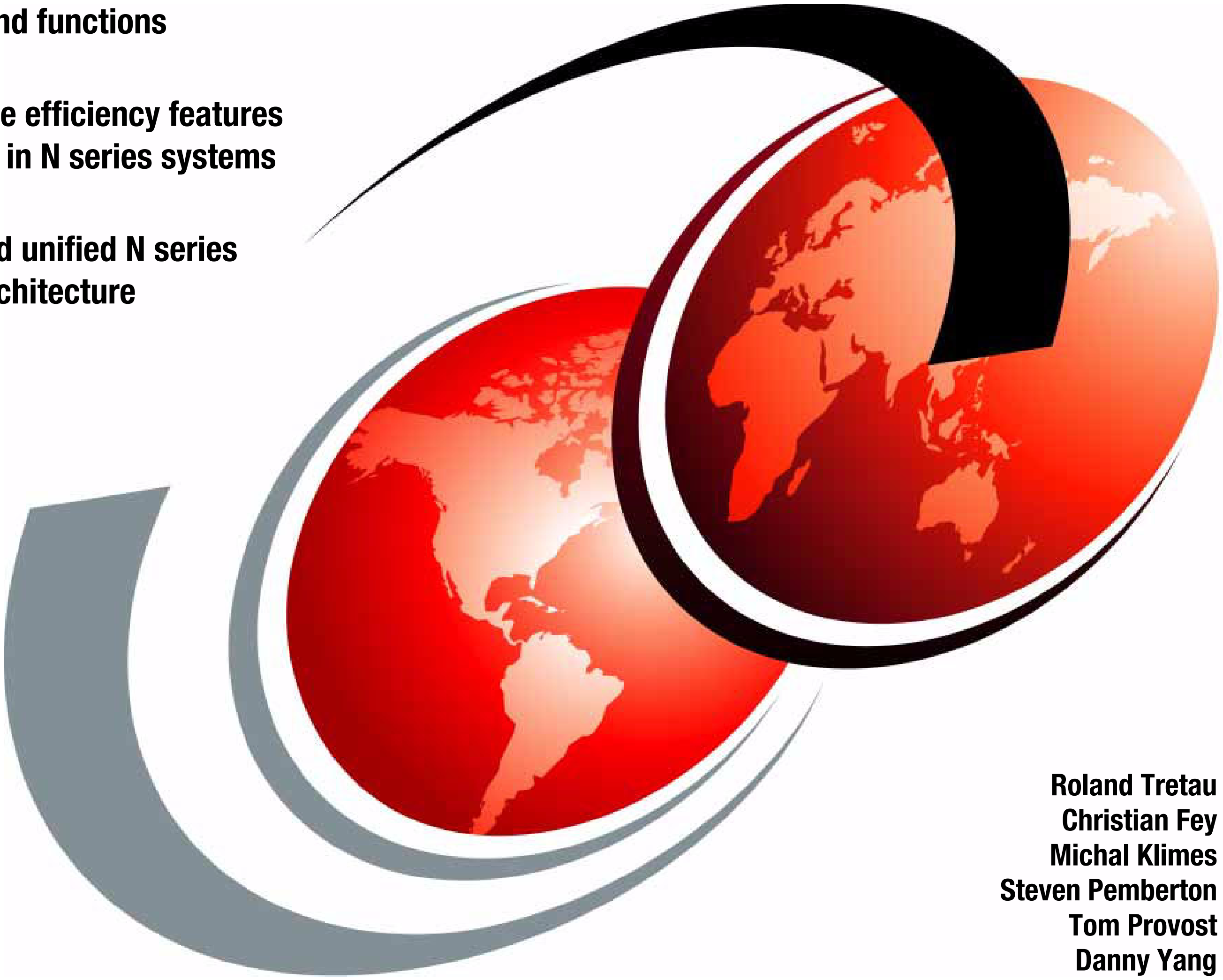
IBM System Storage N series Software Guide

Learn about Data ONTAP 8.2 7-mode features and functions

See storage efficiency features embedded in N series systems

Understand unified N series storage architecture



Roland Tretau
Christian Fey
Michal Klimes
Steven Pemberton
Tom Provost
Danny Yang

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM System Storage N series Software Guide

July 2014

SG24-7129-07

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

Eighth Edition (July 2014)

This edition applies to the IBM System Storage N series portfolio and Data ONTAP 8.2 as of October 2013.

Contents

Figures

Tables

Examples

Notices

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

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

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

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

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

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

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

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

AIX®
DB2®
Domino®
DS8000®
Enterprise Storage Server®
Global Technology Services®
GPFS™
IBM®
Lotus®
Lotus Notes®
Notes®
Real-time Compression™
Real-time Compression Appliance™
Redbooks®
Redpapers™
Redbooks (logo) ®
System Storage®
System x®
Tivoli®
XIV®

The following terms are trademarks of other companies:

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

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

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

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

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

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

Preface

Corporate workgroups, distributed enterprises, and small to medium-sized companies are increasingly seeking to network and consolidate storage to improve availability, share information, reduce costs, and protect and secure information. These organizations require enterprise-class solutions capable of addressing immediate storage needs cost-effectively, while providing an upgrade path for future requirements. Ideally, IT managers want a maximum degree of flexibility to design the architecture that best supports the requirements of multiple types of data and a broad range of applications. IBM® System Storage® N series storage systems and their software capabilities are designed to meet these requirements.

IBM System Storage N series storage systems offer an excellent solution for a broad range of deployment scenarios. IBM System Storage N series storage systems function as a multiprotocol storage device that is designed to allow you to simultaneously serve both file and block-level data across a single network. These activities are demanding procedures that, for some solutions, require multiple, separately managed systems. The flexibility of IBM System Storage N series storage systems, however, allows them to address the storage needs of a wide range of organizations, including distributed enterprises and data centers for midrange enterprises. IBM System Storage N series storage systems also support sites with computer and data-intensive enterprise applications, such as database, data warehousing, workgroup collaboration, and messaging.

This IBM Redbooks® publication explains the software features of the IBM System Storage N series storage systems. This book also covers topics such as installation, setup, and administration of those software features from the IBM System Storage N series storage systems and clients, and provides example scenarios.

This book is a companion to the Redbooks publication, *IBM System Storage N Series Hardware Guide*, SG24-7840, which can be found at the following website:

http://www.redbooks.ibm.com/abstracts/sg247840.html?Open

This book describes the software features on Data ONTAP 8.2 in 7-mode. For information about Cluster Mode, see the Redbooks publication, *IBM System Storage N series Clustered Data ONTAP*, SG24-8200, which can be found at the following website:

http://www.redbooks.ibm.com/abstracts/sg248200.html?Open

Authors

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

Roland Tretau is an Information Systems professional with more than 15 years of experience in the IT industry. He holds Engineering and Business Masters degrees, and is the author of many storage-related IBM Redbooks publications. Roland’s areas of expertise range from project management, market enablement, managing business relationships, product management, and consulting to technical areas including operating systems, storage solutions, and cloud architectures.

Christian Fey is a System Engineer working with IBM Premier Business Partner System Vertrieb Alexander GmbH (SVA) in Germany. His areas of expertise include IBM storage products in N series, IBM GPFS™ and SONAS environments, storage area networks, and storage virtualization solutions. He joined SVA in 2010.

Michael Klimes is an IT Specialist and team leader providing Level 2 and 3 support for IBM storage products in Czech Republic. His expertise spans all recent technologies of the IBM storage portfolio including tape, disk, SAN, and NAS technologies.

Steven Pemberton is a Senior Storage Architect with IBM GTS in Melbourne, Australia. He has broad experience as an IT solution architect, pre-sales specialist, consultant, instructor, and enterprise IT customer. He is a member of the IBM Technical Experts Council for Australia and New Zealand (TEC A/NZ), has multiple industry certifications, and is co-author of seven previous IBM Redbooks.

Tom Provost is a Field Technical Sales Specialist for the IBM Systems and Technology Group in Belgium. Tom has multiple years of experience as an IT professional providing design, implementation, migration, and troubleshooting support for IBM System x®, IBM System Storage, storage software, and virtualization. Tom also is the co-author of several other Redbooks publications and IBM Redpapers™.

Danny Yang is a Consulting PS Professional with IBM Global Technology Services® in IBM Korea. He joined IBM in 2004. He has worked in the South Korea Technical Support Group as a country storage Top-Gun for mid-range storage products since 2008. He has over 10 years of experience in designing and supporting of networks, Operating Systems (Linux, Windows), and storage products. He provides post-sales support for all of mid-range storage products such as N series, V7000 series, DS5000 series, SONAS, VTL, and TS3500 tape libraries.

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

- Bertrand Dufrasne, Uwe Heinrich Mueller, Uwe Schweikhard

  IBM
- Jacky Ben-Bassat, Craig Thompson

  NetApp

Now you can become a published author, too!

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-7129-07
for IBM System Storage N series Software Guide
as created or updated on July 31, 2014.

July 2014, Eighth Edition

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

- Information and changes in Data ONTAP 8.2 have been included.
- The N series hardware and software portfolios have been updated per the October 2013 status quo.

Changed information

- Hardware information for products no longer available has been removed.
- Information only valid for previous versions of Data ONTAP has been removed or modified to highlight differences and improvements in the current Data ONTAP 8.2 release.

For a complete list of the new and changes features in Data ONTAP 8.2, see the *IBM System Storage N series Data ONTAP 8.2 Release Notes for 7-Mode.*

http://www-01.ibm.com/support/docview.wss?crawler=1&uid=ssg1S7004582

Part 1

Introduction

In this part of the book, we introduce the IBM System Storage N series software. The IBM System Storage N series provide a range of reliable, scalable storage solutions for a variety of storage requirements. These capabilities are achieved by using network access protocols such as Network File System (NFS), Common Internet File System (CIFS), HTTP, FTP, Network Data Management Protocol (NDMP), and storage area network technologies, such as iSCSI, Fibre Channel over Ethernet (FCoE), and Fibre Channel Protocol (FCP).

Using built-in Redundant Array of Inexpensive Disks (RAID) technologies, all data is well protected, with options to enhance protection through mirroring, replication, Snapshots, and backup. These storage systems are also characterized by simple management interfaces that make installation, administration, and troubleshooting straightforward. The IBM System Storage N series is designed from the ground up as a stand-alone storage system.

Using this type of flexible storage solution provides the following advantages:

- Heterogeneous unified access for multiprotocol storage environments in a flexible management solution that can handle the unpredictable and explosive growth.
- Versatile, single integrated architecture designed to support concurrent block I/O and file servicing over Ethernet and Fibre Channel SAN infrastructures.
- Comprehensive software suite designed to provide robust system management, copy services, and virtualization technologies.
- Tuning the storage environment to a specific application while maintaining flexibility to increase, decrease, or change access methods with minimal or no disruption.
- Reacting easily and quickly to changing storage requirements. If additional storage is required, you can expand it quickly and non disruptively. When existing storage exists but is deployed incorrectly, you have the capability to reallocate available storage from one application to another quickly and simply.
- Maintaining availability and productivity during upgrades. If outages are necessary, keeping them as short as possible.
- Creating effortless backup/recovery solutions that operate commonly across all data access methods.

- Simplifying your infrastructure with file and block-level services in a single system.
- Changing the deployment of storage resources non disruptively, easily, and quickly. Online storage resource redeployment is possible.
- Implementing the upgrade process easily and quickly. Non-disruptive upgrade is possible.
- Using strong data protection solutions and support for online backup/recovery.

The following topics are covered within this part of the book:

- Overview
- Data ONTAP
- Write Anywhere File Layout
- Aggregates and volumes
- qtrees
- FlexClone volumes
- FlexCache volumes
- FlexShare
- Network configuration
- MultiStore

1

Overview

The IBM System Storage N series offers additional choices to organizations facing the challenges of enterprise data management. The IBM System Storage N series is designed to deliver high-end value with midrange affordability. Built-in enterprise serviceability and manageability features help to support customer efforts to increase reliability, simplify and unify storage infrastructure and maintenance, and deliver exceptional economy.

The following topics are covered:

- Introduction to features
- IBM System Storage N series hardware
- Software licensing structure
- Data ONTAP 8 supported systems

1.1 Introduction to features

In this section, we introduce the IBM System Storage N series and describe its hardware features. The IBM System Storage N series provides a range of reliable, scalable storage solutions for a variety of storage requirements. These capabilities are achieved by using network access protocols such as Network File System (NFS), Common Internet File System (CIFS), HTTP, FTP, and iSCSI, as well as storage area network technologies such as Fibre Channel (FC) and Fibre Channel Over Ethernet (FCoE).

Utilizing built-in Redundant Array of Independent Disks (RAID) technologies, all data is well protected, with options to enhance protection through mirroring, replication, Snapshots, and backup. These storage systems are also characterized by simple management interfaces that make installation, administration, and troubleshooting straightforward.

The N series unified storage solution supports file and block protocols as shown in Figure 1-1. Furthermore, converged networking is supported for all protocols.

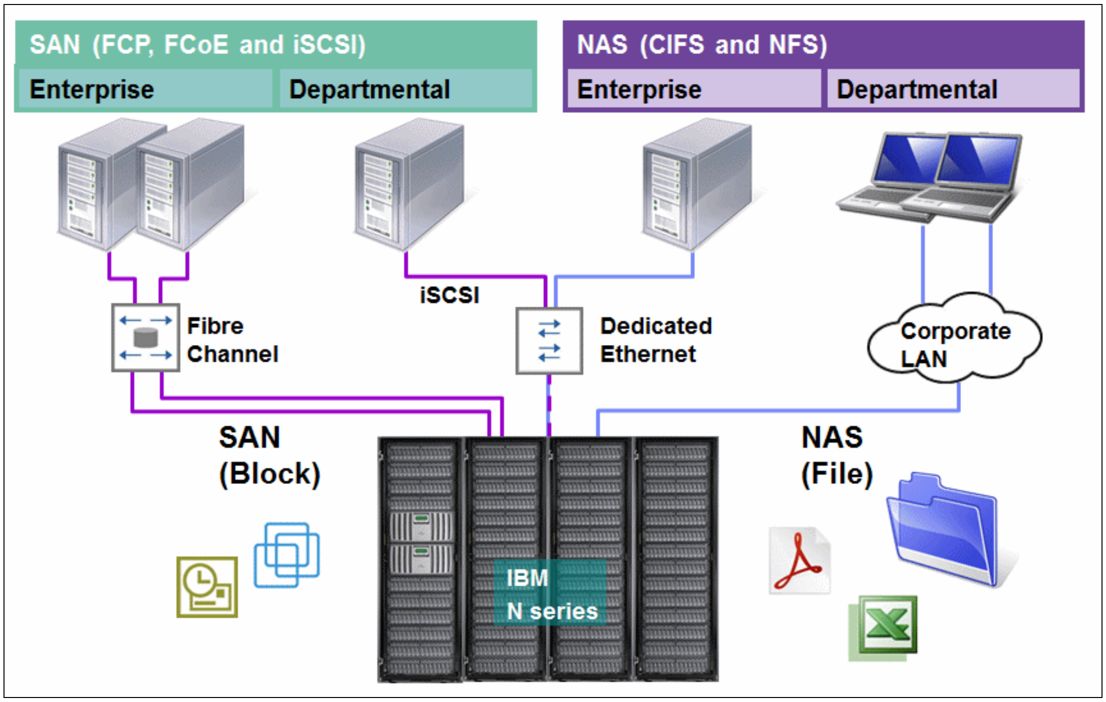


Figure 1-1 Unified storage

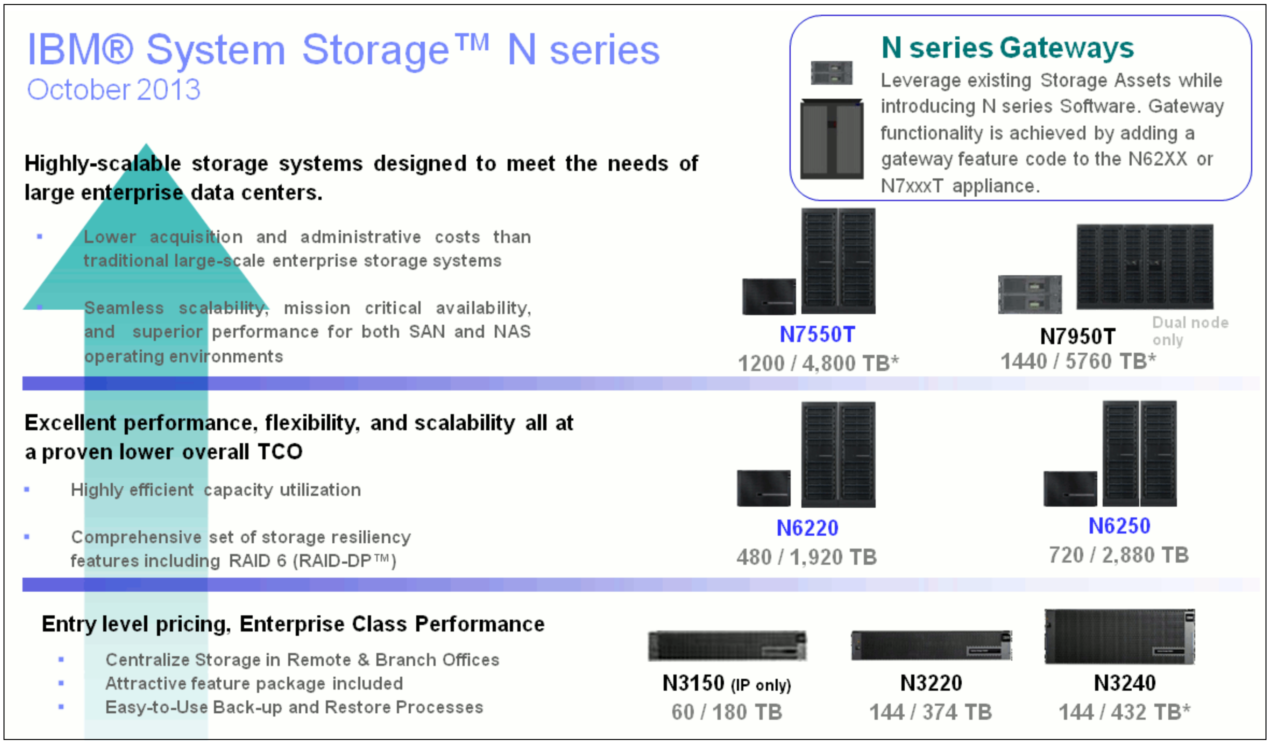
This type of flexible storage solution offers many benefits:

- Heterogeneous unified storage solution, with unified access for multi-protocol storage environments.
- Versatile-single integrated architecture, designed to support concurrent block I/O and file servicing over Ethernet and Fibre Channel SAN infrastructures.
- Comprehensive software suite, designed to provide robust system management, copy services, and virtualization technologies.
- Ease of changing storage requirements, allowing fast, dynamic changes. If additional storage is required, you can expand it quickly and non-disruptively. If existing storage is deployed incorrectly, you have the capability to reallocate available storage from one application to another quickly and easily.
- Maintain availability and productivity during upgrades. If outages are necessary, downtime is kept to a minimum.
- Easily and quickly implement non-disruptive upgrades.
- Create effortless backup and recovery solutions that operate in a common manner across all data access methods.
- Tune the storage environment to a specific application while maintaining its availability and flexibility.
- Change the deployment of storage resources easily, quickly, and non-disruptively. Online storage resource redeployment is possible.
- Achieve robust data protection with support for online backup and recovery.
- Include added value features such as deduplication to optimize space management.

All N series storage systems utilize a single operating system (Data ONTAP) across the entire platform, and offers advanced function software features that provide one of the industry's most flexible storage platforms. It includes comprehensive system management, storage management, onboard copy services, virtualization technologies, disaster recovery, and backup solutions.

1.2 IBM System Storage N series hardware

In the following sections, we describe the N series models available at the time of writing. Figure 1-2 identifies all the N series models released by IBM to date that belong to the N3000, N6000, and N7000 series line.



IBM® System Storage™ N series
October 2013

Highly-scalable storage systems designed to meet the needs of large enterprise data centers.

- Lower acquisition and administrative costs than traditional large-scale enterprise storage systems
- Seamless scalability, mission critical availability, and superior performance for both SAN and NAS operating environments

N series Gateways
Leverage existing Storage Assets while introducing N series Software. Gateway functionality is achieved by adding a gateway feature code to the N62XX or N7xxxT appliance.

N7550T
1200 / 4,800 TB*

N7950T
1440 / 5760 TB*
Dual node only

Excellent performance, flexibility, and scalability all at a proven lower overall TCO

- Highly efficient capacity utilization
- Comprehensive set of storage resiliency features including RAID 6 (RAID-DP™)

N6220
480 / 1,920 TB

N6250
720 / 2,880 TB

Entry level pricing, Enterprise Class Performance

- Centralize Storage in Remote & Branch Offices
- Attractive feature package included
- Easy-to-Use Back-up and Restore Processes

N3150 (IP only)
60 / 180 TB

N3220
144 / 374 TB

N3240
144 / 432 TB*

Figure 1-2 N series hardware portfolio

The following features and benefits are included:

- Data compression:
  - Provides transparent in-line data compression that can store more data in less space, reducing the amount of storage you need to purchase and maintain.
  - Reduces the time and bandwidth required to replicate data during volume SnapMirror transfers.
- Deduplication:
  - Performs block-level data deduplication on NearStore data volumes.
  - Scans and deduplicates volume data automatically, resulting in fast, efficient space savings with minimal impact on operations.
- Data ONTAP:
  - Provides full-featured and multiprotocol data management for both block and file serving environments through N series storage operating system.
  - Simplifies data management through single architecture and user interface, and reduces costs for SAN and NAS deployment.
- Disk sanitization:
  - Obliterates data by overwriting disks with specified byte patterns or random data.
  - Prevents recovery of current data by any known recovery methods.

- FlexCache:
  - Creates a flexible caching layer within your storage infrastructure that automatically adapts to changing usage patterns to eliminate bottlenecks.
  - Improves application response times for large compute farms, speeds data access for remote users, or creates a tiered storage infrastructure that circumvents tedious data management tasks.
- FlexClone:
  - Provides near-instant creation of LUN and volume clones without requiring additional storage capacity.
  - Accelerates test and development, and storage capacity savings.
- FlexShare:
  - Prioritizes storage resource allocation to highest-value workloads on a heavily loaded system.
  - Ensures that best performance is provided to designated high-priority applications.
- FlexVol:
  - Creates flexibly sized LUNs and volumes across a large pool of disks and one or more RAID groups.
  - Enables applications and users to get more space dynamically and non disruptively without IT staff intervention. Enables more productive use of available storage and helps improve performance.
- Gateway:
  - Supports attachment to IBM Enterprise Storage Server® (ESS) series, IBM XIV® Storage System, IBM System Storage IBM DS8000® and DS5000 series and a broad range of IBM, EMC, Hitachi, Fujitsu, and HP storage subsystems.
- MetroCluster:
  - Offers an integrated high-availability/disaster-recovery solution for campus and metro-area deployments.
  - Ensures high data availability when a site failure occurs.
  - Supports Fibre Channel attached storage with SAN Fibre Channel switch, SAS attached storage with Fibre Channel-SAS bridge, or Gateway storage with SAN Fibre Channel switch.
- MultiStore:
  - Partitions a storage system into multiple virtual storage appliances.
  - Enables secure consolidation of multiple domains and file servers.
- NearStore (near-line):
  - Increases the maximum number of concurrent data streams (per storage controller).
  - Enhances backup, data protection, and disaster preparedness by increasing the number of concurrent data streams between two N series systems.
- OnCommand:
  - Enables the consolidation and simplification of shared IT storage management by providing common management services, integration, security and role-based access controls delivering greater flexibility and efficiency.
  - Manages multiple N series systems from a single administrative console.
  - Speeds deployment and consolidated management of multiple N series systems.

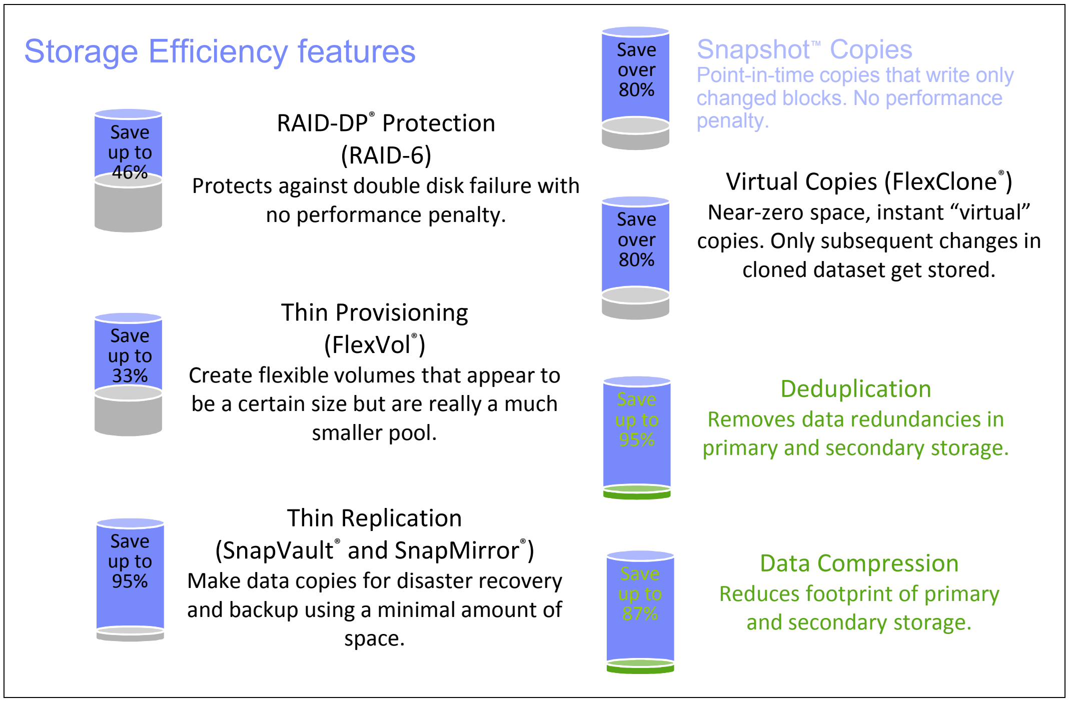
- Flash Cache (Performance Acceleration Module):
  - Improves throughput, and reduces latency for file services and other random read intensive workloads.
  - Offers power savings by consuming less power than adding more disk drives to optimize performance.
- RAID-DP:
  - Offers double parity bit RAID protection (N series RAID 6 implementation).
  - Protects against data loss due to double disk failures and media bit errors occurring during drive rebuild processes.
- SecureAdmin:
  - Authenticates both the administrative user and the N series system, creating a secure, direct communication link to the N series system.
  - Protects administrative logins, passwords, and session commands from cleartext snooping by replacing RSH and Telnet with the strongly encrypted SSH protocol.
- Single Mailbox Recovery for Exchange (SMBR):
  - Enables the recovery of a single mailbox from a Microsoft Exchange Information Store.
  - Extracts a single mailbox or email directly in minutes with SMBR, compared to hours with traditional methods, eliminating the need for staff-intensive, complex, and time-consuming Exchange server and mailbox recovery.
- SnapDrive:
  - Provides host-based data management of N series storage from Microsoft Windows, UNIX, and Linux servers.
  - Simplifies host-consistent Snapshot copy creation and automates error-free restores.
- SnapLock:
  - Write-protects structured application data files within a volume to provide WORM disk storage.
  - Provides storage enabling compliance with government records retention regulations.
- SnapManager:
  - Provides host-based data management of N series storage for databases and business applications.
  - Simplifies application-consistent Snapshot copies, automates error-free data restores, and enables application-aware disaster recovery.
- SnapMirror:
  - Enables automatic, incremental data replication between synchronous or asynchronous systems.
  - Provides flexible, efficient site-to-site mirroring for disaster recovery and data distribution.
- SnapRestore:
  - Restores single files, directories, or entire LUNs and volumes rapidly, from any Snapshot backup.
  - Enables near-instant recovery of files, databases and complete volumes.

- Snapshot:
  - Makes incremental, data-in-place, point-in-time copies of a LUN or volume with minimal performance impact.
  - Enables frequent, nondisruptive, space-efficient and quickly restorable backups.
- SnapVault:
  - Exports Snapshot copies to another N series system, providing an incremental block-level backup solution.
  - Enables cost-effective, long-term retention of rapidly restorable disk-based backups.
- Storage Encryption:
  - Provides support for Full Disk Encryption (FDE) drives in N series disk shelf storage and integration with License Key Managers, including IBM Tivoli® License Key Manager (TLKM).
- SyncMirror:
  - Maintains two online copies of data with RAID-DP protection on each side of the mirror.
  - Protects against all types of hardware outages, including triple disk failure.
- Gateway:
  - Reduce data management complexity in heterogeneous storage environments for data protection and retention.
- Software bundles:
  - Provides flexibility to take advantage of breakthrough capabilities, while maximizing value with a considerable discount.
  - Simplifies ordering of combinations of software features: Windows Bundle, Complete Bundle, and Virtual Bundle.

More details about N series hardware features can be found in the companion book, *IBM System Storage N series Hardware Guide*, SG24-7840. That IBM Redbooks publication can be found at the following website:

`http://www.redbooks.ibm.com/abstracts/sg247840.html?Open`

All N series systems support the storage efficiency features shown in Figure 1-3.



Storage Efficiency features

RAID-DP® Protection (RAID-6)
Protects against double disk failure with no performance penalty.

Thin Provisioning (FlexVol®)
Create flexible volumes that appear to be a certain size but are really a much smaller pool.

Thin Replication (SnapVault® and SnapMirror®)
Make data copies for disaster recovery and backup using a minimal amount of space.

Snapshot™ Copies
Point-in-time copies that write only changed blocks. No performance penalty.

Virtual Copies (FlexClone®)
Near-zero space, instant “virtual” copies. Only subsequent changes in cloned dataset get stored.

Deduplication
Removes data redundancies in primary and secondary storage.

Data Compression
Reduces footprint of primary and secondary storage.

Figure 1-3 Storage efficiency features

1.3 Software licensing structure

This section provides an overview of the software licensing structure.

1.3.1 What is new in Data ONTAP 8.2

Data ONTAP 8.2 introduces licensing changes that affect how you manage the licenses of your storage system:

- How licensed features are mapped to packages has changed.

  Data ONTAP feature licenses are issued as packages, each of which contains multiple features or a single feature. A package requires a license key, and installing the key enables you to access all features in the package.

  For example, installing the key for the SnapManager Suite package on a system entitles the system to use all SnapManager products in the package.

- Data ONTAP no longer displays license states for individual features that are part of a license package.

  Instead, Data ONTAP manages and displays entitlements at the license package level.

- A standard Data ONTAP license, also known as a node-locked license, is issued for a system with a specific serial number and is not valid on other systems.

  Data ONTAP 8.2 treats a license installed prior to upgrading to Data ONTAP 8.2 as a standard license.

- The length of a license key has been increased to 28 characters.

  Licenses installed in a release earlier than Data ONTAP 8.2 continue to work after you upgrade to Data ONTAP 8.2. However, if you need to reinstall a license in Data ONTAP 8.2 you must enter the license key in the new, 28-character format.

- Starting with Data ONTAP 8.2, the following functionality no longer requires a license, but you might need to enable certain options before using it:

  – High-availability (HA) configurations:

    You must set **`options cf.mode`** to **`ha`**.

  – MetroCluster:

    You must enable `options` **`cf.remote_syncmirror.enable`** and set **`options cf.mode`** to **`ha`**.

  – SyncMirror:

    There is no option that you need to enable.

- If you replace or upgrade a controller that is running Data ONTAP 8.2, within a grace period (up to 90 days), the new controller can use the same licensed functionality that the original controller had.

  During the grace period, you should contact your sales representative to obtain a proper license key to install on the new controller.

  During the grace period when the first valid license key is installed, you have 24 hours to complete license installation for all packages that you want the new controller to use. The grace period ends after the 24-hour period, and all previously installed licenses that were associated with the original system serial number will be removed from the new controller.

- Data ONTAP commands for managing licenses have been enhanced.

  The **`license add`** command now enables you to add multiple license keys at a time.

  The **`license show`** command output has a different format and displays additional details.

1.3.2 License packages

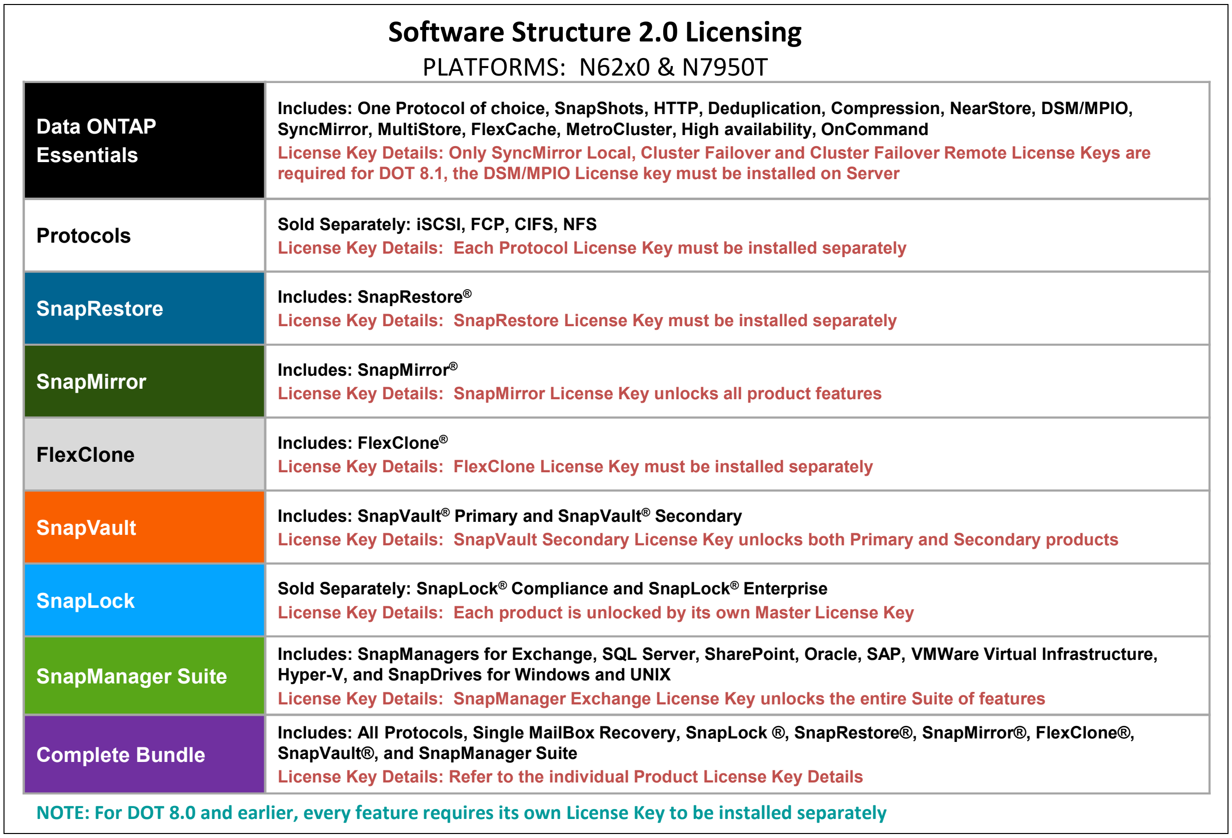
License packages are available for mid-range, high-end, and entry-level systems.

Mid-range and high-end systems

The software structure for mid-range and high-end systems is composed of eight major options:

- Data ONTAP Essentials (including one protocol of choice)
- Protocols (CIFS, NFS, FC, iSCSI)
- SnapRestore
- SnapMirror
- SnapVault
- FlexClone
- SnapLock
- SnapManager Suite

Figure 1-4 provides an overview of the software structure introduced with the availability of Data ONTAP 8.1.



Software Structure 2.0 Licensing

PLATFORMS: N62x0 & N7950T

Data ONTAP Essentials	Includes: One Protocol of choice, SnapShots, HTTP, Deduplication, Compression, NearStore, DSM/MPIO, SyncMirror, MultiStore, FlexCache, MetroCluster, High availability, OnCommand License Key Details: Only SyncMirror Local, Cluster Failover and Cluster Failover Remote License Keys are required for DOT 8.1, the DSM/MPIO License key must be installed on Server
Protocols	Sold Separately: iSCSI, FCP, CIFS, NFS License Key Details: Each Protocol License Key must be installed separately
SnapRestore	Includes: SnapRestore® License Key Details: SnapRestore License Key must be installed separately
SnapMirror	Includes: SnapMirror® License Key Details: SnapMirror License Key unlocks all product features
FlexClone	Includes: FlexClone® License Key Details: FlexClone License Key must be installed separately
SnapVault	Includes: SnapVault® Primary and SnapVault® Secondary License Key Details: SnapVault Secondary License Key unlocks both Primary and Secondary products
SnapLock	Sold Separately: SnapLock® Compliance and SnapLock® Enterprise License Key Details: Each product is unlocked by its own Master License Key
SnapManager Suite	Includes: SnapManagers for Exchange, SQL Server, SharePoint, Oracle, SAP, VMWare Virtual Infrastructure, Hyper-V, and SnapDrives for Windows and UNIX License Key Details: SnapManager Exchange License Key unlocks the entire Suite of features
Complete Bundle	Includes: All Protocols, Single MailBox Recovery, SnapLock ®, SnapRestore®, SnapMirror®, FlexClone®, SnapVault®, and SnapManager Suite License Key Details: Refer to the individual Product License Key Details

NOTE: For DOT 8.0 and earlier, every feature requires its own License Key to be installed separately

Figure 1-4 Software structure for mid-range and enterprise systems

As part of the new packaging effort, and to increase the business flow efficiencies, the 7-mode licensing infrastructure was modified to handle $0 features in a more bundled/packaged manner.

You no longer need to add license keys on your system for most features that are distributed free of cost. For some platforms, features in a given software bundle will only require one license key. Some features are enabled when you add certain other software bundle keys.

Entry-level

The entry-level software structure is very similar to the mid-range and high-end structure outlined in the previous section. The following changes apply:

- All protocols (CIFS, NFS; FC, iSCSI) are included with entry-level systems.
- The gateway feature is not available.
- The MetroCluster feature is not available.

1.3.3 Evaluation licenses

Evaluation license keys can be generated by IBM employees using an internal support process. IBM business partners must work with their IBM contacts to obtain “eval” keys for their customers.

The evaluation license keys provide approximately 90 days of feature use. You can extend an evaluation by requesting and installing a new eval key with a later expiration date.

Once the permanent license key has been purchased, simply install it on the controller to convert from the evaluation license to a production license.

1.4 Data ONTAP 8 supported systems

Table 1-1 provides an overview of systems that support Data ONTAP 8. The current models in N series product portfolio as of October 2013 as shown in grey. Some legacy N series model that are suitable to run Data ONTAP 8 are also shown.

Table 1-1 Supported Data ONTAP 8.x systems

Model	Supported by Data ONTAP versions 8.0 and higher			
	8.0	8.0.[123]	8.1.1	8.2
N3150			X	X
N3220			X	X
N3240			X	X
N3400	X	X	X	X
N5300	X	X	X	X
N5600	X	X	X	X
N6040	X	X	X	X
N6060	X	X	X	X
N6070	X	X	X	X
N6210		X	X	X
N6220			8.1.2 min	X
N6240		X	X	X
N6250			8.1.2 min	X
N6270		X	X	X
N7550T			8.1.2 min	X
N7600	X	X	X	X
N7700	X	X	X	X
N7800	X	X	X	X
N7900	X	X	X	X
N7950T		X	X	X

2

Data ONTAP

This chapter introduces Data ONTAP for the IBM System Storage N series, which is an operating system that has been specifically designed to provide data management tools and technologies in a network-oriented environment. This specialized operating system enables you to make full use of your investment in data and information, and gives you tools to support and manage your fast-growing data and information requirements.

The following topics are covered:

- Data ONTAP for IBM System Storage N series
- Data ONTAP overview
- Data ONTAP approach
- Data ONTAP architecture
- Data ONTAP 8.1 7-mode
- Data ONTAP 8.1 upgrades
- N series Data Motion

2.1 Data ONTAP for IBM System Storage N series

Data ONTAP provides the following benefits:

- Flexibility: Designed to support implementation of efficient storage environments that can adapt to the changing needs of the enterprise
- Ease of management: Enables implementation of customized policies for individual data sets to help simplify overall management and lower administrative burdens
- Enhanced data availability: Offers a range of options to enhance data availability, replication, and disaster recovery capabilities for single and multiple site operations, including clustering and mirroring capabilities
- Multiprotocol support: Supports data consolidation and data sharing

2.1.1 The challenge of managing explosive growth

In today's rapidly changing business climate, your enterprise demands cost-effective, flexible data management solutions that can help handle the unpredictable and explosive growth of storage in your heterogeneous environment. To enable enterprise-wide data management, support business continuance, and improve resource utilization, you need a flexible and scalable storage network solution, one that can also offer options to help reduce complexity and costs.

2.1.2 The solutions

IBM System Storage N series systems employ Data ONTAP, a highly scalable and flexible operating system that is designed to support the use of network filers, such as those in the IBM System Storage N series, in heterogeneous host environments. Data ONTAP offers flexible management and high availability options to support business continuance, thereby helping to reduce storage management complexity in your enterprise. Data ONTAP is designed for use in UNIX, Windows, and Web (http) environments, providing the foundation to build your storage infrastructure.

The innovative Data ONTAP architecture is designed to deliver scalable performance and support a flexible storage environment, allowing IBM System Storage N series systems to serve a range of needs, from small workgroups to enterprise data centers. N series systems can store and serve applications, consolidate data, and support reliable data access throughout the enterprise.

The Data ONTAP operating system can help simplify management and improve storage utilization by combining innovative file-system technology and a microkernel design to enable these features:

- Flexible data management: Data ONTAP supports the implementation of efficient storage environments with flexible volumes that do not require pre-partitioning (Figure 2-1). These capabilities can enable you to tailor data management to the requirements of each data set, respond quickly to changing needs of the enterprise, and help reduce implementation and management overhead.

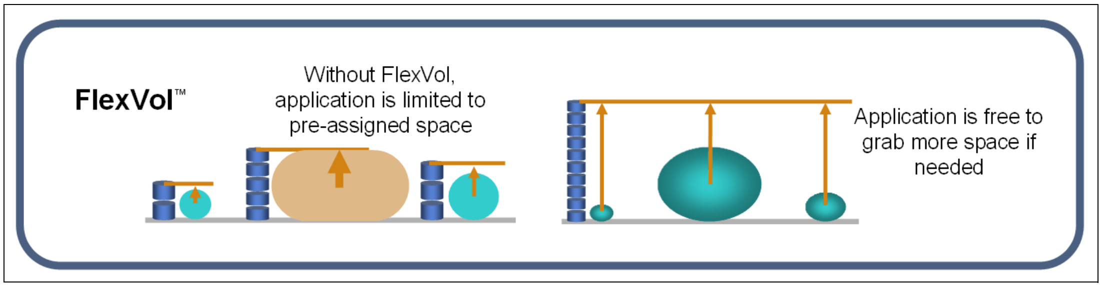

FlexVol™
Without FlexVol, application is limited to pre-assigned space
Application is free to grab more space if needed

Figure 2-1 Flexible volumes - get more space dynamically and non-disruptively

- Impressive scalability: Data ONTAP takes advantage of multiple processors to help deliver attractive performance. You can consolidate multiple terabytes of data onto a single appliance, making IBM System Storage N series systems a good fit for a variety of database, messaging, Web and other applications.
- Heterogeneous access: Data ONTAP provides data access using block-level and file-level protocols from the same hardware platform (Figure 2-2). IBM System Storage N series systems provide block-level data access over a Fibre Channel SAN fabric using FCP and over an IP-based Ethernet network using iSCSI. File access protocols such as NFS, CIFS, HTTP, or FTP provide file-level access over an IP-based Ethernet network. Additionally, SecureShare cross-protocol locking provides heterogeneous data sharing while supporting security, compatibility, and performance.

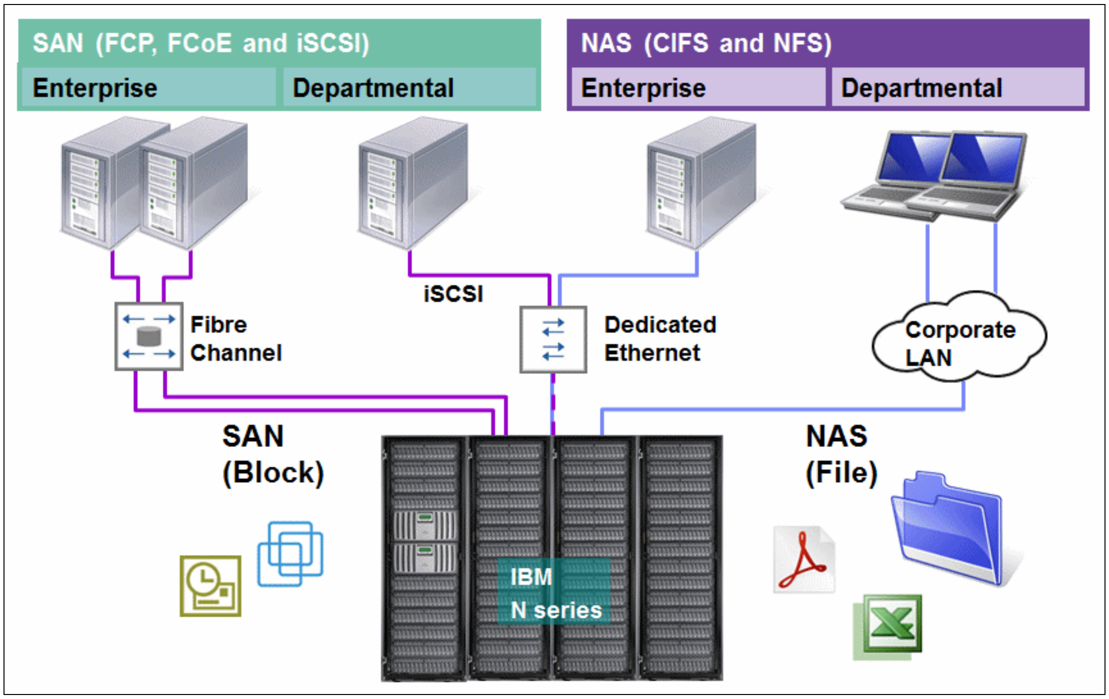


Figure 2-2 Block-level and file-level protocols from the same hardware platform

- Flexibility, availability, and reliability: IBM System Storage N series systems can help reduce costly downtime and improve access to mission-critical data. A combination of standard features and optional software capabilities allows Data ONTAP to support high availability for mission-critical applications. The following features are included.
  - The WAFL (Write Anywhere File Layout) file system: This feature supports a high-level of data availability while providing dynamic and flexible data storage volumes using FlexVol technology, as well as data protection using integrated, nonvolatile RAM and block-level checksum capability.
  - Support for business continuance and data retention: In today's business environment, lengthy disruptions to information access and noncompliance with records retention regulations are of increasing concern. Data ONTAP incorporates several innovative features to support disaster-tolerant data protection and recovery to help improve business continuance. It also provides disk-based, non-erasable, non-rewritable (WORM) data for the retention of reference data.
  - Integrated RAID: Data ONTAP is designed to provide cost-effective protection against disk failures and errors using double-parity RAID to help reduce disruption of service to users.
  - Lowering total cost of ownership: Data ONTAP can help reduce the complexity of deploying, administering and managing the storage infrastructure in your enterprise, enabling greater efficiency and productivity within your organization.
  - Easy to deploy: Data ONTAP can integrate into existing UNIX and Windows environments by utilizing standard naming and authentication services. Additionally, Data ONTAP incorporates a setup wizard to quickly complete basic configuration and installation.
  - Easy to manage: Wizard-based tools guide you through common management operations. DataFabric Manager supports centralized, multi-appliance management throughout your network.
  - Converged network support: N series systems provide full support for converged networks, also referred to as FCoE. The True end-to-end network conversion provides increased efficiency and simplified management and extends the unified architecture benefits as shown in Figure 2-3.

