivoli Storage Manager in the *IBM SONAS Implementation Guide*, SG24-7962.

7.3.3 Network Data Management Protocol (NDMP)

The IBM SONAS system supports NDMP, which is an open standard protocol for network-attached storage (NAS) backup and restore functions. NDMP Version 4, provided by compatible Data Management Applications (DMAs) such as Symantec’s Veritas NetBackup, is supported.

> **NDMP backups with Tivoli Storage Manager:** This section describes NDMP backup by programs other than Tivoli Storage Manager. Although it might be technically possible to use Tivoli Storage Manager and back up with NDMP, Tivoli Storage Manager backup should be done with the SONAS Tivoli Storage Manager client.

Planning for Network Data Management Protocol

Full and incremental backup and restore of file system or independent file set data is provided by capturing data and metadata 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 maximum length of the names of files and directories that are backed up or restored using NDMP is 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.

A single file in the NDMP backup data stream can be restored by using the NDMP Direct Access Restore (DAR) feature if DAR is supported by the Data Management Application (DMA). All metadata, including ACL information, is restored along with the file contents.

The NDMP function is configured and managed on the IBM SONAS system by using the IBM SONAS command-line interface (CLI) from the active Management Node. NDMP alerts are written to the alert log on the IBM SONAS active management node. Viewing an alert log NDMP entry is the only IBM SONAS system NDMP feature that is accessible from the IBM SONAS graphical user interface (GUI). It should be a part of the daily routine when NDMP backup is deployed.

Full or differential backups

Full and differential backup and restore of file system data is provided by capturing all data and all metadata 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. 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.

NDMP-based differential backup triggers an internal parallel scan of all changed files and then makes them available to the NDMP server. NDMP-based backups are done over TCP/IP, and the backup server is responsible for streaming the returned data to tape.

Currently, NDMP for full or differential backups can be done at the directory or file system level, including snapshots.

After a full backup image is created, subsequent backups can be differential images. A differential image consists of all the files that have changed since the previous full image backup. The restore of a differential image automatically restores the differential image after the appropriate full image has been restored.

Advantages of NDMP

The advantages to using the NDMP methodology are discussed here.

Selectable backups

A subset of a file system can be backed up by specifying what is to be backed up at the file set level. This process enables specific directories to be removed from backup consideration if the versions of files that are held in those directories are already backed up, or a backup copy is not necessary.

GPFS scan

As the number of files continues to grow in a single file system, the time that is required for this scan with the traditional “walk the directory tree” method becomes a major obstacle to completing the backup within the backup window. To make the process of identifying files that need to be processed by NDMP-based backup, the GPFS scan is used.

Data consistent backups

The SONAS NDMP support uses the GPFS snapshot to provide a file consistent backup.

HSM

HSM functionality depends on Tivoli Storage Manager agents. If you are using NDMP backups with a non-Tivoli Storage Manager service, you cannot use a Tivoli Storage Manager service to provide HSM storage tiering.

Cumulative differential backups

For most agents, differential backups are cumulative. This means that each differential backup contains all changes that are accumulated since the last full backup. Each successive differential backup contains all the changes from the previous differential backup. This means that each subsequent differential backup takes longer to complete. At some time, a new full backup must be done to reset the differential backup. Periodically doing a complete backup affects the start of the next backup and affects the RPO.

No versioning

Tivoli Storage Manager like version retention of multiple versions of a file is not available through NDMP. There is no way to specify the number of versions for an individual file because retention is managed at the image level. Using the NDMP backup utility, it might be possible to manage versions of files can at the image level.

Single File recovery

Single file recovery is not available through NDMP.

Backup and Recovery metrics

This section describes the backup and recovery metrics as related to NDMP.

RPO

The RPO of backups that are created by the NDMP command depend on the scheduling of the backup and the time it takes to get the backup version to a secure site. Another factor is the time that it takes to complete the backup because the previous backup must complete before starting the next backup. This time becomes a problem when the differential backups must be reset with a full backup. This process is impractical for backing up large file systems.

The real advantage is the ability to maintain a single file system (name space) while removing a set of directories from backup consideration. By doing this process, an acceptable RPO can be accomplished.

RTO

For large-scale systems, this process becomes an issue because recovering petabytes of data can take a long time (high RTO times). This time is even more of a problem with NDMP, especially if there are several differential backups since the last full backup. This is because the full backup must be restored first, followed by the cumulative differential backup, which can contain changes for several days.

The management of files that are included in a backup by excluding files or by not causing a backup of the directories that the files are stored in is one way to lower the attainable RTO. This is achieved by designating specific directories to be backed up leaving the files in other directories without backup copies. This methodology enables both the removal of files for backup consideration and the recovery of critical files (stored in specific directories) to be restored to speed the resumption for production.

If there is accidental deletion and operational corruption, it is possible to restore a file system or file set. However, restore is at the file system or file set granularity, so recovery of a particular file is not possible. Files are restored at the image level, so all files in the image are restored. However, the entire image must be restored. Care must be taken to avoid overwriting files that are not part of the set of times that are needed for restoration with earlier versions.

Recovery through NDMP

The last full backup is recovered first. It is followed by the last incremental backup to complete the recovery.

Guidelines

The following list describes the major challenges for the backup from an NDMP share backup methodology at scale:

- During incremental NDMP backups, file system metadata is scanned to determine which files to back up. If the directory being backed up is very large, there might be a delay before data starts transferring.
- The time that it takes to restore the entire file system for disaster recovery.

Both of these challenges can be overcome by managing files in separate directories, and being selective about what needs to be backed up and recovered.

If it is possible to identify the files that are critical to recover first, it might be advantageous to place those files into a file system that will be recovered first for disaster recovery. This step might significantly lower the RTO.

For some applications that do not modify files, such as imaging, a directory by time period (year, quarter, month, and so on) can help, because the files for the previous period do not need to be “walked”. Backup of the current period is all that needs to be done.

7.4 Snapshots

This section describes GPFS snapshots for data versioning.

7.4.1 Managing snapshots

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 alone. However, running snapshots of a file system’s root file set captures snapshots of its dependent file sets, and enables you to manage the snapshots of independent file sets, independently.

The storage that is needed for maintaining a snapshot is due to the required retention of a copy of all of the data blocks that have been changed or deleted after the time of the snapshot, and is charged against the file system or independent file set quota. Therefore, SONAS snapshots are space-efficient because they only store data of blocks that have changed.

It is not advised to run scheduled snapshots of the file systems and all their independent file sets, because this process is redundant in the case of independent file sets.

Also, when there are many independent file sets in your file system, it is not advised that all file sets apply the same sets of rules. Instead, it is the preferred practice to break up the timing of the activity of snapshots and associated snapshot deletions so that different times can be used to achieve that feature enablement. (More information is provided later in this section.)

Snapshots are read-only. Changes can be made only to the normal, active files and directories, not to the snapshot.

The snapshot function enables a backup or mirror program to run concurrently with user updates, and still obtain a consistent copy of the data as of the time that the snapshot was created. Snapshots also provide an online backup capability that enables 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 independent file sets, or both, by the system administrator.

There are two steps to configure snapshot management for a file system or independent file set. First, create the rule or rules. Then associate the rule or rules with the file system or independent file set.

A user must have the Snapshot Administrator role to do snapshot management functions.

Snapshot rule

A *snapshot rule* indicates the frequency and timing of the creation and deletion 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.

7.4.2 Traditional backup versions compared to snapshot versions

A snapshot of an entire GPFS file system or independent file set can be created to preserve the file system’s or file set’s contents at a single point in time. A snapshot contains a pointer to the file data as it existed at the time of the snapshot creation, while enabling the normal files to continue to be updated. Files that are preserved in a snapshot are read-only. Therefore, snapshot is a way to preserve versions of files as they existed at some time in the past.

The snapshot function can be used to preserve one or more versions of all files in a file system or file set at a particular point in time. These snapshot versions can be used to recover a previous version of one or more files if a file is accidentally deleted or corrupted by an application, much like a file or files are recovered from a traditional backup.

There are some differences worth noting between the traditional backup versions and snapshot versions. The snapshot copy of the file cannot be moved to an off-site location. Note however, that a copy of a snapshot copy can be moved off site.

Another important difference is that the backup version is made when a file has been changed. The snapshot version is taken across the file system or file set at a particular time even if no files (or all files) have been changed.

If a file has not changed, there is still a version of that file that is saved in the snapshot. This means that a file that is not changed is recoverable from any snapshot that was taken even though the file did not change.

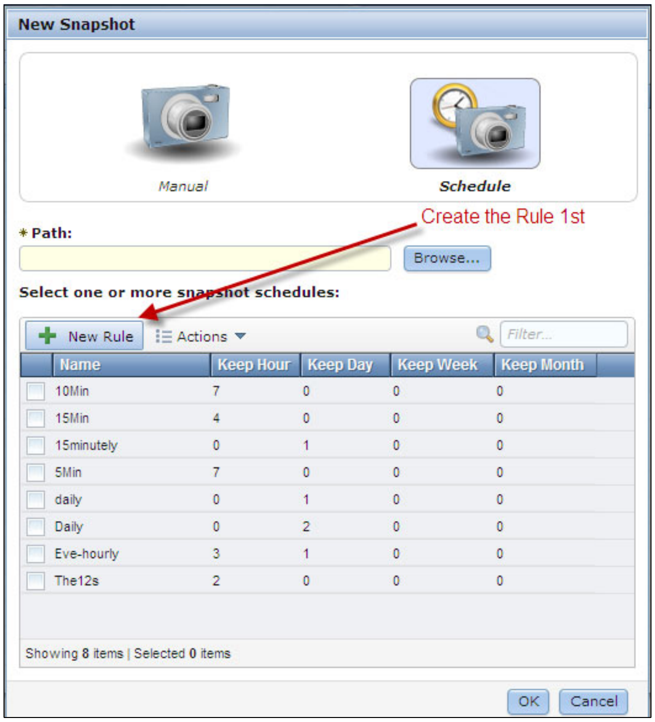
The snapshot version is recoverable if the snapshot has not been deleted. Snapshots take up space in the GPFS file system only when changes are made to the files, where the traditional backup version is stored outside of the SONAS.

Note: Do not rely only on snapshots for important backups. A snapshot might not survive file system corruption or a catastrophic hardware failure, possibly due to an external event.

Example of versioning with snapshots

A customer wants to have a daily backup. Each day's version of the file is kept for seven days. A weekly backup version is kept for eight weeks. After eight weeks, no backup versions are available.

To implement this scenario, you need to create some backup rules and associate the rules with an automatic snapshot policy. Figure 7-12 shows how to use the GUI to create a new rule.



* Path:

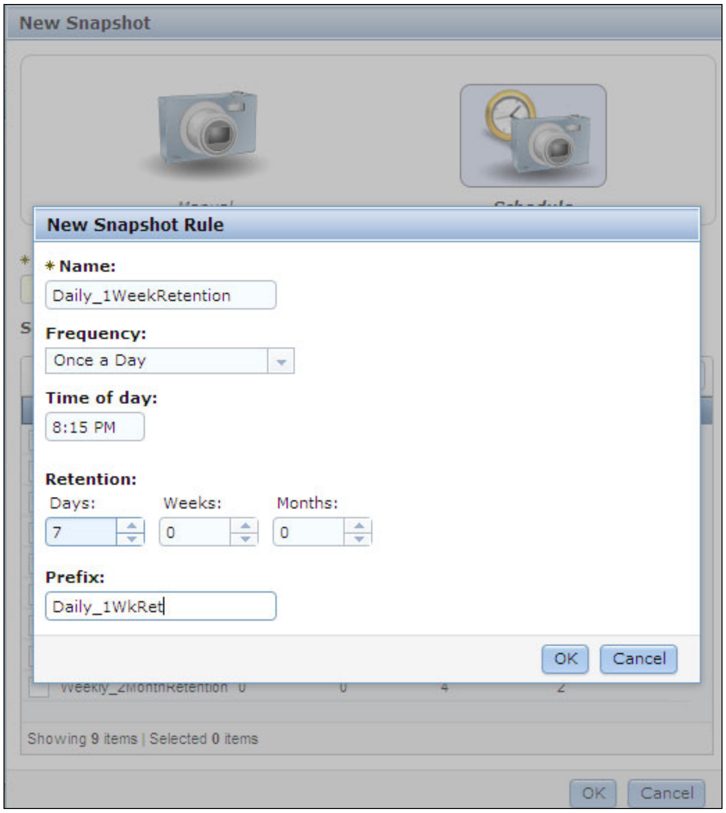
Select one or more snapshot schedules:

	Name	Keep Hour	Keep Day	Keep Week	Keep Month
□	10Min	7	0	0	0
□	15Min	4	0	0	0
□	15minutely	0	1	0	0
□	5Min	7	0	0	0
□	daily	0	1	0	0
□	Daily	0	2	0	0
□	Eve-hourly	3	1	0	0
□	The12s	2	0	0	0

Showing **8** items | Selected **0** items

Figure 7-12 Snapshot Rule Creation Launch from GUI

This scenario requires one snapshot rule to keep daily snapshots for seven days (see Figure 7-13). You can call this rule a `Daily_w_1WeekRetention` rule.



New Snapshot

New Snapshot Rule

* Name: Daily_1WeekRetention

Frequency: Once a Day

Time of day: 8:15 PM

Retention:

Days:	Weeks:	Months:
7	0	0

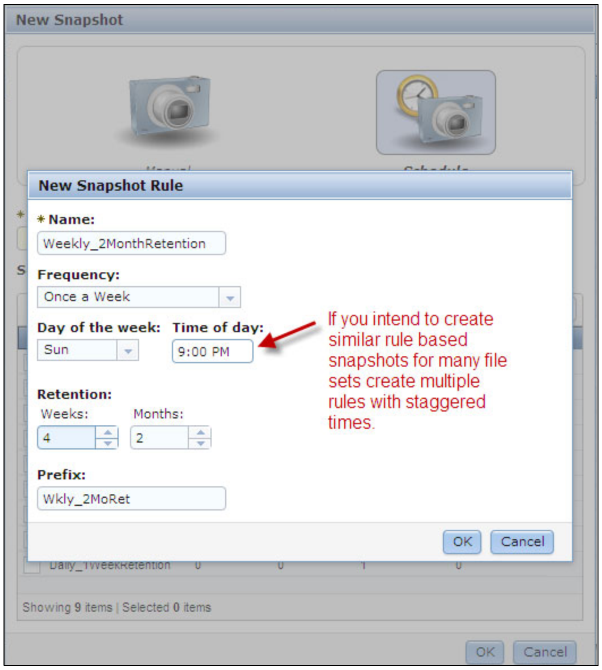
Prefix: Daily_1WkRet

Weekly_2MonthRetention 0 0 4 2

Showing 9 items | Selected 0 items

Figure 7-13 Daily snapshot with one-week retention from the GUI

It also requires a second snapshot rule to capture weekly snapshots with two months of retention (see Figure 7-14). This rule can be called `Weekly_w_2MonthRetention`. You can then create two automated snapshots, one with each rule associated, for each file system you want to have these snapshots.



* Name: Weekly_2MonthRetention

Frequency: Once a Week

Day of the week: Sun Time of day: 9:00 PM

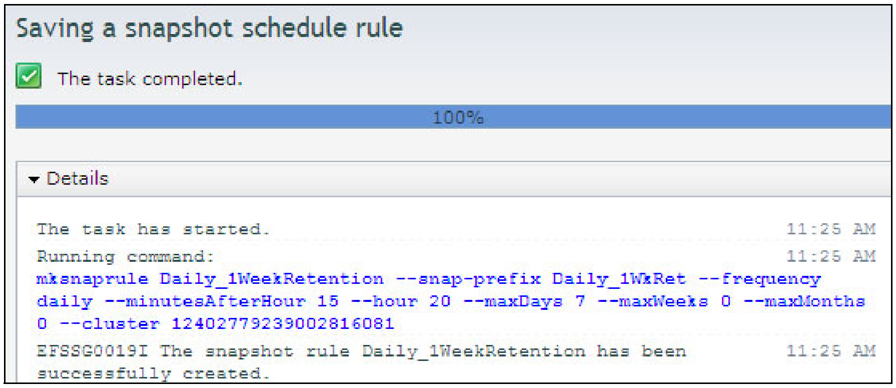
Retention:
Weeks: 4 Months: 2

Prefix: Wkly_2MoRet

If you intend to create similar rule based snapshots for many file sets create multiple rules with staggered times.

Figure 7-14   Weekly snap with two-month retention from GUI

When the snapshot is saved, the CLI command structure that is used to create the rule is displayed (see Figure 7-15).



Saving a snapshot schedule rule

☑ The task completed.

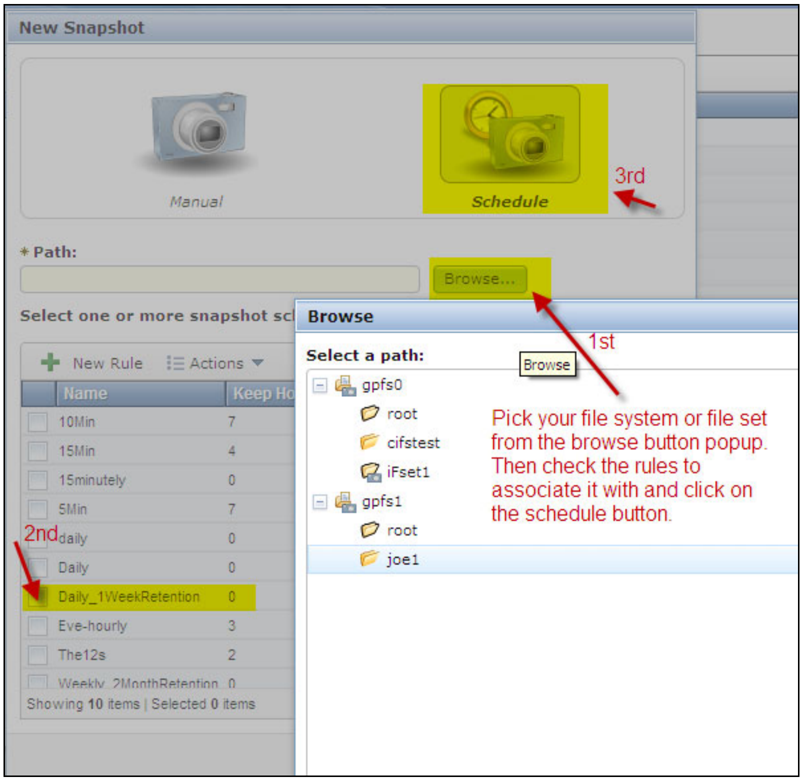
100%

▾ Details

```
The task has started.                                              11:25 AM
Running command:                                                   11:25 AM
mksnaprule Daily_1WeekRetention --snap-prefix Daily_1WkRet --frequency
daily --minutesAfterHour 15 --hour 20 --maxDays 7 --maxWeeks 0 --maxMonths
0 --cluster 12402779239002816081
EFSSG0019I The snapshot rule Daily_1WeekRetention has been         11:25 AM
successfully created.
```

Figure 7-15   CLI for creating a snapshot rule that is associated to scheduled snapshots

When the rule is created, the snapshot can be scheduled and associated with that rule for any file system or independent file set as shown in Figure 7-16.



* Path:

Select one or more snapshot sc

	Name	Keep Ho
□	10Min	7
□	15Min	4
□	15minutely	0
□	5Min	7
□	daily	0
□	Daily	0
□	Daily_1WeekRetention	0
□	Eve-hourly	3
□	The12s	2
□	Weekly_2MonthRetention	0

Showing **10** items | Selected **0** items

Select a path:

Pick your file system or file set from the browse button popup. Then check the rules to associate it with and click on the schedule button.

Figure 7-16 Snapshot with Auto Retention cleanup creation GUI

7.4.3 Snapshot rule preferred practices

Snapshot creation is instantaneous, but snapshot deletion requires a GPFS scan of the entire file system. Because snapshot rules instruct when the snapshot is to be created, tying too many automated snapshots to a single rule can create a GPFS burden at that time in the schedule if too many file systems and file sets use the same rule. There are a few preferred practices to consider.

Effect on capacity

For preferred practice, consider the effect on capacity for snapshots and the effect on performance during snapshot deletions.

SONAS uses space efficient snapshots and only writes physical copies on changed files, but these snapshots apply to all file system or file set data. So the effect can be a tremendous use of capacity if frequent changes are stored long term. At some point, having too many snapshots makes it more difficult to manage what versions had what changes, and its practical value diminishes greatly.

For that reason, it is the preferred practice to keep snapshots only on an as-necessary basis. This is probably a value of less than 50 or 60 snapshots per file set or file system.

Stagger snapshot creation times

It is also preferred practice to stagger your snapshot creation times. If you have 400 file sets, it is not practical or advisable to associate a single `Daily` snapshot rule to all file sets. This configuration not only creates the automatic snapshots to launch at the exact same time for all file sets, but also creates a retention cleanup operation that happens at the same time for each file set.

Remember that a snapshot delete requires a GFPS file system or file set scan, and each file set snapshot does this process independently. You might create a large effect by scanning 400 file sets all at the same time, or with the same time frame. Therefore, *keep as many snapshots as necessary but as few as possible*, and coordinate the time of snapshot creation to stagger the burden of snapshot deletion schedules.

Stagger replication timing

Another preferred practice is to stagger the replication timing from the snapshot schedules and peak client activity hours in heavy workload environments. Because every replication includes an initial snapshot creation, a metadata scan, and incremental data transfer, it is important to monitor performance workloads during these events to understand their effect.

If normal workflow performance is degraded during this process, you might consider adjusting the start times or expanding the RPO objectives to reduce the effect to production data workflow until you can add resources to the cluster where they are needed.

7.4.4 Snapshot configuration example

Here is an example of how to configure a client's snapshot schedules for preferred practices:

- No snapshots are scheduled for the file system itself.
- All independent file sets (including the file system `root` file set) are on an automatic snapshot schedule for semi-daily snapshots, weekly snapshots, and monthly snapshots.

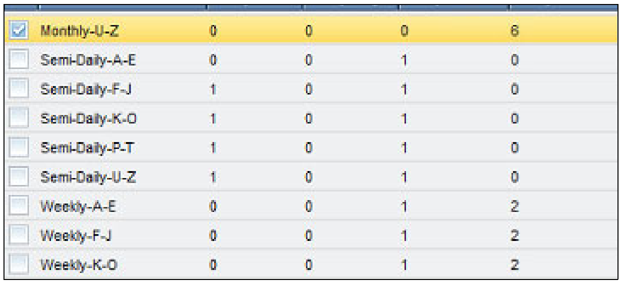
> **Note:** The file system `root` file set snapshot captures all file system and dependent file set data, but not the underlying independent file sets.

- The snapshot rule that is applied to each file set is assigned alphabetically to ensure that not all file sets have applied snapshots or snapshot deletions at the same time. An example is shown in Figure 7-17 on page 250.

  This configuration enables the schedules to stagger and distribute the heavy metadata scan requirements that are associated with snapshots at varying times in the daily schedule.
- A file set, whose name begins with the letters A - E, is assigned a semi-daily, weekly, or monthly snapshot rule that ends in A - E, and so on.
- The semi-daily snapshot takes a snapshot at roughly 5 or 6 am and 5 or 6 pm daily, and these snapshots are naturally and automatically deleted after one week of age.
- The weekly snapshot takes a snapshot at roughly 9 pm on a weekend, and these snapshots are naturally and automatically deleted after five weeks of age.
- The monthly snapshot takes a snapshot at roughly 1 am on a weekend and these snapshots are naturally and automatically deleted after six months of age.

This configuration ensures adequate protection of file set data (restorable by clients), and yet limits the number of active snapshots to roughly 25 for any file set, while mixing up the activation time and distributing the metadata scan burden of deletion schedules.

Figure 7-17 shows snapshot rules assigned alphabetically.



☑	Monthly-U-Z	0	0	0	6
□	Semi-Daily-A-E	0	0	1	0
□	Semi-Daily-F-J	1	0	1	0
□	Semi-Daily-K-O	1	0	1	0
□	Semi-Daily-P-T	1	0	1	0
□	Semi-Daily-U-Z	1	0	1	0
□	Weekly-A-E	0	0	1	2
□	Weekly-F-J	0	0	1	2
□	Weekly-K-O	0	0	1	2

Figure 7-17 Snapshot rules with alphabetical separation

7.5 File clones

Clones are writable, point-in-time, space-efficient copies of individual files. Here are two related use cases.

One use case is for provisioning virtual machines by creating a file for each system by cloning a common base image.

The second use case is for cloning the virtual disk image of an individual system as part of creating a snapshot of the system state. You can clone a file, not a directory, file set, or file system. However, you can snapshot independent file sets and file systems. This feature is a special purpose feature.

7.5.1 Managing clone files

Use the **`mkclone`** CLI command to create a clone, to create multiple clones of the same file with no additional space required, and to create clones of clones. A clone parent file is a read-only copy of the source file that is being cloned and shares disk space for the data that is common with the original source file. The owning user of the original source file can modify the original file without affecting the clones.

As blocks are written to either file, pointers to the new data blocks are created. With cloning, two files that are 90% identical require 55% of the space that is required if the source file is instead copied to a non-clone file.

A parent file for the clone must be designated or created when the clone is created. This clone parent file is permanently immutable, and can either be designated to be the source file, or it can be a unique new file.

Restriction: When a file, even if it is a source file, is designated as the parent file, the source file becomes an immutable parent file, and *can never be modified* from that point forward. An attempt to modify a clone parent results in an `access denied` return code. When a clone parent becomes immutable, it cannot be reverted to being modifiable. It can only be deleted, and then only if it has no clone child.

Creating a clone file

To create a clone file, a CLI user that has the SecurityAdmin or DataAccess role authorization can use the **mkclone** CLI command and specify the source file and target file with the -s or --source and -t or --target options. You can optionally use the -p or --cloneParentFile option to specify a clone parent file that is separate from the source file; if these options are omitted, the source file becomes immutable due to its default designation as the clone parent.

Creating a clone example

The following example creates a separate parent file named someFileParent:

```
$ mkclone -s someFile -t someFileClone -p someFileParent
```

See the man page for more information:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/manpages/mkclone.html

The someFileParent file is created as an immutable clone parent file and cannot be modified. The file content of someFile and someFileClone are stored in the someFileParent file. The someFile and someFileClone files can be modified. However, the someFileParent file is immutable.

A file can have only one immutable parent. You cannot have two immutable clone parents of the same file clone, so the preferred practice is to check the status and clone parent information before creating your clone.

Before running the **mkclone** command, you must run **lsclone *<file-name>*** first to determine whether the source *<file-name>* is already a clone.

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 has a depth value of 2, and so on.

7.6 Management node configuration backup tasks

This section describes backing up the active management node.

The active management node can be backed up by using the **backupmanagementnode** CLI command. 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, use the **backupmanagementnode** CLI command. 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 section of the IBM Knowledge Center:

http://pic.dhe.ibm.com/infocenter/sonasic/sonas1ic/index.jsp

The command creates an archive that includes all the components. To specify components to back up, you must use the --component option.

Only the active management node is backed up with the **backupmanagementnode** CLI command.

> **Important:** The **backupmanagementnode** CLI command does not back up file systems, file system settings, exports, file system pools, file sets, snapshots, or quotas.

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, enter the command that is shown in Figure 7-18.

```
# backupmanagementnode --targethost strg001st001 --targetpath /persist/mgmtbackup
```

Figure 7-18 Backupmanagementnode command example

To back up to a USB flash drive, enter the command that is shown in Figure 7-19.

```
# backupmanagementnode --mount /media/usb
```

Figure 7-19 Example of using the backupmanagementnode command to back up to a USB flash drive

> **Tip:** The **backupmanagementnode** command makes a backup from the local management node, where the command is running, and stores it on another host or server. This command should be used only in configurations where there is a single dedicated management node. For Storwize V7000 Unified and for dual management nodes the backup is automatic, and the command should never be run from the command line.

7.7 External log service

This section shows how to create a system log server definition with the CLI.

You can use the **mksyslogserver** CLI command to create a new system log server configuration for an IBM SONAS system. You can configure at most six external servers that can receive system log (syslog) messages from an IBM SONAS clustered system, using the BSD syslog protocol as defined by Internet Engineering Task Force (IETF) Request for Comments (RFC) 3164.

7.7.1 Log server overview

The **mksyslogserver** CLI command creates a new system log server definition for an IBM SONAS system.

The command syntax is simple and self-explanatory, and log traffic can be secured by firewall rules to isolate an IP and port number. External logging is considered a preferred practice for SONAS and Storwize V7000 Unified.

The syntax in Figure 7-20 is used for the **mksyslogserver** command.

```
mksyslogserver syslogServerName [--ip ipAddress] [--port portNumber] [-c {system ID |
system name}]
```

Figure 7-20 CLI for initiating external logging

7.7.2 Retrieving system log server information by using the CLI

You can use the **lssyslogserver** CLI command to display information about the system log servers of an IBM SONAS system.

Overview

The **lssyslogserver** CLI command retrieves information about the system log servers of an IBM SONAS system from the database and can optionally display the information in colon-delimited format.

The command syntax is simple, and it is good practice to validate that the external log server is collecting logs in your daily system status check.

The command in Figure 7-21 can be submitted without any option to list the system log servers of the SONAS system.

```
lssyslogserver [syslogServerName] [-c {system ID | system name}] [-Y]
```

Figure 7-21 Listing log server information

7.8 Antivirus

With today’s continuing explosive growth in data comes the need for storing the data without compromising data integrity from potential threats that might exist in an enterprise network environment. SONAS is qualified for interoperability with two leading antivirus scan engines: Symantec AntiVirus (SAV) for NAS and McAfee VirusScan Enterprise.

The data that is created or accessed with Network File System (NFS) or Common Internet File System (CIFS) is vulnerable to the potential threats of viruses, worms, Trojan horses, and other forms of malware. Computer viruses mostly target Microsoft operating systems. However, computers that are running other operating systems can be directly or indirectly affected by viruses.

The antivirus connector is a part of the SONAS management software, which communicates with vendor scan engines by using Internet Content Adaptation Protocol (ICAP). There are two main approaches for virus scanning:

- On-access scan. It scans all of the specified files on SONAS when accessed or created. This method has the benefit of ensuring that the files are scanned with the latest virus signature before they are accessed. This approach is more effective at detecting viruses before they are able to compromise data, and this method does not generate heavy network traffic between SONAS and SAV scan engines. This approach is ideal for customers who are using Windows clients and CIFS file I/O.
- Bulk scan. This method enables scanning of all of the specified files on a file system, or a part of file system. This scan is typically performed on the schedule that is defined on the SONAS system.

  The disadvantage in using this method is that the files that are recently updated might not be scanned before being used. Bulk scans can generate heavy network traffic between SONAS and scan engines, and can generate heavy load on a storage system. Also, bulk scans can take significant time to complete, depending on the number of files to be scanned. Storage administrators are likely to use the bulk scans for non-CIFS files (for example, NFS) protection, which are less prone to virus attacks.

When users access a file from SONAS over the network, SONAS initiates the scan of a file in real time, opens a connection with the scan engine, and then passes the file to the scan engine for scanning. The scan engine indicates the scanning results to SONAS after the file is scanned. If the file is infected, the scan engine tries to repair the file, and sends the repaired file to SONAS.

If the file is infected and can be cleaned, a stored version of the infected file is replaced on SONAS with the repaired file that is received from the scan engine. Only the repaired file is passed to the requesting user. If a virus is detected and repair of the file is not possible, SONAS can be configured to quarantine or delete the non-repairable file, and notify the user with an error message that indicates permission is denied.

7.8.1 Integration considerations

The following factors need to be carefully considered before you integrate IBM SONAS with Symantec AntiVirus.

Numbers of Symantec scan engines

Antivirus scanning on SONAS requires a minimum of one scan engine that is configured with Symantec AntiVirus for NAS. However, to take full benefit of load balancing and the high availability feature of SONAS, use a minimum of two scan engines.

SONAS antivirus connector automatically performs load balancing to make sure that the workload is evenly distributed across the scan engines. When a scan engine becomes unavailable, the workload is directed to the remaining operational scan engines. The considerations that are listed here affect the number of scan engines that might be required:

- Total number of files that are stored on the SONAS requiring scanning

  Large numbers of files can be scanned by multiple scan engines using the SONAS antivirus connector load balancing feature.

- Host processor speed and memory configuration

  Fewer scan engines might be needed if processor speeds are faster and more memory is present on each scan engine.

- Network speed

  Faster network speeds support reduced time in transferring larger files to the scan engine for scanning.

Figure 7-22 shows the workflow for antivirus software that is running on SONAS.

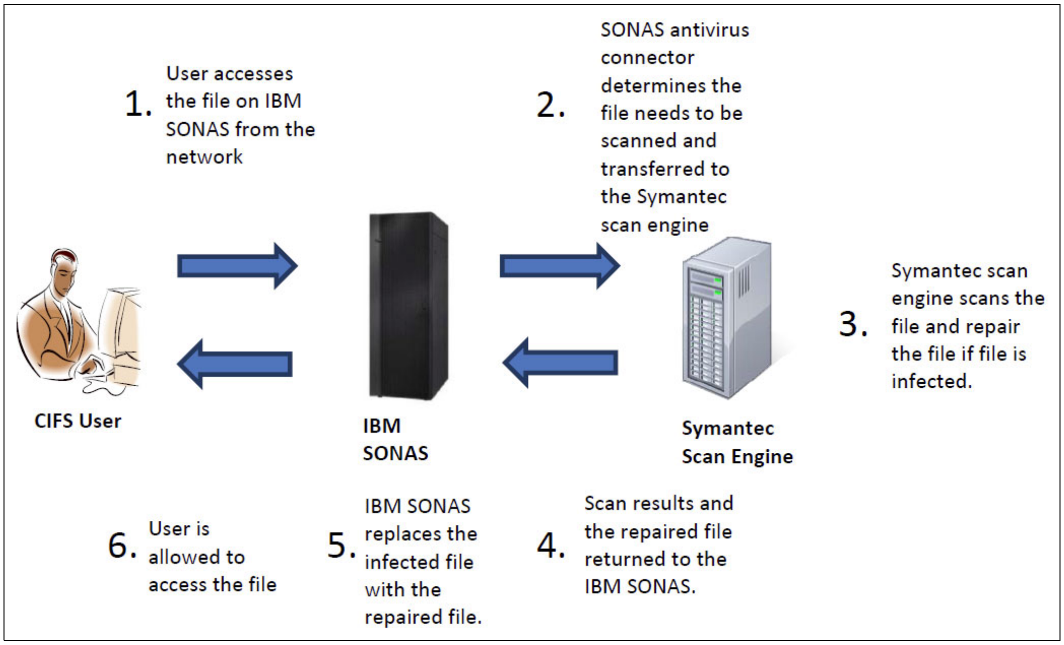

1. User accesses the file on IBM SONAS from the network
2. SONAS antivirus connector determines the file needs to be scanned and transferred to the Symantec scan engine
3. Symantec scan engine scans the file and repair the file if file is infected.
4. Scan results and the repaired file returned to the IBM SONAS.
5. IBM SONAS replaces the infected file with the repaired file.
6. User is allowed to access the file
CIFS User
IBM SONAS
Symantec Scan Engine

Figure 7-22 Workflow for antivirus software

7.8.2 Defining files to be scanned

In SONAS, the administrator can define which files or file types are to be scanned. The administrator can control and decide whether to scan files by exclusion list or inclusion list, or whether to scan all the files regardless of extensions.

The inclusion/exclusion list defines the following behavior:

- If the include list is empty or not defined, the default is that all extensions are included in the scan.

  An excluded list is created to exclude files with specific file extensions from scanning by the Symantec scan engine.
- If an extension is in the include list, only files with that extension are scanned.
- If an extension is in the include and exclude lists, files with that extension are not scanned.

Careful planning is required to create the include and exclude list, because they play an important role in improving performance of the scan process. Not all file extensions need to be scanned due to the nature of the files and file types, which are unlikely to have viruses.

It is important to plan for the action that needs to be taken in case an unrecoverable virus file is identified. IBM SONAS provides the option to quarantine or delete the infected, unrecoverable file.

7.8.3 Defining scan engine pool

At least one scan engine must be registered to provide virus scanning for each SONAS. However, configure a minimum two scan engines in a scan engine pool to use the load-balancing facility that is provided by SONAS to distribute the scan load. Also, it provides the high-availability feature in case one scan engine is not available. SONAS tries to contact the failed scan engine periodically and reinstate it for scanning after it becomes available.

Every time a new antivirus definition file is downloaded by the scan engine, all files that are defined within all scopes must be rescanned before access. The bulk scan feature is a method to proactively scan all of those files during a window when access to the SONAS is at a minimum, reducing the load on the system and network during peak usage times.

7.8.4 Antivirus preferred practices

Antivirus scanning, particularly bulk scanning of large files, can add significant load to several SONAS system resources, and can cause performance bottlenecks. The following guidelines can help you minimize performance effect to the system:

- If on-access or bulk scan produces timeout errors, consider increasing the timeout value of scans by using the `--timeout` option of the **`cfgav`** command. Do not increase the timeout parameter beyond the CIFS client timeout value, which can cause files to become inaccessible to the user.
- Avoid scanning "expensive" items (such as scanning inside the HSM archived files or other containers) to avoid timeout issues.
- Avoid scanning files that are managed on tape (HSM) pools.
- Depending on the scanning performance requirements, the number of interface nodes on which bulk scans are run can be configured by using the `--nodes` option of the **`ctlavbulk`** command. If higher scanning performance is wanted, consider running scans on more interface nodes. To reduce effect to other SONAS resources, consider limiting the number of interface nodes on which bulk scans are run.
- Carefully select file types for scanning. Certain classes of large files are less likely to be prone to virus attacks. By unconfiguring certain types of files by using the `--add-include|--rem-include|--set-include|--set-exclude` options of the **`cfgav`** command, overall antivirus scanning performance can be greatly improved.
- Give similar consideration to selecting scopes for scanning, because some scopes might contain files that will not be accessed, and they are not likely to be prone to the virus attacks.
- Ensure that the storage system has adequate capacity for the client and scan traffic. On-access scans are less likely to add significant load to the storage backend because it is typically scanning data that has either just been written or is just about to be read by the client and therefore can take advantage of caching. Bulk scans can add significant load to the storage back-end.
- After updating the antivirus signature, scan all protected files during off-peak hours to minimize the effect of scanning during peak usage.
- Ensure that the network infrastructure, such as routers, switches, and network cards on both SONAS and scan engines, has adequate capacity. In a typical configuration, use 10-Gigabit Ethernet.
- Use a minimum of two scan engines to use high-availability and load-balancing features for the scanning.
- Ensure that scan nodes have adequate processor and disk performance.
- Run bulk scan after a migration either by Hierarchical Storage Management (HSM) recall or data restoration from the backup server.

- When using multiple scan engines to support scanning of IBM SONAS, consider the following factors:
  - Configure the setting on each scan engine to be identical.
  - Schedule an automatic update of all SAV scan engines to occur at the same time to ensure that virus definitions are identical.
  - Configure virus scan functionality for each identical SONAS system that uses a particular scan engine to avoid inconsistency.

7.8.5 AntiVirus Summary

The ability to effectively protect shared file data against viruses, file loss, and other malicious threats is an important challenge for storage and security administrators who require a trusted and reliable antivirus solution. Not only must the integrity of the data be constantly maintained, the solution must also be scalable to match the continually expanding size and volume of data that is retained on an NAS system.

For specific details about how to configure and manage antivirus services and management details, read about both SONAS supported antivirus solutions in the IBM Knowledge Center:

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

7.9 Failover, failback, and disaster recovery

This section describes business continuance if a primary SONAS site is destroyed.

7.9.1 Disaster recovery

Disaster recovery (DR) is only supported for file systems that are correctly configured in asynchronous replication today. This section describes a process for DR with SONAS. It requires correct authentication of the primary and secondary site to the same directory service, and that those services have the correct authentication extensions for replication.

The target file systems must have adequate spindle and interface node configurations to support the capacity and performance that are required of the user base and all associated data.

The target site also needs adequate networking, access to active directory or LDAP services, NTP, and DNS services to serve the user community at large.

Asynchronous replication disaster recovery

Recovering a file system by using asynchronous replication requires that a replication relationship from the original target site to the original source site is configured and started.

After the source (Site A) site fails, you must set the target site (Site B) as the new source site, replicating back to Site A.

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.

Start the replication by using the **`startrepl`** CLI command, specifying the `--fullsync` option.

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 perform a final replication to catch up Site A. These incremental replications should *not* use the `--fullsync` option.

When data is verified as being replicated accurately to Site A, Site A can be reconfigured as the primary site. Remove any replication tasks that are going from Site B to Site A by using the **rmtask** CLI command.

7.9.2 Disaster recovery testing and high-level process overview

In most cases, do failover testing with sample data before going into production to allow the client to accurately document and test failover capturing all the specific edits and details to Shares, DNS, and so on, as it applies to them.

The following is a summary of the tasks that are typically required to test failover. Prepare for failover before applying the configuration to production. If the cluster is already in production, a small test file system can be assembled to test this configuration for a non-production file system before applying the failover script to a production file system.

Failover plan overview

A high-level plan includes the following tasks:

- Verify that replication is working and that the last replication was successful.
- Stop the automatic replication job, which runs every ½ hour (or as assigned).
- Initiate replication (primary to DR) manually and monitor progress.
- Inform users and application owners as soon as replication is completed.