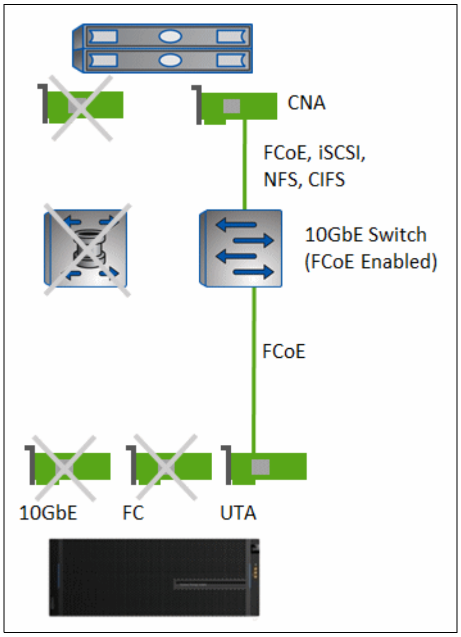


Figure 2-3 FCoE integration

2.2 Data ONTAP overview

Data ONTAP is a robust, tightly coupled, multi-tasking, real-time micro-kernel that minimizes complexity and improves storage system reliability. Data ONTAP has a look and feel similar to UNIX, but it is a proprietary kernel that is produced by NetApp, Inc.

This pre-tuned compact kernel minimizes complexity and improves reliability. In fact, Data ONTAP software is less than 2% of the total size of general-purpose operating systems. One of the real benefits of Data ONTAP is that by maintaining a lightweight workable size, the upgrades, maintenance, acquisition time, and complexity are reduced. Additional benefits of the kernel are as follows:

- No third-party application software is allowed to be installed on Data ONTAP, thereby reducing resource contention and application management impact.
- No third-party scripts or executables are allowed to execute against the kernel, thereby securing the kernel from malicious viruses or poor programming effects. In fact, all external operations are performed through the access services interface of Data ONTAP.
- IBM System Storage N series storage system side software (both standard and optionally enabled) is built directly into the kernel.

At the lowest level, the Data ONTAP kernel comprises three basic elements:

- A network interface driver
- A RAID manager
- The Write Anywhere File Layout (WAFL) file system

The Data ONTAP kernel includes these additional characteristics:

- Its own command set of familiar and unique commands
- A graphical user interface called FilerView
- Support of NTLM (NetLanMan)
- Data Encryption Standard (DES)
- Access services included in the kernel
- Autosupport for “phone home” to IBM with information and events

Designed with the goal of maximizing the throughput between network interfaces and disk drives, the Data ONTAP kernel utilizes the robust Write Anywhere File Layout file system. (See Figure 2-4 for more information about this topic.)

WAFL and RAID were designed together to avoid the performance problems that most file systems experience with RAID and to ensure the highest level of reliability. RAID is integrated into the WAFL file system (as opposed to other approaches, which have some type of volume manager on top of an operating system). Such a design reduces operator errors, operating system and application software release mismatches, patch level mismatches, and so on. This integration results in RAID acting as a performance accelerator for WAFL, rather than inhibiting performance.

2.3 Data ONTAP approach

The Data ONTAP approach also helps improve overall application availability, in that file system operations that normally run on general-purpose application file servers are no longer executed, thus improving general application server availability. It is a clear differentiation when compared with conventional storage subsystems. In these examples, the odds of application server downtime are increased due to the 100% dependency on the application server's operating system and file system software for all I/O operations.

It contrasts significantly with Data ONTAP deployment options, which allow for multiple application servers such that the failure of any one of those application servers does not preclude the other application servers from accessing the data. It is an added benefit that is not measured in N series storage system fault resilient availability.

The robust Data ONTAP software is based on a simple, message-passing kernel that has fewer failure modes than general-purpose operating systems.

2.4 Data ONTAP architecture

Data ONTAP comprises the following components:

- WAFL protection RAID and mirroring
- NVRAM management
- WAFL virtualization
- Snapshot management
- File services
- Block services
- Network layer
- Protocol layer
- System administration

Figure 2-4 illustrates the Data ONTAP architecture.

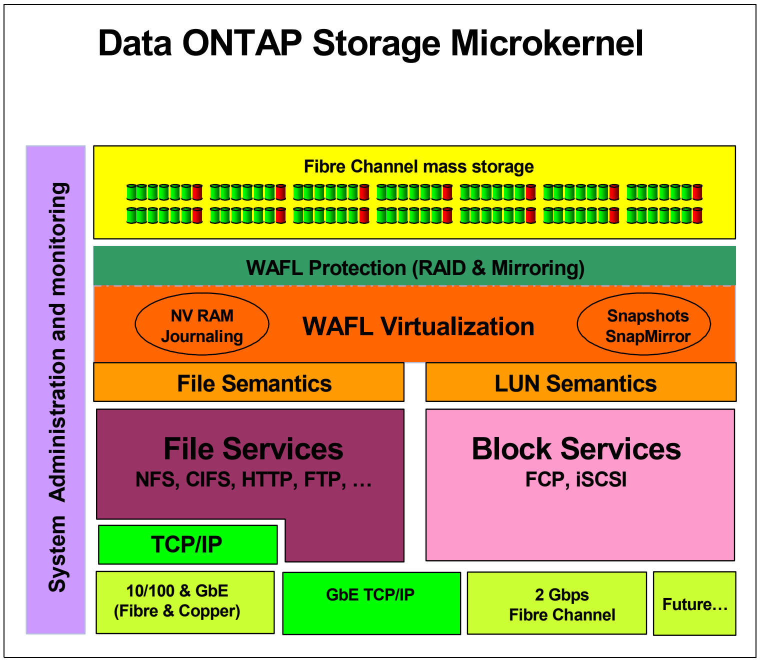


Figure 2-4 Data ONTAP architecture

2.4.1 The Network Interface driver

A Network Interface driver within Data ONTAP is responsible for receiving all incoming Network File System (NFS), Common Internet File System (CIFS), Fibre Channel Protocol (FCP), iSCSI, HTTP, and FTP requests. As each request is received, it is logged in non-volatile RAM (NVRAM). An acknowledgement is immediately sent back to the requestor, and any processing needed to satisfy the request is initiated. After being initiated, this processing runs uninterrupted (and continuously) as long as possible. This approach differs from that of traditional file servers, which employ separate processes for handling the network protocol stack, the remote file system semantics, the local file system, and the disk subsystem.

2.4.2 The RAID manager

Redundant Array of Inexpensive Disks (RAID) technology is designed to protect against loss of data in the event that disk failure occurs. Although RAID technology can be implemented in five different levels (each of which has advantages and disadvantages), levels 1, 3, and 5 are the most commonly used forms.

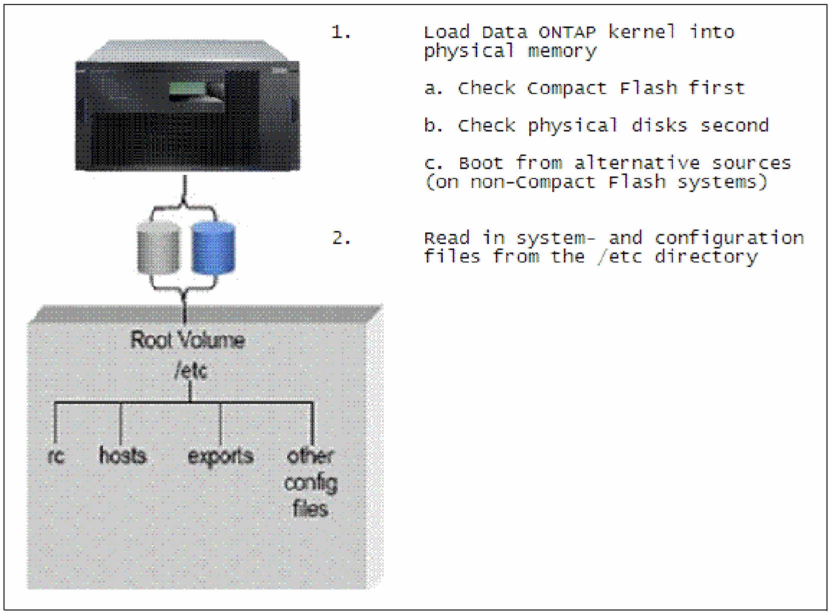
Data ONTAP stores data on disks in disk shelves connected to storage systems or uses storage on third-party storage arrays.

For native storage, Data ONTAP uses RAID-DP or RAID 4 groups to provide parity protection.

For third-party storage, Data ONTAP uses RAID0 groups to optimize performance and storage utilization. The storage arrays provide the parity protection for third-party storage. Data ONTAP RAID groups are organized into plexes, and plexes are organized into aggregates.

2.4.3 Data ONTAP startup

Data ONTAP itself resides on the compact flash and on each of the physical disks. Figure 2-5 shows the boot sequence of Data ONTAP.



1. Load Data ONTAP kernel into physical memory

   a. Check Compact Flash first

   b. Check physical disks second

   c. Boot from alternative sources (on non-Compact Flash systems)

2. Read in system- and configuration files from the /etc directory

Figure 2-5 Boot sequence of Data ONTAP

During this boot sequence, Data ONTAP checks the /etc directory to determine whether an installation was already done. Flash memory also holds a copy of the /etc directory.

2.5 Data ONTAP 8.1 7-mode

A number of features and enhancements have been added or changed in the Data ONTAP 8.1 7-Mode release family. Features that were present in the Data ONTAP 7.x release family are now supported in this release.

Attention:

- The following sections describe the enhancements that you can find in the new Data ONTAP 8.1 7-mode. For more information as well as in-depth explanations of these enhancements, see the *Data ONTAP 8.1 7-Mode Release Notes.*
- Some new and changed features in this release might also be introduced in a maintenance release of an earlier Data ONTAP release family. Before upgrading, be sure to consult with your IBM representative about new Data ONTAP functionality to determine the best solution for your business needs.

2.5.1 New terminology

This section describes changes in the new terminology introduced with DOT 8.x.

Cluster and high-availability terms

The following cluster and high-availability terms changed:

- Cluster: In the Data ONTAP 7.1 release family and earlier releases, refers to an entirely different functionality: a pair of storage systems (sometimes called nodes) configured to serve data for each other if one of the two systems stops functioning.
- HA (high availability): In Data ONTAP 8.x, refers to the recovery capability provided by a pair of nodes (storage systems), called an HA pair, that are configured to serve data for each other if one of the two nodes stops functioning.
- HA pair: In Data ONTAP 8.x, refers to a pair of nodes (storage systems) configured to serve data for each other if one of the two nodes stops functioning. In the Data ONTAP 7.3 and 7.2 release families, this functionality is referred to as an active/active configuration.
- CFO: The term is now used for controller failover rather than cluster failover.

New Data ONTAP terms

The following Data ONTAP terms changed:

- Interface groups (ifgrps): The naming convention for 802.3ad link aggregation was not consistent. In Data ONTAP GX, 802.3ad link aggregation was called *trunks*, while in Data ONTAP 7G link aggregation was called *vifs*. There is now one name, *ifgrps* (for interface groups), for both 7-mode and cluster-mode. Interface groups are the grouping of several physical ports together to provide increased aggregate bandwidth and redundancy.
- FreeBSD: FreeBSD is now the DataONTAP foundation.

2.5.2 New and changed platform and hardware support

Data ONTAP 8.1 7-mode now also provides the following support:

- Support for the N3220 and N3240 system models
- Support for SAS tape drives
- Support for FC/SAS bridge (ATTO for IBM FibreBridge 6500N)
- End of support for EXN2000 storage expansion units with ESH2 modules

2.5.3 Manageability enhancements

This Data ONTAP release provides additional management capabilities using MultiStore and other tools. This section provides an overview of these management capabilities:

- Cache rewarming for Flash Cache modules
- Reallocation scans, read reallocation, and extents with deduplicated volumes
- System health monitoring
- AutoSupport enhancements
- Default 64-bit root aggregate for new systems
- UPS management is no longer supported
- Changes to default aggregate Snapshot reserve for newly created nonmirrored aggregates
- IPv6 support for SP and RLM
- Support for Transport Layer Security protocol
- Licensing changes

2.5.4 Storage resource management enhancements

This Data ONTAP release provides improved performance, resiliency, and management capabilities for storage resources:

- Enhancements to aggregates
- N6210 now supports 500 FlexVol volumes
- Limit on number of subdirectories has been removed
- Enhancements to FlexClone files and FlexClone LUNs

2.5.5 High-availability pair enhancements

This Data ONTAP release includes new features and enhancements for high-availability. This section provides an overview of these features and enhancements:

- MetroCluster system now supports shared-switches configurations.

For more information about these features, see the *Data ONTAP 7-Mode High-Availability Configuration Guide.*

2.5.6 Networking and security protocol enhancements

This Data ONTAP release includes a number of new features and enhancements for networking and security protocol enhancements. This section provides an overview of these features and enhancements:

- Support for IPv6
- Improving TCP network congestion with Appropriate Byte Counting

2.5.7 File access protocol enhancements

This Data ONTAP release includes a number of new features and enhancements for file access and protocols management. This section provides an overview of these features and enhancements:

- Support for SFTP
- Support for SMB 2.0
- Support for FTPS

For more information about these features, see the *Data ONTAP 7-Mode File Access and Protocols Management Guide*.

2.5.8 Data protection enhancements

This Data ONTAP release includes a number of new features and enhancements for data protection enhancements. This section provides an overview of these features and enhancements:

- Support for volume SnapMirror replication between 32-bit and 64-bit volumes
- Changes to default FlexVol volume Snapshot reserve value
- Support for concurrent volume SnapMirror and SMTape backup operations
- Protection of data at rest through Storage Encryption
- Support for SnapLock

For more information about these features, see the Data ONTAP 7-Mode Data Protection Online Backup and Recovery Guide and the *Data ONTAP 7-Mode Data Protection Tape Backup and Recovery Guide.*

2.5.9 Storage efficiency enhancements

This Data ONTAP release includes a number of new features and enhancements for storage efficiency enhancements. This section provides an overview of these features and enhancements:

- Changes to deduplication
- Changes to data compression

2.5.10 MultiStore enhancements

This Data ONTAP release includes a number of new features and MultiStore enhancements:

- Online migration support for vFiler units
- Interactive SSH support for vFiler units
- Data compression support on vFiler units

2.6 Data ONTAP 8.1 upgrades

This section provides information about possible upgrade paths to Data ONTAP 8.1.

> **Important:** Prior to any update, check the release notes of the new code level to ensure that all your required features are available within the new release and that your system supports that specific upgrade.

Be aware that the licence structure of DataONTAP 8.1 has been changes compared to previous releases, for example, compression and deduplication are now volumes features and do not require a designated license anymore. See the release notes for full details. You have the following options:

- Upgrade from Data ONTAP 7.x/8.0 to 8.1:
  - Directly supported within Data ONTAP
  - ONTAP Essentials features not needed in 8.1 are accepted, and the legacy keys are used with a warning message.
- Revert from Data ONTAP 8.1 to 8.0/7.x:
  - Directly supported within Data ONTAP
  - Keys installed at time of upgrade will be available at time of revert. Required legacy licensing keys.
  - Customer does not need to acquire additional license keys, unless additional features were installed after the upgrade and before the revert.
- Downgrade from Data ONTAP 8.1 to 8.0/7.x (for systems that never had a DOT 8.0/7.x version installed):
  - Requires a system wipe and reinstall of ONTAP
  - All 8.0/7.x software pack license keys are available on original license key sheet provided with system and can be installed after downgrade

Starting with Data ONTAP 8.1, features that are $0 will show the keyword ENABLED instead of an actual license key.

For certain features, if you were not previously (prior to DOT 8.1) using them, upon upgrade to DOT 8.1, the feature is shown as ENABLED in the output of the license show command, but you need to take an additional step to “turn on” the feature.

To enable the following features, use the **`options`** command as shown here:

```
sanitization (licensed_feature.disk_sanitization.enable)
FCP (licensed_feature.fcp.enable)
FlexCache (licensed_feature.flexcache_nfs.enable )
iSCSi (licensed_feature.iscsi.enable)
Multistore (licensed_feature.multistore.enable)
Nearstore (licensed_feature.nearstore_option.enable)
```

2.7 N series Data Motion

With the release of Data ONTAP 7.3.3 and later, IBM now offers the IBM System Storage N series Data Motion feature.

2.7.1 Overview of Data Motion

Data Motion is a data migration solution that integrates virtual storage, mirroring, and provisioning software technologies so that you can perform migrations non-disruptively in both physical and virtual environments.

By using the Provisioning Manager interface, you can migrate data from one storage system to another, as long as the data is contained in vFiler units and associated with datasets.

Migration operations are performed transparently, so users are unaware of the migration operation being performed, and non-disruptively, so users retain access to migrated data, and the hosts and applications that access the migrated data do not require reconfiguration.

Figure 2-6 shows the Data Migration process where connected users can have their source data moved transparently without any outages.

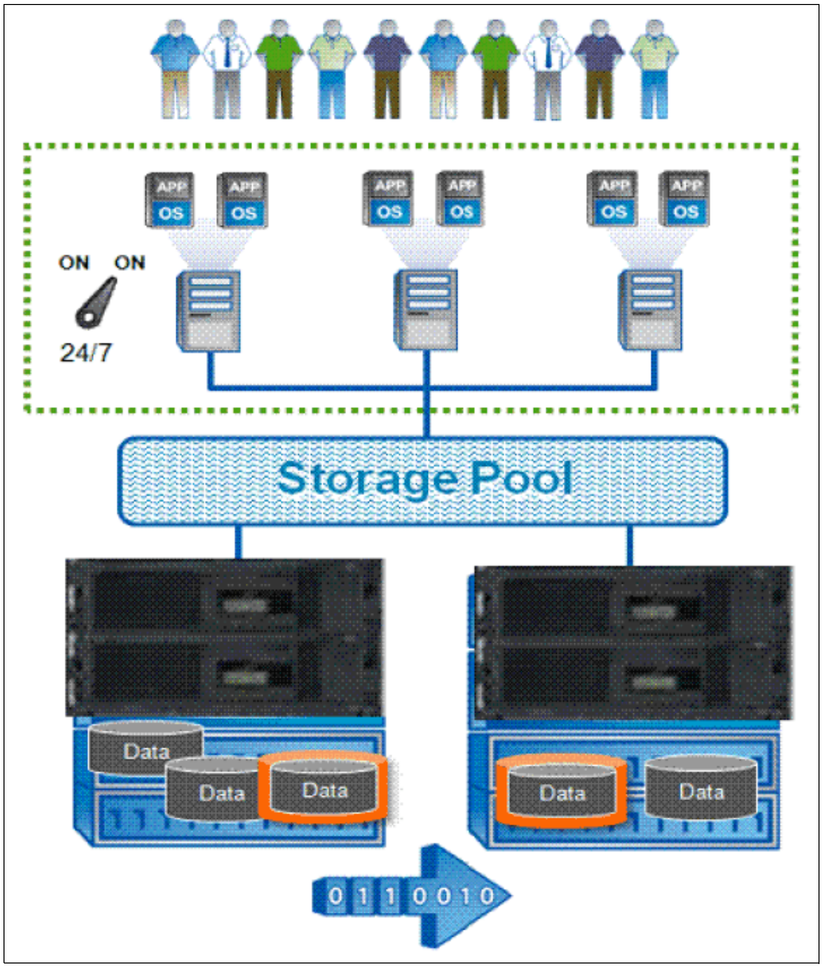


Figure 2-6 N series Data Motion or Data Migration

The application interfaces and documentation commonly refer to the Data Motion capability as “online migration,” “online dataset migration,” or “online vFiler unit migration.”

2.7.2 Business value of Data Motion

Data Motion significantly improves the availability of shared storage infrastructure by avoiding the service outages that are associated with planned activities, such as storage life-cycle management and cost/service-level optimization, thus helping customers to enable an always-on IT environment. The business values of Data Motion are as follows:

- No planned downtime:
  - For storage capacity expansion
  - For scheduled maintenance outages for moving data
  - For technology refresh
- Improved SLA flexibility:
  - On-demand load balancing
  - Adjustable storage tiers
- Application transparency:
  - No performance impact
  - Transaction integrity

For additional information about Data Motion, see the IBM Redbooks publication, *N series Data Motion*, SG24-7900, which can be found at the following website:

http://publib-b.boulder.ibm.com/abstracts/sg247900.html?Open

3

Write Anywhere File Layout

This chapter describes Write Anywhere File Layout (WAFL), which is a file system designed specifically to work in a file server appliance. Our primary focus in this chapter is on the algorithms and data structures that WAFL uses to perform its I/O and to implement *Snapshots* (read-only clones of the active file system). WAFL uses a unique copy-on-write technique to minimize the disk space that Snapshots consume. This chapter also describes how WAFL uses Snapshots to eliminate the need for file system consistency checking after an unclean shutdown.

The file system requirements for a file server storage system are different from those for a general-purpose UNIX/Windows system because a file server storage system must be optimized for network file access and because it must be easy to use.

The following topics are covered:

- Introduction to Write Anywhere File Layout
- Write Anywhere File Layout design
- File system consistency and non-volatile RAM
- Write allocation
- Summary

3.1 Introduction to Write Anywhere File Layout

An appliance is a device designed to perform a particular function. A recent trend in networking has been to provide common services using appliances instead of general purpose computers.

A new type of network appliance is the unified file server appliance. Traditionally, the N series based on WAFL started out as an NFS appliance. The requirements for a file system operating in an NFS appliance are different from those for a general purpose file system: NFS access patterns are different from local access patterns, and the special-purpose nature of an appliance also affects the design. Write Anywhere File Layout (WAFL) is the file system used in all Network Appliance Corporation's file servers.

WAFL was designed to meet four primary requirements:

- It must provide fast NFS service.
- It must support large file systems (tens of GB) that grow dynamically as disks are added.
- It must provide high performance while supporting Redundant Array of Independent Disks (RAID).
- It must restart quickly, even after an unclean shutdown due to power failure or system crash.

The requirement for fast NFS service is obvious, with WAFL's intended use in an NFS appliance. Support for large file systems simplifies system administration by allowing all disk space to belong to a single large partition. Large file systems make RAID desirable because the probability of disk failure increases with the number of disks. Large file systems require special techniques for fast restart because the file system consistency checks for normal UNIX file systems become unacceptably slow as file systems grow.

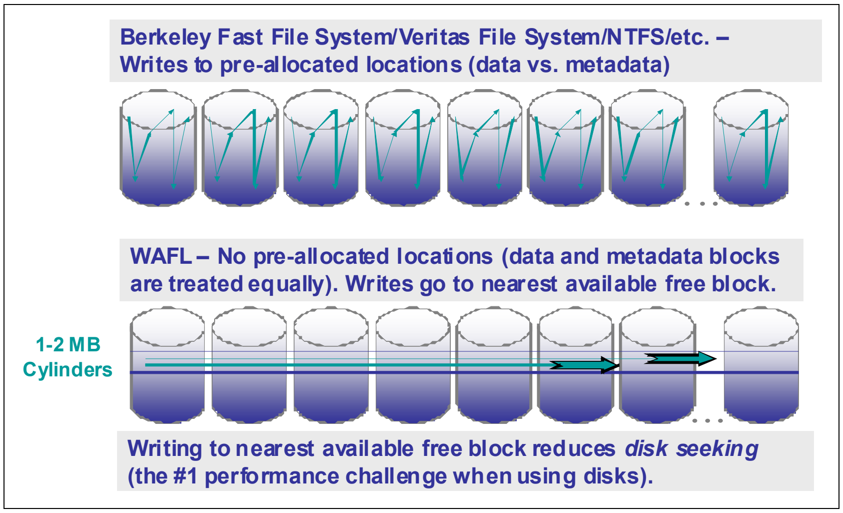
NFS and RAID both strain write performance: NFS because servers must store data safely before replying to NFS requests, and RAID because of the read-modify-write sequence it uses to maintain parity. It led us to use non-volatile RAM to reduce NFS response time and a write-anywhere design that allows WAFL to write to disk locations that minimize RAID's write performance penalty. The write-anywhere design enables Snapshots, which in turn eliminate the requirement for time-consuming consistency checks after power loss or system failure.

In later iterations, unified protocol access including file (NFS, CIFS) and block protocols (FC, iSCSI) have been added. Still, WAFL remains the underlying file system structure for the N series.

3.2 Write Anywhere File Layout design

WAFL is a compatible file system optimized for network file access. It is unique in that it stores sufficient information to make it compatible with a number of different client environments (NFS, CIFS, HTTP, and so on) and is optimized to maximize the reading and writing of disk content while supplying it to various types of network clients.

In many ways, WAFL is similar to other UNIX file systems (UFS), such as the Berkeley Fast File System (FFS) and the TransArc Episode file system (Figure 3-1). WAFL is a block-based file system that uses *inodes* to describe files (it stores all information about a file, directory, file system object except its data, and name).



Berkeley Fast File System/Veritas File System/NTFS/etc. – Writes to pre-allocated locations (data vs. metadata)

WAFL – No pre-allocated locations (data and metadata blocks are treated equally). Writes go to nearest available free block.

Writing to nearest available free block reduces *disk seeking* (the #1 performance challenge when using disks).

Figure 3-1 Write Anywhere File Layout comparison

3.2.1 WAFL overview

WAFL is a UNIX compatible file system optimized for network file access. In many ways WAFL is similar to other UNIX file systems such as the Berkeley Fast File System (FFS) and TransArc's Episode file system. WAFL is a block-based file system that uses inodes to describe files. It uses 4 KB blocks with no fragments.

Each WAFL inode contains 16 block pointers to indicate which blocks belong to the file. Unlike FFS, all the block pointers in a WAFL inode refer to blocks at the same level. Thus, inodes for files smaller than 64 KB use the 16 block pointers to point to data blocks. Inodes for files larger than 64 MB point to indirect blocks which point to actual file data. Inodes for larger files point to doubly indirect blocks. For very small files, data is stored in the inode itself in place of the block pointers.

Figure 3-2 illustrates inode space usage. Each inode contains 16 block pointers, meaning that a single inode can address a file smaller than or equal to 64 KB. If a file exceeds the 64 KB limit, metadata blocks are used to point to actual data, while small files (metadata) are stored directly in the inode file.

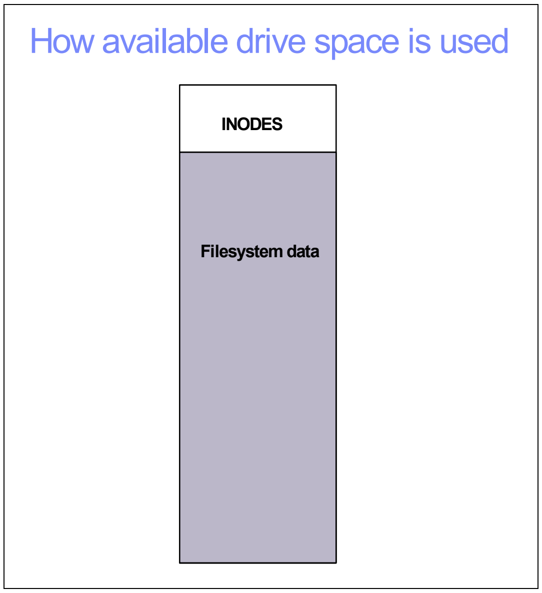


Figure 3-2 The inode space usage

3.2.2 Metadata resides in files

Like Episode, WAFL stores metadata in files. WAFL's three metadata files are the inode file, which contains the inodes for the file system; the block-map file, which identifies free blocks; and the inode-map file, which identifies free inodes. The term map is used instead of bit map because these files use more than one bit for each entry. The block-map file's format is described in detail next (see Figure 3-3).

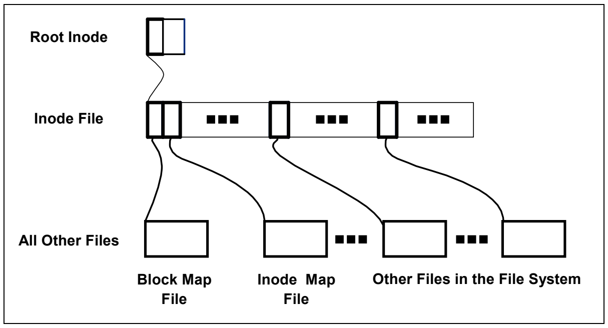


Figure 3-3 Metadata files with regular files underneath

Keeping metadata in files allows WAFL to write metadata blocks anywhere on disk. It is the origin of the name WAFL, which stands for Write Anywhere File Layout. The write-anywhere design allows WAFL to operate efficiently with RAID by scheduling multiple writes to the same RAID stripe whenever possible to avoid the 4-to-1 write penalty that RAID incurs when it updates just one block in a stripe.

Keeping metadata in files makes it easy to increase the size of the file system on the fly. When a new disk is added, the N series server automatically increases the sizes of the metadata files. The system administrator can increase the number of inodes in the file system manually if the default is too small. Finally, the write-anywhere design enables the copy-on-write technique used by Snapshots. For Snapshots to work, WAFL must be able to write all new data, including metadata, to new locations on disk, instead of overwriting the old data. If WAFL stored metadata at fixed locations on disk, it would not be possible.

3.2.3 A tree of blocks

A WAFL file system is best thought of as a *tree of blocks*. At the root of the tree structure is the root inode, as shown in Figure 3-3 on page 33. The *root inode* is a special inode that describes the inode file. The inode file contains the inodes that describe the rest of the files in the file system, including the block-map and inode-map files. The *leaves* of the tree are the data blocks of all the files.

Figure 3-4 here shows a more detailed version of Figure 3-3. It illustrates that files are made up of individual blocks, and that large files have additional layers of indirection between the inode and the actual data blocks. In order for WAFL to boot, it must be able to find the root of this tree, so the only exception to the WAFL write-anywhere rule is that the block containing the root inode must reside at a fixed location on disk where WAFL can find it.

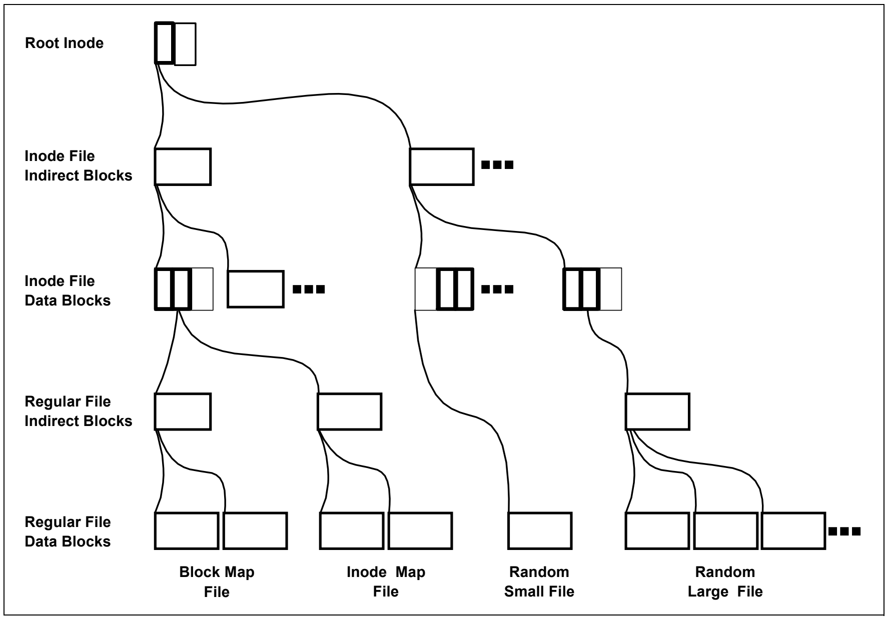


Figure 3-4 Detailed view of the WAFL tree of blocks

3.3 File system consistency and non-volatile RAM

WAFL avoids the need for file system consistency checking after an unclean shutdown by creating a special Snapshot called a consistency point every few seconds. Unlike other Snapshots, a consistency point has no name, and it is not accessible through NFS. However, like all Snapshots, a consistency point is a completely self consistent image of the entire file system. When WAFL restarts, it simply reverts to the most recent consistency point. It allows an N series server to reboot in about a minute even with 20 GB or more of data in its single partition.

Between consistency points, WAFL does write data to disk, but it writes only to blocks that are not in use, so the tree of blocks representing the most recent consistency point remains completely unchanged. WAFL processes hundreds or thousands of NFS requests between consistency points, so the on-disk image of the file system remains the same for many seconds until WAFL writes a new consistency point, at which time the on-disk image advances atomically to a new state that reflects the changes made by the new requests. Although this technique is unusual for a UNIX file system, it is well known for databases. Even in databases, it is unusual to write as many operations at one time as WAFL does in its consistency points.

WAFL uses non-volatile RAM (NVRAM) to keep a log of NFS requests it has processed since the last consistency point. (NVRAM is special memory with batteries that allow it to store data even when system power is off.) After an unclean shutdown, WAFL replays any requests in the log to prevent them from being lost. When an N series server shuts down normally, it creates one last consistency point after suspending NFS service. Thus, on a clean shutdown, the NVRAM does not contain any unprocessed NFS requests, and it is turned off to increase its battery life.

WAFL actually divides the NVRAM into two separate logs. When one log gets full, WAFL switches to the other log and starts writing a consistency point to store the changes from the first log safely on disk. WAFL schedules a consistency point every 10 seconds, even if the log is not full, to prevent the on-disk image of the file system from getting too far out of date.

Logging NFS requests to NVRAM has several advantages over the more common technique of using NVRAM to cache writes at the disk driver layer. Lyon and Sandberg describe the NVRAM write cache technique, which Legato's Prestoserve NFS accelerator uses.

Processing an NFS request and caching the resulting disk writes generally takes much more NVRAM than simply logging the information required to replay the request. For instance, to move a file from one directory to another, the file system must update the contents and inodes of both the source and target directories. In FFS, where blocks are 8 KB each, it uses 32 KB of cache space. WAFL uses about 150 bytes to log the information needed to replay a rename operation. Rename, with its factor of 200 difference in NVRAM usage, is an extreme case, but even for a simple 8 KB write, caching disk blocks will consume 8 KB for the dat a, 8 KB for the inode update, and for large files another 8 KB for the indirect block. WAFL logs just the 8 KB of data along with about 120 bytes of header information. With a typical mix of NFS operations, WAFL can store more than 1000 operations per megabyte of NVRAM.

Using NVRAM as a cache of unwritten disk blocks turns it into an integral part of the disk subsystem. An NVRAM failure can corrupt the file system in ways that **fsck** cannot detect or repair. If something goes wrong with WAFL's NVRAM, WAFL might lose a few NFS requests, but the on-disk image of the file system remains completely self consistent. It is important because NVRAM is reliable, but not as reliable as a RAID disk array.

A final advantage of logging NFS requests is that it improves NFS response times. To reply to an NFS request, a file system without any NVRAM must update its in-memory data structures, allocate disk space for new data, and wait for all modified data to reach disk. A file system with an NVRAM write cache does all the same steps, except that it copies modified data into NVRAM instead of waiting for the data to reach disk. WAFL can reply to an NFS request much more quickly because it need only update its in-memory data structures and log the request. It does not allocate disk space for new data or copy modified data to NVRAM.

3.4 Write allocation

Write performance is especially important for network file servers. Ousterhout observed that as read caches get larger at both the client and server, writes begin to dominate the I/O subsystem. This effect is especially pronounced with NFS which allows very little client-side write caching. The result is that the disks on an NFS server might have 5 times as many write operations as reads.

WAFL's design was motivated largely by a desire to maximize the flexibility of its write allocation policies. This flexibility takes three forms:

- WAFL can write any file system block (except the one containing the root inode) to any location on disk. In FFS, metadata, such as inodes and bit maps, is kept in fixed locations on disk. It prevents FFS from optimizing writes by, for example, putting both the data for a newly updated file and its inode right next to each other on disk. Because WAFL can write metadata anywhere on disk, it can optimize writes more creatively.
- WAFL can write blocks to disk in any order. FFS writes blocks to disk in a carefully determined order so that `fsck` can restore file system consistency after an unclean shutdown. WAFL can write blocks in any order because the on-disk image of the file system changes only when WAFL writes a consistency point. The one constraint is that WAFL must write all the blocks in a new consistency point before it writes the root inode for the consistency point.
- WAFL can allocate disk space for many NFS operations at once in a single write episode. FFS allocates disk space as it processes each NFS request. WAFL gathers up hundreds of NFS requests before scheduling a consistency point, at which time it allocates blocks for all requests in the consistency point at once. Deferring write allocation improves the latency of NFS operations by removing disk allocation from the processing path of the reply, and it avoids wasting time allocating space for blocks that are removed before they reach disk.

These features give WAFL extraordinary flexibility in its write allocation policies. The ability to schedule writes for many requests at once enables more intelligent allocation policies, and the fact that blocks can be written to any location and in any order allows a wide variety of strategies. It is easy to try new block allocation strategies without any change to WAFL's on-disk data structures.

The details of WAFL's write allocation policies are outside the scope of this paper. In short, WAFL improves RAID performance by writing to multiple blocks in the same stripe; WAFL reduces seek time by writing blocks to locations that are near each other on disk; and WAFL reduces head- contention when reading large files by placing sequential blocks in a file on a single disk in the RAID array. Optimizing write allocation is difficult because these goals often conflict.

3.5 Summary

WAFL was developed and became stable surprisingly quickly for a new file system. We attribute this stability in part to the WAFL use of consistency points. Processing file system requests is simple because WAFL updates only in-memory data structures and the NVRAM log. Consistency points eliminate ordering constraints for disk writes, which are a significant source of errors in most file systems. The code that writes consistency points is concentrated in a single file and interacts little with the rest of WAFL.

More importantly, it is much easier to develop high-quality, high-performance system software for an appliance than for a general-purpose operating system. Special purpose file systems also have difficulty achieving good performance and reliability because they are often hosted on general-purpose platforms, which limits their efficiencies and reliability.

Compared with a general-purpose file system, WAFL handles a regular and simple set of requests. A general-purpose file system receives requests from thousands of different applications with a wide variety of different access patterns, and new applications are added frequently. By contrast, WAFL receives requests only from the network-attached storage or SAN client modules of other systems that have been implemented following a strict regime of industry-developed protocol definitions. iSCSI, NFS, FTP, and HTTP all must function the same regardless of which platform they are running on because the protocol that they follow is well constructed. CIFS is only available from a single source, so it too is well constrained.

Of course, applications are the ultimate source of I/O requests, but the client code converts application requests into a regular pattern of network requests, and it filters out error cases before they reach the server. The small number of operations that WAFL supports makes it possible to define and test the entire range of inputs that it is expected to handle.

These advantages apply to any IBM System Storage N series, not just to file server appliances. Network-attached storage only makes sense for protocols that are well defined and widely used, but for such protocols, network-attached storage can provide important advantages over a general-purpose computer.

4

Aggregates and volumes

This chapter introduces the topic of IBM System Storage N series and aggregates, and explains the concept of volumes.

The following topics are covered:

- Overview of aggregates
- Aggregates and the IBM N series storage hierarchy
- Introduction to 64-bit aggregates
- Introduction to flexible volumes

4.1 Overview of aggregates

An aggregate is a collection of physical disk space used as a container, depending on whether you want to take advantage of RAID-level mirroring and the physical layer. If the aggregate is unmirrored, it only contains a plex. A *plex* is a collection of one or more RAID groups that together provide the storage for one or more Write Anywhere File Layout (WAFL) file system volumes. If the SyncMirror feature is licensed and enabled, Data ONTAP adds a second plex to the aggregate, which serves as a RAID-level mirror for the first plex in the aggregate.

When you create an aggregate, Data ONTAP assigns data disks and parity disks to RAID groups depending on the options you choose, such as the size of the RAID group or the level of RAID protection.

Each aggregate possesses its own RAID configuration and set of assigned disks. Within each aggregate, you can create one or more FlexVol volumes. You can increase the usable space in an aggregate by adding disks to existing RAID group or by adding new RAID groups.

Table 4-1 lists the limits of aggregates and volumes.

Table 4-1 Aggregate and volume limits

Items	Data ONTAP 7G	Data ONTAP 8G
Maximum aggregate size	16 TB	100 TB
Maximum FlexVolume size	16 TB	100 TB

Reference: Aggregate and volume maximums depend on the product where Data ONTAP is installed. For additional information, see the *IBM System Storage N series Data ONTAP 8.1 7-Mode Storage Management Guide*, available at the following website:

`http://www.ibm.com/storage/support/nas`

4.1.1 What is new in 8.2

Data ONTAP 8.2 supports several enhancements to aggregates and volumes:

- There is an increased maximum aggregate size for some platforms.
- Flash Pool aggregates support different RAID types and group sizes for SSD cache.
- There is more flexibility for calculating maximum Flash Pool SSD cache size in HA configurations.
- Aggregate Snapshot copy automatic deletion is always enabled.
- Disabled volume guarantees are no longer reenabled automatically after more space becomes available.
- Fractional reserve setting can be set only to 0 or 100.

  If you upgrade a system to Data ONTAP 8.2 with volumes whose fractional reserve settings are between 1 and 99, the fractional reserve setting for those volumes is set to 0.

4.1.2 Mirrored and unmirrored aggregates

Aggregates can be mirrored, which protects from hardware failure or site disaster. Mirrored aggregates are utilized by the optional SyncMirror License. SyncMirror is described in further detail in Chapter 15., “SyncMirror” on page 207.

4.1.3 Unmirrored aggregate

Figure 4-1 shows an unmirrored aggregate (named *aggrA* by the user) that is made up of three RAID-DP groups, which are automatically named by Data ONTAP as rg0, rg1, and rg2.

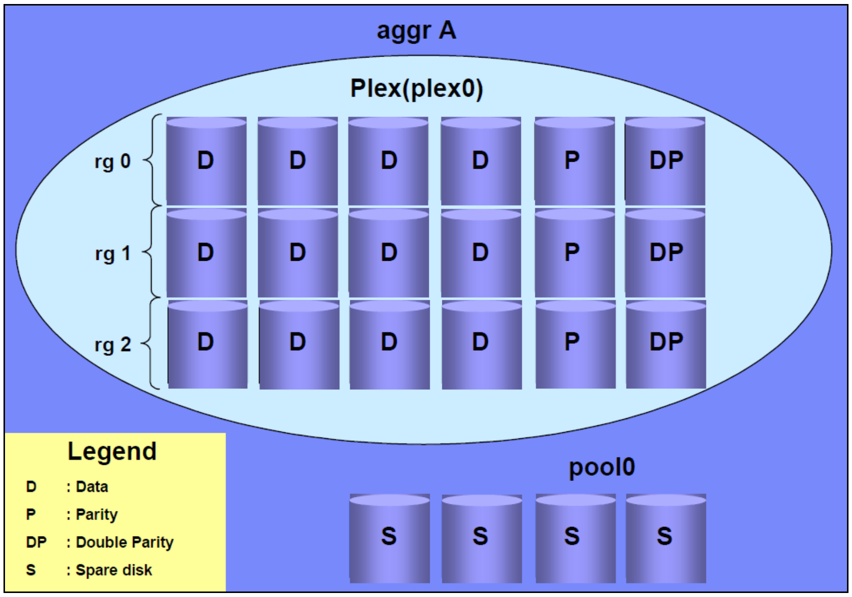


Figure 4-1 An unmirrored aggregate

Notice that RAID-DP requires that both a parity disk and a double-parity disk be in each RAID group. In addition to the disks that have been assigned to a RAID group, there are four hot spare disks in one pool of disks waiting to be assigned.

4.1.4 Mirrored aggregate

A mirrored aggregate consists of two plexes, which provides an even higher level of data redundancy through RAID-level mirroring. In order for an aggregate to be enabled for mirroring, the storage system must have a SyncMirror license for syncmirror_local or cluster_remote installed, and the storage system's disk configuration must support RAID-level mirroring.

When you enable SyncMirror, Data ONTAP divides all the hot spot spare disks into two disk pools to ensure that a single failure does not affect disks in both pools. It allows the creation of mirrored aggregates. Data ONTAP uses disks from one pool to create the first plex, and another pool to create the second plex. A failure that affects one plex will not affect the other plex.

The plexes are physically separated and are updated simultaneously during normal operation. After the plex has a problem is fixed, you can resynchronize the two plexes and reestablish the mirror relationship.

Figure 4-2 shows that SyncMirror is enabled, and plex0 and plex1 contain copies of one or more file systems. There are also hot spare disks in disk shelves and a pool for each sub-container, waiting to be assigned.

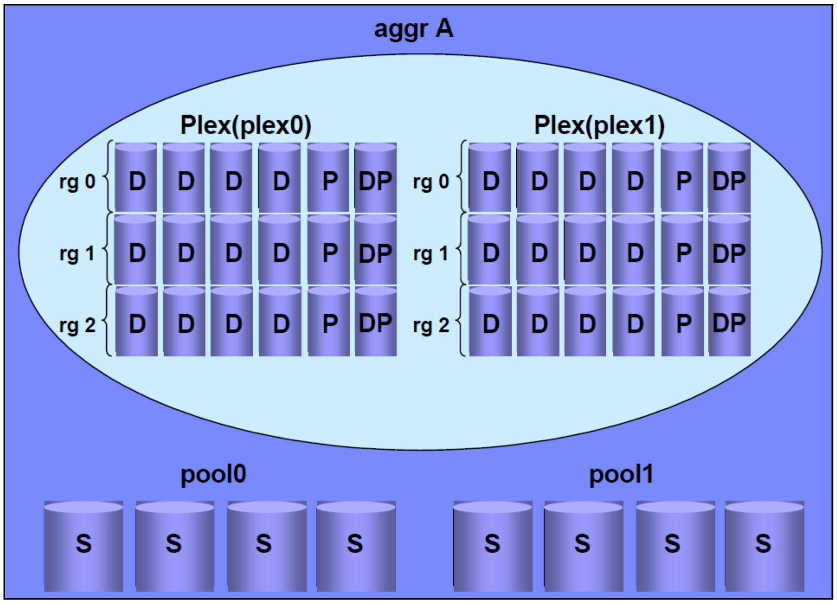


Figure 4-2 A mirrored aggregate

4.2 Aggregates and the IBM N series storage hierarchy

Each IBM System Storage N series volume depends on its containing aggregate for all its physical storage. The way that a volume is associated with its containing aggregate depends on whether the volume is a traditional volume or a flexible volume (FlexVol). In the following sections, we explain these types of volumes in more detail.

4.2.1 Traditional volume

A traditional volume is the collection of physical disk space whose entire contents is used to support a single volume. A traditional volume is tightly coupled with its containing aggregate and both the physical and the logical layer. The only way to increase the size of a traditional volume is to add entire disks to its containing aggregate. It is impossible to decrease the size of a traditional volume.