  Sample Message: `Planned Service Outage: There will be a minor disruption in access. The NAS System is failing over to the disaster recovery site. It will be available in a few minutes. Stand by for an update.`
- Update or change exports and shares to read-only (RO) mode at the primary site, then initiate another replication (Primary to DR) manually and monitor progress.
- Run the **chexports** command as needed to make them RO at the primary site.
- Run the **lsexports** command to verify exports that are on this file system are RO and get a name list.
- To the extent possible, stop all file access:
  - If you have root access, use the **smbstatus** command to get process IDs (PIDs) of users that are actively connected to the previously mentioned export name list for this file system and stop processes that are associated with those PIDs on interface nodes.
  - Use equivalent commands like **showmount** to stop NFS and HTTP access.
- Run the **rmrepltarget** command on the secondary site to ensure that the target is not over written if the primary comes back online.
- Change the Domain Name Server (DNS) entry for the application with the collaboration of the network engineering team:
  - Rename the CName/alias from `PrimarySONAS` to `SecondarySONAS`.
  - Rename the CName/alias from `SecondarySONAS` to `PrimarySONAS`.
- Verify that replication is completed. Change exports and shares to read/write (RW) mode at the DR site.
- Inform users and app owners as soon as exports and shares are updated to RW mode.

- If a problem is discovered, open an incident ticket and follow the troubleshooting steps.
- When service is validated, Start automatic replication from the DR to the primary system. Schedule the replication job to start automatic replication.

Figure 7-23 is a diagram of a disaster recovery operation with SONAS replication.

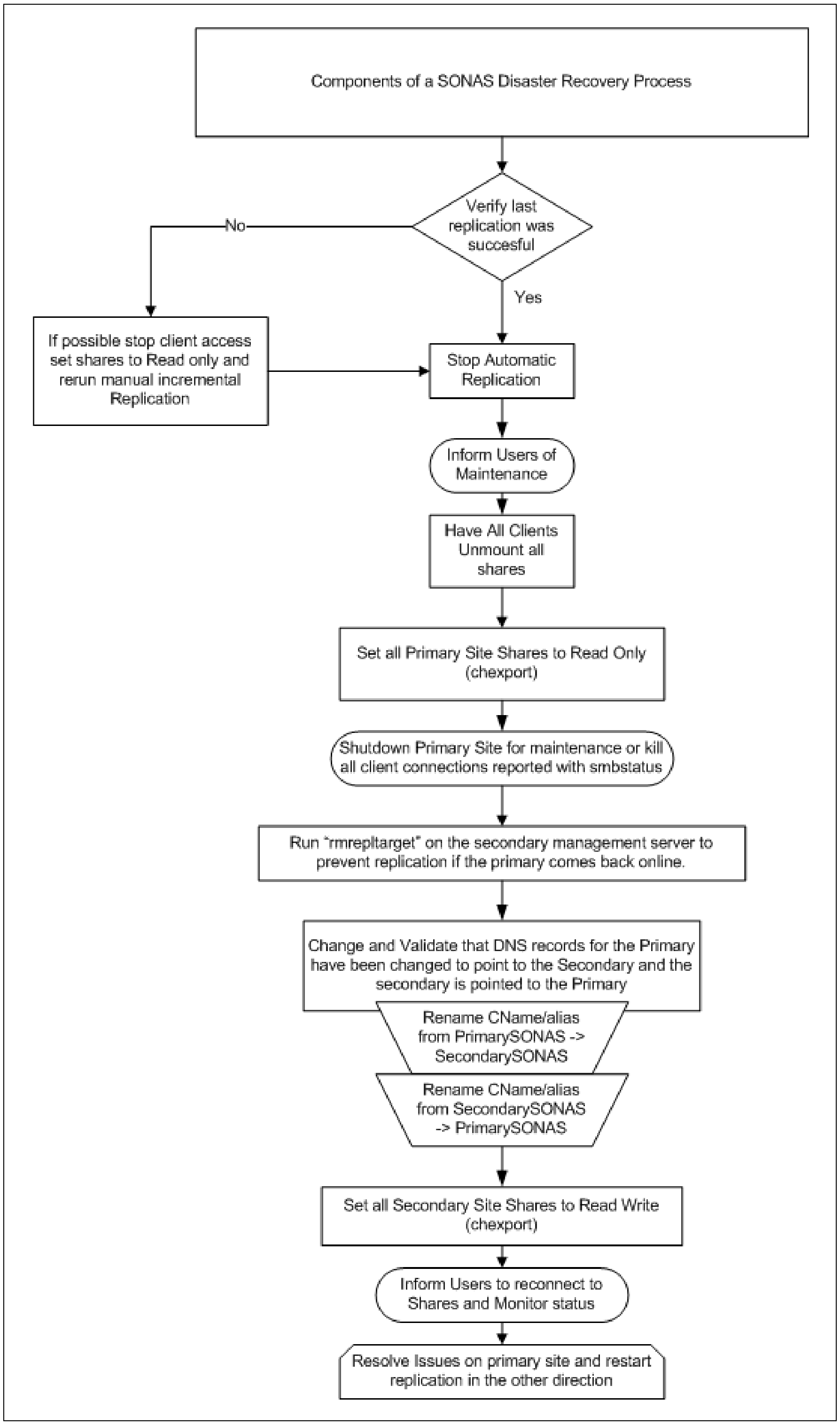

Components of a SONAS Disaster Recovery Process
Verify last replication was succesful
No
Yes
If possible stop client access set shares to Read only and rerun manual incremental Replication
Stop Automatic Replication
Inform Users of Maintenance
Have All Clients Unmount all shares
Set all Primary Site Shares to Read Only (chexport)
Shutdown Primary Site for maintenance or kill all client connections reported with smbstatus
Run "rmrepltarget" on the secondary management server to prevent replication if the primary comes back online.
Change and Validate that DNS records for the Primary have been changed to point to the Secondary and the secondary is pointed to the Primary
Rename CName/alias from PrimarySONAS -> SecondarySONAS
Rename CName/alias from SecondarySONAS -> PrimarySONAS
Set all Secondary Site Shares to Read Write (chexport)
Inform Users to reconnect to Shares and Monitor status
Resolve Issues on primary site and restart replication in the other direction

Figure 7-23 Flow diagram of a disaster recovery operation with SONAS replication

8

Monitoring

This chapter provides preferred practices for managing day-to-day monitoring of IBM Scale Out Network Attached Storage (SONAS) cluster resources by using internal tools, the graphical user interface (GUI), and the command-line interface (CLI). You can monitor components outside of the SONAS appliance, such as Fibre Channel Switches, zone configurations, and gateway storage devices.

These should be managed separately by using documentation and tools that are appropriate and specific to that device. For more concise back-end storage device monitoring tools, review the storage brand product administration guides.

If you are new to your SONAS or IBM Storwize V7000 Unified solution, keep this chapter and use the Daily, Weekly, and Monthly check lists at the end of this chapter as a guide for walking through health checks. As you become more proficient, you will find preferred methods of validating system health and readiness.

As a component of preferred practices, avoid becoming too comfortable with learned processes, and continue to challenge yourself to learn more about system health, logs, and code changes on all of your systems from end-to-end. Provide your suggestions to IBM through your account and technical resources so that IBM can continue to shape products to your needs.

This chapter contains the following information:

- Monitoring daily work
- Weekly tasks to add to daily monitoring tasks
- Monthly checks for trends, growth planning, and maintenance review
- Monitoring with IBM Tivoli Storage Productivity Center

8.1 Monitoring daily work

This section provides step-by-step monitoring suggestions to add to your daily, weekly, or monthly activity routines. It is a high-level consideration of monitoring tools that is intended to help you understand how to make sure that everything is optimally functioning, or point out what might need deeper investigation. It highlights relevant monitoring short cuts and tips along the way, to help you quickly assess your clusters' status.

For more information, see the *IBM SONAS Implementation Guide*, SG24-7962 and *Scale Out Network Attached Storage Monitoring*, SG24-8207 IBM Redbooks publications.

8.1.1 Daily checks

The following list shows the health checks you should make daily:

- Cluster health
- Cluster node health
- Cluster services health
- File system capacity
- File system and independent file set iNode capacity
- Storage pool capacity
- Backup job status
- Replication task status
- Events list review
- User quota checks

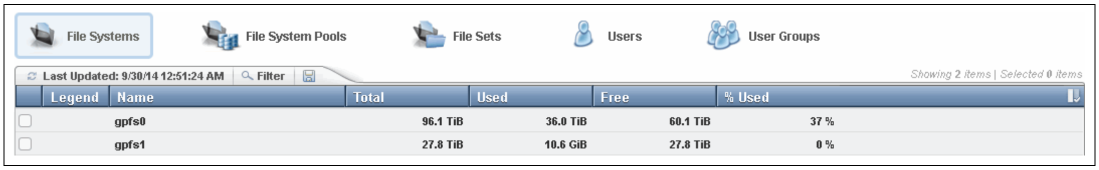
8.1.2 Health and system component check

See the daily monitoring chapter in the *Scale Out Network Attached Storage Monitoring*, SG24-8207 IBM Redbooks publication for step-by-step information about the daily activity routine for health and system component check.

File system capacity

When you run out of file system capacity, the only way to mitigate the situation is to delete data, or add Network Shared Disks (NSDs) to the file system.

Figure 8-1 shows how to run this check from the GUI.


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

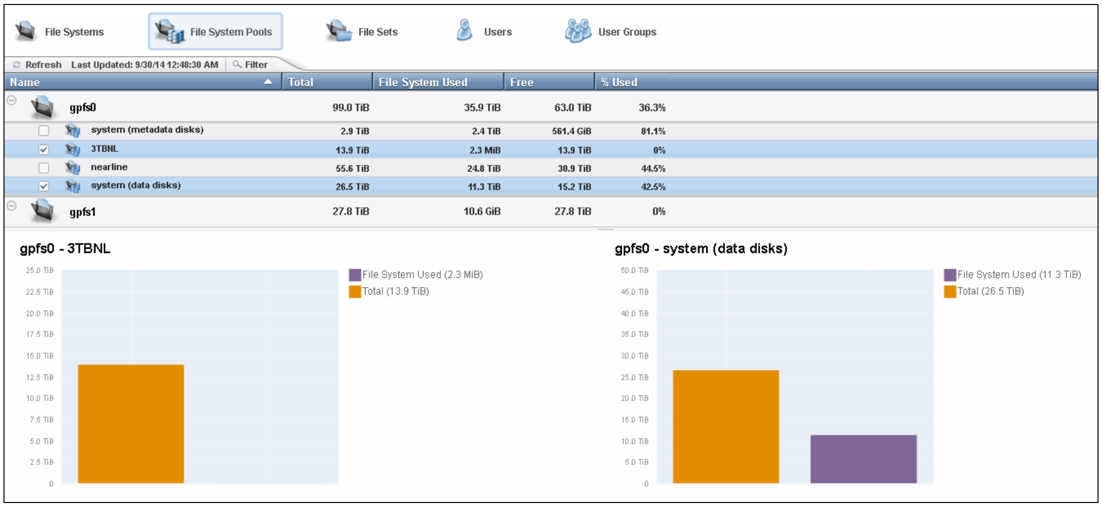
	Legend	Name	Total	Used	Free	% Used
□		gpfs0	96.1 TiB	36.0 TiB	60.1 TiB	37 %
□		gpfs1	27.8 TiB	10.6 GiB	27.8 TiB	0 %

Figure 8-1 GUI showing file system capacity (from the tab selection)

Storage pool capacity

When you run out of storage pool capacity, the only way to mitigate the situation is to migrate data to a different pool, or add NSDs to that specific storage pool.

Figure 8-2 shows how to monitor storage pool capacity from the GUI.



Last Updated: 9/30/14 12:48:30 AM

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 8-2 Monitoring storage pool capacity

Figure 8-3 shows how to monitor storage pool capacity from the CLI.

```
[root@xivsonas.mgmt001st001 ~]# lspool -d gpfs0 -r
EFSSG0015I Refreshing data.
Filesystem Name   Size     Usage Available fragments Available blocks Disk list
gpfs0       system 53.23 TB 0%    2.32 MB            53.23 TB
DCS3700_360080e50002ea78c000014fb521cdbeb;DCS3700_360080e50002ea78c000014fd521cdc32;DCS3700_360080e50002ee7bc00001
304521cdeaa;DCS3700_360080e50002ee7bc00001306521cdef4
EFSSG1000I The command completed successfully.
```

Figure 8-3 Monitoring storage pool capacity from the CLI

Node file system capacity

When the root file systems on nodes in your cluster (or your Tivoli Storage Manager server) fill up, the product can become unstable. The only way to mitigate that situation is to find what is filling the file system and compress or remove it. However, this task does require some caution. File system capacity can be checked from the CLI.

From the root directory of the node, as the root user, run the **df -h** command, as shown in Figure 8-4.

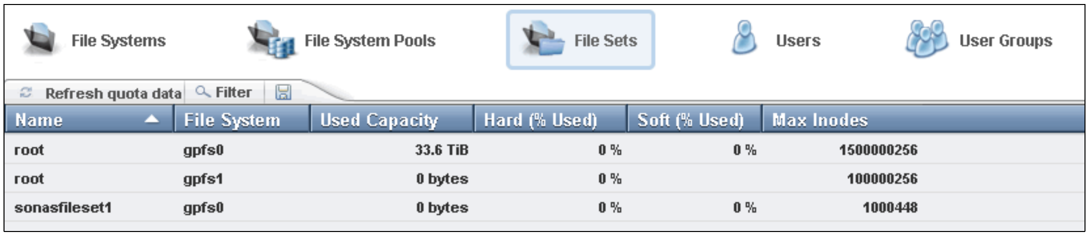
```
[root@xivsonas.mgmt001st001 ~]# df -h
Filesystem            Size  Used Avail Use% Mounted on
/dev/sdb2              19G   11G  7.2G  59% /
tmpfs                  24G  4.0K   24G   1% /dev/shm
/dev/sda1             276G  302M  261G   1% /ftdc
/dev/sdb1             9.1G  150M  8.5G   2% /persist
/dev/sdb5             103G  2.7G   95G   3% /var
/dev/sdb6             9.2G  158M  8.6G   2% /var/ctdb/persistent
/dev/gpfs0             54T  7.0G   54T   1% /ibm/gpfs0
```

Figure 8-4 Sample df -h output

Quotas

When a quota reaches a hard limit, the user cannot write to the quota-protected space unless other data is removed (to free capacity), or the quota is modified (to expand the limitation).

Figure 8-5 shows how to view the quota from the GUI.



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 8-5 Sortable quotas status view in the GUI

Figure 8-6 shows how to check the quota with the CLI. You can optionally use the `--force` option to prevent the display of the confirmation prompt. In the example, quotas are checked on the file system named `gpfs0`.

```
# chkquota gpfs0 --force The system displays information similar to the following
output:
gpfs0: Start quota check
  1 % complete on Wed Dec  1 11:25:09 2010
...
100 % complete on Wed Dec  1 11:26:41 2010
Finished scanning the inodes for gpfs0.
Merging results from scan.
```

Figure 8-6 Forcing quota checks with the CLI

Backup scratch pool tape availability

When the backup server runs out of scratch pool tapes, it fails backups when the last writable tape is full and there is no space left to write data.

The check for the number of available scratch tapes uses the **`RUN Q_SCRATCH`** command, as shown in Figure 8-7. Run this command only from the Tivoli Storage Manager server as the `dsadmin` user.

```
Scratch Tapes
Check for the number of available scratch tapes.

tsm: TSM1CLIENT> RUN Q_SCRATCH
“Query_Scratch” is a user-defined server command script.
```

Figure 8-7 Commands run from the Tivoli Storage Manager server to check quantity of scratch tapes

Backup job status

When backup jobs are failing, the system might not have data that can be efficiently restored from tools outside of the cluster. The only way to mitigate this risk is to determine the reason why backups are failing, close the gap, and catch up with a new incremental backup. Monitoring of backup task success begins with the GUI or the CLI command **`lsjobstatus`**.

To determine whether a Tivoli Storage Manager backup session is running, run the **`lsjobstatus -j backup -r`** command. See Example 8-1 on page 265.

Example 8-1 Example of the output for the lsjobstatus -j backup -r command

```
File system Job  Job id   Status   Start time            End time/Progress    RC
Message
gpfs0 backup   34       running   10/17/12 8:13:49 AM MST    backing up
776/602200 Errors=0 Expire=562/569
EFSSG1000I The command completed successfully
```

The **lsjobstatus** command shows the running and completed jobs for a file system, and the primary node where the job started. By default, only the running jobs are shown, but this display can be modified with the `--done` and `--all` options. The command lists the start and end times for the jobs. See Example 8-2.

Example 8-2 Sample lsjobstatus output with common options

```
[root@sonas1.mgmt001st001 ~]# lsjobstatus -j backup --all
File system Job    Job id Status  Start time              End time/Progress                                     RC
Message
gpfs0       backup 0      done(+) 2/3/12 4:50:30 PM EST    2/3/12 4:56:50 PM EST                                 0
EFSSG1000I
........
gpfs0       backup 854    done(-) 8/25/13 2:00:05 AM EDT   8/25/13 2:21:46 AM EDT                                1
Errors=51 Expire=7962/7962 Backup=559/965 EFSSA0797C Backup partially failed. Please check the logs for details.
gpfs0       backup 855    done(-) 8/26/13 2:00:03 AM EDT   8/26/13 2:27:53 AM EDT                                1
Errors=31 Expire=26/26 Backup=379/759 EFSSA0797C Backup partially failed. Please check the logs for details.
gpfs0       backup 859    done(-) 8/27/13 2:00:05 AM EDT   8/27/13 6:30:38 AM EDT                                1
Errors=32 Expire=27996/27996 Backup=52300/50442 EFSSA0797C Backup partially failed. Please check the logs for
details.
gpfs0       backup 860    running 8/28/13 2:00:05 AM EDT   Errors=6102 Expire=266610/266610 Backup=447529/452892
Errors=6102 Expire=266610/266610 Backup=447529/452892
EFSSG1000I The command completed successfully.
```

Check replication tasks

When replication fails, the changes from the last incremental backup get larger and larger as time goes by, extending the time that it takes to complete the corrective action and the next incremental asynchronous replication. This leaves your system more vulnerable to being inconsistent if you have a disaster recovery requirement. The only way to mitigate the situation is to repair the issues that are causing failure, and catch up with the next incremental replication.

To display the status of the selected asynchronous replication in the management GUI complete the following tasks:

1. Log on to the SONAS GUI.
2. Select **Copy Services** → **Replication**.
3. Right-click an **asynchronous replication**, and then select **Details**.

In the CLI, use the **lsrepl** command to display the status of asynchronous replications, as shown in Example 8-3.

Example 8-3 Sample lsrepl output display

```
$ lsrepl
Filesystem Log ID         Status   Description                                 Last Updated Time
pfs0       20120217023214 FINISHED 8/8 Asynchronous replication process finished   2/17/12 02:33 AM
gpfs0      20120221004341 FINISHED 8/8 Asynchronous replication process finished   2/21/12 00:45 AM
gpfs0      20120217023214 FINISHED 8/8 Asynchronous replication process finished   2/21/12 01:33 AM
gpfs0      20120217023116 FAILED   The replication was aborted due to a critical error.
 Please use '--clearlock' option the next time, to remove the lock file.     2/21/12 03:44 AM
gpfs0      20120221004341 FINISHED 8/8 Asynchronous replication process finished   2/21/12 04:45 AM
gpfs0      20120221004341 FINISHED The replication completed successfully through failover recovery of
node pairs.                       2/21/12 05:45 AM
EFSSG1000I The command completed successfully.
```

You can use the --status option to display the status of processes that are involved in the asynchronous replications. This option can be used to investigate a node name to be used in the **runreplrecover** command. By default, this option displays the number of active (running), available (not banned), and banned replication processes for each node, as shown in the output in Example 8-4.

Example 8-4 lsrepl sample output with the --status option

```
root@st001.mgmt001st001 ~]$ lsrepl gpfs0 --status
Filesystem: gpfs0
Log ID: 20120429064041
Source          Target          Active Procs    Available Procs Total Procs
int001st001     10.0.100.141    2               3               3
int002st001     10.0.100.143    3               3               3
```

You can use the --process option to display the status of active (running), available (not banned), and banned replication processes for each node. See Example 8-5.

Example 8-5 lsrepl sample output with the --process and the --status option

```
[root@st001.mgmt001st001 ~]$ lsrepl gpfs0 —-status --process
Filesystem: gpfs0
Log ID: 20120429064041
Index   Source          Target          Repl Status     Health Status
1       int002st001     10.0.100.143    active          available
2       int002st001     10.0.100.143    active          available
3       int002st001     10.0.100.143    active          available
4       int001st001     10.0.100.141    inactive        available
5       int001st001     10.0.100.141    active          available
6       int001st001     10.0.100.141    active available
```

You can use the --progress option to display information about the transferred size, progress, transfer rate, elapsed time, remaining size, and remaining time for each process running in the asynchronous replication. See Example 8-6.