> **Attention:** With the introduction of Data ONTAP 8.x, the usage of traditional volumes is no longer relevant. All volumes must be flexible volumes.

The aggregate portion of each traditional volume is assigned its own pool of disks that are used to creates its RAID groups, which are organized into one plex. Because traditional volumes are defined by their own set of disks and RAID groups, they exist outside and independently of any other aggregates that might be defined on storage systems.

Figure 4-3 illustrates how a traditional volume is tightly coupled to its containing aggregate. When a traditional volume is created, its size is determined by the amount of disk space requested, the number of disks and their capacity to be used, or a list of disks to be used.

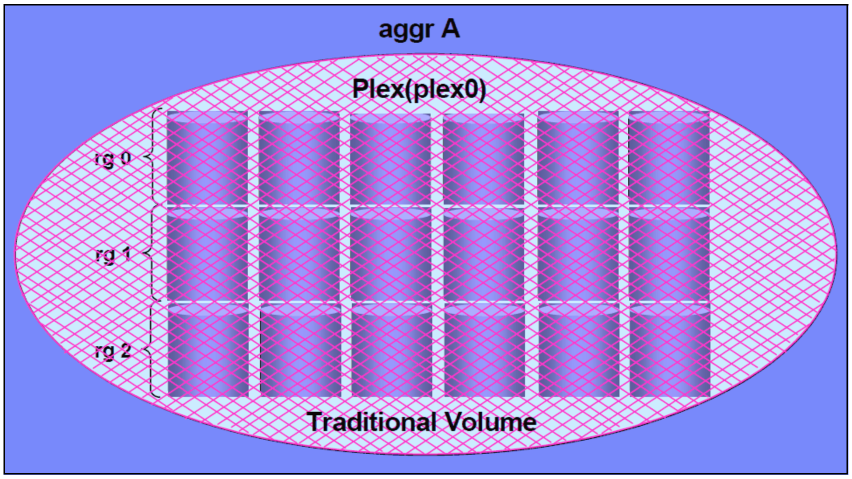


Figure 4-3 Aggregates: Traditional volume

4.2.2 Flexible volume (FlexVol)

A flexible volume, or FlexVol, is the collection of disk space allocated as a subset of the available space within an aggregate. The FlexVol volumes are loosely coupled to their containing aggregates and the logical layers. Because the volume is managed separately from the aggregate, FlexVol volumes provide many more options for managing the size of the volume.

FlexVol volumes offer these advantages:

- You can create FlexVols in an aggregate almost instantaneously. They can be as small as 20 MB or as large as the volume capacity that is supported for the storage system.
- You can increase and decrease the size of a FlexVol in small increments (as small as 4 KB) almost instantaneously.

A FlexVol can share its containing aggregate with other FlexVols. Thus, a single aggregate is the shared source of all the storage used by the FlexVol it contains, as shown in Figure 4-4.

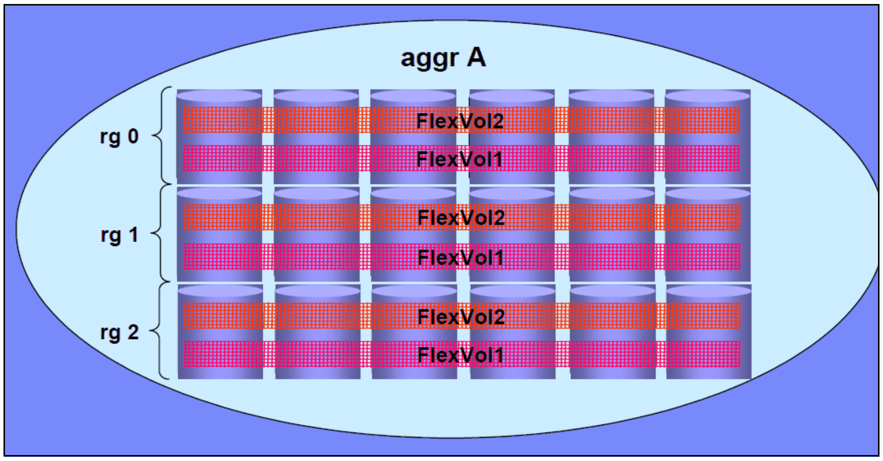


Figure 4-4 Aggregates: FlexVol volume

For more information about Flexible volumes, see 4.4, “Introduction to flexible volumes” on page 51.

4.2.3 Volume properties

Each volume is handled and known by Data ONTAP as a file system. Each file system has its own Snapshot area.

Aggregate and volume limits

In this section, we describe aggregate and volume limits.

Restrictions: Be aware of the following restrictions:

- The minimum FlexVol size is 20 MB.
- The maximum FlexVol size is 16 TB for 32 bit aggregates.
- The maximum FlexVol size is 100 TB for 64 bit aggregates.
- The maximum number of volumes is 200 for N3400.
- The maximum number of volumes is 500 for all other N series systems.

Aggregate and volume maximums depend on the product where Data ONTAP is installed.

Reference: For more information about maximums, see the “Storage limits” section of the *IBM System Storage N series Data ONTAP 8.0 7-Mode Storage Management Guide*, available at the following website:

http://www.ibm.com/storage/support/nas

The **vol status** command, shown in Example 4-1, provides information about volumes, such as name, status, settings, and options.

Example 4-1 The vol status command provides information about volumes

```
itsotuc1> vol status
         Volume State           Status            Options
           vol0 online          raid_dp, flex     root
           vol1 online          raid_dp, flex     create_ucode=on,
                                sis               convert_ucode=on
           vol2 online          raid_dp, flex     create_ucode=on,
                                sis               convert_ucode=on
itsotuc1>
```

The **df** command, shown in Example 4-2, provides information about the file system level of the volumes. The **df** command has many variables; for example, the command **df -h** gives the output a *human* perspective.

Example 4-2 The df command provides file system information

```
itsotuc1> df
Filesystem                kbytes       used      avail capacity  Mounted on
/vol/vol0/             180795672    4486516  176309156       2%  /vol/vol0/
/vol/vol0/.snapshot     45198916      70700   45128216       0%  /vol/vol0/.snapshot
/vol/vol1/                838896        132     838764       0%  /vol/vol1/
/vol/vol1/.snapshot       209680         44     209636       0%  /vol/vol1/.snapshot
/vol/vol2/               1677792        128    1677664       0%  /vol/vol2/
/vol/vol2/.snapshot       419360         44     419316       0%  /vol/vol2/.snapshot

itsotuc1> df -h
Filesystem                 total       used      avail capacity  Mounted on
/vol/vol0/                 172GB     4381MB      168GB       2%  /vol/vol0/
/vol/vol0/.snapshot         43GB       69MB       43GB       0%  /vol/vol0/.snapshot
/vol/vol1/                 819MB      132KB      819MB       0%  /vol/vol1/
/vol/vol1/.snapshot        204MB       44KB      204MB       0%  /vol/vol1/.snapshot
/vol/vol2/                1638MB      128KB     1638MB       0%  /vol/vol2/
/vol/vol2/.snapshot        409MB       44KB      409MB       0%  /vol/vol2/.snapshot

itsotuc1>
```

Space guarantee types

Next we describe flexible volume space guarantee types. The space guarantee type specifies how Data ONTAP allocates storage space (in an aggregate) for a FlexVol. Space guarantee types are volume, none, and file, as explained here:

- Volume:

  With this type, at the creation of the volume, space is allocated from the aggregate for the entire size of the volume. Space is not used, but instead reserved for the specific volume. It is the default setting.

- None:

  With this type, at the creation of the volume, there is no space allocation. Space is allocated as data is written to the volume. Note that you might run out of space before achieving the volume size. There is no support for file and LUN space reservations.

- File:

  With this type, at the creation of the volume there is no space allocation. Rather, space is allocated from the aggregate as data is written to the volume. There *is* support for file and LUN space reservations.

> **Important:** Do *not* set the space guarantee to *none* for volumes in a Common Internet File System (CIFS) environment because out-of-space errors are unexpected in a CIFS environment.

4.2.4 List affiliation of FlexVols and aggregates

The **`vol container volume_name`** command lists the volume and its containing aggregate, as shown in Example 4-3.

Example 4-3 The vol container command

```
itsotuc1> vol container vol0
Volume 'vol0' is contained in aggregate 'aggr0'
itsotuc1> vol container vol1
Volume 'vol1' is contained in aggregate 'aggr0'
itsotuc1> vol container vol2
Volume 'vol2' is contained in aggregate 'aggr0'

itsotuc1>
```

4.3 Introduction to 64-bit aggregates

Data ONTAP 8.0 and later releases support a new aggregate type whose maximum size is much bigger than the size that was supported in earlier releases of Data ONTAP. This aggregate type, called 64-bit aggregate, offers a much larger size threshold while providing all the advantages and capabilities of aggregates, including flexibility and storage efficiency. FlexVol volumes that are created in 64-bit aggregates also have a bigger maximum size threshold. Next, we describe 64-bit aggregates in detail.

4.3.1 Overview

Aggregates and FlexVol volumes are a technology that was introduced in Data ONTAP 7G to give storage administrators greater flexibility in managing the ever-changing storage size requirements while maximizing storage efficiency.

Data ONTAP 8.0 7-mode or later supports a new aggregate type whose size threshold is much bigger than the 16 terabyte (TB) aggregate size threshold that was supported in previous releases of Data ONTAP. This aggregate type, called 64-bit aggregate, offers a much larger size threshold while providing all the advantages and capabilities of aggregates, including flexibility and storage efficiency. 64-bit aggregates also support larger FlexVol volumes contained inside them. The exact size thresholds for 64-bit aggregates and for FlexVol volumes contained inside them are given in Table 4-2 on page 48, “Maximum Aggregate and Volume Sizes.” They range from 30 TB to 100 TB, depending on the model of the storage system.

The aggregates that were supported until now in Data ONTAP 7G, which have a maximum size threshold of 16 TB, will be called 32-bit aggregates from Data ONTAP 8.0 on, and they will continue to be supported with the same 16 TB threshold. When a storage system is upgraded to Data ONTAP 8.0 7-Mode or later, all the existing aggregates show up as 32-bit aggregates.

The default aggregate type that is created in data ONTAP 8.1 7-Mode or later is a 64-bit aggregate. After being created, 64-bit aggregates behave and can be used just like existing 32-bit aggregates. All the processes such as creating and managing FlexVol volumes inside the aggregate and accessing the volumes are the same, with the same commands. All operations that can be performed on FlexVol volumes in 32-bit aggregates are also supported and work with FlexVol volumes in 64-bit aggregates. A 64-bit aggregate of any size can be created as long as it is less than the maximum size threshold stated in this document.

The newer 64-bit aggregates can coexist with new and existing 32-bit aggregates on the storage system. Therefore, you can create a 64-bit aggregate on a storage system that has existing 32-bit aggregates. This coexistence has no impact on the storage system in any way. After the 64-bit aggregate is created, it is just another aggregate on the storage system, with a higher size threshold, and can be used as a regular aggregate. You can also create 32-bit aggregates, if necessary.

Important:

- Starting with Data ONTAP 8.1, the root aggregate is, by default, a 64-bit aggregate. Prior to this release, the root aggregate had to be a 32-bit aggregate.
- Newly created aggregates are 64-bit by default, and new systems are shipped with the root volume in a 64-bit aggregate (default aggregate format).

4.3.2 About aggregate types

Starting with Data ONTAP 8.0, aggregates can be of two types, depending on how they are created. They can be either 32-bit or 64-bit.

The following list outlines the characteristics of the two types of aggregates:

- 32-bit aggregates:
  - 32-bit aggregates can grow to a maximum of 16 TB, depending on the storage system model.
  - FlexVol volumes contained by 32-bit aggregates are called 32-bit volumes, and they have different characteristics than 64-bit volumes.
  - All aggregates created with versions of Data ONTAP earlier than 8.0 are 32-bit aggregates.
  - Volumes enabled for deduplication can grow to 16 TB, depending on the product
- 64-bit aggregates:
  - 64-bit aggregates have a maximum size of up to 100 TB, depending on the storage system model.
  - FlexVol volumes contained by 64-bit aggregates are called 64-bit volumes.
  - Volumes enabled for deduplication can grow to 16 TB, depending on the product.

Important: Starting with Data ONTAP 8.1, N series system have the ability to non-disruptively expand 32-bit aggregates to the 64-bit format, without requiring data to be copied or moved.

4.3.3 The need for 64-bit aggregates

Next, we consider the need for 64 bit aggregates:

- Delivering performance with large-sized SATA drives: The size of disk drives is constantly increasing. As the drives get bigger, the maximum number of disks that can fit in a 16 TB aggregate decreases. The low disk count and therefore the low spindle count can become a performance bottleneck, especially in workloads that are spindle bound for meeting their performance requirements. To alleviate this problem, aggregates that can have more spindles are needed, which means they must have bigger size thresholds.
- Maintaining high storage efficiency: The 16 TB size thresholds of existing 32-bit aggregates can hinder the ability to add new disk drives to the aggregate in fully populated RAID groups, especially when using large-sized SATA drives. In such cases, where the aggregate is not composed of fully populated RAID groups, the storage system does not provide the best possible storage efficiency. 64-bit aggregates with larger size thresholds will let you configure your storage systems to deliver the maximum storage efficiency.
- The need for larger FlexVol volume sizes: Certain applications might need the FlexVol volumes in which they store their data to be bigger than the existing 16 TB threshold. It means that the underlying aggregate must also be larger.

4.3.4 Maximum aggregate and volume sizes

The maximum size of a 64-bit aggregate depends on the storage system model. Table 4-2 lists the maximum aggregate and FlexVol volume sizes for different storage systems in the Data ONTAP 8.1 7-Mode.

Table 4-2 Maximum aggregate and FlexVol volume sizes

Controller model	Maximum aggregate size (TB)	Maximum FlexVol volume size (TB)
N7950T	100	100
N6270	70	70
N3220, N3240	60	60
N6210, N6240	50	50
N3400	30	30

4.3.5 Advantages of 64-bit aggregates

In this section, we show some advantages of using 64-bit aggregates:

- Performance: The larger size of 64-bit aggregates makes it possible to add many more disks to an aggregate than is feasible with 32-bit aggregates. Therefore, for scenarios where the disks are the bottleneck in improving performance, 64-bit aggregates with a higher spindle count can give a performance boost. All the FlexVol volumes that are created inside a 64-bit aggregate span across all the data drives in the aggregate, thus providing more disk spindles for the I/O activity on the FlexVol volumes. However, when considering the performance of a storage system, factors beyond just the disk counts often determine the performance being delivered by the storage system.
- Maintaining higher storage efficiency: 64-bit aggregates give you the ability to add disks to an aggregate in fully populated RAID groups. It gives you maximum storage efficiency while also offering all the data protection benefits of RAID-DP.
- Ease of management: 64-bit aggregates allow more disk shelves to be put into an aggregate because of their larger size. It gives storage administrators the ability to manage a system with fewer aggregates, thus reducing the overhead of managing multiple aggregates.
- Use of all existing Data ONTAP features: 64-bit aggregates have all the same abilities, functionality, and features of aggregates that were supported in Data ONTAP 7G and later. After being created, a 64-bit aggregate is just another aggregate that exists in the storage system and can be used as a regular aggregate.

  All the advantages and features of Data ONTAP, including space efficiency, thin provisioning, deduplication, FlexVol volumes, and many other features can be used on 64-bit aggregates with the same ease and simplicity as on 32-bit aggregates.

  You use the same commands for working with 32-bit and 64-bit aggregates. So, you can use your existing scripts and management tools to manage and use 64-bit aggregates and FlexVol volumes contained in 64-bit aggregates. It makes the process of transitioning to 64-bit aggregates seamless and easy.
- Better scalability for future growth: 64-bit aggregates offer you bigger aggregate and volume sizes, which means more flexibility in building a storage solution that has greater scope for scalability and management ease in the future as your storage needs grow.
- Power and space savings: 64-bit aggregates give you the ability to use large-sized SATA drives in your aggregates while providing the number of disk spindles necessary to maintain performance. You can use fewer large-sized drives, thus reducing the amount of power used by the drives. Fewer drives require less space to house them and also can reduce cooling costs and power consumed by the data center.
- Consolidate data: With 64-bit aggregates, you can create large-sized FlexVol volumes inside the aggregates, which means that you can consolidate your data into a single FlexVol volume if necessary. This consolidation helps you to simplify your data and volume management.

4.3.6 Creating and managing FlexVol volumes in a 64-bit aggregate

The process of creating and managing FlexVol volumes in a 64-bit aggregate is exactly the same as it was in previous releases of Data ONTAP 7G. You can use the **`vol create`** command to create a FlexVol inside a 64-bit aggregate. You can use the **`vol`** family of commands to manage FlexVols.

Attention: The presence of 64-bit aggregates on the storage system prevents that system from being reverted to releases earlier than Data ONTAP 8.0.

How 32-bit and 64-bit volumes differ

Starting in Data ONTAP 8.0, FlexVol volumes are one of two types:

- 32-bit
- 64-bit

The FlexVol volume type depends on the type of their containing aggregate. All FlexVol volumes created using versions of Data ONTAP earlier than 8.0 are 32-bit volumes. 32-bit volumes have a maximum size of 16 TB.

Important: Starting with Data ONTAP 8.1, N series systems support volume SnapMirror replication between 32-bit and 64-bit volumes.

The maximum size of 64-bit volumes is determined by the size of their containing aggregate; up to 100 TB, depending on the storage system model.

Tip: In both types of volumes, the maximum size for LUNs and files is 16 TB. (The term LUNs in this context refers to the LUNs that Data ONTAP serves to clients, not to the array LUNs used for storage on a storage array.)

How FlexVol volumes work

A FlexVol volume is a volume that is loosely coupled to its containing aggregate. A FlexVol volume can share its containing aggregate with other FlexVol volumes. Thus, a single aggregate can be the shared source of all the storage used by all the FlexVol volumes contained by that aggregate.

Because a FlexVol volume is managed separately from the aggregate, you can create small FlexVol volumes (20 MB or larger), and you can increase or decrease the size of FlexVol volumes in increments as small as 4 KB.

When a FlexVol volume is created, it reserves a small amount of extra space (approximately 0.5 percent of its nominal size) from the free space of its containing aggregate. This space is used to store the volume's metadata. Therefore, upon creation, a FlexVol volume with a space guarantee of volume uses free space from the aggregate equal to its size × 1.005.
A newly-created FlexVol volume with a space guarantee of `none` or `file` uses free space equal to .005 × its nominal size.

Reference: FlexVol volumes and traditional volumes have different preferred practices, optimal configurations, and performance characteristics. Make sure that you understand these differences and deploy the configuration that is optimal for your environment.

For additional information, see the *IBM System Storage N series Data ONTAP 8.0 7-Mode Storage Management Guide*, found at this website:

`http://www.ibm.com/storage/support/nas`

4.4 Introduction to flexible volumes

Volumes contain file systems that hold user data, which is accessible using one or more of the access protocols supported by Data ONTAP. Each volume depends on its containing aggregate for all its physical storage, that is, for all storage in the aggregate's disks and RAID groups.

4.4.1 How FlexVol volumes work

Flexible volumes are file systems that hold user data that is accessible through one or more of the access protocols supported by Data ONTAP, including Network File System (NFS), Common Internet File System (CIFS), HTTP, FTP, FCP, and iSCSI. Because each flexible volume is a separate file system, you can create one or more Snapshots of the data in a volume so that multiple, space-efficient, and point-in-time images of the data can be maintained for purposes such as backup and error recovery.

FlexVol technology is a ground breaking technology that comes embedded with Data ONTAP software. FlexVols are independent of the underlying physical storage; they are the logical entities that are sized, resized, managed, and moved independently of the underlying storage.

Volumes remain the primary unit of data management. Flexible volumes refer to logical entities, not (directly) to physical storage, and are transparent to the administrator.

The physical storage supporting flexible volumes is first arranged in RAID groups (either RAID 4 or RAID-DP, which is the default). One or more RAID groups are then combined together into an *aggregate*, as shown in Figure 4-5.

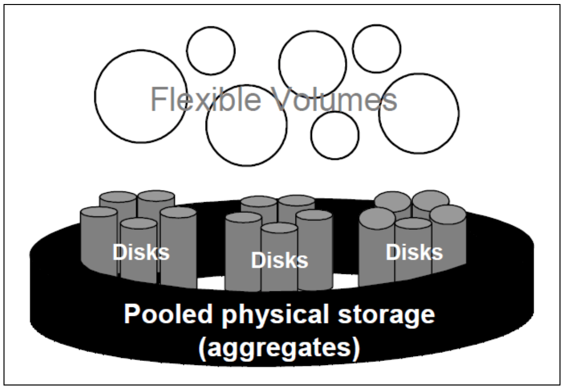


Figure 4-5 Flexible volumes

Each storage appliance can support multiple aggregates, with the maximum number dependent on the capacity of the storage appliance and the version of Data ONTAP.

Each volume depends on its containing aggregate for all its physical storage, that is, for all storage in the aggregate's disks and RAID groups, as shown in Figure 4-6.

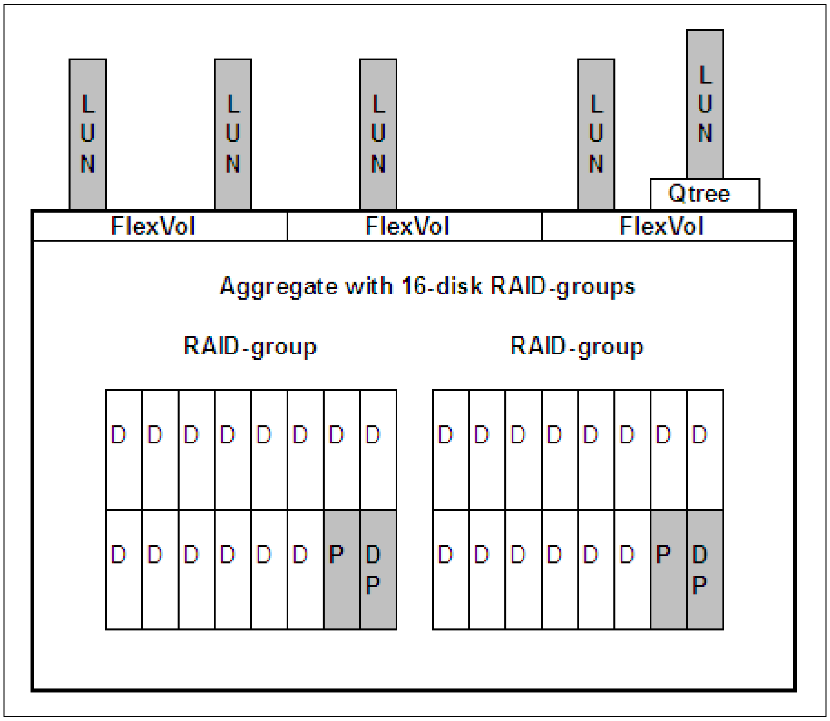
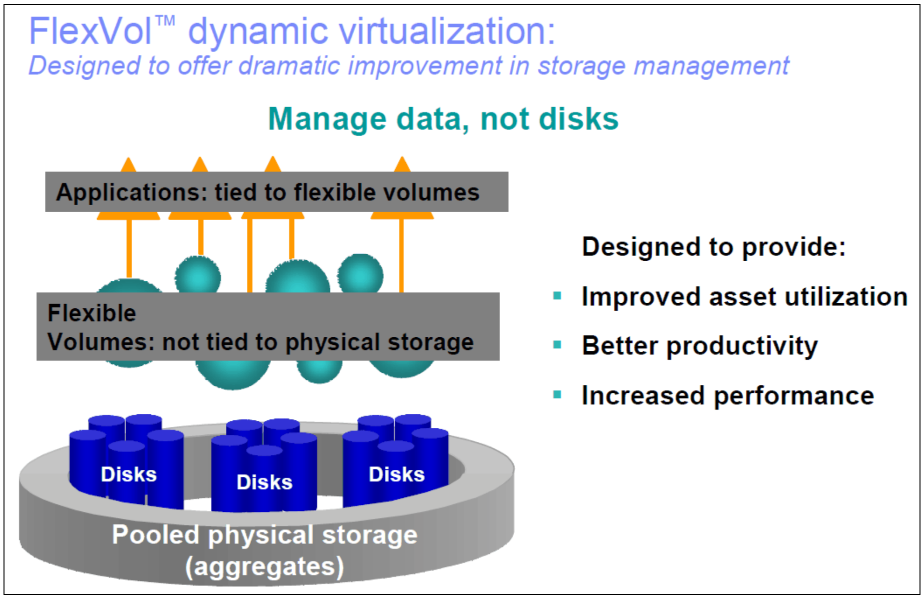


Figure 4-6 Flexible volumes: Storage

Because a FlexVol is managed separately from the aggregate, you can create small FlexVols (20 MB or larger), and you can increase or decrease the size of FlexVols in increments as small as 4 KB.

A FlexVol can share its containing aggregate with other FlexVol volumes. Thus, a single aggregate can be the shared source of all the storage used by all the FlexVol volumes contained by that aggregate. The unused space is managed by the aggregate so that unallocated space in one FlexVol does not impact the space used in another FlexVol within the same aggregate.

As shown in Figure 4-7, an aggregate is defined as a pool of many disks from which space is allocated to volumes. (Volumes are shown in Figure 4-7 as FlexVol and FlexClone entities.)



FlexVol™ dynamic virtualization:

Designed to offer dramatic improvement in storage management

Manage data, not disks

Applications: tied to flexible volumes
Flexible
Volumes: not tied to physical storage
Disks
Disks
Disks
Pooled physical storage
(aggregates)

Designed to provide:

- **Improved asset utilization**
- **Better productivity**
- **Increased performance**

Figure 4-7 An aggregate: A pool of many disks from which space is allocated to volumes

From the administrator’s point of view, volumes remain the primary unit of data management. But transparently to the administrator, flexible volumes now refer to logical entities, not (directly) to physical storage.

4.4.2 Benefits of FlexVol

Flexible volumes are no longer bound by the limitations of the disks on which they reside. A FlexVol volume is simply a *pool* of storage that can be sized based on how much data you want to store in it, rather than on what the size of your disks dictates. A FlexVol can be shrunk or increased dynamically without any downtime. Flexible volumes have all the spindles in the aggregate available to them at all times. For I/O-bound applications, flexible volumes can run much faster than equivalent-sized traditional volumes.

Flexible volumes provide these new benefits while preserving the familiar semantics of volumes and the current set of volume-specific data management and space allocation capabilities. Functions like Snapshot scheduling, quotas, and volume security options are all retained with flexible volumes, and their function and access is unchanged.

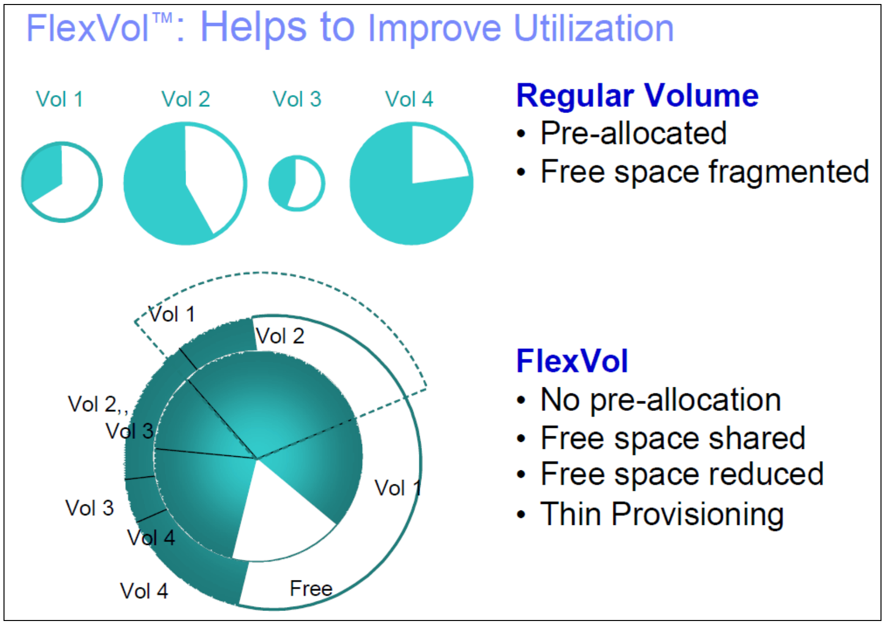
Improved performance

With Data ONTAP, disks are still organized in RAID groups, which consist of a parity disk (two in the case of RAID-DP) and some number of data disks. RAID groups are now usually be combined into aggregates.

In Data ONTAP, RAID groups are combined to create aggregates. Because volumes are still the usual unit of storage management, it is common to include all disks on a single IBM System Storage N series storage system in one aggregate, and then allocate multiple volumes on that one large aggregate. It makes it possible to tap the unused performance capacity of all the disks, making that capacity available to the busiest part of the system. A FlexVol is flexible in changing size because the underlying physical storage does not have to be repartitioned.

Improved utilization

With FlexVol, there is no preallocation required (that is, no special scripts, provisioning, or formatting of physical or logical drives). Free space is aggregated and shared to reduce space waste. Sharing, in turn, increases utilization and reduces total free space (Figure 4-8).



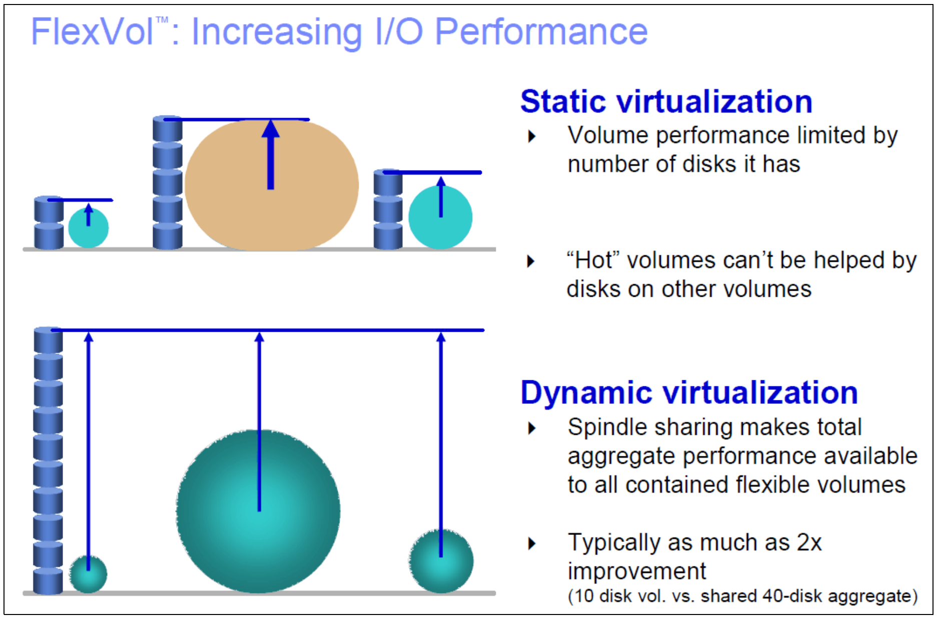
FlexVol™: Helps to Improve Utilization

Regular Volume
- Pre-allocated
- Free space fragmented

FlexVol
- No pre-allocation
- Free space shared
- Free space reduced
- Thin Provisioning

Figure 4-8 FlexVol utilization

Dynamic virtualization, shown in Figure 4-9, reduces storage bottlenecks related to hardware and allows for easy movement of data to idle space and spindles. For I/O-intensive applications, dynamic virtualization helps those applications perform better. Even for those applications or data patterns that have a tight locality of reference, those references are virtually spread among multiple physical disks.



FlexVol™: Increasing I/O Performance

Static virtualization
- Volume performance limited by number of disks it has
- “Hot” volumes can’t be helped by disks on other volumes

Dynamic virtualization
- Spindle sharing makes total aggregate performance available to all contained flexible volumes
- Typically as much as 2x improvement (10 disk vol. vs. shared 40-disk aggregate)

Figure 4-9 FlexVol performance advantages

Figure 4-10 shows how a Flexvol can be used in various ways.

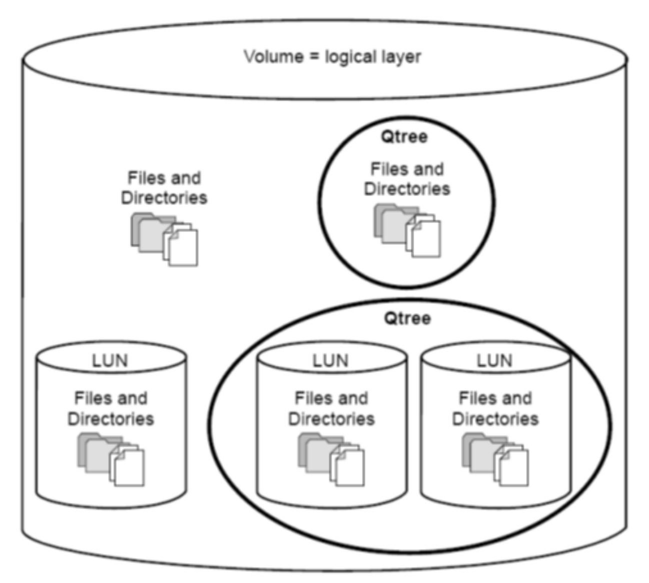
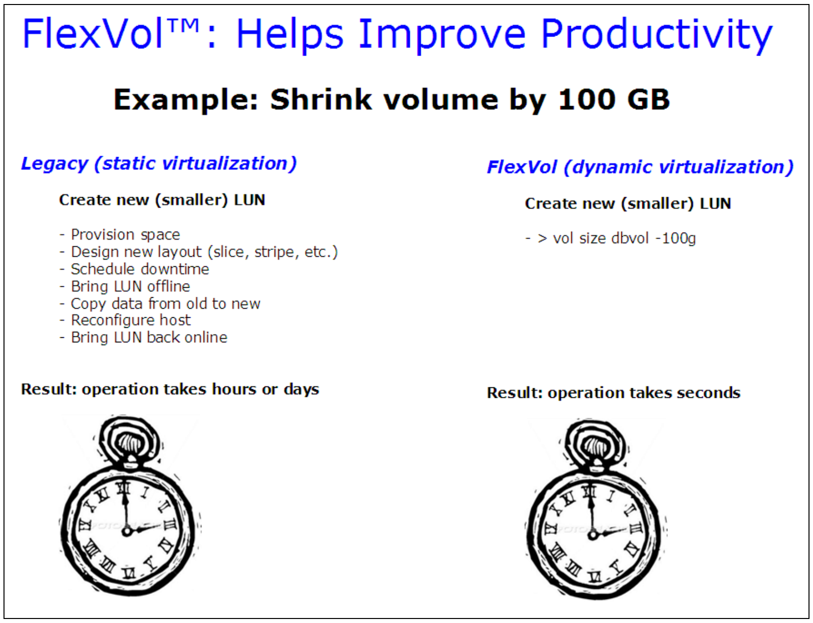


Figure 4-10 FlexVol sample usage

Flexible capacity planning

Flexible Volumes can be resized dynamically (Figure 4-11). Restrictions in size differ between platform types.



FlexVol™: Helps Improve Productivity

Example: Shrink volume by 100 GB

Legacy (static virtualization)

Create new (smaller) LUN

- Provision space
- Design new layout (slice, stripe, etc.)
- Schedule downtime
- Bring LUN offline
- Copy data from old to new
- Reconfigure host
- Bring LUN back online

Result: operation takes hours or days

FlexVol (dynamic virtualization)

Create new (smaller) LUN

- > vol size dbvol -100g

Result: operation takes seconds

Figure 4-11 FlexVol: Dynamic resizing

Administrators can use flexible volumes as a powerful tool for allocation and provisioning of storage resources among various users, groups, and projects. The smallest growth, or shrink increment, is 4 KB (which is 1 block, in WAFL terms). For example, suppose a database grows much faster than originally anticipated. The administrator can reconfigure the relevant flexible volumes at any time during the operation of the system.

The reallocation of storage resources does not require any downtime, and it is transparent to users on a file system or a LUN mapped to a host in a block environment. The effect is nondisruptive to all clients connected to this file system.

When additional physical space is required, the administrator can increase the size of the aggregate by assigning additional disks to it. The new disks become part of the aggregate, and their capacity and I/O bandwidth are available to all of the flexible volumes in the aggregate.

Overall FlexVol capacity can also be overallocated where the set capacity of all the flexible volumes on an aggregate exceeds the total available physical space. Increasing the capacity of a FlexVol does not require changing the capacity of another volume in the aggregate or the aggregate itself.

Attention: Be aware of the following restrictions:

- The minimum FlexVol size is 20 MB.
- The maximum FlexVol size is 16 TB for 32 bit aggregates.
- The maximum FlexVol size is 100 TB for 64 bit aggregates.
- The maximum number of volumes is 200 for N3400.
- The maximum number of volumes is 500 for all other N series systems.

Storage limits: For more information about maximums, see the "Storage limits" section of the *IBM System Storage N series Data ONTAP 8.0 7-Mode Storage Management Guide*, which can be found at the following website:

http://www.ibm.com/storage/support/nas

4.4.3 64-bit FlexVol volumes

Starting in Data ONTAP 8.0, FlexVol volumes are one of two types, 32-bit or 64-bit, depending on the type of their containing aggregate. All FlexVol volumes created using versions of Data ONTAP earlier than 8.0 are 32-bit volumes.

How 32-bit and 64-bit volumes differ

32-bit volumes have a maximum size of 16 TB. The maximum size of 64-bit volumes is determined by the size of their containing aggregate up to 100 TB, depending on the storage system model as shown in Table 4-3.

In both types of volumes, the maximum size for LUNs and files is16 TB. The term LUNs in this context refer to the LUNs that Data ONTAP serves to clients, not to the array LUNs used for storage on a storage array.

Tips:

- For best performance, if you want to create a large number of small files in a volume, use a 32-bit volume.
- Root FlexVols must be 32-bit volumes.

Limits for 32-bit and 64-bit volumes differ

Table 4-3 shows the maximum limits for 32-bit and 64-bit volumes available with Data ONTAP.

Table 4-3 Maximum aggregate and FlexVol volume sizes

Controller model	Maximum aggregate size (TB)	Maximum FlexVol volume size (TB)
N7950T	100	100
N6270	70	70
N3220, N3240	60	60
N6210, N6240	50	50
N3400	30	30

5

qtrees

This chapter describes qtrees, what they are, why you need them, and their management. It also provides some examples of configuring file system quotas using qtrees.

The following topics are covered:

- What qtrees are
- What qtrees do
- Working with qtrees

5.1 What qtrees are

A qtree is a special type of directory, existing only in the root of an N series FlexVol or traditional volume). Because a qtree can only exist in the volume root, qtrees cannot be nested. However, multiple normal directories can be created within a qtree. A storage controller can contain many qtrees (about 4994 per volume).

A qtree directory has a number of features that differentiate it from a normal directory.

In earlier versions of Data ONTAP, before the introduction of FlexVols, the qtree was the only method to safely sub-provision capacity within a traditional volume. By setting a quota at the “quota-tree” level (thus the “q” prefix), the administrator could effectively share capacity between multiple users or applications with a traditional volume.

However, qtrees also provide a number of other features that mean that they remain useful, even when using FlexVol volumes (which can be thought of as a method of sub-provisioning within an aggregate).

For example: qtrees can help to reduce the number of FlexVols required, by allowing multiple users or applications to safely share capacity within a common volume.

5.2 What qtrees do

The qtrees interact with a number of other features within Data ONTAP, as described in the following sections.

5.2.1 Security styles

Although all NAS data that is stored on the N series is saved in the WAFL file system, this file system can present different security metadata depending on your needs.

Whenever you create a new volume or qtree, the system will configure it with the default security style. It is either NTFS or UNIX. Note that the file system is still WAFL, which is just presenting a compatible security model to the clients.

The default security style for new volumes is set with the following command:

```
options wafl.default_security_style [ ntfs | unix | mixed ]
```

When a qtree is created, it will inherit the security style of the containing volume.

Generally, you would set the qtree security style to match the OS of the NAS clients accessing it. For example, use NTFS security style for CIFS access from Windows clients, and UNIX security style for NFS access from UNIX clients.

5.2.2 Oplocks

A qtree can be used to enabled or disable oplock support for any CIFS shares that it contains. It might be necessary if, for example, you are using a database that requires CIFS oplocks to be disabled. Then you can set CIFS oplocks to Off for that project's qtree, while allowing other projects to still use CIFS oplocks.

CIFS oplocks can be enabled or disabled in a number of ways; globally, per share, or per qtree. For more information about the interaction between these settings, see the *Data ONTAP 8.1 Commands: Manual Page Reference, Volume 1*:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003778

5.2.3 Quotas

You can limit the size of the data used by a particular project, by placing all of that project's files into a qtree and applying a tree quota to the qtree.

When first created, a qtree does not have any restrictions on disk space or the number of files (apart from the capacity of the containing volume and/or aggregate). If desired, you can impose a hard or soft quota at the qtree level.

Note that you cannot apply a quota to a non-qtree directory. If you do need to apply quotas to normal directories, then you will need to investigate an external quota management system (consult your IBM representative for advice).

For more information about setting qtree quotas, see the *Data ONTAP 8.1 7-Mode Storage Management Guide*:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003777

5.2.4 SnapMirror

The SnapMirror feature comes in two types, replicating either at the volume-level or at the qtree-level.

Although volume SnapMirror (VSM) can have some technical advantages, it also has some restrictions, such as not being able to replicate between dislike volume types. For example, you cannot use VSM between a FlexVol and a traditional volume.

In comparison, qtree SnapMirror (QSM) is able to replicate a specific qtree (not the entire volume) between any type of volume.

5.2.5 SnapVault

The SnapVault feature, which is used to copy and retain Snapshots on a secondary N series controller, is built on the assumption that you will be using qtrees.

SnapVault creates a new local Snapshot on the primary controller, and then transfers the changed data to the secondary controller. On the secondary controller, a Snapshot is made of the new data and usually retained for an extended period.

Much like the QSM feature, SnapVault is designed to back up at the qtree level.

5.3 Working with qtrees

Use the command shown in Example 5-1 to list the existing qtrees (and volumes).

Example 5-1 The qtree status command

```
nas1> qtree status

Volume   Tree     Style Oplocks  Status
-------- -------- ----- -------- ---------
vol0              unix  enabled  normal
vol1              unix  enabled  normal
vol1     qtree1   unix  enabled  normal
vol1     qtree2   unix  enabled  normal
vol1     qtree3   unix  enabled  normal
```

Use the command shown in Example 5-2 to create a new qtree.

Example 5-2 The qtree create command

```
nas1> qtree create /vol/vol1/qtree4
```

Use the command shown in Example 5-3 to change a qtree's security style.

Example 5-3 The qtree security command

```
nas1> qtree security /vol/vol1/qtree4 ntfs

Mon Jun  4 20:16:06 GMT [wafl.quota.sec.change:notice]: security style for
/vol/vol1/qtree4 changed from unix to ntfs
```

Use the command shown in Example 5-4 to delete a qtree.

Example 5-4 The qtree delete command

```
nas1> priv set advanced

Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.

nas1*> qtree delete /vol/vol1/qtree4

Mon Jun  4 20:21:16 GMT [wafl.Qtree.delete.start:notice]: Deleting qtree
/vol/vol1/qtree4.
Mon Jun  4 20:21:16 GMT [wafl.Qtree.delete.success:notice]: Successfully deleted
qtree /vol/vol1/qtree4.
```

A qtree can also be deleted from a CIFS or NFS attached as if it were a normal directory (for example, with the **rmdir** command).

6

FlexClone volumes

This chapter introduces FlexClones and helps storage system administrators learn about the full value that FlexClone volumes can bring to their operations.

The following topics are covered:

- Introduction to FlexClone volumes
- FlexClone operation
- Practical applications of FlexClone
- FlexClone performance
- Creating a FlexClone
- Accessing FlexClone volumes
- Splitting FlexClone volumes
- Summary

6.1 Introduction to FlexClone volumes

In this chapter, we describe a feature that allows IBM System Storage N series administrators to instantly create clones of a flexible volume (FlexVol). A FlexClone volume is a writable point-in-time image of a FlexVol volume or another FlexClone volume, as shown in Figure 6-1.

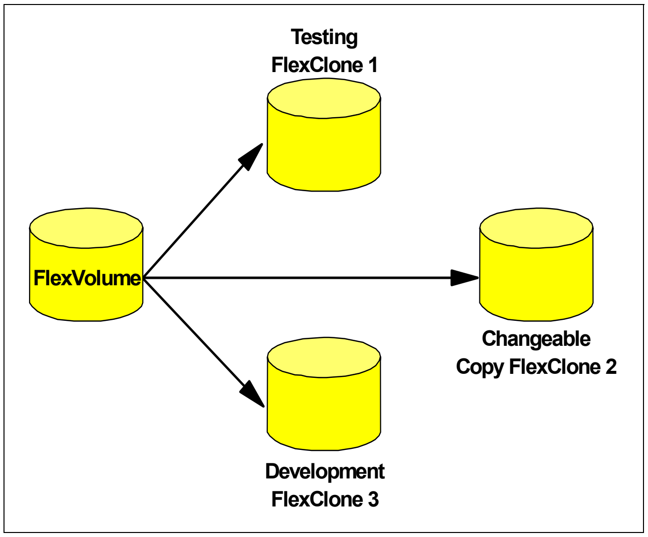
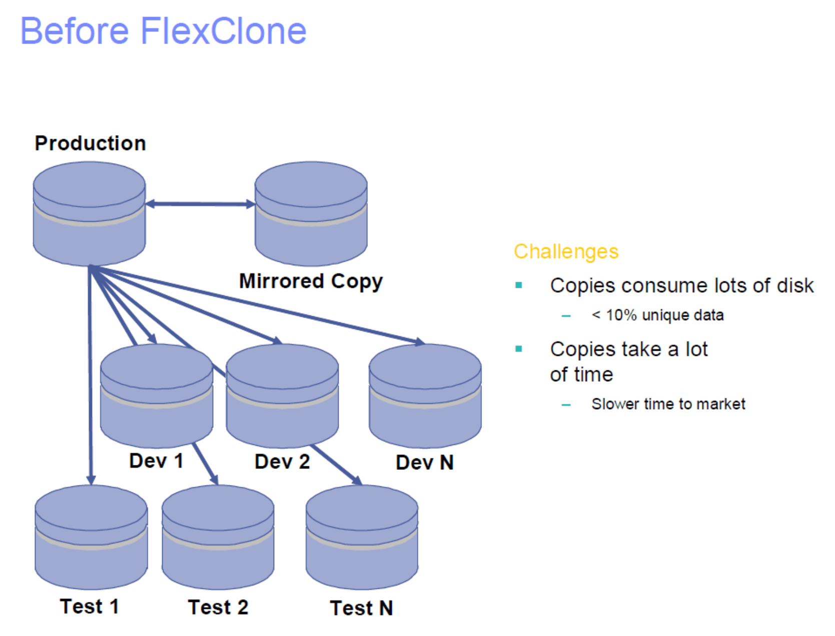


Figure 6-1 FlexClone usage

FlexClone volumes add a new level of agility and efficiency to storage operations. They take only a few seconds to create, and are created without interrupting access to the parent FlexVol volume.

FlexClone volumes use space efficiently, utilizing the Data ONTAP architecture to store only data that changes between the parent and the clone. It is a significant potential saving in dollars, space, and energy. In addition to all these benefits, FlexClone volumes have the same high performance as other kinds of volumes, as shown in Figure 6-2.



Before FlexClone

Challenges

- Copies consume lots of disk
  - < 10% unique data
- Copies take a lot of time
  - Slower time to market

Figure 6-2 Before FlexClone

Conceptually, FlexClone volumes are useful for any situation where testing or development occurs, any situation where progress is made by locking in incremental improvements, and any situation where there is a need to distribute data in changeable form without endangering the integrity of the original, as shown in Figure 6-3.

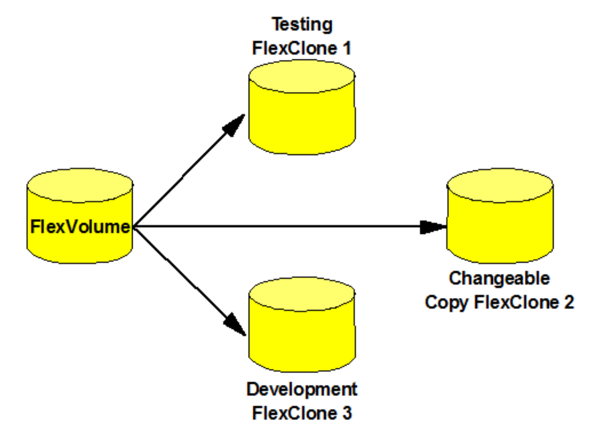
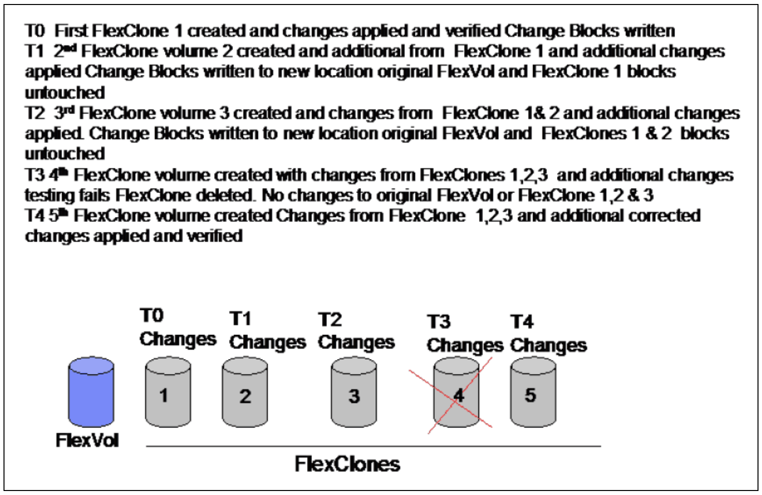


Figure 6-3 Testing with FlexClone volumes

For example, imagine a situation where the IT staff must make substantive changes to a production environment, as shown in Figure 6-4. However, the cost and risk involved are too high to do it on the production volume. Ideally, there will be an instant writable copy of the production system available at minimal cost in terms of storage and service interruptions.



T0 First FlexClone 1 created and changes applied and verified Change Blocks written
T1 2nd FlexClone volume 2 created and additional from FlexClone 1 and additional changes applied Change Blocks written to new location original FlexVol and FlexClone 1 blocks untouched
T2 3rd FlexClone volume 3 created and changes from FlexClone 1& 2 and additional changes applied. Change Blocks written to new location original FlexVol and FlexClones 1 & 2 blocks untouched
T3 4th FlexClone volume created with changes from FlexClones 1,2,3 and additional changes testing fails FlexClone deleted. No changes to original FlexVol or FlexClone 1,2 & 3
T4 5th FlexClone volume created Changes from FlexClone 1,2,3 and additional corrected changes applied and verified

Figure 6-4 FlexClone example

Figure 6-4 shows the following process:

- T0: The first FlexClone 1 is created and changes are applied. The verified change blocks are written.
- T1: The second FlexClone volume 2 is created, and changes from FlexClone 1 and additional changes are applied. Change blocks are written to the new location. (The original FlexVol and FlexClone 1 blocks are untouched.)
- T2: The third FlexClone volume 3 is created and changes from FlexClone1 and FlexClone 2 are applied, along with additional changes. Change blocks are written to the new location. (The original FlexVol, FlexClone 1, and FlexClone 2 blocks are untouched.)
- T3: The fourth FlexClone volume is created with changes from FlexClone1, FlexClone 2, and FlexClone 3, and additional changes are applied. The testing fails, and the FlexClone is deleted. (No changes are made to the original FlexVol, FlexClone 1, FlexClone 2, or FlexClone 3.)
- T4: The fifth FlexClone volume is created from changes from FlexClone 1, FlexClone 2, and FlexClone 3, and additional changes are applied and verified.

By using FlexClone volumes, the IT staff gets an instant point-in-time copy of the production data that is created transparently and uses only enough space to hold the desired changes. The staff can try out upgrades using FlexClone volumes.

At every point where the IT staff make solid progress, they clone their working FlexClone volume to lock in the successes. At any point where they get stuck, they just destroy the working clone and go back to the point of their last success. When everything is finally working as planned, the staff can either split off the clone to replace the current production volumes or codify the successful upgrade process to use on the production system during the next maintenance window.

To summarize, FlexClone allows you to make the necessary changes to your infrastructure without worrying about crashing your production systems or making untested changes on the system under tight maintenance window deadlines. The results are less risk, less stress, and higher levels of service for IT customers.

6.2 FlexClone operation

FlexClone volumes have all the capabilities of a FlexVol volume, including growing, shrinking, and being the source of a Snapshot or even another FlexClone volume. The technology that makes it all possible is integral to how Data ONTAP manages storage.

IBM System Storage N series storage systems use a Write Anywhere File Layout (WAFL) to manage disk storage. FlexClone writes use free blocks in the aggregate. Any new data that gets written to the volume does not need to go on a specific spot on the disk. It can be written anywhere. WAFL then updates the metadata to integrate the newly written data into the correct place in the file system.

If the new data is meant to replace older data, and the older data is not part of a Snapshot, WAFL marks the blocks containing the old data as reusable. It can happen asynchronously and does not affect performance. Snapshots work by making a copy of the metadata associated with the volume. Data ONTAP preserves pointers to all the disk blocks currently in use at the time that the Snapshot is created.

When a file is changed, the Snapshot still points to the disk blocks where the file existed before it was modified, and changes are written to new disk blocks. As data is changed in the parent FlexVol, the original data blocks stay associated with the Snapshot, rather than getting marked for reuse.

All the metadata updates are just pointer changes, and the storage system takes advantage of locality of reference, non-volatile RAM (NVRAM), and RAID technology to keep everything fast and reliable. Figure 6-5 provides a graphical illustration of how it works.

FlexClone reads are satisfied in either of the following ways:

- Blocks that are written to the FlexClone
- The parent FlexVol, if data is unchanged since the Snapshot

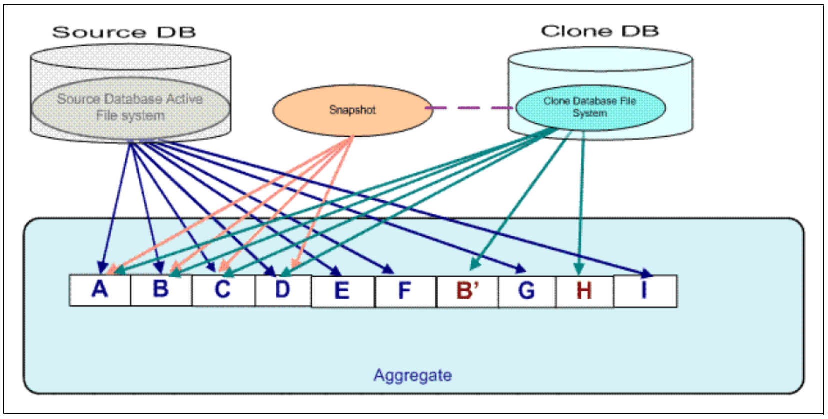

Source DB
Clone DB
Source Database Active File system
Snapshot
Clone Database File System
A
B
C
D
E
F
B’
G
H
I
Aggregate

Figure 6-5 Snapshot

You can think of a FlexClone volume as a transparent writable layer in front of the Snapshot (Figure 6-6). A FlexClone volume is writable, so it needs some physical space to store the data that is written to the clone. It uses the same mechanism used by Snapshot copies to get available blocks from the containing aggregate.

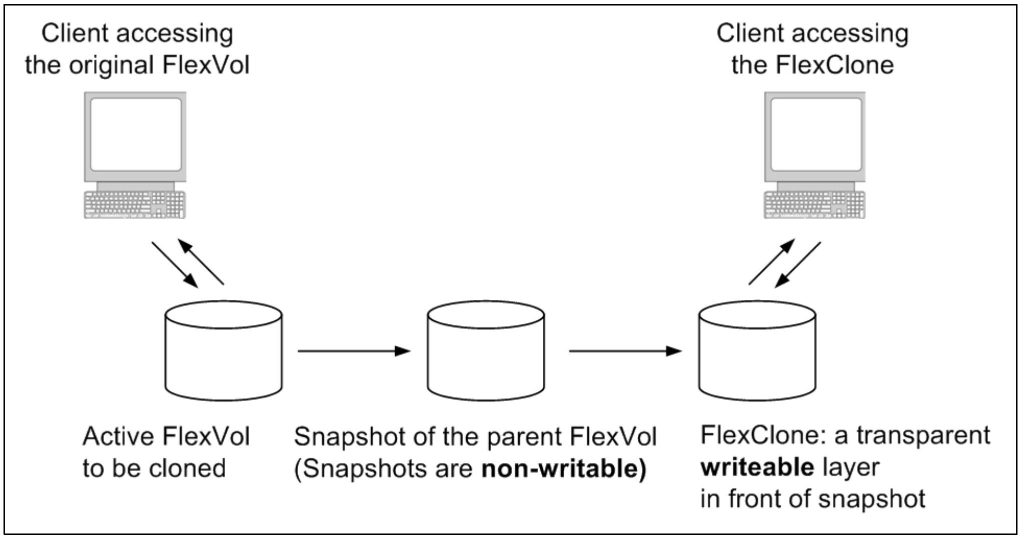
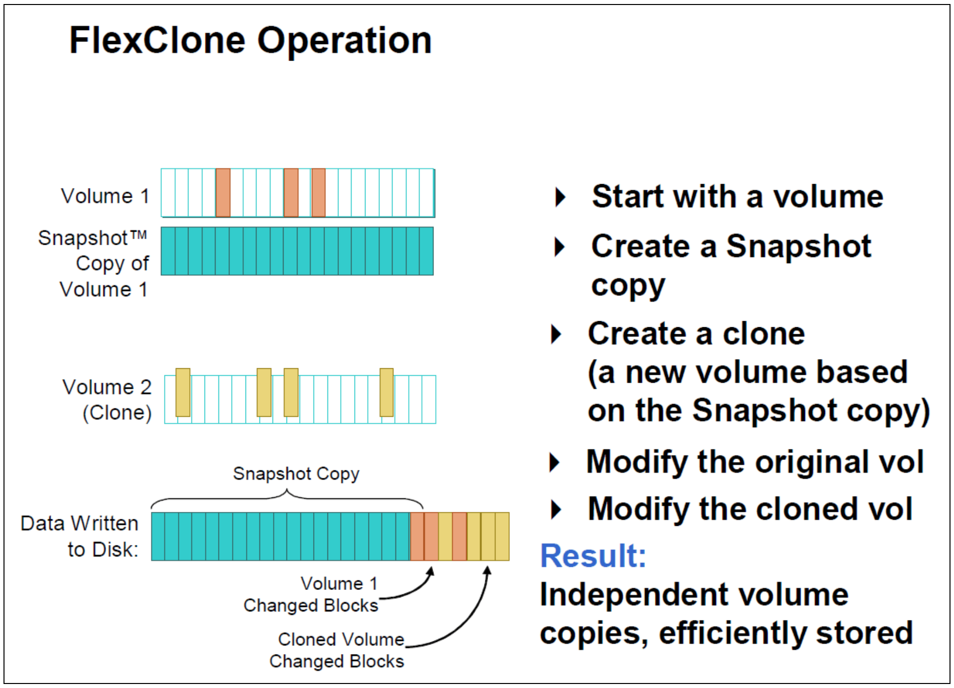

Client accessing the original FlexVol
Client accessing the FlexClone
Active FlexVol to be cloned
Snapshot of the parent FlexVol (Snapshots are **non-writable)**
FlexClone: a transparent **writeable** layer in front of snapshot

Figure 6-6 Think of a FlexClone volume as a transparent writable layer in front of a Snapshot

A Snapshot simply links to existing data that was overwritten in the parent. In contrast, a FlexClone volume stores the data written to it on disk (using WAFL) and then links to the new data as well (Figure 6-7). The disk space associated with the Snapshot and FlexClone is accounted for separately from the data in the parent FlexVol.



FlexClone Operation

- **Start with a volume**
- **Create a Snapshot copy**
- **Create a clone (a new volume based on the Snapshot copy)**
- **Modify the original vol**
- **Modify the cloned vol**

Result:
Independent volume copies, efficiently stored

Figure 6-7 FlexClone operation

When a FlexClone volume is first created, it needs to know the parent FlexVol and also the Snapshot of the parent to use as its base. The Snapshot can already exist, or it can be created automatically as part of the cloning operation.

The FlexClone volume takes a copy of the Snapshot metadata and then updates its metadata as the clone volume is created. Creating the FlexClone volume takes just a few moments, because the copied metadata is small compared with the actual data.

The parent FlexVol can change independently of the FlexClone volume because the Snapshot is there to keep track of the changes and prevent the original parent's blocks from being reused while the Snapshot exists. The same Snapshot is read-only and can be efficiently reused as the base for multiple FlexClone volumes.

Space is used efficiently, because the only new disk space used is either associated with the small amounts of metadata, updates, or additions to either the parent FlexVol or the FlexClone volume.

FlexClone volumes appear to the storage administrator just like a FlexVol, that is, they look like a regular volume and have all of the same properties and capabilities. Using the CLI, FilerView, or DataFabric Manager, you can manage volumes, Snapshot copies, and FlexClone volumes, including getting their status (as shown in Example 6-1) and seeing the relationships between the parent, Snapshot, and clone.

Example 6-1 The vol status command

```
itsotuc1> vol status cifs_vol1_clone5
         Volume State           Status            Options
cifs_vol1_clone5 online          raid_dp, flex     create_ucode=on,
                                sis               convert_ucode=on
                Clone, backed by volume 'cifs_vol1', snapshot
'clone_cifs_vol1_clone5.1'
                         Volume UUID: eb289e50-5c00-11e0-b9d8-00a098098a07
                Containing aggregate: 'aggr1'

itsotuc1>
```

The CLI is required to create and split a FlexClone volume. FlexClone volumes are treated just like a FlexVol for most operations. The main limitation is that Data ONTAP forbids operations that will destroy the parent FlexVol or base Snapshot while dependent FlexClone volumes exist.

Other caveats are that management information in external files (for example, /etc) associated with the parent FlexVol is not copied, quotas for the clone volume get reset rather than added to the parent FlexVol, and LUNs in the cloned volume are automatically marked offline until they are uniquely mapped to a host system. Lastly, splitting the FlexClone volume from the parent volume to create a fully independent volume requires adequate free space in the aggregate to copy shared blocks.

6.3 Practical applications of FlexClone

FlexClone technology enables multiple, instant data set clones with no storage impact. It provides dramatic improvements for application test and development environments, and is tightly integrated with file system technology and a microkernel design in a way that renders competitive methods archaic.

FlexClone volumes are ideal for managing production data sets. They allow effortless error containment for bug fixing and development. They simplify platform upgrades for Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) applications. Instant FlexClone volumes provide data for multiple simulations against large data sets for ECAD, MCAD, and Seismic applications, without unnecessary duplication or waste of physical space.

The ability to split a FlexClone volume from its parent allows administrators to easily create new permanent, independent volumes for forking project data. FlexClone volumes have their limits, but the real range of applications is limited only by imagination. Table 6-1 lists a few of the more common examples.

Table 6-1 Application of FlexClone

Application area	Benefits
Application testing	You can make the necessary changes to infrastructure without worrying about crashing production systems, avoid making untested changes on the system under tight maintenance window deadlines, and experience less risk, less stress, and higher service level agreements.
Data mining	Data mining operations and software can be implemented more flexibly because both reads and writes are allowed.
Parallel processing	Multiple FlexClone volumes of a single milestone/production data set can be used by parallel processing applications across multiple servers to get results more quickly.
Online backup	You can resume immediately read-write workload upon discovering corruption in the production data set by mounting the clone instead. Use database features such as IBM DB2® write-suspend or Oracle hot backup mode to transparently prepare the database volumes for cloning by delaying write activity to the database. It is necessary because databases must maintain a point of consistency.
System deployment	You can maintain a template environment and use FlexClone volumes to build and deploy either identical or variation environments, create a test template that is cloned as needed for predictable testing, and have faster and more efficient migration using the Data ONTAP SnapMirror feature in combination with FlexClone volumes.
IT operations	You can maintain multiple copies of production systems (live, development, test, reporting, and so on) and refresh working FlexClone volumes regularly to work on data as close to live production systems as practical.

6.4 FlexClone performance

The performance of FlexClone volumes is nearly identical to the performance of flexible volumes. It is due to the way that cloning is tightly integrated with WAFL and the IBM System Storage N series architecture. Unlike other implementations of cloning technology, FlexClone volumes are implemented as a simple extension to existing core mechanisms.

The impact of cloning operations on other system activity will also be relatively light and transitory. The FlexClone create operation is nearly identical to creating a Snapshot. Some CPU, memory, and disk resources are used during the operation, which usually completes in seconds. The clone metadata is held in memory like a regular volume, so the impact on storage system memory consumption is identical to having another volume available. After the clone creation completes, all ongoing accesses to the clone are nearly identical to accessing a regular volume.

Splitting the FlexClone to create a fully independent volume also uses resources. While the split is occurring, free blocks in the aggregate are being copied (Figure 6-8). It incurs disk I/O operations and can potentially compete with other disk operations in the aggregate.

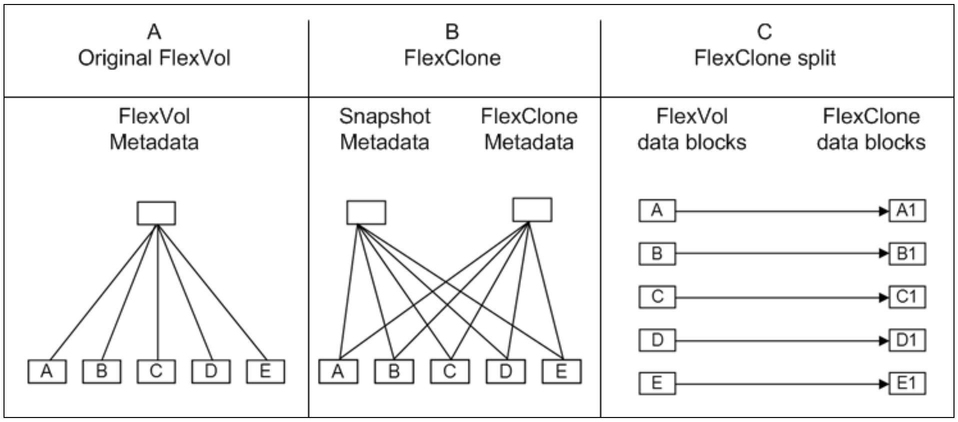


Figure 6-8 FlexClone split

The copy operation also uses CPU and memory resources, which might impact the performance of a fully loaded storage system. Data ONTAP addresses these potential issues by completing the split operation in the background, and sets priorities in a way that does not significantly impact foreground operations. It is also possible to manually stop and restart the split operation if some critical job requires the full resources of the storage system.

The final area to consider is the impact on disk usage from frequent operations where FlexClone volumes are split off and used to replace the parent FlexVol volume. The split volume is allocated free blocks in the aggregate, taking contiguous chunks as they are available. If there is ample free space in the aggregate, the blocks allocated to the split volume will be mostly contiguous. If the split is used to replace the original volume, the blocks associated with the destroyed original volume will become available and create a potentially large free area within the aggregate. That free area will also be mostly contiguous.

In cases where many simultaneous volume operations reduce contiguous regions for the volumes, Data ONTAP uses a block reallocation functionality. The **`reallocate`** command makes defragmentation and sequential reallocation even more flexible and effective. It reduces any impact of frequent clone split and replace operations, and optimizes performance after other disk operations (for example, adding disks to an aggregate) that might unbalance block allocations.

6.5 Creating a FlexClone

Our demonstration in this section uses the Data ONTAP CLI.

Perform the following steps:

1. When SnapClones are created, the **`vol clone create`** command automatically creates a new Snapshot for the clone if none is specified. A Snapshot can also be specified for the **`vol clone create`** command**.** We create a new Snapshot by issuing the **`snap create <volume_name> <snap_name>`** command, as shown in Example 6-2.

Example 6-2 Create and list Snapshots

```
itsotuc1> snap create cifs_vol1 cifs_vol1_snap1
itsotuc1> snap list cifs_vol1
Volume cifs_vol1
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
 27% (27%)    0% ( 0%)  Mar 29 22:51  cifs_vol1_snap1
 44% (29%)    0% ( 0%)  Mar 29 20:00  hourly.0
 57% (35%)    0% ( 0%)  Mar 29 16:00  hourly.1
 63% (29%)    0% ( 0%)  Mar 29 12:00  hourly.2
 68% (29%)    0% ( 0%)  Mar 29 08:00  hourly.3
 71% (26%)    0% ( 0%)  Mar 29 00:00  nightly.0

itsotuc1>
```

2. Run the **`vol status <volume_name>`** command on your volume to obtain the current status of the volume and to identify the aggregate to which it belongs (Example 6-3).

Example 6-3 The vol status command

```
itsotuc1> vol status cifs_vol1
         Volume State           Status            Options
      cifs_vol1 online          raid_dp, flex     create_ucode=on,
                                                  convert_ucode=on
                         Volume UUID: 2321fa5a-5b24-11e0-aade-00a098098a07
                Containing aggregate: 'aggr1'
itsotuc1>
itsotuc1> vol container cifs_vol1
Volume 'cifs_vol1' is contained in aggregate 'aggr1'

itsotuc1>
```

3. Use the **df -g** command to check the available disk space on your volume (Example 6-4).

Example 6-4 df -g command (output modified for clarity)

```
itsotuc1> df -g
Filesystem              total       used      avail capacity  Mounted on
/vol/cifs_vol1/           8GB        0GB        7GB       0%  /vol/cifs_vol1/
/vol/cifs_vol1/.snapshot 2GB 0GB 1GB 0% /vol/cifs_vol1/.snapshot
/vol/vol0/              164GB        4GB      159GB       3%  /vol/vol0/
/vol/vol0/.snapshot      41GB        0GB       41GB       0%  /vol/vol0/.snapshot

itsotuc1>
```

4. Use the **df -Ag** command to check the available disk space on the aggregates (Example 6-5).

Example 6-5 df -Ag command

```
itsotuc1> df -Ag
Aggregate                total       used      avail capacity
aggr1                    227GB       10GB      217GB       4%
aggr1/.snapshot           11GB        0GB       11GB       0%
aggr0                    227GB      206GB       20GB      91%
aggr0/.snapshot           11GB        0GB       11GB       3%
itsotuc1>
```

5. Clone the existing volume by issuing the following command:

   `vol clone create <cloneVol> -s volume -b <parentVol> <parentSnap>`

   Where:

   - **.<cloneVol>** is the name of the new clone volume.
   - **-s volume** is the space guarantee for the volume.
   - **-b <parentVol>** is the volume to be cloned.
   - **<parentSnap>** is the parent Snapshot (can be omitted and a new Snapshot for the clone will be automatically be created).

6. Create the volume and check the status using the **vol status** command as shown in Example 6-6.

Example 6-6 The vol clone create command

```
itsotuc1> vol clone create cifs_vol1_clone1 -s volume -b cifs_vol1
cifs_vol1_snap1
Tue Mar 29 23:26:51 GMT [wafl.volume.clone.created:info]: Volume clone
cifs_vol1_clone1 of volume cifs_vol1 was created successfully.
Creation of clone volume 'cifs_vol1_clone1' has completed.

itsotuc1> vol status cifs_vol1_clone1
         Volume State           Status           Options
cifs_vol1_clone1 online          raid_dp, flex     create_ucode=on,
                                                  convert_ucode=on
                Clone, backed by volume 'cifs_vol1', snapshot 'cifs_vol1_snap1'
                         Volume UUID: 5b069910-5bee-11e0-b9d8-00a098098a07
                Containing aggregate: 'aggr1'
itsotuc1>
```

The **snap list** command shows that the snapshot we selected for our clone cifs_vol1_snap1 is busy. This Snapshot is the source for or new clone and hence it cannot be deleted before the SnapClone is deleted. See Example 6-7.

Example 6-7 Status of FlexClone creation

```
itsotuc1> snap list cifs_vol1
Volume cifs_vol1
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
 10% (10%)    0% ( 0%)  Mar 29 22:51  cifs_vol1_snap1 (busy,vclone)
 18% (10%)    0% ( 0%)  Mar 29 20:00  hourly.0
 27% (13%)    0% ( 0%)  Mar 29 16:00  hourly.1
 32% (10%)    0% ( 0%)  Mar 29 12:00  hourly.2
 37% (10%)    0% ( 0%)  Mar 29 08:00  hourly.3
 41% ( 9%)    0% ( 0%)  Mar 29 00:00  nightly.0

itsotuc1>
```

6.6 Accessing FlexClone volumes

In the previous sections, we demonstrated how to create FlexClone volumes based on a parent volume and a Snapshot.

In this section, we explain how to connect to a FlexClone volume.

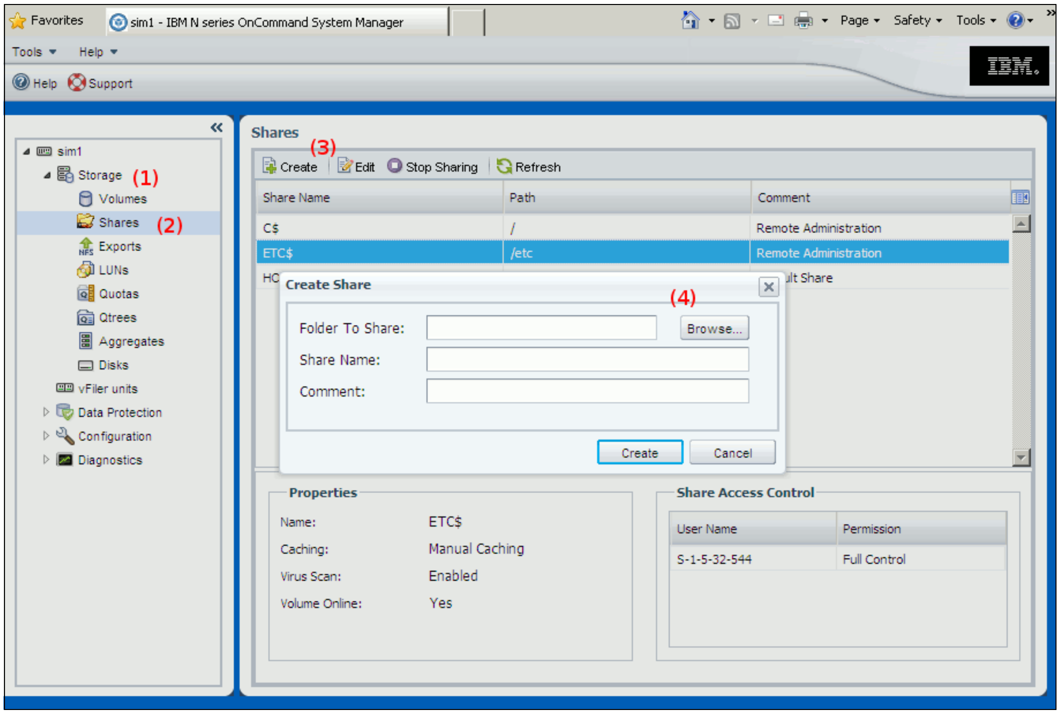
To connect to a FlexClone volume from a Window 2008 server, do the following steps:

1. Use the **vol status** command to check which volumes are available as shown in Example 6-8.

Example 6-8 Listing volumes

```
itsotuc1> vol status
         Volume State           Status           Options
      cifs_vol1 online          raid_dp, flex    create_ucode=on,
                                sis              convert_ucode=on
           vol0 online          raid_dp, flex    root
cifs_vol1_clone1 online          raid_dp, flex     create_ucode=on,
                                sis              convert_ucode=on
cifs_vol1_clone2 online          raid_dp, flex     create_ucode=on,
                                sis              convert_ucode=on
cifs_vol1_clone3 online          raid_dp, flex     create_ucode=on,
                                sis              convert_ucode=on
cifs_vol1_clone4 online          raid_dp, flex     create_ucode=on,
                                sis              convert_ucode=on
cifs_vol1_clone5 online          raid_dp, flex     create_ucode=on,
                                sis              convert_ucode=on
itsotuc1> vol status cifs_vol1_clone5
         Volume State           Status           Options
cifs_vol1_clone5 online          raid_dp, flex     create_ucode=on,
                                sis              convert_ucode=on
                Clone, backed by volume 'cifs_vol1', snapshot
'clone_cifs_vol1_clone5.1'
                         Volume UUID: eb289e50-5c00-11e0-b9d8-00a098098a07
                Containing aggregate: 'aggr1'
itsotuc1>
```

2. In order to connect to cifs_vol1_clone5, first create a share for our FlexClone volume. We use System Manager for this task as shown in Figure 6-9.
   Click **Storage** → **Shares** → **Create** → **Browse**.



Share Name	Path	Comment
C$	/	Remote Administration
ETC$	/etc	Remote Administration
HO		ult Share

Create Share

Folder To Share: Browse...

Share Name:

Comment:

Create Cancel

Properties

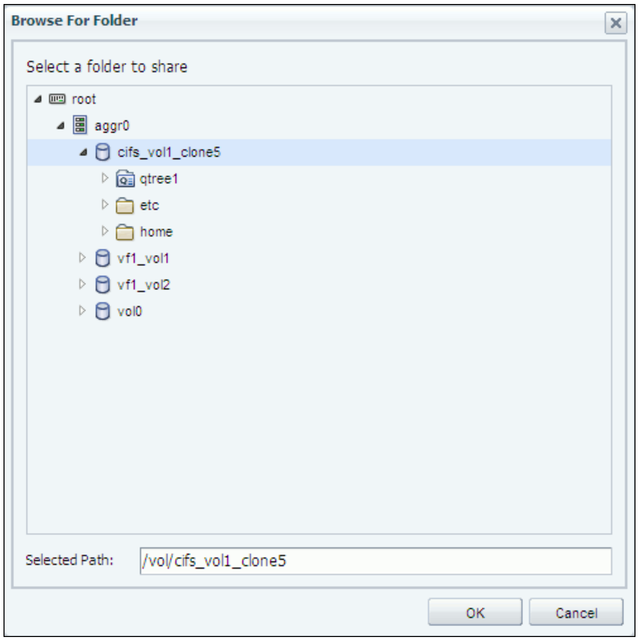
Name:	ETC$
Caching:	Manual Caching
Virus Scan:	Enabled
Volume Online:	Yes

Share Access Control

User Name	Permission
S-1-5-32-544	Full Control

Figure 6-9 Creating a CIFS share on our FlexClone volume

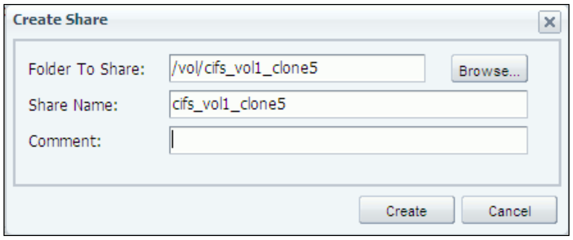
3. Use the Browse window to select **`cifs_vol1_clone5`** to share as shown in Figure 6-10. Click **OK**.



Selected Path: /vol/cifs_vol1_clone5

Figure 6-10 Select the location for the CIFS share

4. The location details will then update in the previous dialog, as shown in Figure 6-12. Click **Create**.



Folder To Share: /vol/cifs_vol1_clone5

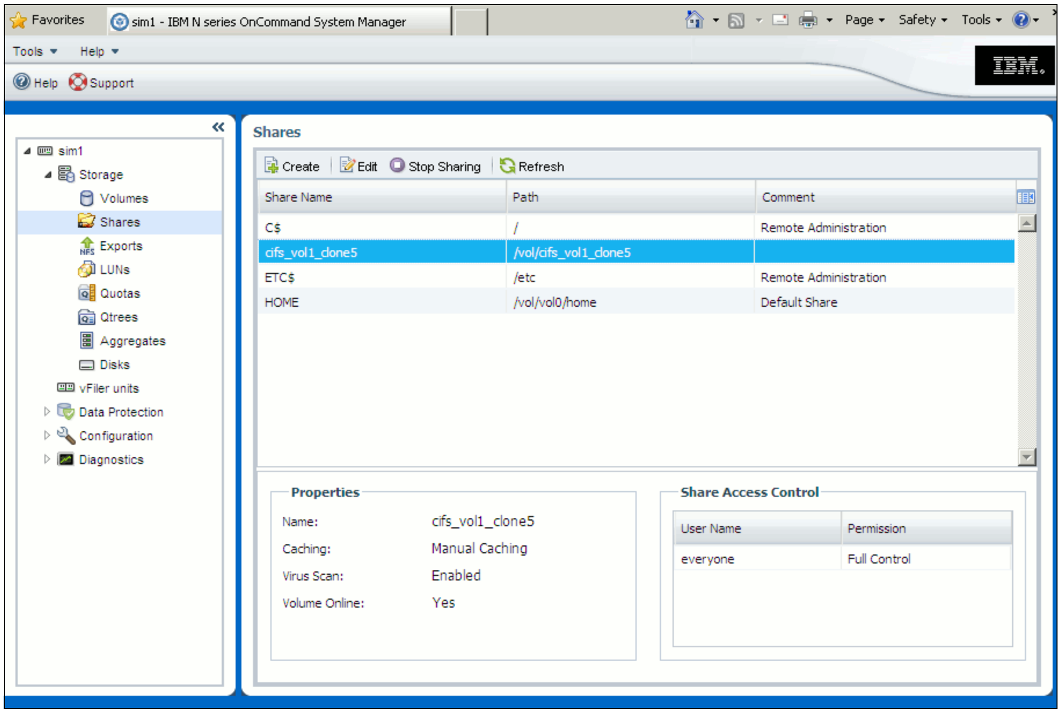
Share Name: cifs_vol1_clone5

Comment:

Figure 6-11 Preparing to create a CIFS share

5. Observe that our CIFS share creation is successful as shown in Figure 6-12.

   Normally, next you would select **Edit** to modify the share settings, such as access permissions, but we will leave the default values in this example.



Share Name	Path	Comment
C$	/	Remote Administration
cifs_vol1_clone5	/vol/cifs_vol1_clone5	
ETC$	/etc	Remote Administration
HOME	/vol/vol0/home	Default Share

Properties

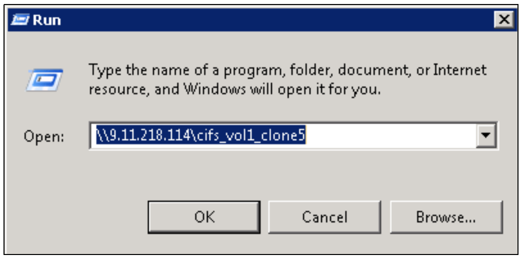
Name:	cifs_vol1_clone5
Caching:	Manual Caching
Virus Scan:	Enabled
Volume Online:	Yes

Share Access Control

User Name	Permission
everyone	Full Control

Figure 6-12 Add CIFS share successful

6. From our Windows 2008 server, click **Start** → **Run** and type the name of the CIFS share as shown in Figure 6-13.

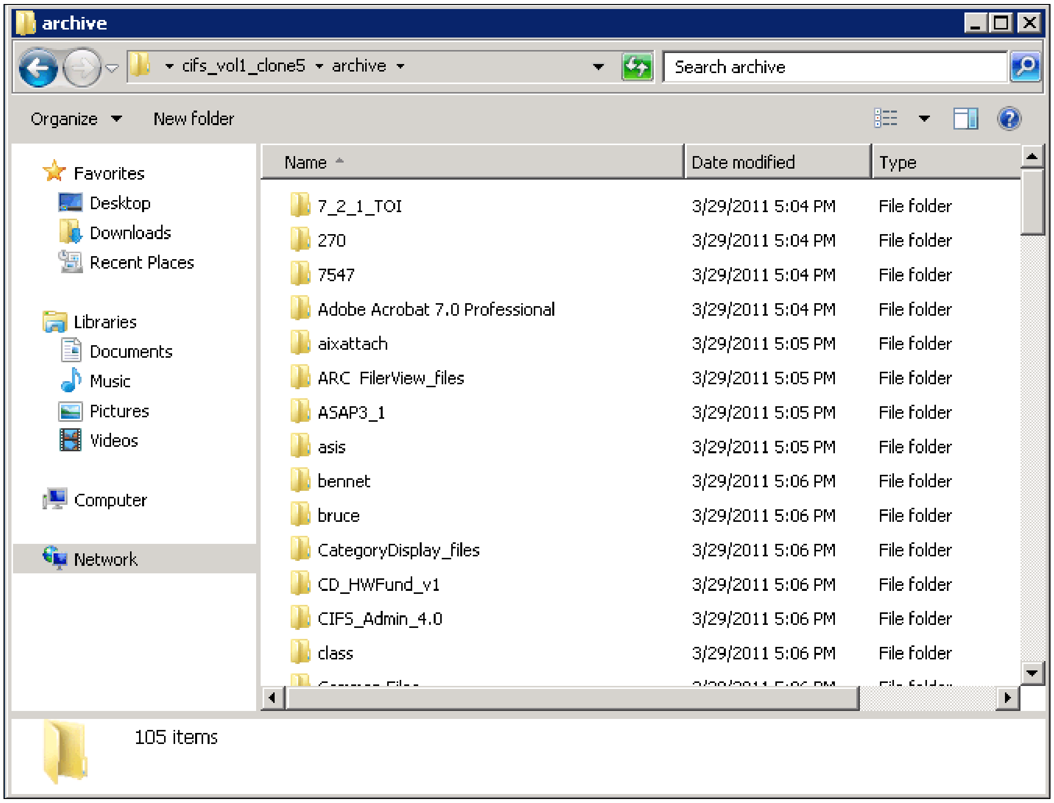


Type the name of a program, folder, document, or Internet resource, and Windows will open it for you.

Open: \\9.11.218.114\cifs_vol1_clone5

Figure 6-13 Connecting the FlexClone CIFS share from a Windows 2008 server

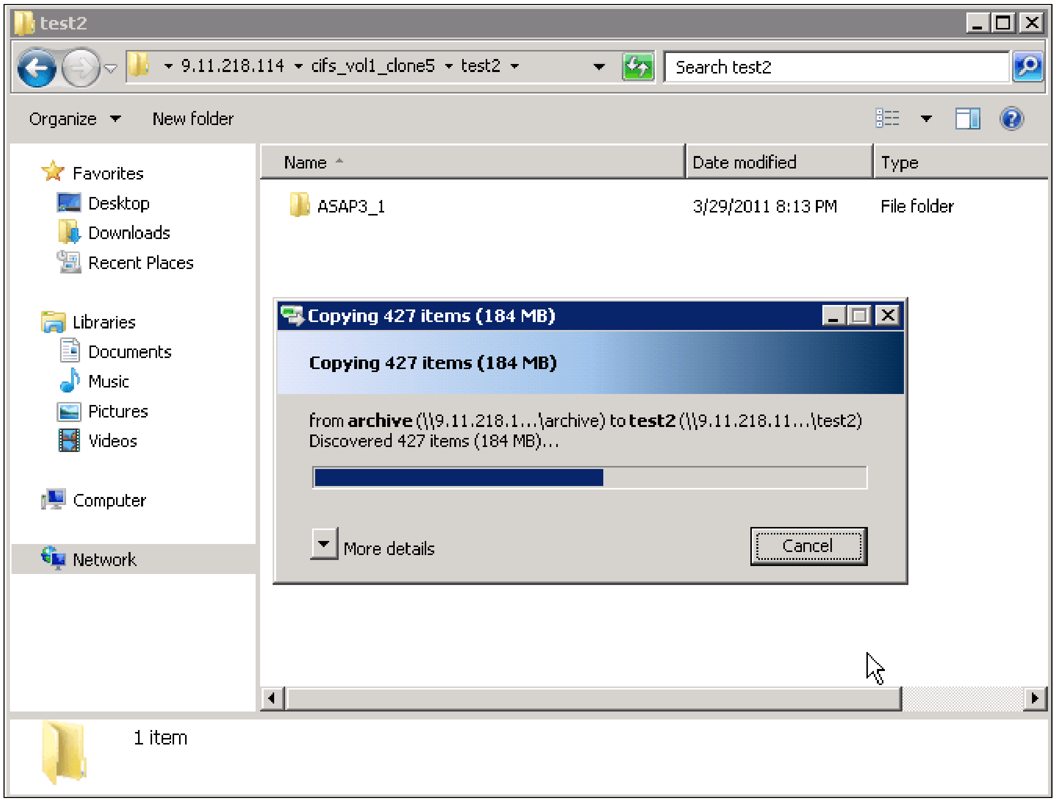
7. Open the CIFS share and verify its content as shown in Figure 6-14.



Name	Date modified	Type
7_2_1_TOI	3/29/2011 5:04 PM	File folder
270	3/29/2011 5:04 PM	File folder
7547	3/29/2011 5:04 PM	File folder
Adobe Acrobat 7.0 Professional	3/29/2011 5:04 PM	File folder
aixattach	3/29/2011 5:05 PM	File folder
ARC FilerView_files	3/29/2011 5:05 PM	File folder
ASAP3_1	3/29/2011 5:05 PM	File folder
asis	3/29/2011 5:05 PM	File folder
bennet	3/29/2011 5:06 PM	File folder
bruce	3/29/2011 5:06 PM	File folder
CategoryDisplay_files	3/29/2011 5:06 PM	File folder
CD_HWFund_v1	3/29/2011 5:06 PM	File folder
CIFS_Admin_4.0	3/29/2011 5:06 PM	File folder
class	3/29/2011 5:06 PM	File folder

Figure 6-14 Verifying that our data from the FlexClone is available

8. Write to the FlexClone CIFS share as shown in Figure 6-15.



ASAP3_1 3/29/2011 8:13 PM File folder

Copying 427 items (184 MB)

Copying 427 items (184 MB)

from **archive** (\\9.11.218.1...\archive) to **test2** (\\9.11.218.11...\test2)
Discovered 427 items (184 MB)...

1 item

Figure 6-15 Writing data to the FlexClone CIFS share

We have successfully created a writable FlexClone copy and accessed it from a Windows 2008 server.

6.7 Splitting FlexClone volumes

This section describes how a FlexClone volume can be split from its parent volume.

There are some limitations to FlexClone volumes, and it is likely that at some point in time, an IBM N series administrator will want to get the full benefit of a FlexVol volume. In order to obtain it, the FlexClone volume can be split from its parent volume and Snapshot.

Reference: For more information about the limitations of FlexClone volumes, see the *IBM System Storage N series Data ONTAP 8.0 7-Mode File Access and Protocols Management Guide*, available at:

http://www.ibm.com/storage/support/nas

FlexClone volumes, while being FlexClones, do not take up their own space in the containing aggregate. A FlexClone volume is based on pointers to a Snapshot, and the only space it requires from the aggregate is the delta from what has been written to the writable FlexClone volume since it was created.

We now demonstrate how to split a FlexClone volume from its parent FlexVol volume, so that it becomes an individual FlexVol volume.

Example 6-9 shows how we check the available space in the containing aggregate aggr1 and in the FlexClone volume cifs_vol1_clone5.

Example 6-9   Checking space in aggr1 before split

```
itsotuc1> df -Ag aggr1
Aggregate                total       used      avail capacity
aggr1                    227GB       75GB      151GB      33%
aggr1/.snapshot           11GB        0GB       11GB       0%
itsotuc1>
itsotuc1> df -g cifs_vol1_clone5
Filesystem               total       used      avail capacity  Mounted on
/vol/cifs_vol1_clone5/       24GB       22GB        1GB      92%
/vol/cifs_vol1_clone5/
/vol/cifs_vol1_clone5/.snapshot        6GB        0GB        5GB       0%
/vol/cifs_vol1_clone5/.snapshot

itsotuc1>
```

Now we split the clone as shown in Example 6-10.

Example 6-10   Splitting the clone from its parent

```
itsotuc1> vol status cifs_vol1_clone5
         Volume State           Status            Options
cifs_vol1_clone5 online          raid_dp, flex     create_ucode=on,
                                sis               convert_ucode=on
                Clone, backed by volume 'cifs_vol1', snapshot
'clone_cifs_vol1_clone5.1'
                         Volume UUID: eb289e50-5c00-11e0-b9d8-00a098098a07
                Containing aggregate: 'aggr1'

itsotuc1> vol clone split start cifs_vol1_clone5
Wed Mar 30 04:05:56 GMT [wafl.volume.clone.split.started:info]: Clone split was
started for volume cifs_vol1_clone5
Wed Mar 30 04:05:56 GMT [wafl.scan.start:info]: Starting volume clone split on
volume cifs_vol1_clone5.
Clone volume 'cifs_vol1_clone5' will be split from its parent.
Monitor system log or use 'vol clone split status' for progress.

itsotuc1> vol clone split status
Volume 'cifs_vol1_clone5', 13567 of 135681 inodes processed (9%)
        980097 blocks scanned. 974729 blocks updated.

itsotuc1>
```

Splitting a FlexClone volume can take hours or even days it the FlexClone volume is very large. Check the IBM N series syslog or check the actual progress with the CLI command, **`vol clone split status`.** In our case, the 24 GB volume split took about 15 minutes.

As soon as the FlexClone split action is started, we can immediately see the result in available capacity for aggr1 as shown in Example 6-11.

Example 6-11 Checking space in aggr1 after split

```
itsotuc1> df -Ag aggr1
Aggregate                total       used      avail capacity
aggr1                    227GB       97GB      129GB      43%
aggr1/.snapshot           11GB        0GB       11GB       0%

itsotuc1> df -g cifs_vol1_clone5
Filesystem               total       used      avail capacity  Mounted on
/vol/cifs_vol1_clone5/       24GB       22GB        1GB      93%
/vol/cifs_vol1_clone5/
/vol/cifs_vol1_clone5/.snapshot        6GB        0GB        6GB       0%
/vol/cifs_vol1_clone5/.snapshot

itsotuc1>
```

By observing the output from **`df -Ag`** and from **`df -g cifs_vol1_clone5`**, we can see that the available space in the aggregate immediately reduces with 22 GB when clone split is initiated. This is exactly how much space our new FlexVol volume (it is no longer a FlexClone) takes up of space in the aggregate.

6.8 Summary

Storage administrators now have access to greater flexibility and performance. Flexible volumes, aggregates, and RAID-DP provide unparalleled levels of storage virtualization, enabling IT staff to economically manage and protect enterprise data without compromise. FlexClone volumes are one of the many powerful features that make it possible, providing instantaneous writable volume copies that use only as much storage as necessary to hold new data.