Example 8-6 lsrepl sample output with the --progress option

```
   [root@st001.mgmt001st001 ~]$ lsrepl gpfs0 --progress
   Filesystem: gpfs0
   Log ID: 20120429064041
   Mon Apr 16 07:56:46 CEST 2012
    interval 1 sec, remain loop: 26
    display rsync progress information
   ===============================================================================
   =

   PROC #: NODE-PAIR <HEALTH_STATUS>
   FILE-PATH
   FILE:              XFER-SIZE(TOTAL)          PROG(%)   XFER-RATE    ELAPSED
   REMAIN(TIME)
   -------------------------------------------------------------------------------
   -
   Proc 1: int002st001->10.0.100.144 <available>
   dir/file3          65,536,000(500.00MB)      12.50%    10.79MB/s    0:00:07
   437.50MB(0:00:41)
   Proc 2: int001st001->10.0.100.143 <available>
   dir/file4          98,435,072(500.00MB)      18.77%     7.16MB/s    0:00:10
```

```
406.12MB(0:00:58)
Proc 3: int003st001->10.0.100.145 <available>
dir/file5         75,202,560(500.00MB)      14.34%     6.51MB/s    0:00:08
428.28MB(0:01:07)
Proc 4: mgmt002st001->10.0.100.141 <available>
dir/file1         43,548,672(500.00MB)       8.31%     6.74MB/s    0:00:06
458.46MB(0:01:09)
Proc 5: mgmt001st001->10.0.100.142 <available>
dir/file2        115,736,576(500.00MB)      22.07%     9.50MB/s    0:00:13
389.62MB(0:00:42)
--------------------------------------------------------------------------------
-
Overall Progress Information: 0 of 8 files comp
XFER-SIZE(TOTAL)    PROG(%)    XFER-RATE    ELAPSED    REMAIN(TIME)
80MB(  2.45GB)      15.09%     41.36MB/s    0:00:10    2.08GB(0:01:06)
```

Monitoring health from the CLI

See the chapter about monitoring as a daily administration task from *Scale Out Network Attached Storage Monitoring*, SG24-8207 for CLI commands that can be used to monitor the cluster.

SONAS log review

SONAS log review is important for understanding the status and diagnosing any issues. With a quick search for `logs` in the SONAS section of the IBM Knowledge Center, you can find the commands for monitoring different logs. It is useful and helps to explain the logging system, and provide information about monitoring logs from a different perspective.

The SONAS section of the IBM Knowledge Center can be found on the following website:

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

The *Using IBM SONAS logs* topic provides a good starting point:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_trbl_logs.html

It can also be helpful to become familiar with common log entries when there are no issues. This can help you avoid spending time looking at entries that do not indicate problems.

The **showlog** CLI command is useful for troubleshooting or validating job or task success. It can be used to show the log file for the specified job. Example 8-7 shows common uses of the **showlog** command.

Example 8-7   Common use of the showlog command

```
showlog 15
 - Shows the log for job with jobID 15.

showlog backup:gpfs0
 - Shows the backup log for the latest backup job done for file system gpfs0.

showlog 15 -count 20
 - Shows only last 20 lines of the log for 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 date and time specified.
```

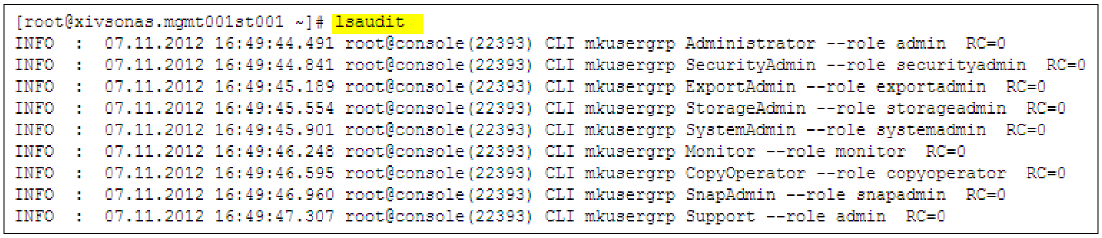
Example 8-8 shows common uses of the **showerrors** CLI command.

Example 8-8 Sample syntax of the showerrors command

```
showerrors 15
 - Shows the error log for job with jobID 15.

showerrors backup:gpfs0
 - Shows the backup error log for the latest backup job done for file system
gpfs0.

showerrors 15 -count 20
 - Shows only the last 20 lines of the error log for job with jobID 15.

showerrors backup:gpfs0 -t 03.05.2011 14:18:21.184
 - Shows the backup error log taken of file system gpfs0 at the date and time
specified.
```

Output from the lsaudit CLI command

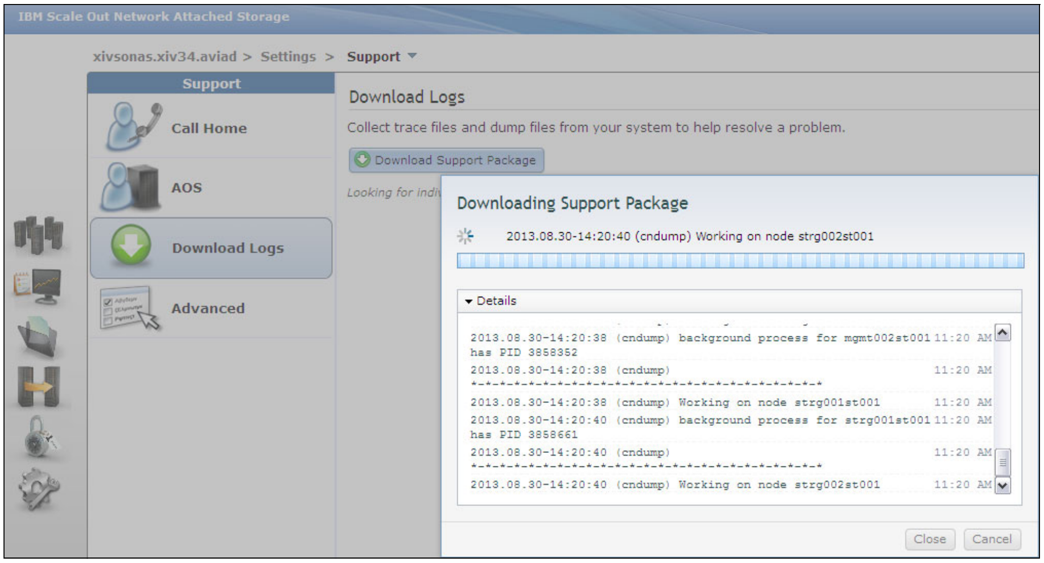
Audit logs help to track the commands that are run in the GUI and the CLI by the SONAS administrators on the system. Figure 8-8 shows sample output from the **lsaudit** command.



```
[root@xivsonas.mgmt001st001 ~]# lsaudit
INFO  :  07.11.2012 16:49:44.491 root@console(22393) CLI mkusergrp Administrator --role admin  RC=0
INFO  :  07.11.2012 16:49:44.841 root@console(22393) CLI mkusergrp SecurityAdmin --role securityadmin  RC=0
INFO  :  07.11.2012 16:49:45.189 root@console(22393) CLI mkusergrp ExportAdmin --role exportadmin  RC=0
INFO  :  07.11.2012 16:49:45.554 root@console(22393) CLI mkusergrp StorageAdmin --role storageadmin  RC=0
INFO  :  07.11.2012 16:49:45.901 root@console(22393) CLI mkusergrp SystemAdmin --role systemadmin  RC=0
INFO  :  07.11.2012 16:49:46.248 root@console(22393) CLI mkusergrp Monitor --role monitor  RC=0
INFO  :  07.11.2012 16:49:46.595 root@console(22393) CLI mkusergrp CopyOperator --role copyoperator  RC=0
INFO  :  07.11.2012 16:49:46.960 root@console(22393) CLI mkusergrp SnapAdmin --role snapadmin  RC=0
INFO  :  07.11.2012 16:49:47.307 root@console(22393) CLI mkusergrp Support --role admin  RC=0
```

Figure 8-8 The lsaudit command output

Audit logs are valuable for tracking what administrators run from the GUI or the CLI, and when they run it. One of the reasons administrators should get personal accounts with administrator privileges rather than using the default `Admin` account is to make sure that the audit log defines the person who issued a command. Otherwise, it might show a group that shares the account. You can also download the audit logs from the system in the Support view in the Settings menu, as shown in Figure 8-9.

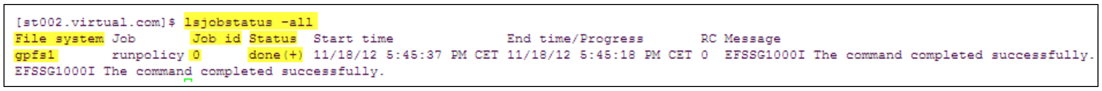

IBM Scale Out Network Attached Storage
xivsonas.xiv34.aviad > Settings > Support
Support
Call Home
AOS
Download Logs
Advanced
Download Logs
Collect trace files and dump files from your system to help resolve a problem.
Download Support Package
Looking for indi
Downloading Support Package
Details
Close
Cancel

```
2013.08.30-14:20:40 (cndump) Working on node strg002st001

2013.08.30-14:20:38 (cndump) background process for mgmt002st001 11:20 AM
has PID 3858352
2013.08.30-14:20:38 (cndump)                                      11:20 AM
*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2013.08.30-14:20:38 (cndump) Working on node strg001st001         11:20 AM
2013.08.30-14:20:40 (cndump) background process for strg001st001 11:20 AM
has PID 3858661
2013.08.30-14:20:40 (cndump)                                      11:20 AM
*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*
2013.08.30-14:20:40 (cndump) Working on node strg002st001         11:20 AM
```

Figure 8-9 Image of the log download from the Support menu

Output from the lsjobstatus CLI command

The **`lsjobststaus`** command displays the status of jobs that are currently running or already finished. For example, you can check whether an ILM policy or backup job completed, or if they completed with errors. See Figure 8-10.



```
[st002.virtual.com]$ lsjobstatus -all
File system Job       Job id Status  Start time              End time/Progress       RC Message
gpfs1       runpolicy 0      done(+) 11/18/12 5:45:37 PM CET 11/18/12 5:45:18 PM CET 0  EFSSG1000I The command completed successfully.
EFSSG1000I The command completed successfully.
```

Figure 8-10 The lsjobstatus --all command

The **`lsjobstatus`** command can be used to find the Job ID of a specific process. This information can be used to get more details about any specific job or task. It can help identify trace information that is reported when you are using the **`showlog`** command.

The **`lsjobstatus`** command can be used to list the running and completed backups. Specify a file system to show the completed and running backups for only that file system as shown in Example 8-9. The output of the **`lsjobstatus`** CLI command shows only backup session results for the past seven days.

Example 8-9 List the backups for file system gpfs0

```
[root@examplemgmt.mgmt001st001 ~]# lsjobstatus 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.
```

Output from the lsbackupfs CLI command

The **lsbackupfs** CLI command displays backup configurations as shown in Example 8-10. For each file system, the display includes the following information:

- 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 recently completed backup
- The status message from the last backup
- The last update

Example 8-10 Example of a backup that started on 1/20/2010 at 2 a.m.

```
# 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 2:00AM
```

Monitoring capacity

See the chapter about monitoring as a daily administration task from *Scale Out Network Attached Storage Monitoring*, SG24-8207 for information about monitoring capacity-related tasks.

8.1.3 Monitoring inode use and availability

See the chapter about monitoring as a daily administration task from *Scale Out Network Attached Storage Monitoring*, SG24-8207 for information about monitoring inode use and availability.

8.2 Weekly tasks to add to daily monitoring tasks

> **Tip:** Every file system and independent file set has a root file set that reports cache, snapshots, capacity, and inode consumption. When they run out of inodes, data stops writing. In that case, the **chfs** or **chfset** command can be used to increase the number of maximum and pre-allocated inodes.

Along with daily tasks, you can add the following things to check weekly (preferably before the weekend):

- All daily checks
- SONAS cluster node root file system capacity
- SONAS audit log review
- SONAS storage node NSD resource and workload balance analysis
- SONAS interface node network saturation levels
- SONAS back-end storage health reviews (gateway solutions)
- Tivoli Storage Manager server root file system capacity
- Tivoli Storage Manager server scratch tape count
- Tivoli Storage Manager server error and log review
- Open technical service request progress validations

8.2.1 Cluster node root file system capacity

When the root file system of one of the nodes fills up, it is possible for that node to either crash or become unstable. Occasional monitoring of this capacity can prevent surprises. The only way to mitigate an issue from occurring is to preemptively search (with IBM support) the file systems that are in dangerously high watermarks. Do this to help understand what options are preferred for freeing required capacity. See Example 8-11.

Example 8-11 Sample output of cndsh df -h command run on every cluster node

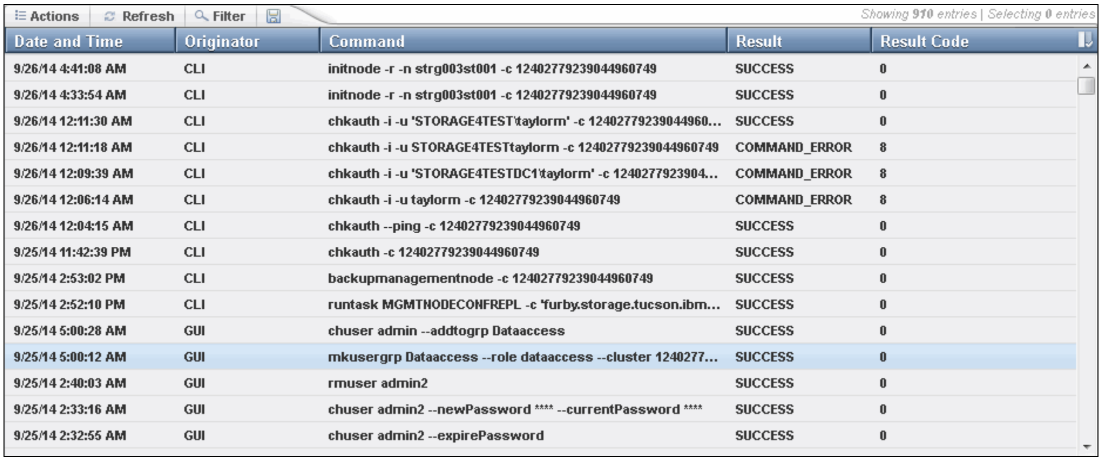
```
[root@xivsonas.mgmt001st001 ~]# cndsh df -h
mgmt001st001:  Filesystem           Size  Used Avail Use% Mounted on
mgmt001st001:  /dev/sdb2             19G   11G  7.2G  59% /
mgmt001st001:  tmpfs                 24G  4.0K   24G   1% /dev/shm
mgmt001st001:  /dev/sda1            276G  302M  261G   1% /ftdc
mgmt001st001:  /dev/sdb1            9.1G  150M  8.5G   2% /persist
mgmt001st001:  /dev/sdb5            103G  2.7G   95G   3% /var
mgmt001st001:  /dev/sdb6            9.2G  158M  8.6G   2% /var/ctdb/persistent
mgmt001st001:  /dev/gpfs0            54T  7.0G   54T   1% /ibm/gpfs0
mgmt002st001:  Filesystem           Size  Used Avail Use% Mounted on
mgmt002st001:  /dev/sda2             19G  9.3G  8.2G  54% /
mgmt002st001:  tmpfs                 24G  4.0K   24G   1% /dev/shm
mgmt002st001:  /dev/sda7            103G  222M   98G   1% /ftdc
mgmt002st001:  /dev/sda1            9.1G  150M  8.5G   2% /persist
mgmt002st001:  /dev/sda5            103G  2.4G   96G   3% /var
mgmt002st001:  /dev/sda6            9.2G  158M  8.6G   2% /var/ctdb/persistent
mgmt002st001:  /dev/gpfs0            54T  7.0G   54T   1% /ibm/gpfs0
strg001st001:  Filesystem           Size  Used Avail Use% Mounted on
strg001st001:  /dev/sda2             19G  5.2G   13G  30% /
strg001st001:  tmpfs                3.8G     0  3.8G   0% /dev/shm
strg001st001:  /dev/sda7            103G  213M   98G   1% /ftdc
strg001st001:  /dev/sda1            9.1G  150M  8.5G   2% /persist
strg001st001:  /dev/sda5            103G  1.8G   96G   2% /var
strg001st001:  /dev/sda6            9.2G  149M  8.6G   2% /var/ctdb/persistent
strg002st001:  Filesystem           Size  Used Avail Use% Mounted on
strg002st001:  /dev/sda2             19G  5.2G   13G  30% /
strg002st001:  tmpfs                3.8G     0  3.8G   0% /dev/shm
strg002st001:  /dev/sda7            103G  213M   98G   1% /ftdc
strg002st001:  /dev/sda1            9.1G  150M  8.5G   2% /persist
strg002st001:  /dev/sda5            103G  1.7G   96G   2% /var
strg002st001:  /dev/sda6              2G  149M  8.6G   2% /var/ctdb/persistent
```

8.2.2 Audit log review

When working with a team of storage administrators on a large SONAS solution, it can be helpful to review the work that is done by the team by auditing the logs weekly. This information can be useful for cross training, measuring trends of common work patterns, or evaluating training opportunities or process improvements.

Audit logs can be captured from either the GUI or the CLI. The following examples explain how this information can be obtained.

Figure 8-11 shows how to review the audit logs with the GUI.



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 8-11 GUI access to the Admin Audit Log

Example 8-12 shows CLI access to the audit log by using the **`lsaudit`** CLI command.

Example 8-12 CLI access to the audit log

```
[root@xivsonas.mgmt001st001 ~]# lsaudit
INFO  :  01.08.2013 21:29:27.353 root@unknown(60715) CLI mkusergrp Administrator --role admin  RC=0
INFO  :  01.08.2013 21:29:27.699 root@unknown(60715) CLI mkusergrp SecurityAdmin --role securityadmin  RC=0
INFO  :  01.08.2013 21:29:28.022 root@unknown(60715) CLI mkusergrp ExportAdmin --role exportadmin  RC=0
INFO  :  01.08.2013 21:29:28.351 root@unknown(60715) CLI mkusergrp StorageAdmin --role storageadmin  RC=0
INFO  :  01.08.2013 21:29:28.679 root@unknown(60715) CLI mkusergrp SystemAdmin --role systemadmin  RC=0
INFO  :  01.08.2013 21:29:29.008 root@unknown(60715) CLI mkusergrp Monitor --role monitor  RC=0
INFO  :  01.08.2013 21:29:29.333 root@unknown(60715) CLI mkusergrp CopyOperator --role copyoperator  RC=0
INFO  :  01.08.2013 21:29:29.656 root@unknown(60715) CLI mkusergrp SnapAdmin --role snapadmin  RC=0
INFO  :  01.08.2013 21:29:30.003 root@unknown(60715) CLI mkusergrp Support --role admin  RC=0
```

8.2.3 Monitoring performance and workload balance

When the performance becomes sluggish, or complaints are made about performance, your first response should be to capture a clear and concise articulation of the following information:

- Who is complaining
- What they are complaining about
- What details summarize the user, client, share, storage, data, and timing
- An accurate description of complaint

If you understand the performance indicators for normal behavior, it can help lead you to the points of interest in serious deviations. More detail is provided later in this chapter.

Performance monitoring is always a complex question. There are several components that offer simplified, high-level reference. However, realize that a client issue can (sometimes) be realized at the client system itself, and for that reason it can prove advantageous to develop methods or check points for quickly validating client concerns.

The easiest way to evaluate the SONAS system performance is to use the built-in Performance Center in SONAS. It collects the status of SONAS environment components every second, and enables you to look back on system performance averages historically.

The Performance Center is provided in both the GUI and CLI, and they use the same status log data when it is collected. You can view performance graphs of a specific part of your system in near real-time for cluster operations, front-end and back-end devices, and protocol services.

In cases of more complex configurations, such as SONAS gateways, it can be useful to use a centralized monitoring system, such as Tivoli Storage Productivity Center, to have a whole view of the storage environment.

Performance monitoring is somewhat basic for NAS systems in Tivoli Storage Productivity Center. However, it facilitates centralized monitoring of the block storage systems, the SAN network devices, and the SONAS gateway solution. These tools can chart relevant drops in performance that can help triangulate root cause indicators.

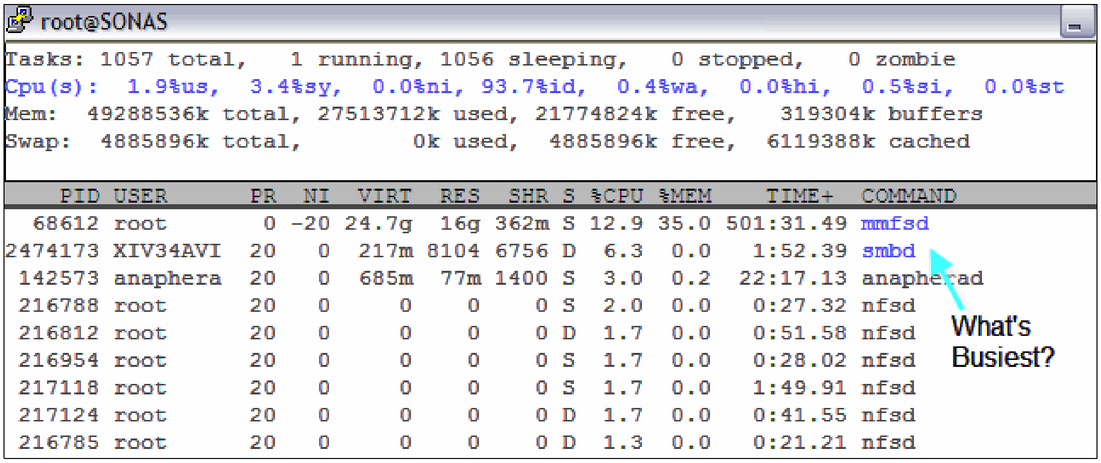
8.2.4 Performance monitoring with the SONAS GUI

See the daily monitoring information in the *Scale Out Network Attached Storage Monitoring*, SG24-8207 IBM Redbooks publication for information about performance monitoring-related tasks.