FlexClone volumes enable and simplify many operations. Application testing benefits from less risk, less stress, and higher service levels by using FlexClone volumes to try out changes on clone volumes and upgrade under tight maintenance windows by simply swapping tested FlexClone volumes for the originals.

Data mining and parallel processing benefit by using multiple writable FlexClone volumes from a single data set, all without using more physical storage than needed to hold the updates.

FlexClone volumes can be used as online backup and disaster recovery volumes immediately resuming read-write operation if a problem occurs. System deployment becomes much easier by cloning template volumes for testing and rollout. IT operations benefit from multiple copies of the production system that can be used for testing and development and refreshed as needed to more closely mirror the live data.

7

FlexCache volumes

This chapter provides an introduction to FlexCache, a feature of Data ONTAP that implements file caching for NFS environments.

The following topics are covered:

- Introduction to FlexCache
- How FlexCache works

7.1 Introduction to FlexCache

FlexCache is a caching technology that provides a cache architecture at the storage protocol layer. Similar to the way a cache in the memory architecture of a computer system improves performance, FlexCache improves performance in your NFS environments by scaling out cache volumes for increased IOPs, bringing data closer to your hosts for decreased latencies, off-loading overburdened storage controllers, or a combination of all of these.

Restriction: FlexCache is only available for the NFS protocol.

For more information about the FlexCache feature, see the Data ONTAP 8.1 7-mode Storage Management Guide:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003777

7.2 How FlexCache works

This section briefly describes how FlexCache works. It explains caching and shows how FlexCache handles reads and writes from the hosts.

7.2.1 Caching

A cache is a temporary storage location that resides between a host and a source of data. The main objective of a cache is to store frequently accessed portions of a source of data in a way that allows the data to be served faster and/or more efficiently than it would be by fetching the data from the source. Caches are beneficial in read-intensive environments in which data is accessed more than once and/or is shared by multiple hosts.

A cache can serve data faster in one of two ways:

- The cache system is faster than the system with the data source. It can be achieved through faster storage in the cache (for example, FC versus SATA), increased processing power in the cache, and increased (or faster) memory in the cache.
- The storage space for the cache is physically closer to the host, so it does not take as long to reach the data.

Caches are implemented with different architectures, policies, and semantics so that the integrity of the data is protected as it is stored in the cache and served to the host. The following sections describe the way FlexCache implements its cache architecture.

7.2.2 Reads

Caches are populated as a host reads data from the source (see Figure 7-1). On the first read of any data (1), the cache has to fetch the data from the original source (2). The data is returned to the cache (3), stored in the cache, and then passed back to the host (4). As reads are passed through a cache, the cache fills up by storing the requested data.

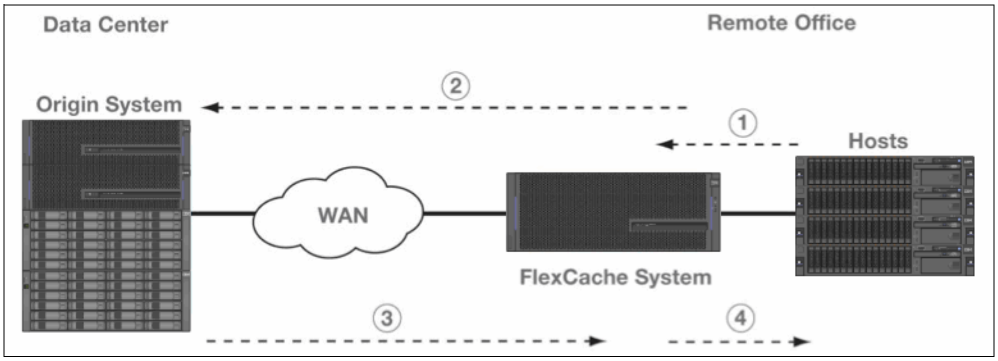


Figure 7-1 First read of data

Any subsequent accessing of data (see Figure 7-2) that is already stored in the cache (1) can be served immediately back to the host (2) without spending time and resources accessing the original source of the data. It is the primary advantage of a cache serving frequently accessed data directly to a host without having to fetch the data from the original source.

However, you might be thinking, "What if the data changed on the origin system? Does the FlexCache system still serve the data stored in its cache?" It is possible that the FlexCache system could be storing data that has changed at the origin system (called stale data). However, although it is true that the FlexCache system can store stale data, policies exist that allow you to control and manage how the FlexCache system handles stale data.

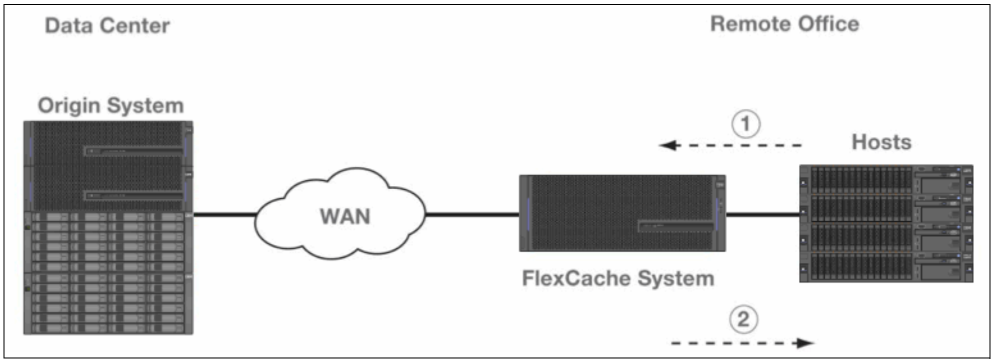


Figure 7-2 Subsequent read of the same data

7.2.3 Writes

In a FlexCache system (see Figure 7-3), all writes from a host (1) are passed directly through the cache volume to the origin volume (2). The origin volume responds to the FlexCache volume when it assumes responsibility for the new or changed data (3); only then does the FlexCache volume acknowledge the result of the write to the host (4). It is called a write-through cache.

A write-through cache is a cache that does not respond to the host until it receives a response from the next subsystem in the line. In other words, the FlexCache cache volume does not respond to the host until the origin volume acknowledges receipt of the data, thus helping to keep the data safe and sound.

It is in contrast to a write-back cache, which responds to the host immediately before verifying that the data can be successfully passed to the next subsystem. When a write-back cache accepts responsibility for data and responds to the host before acknowledging receipt of the next subsystem, the write-back cache must protect the data until it is written to physical media (such as disk).

Data in this state is called dirty data, and it must be protected from system failures such as power loss (in such a way that when power is restored, the dirty data is still accessible and ready to be stored on disk). A FlexCache system is not a write-back cache, and therefore it does not store dirty data. Because dirty data is never stored in a cache volume, it is not imperative to protect a FlexCache system from power failure. A power failure does not result in data loss or data corruption. A power failure results only in the loss of the host's NFS mount point.

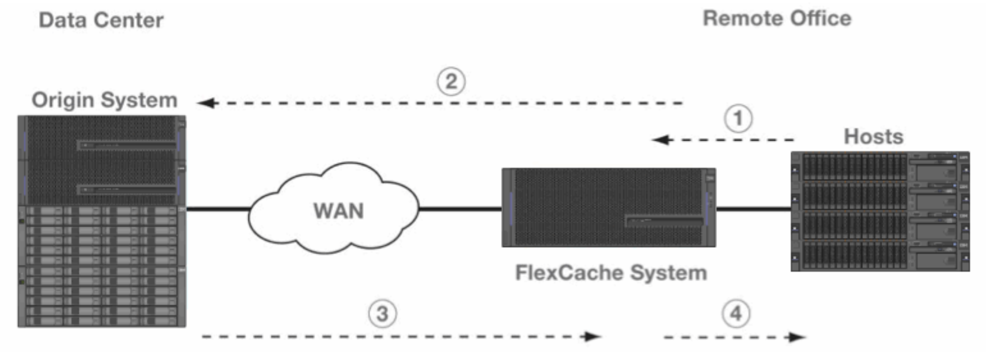


Figure 7-3 Host writes data to a write-through FlexCache system

7.2.4 Caching granularity

Over time, a cache volume becomes a collection of bits and pieces from various files, various directories, and various other objects. The collection of data in a cache volume is never intended to represent a fully functional set of the data at any particular point in time. It is called a sparse volume. A sparse volume occurs because applications and operating systems request only the portions of data needed at that moment in time; that is, an entire file is not read from the origin for every file access.

There are several types of objects that can be read from a host. A cache volume caches files, directories, and symbolic links. Each object is stored at 4K-block-level granularity. In other words, if a file on the origin volume is 400K but the host requests only the first 8,000 bytes of the file, then FlexCache requests and stores only the first two 4K blocks of the file. It maximizes disk space efficiency. Also, for every file that contains data blocks in the cache volume, the file attributes are cached (meaning that the entire file is considered to be cached). If a file or piece of data that does not exist in the cache is requested from the cache volume, that data is retrieved from the origin volume. Data continues to be stored in the cache volume until the cache volume or its aggregate runs out of space.

When writes come from the host for data that is in the cache volume, they must be sent through to the origin volume. If the new data is part of a file that is in the cache, then the entire file is invalidated, meaning that every block of data stored in the cache for that file is no longer used. It can result in invalidating more data than necessary; for example, if you changed only one block of a file but the cache stored 10 blocks of the file, none of the 10 blocks will be used for future reads. Because of this design point, data sets consisting of large files that are frequently updated might not be good candidates for a FlexCache implementation. Also note that the newly written data is not stored in the cache volume; data is stored in the cache volume only as a result of a read through the cache.

While writes to files and directories clearly invalidate the cached object as described earlier, changing the permissions of a file or directory is not as obvious. When the permissions of a file are changed from the host, the file is not invalidated in the cache volume. However, if the permissions of a directory are changed, then the directory object is invalidated in the cache volume.

8

FlexShare

This chapter introduces FlexShare, which is a Data ONTAP feature that allows administrators to prioritize how system resources are used.

FlexShare, a built-in feature of Data ONTAP, allows storage administrators to accomplish these tasks with ease and flexibility. Using FlexShare, storage administrators can host different applications confidently on a single storage system without impacting critical applications, resulting in reduced costs and simplified storage management.

The following topics are covered:

- Introduction to FlexShare
- FlexShare concept
- Benefits of using FlexShare
- When to use FlexShare
- Supported configurations
- Using FlexShare in cluster storage systems
- Setting up FlexShare
- FlexShare usage examples
- FlexShare preferred practices
- FlexShare administration
- Understanding FlexShare behavior and troubleshooting
- Summary

8.1 Introduction to FlexShare

IBM System Storage N series FlexShare is a control-of-service tool designed to give administrators the control to prioritize applications based on how critical they are to the business. It also provides a priority mechanism to give preferential treatment to higher priority tasks using the methods described in this section.

Priorities are assigned to volumes to assign relative priorities between these possibilities:

- Using different volumes:

  For example, you can specify that operations on volume cifs_vol3 are more important than operations on volume cifs_vol2 and other volumes.

- Client data accesses and system operations:

  You can specify that client accesses are more important than SnapMirror operations (Figure 8-1).

- Cache utilization options:

  You can configure the cache to retain data in cache or reuse the cache depending on workload characteristics. Optimizing cache usage can significantly increase performance for data that is frequently read or written.

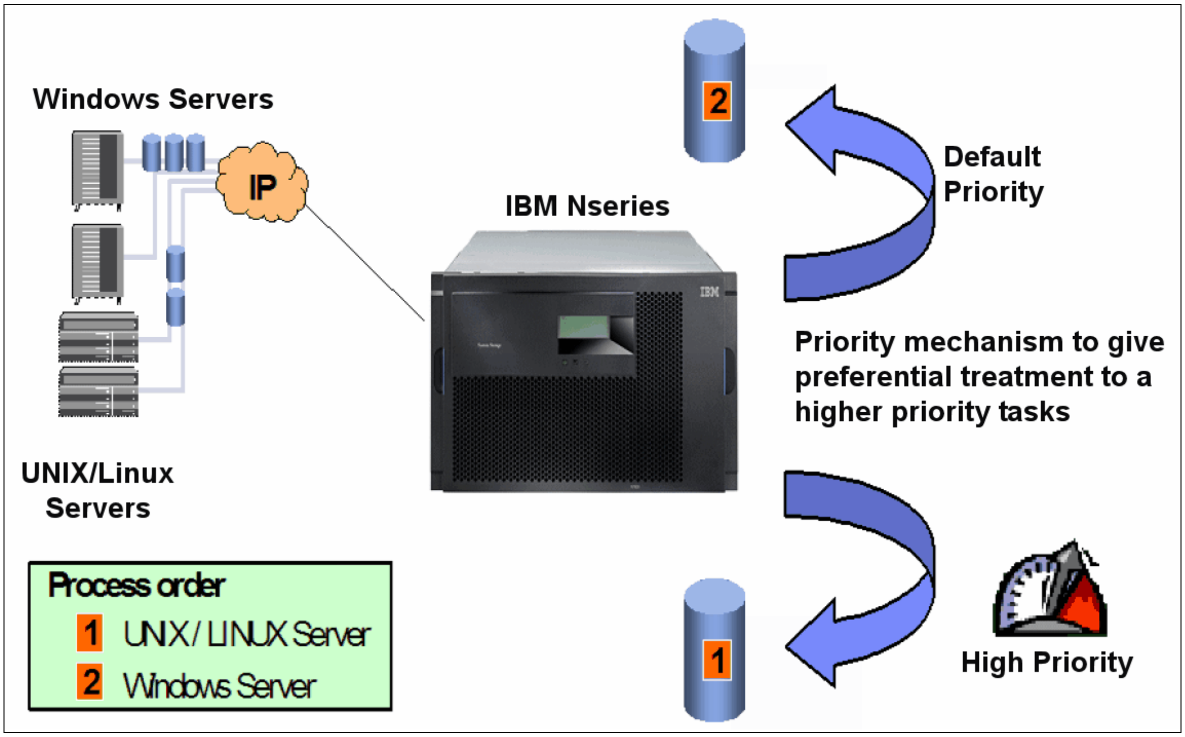


Figure 8-1 Assigning priority

For more information about the FlexShare feature, see the *Data ONTAP 8.1 7-Mode System Administration Guide*:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003720

8.2 FlexShare concept

If your storage system consistently provides the performance required for your environment, you do not need FlexShare. However, if your storage system sometimes does not deliver sufficient performance to some of its users, you can use FlexShare to increase your control over storage system resources to ensure that those resources are being used more effectively than before.

FlexShare provides the ability to assign priorities to different volumes and the ability to configure certain per-volume attributes, including user versus system priority and cache policies.

Table 8-1 lists important user and system operations.

Table 8-1 User operations and system operations

User operations	System operations
Data access operations using: ► Network File System (NFS) ► Common Internet File System (CIFS) ► iSCSI ► FCP ► HTTP ► FTP	SnapMirror SnapVault WAFL scanners **`vol clone`** and **`vol split`** commands NDMP

8.2.1 Queues

FlexShare maintains different queues for each volume that has a configured priority setting. FlexShare populates queues for each volume with WAFL operations as they are submitted for execution. The queues are only used when the FlexShare service is on. Any volume that does not have a priority assigned is in the default queue.

8.2.2 Buffer cache policies

Data ONTAP uses the cache to store buffers in memory for rapid access. When the cache is full and space is required for a new buffer, Data ONTAP uses a modified least recently used (LRU) algorithm to determine which buffers must be discarded from the cache.

FlexShare can modify how the default buffer cache policy behaves by providing hints for the buffers associated with a volume. FlexShare provided hints to Data ONTAP by specifying which information must be kept in the cache and which information must be reused.

FlexShare caching policies, if configured properly based on application workload, can enhance overall system performance. The buffer cache policy configuration is based on a per-volume setting.

8.2.3 How FlexShare schedules WAFL operations

FlexShare impacts the order in which WAFL operations are processed by the storage system. FlexShare determines the order in which WAFL operations will be processed, based on the priority configuration. When the FlexShare service is on, the prioritization processing is always in effect.

Volume level priorities

The impact of FlexShare volume-level priority can be understood by comparing one storage system with the FlexShare service off with a second storage system with the FlexShare service on.

FlexShare prioritizes processing resources for key services when the system is under heavy load. It does not provide guarantees on the availability of resources or how long particular operations will take to complete. FlexShare provides a priority mechanism to give preferential treatment to higher priority tasks

When the FlexShare service is off, the system processes the requests in the order in which they arrive. Figure 8-2 shows the order in which tasks arrive to be processed and the order in which they are processed by a storage system. The order of tasks processed is exactly the same as the order in which tasks arrive.

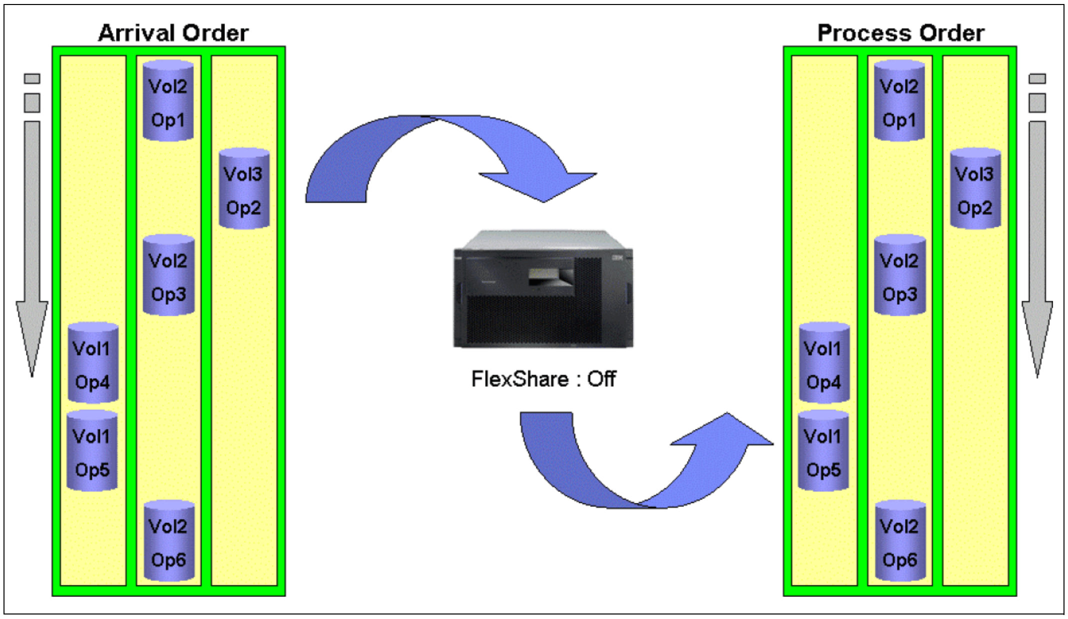


Figure 8-2 FlexShare: Off status

When FlexShare service is on, FlexShare chooses the order in which tasks are processed to best meet the priority configuration. On average, FlexShare is more likely to choose higher priority operations to be processed before lower priority operations.

Figure 8-3 shows how FlexShare can impact the order in which tasks are processed based on the priority-level configurations. The diagram illustrates an ordering of tasks when FlexShare service is on. The order in which tasks arrive is different from the order in which tasks are processed by the storage system. FlexShare orders tasks for processing taking into account the priority configuration. In this example, vol1 is a higher priority configuration than the other volumes.

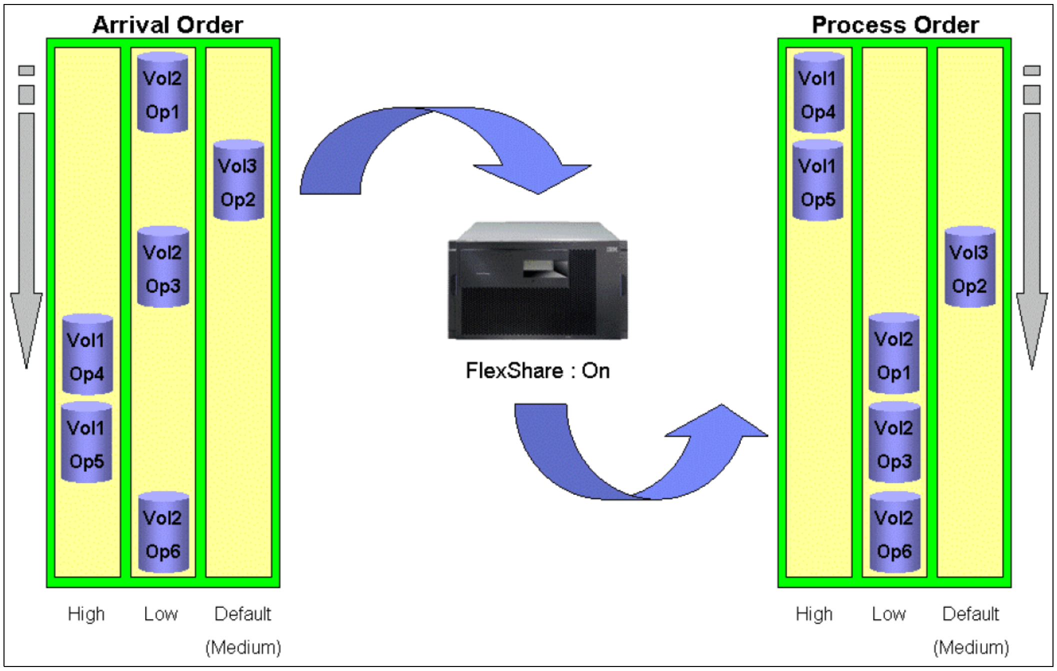


Figure 8-3 FlexShare: On status

Volume level and system priorities

FlexShare orders WAFL operations to be processed based on the following items:

1. The configured volume priority
2. The configured user versus system priority

The order of these steps is important in determining when WAFL operations are executed. First, the WAFL operations are prioritized based on the volume priorities. The priory of queue is the first factor that is considered. Then the WAFL operations are prioritized based on the configured user versus system priority. The operations in the individual queue are ordered with respect to the user versus system priority.

Figure 8-4 shows how FlexShare chooses WAFL operations to execute based on the priority level and system configurations. Vol1 is configured with a high priority level and low system priority. Vol2 is configured with a low priority level and medium system priority. Vol1 and Vol2 are the only volumes that have FlexShare priority configurations, and as a result have dedicated queues.

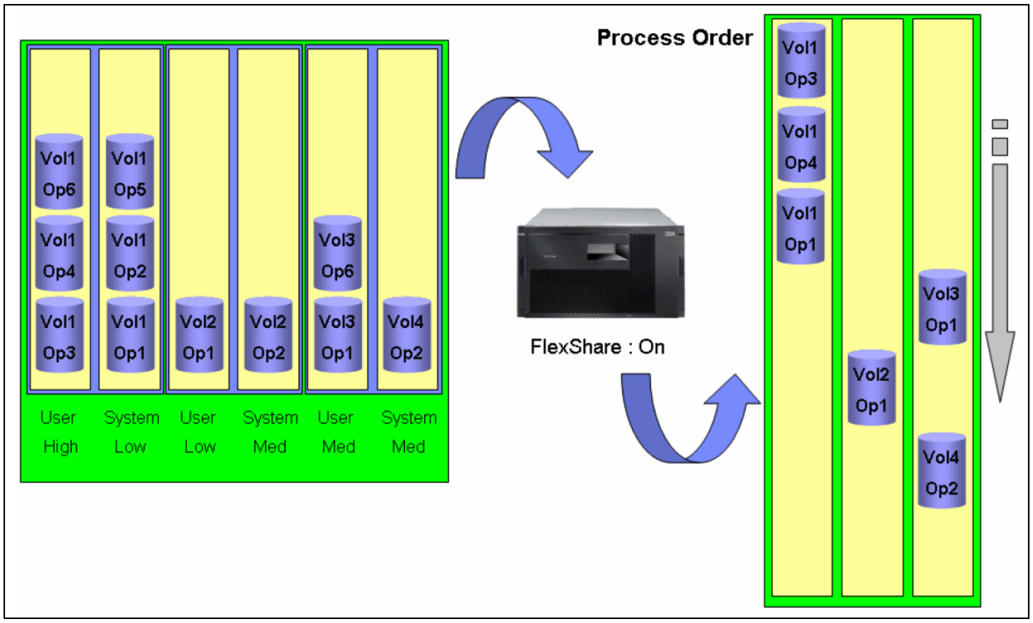


Figure 8-4 FlexShare priorities

Volume-level priorities and scheduling

FlexShare does not provide a schedule mechanism. However, it is possible to set up volume priority time using cron jobs and scripts on UNIX, as shown in Figure 8-5.

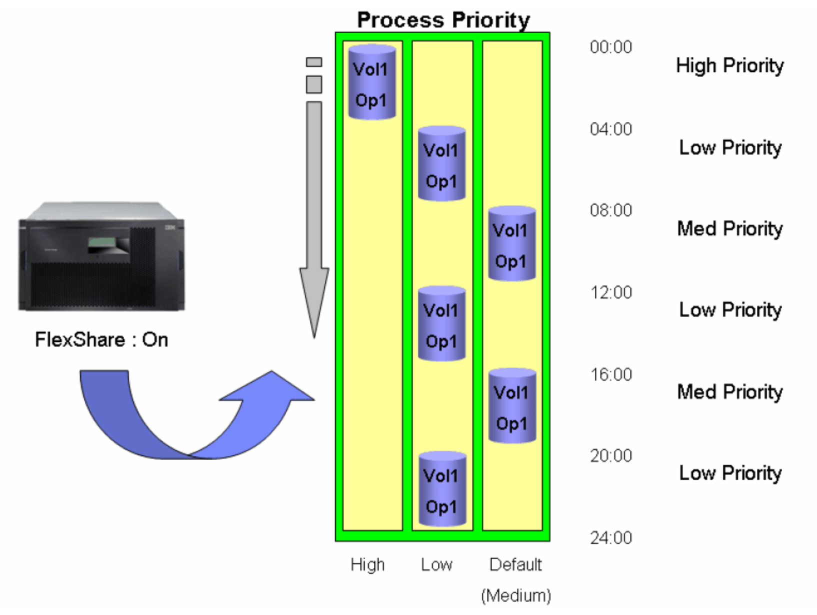


Figure 8-5 Volume level priority and scheduling

Considerations:

- FlexShare does not impact the running of WAFL operations. After a WAFL operation is dispatched to execute, FlexShare works with the WAFL operation until it is complete.
- If there is a WAFL operation that has been dispatched or is already in progress, FlexShare will not interrupt that WAFL operation, even if higher priority WAFL operations arrive in the system.
- FlexShare only controls the *order* in which WAFL operations are dispatched to be processed, but after they are dispatched, they are out of the control of FlexShare.

8.2.4 How FlexShare manages system resources

FlexShare prioritization controls the following systems resources:

- CPU: Based on volume priority level
- Disk I/O: Automatically based on volume priority level
- Non-volatile RAM (NVRAM): Automatically based on volume priority level
- Memory: Based on a per-volume cache policy

8.3 Benefits of using FlexShare

The use of FlexShare can result in many benefits, such as simplification of storage management, reduction in cost, and flexibility, as described here:

- Simplification of storage management:
  - Reduces the number of storage systems that need to be managed by enabling consolidation
  - Provides a simple mechanism for managing performance of consolidated environments
  - Easy to administer using the same command-line interface (CLI)
- Reduction in costs:
  - Allows increased capacity and processing utilization per storage system without impact to critical applications
  - No special hardware and software required
  - No additional license required
- Flexibility:
  - Can be easily customized to meet performance requirements of different workloads.

8.4 When to use FlexShare

If your storage system consistently provides the performance required for your environment, then you do not need FlexShare. If, however, your storage system sometimes does not deliver sufficient performance to some of its users, you can use FlexShare to increase your control over storage system resources to ensure that those resources are being used most effectively for your environment.

FlexShare is designed to change performance characteristics when the storage system is under load. If the storage system is not under load, it is expected that the FlexShare impact will be minimal and can even be unnoticeable.

The following examples show how you can use FlexShare to set priorities for system resource usage:

- You have a mission-critical database on the same storage system as user home directories. You can use FlexShare to ensure that database accesses are assigned a higher priority than accesses to home directories (Example 8-1).

  Example 8-1 Setting priority levels

  ```
  itsotuc1> priority set volume cifs_vol3 level=veryhigh

  itsotuc1> priority show volume -v cifs_vol3
  Volume: cifs_vol3
          Enabled: on
            Level: VeryHigh
           System: Medium
            Cache: n/a

  itsotuc1>
  ```

- You want to reduce the impact of system operations, such as SnapMirror operations, on client data accesses. You can use FlexShare to ensure that client accesses are assigned a higher priority than system operations.
- You have volumes with different caching requirements. For example, if you have a database log volume that does not need to be cached after writing, or a heavily accessed volume that must remain cached as much as possible, you can use the cache buffer policy hint to help Data ONTAP determine how to manage the cache buffers for those volumes.

Even though FlexShare enables you to construct a priority policy that helps Data ONTAP manage system resources optimally for you, it does not provide any performance guarantees. With FlexShare-enabled Data ONTAP, critical application operations will get queued first a majority of the time, but 100% priority queuing is not guaranteed.

8.5 Supported configurations

FlexShare is available starting with Data ONTAP V7.2 and is a no-charge feature. It is supported in the following environments:

- High availability (HA) (clustered failover)
- All data access protocols: NFS, CIFS, FCP, iSCSI, HTTP, and FTP
- Both volume types, such as traditional volumes and flexible volumes
- SnapMirror
- SnapVault
- SnapLock

8.6 Using FlexShare in cluster storage systems

If you use FlexShare on active/active storage systems, you must ensure that FlexShare is enabled or disabled on *both* nodes.

After a takeover occurs, as shown in Figure 8-6, the FlexShare priorities that you have set for volumes on the node that was taken over are still operational, and the takeover node creates a new priority policy by merging the policies configured on each node.

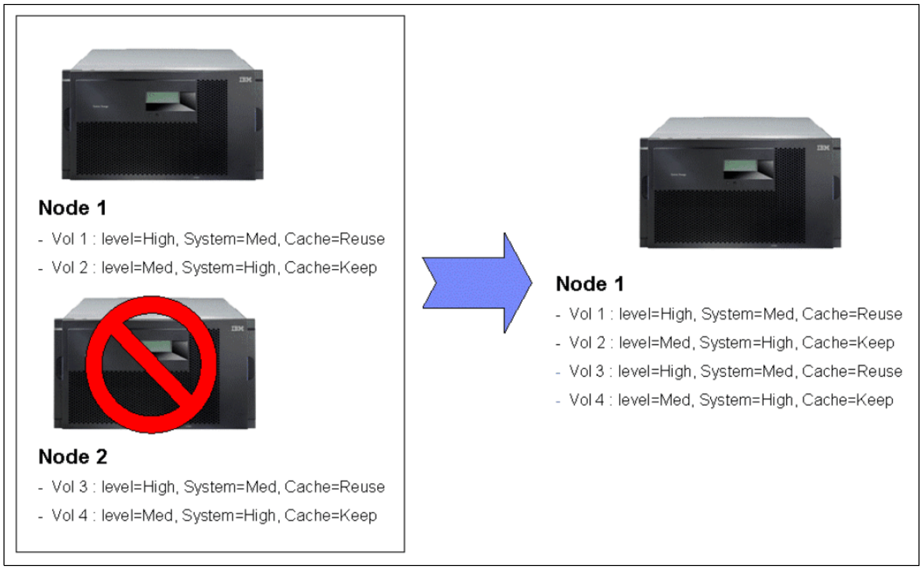

Node 1
- Vol 1 : level=High, System=Med, Cache=Reuse
- Vol 2 : level=Med, System=High, Cache=Keep
Node 2
- Vol 3 : level=High, System=Med, Cache=Reuse
- Vol 4 : level=Med, System=High, Cache=Keep
Node 1
- Vol 1 : level=High, System=Med, Cache=Reuse
- Vol 2 : level=Med, System=High, Cache=Keep
- Vol 3 : level=High, System=Med, Cache=Reuse
- Vol 4 : level=Med, System=High, Cache=Keep

Figure 8-6 Takeover priority in cluster storage systems

8.7 Setting up FlexShare

FlexShare can be administered through the CLI or the Manage ONTAP API. This section provides an overview of the typical commands and options for the CLI.

The following default values are assigned when FlexShare is initially enabled:

- Volume level: Medium
- System: Medium
- Cache: Default

However, the FlexShare configuration can be dynamically changed at any time when the system is running. Configuration changes take effect as soon as they are issued in the system and have no processing impact. Configuration changes stay active across system reboots.

8.7.1 FlexShare CLI overview

The **priority** command is the CLI command that provides all configuration and status information related to FlexShare. Issue the command with any arguments to display the priority command options (Example 8-2).

Example 8-2 Displaying priority options

```
itsotuc1> priority
The following commands are available; for more information
type "priority help <command>"
delete              off                 set                 show
help                on

itsotuc1>
```

Use the **help** option to obtain more information about a command (Example 8-3).

Example 8-3 Option help for priority

```
itsotuc1> priority help on
priority on
- Start priority policy management.

itsotuc1>
```

8.7.2 Enabling FlexShare service

To check the status of the FlexShare service, use the **show** option to verify that the FlexShare service is off by default (Example 8-4).

Example 8-4 Option show for priority

```
itsotuc1> priority show
Priority scheduler is stopped.

Priority scheduler system settings:
        io_concurrency: 8
        enabled_components: all

itsotuc1>
```

Tip: The **io_concurrency** setting represents the average number of concurrent suspended operations per disk for a volume.

Disks have a maximum number of concurrent I/O operations that they can support. The limits vary according to disk type. FlexShare limits the number of concurrent I/O operations per volume based on various values, including the volume priority and the disk type.

To enable FlexShare service, use the **on** option. To verify the status, use the **show** option again (Example 8-5).

Example 8-5 Option on for priority and verifying the status

```
itsotuc1> priority on
Wed Apr  6 21:10:42 GMT [wafl.priority.enable:info]: Priority scheduling is being
enabled
Priority scheduler starting.

itsotuc1> priority show
Priority scheduler is running.

Priority scheduler system settings:
        io_concurrency: 8
        enabled_components: all

itsotuc1>
```

To disable FlexShare service, use the **off** option (Example 8-6).

> **Important:** This option takes effect across the entire system, so use caution when changing its value and be sure to monitor system performance to ensure that performance has improved.

Example 8-6 Option off for priority and verifying the status

```
itsotuc1> priority off
Wed Apr  6 21:11:34 GMT [wafl.priority.disable:info]: Priority scheduling is being
disabled
Priority scheduler has stopped.

itsotuc1> priority show
Priority scheduler is stopped.

Priority scheduler system settings:
        io_concurrency: 8
        enabled_components: all

itsotuc1>
```

8.7.3 Priority settings

The **set** command is used to configure volume priorities. Configuration for level, system, and cache can be specified.

The level option is configured on a per-volume basis. A volume with a higher priority level is given more resources than a volume with a lower priority level.

The system option is configured on a per-volume basis. It controls the balance of system versus user priority given to a volume.

Valid options for level and system include:

- VeryHigh
- High
- Medium
- Low
- VeryLow

All the volumes with priority configurations inherit the default settings unless explicitly configured (Example 8-7).

Example 8-7 Default priorities

```
itsotuc1> priority show default -v
Default:
          Level: Medium
         System: Medium

itsotuc1>
```

Example 8-8 shows how to change the volume level priority, first to high, then to low.

Example 8-8 Change volume level priority

```
itsotuc1> priority set volume cifs_vol3 level=High

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: on
          Level: High
         System: Medium
          Cache: n/a

itsotuc1>
```

The default configuration can be modified by using the `set` command (Example 8-9).

Example 8-9 Change default system priority

```
itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: on
          Level: High
         System: Medium
          Cache: n/a

itsotuc1> priority show default -v
Default:
          Level: Medium
         System: Medium
itsotuc1> priority set default system=low

itsotuc1> priority show default -v
Default:
          Level: Medium
         System: Low

itsotuc1>
```

Priority configuration can be deleted or temporarily disabled by using the `delete` and `off` commands (Example 8-10).

Example 8-10 Delete priority configuration

```
itsotuc1> priority delete volume cifs_vol3

itsotuc1> priority show volume cifs_vol3
Unable to find priority scheduling information for 'cifs_vol3'

itsotuc1> priority set volume cifs_vol3 service=off

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: Medium
         System: Low
          Cache: n/a

itsotuc1>
```

The cache option is configured on a per-volume basis. It controls the buffer cache policy for the volume. Valid options for cache include these:

reuse This value tells Data ONTAP to make buffers from this volume available for reuse quickly. You can use this value for volumes that are written but rarely read, such as database log volumes, or volumes for which the data set is so large that keeping the cache buffers will probably not increase the hit rate.

keep This value tells Data ONTAP to wait as long as possible before reusing the cache buffers. This value can improve performance for a volume that is accessed frequently, with a high incidence of multiple accesses to the same cache buffers.

default This value tells Data ONTAP to use the default system cache buffer policy for this volume.

Example 8-11 shows how to change and how to verify the cache priority.

Example 8-11 Change and verify cache priority

```
itsotuc1> priority show default -v
Default:
          Level: Medium
         System: Low

itsotuc1> priority set volume cifs_vol3 cache=keep

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: Medium
         System: Low
          Cache: keep

itsotuc1> priority set volume cifs_vol3 level=high system=low cache=reuse

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: High
         System: Low
          Cache: reuse

itsotuc1> priority set volume cifs_vol3 cache=default

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: High
         System: Low
          Cache: n/a
itsotuc1>
```

8.8 FlexShare usage examples

This section provides examples of how you can use FlexShare to your advantage:

- In a consolidated environment
- With mixed storage, including FC and SATA

8.8.1 Consolidating different workloads

FlexShare enables storage administrators to consolidate different applications and data sets on a single storage system without impacting critical applications.

Examples of data sets that can be consolidated include these:

- Mail
- Database
- Home directories

A highly beneficial use case is to consolidate an application (such as database or mail) and home directories on the same storage system. The application can be set to a higher priority than the home directories. This prioritization protects the application workload to be preferentially treated in case the system is overloaded.

A priority configuration must be set for all the volumes of the application and all the volumes for the home directories.

Figure 8-7 shows database and home directory data residing on the same storage system, and the database is given higher priority.

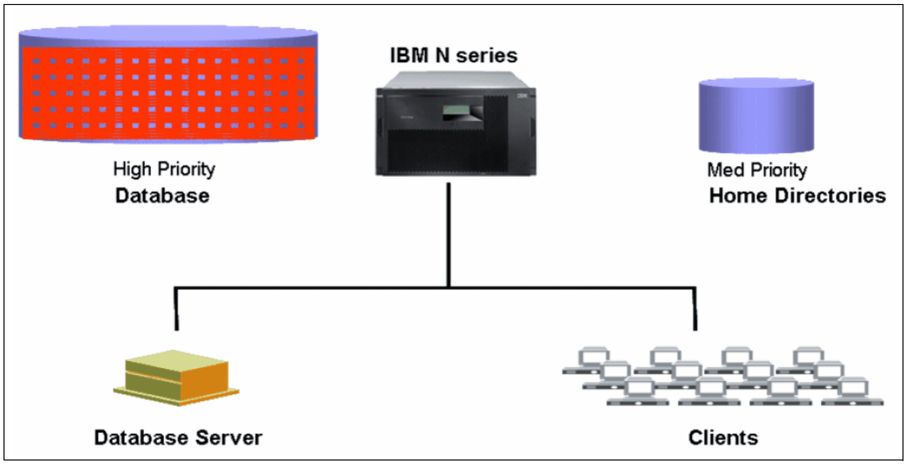


Figure 8-7 FlexShare example: Consolidated environment

Using the FlexShare service, a high-priority volume can use more system resources than other volumes (Figure 8-8).

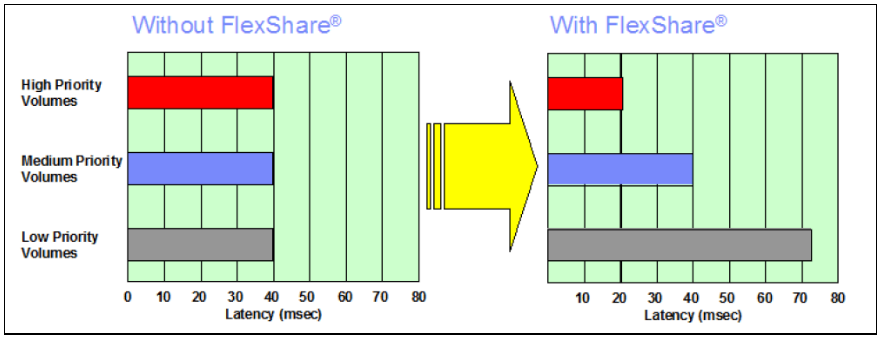


Figure 8-8 Example of FlexShare service on

However, FlexShare does not provide any performance guarantees.

8.8.2 Mixed storage including FC disks and SATA disks

It is mandatory to perform an application workload analysis to determine IOP demand and latency requirements before deploying a primary application in a storage system.

Note the following general information about mixed FC and SATA:

- FC and SATA disks must be in different shelves and on separate loops.
- Aggregates or traditional volumes cannot span both FC and SATA disks.

FlexShare enables storage administrators to prioritize data access in a mixed storage environment that includes FC and SATA disks so that high-end storage is utilized to its full extent. Storage administrators can choose to prioritize volumes on FC disks over volumes on SATA disks (Figure 8-9).

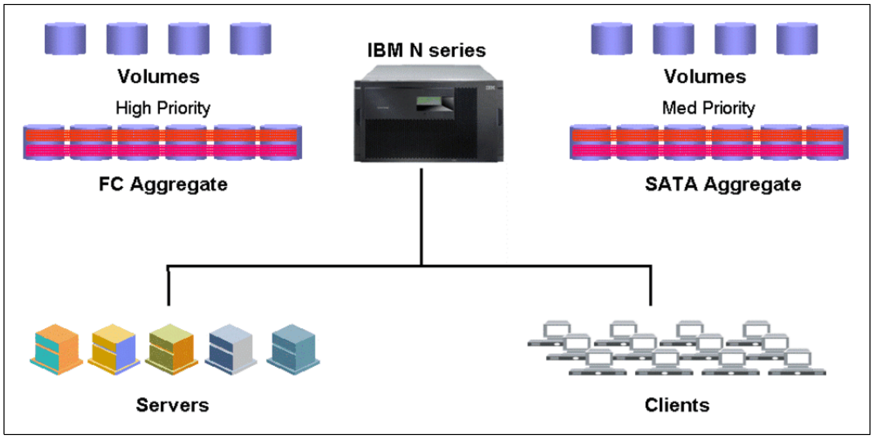


Figure 8-9 FlexShare example: Mixed storage including FC and SATA disks

8.9 FlexShare preferred practices

Following the preferred practices that we outline in this section can help ensure that your FlexShare configuration meets the highest level of performance and robustness.

8.9.1 Setting a priority configuration for all volumes in an aggregate

While volumes in an aggregate can have different priority configurations, it is important to set a priority configuration for all volumes in an aggregate explicitly. If any volume in an aggregate requires a priority configuration, set the priority configuration for all volumes in the aggregate explicitly.

Setting individual priorities is required because the performance of the storage system is more balanced if all volumes have the priority configuration. This preferred practice is based on what happens when some volumes in an aggregate have priorities and others do not. Volumes that do not have a priority configuration are treated with the default priority. The default priority processes WAFL operations from a common default processing bucket. As a result, all the tasks from the default priority volumes are processed by the same bucket, which can result in undesired performance constraints on the default priority volumes.

For example, consider a large aggregate that has 100 volumes. The large aggregate enables the 100 volumes to use all the available disk resources in the aggregate. One volume in the aggregate has a high priority configured and the other 99 volumes do not have an explicit priority configuration. FlexShare creates an independent processing bucket to prioritize operations for the high priority volume. The 99 remaining volumes are serviced by the default bucket. With this configuration, the 99 volumes can be strained easily for resource time.

Now, consider modifying the example to meet the preferred practice. The configuration will consist of 99 volumes with a medium priority and one volume with a high priority. In this case, FlexShare creates a dedicated processing bucket for all the 100 volumes, that is, one with high priority and 99 with medium priority. This configuration results in better load distribution across all the volumes.

Figure 8-10 shows the difference in the FlexShare processing buckets between a non-optimal priority configuration and an optimal priority configuration when following this preferred practice. The aggregate has 100 volumes, labeled vol1 to vol100. In this example, vol1 has a high priority configuration.

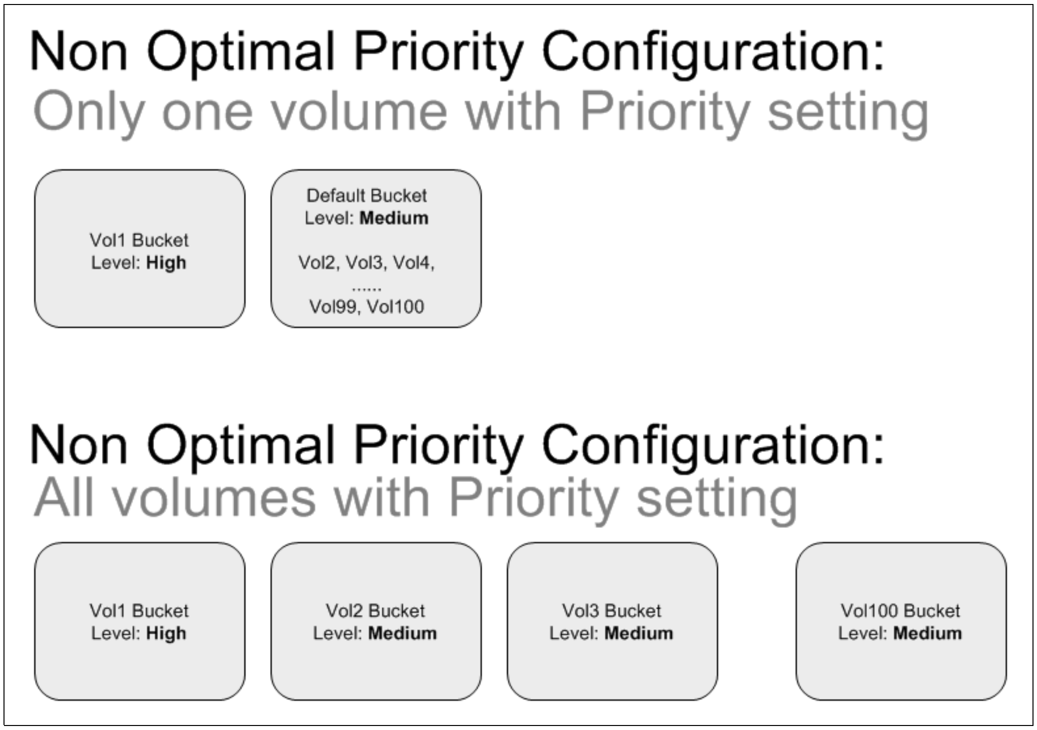


Figure 8-10 Priority configuration with processing buckets

In the non-optimal case, the default bucket processes WAFL operations for vol2 through vol100. In the preferred practice configuration, each volume in the aggregate has its own processing bucket.

8.9.2 Configuring Active/Active configurations consistently

There are some important precautions that administrators need to take into account in a Active/Active deployment:

- Both nodes of a cluster must have the same global priority on or off setting.
- The priority configuration of the individual nodes in a cluster need to be configured to meet the desired behavior in case a cluster failover occurs.