In addition to these tasks, in some cases, privileged user access can be authorized by support for advanced diagnostic review.

Top

Top is a simple and useful tool for monitoring the workload and performance in the interface and storage nodes, and at clients. So, if you run the **top** command on either interface nodes or storage nodes, you can see how hard nodes are working, and which daemons or processes are running on the top of the list, as shown in Figure 8-12.



```
Tasks: 1057 total,   1 running, 1056 sleeping,   0 stopped,   0 zombie
Cpu(s):  1.9%us,  3.4%sy,  0.0%ni, 93.7%id,  0.4%wa,  0.0%hi,  0.5%si,  0.0%st
Mem:  49288536k total, 27513712k used, 21774824k free,   319304k buffers
Swap:  4885896k total,        0k used,  4885896k free,  6119388k cached

    PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
  68612 root       0 -20 24.7g  16g 362m S 12.9 35.0 501:31.49 mmfsd
2474173 XIV34AVI  20   0  217m 8104 6756 D  6.3  0.0   1:52.39 smbd
 142573 anaphera  20   0  685m  77m 1400 S  3.0  0.2  22:17.13 anapherad
 216788 root      20   0     0    0    0 S  2.0  0.0   0:27.32 nfsd
 216812 root      20   0     0    0    0 D  1.7  0.0   0:51.58 nfsd
 216954 root      20   0     0    0    0 S  1.7  0.0   0:28.02 nfsd
 217118 root      20   0     0    0    0 S  1.7  0.0   1:49.91 nfsd
 217124 root      20   0     0    0    0 D  1.7  0.0   0:41.55 nfsd
 216785 root      20   0     0    0    0 D  1.3  0.0   0:21.21 nfsd
```

What's Busiest?

Figure 8-12 Top command output

For example, the report provides the following information:

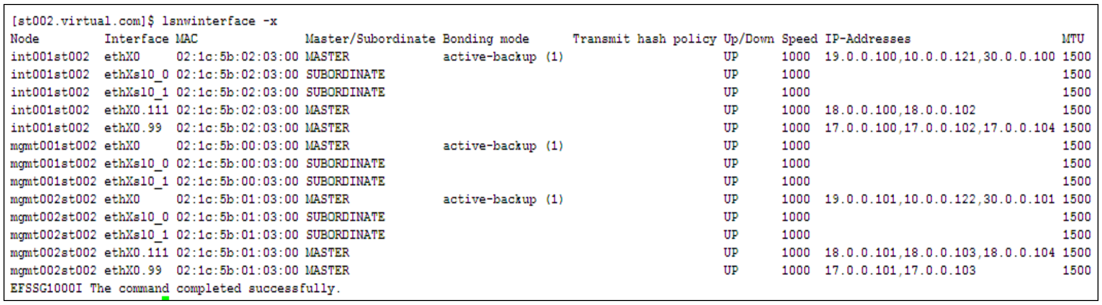
- The **mmfsd** activity shows the GPFS demon in every node.
- The **smbd** activity shows the heavy CIFS shot in the front end.
- The **nfsd** activity shows the heavy NFS shot in the front end.

Front-end performance

Ensure that as many interface nodes as possible are sharing the network, and that all subnet ports are up. The easiest way to check these configurations is the **lsnwinterface** command.

The lsnwinterface -x command

Figure 8-13 shows an example of the **`lsnwinterface -x`** command.



```
[st002.virtual.com]$ lsnwinterface -x
Node        Interface MAC               Master/Subordinate Bonding mode      Transmit hash policy Up/Down Speed IP-Addresses                          MTU
int001st002  ethX0     02:1c:5b:02:03:00 MASTER             active-backup (1)                      UP      1000  19.0.0.100,10.0.0.121,30.0.0.100 1500
int001st002  ethXsl0_0 02:1c:5b:02:03:00 SUBORDINATE                                               UP      1000                                   1500
int001st002  ethXsl0_1 02:1c:5b:02:03:00 SUBORDINATE                                               UP      1000                                   1500
int001st002  ethX0.111 02:1c:5b:02:03:00 MASTER                                                    UP      1000  18.0.0.100,18.0.0.102            1500
int001st002  ethX0.99  02:1c:5b:02:03:00 MASTER                                                    UP      1000  17.0.0.100,17.0.0.102,17.0.0.104 1500
mgmt001st002 ethX0     02:1c:5b:00:03:00 MASTER             active-backup (1)                      UP      1000                                   1500
mgmt001st002 ethXsl0_0 02:1c:5b:00:03:00 SUBORDINATE                                               UP      1000                                   1500
mgmt001st002 ethXsl0_1 02:1c:5b:00:03:00 SUBORDINATE                                               UP      1000                                   1500
mgmt002st002 ethX0     02:1c:5b:01:03:00 MASTER             active-backup (1)                      UP      1000  19.0.0.101,10.0.0.122,30.0.0.101 1500
mgmt002st002 ethXsl0_0 02:1c:5b:01:03:00 SUBORDINATE                                               UP      1000                                   1500
mgmt002st002 ethXsl0_1 02:1c:5b:01:03:00 SUBORDINATE                                               UP      1000                                   1500
mgmt002st002 ethX0.111 02:1c:5b:01:03:00 MASTER                                                    UP      1000  18.0.0.101,18.0.0.103,18.0.0.104 1500
mgmt002st002 ethX0.99  02:1c:5b:01:03:00 MASTER                                                    UP      1000  17.0.0.101,17.0.0.103            1500
EFSSG1000I The command completed successfully.
```

Figure 8-13 lsnwinterface output

The **`lsnwinterface`** command shows that Internet Protocol (IP) addresses that are assigned for external client access are evenly distributed across all interface nodes.

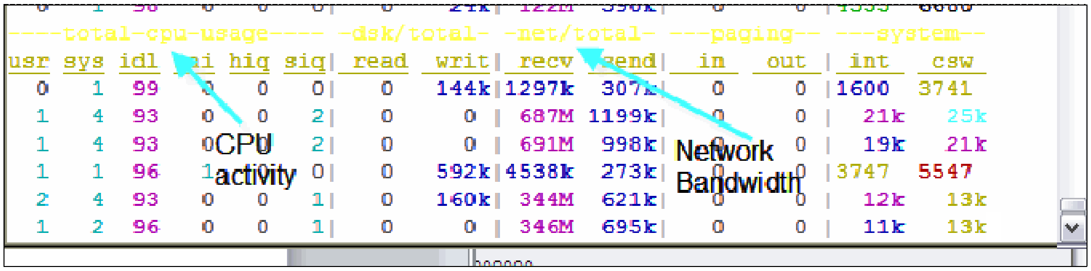
The **`lsnwinterface -x`** command shows that all subordinate ports on your network bonds are up, and that the distribution of IPs is evenly balanced.

Make sure that the interfaces are up and IP addresses are balanced across all of the active interface nodes in detail.

In the next step, you can go deeper in monitoring methods. With `root` access, you can use non-SONAS-specific commands, such as **`sdstat`**, to monitor the front-end performance, as shown in Figure 8-14.

The sdstat -a 1 command

Figure 8-14 shows an example of the **`sdstat`** command output.



```
----total-cpu-usage---- -dsk/total- -net/total- ---paging-- ---system--
usr sys idl wai hiq siq| read  writ| recv  send|  in   out | int   csw
  0   1  99   0   0   0|   0   144k|1297k  307k|   0     0 |1600  3741
  1   4  93   0   0   2|   0     0 | 687M 1199k|   0     0 |  21k   25k
  1   4  93   0   0   2|   0     0 | 691M  998k|   0     0 |  19k   21k
  1   1  96   1   0   0|   0   592k|4538k  273k|   0     0 |3747  5547
  2   4  93   0   0   1|   0   160k| 344M  621k|   0     0 |  12k   13k
  1   2  96   0   0   1|   0     0 | 346M  695k|   0     0 |  11k   13k
```

Figure 8-14 Example sdstat command output

If you run this command on each interface node in parallel, you can compare the use of nodes. This command shows information about processor use, disk, network (send and receive), paging, and system type information.

The send and receive date can show if any interface node was more loaded by clients. If you have multiple IPs on each node, you can move one of the IPs to a different node that is less busy.

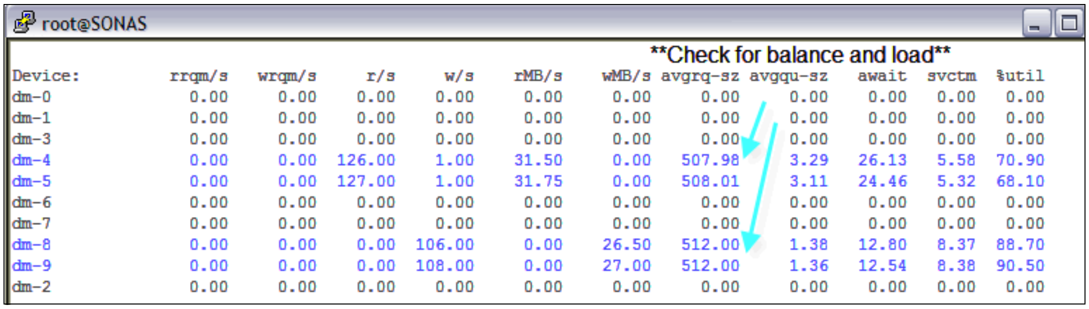
> **Tip:** If this view shows that you have reached the physical limit of network interfaces, you might need either more NICs on your interface nodes or more interface nodes to maximize your network capabilities. Low processor idle statistics can also indicate that you might not have enough interface nodes to manage the current workload.

Back-end performance

To monitor the back-end system performance, one commonly used tool is the **`iostat`** command. It also requires `root` access on the SONAS system:

```
iostat -xm /dev/dm* 1
```

This command shows the dynamic multipath device activity (reads, writes, queue-size, I/O wait, and device use), as shown in Figure 8-15.



```
Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await  svctm  %util
dm-0              0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
dm-1              0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
dm-3              0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
dm-4              0.00     0.00  126.00    1.00    31.50     0.00   507.98     3.29   26.13   5.58  70.90
dm-5              0.00     0.00  127.00    1.00    31.75     0.00   508.01     3.11   24.46   5.32  68.10
dm-6              0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
dm-7              0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
dm-8              0.00     0.00    0.00  106.00     0.00    26.50   512.00     1.38   12.80   8.37  88.70
dm-9              0.00     0.00    0.00  108.00     0.00    27.00   512.00     1.36   12.54   8.38  90.50
dm-2              0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
```

Figure 8-15 iostat -xm /dev/dm 1 command output*

In this example, you should see a fairly even I/O pattern, which means there are four NSDs working together in the target file system.

If the utilization of devices is fairly high, it might be an indication that you have not provided enough spindles behind the NSDs. Alternatively, it might mean that you have not provided enough NSDs to distributed workload for this workload, or that the file system NSD or parameter settings are not optimal for the I/O in the client workload pattern.

> **Remember:** If percent use and average queue size are high on NSDs of file system, it is a good indicator to add more disk to the file system.

Weekly checkups on the Tivoli Storage Manager server

Run the following weekly checkups on the Tivoli Storage Manager server to monitor scheduled operations and to ensure that client and server scheduled operations are completing successfully, and that no problems exist:

- Tivoli Storage Manager Server root file system capacity

  Ensure that the `Root`, `Database`, and `Log` directories do not fill up.

- Tivoli Storage Manager Server error log review and scratch tape count
- Client events

  Check that the backup and archive schedules did not fail using the following Tivoli Storage Manager command. This check is only for Tivoli Storage Manager server initiated schedules:

  ```
  tsm: TSM1URMC> Query  Event  *  *  Type=Admin BEGINDate=-1 BEGINTime=17:00
  (Format=Detail)
  ```

  In this case, the first asterisk (*) is for the domain name. The second asterisk (*) is for the schedule name.

- Administrative events

  Check that the administrative command schedules did not fail using the following Tivoli Storage Manager command:

  ```
  tsm: TSM1URMC> Query  Event  *  Type=Administrative BEGINDate=TODAY
  (Format=Detail)
  ```

  The asterisk (*) is for the schedule name.

- Scratch tapes

  Check the number of available scratch tapes using the following Tivoli Storage Manager command. Note that `Query_Scratch` is a user-defined server command script.

  ```
  tsm: TSM1URMC> RUN Q_SCRATCH
  ```

- Read-only tapes

  Check for any tapes with access of `read-only` using the following Tivoli Storage Manager command. If any tapes are in read only (RO) mode, check the Tivoli Storage Manager Server activity log for related errors:

  ```
  tsm> Query VOLume ACCess=READOnly
  ```

- Unavailable tape

  Check for any tapes with access of `unavailable` using the following Tivoli Storage Manager command. If any tapes are in this mode, check the Tivoli Storage Manager Server activity log for related errors and take appropriate actions:

  ```
  tsm: TSM1URMC> Query VOLume ACCess=UNAVailable
  ```

- Open technical service request progress validations

  Any SONAS-related technical service calls on the NAS solution or its clients should be logged and tracked for team review, knowledge transfer, and post-event process improvement. With every platform of complex technology, the quickest road to improvement comes from study of failures, issues, and events.

8.3 Monthly checks for trends, growth planning, and maintenance review

The following are monthly checks you should perform on your SONAS for monitoring and planning purposes:

- File system growth trend analysis
- File set growth trend analysis
- File system storage pool growth trend analysis:
  - Capturing growth trends and analyzing storage pool and file set growth is important for planning incremental growth. This information is obtainable from **`cndump`** files.
  - Capturing dumps once a month and charting valuable statistics from the data can help you understand trends that prevent unplanned surprises and off-hours calls to service.
  - The output of the **`cndump`** is a compressed collection of log files and command outputs that support services use to analyze your SONAS cluster health. Take the time to become familiar with its output and value for collecting the previously mentioned points of interest. A monthly review schedule seems adequate for most high-end clients. It might also be a good tool for reviewing these trends with your technical advisors or IBM account teams for planning future growth demands.

- System performance trend analysis

  Cataloging some basic performance watermarks and trends from your weekly performance reviews adds value to your monthly report and trend analysis. Actions for expansion consideration should be considered if and when you near saturation points on the front- or back-end devices in your SONAS cluster.

- User and client satisfaction survey

  At the end of the day, the clients who use the services day-in and day-out are the best resources for consistent feedback on solution satisfaction. It might be beneficial to establish a high-level, simplified service satisfaction survey for your client base that can be reviewed on a monthly or semiannual basis. Any sense of dissatisfaction can be further escalated and better understood for focused, effective response.

- IBM technical advisors and other vendor bug and patch advisory meetings

  Most IBM account teams are happy to meet monthly to discuss or review solution feedback, growth analysis, needs changes, and so on. In many cases, clients have assigned technical advisors that actively follow their needs and product developments to keep both sides well-informed.

- Maintenance and process improvement considerations review:

  - Internal NAS team meetings are a common place for reviewing monthly trend analysis, cluster growth trend maintenance plans, and to discuss ideas for process improvements.
  - A strong, informed team with continuous process improvement, and a stable NAS solution with adequate capacity, are the keys to success with a Scale Out NAS requirement.

8.4 Monitoring with IBM Tivoli Storage Productivity Center

For a complete guide to using Tivoli Productivity Center to monitor SONAS, see the *Scale Out Network Attached Storage Monitoring*, SG24-8207 IBM Redbooks publication.

8.4.1 Summary for monitoring SONAS daily, weekly, and monthly

This checklist summarizes the preferred practices for SONAS monitoring:

Daily Checks

□ Cluster health

□ Cluster node health

□ Cluster services health

□ File system capacity

□ File system and independent file set iNode capacity

□ Storage pool capacity

□ Backup job status

□ Replication task status

□ Events list review

□ User quota checks

Weekly Checks (preferably prior to weekend)

- □ All daily checks
- □ SONAS cluster node root file system capacity
- □ SONAS audit log review
- □ SONAS storage node NSD resource and workload balance analysis
- □ SONAS interface node network saturation levels
- □ SONAS back-end storage health reviews (gateway solutions)
- □ Tivoli Storage Manager server root file system capacity
- □ Tivoli Storage Manager server scratch tape count
- □ Tivoli Storage Manager server error and log review
- □ Open technical service request progress validations

Monthly checks (trends, growth planning, and maintenance review)

- □ File system growth trend analysis
- □ File set growth trend analysis
- □ File system storage pool growth trend analysis
- □ System performance trend analysis
- □ User and client satisfaction survey
- □ IBM TA and other vendor bug and patch advisory meetings
- □ Process improvement considerations review
- □ Maintenance and project planning meetings

9

Troubleshooting and support

This chapter introduces how to open a problem management record (PMR) for software and hardware support and how to use the IBM Scale Out Network Attached Storage (SONAS) IBM Knowledge Center to provide troubleshooting guidance. It also explains how to prepare for correct, efficient service escalation to get correct support attention quickly. It explains how to work with IBM teams such as technical advisors and development when service challenges are at a heightened sense of concern.

Also see the Start here for troubleshooting topic in the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/sonas_starthere.html

This chapter describes the following topics:

- Creating PMRs
- Network troubleshooting
- Troubleshooting authentication issues

9.1 Creating PMRs

In any case where there is a technical product issue, a problem ticket is required for official service and troubleshooting assistance. This ticket is called a PMR, and it is a requirement to open one to get assistance with diagnostics in SONAS. IBM uses the PMR system to track support activity for client-owned IBM products and applications. SONAS call home services might automatically create a PMR, or you can manually create a PMR to track maintenance issues and resolutions from onset to close.

In some cases, *code defects* are opened by the SONAS support team at the same time as PMR activity if the problem is suspect as a coding defect, if an anomaly is discovered, or even if an error exists in documentation. This section walks you through the process to create a hardware and a software PMR. This section provides the phone numbers and test scenarios to guide you through the process.

As issues develop, it is the client's responsibility to manage the understanding of severity levels. It is encouraged that they keep their assigned technical advisor informed of issues that arise to ensure the highest level of cooperative awareness, escalation, resolution, and ensure that further risk mitigation is properly applied. The IBM team that is assigned to every SONAS is committed to client success, and is helpful at escalating to any level to help maintain the highest level of customer satisfaction.

Note that in the case of a SONAS that is installed as a gateway with separate storage (such as IBM XIV, IBM Storwize V7000, or IBM DS8000), there might be a need to establish a PMR for either the SONAS stack or the storage stack independently. The products in a gateway configuration are separate, and are not bundled under the same warranty or serial number.

9.1.1 Software PMR

Complete the following steps to open a software PMR:

1. Call IBM at 1-800-IBM-SERV, which is 1-800-426-7378.

   You are prompted to respond if you are calling about a Lenovo product or an IBM product:

   – Press 1 for a Lenovo product.
   – Press 2 for an IBM product.

   In this scenario, press 2.

2. You are asked if you have a five-digit premium service number. This number does not apply in this case.

3. You are asked if it is a hardware or software problem:

   – Press 1 for a hardware problem.
   – Press 2 for a software problem.

   In this scenario, press 2.

4. You are asked if the system is IBM AIX or Other:

   – Press 1 for AIX.
   – Press 2 for Other.

   In this scenario, press 2.

5. You are transferred to a live representative, who asks for the following information:
   - Customer ID, which is your 7-digit customer number.
   - Component ID, which is 5639SN100 for SONAS.
   - Software version of SONAS that you are running, for example 1.3.1.1. If you do not know this answer, it does not matter.
   - Your name, phone number, and email address.
   - Severity of the problem.

A PMR is opened based on the information you provide.

9.1.2 Hardware PMR

The following steps take you through the prompts to open a hardware PMR.

You go through the same phone steps as in, 9.1.1, “Software PMR” on page 280, except that you select **1** for hardware. Then you give the machine type 2851, and the serial number of any of your SONAS servers:

1. Call IBM at 1-800-IBM-SERV, which is 1-800-426-7378.

   You are prompted to respond if you are calling about a Lenovo product or an IBM product:
   - Press 1 for a Lenovo product.
   - Press 2 for an IBM product.

     In this scenario, press 2.
2. You are asked if you have a five-digit premium service number. This number does not apply in this case.
3. You are asked if it is a hardware or software problem:
   - Press 1 for a hardware problem.
   - Press 2 for a software problem.