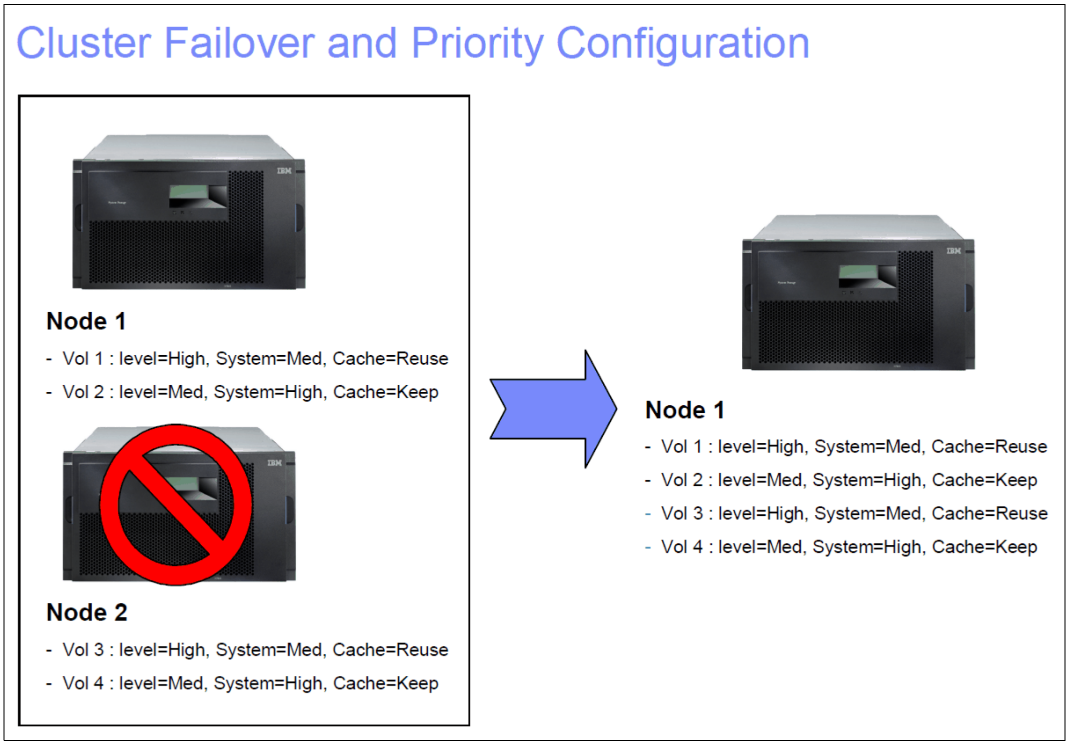
Priority setting

Set the service on or off identically on both nodes. Verify the configuration using the `priority show` command.

Cluster failover

In the event of a cluster failover, the priority schedules are merged. The priority configuration from the failed cluster node is inherited by the healthy cluster node after the cluster failover.

The priority configurations must take into account that the priorities are merged in the event of a cluster failover. In planning the priority configuration, the administrator needs to consider, if all the volumes from both storage systems were hosted on a single storage system, how must their relative priority be? The best approach is to compile a complete list of volumes in the cluster, prioritize among them, and then set the priority configuration. See Figure 8-11.



Cluster Failover and Priority Configuration

Node 1

- Vol 1 : level=High, System=Med, Cache=Reuse
- Vol 2 : level=Med, System=High, Cache=Keep

Node 2

- Vol 3 : level=High, System=Med, Cache=Reuse
- Vol 4 : level=Med, System=High, Cache=Keep

Node 1

- Vol 1 : level=High, System=Med, Cache=Reuse
- Vol 2 : level=Med, System=High, Cache=Keep
- Vol 3 : level=High, System=Med, Cache=Reuse
- Vol 4 : level=Med, System=High, Cache=Keep

Figure 8-11 Cluster failover and priority configuration

Prior to cluster failover, each node has its own independent priority configuration for its volumes. After Node 2 fails, Node1 acquires the priority configuration from the original Node1, merging Node1 and Node 2's priority configuration. After a failback, the priority configuration will be exactly like it was before the cluster failover.

8.9.3 Setting volume cache usage appropriately

A properly configured buffer cache policy can improve the cache hit rate, significantly improving overall system performance. This section explains some guidelines to make sure that the buffer cache policy is configured optimally for the environment.

Selecting the correct workloads for keep versus reuse

Configure data sets that will benefit from caching with a keep policy. Data sets that make good candidates for a keep buffer cache policy typically have active read or write workloads to a small working set relative to the storage system's buffer cache. A database that is frequently accessed with queries involving the same tables can make a good candidate for a keep policy.

It is equally important to identify and properly configure data sets that are not going to benefit from caching with a *reuse* policy (Example 8-12). This reuse policy allows space in the cache to be used optimally for the data sets that will benefit from caching. Data sets that are read once or infrequently must use the reuse policy. For example, a volume with database logs is generally written sequentially, but infrequently read. Therefore, caching database logs generally does not improve the cache hit rate. As a result, database log volumes often make good candidates for a reuse policy.

Example 8-12 Reuse policy

```
itsotuc1> priority set volume cifs_vol3 level=high system=low cache=reuse

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: High
         System: Low
          Cache: reuse

itsotuc1>
```

Importance of not overallocating the number of keep volumes

The benefit of having the keep buffer cache policy is that critical data can achieve a high percent of cache hits and also minimize the amount of data that is swapped in and out of the cache. To minimize the amount of data that is swapped in and out, it is important that the active data sets that are configured with a keep policy be smaller than the available cache size. If the active working data sets with the *keep* setting are larger than the available cache, all of the data cannot fit in the cache. As a result, the benefit of the keep policy will be diminished.

8.9.4 Tuning for SnapMirror and backup operations

SnapMirror and backup operations, including NDMP, are system operations that must be prioritized by configuring the system level for a volume (Figure 8-12).

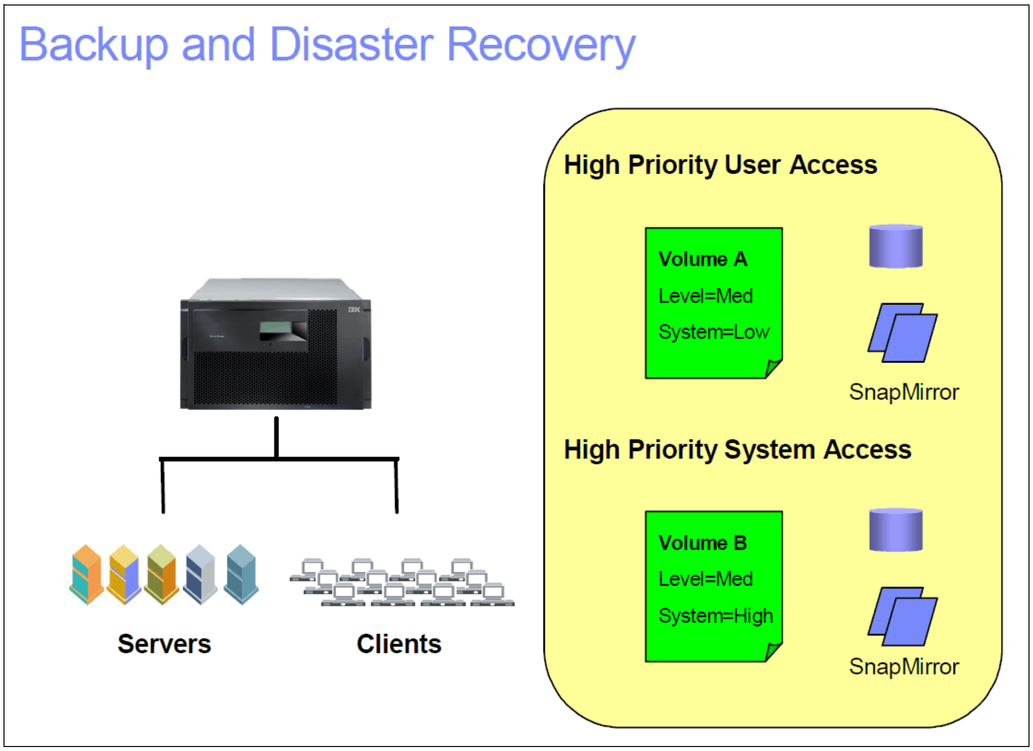


Figure 8-12 SnapMirror prioritization

The per volume system setting impacts all system activities, including SnapMirror and backup operations. FlexShare treats all SnapMirror and backup operations pertaining to a volume as a group, not as individual entities. For example, if a volume has many QSM relationships, the group of QSM relationships is prioritized by FlexShare, not each individual QSM relationship. In other words, FlexShare does not prioritize individual SnapMirror transfers. All SnapMirror transfers for a volume are prioritized together.

In some cases, storage administrators might want to control the SnapMirror or backup operations priority for an entire storage system in a generic way, without having to configure individual volume system priorities. In other cases, individual volumes will have varying requirements and will demand that the system priority be set individually for particular volumes

Understanding expected behavior

Storage administrators will be interested in prioritizing user activity compared to system activity. Some will want to give higher priority to user activity while minimizing SnapMirror and backup operation impact. This prioritization can be accomplished by setting the system priority to be lower. Keep in mind that when the system priority is reduced, the amount of time that SnapMirror transfers or other backup operations take can increase. For example, if there is a lot of user activity and the system priority is low, then the user activity is prioritized for processing above the system activity. As a result, the system activity takes longer to complete.

If you have strict timelines for particular SnapMirror and backup operations to complete, you need to tune the system priorities with caution. Closely monitor and tune the priority configuration to meet the desired behavior.

Configuring SnapMirror or backup operation priority across a storage system

Storage administrators can set global priority for SnapMirror or backup operations across an entire storage system. For example, a storage administrator might want to give higher priority to user activity compared to system activity. To accomplish this task, configure the default system value to meet the desired behavior. The default system value applies to the default bucket. All volumes that have priority configuration need to be configured explicitly to meet the user versus system configuration.

Attention: Priority configuration options that are not configured explicitly inherit the original default settings (that is, level: Medium, system: Medium, and cache: Default).

Configuring SnapMirror or backup operation priority for a volume

For volumes that have different requirements from the global system priority configuration, storage administrators will want to manually change the configuration on a per volume basis. Many environments will want to take advantage of this level of control that FlexShare provides. By configuring system priorities individually, storage administrators can give SnapMirror or backup operations different levels of priority.

For example, imagine two volumes have the same priority level but have different requirements for backup. User access to VolA is critical at all hours of the day, but user access to VolB is critical only during peak hours. The data in VolA is copied with SnapMirror hourly. The data in VolB is copied with SnapMirror nightly during off-peak hours. A delay in the amount of time it takes the backup to complete for VolA is a trade-off that has been considered and is acceptable. The backup window for VolB happens during the off-peak hours, and it is essential that the backup finishes before critical users come online. For this situation, an administrator can choose to have different system policies for VolA and VolB.

In Figure 8-13, it makes sense to have different system priority configurations for different volumes. It is essential for VolB SnapMirror operations to finish in a timely manner. Therefore, the system priority is higher. VolA SnapMirror operations can take place at a slower rate because user activity has a higher priority.

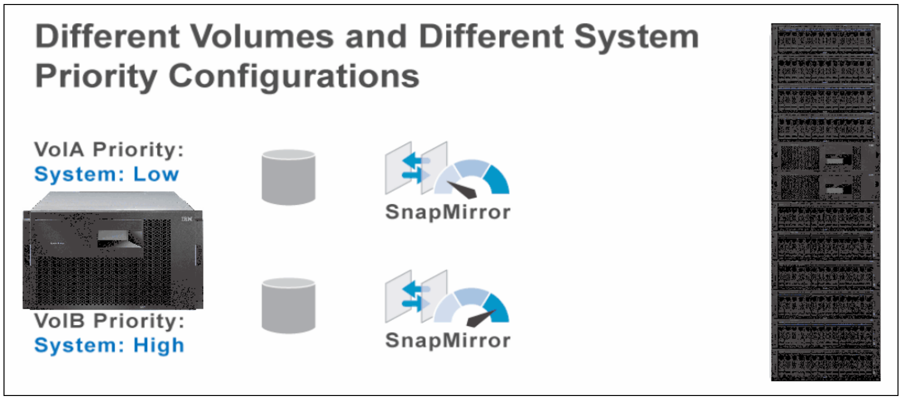


Figure 8-13 Different volumes and different system priority configurations

8.10 FlexShare administration

FlexShare can be administered using the CLI or the Manage ONTAP API. This section describes the important configuration and status commands for the CLI, the CLI commands that impact FlexShare configuration, and details about the Manage ONTAP API. The content in this section provides an overview of the typical commands and options.

Example 8-13 shows the default values that are assigned when FlexShare is initially enabled:

- Volume Level: Medium
- System: Medium

Example 8-13 Default values

```
itsotuc1> priority show default -v
Default:
          Level: Medium
         System: Medium

itsotuc1> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by IBM
         personnel.

itsotuc1*> priority show default -v
Default:
          Level: Medium
         System: Medium
User read limit: n/a
 Sys read limit: n/a
    NVLOG limit: n/a%

itsotuc1>
```

A FlexShare configuration can be dynamically changed at any time the system is running. Configuration changes take effect as soon as they are issued on the system. There is no processing impact when changing the configuration options. Configuration changes stay active across system reboots. The default values assigned by FlexShare can be modified too.

> **Tip:** Working with FlexScale and the priority family of commands is best done from Data ONTAP CLI advance mode using the command **`priv set advanced`**, which gives access to extended commands, counters, and output.

8.10.1 FlexShare CLI overview

The **`priority`** command is the CLI command that provides all configuration and status information related to FlexShare.

Basics

Issue the **`priority`** command without any arguments to display the priority command options (Example 8-14).

Example 8-14 The priority command settings

```
itsotuc1> priority
The following commands are available; for more information
type "priority help <command>"
delete              off                 set                 show
help                on

itsotuc1>
```

Use the **`help`** option to obtain more information about a command (Example 8-15).

Example 8-15 The help option

```
itsotuc1> priority help on
priority on
- Start priority policy management.

itsotuc1>
```

See the **`na_priority`** man page for more information about the **`priority`** command (Example 8-16).

Example 8-16 The na_priority man page (output shortened for clarity)

```
itsotuc1> man na_priority
na_priority(1)                                    na_priority(1)

NAME
       na_priority - commands for managing priority resources.

SYNOPSIS
       priority command argument ...

DESCRIPTION
       The  priority family of commands manages resource policies
       for the appliance.  These policies are especially applica-
```

Enabling service

To see the status of the FlexShare service, use the **show** option (Example 8-17).

Example 8-17 The show command

```
itsotuc1> priority show
Priority scheduler is stopped.

Priority scheduler system settings:
        io_concurrency: 8
        enabled_components: all

itsotuc1>
```

The FlexShare service is *off*, by default.

Important: The **io_concurrency** setting that is displayed in the **priority show** output represents the average number of concurrent suspended operations per disk for a volume. This setting is an advanced option and must not be modified unless advised by support personnel.

To enable FlexShare service, use the **on** option (Example 8-18).

Example 8-18 Setting priority on

```
itsotuc1> priority on
Wed Apr  6 21:32:37 GMT [wafl.priority.enable:info]: Priority scheduling is being
enabled
Priority scheduler starting.

itsotuc1>
```

To verify that the FlexShare service is enabled, use the **show** option (Example 8-19).

Example 8-19 Verifying FlexShare is enabled

```
itsotuc1> priority show
Priority scheduler is running.

Priority scheduler system settings:
        io_concurrency: 8
        enabled_components: all

itsotuc1>
```

To disable FlexShare service, use the **off** option (Example 8-20).

Example 8-20 Disabling FlexShare

```
itsotuc1> priority off
Wed Apr  6 21:33:43 GMT [wafl.priority.disable:info]: Priority scheduling is being
disabled
Priority scheduler has stopped.

itsotuc1>
```

Priority settings

The `set` command is used to configure volume priorities. Configuration for level, system, and cache can be specified. At least one configuration option from level, system, or cache must be specified. Options that are not explicitly set inherit the default setting.

The level option is configured on a per volume basis. A volume with a higher priority level will be given more resources than a volume with a lower priority level.

The system option is configured on a per volume basis. It controls the balance of system versus user priority given to a volume.

Valid level and system options include:

- VeryHigh
- High
- Medium
- Low
- VeryLow

The system option can also take a number as a numeric percentage from 1 to 100 for the system priority.

The cache option is configured on a per volume basis. It controls the buffer cache policy for the volume. Valid cache options include:

- **`reuse`**
- **`keep`**
- **`default`**

Example 8-21 sets the volume level priority to *High*. The volume inherits the default settings for system and cache.

Example 8-21 Setting volume priority

```
itsotuc1> priority set volume cifs_vol3 level=high

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: High
         System: Low
          Cache: reuse

itsotuc1>
```

Example 8-22 explicitly sets the level, system, and cache options.

Example 8-22 Explicitly setting priorities

```
itsotuc1> priority set volume cifs_vol3 level=Low system=Low cache=reuse

itsotuc1> priority show volume -v cifs_vol3
Volume: cifs_vol3
        Enabled: off
          Level: Low
         System: Low
          Cache: reuse

itsotuc1>
```

FlexShare maintains a default configuration that applies to the *default* processing bucket. All the volumes with priority configurations inherit the default settings unless configured explicitly (Example 8-23).

Example 8-23 Priority default configuration

```
itsotuc1> priority show default -v
Default:
         Level: Medium
        System: Medium

itsotuc1>
```

The default configuration can be modified, if desired. Default values for level, system, nvlog_limit, system_read_limit, and user_read_limit can be modified (Example 8-24).

Example 8-24 Modifying default configuration

```
itsotuc1> priority set default system=low

itsotuc1> priority show default -v
Default:
         Level: Medium
        System: Low

itsotuc1>
```

A priority configuration can be deleted, if desired (Example 8-25).

Example 8-25 Deleting priority configuration

```
itsotuc1> priority delete volume cifs_vol3

itsotuc1> priority show volume cifs_vol3
Unable to find priority scheduling information for 'cifs_vol3'
itsotuc1>
```

8.10.2 Expected behavior with other CLI commands

Table 8-2 shows how common CLI commands impact the FlexShare priority configuration.

Table 8-2 Command reference

CLI command	Priority configuration outcome
`vol rename`	The priority configuration is unchanged.
`vol copy`	The destination volume will be assigned the default priority configuration. The source volume’s priority configuration will be unchanged.
`vol clone`	The cloned volume will be assigned the default priority configuration. The source volume’s priority configuration will be unchanged.
`vol online/vol offline`	The priority configuration is unchanged. A volume online or offline will automatically trigger FlexShare to rebalance system resource limits based on the current online volumes in the aggregate.
`vol destroy`	The priority configuration for the volume is permanently removed.

8.10.3 Managing Data ONTAP API

The FlexShare configuration and status can be administered using the Manage ONTAP API. The complete functionality to configure and retrieve the status is available from the Manage ONTAP API.

8.10.4 Upgrading and reverting

FlexShare configuration is safely stored in the Data ONTAP registry. An upgrade preserves the FlexShare priority configuration and the configuration becomes active automatically. If the Data ONTAP version is reverted to a previous version that does not support FlexShare, the FlexShare configuration is ignored without any impact.

8.11 Understanding FlexShare behavior and troubleshooting

The previous sections provide information about how to plan and configure FlexShare. This section focuses on aspects after FlexShare has been configured:

- Using counters to analyze FlexShare behavior
- Troubleshooting
- Maintaining optimal priority configurations

8.11.1 Using counters to analyze FlexShare behavior

FlexShare has a number of advanced diagnostic *counters* that are useful in analyzing FlexShare behavior. These counters provide valuable insight into how FlexShare is operating. They are advanced and only available in the advanced mode.

> **Tip:** The FlexShare counters are only available in the advanced mode.

Counter terminology

The following terminology is frequently used in the counters:

- *Pending*: Waiting to run in FlexShare
- *Scheduled*: Dispatched by FlexShare to WAFL
- *Queued*: Waiting to be scheduled, received by FlexShare, and intentionally being queued

Commands

To retrieve the counters that we describe in this section, issue the following commands:

- **`stats show prisched`**
- **`stats show priorityqueue`**

Counters from **`stats show prisched`** are referred to as *prisched* object counters. The prisched object counters provide information about the total number of operations queued by FlexShare.

Counters from **`stats show priorityqueue`** are referred to as *priorityqueue* object counters. The priorityqueue object counters provide detailed information about each individual processing bucket, or priority queue, including its configuration and performance statistics.

Example 8-26 provides sample output of the FlexShare counters.

Example 8-26 FlexShare counters

```
itsotuc1*> stats show prisched
prisched:prisched:queued:0
prisched:prisched:queued_max:10

itsotuc1*> stats show priorityqueue
priorityqueue:cifs_vol3:weight:50
priorityqueue:cifs_vol3:usr_weight:22
priorityqueue:cifs_vol3:usr_sched_total:0/s
priorityqueue:cifs_vol3:usr_pending:0
priorityqueue:cifs_vol3:avg_usr_pending_ms:0ms
priorityqueue:cifs_vol3:usr_queued_total:0/s
priorityqueue:cifs_vol3:sys_sched_total:4/s
priorityqueue:cifs_vol3:sys_pending:0
priorityqueue:cifs_vol3:avg_sys_pending_ms:0.00ms
priorityqueue:cifs_vol3:sys_queued_total:4/s
priorityqueue:cifs_vol3:usr_read_limit:3
priorityqueue:cifs_vol3:max_user_reads:0
priorityqueue:cifs_vol3:sys_read_limit:6
priorityqueue:cifs_vol3:max_sys_reads:0
priorityqueue:cifs_vol3:usr_read_limit_hit:0
priorityqueue:cifs_vol3:sys_read_limit_hit:0
priorityqueue:cifs_vol3:nvlog_limit:120121344
priorityqueue:cifs_vol3:nvlog_used_max:0
priorityqueue:cifs_vol3:nvlog_limit_full:0
priorityqueue:(default):weight:50
priorityqueue:(default):usr_weight:50
priorityqueue:(default):usr_sched_total:0/s
priorityqueue:(default):usr_pending:0
priorityqueue:(default):avg_usr_pending_ms:0ms
priorityqueue:(default):usr_queued_total:0/s
priorityqueue:(default):sys_sched_total:54/s
priorityqueue:(default):sys_pending:0
priorityqueue:(default):avg_sys_pending_ms:0.00ms
priorityqueue:(default):sys_queued_total:54/s
priorityqueue:(default):usr_read_limit:28
priorityqueue:(default):max_user_reads:0
priorityqueue:(default):sys_read_limit:28
priorityqueue:(default):max_sys_reads:1
priorityqueue:(default):usr_read_limit_hit:0
priorityqueue:(default):sys_read_limit_hit:0
priorityqueue:(default):nvlog_limit:120121344
priorityqueue:(default):nvlog_used_max:0
priorityqueue:(default):nvlog_limit_full:0

itsotuc1*>
```

Basic information about counters

The FlexShare counters fall into two categories (see Table 8-3):

- *Configuration* counters provide information about internal configuration, including how FlexShare translates user configured priority settings and limits on system resources.
- *Performance* counters provide information about how the system is performing.

The priorityqueue object refers to an instance name, which is the priority queue name. The priority queue name is either the volume name (or default) for the default priority queue.

Table 8-3 Priority queue counters

Object	Counter	Description
priorityqueue	weight	The relative weight of this queue compared to other queues. The value can be in the range of 0 to 100.
priorityqueue	usr_weight	The relative weight of user operations compared to system operations. The value can be in the range of 0 to 100.
priorityqueue	nvlog_limit	The maximum amount of NVLOG, measured in bytes, the queue can use during a CP.
priorityqueue	usr_read_limit	The maximum number of concurrent user reads allowed.
priorityqueue	sys_read_limit	The maximum number of concurrent system reads allowed. (See also max_sys_reads.)
prisched	queued	The number of operations currently queued in FlexShare waiting to be scheduled.
prisched	queued_max	The maximum number of operations queued in FlexShare at the same time.
priorityqueue	nvlog_used_max	The maximum amount of NVLOG the queue has used during a CP. (See also nvlog_limit.)
priorityqueue	max_user_reads	The maximum number of user reads that have been outstanding on the queue since FlexShare was enabled or the queue was created.
priorityqueue	max_sys_reads	The maximum number of system reads that have been outstanding on the queue since FlexShare was enabled or the queue was created.
priorityqueue	usr_sched_total	The total number of scheduled user operations per second.
priorityqueue	usr_queued_total	The total number of queued user operations per second.
priorityqueue	avg_usr_pending_ms	The average pending time for user operations in milliseconds.
priorityqueue	usr_pending	The current number of pending user operations.
priorityqueue	sys_sched_total	The total number of scheduled system operations per second.
priorityqueue	sys_queued_total	The total number of queued system operations per second.

Object	Counter	Description
priorityqueue	avg_sys_pending_ms	The average pending time for system operations in milliseconds.
priorityqueue	sys_pending	The current number of pending system operations.

8.11.2 Troubleshooting

The motivation of most, if not all, FlexShare troubleshooting is to validate that the FlexShare configuration is impacting the appropriate tasks and with the appropriate level of priority. There are a number of tips that can assist in any troubleshooting effort.

When FlexShare impact is expected

FlexShare is designed to change performance characteristics when the storage system is under load. If the storage system is not under load, it is expected that the FlexShare impact will be minimal and might even be unnoticeable.

Knowing how FlexShare works and assessing the expected behavior are important first steps in any FlexShare diagnosis.

Using the diagnostic counters

The diagnostic counters described in 8.11.1, “Using counters to analyze FlexShare behavior” on page 114 provide the most in-depth analysis into how FlexShare is internally configured and how FlexShare is performing.

Review the counters and look for the following cases:

- General system performance:
  - Check usr_sched_total and sys_sched_total for each queue to see how many operations are being dispatched by FlexShare to WAFL per second. The sum of usr_sched_total and sys_sched_total for each queue provides the total number of operations being scheduled per second for the queue. Reviewing this information gives a general overview of how many operations are executing relative to each queue.
  - Review the avg_usr_pending_ms and avg_sys_pending_ms counters. Higher priority volumes typically have values of zero or close to zero for avg_usr_pending_ms and avg_sys_pending_ms counters. Lower priority volumes can expect to have higher avg_usr_pending_ms and avg_sys_pending_ms, especially when the system is under load.
- Read performance troubleshooting:

  Compare max_user_reads with usr_read_limit and compare max_sys_reads with sys_read_limit to see if the storage system is frequently running into an I/O limitation. Volumes with lower priority are more likely to reach the respective thresholds for read and write operations. A storage system can have no volumes encountering a read or write threshold if FlexShare determines that the current system performance does not require restrictions on I/O performance.

- Write performance troubleshooting:

  Compare nvlog_used_max with nvlog_limit to see whether the NVLOG throttling is impacting writes (Example 8-27). If FlexShare is restricting write performance of a particular queue, the nvlog_used_max will be greater than or equal to the nvlog_limit for the respective queue.

  Example 8-27 Output command using stats show priorityqueue

  ```
  priorityqueue:(default):nvlog_limit::22020151
  priorityqueue:(default):nvlog_used_max:60848
  ```

- User versus system troubleshooting:

  Check avg_usr_pending_ms and avg_sys_pending_ms to see how FlexShare is preferentially processing user versus system operations for an individual queue. Volumes with higher system priority can expect the avg_sys_pending_ms will be smaller than avg_usr_pending_ms. Volumes with lower system priority can expect that the avg_usr_pending_ms will be smaller than avg_sys_pending_ms. This behavior will be more noticeable when the queue has many simultaneous operations arriving in the system.

8.11.3 FlexShare off versus on

In some troubleshooting scenarios, it might be a useful option to observe the difference when FlexShare is off versus when FlexShare is on. The administrator needs to assess if this option is a viable troubleshooting option for the environment.

Perform the following steps to isolate potential problems on a storage system:

1. Review the system performance characteristics when the FlexShare service is turned off:
   a. Outline the FlexShare configuration that yields the desired system priority configuration.
   b. Outline the expected performance changes.
2. Turn the FlexShare service on:
   a. Verify that the FlexShare configuration matches the designed configuration from Step 1a.
3. Review the system performance characteristics when the FlexShare service is turned on:
   a. Review the diagnostic counters, paying particular attention to the difference in volumes that have priority configurations.
   b. Identify any performance changes in the system performance.
   c. Verify if the performance changes meet the expectations from Step 1b.

If there are no performance bottlenecks on the storage system in Step 1 on page 118, it is unlikely that any major changes will occur when the FlexShare service is enabled.

8.11.4 Maintaining priority configurations

Maintaining a storage system to perform at its optimal performance level is an ongoing task. Meeting the existing priority requirements can sometimes take a few iterations to optimize for an environment. In addition, as existing priority requirements change due to new application deployments or data consolidations, a storage administrator will need to tune appropriately the FlexShare priority configurations.

Adhering to a systematic methodology for tuning priority configurations will result in the fewest misconfigurations. Perform the following steps to tune the FlexShare priority configuration:

1. Review the current system performance characteristics:
   a. Review the existing FlexShare priority configuration.
   b. Review the diagnostic counters.
   c. Outline the FlexShare configuration changes that are required to yield the desired system priority configuration.
   d. Outline the expected performance changes.
2. Make FlexShare configuration changes.
3. Review the effect of the FlexShare configuration changes:
   a. Review the FlexShare priority configuration to make sure it matches the outlined plan from Step 1c.
   b. Review the diagnostic counters, paying particular attention to the differences from Step 1b.
4. Assess if priority performance meets desired goals:
   a. If yes: Plan to re-evaluate priority configuration tuning at recurring times in the future and after major changes, including new application deployments and data consolidations.
   b. If no: Review the desired goals to make sure that they are realistic. Go back to Step 1.

8.12 Summary

FlexShare is a powerful Data ONTAP feature that enables storage administrators to implement workload prioritization on a storage system. It provides administrators with the ability to configure volume priority levels, user versus system priorities, and caching policies. FlexShare has significant intelligence to control and protect critical system resources.

The information that we present here provides details about the FlexShare design, administration, preferred practices, troubleshooting, and high benefit use cases. Administrators are encouraged to review this material and understand the impact of configuring FlexShare. After assessing an environment's performance objectives and reviewing this material, the storage administrator will be better prepared to configure FlexShare and obtain optimal storage systems performance in their environment.

9

Network configuration

Your storage system supports physical network interfaces, such as Ethernet and Gigabit Ethernet interfaces, and virtual network interfaces, such as interface group and Virtual Local Area Network (VLAN). Each of these network interface types has its own naming convention.

The storage system supports the following types of physical network interfaces:

- 10/100/1000 Ethernet
- Gigabit Ethernet (GbE)
- 10 Gigabit Ethernet

In addition, some storage system models have a physical network interface named e0M. It is a low-bandwidth interface of 100 Mbps and is used only for Data ONTAP management activities, such as running a Telnet, SSH, or RSH session.

The following topics are covered:

- Network interfaces
- Configuring network interfaces
- How routing data in Data ONTAP works
- Interface groups
- Ways to improve your storage system's performance

9.1 Network interfaces

This section describes the network capabilities and interfaces on your N series storage system.

9.1.1 Maximum number of network interfaces

Beginning with Data ONTAP 7.3, a storage system can contain 256 to 1,024 network interfaces per system, depending on the storage system model, system memory, and whether they are in an HA pair.

9.1.2 Maximum number of interface groups

The number of physical interfaces depends on the storage system model. Each storage system can support up to 16 interface groups. For the maximum number of network interfaces that each system can support, the total number of interfaces can include physical, interface group, VLAN, VH, and loopback interfaces.

9.2 Configuring network interfaces

You can configure network interfaces either during system setup or when the storage system is operating. When the storage system is operating, you can use the **ifconfig** command to assign or modify configuration values of your network interfaces.

During system setup, you can configure the IP addresses for the network interfaces. An **ifconfig** command is included in the /etc/rc file of the root volume for each network interface that you configured during the system setup. After your storage system has been set up, the **ifconfig** commands in the /etc/rc file are used to configure the network interfaces on subsequent storage system reboots.

You can use the **ifconfig** command to change values of parameters for a network interface when your storage system is operating. However, such changes are not automatically included in the /etc/rc file. If you want your configuration modifications to be persistent after a reboot, you must include the **ifconfig** command values in the /etc/rc file.

9.2.1 Configuring a partner interface in an HA pair

To prepare for a successful takeover in an HA configuration, you can map a network interface to an IP address or to another network interface on the partner node. During a takeover, the network interface on the surviving node assumes the identity of the partner interface.

When specifying the partner IP address, you must ensure that both the local network interface and the partner? network interface are attached to the same network segment or network switch.

Depending on the partner configuration that you want to specify, enter one of the following commands.

To specify a partner IP address, enter the following command:

```
ifconfig interface_name partner address
```

interface_name is the name of the network interface, and *address* is the partner IP address.

To set the partner interface name, enter the following command:

```
ifconfig interface_name partner partner_interface
```

partner_interface is the name of the partner network interface.

9.2.2 Enabling or disabling automatic takeover for a network interface

You can enable or disable negotiated failover for a network interface to trigger automatic takeover if the interface experiences a persistent failure. You can use the **nfo** option of the **ifconfig** command to enable or disable negotiated failover.

You can specify the nfo option for an interface group. However, you cannot specify the nfo option for any underlying physical interface of the interface group.

To enable takeover on interface failures, enter the following command:

```
options cf.takeover.on_network_interface_failure on
```

To enable or disable negotiated failover, enter the following command:

```
ifconfig interface_name {nfo|-nfo}
```

interface_name is the name of the network interface.

nfo enables negotiated failover.

-nfo disables negotiated failover.

9.3 How routing data in Data ONTAP works

You can have Data ONTAP route its own outbound packets to network interfaces. Although your storage system can have multiple network interfaces, it does not function as a router. However, it can route its outbound packets.

9.3.1 Routing mechanisms

Data ONTAP uses two routing mechanisms:

- Fast path: Data ONTAP uses this mechanism to route NFS packets over UDP and to route all TCP traffic. By default, fast path is enabled on the storage system.
- Routing table: To route IP traffic that does not use fast path, Data ONTAP uses the information available in the local routing table. The routing table contains the routes that have been established and are currently in use, as well as the default route specification.

9.3.2 How fast path works

Fast path is an alternative routing mechanism to the routing table. In fast path, the responses to incoming network traffic are sent back by using the same interface as the incoming traffic. By avoiding the routing table lookup, fast path provides a quick access to data.

If fast path is enabled on an interface group and a physical interface in that group receives an incoming request, the same physical interface might not send a response to the request. Instead, any other physical interface in an interface group can send the response.

How fast path works with NFS/UDP

NFS/UDP traffic uses fast path only when sending a reply to a request. The reply packet is sent out on the same network interface that received the request packet. For example, a storage system named toaster uses the toaster-e1 interface to send reply packets in response to NFS/UDP requests received on the toaster-e1 interface.

Fast path is used only in NFS/UDP. However, fast path is not used in other UDP-based NFS services such as portmapper, mountd, and nlm.

How fast path works with TCP

In a TCP connection, fast path is disabled on the third retransmission and the consecutive retransmissions of the same data packet. If Data ONTAP initiates a connection, Data ONTAP can use fast path on every TCP packet transmitted, except the first SYN packet. The network interface that is used to transmit a packet is the same interface that received the last packet.

9.4 Interface groups

An interface group is a feature in Data ONTAP that implements link aggregation on your storage system. Interface groups provide a mechanism to group together multiple network interfaces (links) into one logical interface (aggregate). After an interface group is created, it is indistinguishable from a physical network interface.

Data ONTAP connects with networks through physical interfaces (or links). The most common interface is an Ethernet port, such as e0a, e0b, e0c, and e0d.

IEEE 802.3ad link aggregation is now supported by using interface groups. They can be single mode or multimode. In a single mode interface group, one interface is active while the other interface is on standby. In single mode, a failure signals the inactive interface to take over and maintain the connection with the switch.

In multimode, all interfaces are active and share the same MAC address. Multimode operation has two types of operation:

- Static: ‘multi’
- Dynamic: ‘lacp’

The **ifgrp** command refers to this setting as **multi**. The multimode static interface group implementation complies with the IEEE 802.3ad static standard, while the multimode dynamic life is compliant with the IEEE 802.3ad dynamic standard, also called Link Aggregation Control Protocol or LACP. Dynamic multimode interface groups can detect the loss of link status, as well as a loss of data flow. However, a compatible switch must be used to implement the dynamic multimode configuration. Example 9-1 shows options available to the **ifgrp** command.

Example 9-1 The ifgrp command

```
TUCSON1> ifgrp
Usage:
        ifgrp create [single|multi|lacp] <ifgrp_name> -b [rr|mac|ip]
[<interface_list>]
        ifgrp add <ifgrp_name> <interface_list>
        ifgrp delete <ifgrp_name> <interface_name>
        ifgrp destroy <ifgrp_name>
        ifgrp {favor|nofavor} <interface>
        ifgrp status [<ifgrp_name>]
        ifgrp stat <ifgrp_name> [interval]
```

9.4.1 Types of interface groups

You can create three different types of interface groups on your storage system. These are single mode, static multimode, and dynamic multimode interface groups.

Each interface group provides different levels of fault tolerance. Multimode interface groups provide methods for load balancing network traffic.

Single mode interface group

In a single mode interface group, only one of the interfaces in the interface group is active. The other interfaces are on standby, ready to take over if the active interface fails. All interfaces in a single mode interface group share a common MAC address.

There can be more than one interface on standby in a single mode interface group. If an active interface fails, your storage system randomly picks one of the standby interfaces to be the next active link. The active link is monitored and link failover is controlled by the storage system; therefore, single mode interface group does not require any switch configuration. Single mode interface groups also do not require a switch that supports link aggregation.

If a single mode interface group spans over multiple switches, you must connect the switches with an inter-switch link (ISL). For a single mode interface group, the switch ports must be in the same broadcast domain (for example, a LAN or a VLAN). Link-monitoring ARP packets, which have a source address of 0.0.0.0, are sent over the ports of a single mode interface group to detect whether the ports are in the same broadcast domain.

Static multimode interface group

The static multimode interface group implementation in Data ONTAP is in compliance with IEEE 802.3ad (static). Any switch that supports aggregates, but does not have control packet exchange for configuring an aggregate, can be used with static multimode interface groups.

Static multimode interface groups do not use IEEE 802.3ad (dynamic), also known as Link Aggregation Control Protocol (LACP) or Port Aggregation Protocol (PAgP), the proprietary link aggregation protocol from Cisco.

Dynamic multimode interface group (LACP)

Dynamic multimode interface groups, also known as Link Aggregation Control Protocol (LACP), can detect the loss of link status and the inability of the storage controller to communicate with the direct-attached switch port. Implementation of LACP enables dynamic multimode interface groups that are compatible with HA environments.

Dynamic multimode interface group implementation in Data ONTAP is in compliance with IEEE 802.3 AD (802.1 AX). Data ONTAP does not support Port Aggregation Protocol (PAgP), which is a proprietary link aggregation protocol from Cisco.

A dynamic multimode interface group requires a switch that supports LACP.

Data ONTAP implements LACP in nonconfigurable active mode that works well with switches that are configured in either active or passive mode. Data ONTAP implements the long and short LACP timers for use with nonconfigurable values (3 seconds and 90 seconds, as specified in IEEE 802.3 AD (802.1AX).

The Data ONTAP load-balancing algorithm determines the member port to be used to transmit outbound traffic and does not control how inbound frames are received. The switch determines the member (individual physical port) of its port channel group to be used for transmission, based on the load-balancing algorithm configured in the switch's port channel group. Therefore, the switch configuration determines the member port (individual physical port) of the storage system to receive traffic. For more information about configuring the switch, see the switch vendor's documentation.

9.4.2 Load balancing in multimode interface groups

You can ensure that all interfaces of a multimode interface group are equally utilized for outgoing traffic by using the IP address, MAC address, round-robin, or port based load-balancing methods to distribute network traffic equally over the network ports of a multimode interface group.

The load-balancing method for a multimode interface group can be specified only when the interface group is created. If no method is specified, the IP address based load-balancing method is used.

9.5 Ways to improve your storage system's performance

You can improve your storage system's performance by performing certain configuration procedures, such as using interface groups, correcting duplex mismatches, and upgrading to Ethernet interfaces.

The following configuration procedures can improve the performance of your storage system:

1. Using static or dynamic multimode interface groups to aggregate the bandwidth of multiple interfaces
2. Using jumbo frames with your network interfaces to reduce CPU processing overhead
3. Upgrading to a faster network interface
4. Correcting duplex mismatches on 10Base-T or 100Base-T Ethernet networks
5. Using iSCSI multiconnection sessions to balance the load across interfaces
6. Enabling fast path on your storage system
7. Blocking data traffic on the dedicated management interface

10

MultiStore

This chapter introduces MultiStore and vFiler, an optional software solution that enables secure, multiprotocol storage consolidation across enterprises. It provides secure partitioning of network and storage resources and enables storage consolidation for multi-domain and multi-server configurations. In addition, it reduces management costs by reducing the number of storage systems that storage administrators must administer.

The following topics are covered:

- Introduction to vFiler
- vFiler benefits
- vFiler scenarios

10.1 Introduction to vFiler

vFiler is the logical partitioning of the resources of an IBM System Storage N series storage system, as shown in Figure 10-1. These resources include network addresses and storage, such as volumes and qtrees. It means that different departments can share the network and storage resources on a single IBM System Storage N series storage system, while maintaining independent domains.

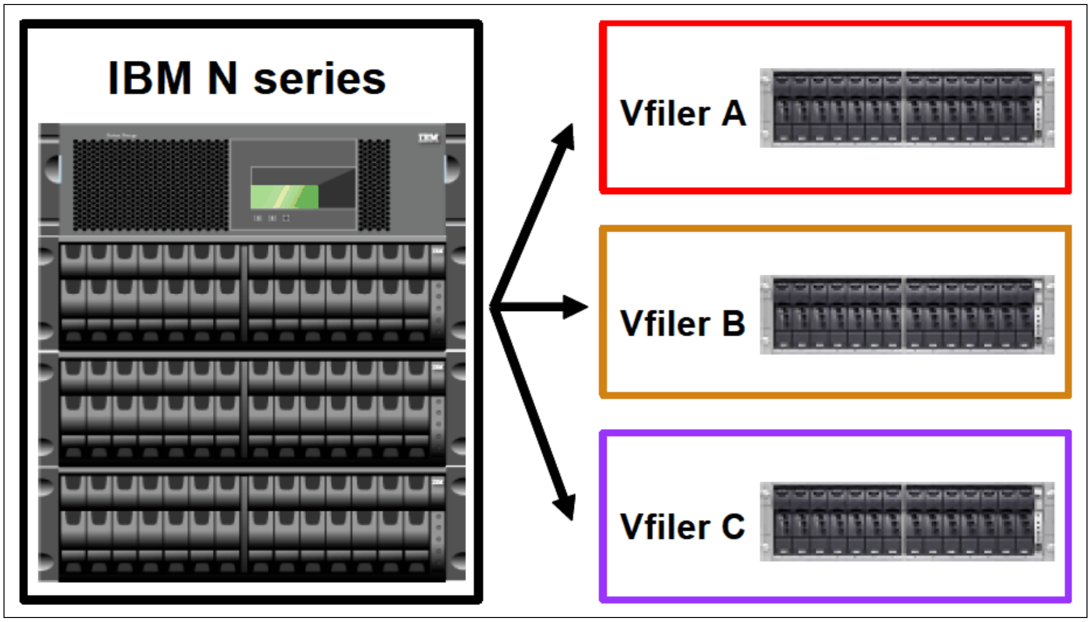


Figure 10-1 The vFiler concept

In particular, each vFiler has its own security domain, and each of the departments that sees its own vFiler is entirely unaware that it is sharing the same physical IBM System Storage N series storage system with other departments. It means that no data flow must exist between vFilers.

Using vFiler, storage and networking resources can be effectively partitioned and dynamically assigned to virtual storage systems. It virtualizes the physical resources and moves beyond the logical architectural limitations inherent in a single physical storage system.

10.1.1 Number of vFiler units allowed

There are limits to the number of vFiler units allowed in a storage system that has the MultiStore license enabled. You can usually have a maximum of 65 vFiler units on a storage system. However, the maximum limit depends on the memory capacity of the hosting storage system.

You can create 64 vFiler units on a storage system. The 65th vFiler unit is vfiler0, which is created automatically when MultiStore is licensed on the storage system. The default vFiler unit exists as long as MultiStore is licensed.

In an HA pair, you can create up to 64 vFiler units on each node of the HA pair, for a maximum of 130 vFiler units in the HA pair.

Attention: These limits can be exceeded only during a takeover scenario, when one storage system takes over the resources of a vFiler unit in another storage system.

Up to 65 vFilers can be created and hosted on a single storage system, each serving data as a storage system does, as shown in Table 10-1.

Table 10-1 Capacity

Storage subsystem capacity	vFiler limit
1 GB or more	26
2 GB or more	65

Tips:

- You can create a maximum of 16 vFiler units in N3400 systems.
- You can use the `sysconfig -v` command to verify the memory size of your system.

10.1.2 vFiler supported protocols

The following protocols are supported on the vFiler:

Restriction: The Fibre Channel (FC) protocol is not supported in vFiler environments.

- Network File System (NFS):

  A vFiler supports the same NFS functions as an IBM System Storage N series storage system. There are some vFiler-based NFS commands supported for each vFiler.
- Common Internet File System (CIFS):

  A vFiler provides the same CIFS support as an IBM System Storage N series storage system.
- iSCSI:

  A vFiler provides the same iSCSI support as an IBM System Storage N series storage system.
- RSH:

  The RSH protocol can be enabled or disabled for the vFilers. If enabled, it is configured for vFiler use just as it is for the physical IBM System Storage N series storage system. However, the set of commands available is severely restricted when RSH is used.

  Tip: You cannot Telnet to a vFiler. You can use RSH and MicroSoft tools such as MMC, Server Manager, and User Manager, which allow you to handle shares, local groups, and so on.

10.1.3 Using vFiler

vFiler is typically used for the following reasons:

- You want to consolidate multiple servers to one storage system.
- You use the storage system to host data for multiple customers, such as clients of a service provider or different organizations within an enterprise (Figure 10-2).

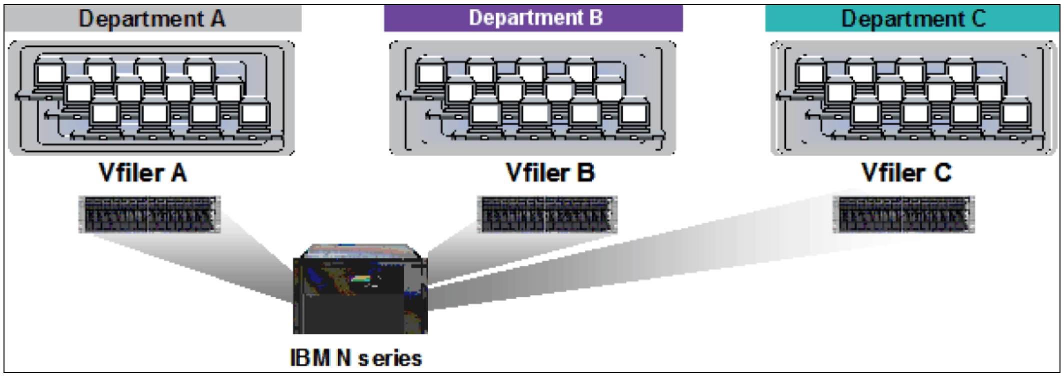


Figure 10-2 vFiler for storage consolidation

- You use the SnapMirror technology to migrate data from one storage system to another transparently (Figure 10-3).