   In this scenario, press 1.
4. You are asked if the system is an AIX or Other:
   - Press 1 for AIX.
   - Press 2 for Other.

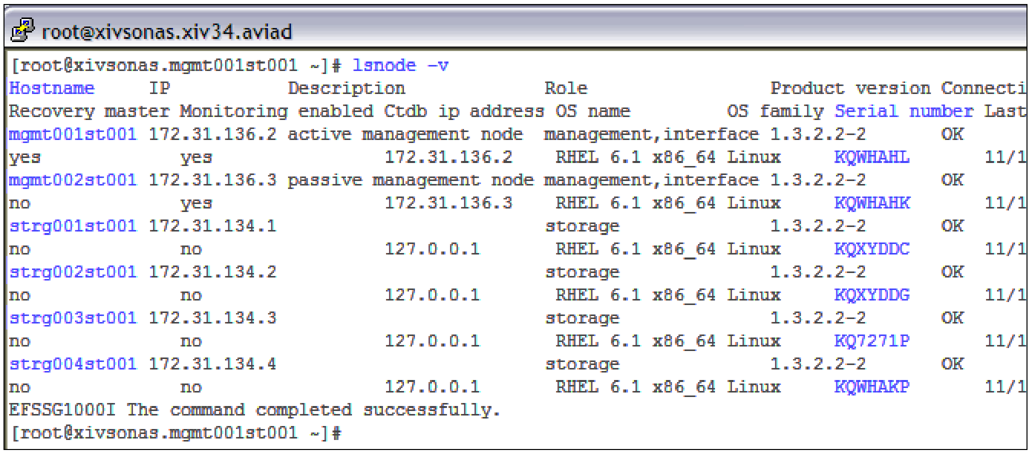
   In this scenario, press 2.
5. You are transferred to a live representative, who asks for the following information:
   - Customer ID, which is your 7-digit customer number
   - Machine type, which is 2851, and a serial number for any of the SONAS servers
   - Your name, phone number, and email address
   - Severity of the problem

You can easily get the serial number by using the **`lsnode -v`** command as shown in Figure 9-1 on page 282. Notice that under the Serial Number column for mgmt001st001, you see KQRGMZV.

This machine type and serial number is required to pass hardware entitlement. After you pass, IBM opens a hardware PMR, which gets routed to SONAS L1 support.

If you need an IBM service support representative (SSR) dispatched to come on site, this request must always be made with a hardware PMR. SSR Dispatch by using a software PMR cannot be done.

Serial numbers of the SONAS nodes can be obtained with the **`lsnode -v`** command, as shown in Figure 9-1.



root@xivsonas.xiv34.aviad

```
[root@xivsonas.mgmt001st001 ~]# lsnode -v
Hostname     IP           Description             Role                Product version Connecti
Recovery master Monitoring enabled Ctdb ip address OS name         OS family Serial number Last
mgmt001st001 172.31.136.2 active management node  management,interface 1.3.2.2-2       OK
yes             yes                172.31.136.2    RHEL 6.1 x86_64 Linux      KQWHAHL       11/1
mgmt002st001 172.31.136.3 passive management node management,interface 1.3.2.2-2       OK
no              yes                172.31.136.3    RHEL 6.1 x86_64 Linux      KQWHAHK       11/1
strg001st001 172.31.134.1                         storage             1.3.2.2-2       OK
no              no                 127.0.0.1       RHEL 6.1 x86_64 Linux      KQXYDDC       11/1
strg002st001 172.31.134.2                         storage             1.3.2.2-2       OK
no              no                 127.0.0.1       RHEL 6.1 x86_64 Linux      KQXYDDG       11/1
strg003st001 172.31.134.3                         storage             1.3.2.2-2       OK
no              no                 127.0.0.1       RHEL 6.1 x86_64 Linux      KQ7271P       11/1
strg004st001 172.31.134.4                         storage             1.3.2.2-2       OK
no              no                 127.0.0.1       RHEL 6.1 x86_64 Linux      KQWHAKP       11/1
EFSSG1000I The command completed successfully.
[root@xivsonas.mgmt001st001 ~]#
```

Figure 9-1 Use the lsnode -v command output to display a SONAS serial number for a PMR

9.2 Network troubleshooting

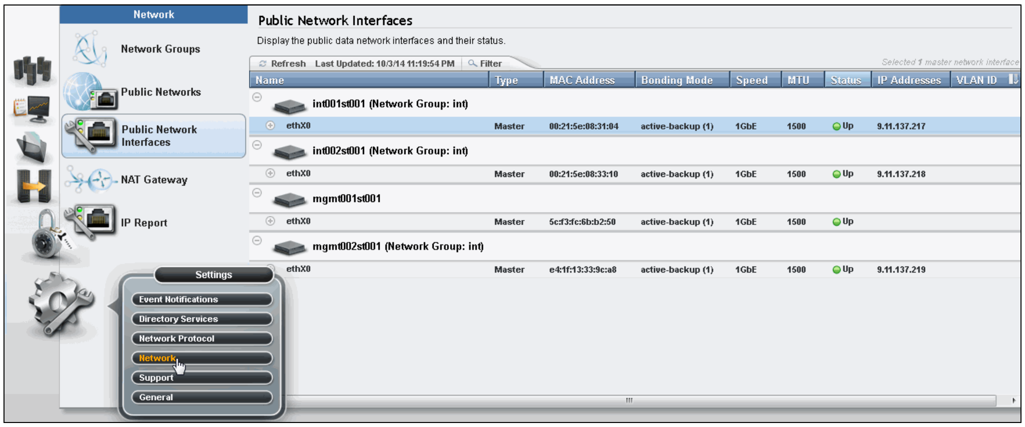
This section describes methods to gather diagnostic information about your SONAS network.

9.2.1 Checking network interface availability

You have several options for checking network availability by using the IBM SONAS graphical user interface (GUI) or the command-line interface (CLI):

1. Complete the following steps to use the GUI:

   a. In the GUI, select **Settings** → **Network** → **Public Network Interface**. The status should be up for each network interface for all the nodes, as shown in Figure 9-2.



Public Network Interfaces

Display the public data network interfaces and their status.

Last Updated: 10/3/14 11:19:54 PM

Selected 1 master network interface

Name	Type	MAC Address	Bonding Mode	Speed	MTU	Status	IP Addresses	VLAN ID
int001st001 (Network Group: int)								
ethX0	Master	00:21:5e:08:31:04	active-backup (1)	1GbE	1500	● Up	9.11.137.217	
int002st001 (Network Group: int)								
ethX0	Master	00:21:5e:08:33:10	active-backup (1)	1GbE	1500	● Up	9.11.137.218	
mgmt001st001								
ethX0	Master	5c:f3:fc:6b:b2:50	active-backup (1)	1GbE	1500	● Up		
mgmt002st001 (Network Group: int)								
ethX0	Master	e4:1f:13:33:9c:a8	active-backup (1)	1GbE	1500	● Up	9.11.137.219	

Figure 9-2 Public Network Interfaces status check

   b. By expanding each of the network interfaces, you can view the status and other properties of the subordinate interface.

2. In the CLI, check the status of the interface `ethX0` (the interface of interface nodes to the customer network):

   a. Open the CLI.

   b. Use the **`lsnwinterface`** command to display the status for the wanted Internet Protocol (IP) addresses.

      The **`#lsnwinterface -x`** command shows the port status with the subordinate port, as highlighted in Example 9-1.

      In the Up/Down column, `UP` indicates a valid connection.

      Int001st001 shows one of six 1 gigabit Ethernet (GbE) subordinate ports (1000) available to keep networking alive on that node, where `int002st001` shows one of two 10 GbE ports (10000) alive.

      The **`lsnwinterface`** command also displays all IP addresses that are actively assigned to Interface nodes, as highlighted on `int003st001`.

Example 9-1 Sample lsnwinterface command output

```
Node Interface MAC Master/Subordinate Bonding mode Up/Down IP-Addresses Speed MTU
int001st001 ethX0 00:21:5e:08:8b:44 MASTER balance-alb (6) UP 9.11.136.52 N/A 1500
int001st001 ethXsl0_0 00:15:17:cc:7b:06 SUBORDINATE DOWN 1000 1500
int001st001 ethXsl0_1 00:21:5e:08:8b:46 SUBORDINATE DOWN 1000 1500
int001st001 ethXsl0_2 00:21:5e:08:8b:44 SUBORDINATE UP 1000 1500
int001st001 ethXsl0_3 00:15:17:cc:7b:04 SUBORDINATE DOWN 1000 1500
int001st001 ethXsl0_4 00:15:17:cc:7b:07 SUBORDINATE DOWN 1000 1500
int001st001 ethXsl0_5 00:15:17:cc:7b:05 SUBORDINATE DOWN 1000 1500
int002st001 eth2 00:15:17:cc:79:cd DOWN N/A 1500
int002st001 eth3 00:15:17:cc:79:cc DOWN N/A 1500
int002st001 eth4 00:15:17:cc:79:cf DOWN N/A 1500
int002st001 eth5 00:15:17:cc:79:ce DOWN N/A 1500
int002st001 eth8 00:21:5e:08:88:b8 DOWN N/A 1500
int002st001 eth9 00:21:5e:08:88:ba DOWN N/A 1500
int002st001 ethX0 00:c0:dd:11:36:98 MASTER active-backup (1) UP 9.11.136.51 N/A
1500
int002st001 ethXsl0_0 00:c0:dd:11:36:98 SUBORDINATE DOWN 10000 1500
int002st001 ethXsl0_1 00:c0:dd:11:36:98 SUBORDINATE UP 10000 1500
int003st001 eth2 00:15:17:c5:cf:a5 DOWN N/A 1500
int003st001 eth3 00:15:17:c5:cf:a4 DOWN N/A 1500
int003st001 eth4 00:15:17:c5:cf:a7 DOWN N/A 1500
int003st001 eth5 00:15:17:c5:cf:a6 DOWN N/A 1500
int003st001 eth8 e4:1f:13:8d:d1:6c DOWN N/A 1500
int003st001 eth9 e4:1f:13:8d:d1:6e DOWN N/A 1500
int003st001 ethX0 00:c0:dd:12:1e:58 MASTER active-backup (1) UP 9.11.136.50 N/A
1500
int003st001 ethXsl0_0 00:c0:dd:12:1e:58 SUBORDINATE DOWN 10000 1500
int003st001 ethXsl0_1 00:c0:dd:12:1e:58 SUBORDINATE UP 10000 1500
mgmt001st001 eth3 00:21:5e:08:8b:be DOWN N/A 1500
mgmt001st001 ethX0 00:21:5e:08:8b:bc UP N/A 1500
```

3. If the network interface is not available, perform a visual inspection of the cabling to ensure that it is plugged in. For example, if you have no system connectivity between nodes and switches, check the external Ethernet cabling. If that cabling is in place, next check the internal InfiniBand cabling. If the cabling is all good, you then need to work upstream. Check intranet availability, for example, or external Internet availability. If none of these checks leads to a resolution of the problem, contact your next level of support.

9.2.2 Collecting network data

This section explains how to collect network data.

The starttrace command

The **starttrace** command declares a trace to monitor network traffic, system calls, or both. The monitoring can generate a huge amount of log data in a short time. The following conditions are checked and must be met before the **starttrace** command is accepted:

- Nodes must have at least 1.1 gigabytes (GB) free storage in the `/var/tmp` folder.
- The management node must have at least (1.2 * <*number of nodes*> * 0.66) GB plus 100 megabytes (MB) contingency of storage capacity in its `/ftdc` folder to store the collected and compressed log files from the interface nodes after the tracing ends.

The **starttrace** command has the following options: `--cifs`, `--nfs`, `--network`, `--gpfs`, `--client`, `--systemcalls`, `--restart-service`, `--duration` <*duration*>. For more information, see the starttrace topic in the SONAS section of the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/manpages/starttrace.html

Example 9-2 shows an example of running the **starttrace** command for network diagnostics.

Example 9-2 Example of running a starttrace command to capture network diagnostics

```
[furby.storage.tucson.ibm.com]$ starttrace --network --client 9.11.213.38
--duration 1
```

In Example 9-2, a trace is run for one minute (the duration that is specified), and the compressed file is placed in the `/ftdc` directory on completion with the following naming convention:

`trace_<date-time-stamp_network>_<Trace-ID-number>_root.gzp`

The lstrace command

The **lstrace** command lists all known traces that were previously created by the **starttrace** command and *which are not yet finished*. Example 9-3 is the output from the network trace followed by an **lstrace** command.

Example 9-3 Sample output from running a network trace followed by an lstrace command

```
[furby.storage.tucson.ibm.com]$ starttrace --network --client 9.11.213.38
--duration 1
TraceID=5248875006284398592
Logfilename=trace_20141003134237_network_5248875006284398592_admin.tgz
EFSSG1000I The command completed successfully.

[furby.storage.tucson.ibm.com]$ lstrace
TraceID             User  Cifs Network Nfs Syscalls Starttime      Endtime
Logfilename                                                Size ClientIP
5248875006284398592 admin no   yes     no  no       20141003134237 20141003134337
trace_20141003134237_network_5248875006284398592_admin.tgz 0    9.11.213.38
EFSSG1000I The command completed successfully.
```

The **lstrace** command does not list a trace that has completed. To see logs for traces that were previously run, you must have admin-privileged access to the GUI.

9.2.3 Working with the network and protocol trace commands

This section provides additional examples of network and protocol trace commands.

Starting a trace

In Example 9-4, the **starttrace** command is initiated to collect all Common Internet File System (CIFS) and network traces from the client 9.11.213.38.

Example 9-4 An example of starting network tracing with the CLI

```
[furby.storage.tucson.ibm.com]$ starttrace --cifs --client 9.11.213.38 --duration
1
Caution. The ip address 9.11.213.38 has a cifs connection.
Starting traces for the given clients will stop their connections prior to
tracing.
Open files on these connections might get corrupted so please close them first.
Do you really want to perform the operation (yes/no - default no):yes
TraceID=5248878942638702592
Logfilename=trace_20141003134632_cifs_5248878942638702592_admin.tgz
EFSSG1000I The command completed successfully.
```

Listing running traces

Example 9-5 shows how to list running traces using the **lstrace** command.

Example 9-5 Listing running traces

```
#[furby.storage.tucson.ibm.com]$ lstrace
TraceID             User  Cifs Network Nfs Syscalls Starttime      Endtime
Logfilename                                             Size ClientIP
5248878942638702592 admin yes  no      no  no       20141003134632 20141003134732
trace_20141003134632_cifs_5248878942638702592_admin.tgz 0    9.11.213.38
EFSSG1000I The command completed successfully.
```

The TraceID identifies the ID number of the trace to show. This ID must be used to stop the trace.

The Starttime and Endtime values identify the data collection duration. The default value is 10 minutes.

The Logfilename identifies the file name of the collected data. The file can be found in the /ftdc directory, or you can download it from the GUI. See “Downloading the trace files using the GUI” on page 286.

Stopping a running trace

The **stoptrace** command stops traces that were created by the **starttrace** command. You can stop a dedicated TraceID or all traces by using this command.

Example 9-6 shows an example of the **stoptrace** command.

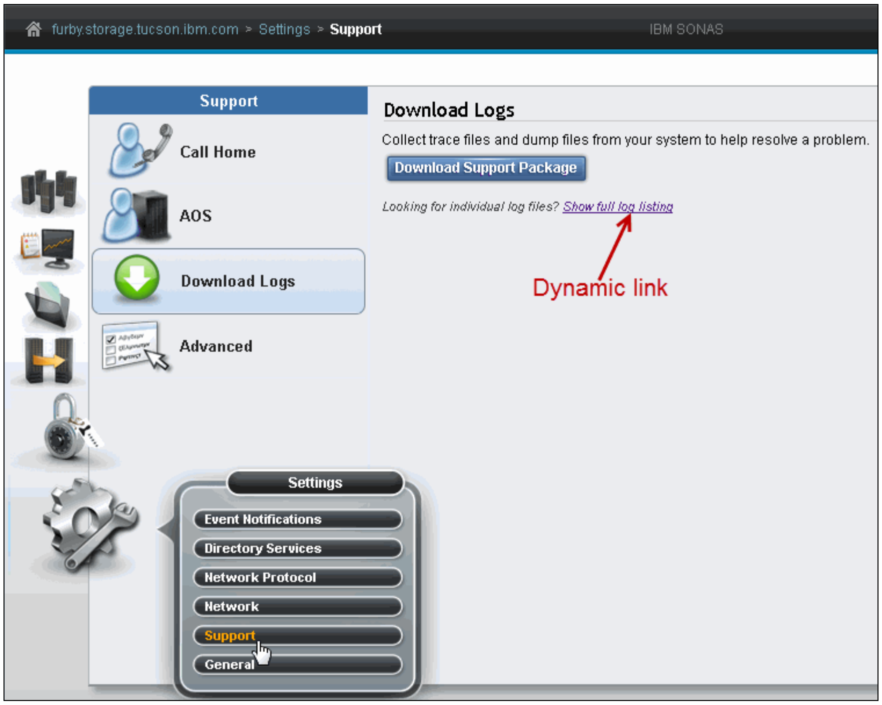
Example 9-6 Stopping a trace example

```
# stoptrace 4266819943737196544
EFSSG1000I The command completed successfully.
```

Downloading the trace files using the GUI

Complete the following steps to download the trace files using the GUI:

1. Log in to the GUI and select **Support** → **Download Logs**, as shown in Figure 9-3.



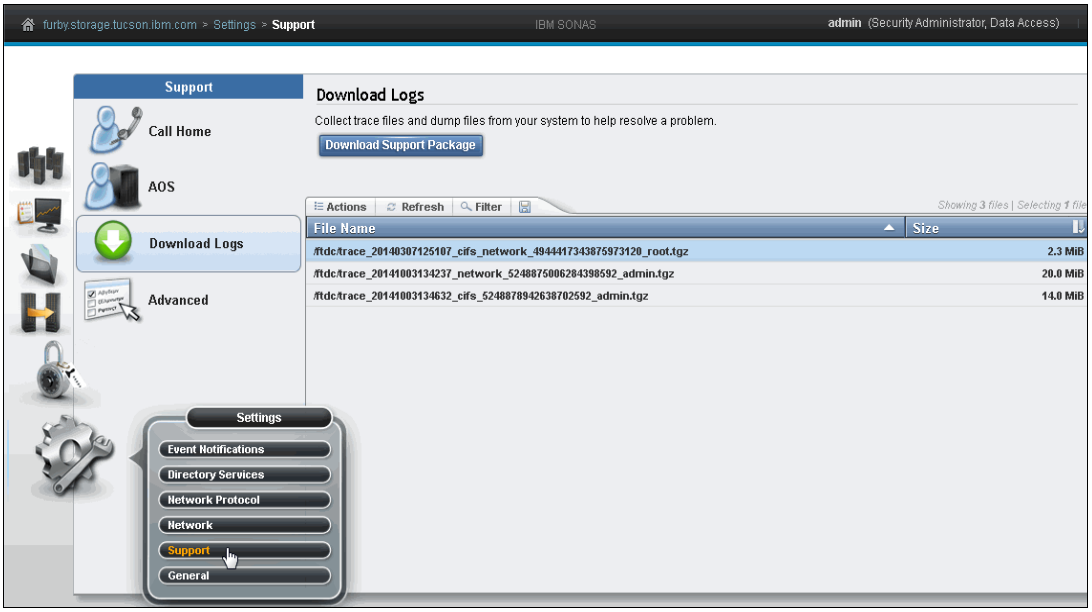
Download Logs

Collect trace files and dump files from your system to help resolve a problem.

Looking for individual log files? Show full log listing

Figure 9-3 GUI expansion of the full log view

2. Click the **Show full log listing** link, as shown in Figure 9-3. The list of downloadable logs displays, as shown in Figure 9-4.



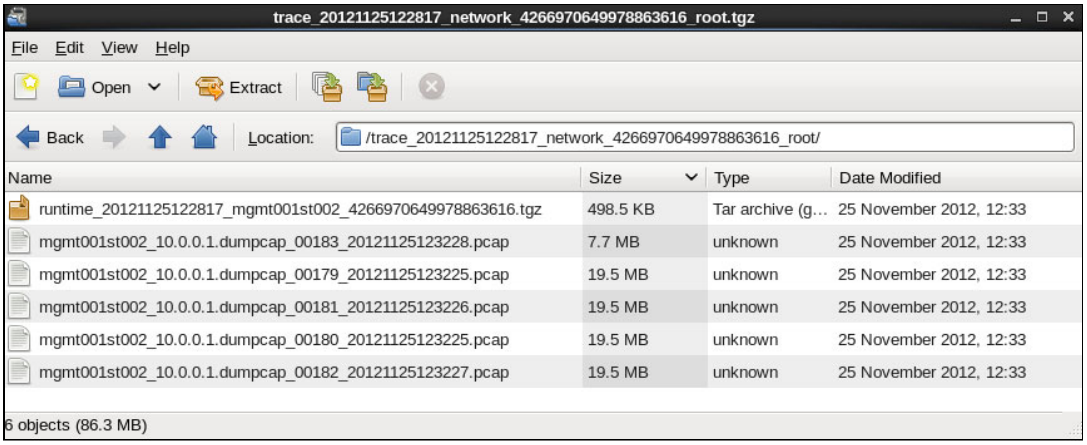
Download Logs

Collect trace files and dump files from your system to help resolve a problem.