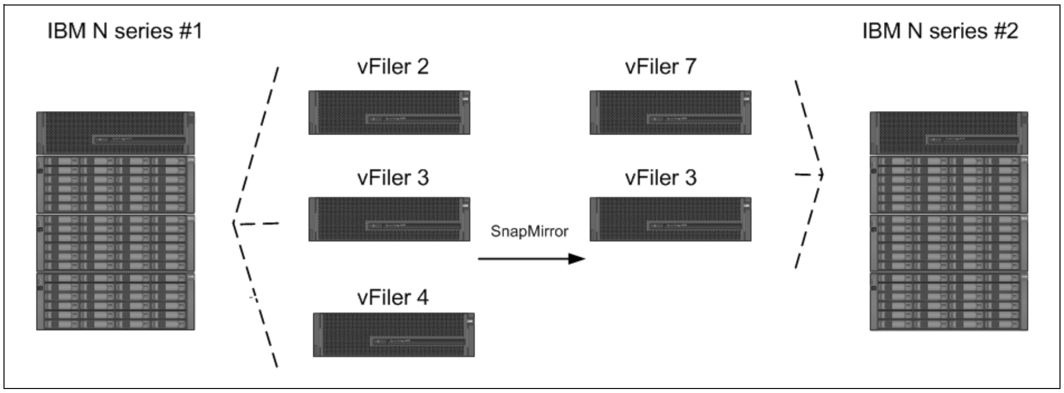


Figure 10-3 vFiler for migration with SnapMirror

10.2 vFiler benefits

A vFiler, logically separating the resources of a physical storage system, provides the following benefits:

- Virtualization:

  Virtualization provides a layer of abstraction, decoupling the physical resources like CPU and system memory of the physical storage system. It provides a logical view of both storage and computing resources. Virtualization hides the complexity and simplifies storage and system management.

- Consolidation:

  A vFiler provides an efficient architecture for consolidating multiple physical storage systems into a smaller number of systems. From the user's perspective, each vFiler appears as a separate physical storage system with a unique IP address.

- Security:

  A vFiler provides a confined environment (Figure 10-4). The data owned by a vFiler cannot be accessed by any other vFilers, even though they are hosted on the same physical storage system. All requests for data access owned by a vFiler are tagged with its context, making it impossible for unauthorized access to data to occur.

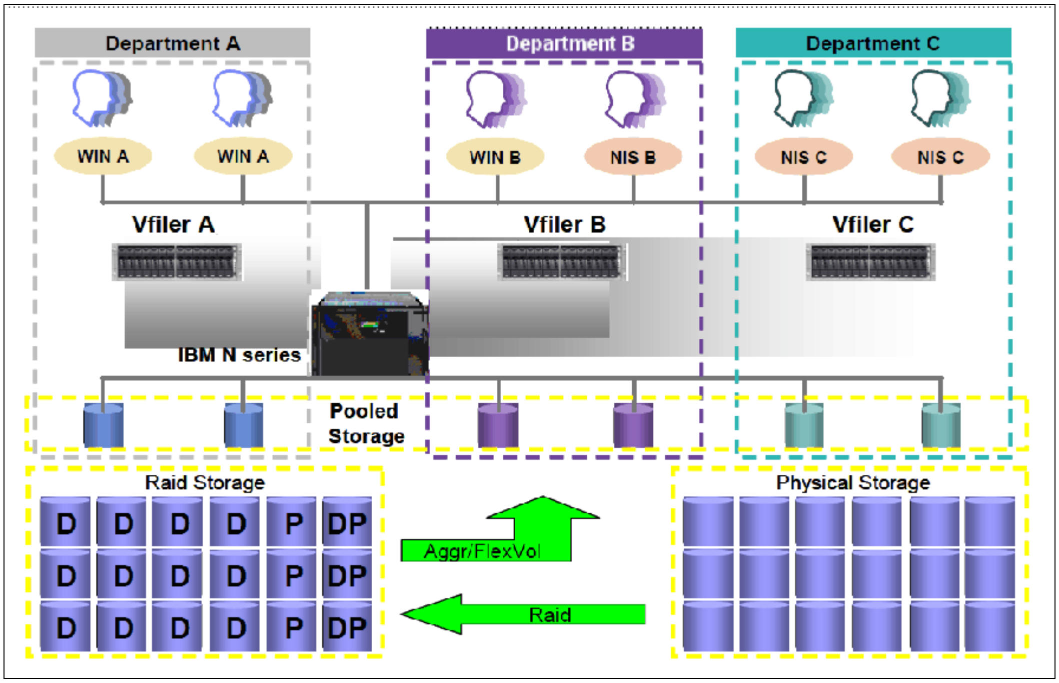


Figure 10-4 vFiler for service provider and enterprises

- Delegation of management:

  A vFiler provides the ability to delegate management access based on roles. The vFiler administrators can have different access rights compared with physical system storage administrators. It provides another layer of security.

- Disaster recovery:

  A vFiler provides an easy to deploy and manage disaster recovery solution that improves recovery time and lowers management costs. The use of virtualization technology removes the requirement that the primary and backup systems must be identical.

- Workload management:

  A vFiler provides an efficient mechanism to perform workload management by migrating vFiler across storage systems. Upgrading hardware or removing old hardware can be accomplished with minimal effort and with zero disruption of data access and zero configuration changes in the clients.

10.3 vFiler scenarios

In this section, we provide two scenarios:

- A vFiler migration with SnapMirror
- A vFiler disaster recovery (DR) solution

10.3.1 vFiler migration with SnapMirror

SnapMirror integration provides the ability to manage SnapMirror relationships within a vFiler. A vFiler can own a volume or qtree (the sources or the targets of SnapMirror relationships). The SnapMirror relationship is maintained only when the vFiler is migrated or failed over if the volume or qtree is the *source* of the SnapMirror relationship. If the volume or the qtree is the *target*, then the SnapMirror relationship will be broken off.

As shown in Figure 10-5, when the vFiler A hosted on IBM System Storage N series storage system 1 was migrated to IBM System Storage N series storage system 3, the existing SnapMirror relationship on a volume owned by vFiler A was also migrated to IBM System Storage N series storage system 3. Figure 10-5 shows the SnapMirror relationship being maintained after migration.

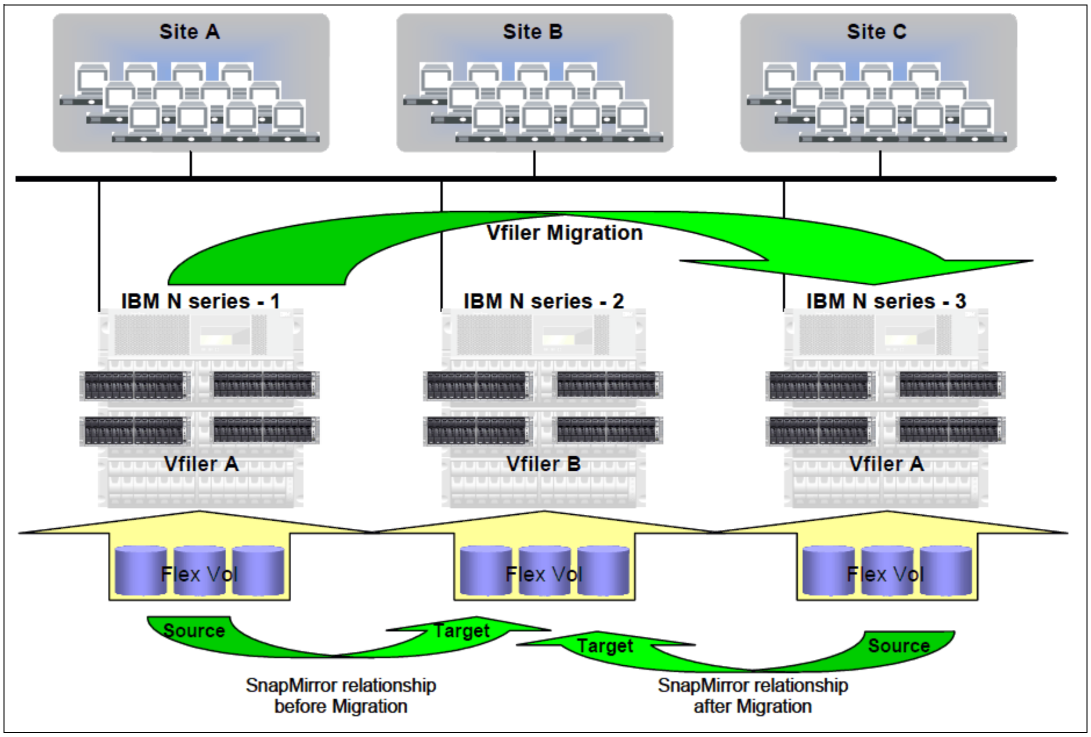


Figure 10-5 vFiler migration with SnapMirror

10.3.2 vFiler disaster recovery

vFiler is typically used in the disaster recovery scenario described in this section.

Basic disaster recovery

Basic disaster recovery (DR) can be configured over either LAN or WAN. DR is preferably configured over WAN to accommodate site failures. In this configuration, the target or the secondary site acts as both a backup site and a disaster site. Figure 10-6 shows basic DR.

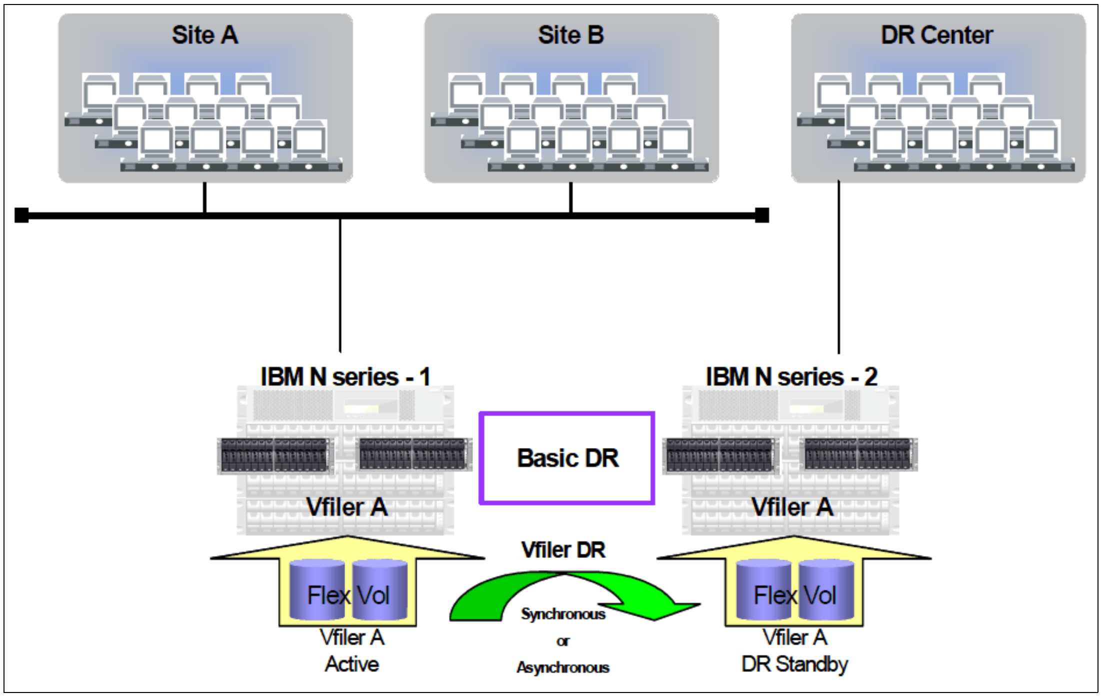


Figure 10-6 vFiler DR: Basic DR

The network traffic over the WAN is directly proportional to the rate of change of data on the primary or the source storage system. If the SnapMirror transfers are scheduled at a low frequency, the loss of data is great. If there is a synchronous SnapMirror, this setup impacts the network and CPU resources, especially if the DR is configured over the WAN.

Disaster recovery with backup

This scenario adds an additional layer of data protection compared with the basic DR configuration. In this configuration, shown in Figure 10-7, data is backed up using SnapMirror in synchronous mode at the primary site over LAN. DR can be configured in asynchronous mode over the WAN to the DR site, better utilizing network and CPU resources.

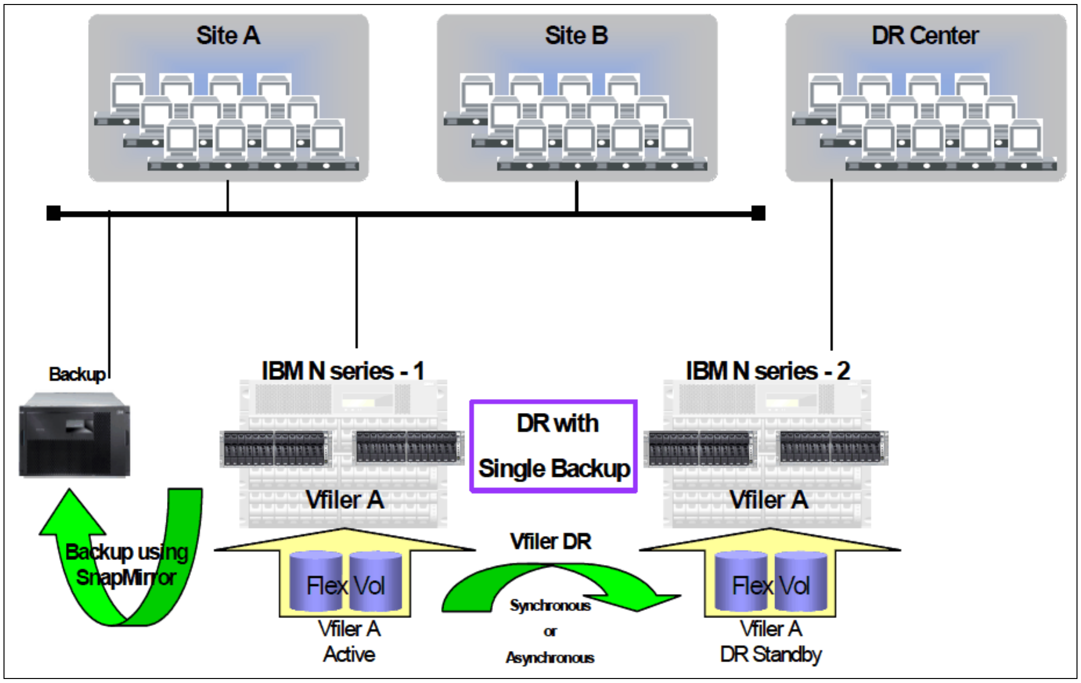


Figure 10-7 vFiler DR with single backup

Disaster recovery with two backups

In addition to the advantages of having a single backup, this scenario, as shown in Figure 10-8, provides an additional layer of data protection by backing up the data to secondary storage at the DR location. This architecture eliminates the single point of failure of the storage system on the DR site.

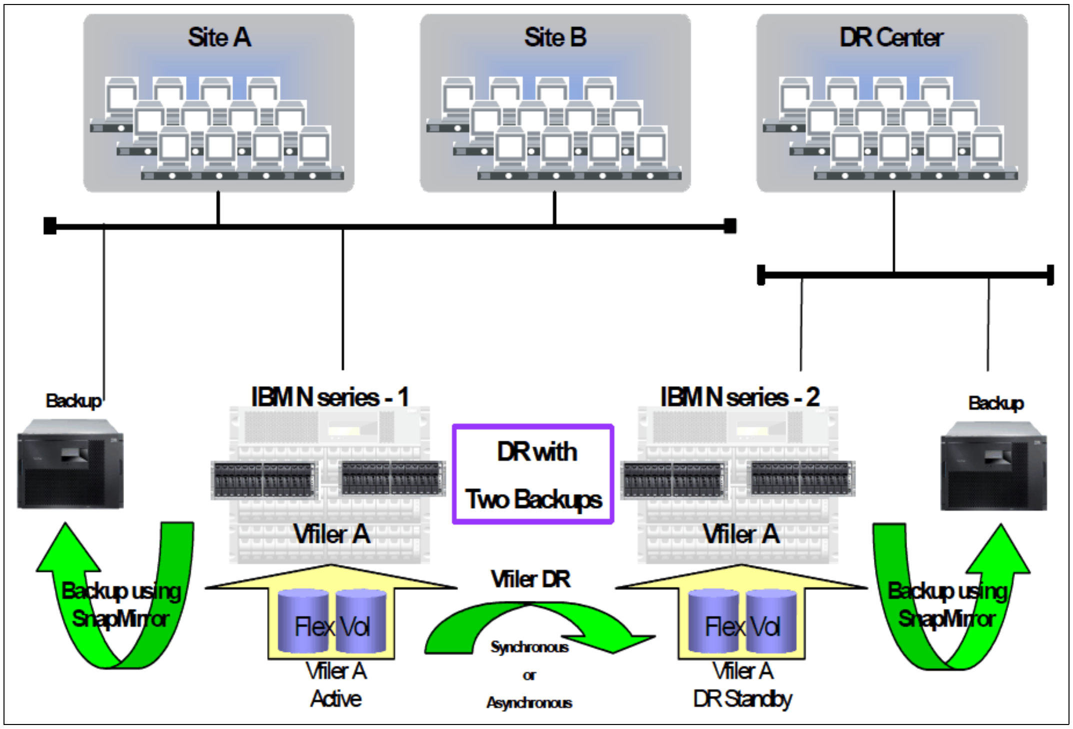


Figure 10-8 vFiler DR with two backups

> **Attention:** If DR is configured in asynchronous mode, data can be copied to the IBM System Storage N series storage system on the DR site using SnapMirror only in asynchronous mode. However, if DR is configured in synchronous mode, data can be copied to the IBM System Storage N series storage system on the DR site using SnapMirror in either synchronous or asynchronous mode.

The cost and data availability are directly proportional to each other. As the requirement for data availability increases, the cost for providing the solution also increases. Figure 10-9, which shows this relationship, is not drawn to scale and is not an indication of the proportion with which the cost increases as the need for data availability increases.

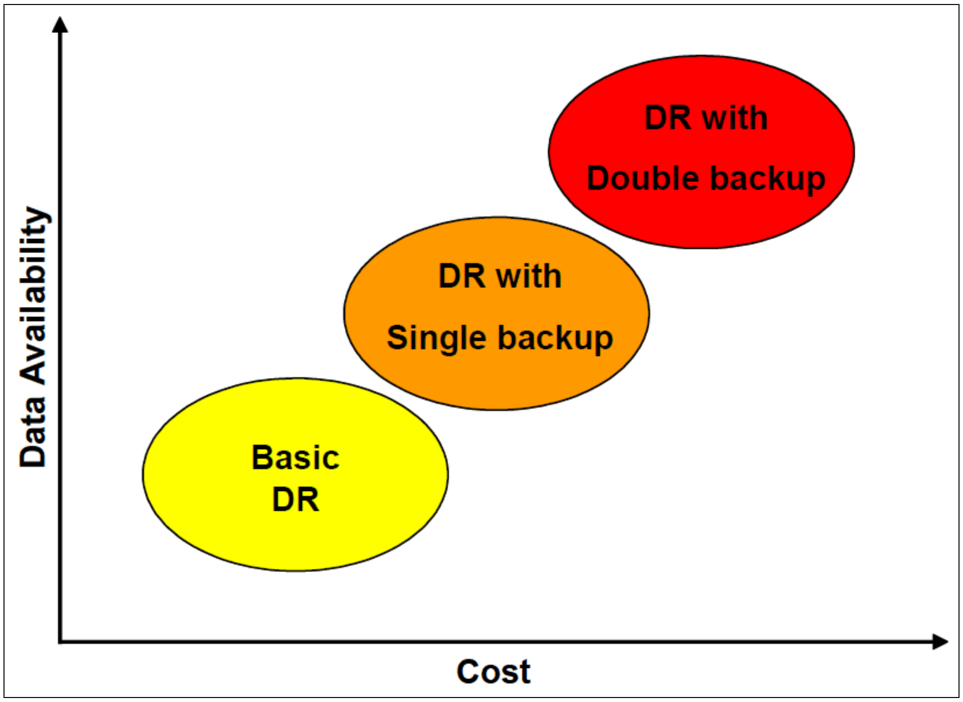


Figure 10-9 Relationship between data availability and cost

NDMP integration

Network Data Management Protocol (NDMP) supports vFiler units and physical storage units. It enables you to run an NDMP server on each vFiler unit, execute an **ndmpcopy** command within a vFiler unit, and perform backups on one or more vFiler units. Because each vFiler unit has its own NDMP server, NDMP enables you to back up or restore each vFiler unit independently, and you can set NDMP options on each vFiler unit as well.

Part 2

Data protection

Data ONTAP 8 data protection services use N series storage efficiency technologies and provide integration with leading-edge data protection applications. Your company can reduce both storage and management costs for a significantly lower total cost of ownership.

In this part of the book, you can learn more about our data protection features. The following topics are covered:

- Snapshot
- SnapRestore
- SnapMirror
- SnapLock
- SyncMirror
- MetroCluster

11

Snapshot

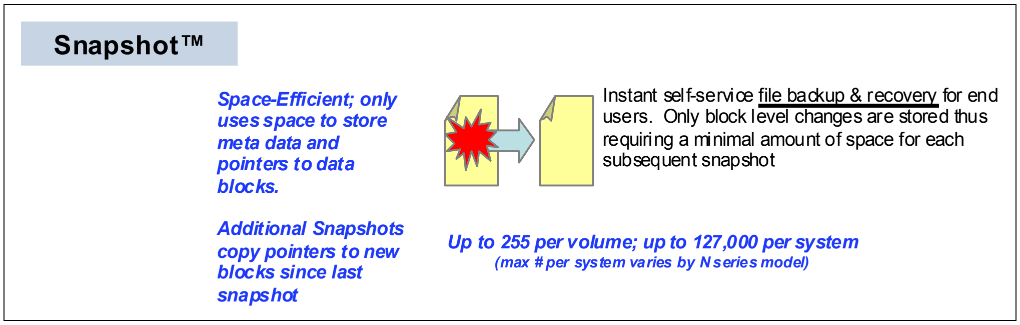
This chapter introduces the snap family of commands, which provides a means to create and manage Snapshot copies in each volume or aggregate.

The following topics are covered:

- Introduction to Snapshot
- Snapshot process: Basic operation
- Understanding Snapshots in detail
- Snapshot data structures and algorithms

11.1 Introduction to Snapshot

A Snapshot, as shown in Figure 11-1, is a read-only copy of the entire file system, as of the time the Snapshot was created. The filer creates Snapshots very quickly without consuming any disk space. The existing data remains in place; future writes to those blocks are redirected to new locations. Only as blocks in the active file system are modified and written to new locations on disk does the Snapshot begin to consume extra space.



Snapshot™

Space-Efficient; only uses space to store meta data and pointers to data blocks.

Instant self-service file backup & recovery for end users. Only block level changes are stored thus requiring a minimal amount of space for each subsequent snapshot

Additional Snapshots copy pointers to new blocks since last snapshot

Up to 255 per volume; up to 127,000 per system

(max # per system varies by N series model)

Figure 11-1 Snapshot

Volume Snapshots are exported to all CIFS or NFS clients. They can be accessed from each directory in the file system. From any directory, a user can access the set of Snapshots from a hidden sub-directory that appears to a CIFS client as *~snapshot* and to an NFS client as *.snapshot*. These hidden sub-directories are special in that they can be accessed from every directory, but they only show up in directory listings at an NFS mount point or at the root of CIFS share

Each volume on the filer can have up to 255 Snapshots at one time. Each aggregate on the filer can have up to 10 Snapshots at one time if Snapshot autodelete is enabled on that aggregate. If autodelete is not enabled, the aggregate can have up to 255 Snapshots. Because of the technique used to update disk blocks, deleting a Snapshot will generally not free as much space as its size would seem to indicate.

Blocks in the Snapshot can be shared with other Snapshots, or with the active file system, and thus might be unavailable for reuse even after the Snapshot is deleted.

11.1.1 Snap commands

If executed on a vFiler, the snap commands can only operate on volumes of which the vFiler has exclusive ownership. Manipulating Snapshots in shared volumes can only be performed on the physical filer. Operations on aggregate Snapshots are unavailable on vFilers and must be performed on the physical filer. For the rest of this section, if the snap commands are executed on a vFiler, all volume names passed on the command line must belong to the vFiler exclusively.

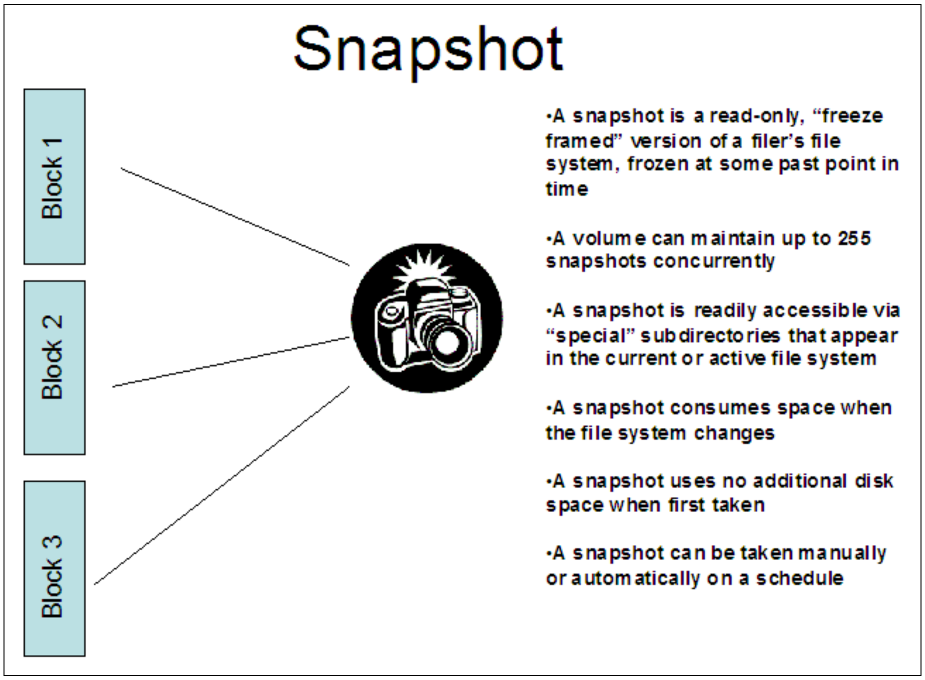
The snap commands are persistent across reboots. Do not include snap commands in the /etc/rc file. If you include a snap command in the /etc/rc file, the same snap command that you enter through the command line interface does not persist across a reboot and is overridden by the one in the /etc/rc file.

11.1.2 Automatic Snapshots

Automatic Snapshots can be scheduled to occur weekly, daily, or hourly. Weekly Snapshots are named weekly.N, where N is “0” for the most recent Snapshot, “1” for the next most recent, and so on. Daily Snapshots are named as daily.N; and hourly Snapshots as hourly.N.

Whenever a new Snapshot of a particular type is created and the number of existing Snapshots of that type exceeds the limit specified by the **sched** option described next, then the oldest Snapshot is deleted and the existing ones are renamed. If, for example, you specified that a maximum of 8 hourly Snapshots were to be saved using the **sched** command, then on the hour, hourly.7 would be deleted, hourly.0 would be renamed to hourly.1, and so on. If deletion of the oldest Snapshot fails because it is busy, the oldest Snapshot is renamed to scheduled_snap_busy.N, where N is a unique number identifying the Snapshot. When the Snapshot is no longer busy, it will be deleted. Do not use Snapshot names of this form for other purposes, because they can be deleted automatically.

Snapshot features are shown in Figure 11-2.



Snapshot

- A snapshot is a read-only, “freeze framed” version of a filer’s file system, frozen at some past point in time
- A volume can maintain up to 255 snapshots concurrently
- A snapshot is readily accessible via “special” subdirectories that appear in the current or active file system
- A snapshot consumes space when the file system changes
- A snapshot uses no additional disk space when first taken
- A snapshot can be taken manually or automatically on a schedule

Figure 11-2 Snapshot features

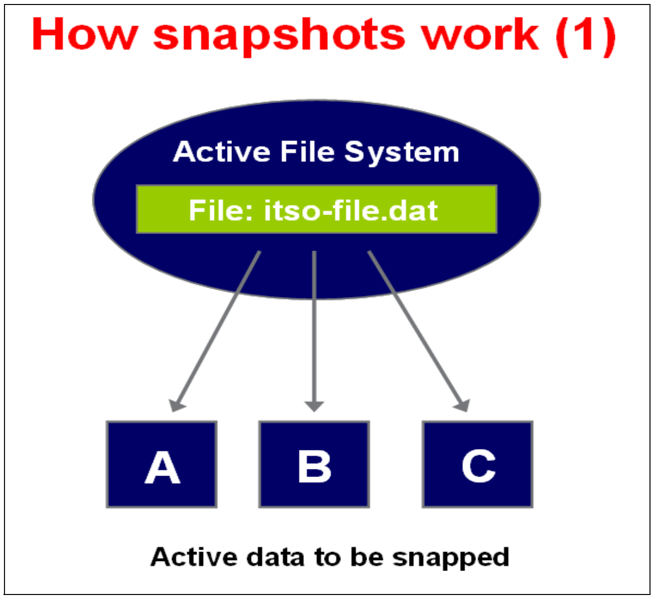
Snapshots use a redirect-on-write technique to avoid duplicating disk blocks that are the same in a Snapshot as in the active file system. Only when blocks in the active file system are modified or removed do Snapshots containing those blocks begin to consume disk space.

Users can access Snapshots to recover files that they have accidentally changed or removed, and system administrators can use Snapshots to create backups safely from a running system. In addition, WAFL uses Snapshots internally so that it can restart quickly even after an unclean system shutdown.

11.2 Snapshot process: Basic operation

The basic operation of the Snapshot process proceeds as follows:

1. Snapshots are performed from active data on the file system (Figure 11-3).



How snapshots work (1)

Active File System

File: itso-file.dat

A

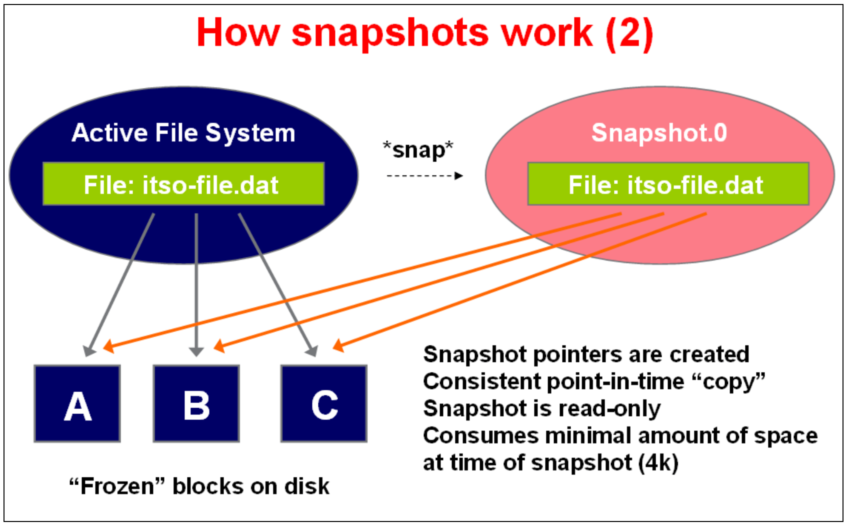
B

C

Active data to be snapped

Figure 11-3 Identify active data to be snapped

2. When an initial Snapshot is done, no initial data is copied. Instead, pointers are created to the original blocks for recording a point-in-time state of these blocks (Figure 11-4). These pointers are contained within metadata.



How snapshots work (2)

Active File System

File: itso-file.dat

snap

Snapshot.0

File: itso-file.dat

A

B

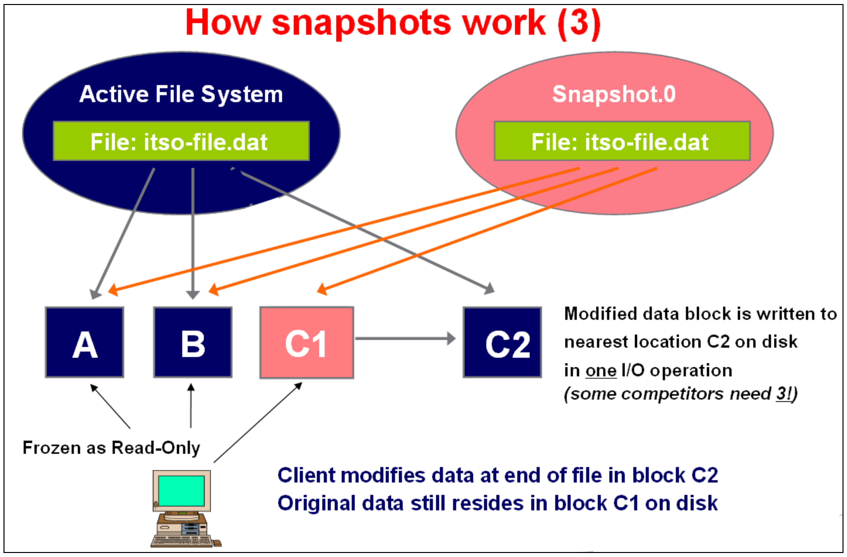
C

“Frozen” blocks on disk

Snapshot pointers are created
Consistent point-in-time “copy”
Snapshot is read-only
Consumes minimal amount of space
at time of snapshot (4k)

Figure 11-4 Pointers are created

3. When a request to block C occurs, the original block C1 is frozen to maintain a point-in-time copy, and the modified block C2 is written to another location on disk and becomes the active block (Figure 11-5).

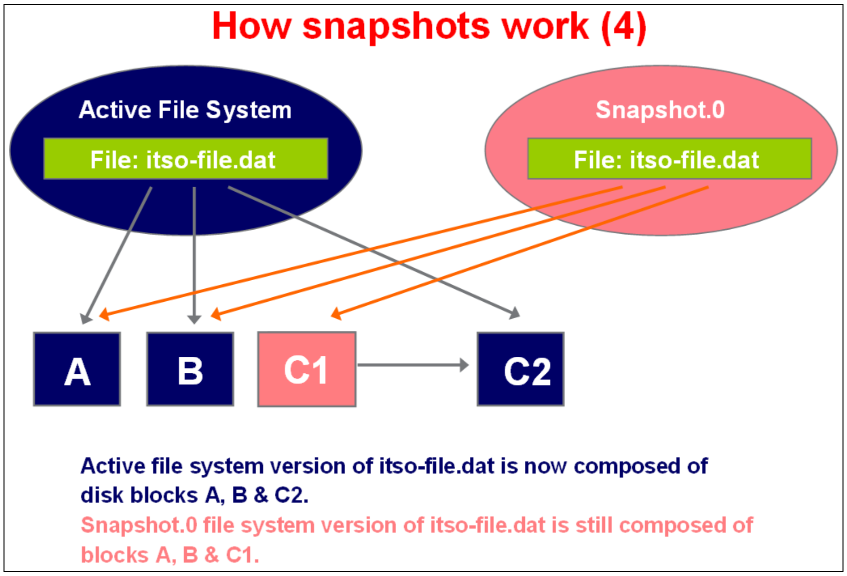


Modified data block is written to nearest location C2 on disk in one I/O operation *(some competitors need 3!)*

Client modifies data at end of file in block C2
Original data still resides in block C1 on disk

Figure 11-5 Modified block written to a location on disk becomes the active block

4. The final result is that the Snapshot now consumes 4 K + C1 of space. Active points for the point-in-time Snapshot are unmodified blocks A, B, and point-in-time copy C1 (Figure 11-6).



Active file system version of itso-file.dat is now composed of disk blocks A, B & C2.
Snapshot.0 file system version of itso-file.dat is still composed of blocks A, B & C1.

Figure 11-6 Final result showing active points

11.3 Understanding Snapshots in detail

A small percentage of the drive's available space is used to store file-system-related data and can be considered as impacting the system. A file system splits the remaining space into small, consistently sized segments. In the UNIX world, these segments are known as *inodes*.

Understanding that the WAFL file system is a tree of blocks rooted by the root inode is the key to understanding Snapshots. To create a virtual copy of this tree of blocks, WAFL simply duplicates the root inode. Figure 11-7 illustrates how this works.

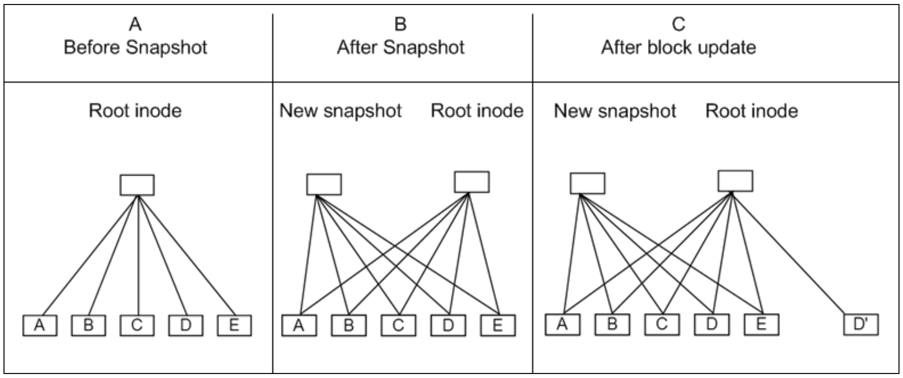


Figure 11-7 WAFL creates a Snapshot by duplicating the root inode

Column A in Figure 11-7 represents the basic situation before the Snapshot.

Column B in Figure 11-7 shows how WAFL creates a new Snapshot by making a duplicate copy of the root inode. This duplicate inode becomes the root of a tree of blocks representing the Snapshot, just as the root inode represents the active file system. When the Snapshot inode is created, it points to exactly the same disk blocks as the root inode. Thus, a brand-new Snapshot consumes no disk space except for the Snapshot inode itself.

Column C in Figure 11-7 shows what happens when a user modifies data block D. WAFL writes the new data to block D on disk and changes the active file system to point to the new block. The Snapshot still references the original block D, which is unmodified on disk.

Over time, as files in the active file system are modified or deleted, the Snapshot references more and more blocks that are no longer used in the active file system. The rate at which files change determines how long Snapshots can be kept online before they consume an unacceptable amount of disk space.

WAFL Snapshots duplicate the root inode instead of copying the entire inode file. It reduces considerable disk I/O and saves a lot of disk space. By duplicating only the root inode, WAFL creates Snapshots quickly and with little disk I/O. Snapshot performance is important because WAFL creates a Snapshot every few seconds to allow quick recovery after unclean system shutdowns.

The transition from column B in Figure 11-7 on page 144 to column C is illustrated in more detail in Figure 11-8 here. When a disk block is modified and its contents written to a new location, the block's parent must be modified to reflect the new location. The parent's parent, in turn, must also be written to a new location, and so on up to the root of the tree.

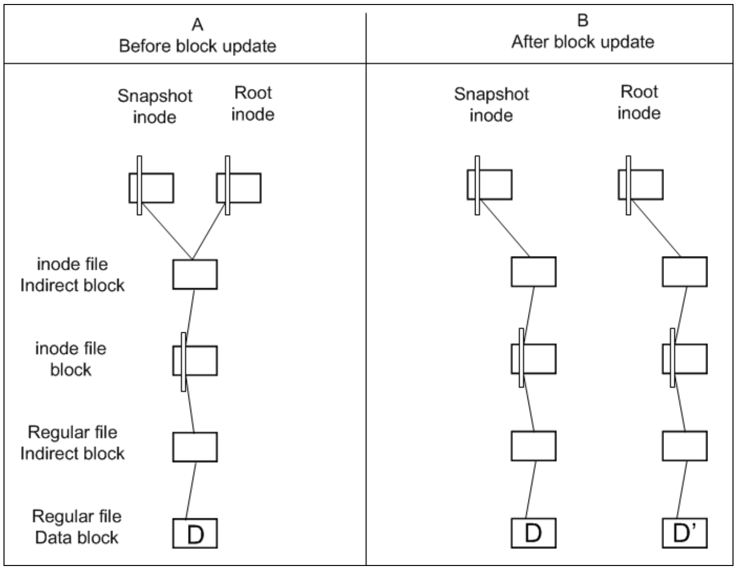


Figure 11-8 Block updates

WAFL might be inefficient if it wrote this many blocks for each Network File System (NFS) write request. Instead, WAFL gathers up many hundreds of NFS requests before scheduling a write episode. During a write episode, WAFL allocates disk space for all the unclean data in the cache and schedules the required disk I/O. As a result, commonly modified blocks (such as indirect blocks and blocks in the inode file) are written only once per write episode instead of once per NFS request.

11.3.1 How Snapshot copies consume disk space

Snapshot copies minimize disk consumption by preserving individual blocks rather than whole files. Snapshot copies begin to consume extra space only when files in the active file system are changed or deleted. When it happens, the original file blocks are still preserved as part of one or more Snapshot copies.

In the active file system, the changed blocks are rewritten to different locations on the disk or removed as active file blocks entirely. As a result, in addition to the disk space used by blocks in the modified active file system, disk space used by the original blocks is still reserved to reflect the status of the active file system before the change.

Figure 11-9 shows disk space usage for a Snapshot copy.

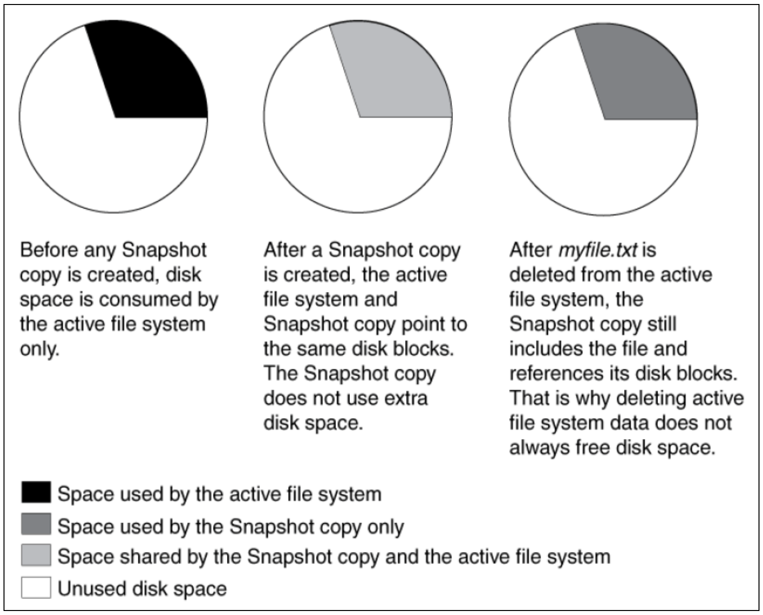

Before any Snapshot copy is created, disk space is consumed by the active file system only.
After a Snapshot copy is created, the active file system and Snapshot copy point to the same disk blocks. The Snapshot copy does not use extra disk space.
After *myfile.txt* is deleted from the active file system, the Snapshot copy still includes the file and references its disk blocks. That is why deleting active file system data does not always free disk space.
■ Space used by the active file system
▩ Space used by the Snapshot copy only
▨ Space shared by the Snapshot copy and the active file system
□ Unused disk space

Figure 11-9 How Snapshot copies consume disk space

11.3.2 How changing file content consumes disk space

A given file can be part of a Snapshot copy. The changes to such a file are written to new blocks. Therefore, the blocks within the Snapshot copy and the new (changed or added) blocks both use space within the volume.

Changing the contents of the myfile.txt file creates a situation where the new data written to myfile.txt cannot be stored in the same disk blocks as the current contents because the Snapshot copy is using those disk blocks to store the old version of myfile.txt. Instead, the new data is written to new disk blocks. As the following illustration shows, there are now two separate copies of myfile.txt on disk a new copy in the active file system and an old one in the Snapshot copy.

Figure 11-10 shows how changing file content consumes disk space.

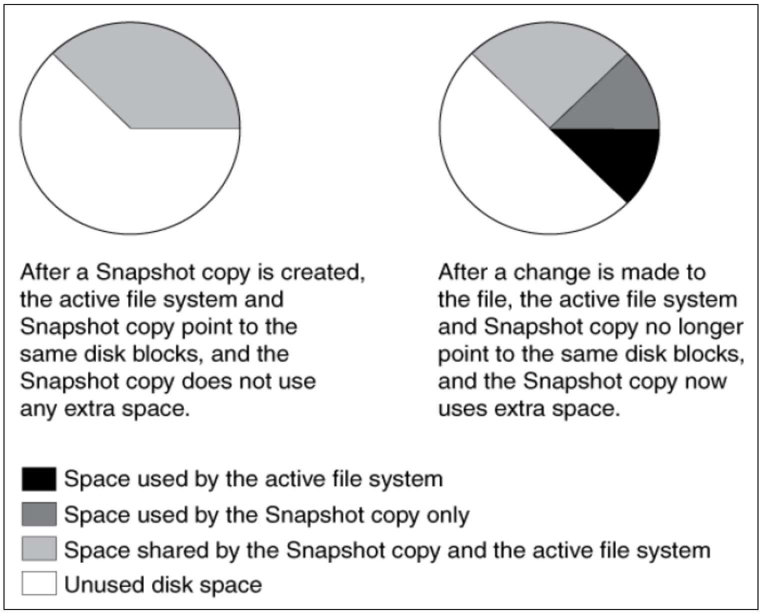

After a Snapshot copy is created, the active file system and Snapshot copy point to the same disk blocks, and the Snapshot copy does not use any extra space.
After a change is made to the file, the active file system and Snapshot copy no longer point to the same disk blocks, and the Snapshot copy now uses extra space.
■ Space used by the active file system
▩ Space used by the Snapshot copy only
▨ Space shared by the Snapshot copy and the active file system
□ Unused disk space

Figure 11-10 How changing file content consumes disk space

11.3.3 What the Snapshot copy reserve is

The Snapshot copy reserve sets a specific percent of disk space for Snapshot copies. The **snap reserve** command is used to set the size of the indicated volume's Snapshot reserve to percent. The **snap reserve** command prints the percentage of disk space reserved for Snapshots for each of the volumes in the system. With no percent argument, the percentage of disk space that is reserved for Snapshots in the indicated volume is displayed (Example 11-1).The Snapshot copy reserve can be used only by Snapshot copies, not by the active file system.

If the active file system runs out of disk space, any disk space still remaining in the Snapshot copy reserve is not available for use by the active file system.

Example 11-1 Snap reserve command

```
itsonas1*> snap reserve LUN1 40
itsonas1*> snap reserve LUN1
Volume LUN1: current snapshot reserve is 40% or 24189816 k-bytes.
itsonas1
*>
```

Tip: Although the active file system cannot consume disk space reserved for Snapshot copies, Snapshot copies can exceed the Snapshot copy reserve and consume disk space normally available to the active file system.

Managing the Snapshot copy reserve involves the following tasks:

- Ensuring that enough disk space is set aside for Snapshot copies so that they do not consume active file system space
- Keeping disk space consumed by Snapshot copies below the Snapshot copy reserve
- Ensuring that the Snapshot copy reserve is not so large that it wastes space that can be used by the active file system

Use of deleted active file disk space

When enough disk space is available for Snapshot copies in the Snapshot copy reserve, deleting files in the active file system frees disk space for new files, while the Snapshot copies that reference those files consume only the space in the Snapshot copy reserve.