File Name	Size
/ftdc/trace_20140307125107_cifs_network_4944417343875973120_root.tgz	2.3 MiB
/ftdc/trace_20141003134237_network_5248875006284398592_admin.tgz	20.0 MiB
/ftdc/trace_20141003134632_cifs_5248878942638702592_admin.tgz	14.0 MiB

Figure 9-4 Sample GUI view of the list of downloadable logs full list

The compressed file contains configuration files, log files, and the network dump files, as shown in Figure 9-5.



trace_20121125122817_network_4266970649978863616_root.tgz

/trace_20121125122817_network_4266970649978863616_root/

Name	Size	Type	Date Modified
runtime_20121125122817_mgmt001st002_4266970649978863616.tgz	498.5 KB	Tar archive (g…	25 November 2012, 12:33
mgmt001st002_10.0.0.1.dumpcap_00183_20121125123228.pcap	7.7 MB	unknown	25 November 2012, 12:33
mgmt001st002_10.0.0.1.dumpcap_00179_20121125123225.pcap	19.5 MB	unknown	25 November 2012, 12:33
mgmt001st002_10.0.0.1.dumpcap_00181_20121125123226.pcap	19.5 MB	unknown	25 November 2012, 12:33
mgmt001st002_10.0.0.1.dumpcap_00180_20121125123225.pcap	19.5 MB	unknown	25 November 2012, 12:33
mgmt001st002_10.0.0.1.dumpcap_00182_20121125123227.pcap	19.5 MB	unknown	25 November 2012, 12:33

6 objects (86.3 MB)

Figure 9-5 Sample compressed file view

The .pcap files can be used with any network analysis program, such as Wireshark.

Using the dig command

The domain information groper (**dig**) command is a flexible tool for interrogating domain name server (DNS) name servers. It does DNS lookups and displays the answers that are returned from the queried name servers, A records, and MX records. Example 9-7 shows a sample **dig** command output.

Example 9-7 Sample dig command output

```
$ dig ADS.virtual.com

; <<>> DiG 9.7.3-P3-RedHat-9.7.3-2.el6_1.P3.3 <<>> ADS.virtual.com
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 19433
;; flags: qr aa rd ra; QUERY: 1, ANSWER: 3, AUTHORITY: 0, ADDITIONAL: 0

;; QUESTION SECTION:
;ADS.virtual.com.INA

;; ANSWER SECTION:
ADS.virtual.com.3600INA10.0.0.100
ADS.virtual.com.3600INA10.0.2.100
ADS.virtual.com.3600INA10.0.1.100

;; Query time: 0 msec
;; SERVER: 10.0.0.100#53(10.0.0.100)
;; WHEN: Tue Dec  4 18:53:09 2012
;; MSG SIZE  rcvd: 81

$ dig +nocmd ADS.virtual.com any +multiline + noall +answer
ADS.virtual.com.3600 INA 10.0.2.100
ADS.virtual.com.3600 INA 10.0.0.100
ADS.virtual.com.3600 INA 10.0.1.100
```

9.2.4 Troubleshooting GPFS and file system issues

Troubleshooting can go in many directions. In general, if you suspect that there might be an issue, open a PMR on the software to get the correct level of attention and guidance.

Before you open the PMR, download a support package. See the *IBM SONAS Implementation Guide*, SG24-7962 IBM Redbooks publication for details about collecting logs for support.

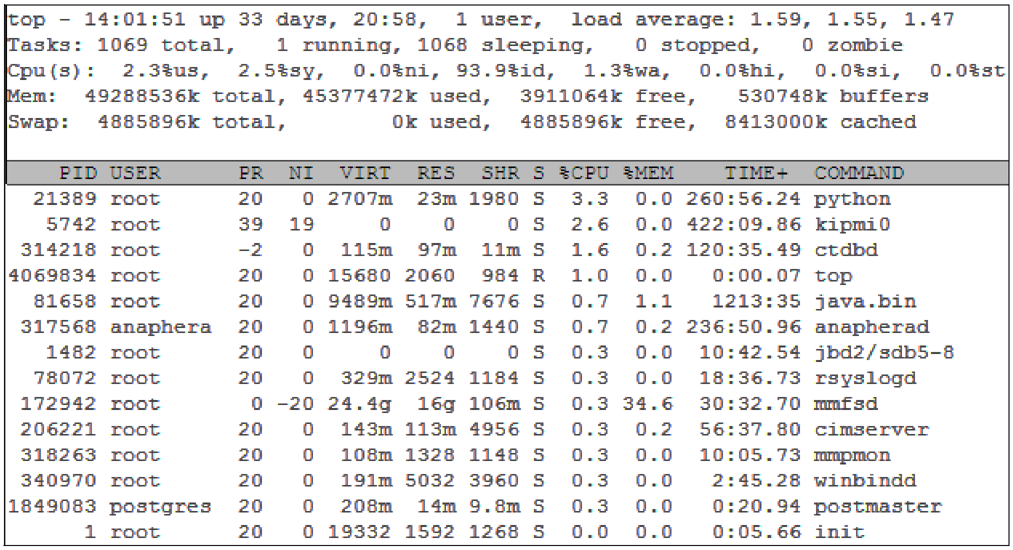
Complete the following steps to collect support logs:

1. Look at the Event log in the GUI and collect any errors or indicators of issues that you might be having.
2. Collect the log `Warnings` or `Severe Issues Report`.

   Run the **`lslog -l SEVERE`** and **`lslog -l WARNING`** commands and use the **`grep`** command or search the `/var/log/messages` file for `gpfs` to find fault indicators or long input/output (I/O) waiters. Lists logs (**`lslog`**) for the specified log level and higher (`FINEST`, `FINER`, `FINE`, `CONFIG`, `INFO`, `WARNING`, or `SEVERE`).
3. Look at the node performance with the **`top`** and **`iostat`** commands as a previleged root user.

   The **`top`** command shows you how hard the node is working, and what processes and applications are working the hardest (it is often an easy way to see what is struggling). This command requires root-privileged access on cluster nodes, but it can provide a good description of what a node is spending its time and resources doing.

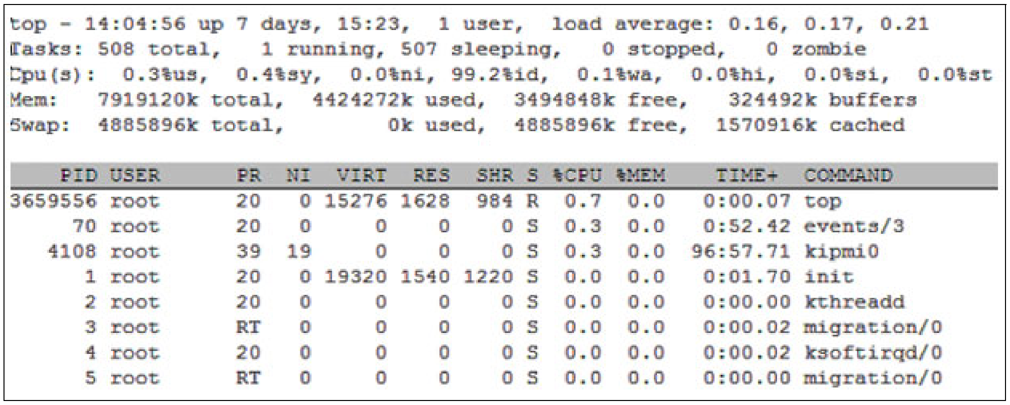
   Figure 9-6 shows sample output from the **`top`** command on an interface node.



```
top - 14:01:51 up 33 days, 20:58,  1 user,  load average: 1.59, 1.55, 1.47
Tasks: 1069 total,   1 running, 1068 sleeping,   0 stopped,   0 zombie
Cpu(s):  2.3%us,  2.5%sy,  0.0%ni, 93.9%id,  1.3%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  49288536k total, 45377472k used,  3911064k free,   530748k buffers
Swap:  4885896k total,        0k used,  4885896k free,  8413000k cached

    PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
  21389 root      20   0 2707m  23m 1980 S  3.3  0.0 260:56.24 python
   5742 root      39  19     0    0    0 S  2.6  0.0 422:09.86 kipmi0
 314218 root      -2   0  115m  97m  11m S  1.6  0.2 120:35.49 ctdbd
4069834 root      20   0 15680 2060  984 R  1.0  0.0   0:00.07 top
  81658 root      20   0 9489m 517m 7676 S  0.7  1.1   1213:35 java.bin
 317568 anaphera  20   0 1196m  82m 1440 S  0.7  0.2 236:50.96 anapherad
   1482 root      20   0     0    0    0 S  0.3  0.0  10:42.54 jbd2/sdb5-8
  78072 root      20   0  329m 2524 1184 S  0.3  0.0  18:36.73 rsyslogd
 172942 root       0 -20 24.4g  16g 106m S  0.3 34.6  30:32.70 mmfsd
 206221 root      20   0  143m 113m 4956 S  0.3  0.2  56:37.80 cimserver
 318263 root      20   0  108m 1328 1148 S  0.3  0.0  10:05.73 mmpmon
 340970 root      20   0  191m 5032 3960 S  0.3  0.0   2:45.28 winbindd
1849083 postgres  20   0  208m  14m 9.8m S  0.3  0.0   0:20.94 postmaster
      1 root      20   0 19332 1592 1268 S  0.0  0.0   0:05.66 init
```

Figure 9-6 Sample output from the top command on an interface node

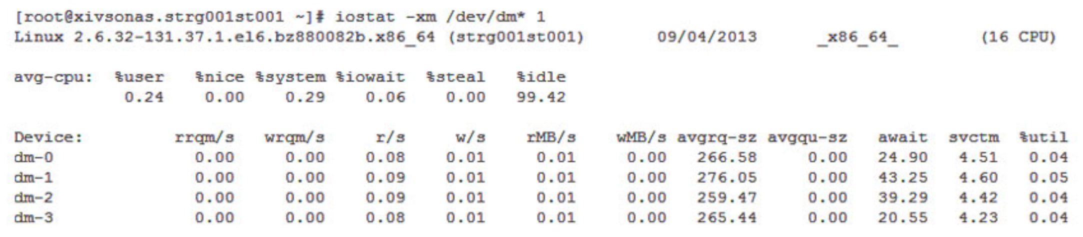
Figure 9-7 shows sample output from the **top** command on a storage node.



```
top - 14:04:56 up 7 days, 15:23,  1 user,  load average: 0.16, 0.17, 0.21
Tasks: 508 total,   1 running, 507 sleeping,   0 stopped,   0 zombie
Cpu(s):  0.3%us,  0.4%sy,  0.0%ni, 99.2%id,  0.1%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:   7919120k total,  4424272k used,  3494848k free,   324492k buffers
Swap:  4885896k total,        0k used,  4885896k free,  1570916k cached

    PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
3659556 root      20   0 15276 1628  984 R  0.7  0.0   0:00.07 top
     70 root      20   0     0    0    0 S  0.3  0.0   0:52.42 events/3
   4108 root      39  19     0    0    0 S  0.3  0.0  96:57.71 kipmi0
      1 root      20   0 19320 1540 1220 S  0.0  0.0   0:01.70 init
      2 root      20   0     0    0    0 S  0.0  0.0   0:00.00 kthreadd
      3 root      RT   0     0    0    0 S  0.0  0.0   0:00.02 migration/0
      4 root      20   0     0    0    0 S  0.0  0.0   0:00.02 ksoftirqd/0
      5 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 migration/0
```

Figure 9-7 Sample output from the command on a storage node

The **iostat -xm /dev/dm* 1** command shows you 1-second updates on how busy each multipath device is on the storage node. When all of the devices are consistently hitting 100% busy, it can be an indication that the solution is disk-bound or back-end bottlenecked. See Figure 9-8.



```
[root@xivsonas.strg001st001 ~]# iostat -xm /dev/dm* 1
Linux 2.6.32-131.37.1.el6.bz880082b.x86_64 (strg001st001)       09/04/2013      _x86_64_        (16 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.24    0.00    0.29    0.06    0.00   99.42

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await  svctm  %util
dm-0              0.00     0.00    0.08    0.01     0.01     0.00   266.58     0.00   24.90   4.51   0.04
dm-1              0.00     0.00    0.09    0.01     0.01     0.00   276.05     0.00   43.25   4.60   0.05
dm-2              0.00     0.00    0.09    0.01     0.01     0.00   259.47     0.00   39.29   4.42   0.04
dm-3              0.00     0.00    0.08    0.01     0.01     0.00   265.44     0.00   20.55   4.23   0.04
```

Figure 9-8 Sample iostat output from a SONAS storage node with four multipath storage NSDs

Analyze the GPFS logs

Contact IBM support for analysis of IBM General Parallel File System (IBM GPFS) log entries. Use this procedure when reviewing GPFS log entries:

1. Download and decompress the support package.
2. Review the log file `/var/adm/ras/mmfs.log.latest`. The log file is ordered from oldest to newest, so the end of the log has the current GPFS information.

The GPFS log is a complex raw log file for GPFS. If you are unable to understand the conditions that are listed in the log, contact IBM support. If you already have advanced GPFS skills, this might be a good source of information.

9.2.5 Capturing GPFS traces

You might receive feedback from the GPFS L2 or L3 team to run a GPFS trace or **gpfs.snap** so that they can review traces on events to get detailed information and better understand issues. In this case, they might request that you activate GPFS tracing on the affected cluster.

> **Attention:** Use this command only under the direction of your IBM service representative, because it can affect the cluster performance.

On the active management node, enter the following commands:

```
#mmtracectl --set --trace=def --trace-recycle=global
#mmtracectl --start
```

The first command updates the GPFS configuration with trace parameters. The second command starts tracing on all nodes. Trace data is saved in the `/ftdc/dumps/gpfs` directory on each node. These traces do not get listed in the GUI download list, and must be manually collected with root-level privileges.

Leave these settings in place until you have another event (Network File System (NFS) exports failover not working, clustered trivial database (CTDB) flapping, GPFS-related performance issue, and so on). When you have an event that is captured by the trace, you *must* turn tracing off by using the following **`--stop`** command (do *not* run traces unless directed to do so by support):

```
#mmtracectl --stop
```

When tracing has stopped, run the following command:

```
#gpfs.snap
```

The snap capture with the current GPFS data and the trace files that were generated are found in the `/tmp/gpfs.snapOut/` directory. Upload the snap data `.tar` file to the IBM-directed File Transfer Protocol (FTP) server.

After the cause of any issues is determined, you can run the following command to clear out the trace parameters:

```
#mmtracectl --off
```

> **Attention:** An **`mmtrace`** is not a trivial data collection, and must be correctly stopped and shut off when collection is completed. Leaving it running under heavy workloads too long can create stability issues with SONAS node devices. You must run *both* of the following commands to correctly shut down the GPFS traces:
>
> ```
> #mmtracectl --stop
> #mmtracectl --off
> ```

9.2.6 Recovering from a failed file system

IBM authorized service providers can use this procedure to recover a GPFS file system after a storage unit failure has been fully addressed. Such an occurrence is rare, and is unlikely to occur. You can use this procedure to become familiar with the concepts of file system recovery if such a need is driven by IBM SONAS support.

> **Important:** The IBM 2851-DR1/DE1 information applies to hardware configurations purchased before the 1.4 release. As of SONAS 1.4, for any new configurations that are sent from manufacturing, DR1/DE1s have been withdrawn and replaced by DR2/DE2 storage models.

This task has the following prerequisites:

- You must be running this procedure on one of the storage nodes.
- All storage node recovery steps must be completed.
- The storage node must be fully functional.
- All storage devices must be available.

For storage node recovery, see *Troubleshooting the System x3650 server* and *Removing a node to perform a maintenance action* in the SONAS section of the IBM Knowledge Center. This procedure provides steps to recover a GPFS file system on a storage node after a total failure of the storage unit.

Note: The **mmnsddiscover** and **mmchdisk** commands are used in the following procedure.

An administrative user with the privileged role can use the **servicecmd** service command to run these commands. For more information about the **servicecmd** command, see *Service provider commands reference* in the SONAS section of the IBM Knowledge Center.

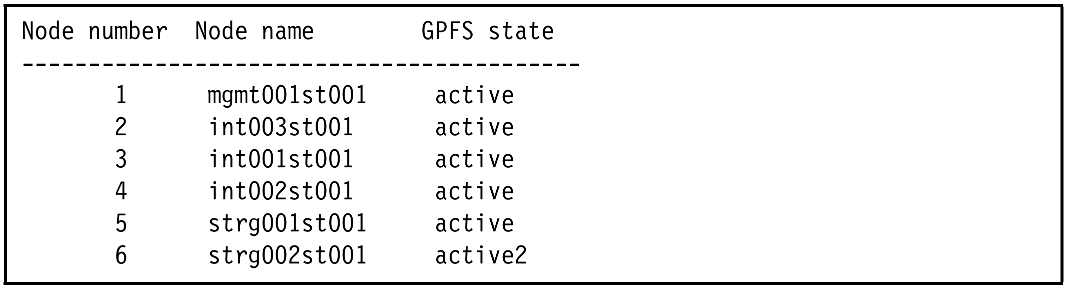
Important: Because, in this state, no I/O can be done by GPFS, it can be assumed for these procedures that a storage unit failure caused a GPFS file system to unmount.

After you ensure that the preceding prerequisites are met, complete the following steps:

1. Verify that GPFS is running across the IBM SONAS cluster by using the **mmgetstate** command:

   `#mmgetstate -a`

   The system displays information similar to that shown in Figure 9-9.



```
Node number  Node name        GPFS state
------------------------------------------
       1      mgmt001st001     active
       2      int003st001      active
       3      int001st001      active
       4      int002st001      active
       5      strg001st001     active
       6      strg002st001     active2
```

Figure 9-9 Sample information from the mmgetstate -a command

2. Verify that all storage devices are accessible to Linux by using the **multipath** command, which shows that all devices are active. In addition, verify that there are no faulty devices.
3. Run the **#multipath -ll** command.

   The system displays information similar to Example 9-8.

Example 9-8 Results of the multipath -ll command when run on a SONAS storage node

```
[root@yourmachine.strg001st001 ~]# multipath -ll
array1_sas_89360007 (360001ff070e9c0000000001989360007) dm-0 IBM,2851-DR1
[size=3.1T][features=1 queue_if_no_path][hwhandler=0][rw]
\_ round-robin 0 [prio=50][active]
 \_ 6:0:0:0 sdb 8:16  [active][ready]
\_ round-robin 0 [prio=10][enabled]
 \_ 8:0:0:0 sdg 8:96  [active][ready]
array1_sas_89380009 (360001ff070e9c0000000001b89380009) dm-2 IBM,2851-DR1
[size=3.1T][features=1 queue_if_no_path][hwhandler=0][rw]
\_ round-robin 0 [prio=50][active]
 \_ 6:0:0:2 sdd 8:48  [active][ready]
\_ round-robin 0 [prio=10][enabled]
 \_ 8:0:0:2 sdi 8:128 [active][ready]
```

4. With GPFS functioning normally on all nodes in the cluster, ensure that GPFS detects the devices with the **mmnsddiscover** command:

   `#mmnsddiscover -a -N all`

   If GPFS is *not* functioning normally on all nodes, run the **mmnsddiscover** command with the list of functioning nodes:

   `#mmnsddiscover -N <node>,...`

   In this example, <*node*> lists the names of all of the functioning nodes.

   The system displays information similar to Figure 9-10.

```
mmnsddiscover:  Attempting to rediscover the disks.  This may take a while
...
mmnsddiscover:  Finished.
```

Figure 9-10 Sample mmnsddiscover command output

5. Run the **mmlsnsd** command and verify that system devices are listed under the Device column in the output. Devices should have names that begin with `/dev`.

   If device names do not display, GPFS cannot access the devices.

   > **Note:** This process can take several minutes to complete.

   The system displays information similar to Figure 9-11.

```
#mmlsnsd -M

Disk name    NSD volume ID      Device         Node name               Remarks
---------------------------------------------------------------------------------------
 array0_sas_888f0013 AC1F86024B4E2048   /dev/mpath/array0_sas_888f0013 strg001st001 server node
 array0_sas_888f0013 AC1F86024B4E2048   /dev/mpath/array0_sas_888f0013 strg002st001 server node
 array0_sas_88910015 AC1F86024B4E2044   /dev/mpath/array0_sas_88910015 strg001st001 server node
```

Figure 9-11 Sample output for the mmlsnsd -M command

6. Run the **mmlsdisk** command to display the state of the disks:

   `#mmlsdisk fs_name`

   The system displays information that is similar to Figure 9-12.

```
    disk         driver   sector failure holds    holds                             storage
    name         type       size   group metadata data  status        availability pool
    ------------ -------- ------ ------- -------- ----- ------------- ------------ ------------
    array0_sas_888f0013 nsd         512       2 yes      yes   ready         up           system
    array0_sas_88910015 nsd         512       2 yes      yes   ready         up           system
    array0_sas_88930017 nsd         512       2 yes      yes   ready         up           system
```

Figure 9-12 Sample output for the mmlsdisk fs_name command

7. If all of the disks have the availability status `up`, go to the next step.

   Otherwise, you must run the **mmchdisk** command:

   `#mmchdisk fs_name start -a`

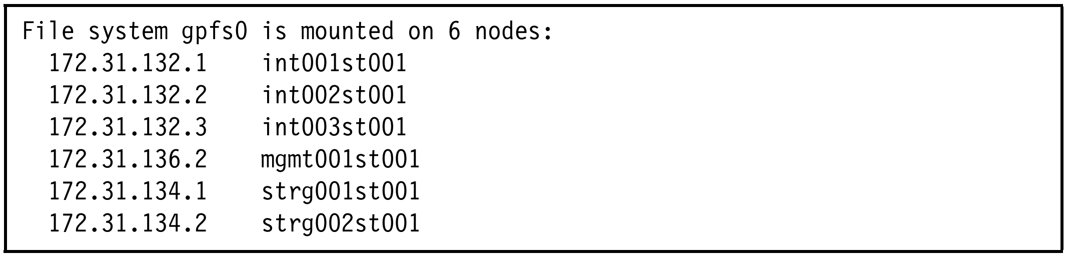
Important: If you need to run the **mmchdisk** command, be sure to rerun the **mmlsdisk** command to verify the availability of all of the disks before you go to the next step.

This process can take several minutes to complete.

8. Verify that the node has a file system mounted by running the **mmlsmount** command:

   ```
   #mmlsmount fs_name -L
   ```

   The system displays information similar to Figure 9-13.



```
File system gpfs0 is mounted on 6 nodes:
  172.31.132.1    int001st001
  172.31.132.2    int002st001
  172.31.132.3    int003st001
  172.31.136.2    mgmt001st001
  172.31.134.1    strg001st001
  172.31.134.2    strg002st001
```

Figure 9-13 Sample output for the mmlsmount fs_name -L command

   If the file system is not mounted, skip the next step and go to step 10. Otherwise, continue with the next step.

9. Check the Linux system log (`/var/log/messages`) on all nodes in the cluster for the presence of `MMFS_FSSTRUCT` errors.

   To complete this process, you might need to restart the node where the GPFS file system is mounted. You can often avoid a node restart by correctly halting the following processes or services. Even after stopping all the processes and services, the file system can remain mounted:

   a. Stop NFS and CIFS servers by using the **service nfs stop** and **service smb stop** commands.

   b. Stop the Tivoli Storage Manager and Hierarchical Storage Manager (HSM) processes by using the **disengages stop** command.

   c. Stop the Tivoli Storage Manager backup process. You can use the **stopbackup** command.

   d. Stop the Network Data Management Protocol (NDMP)-based backup.

   e. Stop the asynchronous replication.

   f. Before unmounting GPFS on all nodes, scan `/proc/mounts` for bind mounts that point to the lost GPFS. These mounts need to be unmounted. If the system still does not unmount, use the **mmshutdown** and **mmstartup** commands.

   g. Try to unmount the GPFS file system. You can use the **unmountfs** command.

   h. If it does not unmount, you can use the **fuser -m** command to determine what processes still hold open GPFS file descriptors. First are processes that just have a file open in GPFS, and second are processes that have their working directory in GPFS. If this is an interactive user shell, change the directory out of GPFS.

10. Run the **fuser** command on all nodes in the cluster by using the following command:

    ```
    #cndsh -N all fuser -m <fs_name>ii
    ```

11. If all attempts to stop these services do not allow the file system to be unmounted, change the auto mounting of the file system to no, and restart the node. (See step 12):

```
#mmchfs <fs_name> -A no8
```

Issue the **mmfsck** command as follows:

```
#mmfsck fs_name -v -n > /tmp/mmfsck_no.out 2>&1
```

Review the output file (/tmp/mmfsck_no.out) for errors.

If the file contains a message that Lost blocks were found, some missing file system blocks are normal. In this situation, go to step 13.

However, if the **mmfsck** command reports more severe errors, and you are certain that running the command does no further harm to the file system, continue with the following step. Otherwise, contact IBM support.

12. Check the messages output by running the **mmfsck** command and viewing the exit code. A successful run of the **mmfsck** command generates the exit code 0. Issue the **mmfsck** command and save the output to a file:

```
#mmfsck fs_name -v -y > /tmp/mmfsck_yes.out 2>&1
```

Immediately after the command completes, verify its exit status by entering echo $?. It should report the value 0.

> **Verify:** If the exit value is not 0, continue to run the command until it reports the value 0. Also, check the output file /tmp/mmfsck_yes.out and verify that the **mmfsck** command reports that all errors were corrected.
>
> **Important:** On a large file system, the **mmfsck** command with the **-y** option can require several hours to run.

13. If the **mmfsck** command completes successfully, you can mount the file system across the cluster by using the IBM SONAS GUI, or you can issue the **mmmount** command:

```
#mmmount fs_name -a
```

> **Note:** For more information about recovering from unmounted file systems, see the *Resolving problems with missing mounted file systems* topic in the SONAS section of the IBM Knowledge Center:
>
> http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/trbl_gpfs_mount_resolving_missing.html

14. If you had to use the **mmchfs** command in step 11 to disable the automatic mounting of the file system when GPFS is started, undo that change now by running the following command:

```
#mmchfs fs_name -A yes
```

9.2.7 Troubleshooting authentication issues

Use the following commands for basic authentication troubleshooting:

1. Do a general check:

   ```
   #chkauth -c <cluster-id>
   ```

2. Check on a specific user:

   ```
   #chkauth -c <cluster-id> -i -u <sonas-domain-name>\\<User-name>
   ```

3. Ping the authentication server:

   ```
   #chkauth -c <cluster-id> -p
   ```

4. Test the secrets:

   ```
   #chkauth -c <cluster-id> -t
   ```

5. List the user on the Active Directory server:

   ```
   #chkauth -c <cluster-id> -u <sonas-domain-name>\\<User-name> -p <password>
   ```

9.3 Troubleshooting authentication issues

This section describes some methods to debug an authentication issue. If your issue is not similar to any of the issues described earlier in this chapter, use one of these commands to check what might be going wrong.

9.3.1 CLI commands to check

This section lists some of the commands that you can run to determine whether output is as expected, or if something is wrong.

List the authentication configured

Check the authentication that is configured by using the **lsauth** command. This command provides information about the configuration. You can check for the different parameters and determine if they are set correctly, depending on the authentication that is configured. Example 9-9 shows an example for Active Directory (AD) authentication. Similarly, for all other authentication methods, you can check for the parameters.

Example 9-9 Checking authentication that is configured on the cluster

```
# lsauth
AUTH_TYPE = ad
idMapConfig = 10000000-299999999,1000000
domain = SONAS
idMappingMethod = auto
clusterName = bhandar
userName = Administrator
adHost = SONAS-PUNE.SONAS.COM
passwordServer = *
realm = SONAS.COM
EFSSG1000I The command completed successfully.
```

Check the ID mapping for users and groups

Run the **chkauth** command to check the user information details, such as user ID (UID) and group ID (GID) for a user or group, as shown in Example 9-10.

Example 9-10 Check user information by using the chkauth command

```
# chkauth -c st002.vsofs1.com -i -u VSOFS1\\testsfuuser2 -p Dcw2k3dom01
Command_Output_Data     UID GID      Home_Directory            Template_Shell
FETCH USER INFO SUCCEED 250 10000011 /var/opt/IBM/sofs/scproot /usr/bin/rssh
```

Check for node synchronization

Run the **chkauth** command to check whether the nodes are in synchronization, as shown in Example 9-11.

Example 9-11 Check node synchronization using chkauth

```
# chkauth
ALL NODES IN CLUSTER ARE IN SYNC WITH EACH OTHER
EFSSG1000I The command completed successfully.
```

Check if a user is able to authenticate successfully

Run the **chkauth** command to check whether a user is able to authenticate with the authentication server, as shown in Example 9-12.

Example 9-12 Check if user can authenticate with server using ckhauth

```
# chkauth -c st002.vsofs1.com -a -u VSOFS1\\testsfuuser2 -p Dcw2k3dom01
Command_Output_Data       UID GID Home_Directory Template_Shell
       AUTHENTICATE USER SUCCEED
```

Check if the authentication server is reachable

Run the **chkauth** command to check whether the authentication server is reachable, as shown in Example 9-13.

Example 9-13 Check if authenticate server is reachable using ckhauth

```
# chkauth -c st002.vsofs1.com -p
Command_Output_Data                UID GID Home_Directory Template_Shell
       PING AUTHENTICATION SERVER SUCCEED
```

9.3.2 Logs to check

The logs contain useful information to help resolve errors and determine what has happened in the SONAS.

System logs

Check the system logs to see if there are any errors. The **lslog** CLI command displays the system log.

Audit logs

To check what commands recently ran, and the command parameters, you can run the **lsaudit** CLI command. This command shows all the commands that were run. You might want to see the sequence of commands run, whether any of them were incorrect, and so on.

9.3.3 More logs to collect and check

If the preceding logs do not help, you can contact IBM support. It is advisable to collect these logs, which help for further analysis or debugging:

- Samba debug 10 logs

  Run the following commands to collect the Samba logs:

  ```
  starttrace -cifs -client <client ip address>
  ```

  Re-create the issue on a Windows client, and then run the following command:

  ```
  stoptrace #traceid
  ```

  See the online SONAS documentation for the **starttrace** and **stoptrace** commands before you use them.

- UID and GID information for the user

  Along with the preceding logs, also collect UID and GID information for the users you see problems with. You can run the **chkauth** command to get the information:

  ```
  # chkauth -i -u <Username>
  ```

- Collect **cndump** information.

9.3.4 Active Directory users denied access due to mapping entry in NIS

Even after AD + Network Information Service (NIS) is successfully configured, some Active Directory users are denied access to the IBM SONAS share.

How to debug this issue

If, after you confirm that sufficient access control lists (ACLs) exist for data access, data is inaccessible, check if the UID for that user is correctly set in Active Directory. Use the following command to check whether the user or group has a UID or GID assigned:

```
#chkauth -i -u "<DOMAIN>\\<username>"
```

Example 9-14 shows that the **chkauth** command does not show the UID or GID for the user (in our example, autouser3). Even if the SONAS cluster can resolve the users on the domain controller, it can be that the user is present in the Active Directory server but not in the NIS server.

Example 9-14 Failed to get IDMapping for user

```
# chkauth -i -u "SONAS\\autouser3"
EFSSG0002C Command exception found: FAILED TO FETCH USER INFO
```

Conclusion

There are several possible reasons for this failure.

Access is denied for users who are present only in the AD server but not present in NIS server

There are multiple ways to resolve this issue:

- Define a corresponding user in the NIS server. See your internal process for creating a new user in NIS.
- Based on your requirements, use one of the three options available to configure the wanted option to `--userMap` while configuring authentication by using the **`cfgnis`** command. This is the default behavior for the **`cfgnis`** command. However, there is still a requirement for a valid user map for every valid user.

UNIX-style names do not allow spaces or special characters in the name

For mapping Active Directory users or groups to NIS users, consider the following conversion on the NIS server:

- Convert all uppercase characters to lowercase.
- Replace all blank spaces with underscores.

For information about special characters to be avoided in UNIX names, see the Limitations topics in the SONAS section of the IBM Knowledge Center:

http://www.ibm.com/support/knowledgecenter/STAV45/com.ibm.sonas.doc/adm_limitations.html

To correct this issue, define the corresponding user in NIS server as in the UNIX-style name conventions.

When the corresponding user is defined in the NIS server as in the UNIX-style name conventions, in this case `auto_user4`, you can use the **`chkauth`** command as shown in Example 9-15 to verify it.

Example 9-15 Successfully displays IDMapping for user

```
# chkauth -i -u "SONAS\\AUTO user4"
Command_Output_Data       UID   GID  Home_Directory           Template_Shell
FETCH USER INFO SUCCEEDED 21015 2100 /var/opt/IBM/sofs/scproot /usr/bin/rssh
EFSSG1000I The command completed successfully.
```

If the default, which is *deny* access, was used with `--userMap` in **`cfgnis`**, for those users who are not in NIS, or improperly placed in NIS, access is denied. However, if the `--userMap` option chosen was either `AD_domain:AUTO` or `AD_domain:DEFAULT` (with a specified `GUEST` account), those unmapped users are either provided auto-incremented UIDs by SONAS, or they are mapped to a `guest` user predefined in AD. In such cases, even though the user might be able to access the shares, they might not have the level of access that they expect.

Authentication preferred practices

The preferred practice is to define all of the users in NIS server with appropriate UNIX-style naming conventions for all of the AD users without uppercase, spaces, or special characters in their names. This needs to be done before storing any data or accessing any data in the cluster share.

9.3.5 NFS share troubleshooting

Sometimes NFS clients fail to mount NFS shares after a client IP change.

Use this information to resolve a `refused mount` or `Stale NFS file handle` response to an attempt to mount Network File System (NFS) shares after a client IP change.

After a client IP change, a **`df -h`** command returns no results, as shown in Example 9-16.

Example 9-16 Command issues with refused mounts or stale NFS file handles.

```
Filesystem            Size  Used Avail Use% Mounted on
      machinename: filename:  -     -     -   -  /sharename
Also, you can see the following error from the ls command:
ls: .: Stale NFS file handle
Also, the hosting node in the IBM SONAS system displays the following error:
 int002st001 mountd[3055867]: refused  mount request from hostname for sharename
(/): not  exported
```

If you receive one of these errors, perform the following steps:

1. Access the node that hosts the active management node role with root privileges. Then run the following command to flush the NFS cache in each node:

   ```
   onnode all /usr/sbin/exportfs -a
   ```

   Verify that the NFS mount is successful. If the problem persists, restart the NFS service on the node that is refusing the mount requests from that client.
2. Verify that the NFS share mount is successful.

9.3.6 CIFS share troubleshooting

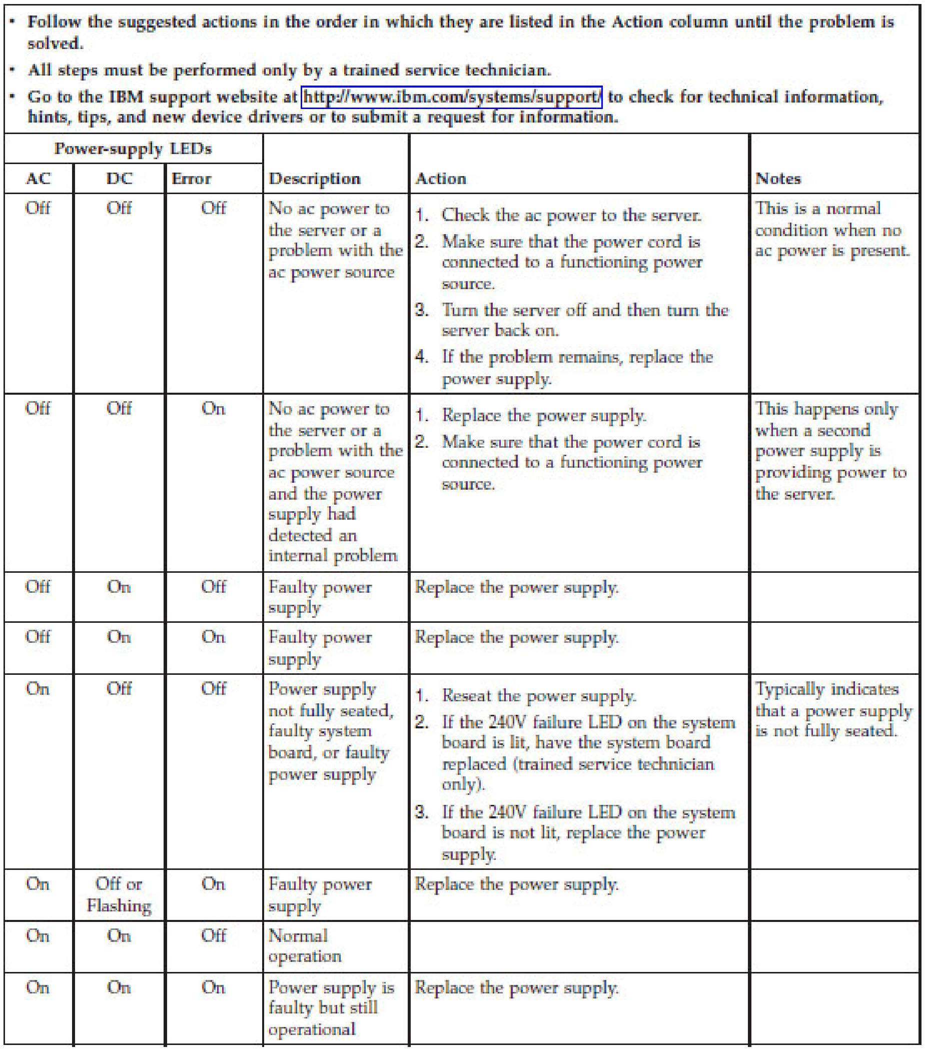
The following limitations do not affect many implementations. However, you might need to consider them if you are having trouble with CIFS shares:

- Alternative data streams are not supported. One example is an New Technology File System (NTFS) alternative data stream from a Mac OS X operating system.
- Server-side file encryption is not supported.
- Level 2 opportunistic locks (oplocks) are currently not supported. This means that level 2 oplock requests are not granted from SONAS CIFS shares.
- Symbolic links cannot be stored or changed and are not reported as symbolic links, but symbolic links that are created with NFS are respected as long they point to a target under the same exported directory.
- Server Message Block (SMB) signing for attached clients is not supported.
- Secure Sockets layer (SSL) communication to Active Directory is not supported.
- IBM SONAS acting as the distributed file system (DFS) root is not supported.
- Windows Internet Name Service (WINS) is not supported.
- Retrieving quota information using `NT_TRANSACT_QUERY_QUOTA` is not supported.
- Setting Quota information using `NT_TRANSACT_SET_QUOTA` is not supported.
- The IBM SONAS system does not grant durable or persistent file handles.

- CIFS UNIX Extensions are not supported.
- Managing the IBM SONAS system using the Microsoft Management Console Computer Management Snap-in is not supported, with the following exceptions:
  - Listing shares and exports
  - Changing share or export permissions
  - Users must be granted permissions to traverse all of the parent folders of an export to enable access to a CIFS export.

9.3.7 Power supply LEDs

It is understood that the system automatically detects issues with node hardware as they are introduced to the cluster. However, it is a preferred practice to have your data center representative perform a weekly walk-through to examine the SONAS frames and examine the cabinets for any error, such as light-emitting diode (LED) warnings. The SONAS node power supply LED indicator guide in Figure 9-14 can help you understand how to validate issues with node power supplies.



- **Follow the suggested actions in the order in which they are listed in the Action column until the problem is solved.**
- **All steps must be performed only by a trained service technician.**
- **Go to the IBM support website at http://www.ibm.com/systems/support/ to check for technical information, hints, tips, and new device drivers or to submit a request for information.**

Power-supply LEDs					
AC	**DC**	**Error**	**Description**	**Action**	**Notes**
Off	Off	Off	No ac power to the server or a problem with the ac power source	1. Check the ac power to the server. 2. Make sure that the power cord is connected to a functioning power source. 3. Turn the server off and then turn the server back on. 4. If the problem remains, replace the power supply.	This is a normal condition when no ac power is present.
Off	Off	On	No ac power to the server or a problem with the ac power source and the power supply had detected an internal problem	1. Replace the power supply. 2. Make sure that the power cord is connected to a functioning power source.	This happens only when a second power supply is providing power to the server.
Off	On	Off	Faulty power supply	Replace the power supply.	
Off	On	On	Faulty power supply	Replace the power supply.	
On	Off	Off	Power supply not fully seated, faulty system board, or faulty power supply	1. Reseat the power supply. 2. If the 240V failure LED on the system board is lit, have the system board replaced (trained service technician only). 3. If the 240V failure LED on the system board is not lit, replace the power supply.	Typically indicates that a power supply is not fully seated.
On	Off or Flashing	On	Faulty power supply	Replace the power supply.	
On	On	Off	Normal operation		
On	On	On	Power supply is faulty but still operational	Replace the power supply.	

Figure 9-14 Power supply LED fault light indicators

Related publications

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

IBM Redbooks

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

- *IBM SONAS Implementation Guide*, SG24-7962
- *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, on the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *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) 1.5.2 product documentation

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

- Support for SONAS

  http://www.ibm.com/support/entry/portal/overview/hardware/system_storage/network_attached_storage_(nas)/sonas/scale_out_network_attached_storage

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

IBM SONAS Best Practices

SG24-8051-00

ISBN 0738439770

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

SG24-8051-00

ISBN 0738439770

Printed in U.S.A.