About this task: If Data ONTAP created a Snapshot copy when the disks were full, then deleting files from the active file system does not create any free space because everything in the active file system is also referenced by the newly created Snapshot copy. Data ONTAP has to delete the Snapshot copy before it can create any new files.

The following topics describe how disk space being freed by deleting files in the active file system ends up in the Snapshot copy. If Data ONTAP creates a Snapshot copy when the active file system is full and there is still space remaining in the Snapshot reserve, the output from the **df** command (Example 11-2) displays statistics about the amount of disk space on a volume.

Example 11-2 Command output - space freed by deleting files in active file system ends up in the Snapshot copy.

```
itsonas1*>  df /vol/LUN1
Filesystem              kbytes       used      avail   capacity
/vol/LUN1/            3000000       300000    0       100%
/vol/LUN1/.snapshot   1000000       1000000   500000  50%
itsonas1*>
```

If you delete 100,000 KB (0.1 GB) of files, the disk space used by these files is no longer part of the active file system, so the space is reassigned to the Snapshot copies instead.

Data ONTAP reassigns 100,000 KB (0.1 GB) of space from the active file system to the Snapshot reserve. Because there was reserve space for Snapshot copies, deleting files from the active file system freed space for new files. If you enter the command again, the output **df** command is displayed (Example 11-3).

Example 11-3 Command output - reassigned 01.GB space from active file system to Snapshot reserve

```
itsonas1*>  df /vol/LUN1
Filesystem            kbytes        used      avail   capacity
/vol/LUN1/            3000000       2900000   100000  97%
/vol/LUN1/.snapshot   1000000       600000    400000  60%
itsonas1*>
```

Snapshot copies can exceed reserve

There is no way to prevent Snapshot copies from consuming disk space greater than the amount reserved for them. Therefore, it is important to reserve enough disk space for Snapshot copies so that the active file system always has space available to create new files or modify existing ones.

Consider what happens if all files in the active file system are deleted. Before the deletion, the **df** command output is listed in Example 11-4.

Example 11-4 Command output before deletion of all files in the active file system

```
itsonas1*>  df /vol/LUN1
Filesystem              kbytes       used      avail   capacity
/vol/LUN1/            3000000      300000    0       100%
/vol/LUN1/.snapshot   1000000      1000000   500000  50%
itsonas1*>
```

After the deletion of all files in an active file systems, the entire 3,000,000 KB (3 GB) in the active file system is still being used by Snapshot copies, along with the 500,000 KB (0.5 GB) that was being used by Snapshot copies before, making a total of 3,500,000 KB (3.5 GB) of Snapshot copy data. It is 2,500,000 KB (2.5 GB) more than the space reserved for Snapshot copies; therefore, 2.5 GB of space that might be available to the active file system is now unavailable to it. The post-deletion output of the **df** command (Example 11-5) lists this unavailable space as used even though no files are stored in the active file system.

Example 11-5 Command output after deletion of all files in the active file system

```
itsonas1*>  df /vol/LUN1
Filesystem            kbytes       used      avail   capacity
/vol/LUN1/            3000000      2500000   500000  83%
/vol/LUN1/.snapshot   1000000      3500000   0       350%
itsonas1*>
```

Recovery of disk space for file system use

Whenever Snapshot copies consume more than 100% of the Snapshot reserve, the system is in danger of becoming full. In this case, you can create files only after you delete enough Snapshot copies.

If 500,000 KB (0.5 GB) of data is added to the active file system, a **df** command generates the output shown in Example 11-6.

Example 11-6 Command output after 500,000 KB of data is added to the active filesystem

```
itsonas1*>  df /vol/LUN1
Filesystem            kbytes       used      avail   capacity
/vol/LUN1/            3000000      2500000   0       100%
/vol/LUN1/.snapshot   1000000      3500000   0       350%
itsonas1*>
```

As soon as Data ONTAP creates a new Snapshot copy, every disk block in the file system is referenced by some Snapshot copy. Therefore, no matter how many files you delete from the active file system, there is still no room to add any more. The only way to recover from this situation is to delete enough Snapshot copies to free more disk space.

What file folding means and how it saves disk space

File folding describes the process of checking the data in the most recent Snapshot copy, and if this data is identical to the Snapshot copy currently being created, by referencing the previous Snapshot copy instead of taking up disk space writing the same data in the new Snapshot copy.

File folding saves disk space by sharing unchanged file blocks between the active version of the file and the version of the file in the latest Snapshot copy, if any.

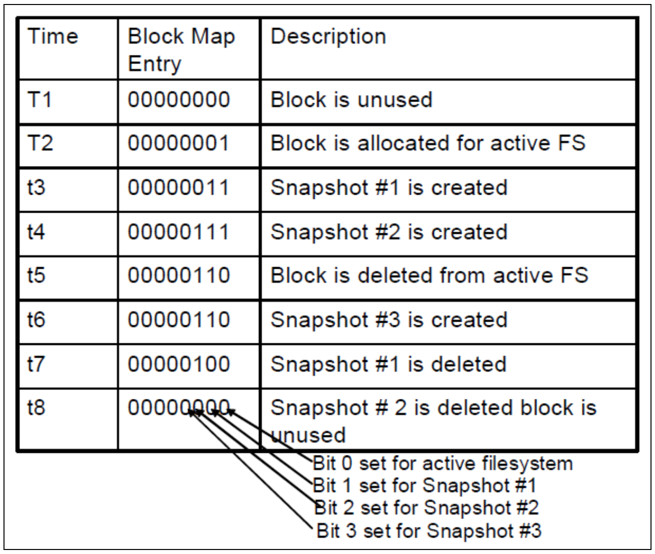
The system must compare block contents when folding a file, so file folding might affect system performance.

If the folding process reaches a maximum limit on memory usage, it is suspended. When memory usage falls below the limit, the processes that were halted are restarted.

11.4 Snapshot data structures and algorithms

Most file systems keep track of free blocks by using a bit map with one bit per disk block. If the bit is set, then the block is in use. However, this technique does not work for WAFL because many Snapshots can reference a block at the same time.

Figure 11-11 shows the life cycle of a typical block-map entry. At time T1, the block-map entry is completely clear, indicating that the block is available. At time T2, WAFL allocates the block and stores file data in it.



Time	Block Map Entry	Description
T1	00000000	Block is unused
T2	00000001	Block is allocated for active FS
t3	00000011	Snapshot #1 is created
t4	00000111	Snapshot #2 is created
t5	00000110	Block is deleted from active FS
t6	00000110	Snapshot #3 is created
t7	00000100	Snapshot #1 is deleted
t8	00000000	Snapshot # 2 is deleted block is unused

Bit 0 set for active filesystem
Bit 1 set for Snapshot #1
Bit 2 set for Snapshot #2
Bit 3 set for Snapshot #3

Figure 11-11 Life cycle of a block-map file entry

When Snapshots are created, at times t3 and t4, WAFL copies the active file system bit into the bit indicating membership in the Snapshot. The block is deleted from the active file system at time t5. It can occur either because the file containing the block is removed or because the contents of the block are updated and the new contents are written to a new location on disk.

The block cannot be reused, however, until no Snapshot references it. In Figure 11-11, it occurs at time t8, after both Snapshots that reference the block have been removed.

11.4.1 Creating a Snapshot

The challenge in writing a Snapshot to disk is to avoid locking out incoming NFS requests. The problem is that new NFS requests might need to change cached data that is part of the Snapshot and that must remain unchanged until it reaches disk.

An easy way to create a Snapshot is to suspend NFS processing, write the Snapshot, and then resume NFS processing. However, writing a Snapshot can take more than a second, which is too long for an NFS server to stop responding. (Remember that WAFL creates a consistency point Snapshot at least every 10 seconds, so performance is critical.)

The WAFL technique for keeping Snapshot data self-consistent is to mark all the unclean data in the cache as IN_Snapshot. The rule during Snapshot creation is that data marked IN_Snapshot must not be modified, and data not marked IN_Snapshot must not be flushed to disk. NFS requests can read all file system data and can modify data that is not IN_Snapshot, but processing for requests that must modify IN_Snapshot data must be deferred.

To avoid locking out NFS requests, WAFL must flush IN_Snapshot data as quickly as possible. To do this, WAFL performs the following tasks:

1. Allocates disk space for all files with IN_Snapshot blocks.

   WAFL caches inode data in two places:

   - In a special cache of in-core inodes
   - In disk buffers belonging to the inode file

   When it finishes write allocating a file, WAFL copies the newly updated inode information from the inode cache into the appropriate inode file disk buffer and clears the IN_Snapshot bit on the in-core inode.

   When this step is complete, no inodes for regular files are marked IN_Snapshot, and most NFS operations can continue without blocking. Fortunately, this step can be done quickly because it requires no disk I/O.

2. Updates the block-map file.

   For each block-map entry, WAFL copies the bit for the active file system to the bit for the new Snapshot.

3. Writes all IN_Snapshot disk buffers in cache to their newly allocated locations on disk.

   As soon as a particular buffer is flushed, WAFL restarts any NFS requests waiting to modify it.

4. Duplicates the root inode to create an inode that represents the new Snapshot and turns the root inode's IN_Snapshot bit off.

   The new Snapshot inode must not reach the disk until after all other blocks in the Snapshot have been written. If this rule were not followed, an unexpected system shutdown can leave the Snapshot in an inconsistent state.

After the new inode has been written, no more IN_Snapshot data exists in cache, and any NFS requests that are still suspended can be processed. Under normal loads, WAFL performs these four steps in less than a second. Step 1 can generally be done in just a few hundredths of a second, and after WAFL completes it, few NFS operations need to be delayed.

11.4.2 Deleting a Snapshot

Deleting a Snapshot is a simple task. WAFL simply zeros the root inode representing the Snapshot and clears the bit representing the Snapshot in each block-map entry.

When creating Snapshots from LUNs, the task can be accomplished by using SnapDrive software from the host and running the **snap delete** command from the Data ONTAP command-line interface (CLI), or using System Manager.

To delete a Snapshot using the **snap delete** command, run the following command:

```
snap delete volume_name snapshot_name
```

The various parts of this expression have the following meanings:

- The volume_name is the name of the volume that contains the Snapshot to delete.
- The snapshot_name is the specific Snapshot to delete.

12

SnapRestore

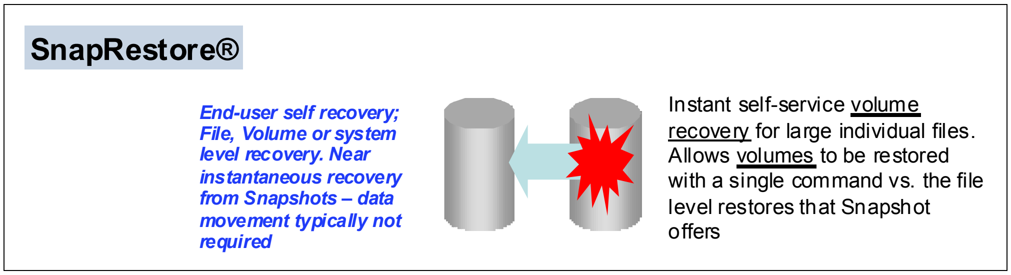
This chapter introduces the snap family of commands, which provides a means to create and manage Snapshot copies in each volume or aggregate.Introducing Snapshots.

The following topics are covered:

- SnapRestore at a glance
- Introduction to SnapRestore
- SnapRestore operation
- SnapRestore: Details of operation
- Examples
- SnapRestore for databases

12.1 SnapRestore at a glance

This chapter introduces the IBM System Storage N series SnapRestore as shown in Figure 12-1.



SnapRestore®

End-user self recovery; File, Volume or system level recovery. Near instantaneous recovery from Snapshots – data movement typically not required

Instant self-service volume recovery for large individual files. Allows volumes to be restored with a single command vs. the file level restores that Snapshot offers

Figure 12-1 SnapRestore

You can use the SnapRestore feature to recover data that is no longer available, or if you are testing a volume or file and want to restore that volume or file to pre-test conditions.

SnapRestore enables you to quickly revert a local volume or file to the state it was in when a particular Snapshot copy was taken. In most cases, reverting a file or volume is much faster than restoring files from tape or copying files from a Snapshot copy to the active file system.

After you select a Snapshot copy for reversion, the Data ONTAP reverts the specified file or the volume to the data and timestamps that it contained when the selected Snapshot copy was taken. Data that was written after the selected Snapshot copy was taken is lost.

Tip: If the volume you select to revert is a root volume, the system reboots.

SnapRestore reverts only the file contents. It does not revert attributes of a volume. For example, the Snapshot copy schedule, volume option settings, RAID group size, and maximum number of files per volume remain unchanged after the reversion.

You use SnapRestore to recover from data corruption. If a primary system application corrupts data files in a volume, you can revert the volume or specified files in the volume to a Snapshot copy taken before the data corruption.

SnapRestore performs Snapshot copy restoration more quickly, using less disk space, than an administrator can achieve by manually copying volumes, qtrees, directories, or large files to be restored from the Snapshot copy.

You must purchase and install the license code before you can use SnapRestore.

On the server, enter the following command:

```
license add xxxxxxx
```

xxxxxxx is the license code you purchased. This setting persists across reboots.

12.2 Introduction to SnapRestore

SnapRestore software makes recovering your data fast and easy. Without SnapRestore, you have to restore files from tape, use a third-party software application, or copy files from a Snapshot to the active file system.

Using these methods takes longer than reverting a volume, and can take longer than reverting a single file (Figure 12-2). It is because with SnapRestore, no data is copied; instead, the file system is restored to an earlier state.

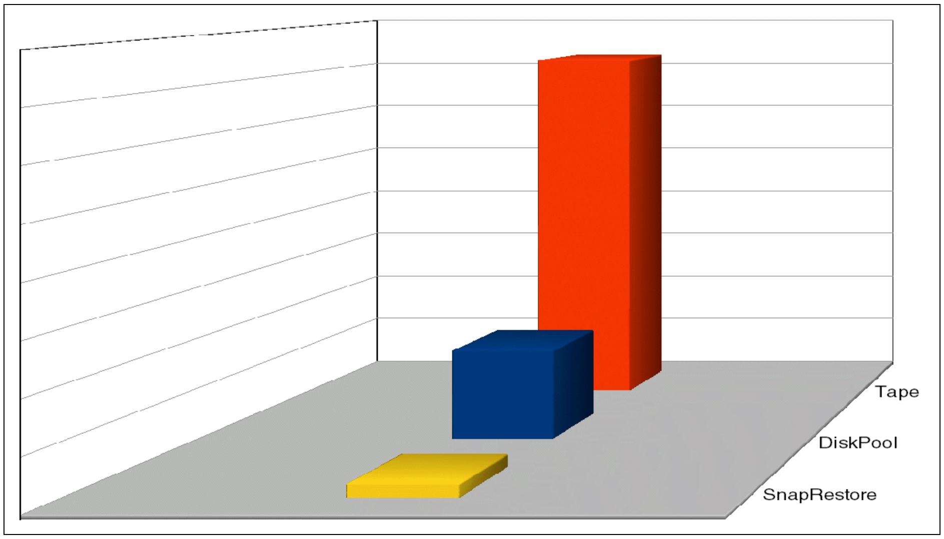


Figure 12-2 Restore time

12.2.1 Cost and storage efficiency

Snapshot technology makes extremely efficient use of storage by storing only block-level changes between each successive Snapshot. Because the Snapshot process is automatic and virtually instantaneous, backups are significantly faster and simpler. SnapRestore software uses Snapshot technology to perform near-instantaneous data restoration. In contrast, alternative storage solutions copy all of the data and require much more time and disk storage for the backup-and-restore operations.

SnapRestore reduces the burden on staffing resources. Whether your business employs a small group of users or an enterprise-scale user community and IT support team, SnapRestore's easy single-command restoration eliminates complexity and reduces errors. Using SnapRestore requires no special training or expertise.

12.2.2 Data restoration

With SnapRestore, data can be restored from any one of the Snapshots stored on the file system. It allows an application development team, for example, to revert to Snapshots from various stages of their design, or test engineers to quickly and easily return data to a baseline state. Restoring to the base environment takes only seconds, and the restored environment is identical to the point at which the Snapshot was created.

12.2.3 A possible SnapRestore use

As an example of a possible SnapRestore use, let us assume that Snapshots are taken every night at midnight. On Friday, you discover that an upgrade that you performed on your application on Thursday accidentally caused some serious corruption to some data stored on that volume. You want to revert your environment to the Snapshot that was taken on Wednesday night.

To accomplish this task, you can invoke the **`snap restore`** command and specify the Wednesday night Snapshot. This action restores the entire volume in seconds and reverts the active file system to the state that it was in on Wednesday night.

Upon further examination, you will notice that the Wednesday Snapshot is still in the Snapshot list, but the Thursday Snapshot is now missing.

You can also accomplish this task by specifying a single file or LUN. However, to make it happen quickly, the file or LUN must be removed first from the active file system. If it is not removed, the blocks for that file or LUN are already defined, so the storage system turns the request into a copy operation instead of an instantaneous revert. So single file SnapRestore (SFSR) becomes the same operation as copying the data from the Snapshot into the active file system.

12.2.4 Flexible restore

SnapRestore allows a customer to quickly restore an entire volume, a single file, or a LUN by simply having any of the available Snapshot images overwrite the existing data in the active file system. Using SnapRestore, you can roll back data to the instant a Snapshot was taken.

> **Attention:** Rolling back in time, by restoring from a Snapshot, has the effect of wiping out any changes made to the volume after the Snapshot.

12.2.5 Additional benefits

SnapRestore includes these additional benefits:

- Simple, single-command operation: There is no special expertise required, so the chance of operator error is greatly reduced.
- File or full volume restore: You can choose to restore only specific files or the entire volume.
- Multiple recovery points: You can restore the most recent clean copy from any Snapshot.
- Unsurpassed reliability: SnapRestore is far more dependable than traditional data restoration methods.
- Fast restoration of databases: This function is especially useful when the recovery time of databases is critical.
- Reduced dependency on operator or administrator to make a file restore: The file owner can restore the file.
- There is quick recovery from virus attacks.
- SnapRestore is based on Snapshot technology.
- There is a possible reduction in dependency upon tape.
- There can be data recovery after a user error or application error.

12.3 SnapRestore operation

After you select a Snapshot for reversion, the IBM System Storage N series storage system restores the volume or file that contain the same data and time stamps as they did when the Snapshot was taken. As mentioned, all data that existed before the reversion is overwritten.

Important: You cannot undo a SnapRestore reversion to change the volume back to the state that it was in prior to the reversion.

12.3.1 What SnapRestore reverts

SnapRestore only reverts file contents. It does *not* revert attributes of a volume, such as the Snapshot schedule, volume option settings, RAID group size, and maximum number of files per volume.

However, option settings applicable to the entire storage system might be reverted, because the option settings are stored in a registry in the /etc directory on the root volume. If you revert the root volume, the registry is reverted to the version that was in use at Snapshot creation time.

You can revert a volume to a Snapshot taken when the storage system was running a different Data ONTAP version as well. Note, however, that doing so can cause problems because of version incompatibilities.

Important:

- You cannot revert a volume to recover a deleted Snapshot.
- After you revert a volume to a specific Snapshot, you will lose Snapshots that are more recent than the Snapshot used for the volume reversion.

12.3.2 Applying SnapRestore

SnapRestore is a data recovery facility available for IBM N series.

SnapRestore scenarios

SnapRestore can be used in the following scenarios:

- Disaster recovery
- Database corruption recovery
- Application testing, such as a development environment using large data files

If a client application corrupts data files in a volume, you can revert the volume to a Snapshot taken before the data corruption.

Prerequisites for using SnapRestore

The prerequisites for using SnapRestore are as follows:

- SnapRestore must be licensed on the system.
- There must be at least one Snapshot on the system that you select to revert.
- The volume to be reverted must be online and must not be a mirror that is used for data replication.
- The LUN must be unmounted before using SnapRestore to revert the volume containing the LUN or to revert a single file. SnapRestore of the LUN. For a single file SnapRestore, the LUN must also be offline.

Licensing:

- SnapRestore is a licensed feature and requires a license code. A SnapRestore license must be purchased and installed before Snapshots can be restored.
- Snapshots can be accessed as read only without buying the optional SnapRestore license. Then single files from within the Snapshot can be restored.

12.4 SnapRestore: Details of operation

In this section, we go into more detail about SnapRestore operation.

As shown in Figure 12-3, at time state 0, the first Snapshot is taken and it points to the 4 K blocks that are equivalent to those in the active file system. No additional space is used at this time by Snapshot 1 because modifications to the active file system blocks have not occurred.

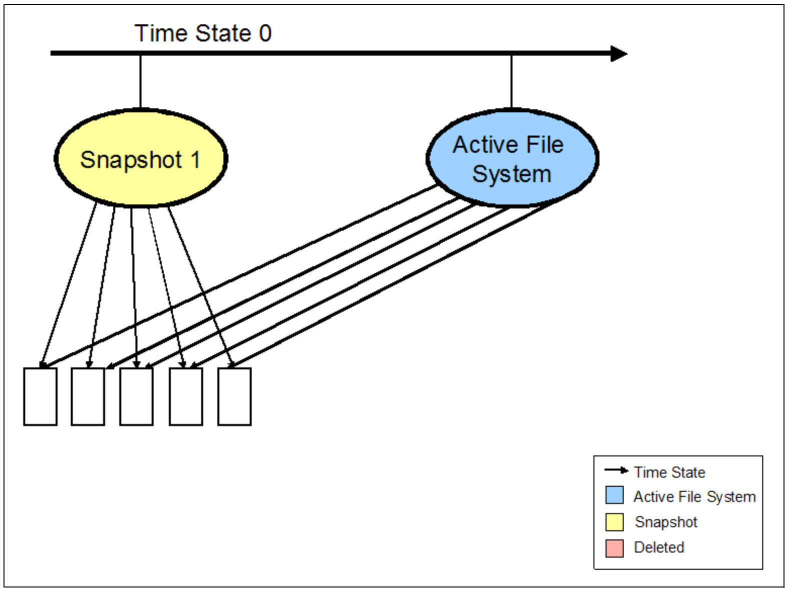


Figure 12-3 Snapshot 1

Some time goes by and new files are added with new blocks and modifications to files and their existing blocks are made, as shown in Figure 12-4. Snapshot 1 now points to blocks and the file system as it appeared in time state 0. Notice that one of the blocks, A1, has not been modified and is still part of the active file system.

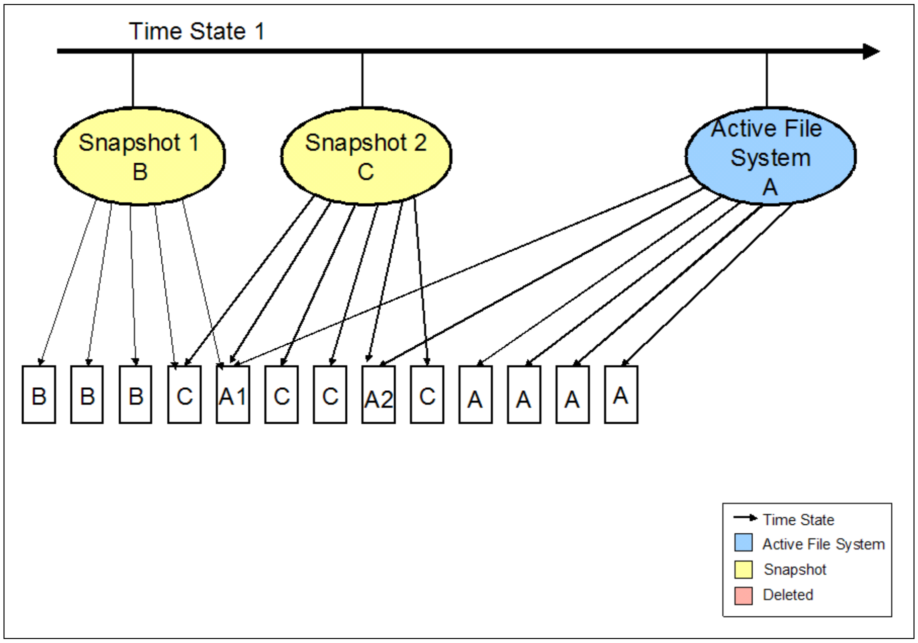


Figure 12-4   Snapshot 2

Snapshot 2 reflects a Snapshot of file modifications and adds since time state 0. Notice that it still points to active file system blocks A1 and A2.

More time goes by and more files are added with new blocks, and modifications to files and their existing blocks are done, as shown in Figure 12-5. Snapshot 1 now points to blocks and the file system as it appeared in time state 0. Snapshot 2 reflects a Snapshot of file modifications and adds since time state 0. Snapshot 3 reflects modifications and adds since time state 1 and Snapshot 2. Note that Snapshot 1 no longer points to any Active file system blocks.

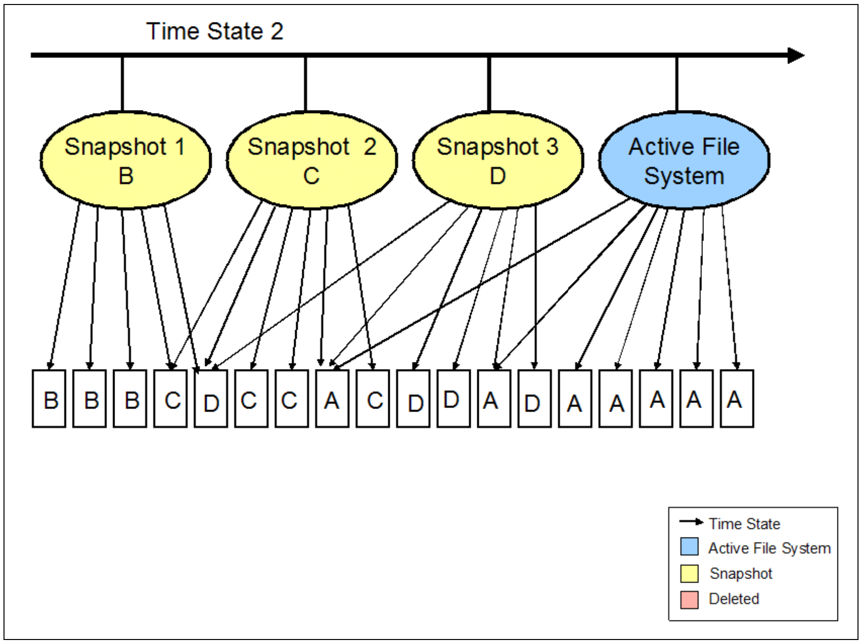


Figure 12-5 Snapshot 3

Snapshot 4, as shown in Figure 12-6, brings us to time state 3. It reflects adds or modifications of 4 K blocks. Notice that the first two Snapshots no longer reflect any of the active file system blocks.

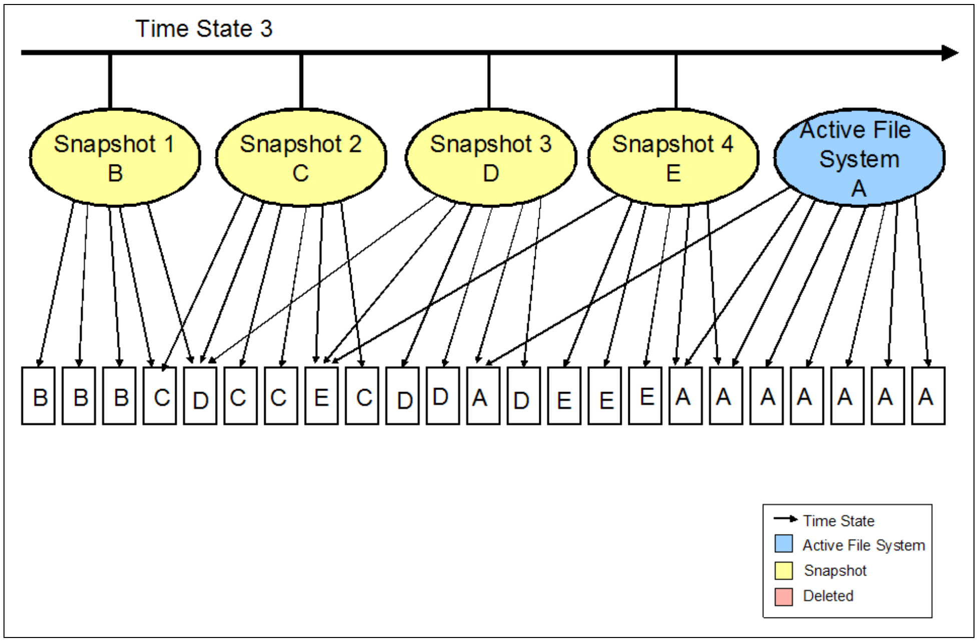


Figure 12-6 Subsequent Snapshots

As shown in Figure 12-7, the customer has discovered that it has file system corruption due to a virus and must revert to the point-in-time Snapshot of time state 2 and Snapshot 3. The active file system becomes Snapshot 3. Blocks that were previously pointed to only by Snapshot 4 or the active file system are freed up for writes again. In addition, blocks that were pointed to only by Snapshot 4 and the previous active file system are also freed up again.

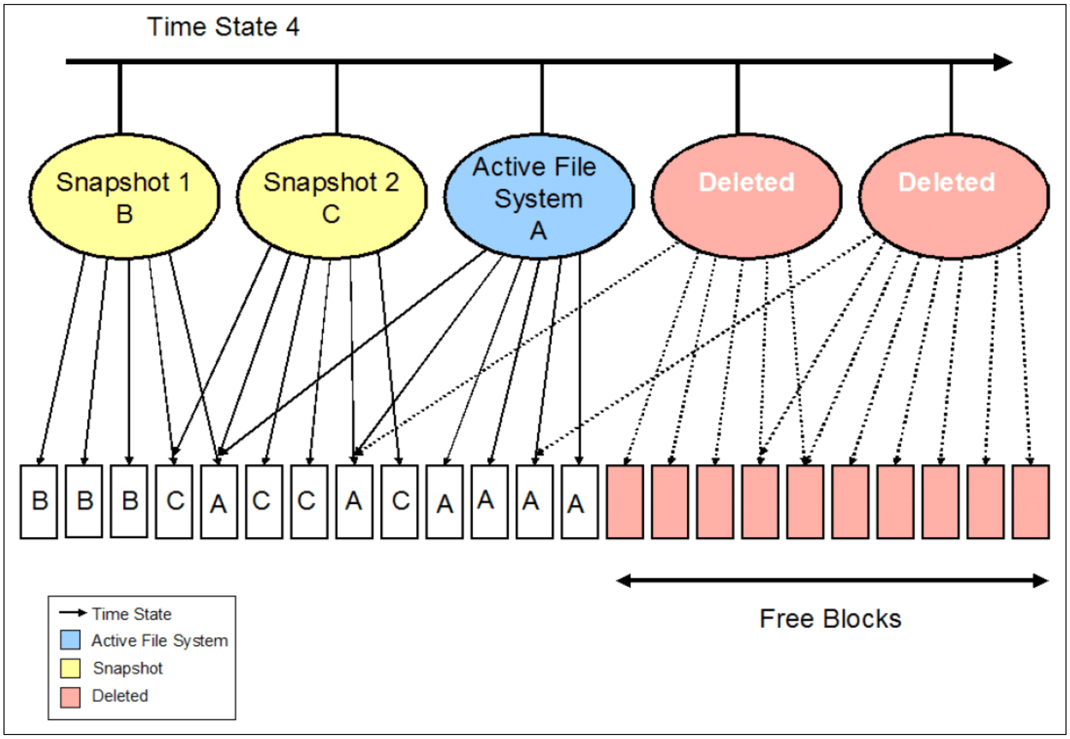


Figure 12-7 Reversion to Snapshot 3

In Figure 12-8, we compare time state 4 and the reversion to Snapshot 3 and time state 1, which reflects the active file system before Snapshot 3. As you can see, they are the same.

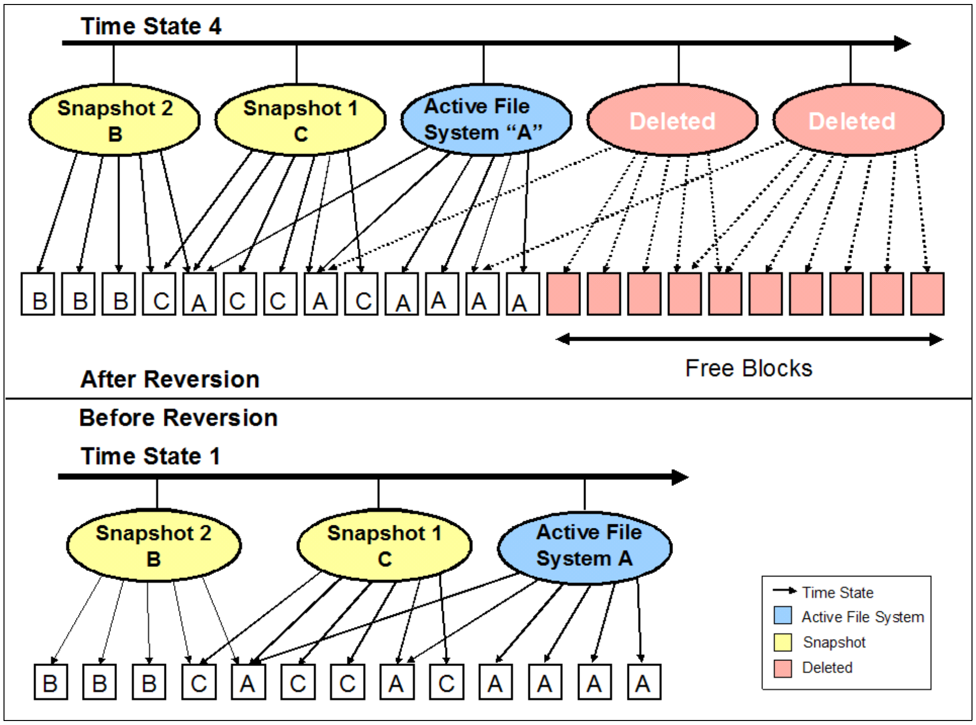


Figure 12-8 Comparison of reversion to Snapshot 3 time state 4 and active file system at time state 1

12.5 Examples

In this section, we cover common examples of using SnapRestore. IBM N series FilerView does not support restore of Snapshots, so we use the Data ONTAP Command Line Interface (CLI).

12.5.1 SnapRestore command syntax

The **snap restore** command's syntax is as follows:

```
snap restore [ -f ] [ -t vol | file ] [ -s snap_shot_name ]
[ -r restore_as_new_path ] vol_name restore_from_path
```

This command reverts a volume to a specified Snapshot, or reverts a single file to a revision from a specified Snapshot.

The **snap restore** command is only available if your storage system has the SnapRestore license. If you do not specify a Snapshot, the IBM System Storage N series storage system prompts you for the Snapshot.

Before reverting the volume or file, the user must confirm the operation. The **-f** option suppresses this confirmation step.

If the **-t** option is specified, it must be followed by **vol** or **file** to indicate which type of SnapRestore is to performed.

A volume cannot have both a volume SnapRestore and a single file SnapRestore executing simultaneously. Multiple single file SnapRestores can be in progress simultaneously.

Considerations:

- Network File System (NFS) users need to unmount the files and directories in the volume before a reversion. If they do not unmount the files and directories, they might get a `stale file handle` error message after the volume reversion.
- For additional information about the **`snap restore`** command and its parameters, see the *Data ONTAP 8.0 7-Mode Data Protection Online Backup and Recovery Guide*, available at this website:

  `http://www.ibm.com/storage/support/nas`

12.5.2 Considerations before performing SnapRestore

Consider the following items before performing SnapRestore:

- When the restore_as_path parameter is specified within the command syntax, the path must be a full path to a file name and must be in the same volume as the volume used for the restore.
- The volume used for restoring the file must be online and must not be a mirror.
- When the restore_as_path parameter is specified within the command syntax, the path must be a full path to a file name and must be in the same volume as the volume used for the restore.
- Files other than normal files and LUNs are not restored. This includes directories (and their contents), and files with NT streams.
- If there is not enough space in the volume, the single file SnapRestore will not start.
- If the file already exists (in the active file system), it will be overwritten with the version in the Snapshot.
- It can take up to several minutes for **`snap`** command output. During this time client exclusive oplocks are revoked and hard exclusive locks like the DOS compatibility lock are invalidated.
- After the **`snap`** command returns, the file restore will proceed in the background. During this time, any operation that tries to change the file will be suspended until the restore is done. Also, other single file SnapRestores can be executed.
- It is possible for the single file SnapRestore to be aborted if we run out of disk space during the operation. When it happens, the time stamp of the file being restored is updated. Thus, it will not be the same as the time stamp of the file in the Snapshot.
- An in-progress restore can be aborted by removing the file. For NFS users, the last link to the file must be removed.
- The Snapshot used for the restore cannot be deleted. New Snapshots cannot be created while a single file SnapRestore is in progress. Scheduled Snapshots on the volume will be suspended for the duration of the restore.
- Tree, user, and group quota limits are not enforced for the owner, group, and tree in which the file is being restored. Thus, if the user, group, or tree quotas are exceeded, /etc/quotas must be altered after the single file SnapRestore operation has completed. Then quota resize must be run.
- When the restore completes, the file’s attributes (size, permissions, ownership, and so on) must be identical to those in the Snapshot.
- If the system is halted or crashes while a single file SnapRestore is in progress, then the operation is restarted upon reboot.

12.5.3 The process for restoring data

There is no FilerView wizard to restore files. The volume used for restoring the file must be online and must not be a mirror.

Restoring data from the command line

We cover three methods of restoring data from the command line:

- Single file restore using SnapRestore
- Volume restore using SnapRestore
- FCP LUN restore using SnapRestore
- Mounting the .snapshot directory using UNIX and Windows clients

Listing Snapshots from the Data ONTAP CLI

Consider a situation where a CIFS-share, NFS-share, file, or LUN is recognized as being damaged and a SnapRestore from an existing Snapshot is being considered. Then we need to know which Snapshots exist, so that we can pick the right one to restore from.

Use the **`snap list`** command to check existing Snapshots for a given volume (Example 12-1).

Example 12-1 Snapshot listing

```
itsotuc1> snap list cifs_vol3
Volume cifs_vol3
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  0% ( 0%)    0% ( 0%)  Mar 31 12:00  hourly.0
  0% ( 0%)    0% ( 0%)  Mar 31 08:00  hourly.1
  0% ( 0%)    0% ( 0%)  Mar 31 00:00  nightly.0
  0% ( 0%)    0% ( 0%)  Mar 30 12:00  hourly.2
  0% ( 0%)    0% ( 0%)  Mar 30 08:00  hourly.3

itsotuc1>
```

The Snapshot schedule is determined at creation of the volume. The default Snapshot schedule saves six hourly and two nightly Snapshots.

12.5.4 SnapRestore volume restore

The SnapRestore process for restoring an entire volume is as follows:

1. Open a command line interface on your IBM System Storage N series storage system.
2. Check which Snapshots exist for a given volume with the command:
   `snap list <vol_name>`
3. Restore the Snapshot to an online volume with the command:
   `snap restore -t vol -s <snapshot_name> <vol_name>`

   Where:

   `-t vol` specifies the volume name to revert.

   `-s <snapshot_name>` specifies the name of the Snapshot copy from which to revert

   `<vol_name>` is the name of the damaged volume to be reverted.

4. Verify the status of the restored volume with the command **`vol status <vol_name>`**.
5. After restoring the Snapshot, you might want to check the remaining Snapshots available for the volume using the **`snap list <vol_name>`**. All Snapshots more recent than the one to which we reverted do not exist any more.

SnapRestore volume restore considerations

Important: The volume must be online and must not be a mirror. If you are reverting the root volume, the IBM System Storage N series storage system will be rebooted.

Non-root volumes do not require a reboot. When reverting a non-root volume, all ongoing access to the volume must be terminated, just as it is done when a volume is brought offline. See the man page for the **`vol offline`** command for a description of circumstances that will prevent access to the volume from being terminated and thus prevent the volume from being reverted.

After the reversion, the volume is in the same state as it was when the Snapshot was taken.

Tip: When a volume containing LUNs is restored, the LUNs must be taken offline or be unmapped prior to recovery. Using SnapRestore on a volume that contains LUNs without stopping all host access to those LUNs can cause data corruption and system errors.

SnapRestore volume restore example

We use the following SnapRestore command to restore our damaged volume from a Snapshot:

```
snap restore -t vol -s hourly.2 cifs_vol3
```

Where:

`-t vol` specifies that we are restoring an entire volume.
`-s hourly.2` specifies the Snapshot we are restoring from.
`cifs_vol3` specifies the damaged volume we are restoring to.

In Example 12-2, we show the result the SnapRestore volume restore process where we revert our active volume from the hourly.2 Snapshot.

Example 12-2 SnapRestore process for restoring a volume

```
itsotuc1> date
Wed Mar 30 20:58:04 GMT 2011

itsotuc1> snap list cifs_vol3
Volume cifs_vol3
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  0% ( 0%)    0% ( 0%)  Mar 30 20:00  hourly.0
 20% (20%)   15% (15%)  Mar 30 16:00  hourly.1
 20% ( 0%)   15% ( 0%)  Mar 30 12:00  hourly.2
 20% ( 0%)   15% ( 0%)  Mar 30 08:00  hourly.3

itsotuc1> snap restore -t vol -s hourly.2 cifs_vol3
```

```
WARNING! This will revert the volume to a previous Snapshot.
All modifications to the volume after the Snapshot will be
irrevocably lost.

Volume cifs_vol3 will be made restricted briefly before coming back online.

Are you sure you want to do this? y

You have selected volume cifs_vol3, Snapshot hourly.2

Proceed with revert? y
Share cifs_vol1_clone5 disabled while volume cifs_vol3 is offline.
Wed Mar 30 20:59:19 GMT [wafl.snaprestore.revert:notice]: Reverting volume
cifs_vol3 to a previous Snapshot.
Share cifs_vol1_clone5 activated.
Wed Mar 30 20:59:19 GMT [cifs.shares.activated:info]: Activated 1 CIFS share on
the volume cifs_vol3.
Volume cifs_vol3: revert successful.

itsotuc1> vol status cifs_vol3
         Volume State          Status           Options
      cifs_vol3 online          raid_dp, flex    create_ucode=on,
                                sis              convert_ucode=on
                        Volume UUID: eb289e50-5c00-11e0-b9d8-00a098098a07
                Containing aggregate: 'aggr1'

itsotuc1> snap list cifs_vol3
Volume cifs_vol3
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  0% ( 0%)    0% ( 0%)  Mar 30 12:00  hourly.0
  0% ( 0%)    0% ( 0%)  Mar 30 08:00  hourly.1

itsotuc1>
```

Tip: After you revert the volume to the `hourly.2` Snapshot, you no longer have access to more recent Snapshots, such as the `hourly.0` and `hourly.1` Snapshots. It is because, at the creation time of the `hourly.2` Snapshot, the `hourly.0` and `hourly.1` Snapshots did not exist. Naming of the remaining Snapshots AFTER SnapRestore is also reverted.

12.5.5 SnapRestore single file restore

The SnapRestore process for restoring a lost or damaged file is as follows:

1. Open a command line interface on your IBM System Storage N series storage system.
2. Check which Snapshots exist for a given volume with the command:

   `snap list <vol_name>`

3. Restore the Snapshot to an online volume with the command:

   ```
   snap restore -t file -s <snapshot_name>
                        -r <restore_as_new_path> <path_and_file_name>
   ```

   Where:

   -t file specifies that we are restoring a single file.

   -s <snapshot_name> specifies the name of the Snapshot copy from which to revert.

   -r <restore_as_new_path> is an optional new path to restore the file to.

   <path_and_file_name> specifies the path and file we are restoring.

4. Verify the status of the restored file from the NFS-share.

SnapRestore single file restore example

In the situation described next, we need to restore a single file from a Snapshot. We choose *not* to mount the .snapshot directory of the volume, but to use SnapRestore for this purpose.

The file that we accidently lost is in use on an NFS-share where it has accidently been deleted. The file is called 13_SnapRestore_lab.doc and is located in the directory /mnt/archive /data1/Redbooks of our UNIX server. We know that the file exists in the Snapshot nfs_vol1_snap2 of our volume, which is called nfs_vol1.

We use the following SnapRestore command to restore our lost file volume from a Snapshot:

```
snap restore -t file -s nfs_vol1_snap2
/vol/nfs_vol1/data1/Redbooks/13_SnapRestore_lab.doc
```

Where:

-t file specifies that we are restoring a single file.

-s nfs_vol1_snap2 specifies the Snapshot that we are restoring from.

/vol/nfs_vol1/data1/Redbooks/13_SnapRestore_lab.doc specifies the file that we are restoring.

Example 12-3 shows how the NFS-share is missing the file 13_SnapRestore_lab.doc, which was accidently deleted. Before doing so, we mount the NFS-share.

Example 12-3 Mounting and checking the NFS-share where we lost a file (mount output modified for clarity)

```
[root@localhost ~]# mount -t nfs 9.11.218.114:/vol/nfs_vol1 /mnt/archive

[root@localhost Redbooks]# mount
9.11.218.114:/vol/nfs_vol1 on /mnt/archive type nfs (rw,addr=9.11.218.114)

[root@localhost Redbooks]# pwd
/mnt/archive/data1/Redbooks

[root@localhost Redbooks]# ls -l
total 8
-rw-r--r-- 1 root root 1112 Mar 31 10:50 11_FlexScale.doc
-rw-r--r-- 1 root root  293 Mar 31 10:50 12_FlexVol.doc

[root@localhost Redbooks#
```

Now we move to the N series Data ONTAP CLI where we perform SnapRestore.

Example 12-4 shows how we perform a SnapRestore command to retrieve the lost file from a Snapshot.

Example 12-4 SnapRestore is performed to retrieve the lost file

```
itsotuc1> snap list nfs_vol1
Volume nfs_vol1
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  7% ( 7%)    0% ( 0%)  Mar 31 18:36  nfs_vol1_snap2
 20% (15%)    0% ( 0%)  Mar 31 17:57  nfs_vol1_snap1
 47% (38%)    0% ( 0%)  Mar 31 16:00  hourly.0
 49% ( 7%)    0% ( 0%)  Mar 31 12:00  hourly.1
 51% ( 7%)    0% ( 0%)  Mar 31 08:00  hourly.2
 52% ( 7%)    0% ( 0%)  Mar 31 00:00  nightly.0

itsotuc1> snap restore -t file -s nfs_vol1_snap2
/vol/nfs_vol1/data1/Redbooks/13_SnapRestore_lab.doc

WARNING! This will restore a file from a Snapshot into the active
filesystem.  If the file already exists in the active filesystem,
it will be overwritten with the contents from the Snapshot.

Are you sure you want to do this? y

You have selected file /vol/nfs_vol1/data1/Redbooks/13_SnapRestore_lab.doc,
Snapshot nfs_vol1_snap2

Proceed with restore? y

itsotuc1>
```

Example 12-5 shows how the lost file is now retrieved after Snaprestore on the lost file has been submitted.

Example 12-5 The lost file is now retrieved on our NFS share

```
[root@localhost Redbooks]# pwd
/mnt/archive/data1/Redbooks

[root@localhost Redbooks]s# ls -l
total 24
-rw-r--r-- 1 root root  1112 Mar 31 10:50 11_FlexScale.doc
-rw-r--r-- 1 root root   293 Mar 31 10:50 12_FlexVol.doc
-rw-r--r-- 1 root root 14100 Mar 31 10:50 13_SnapRestore_lab.doc

[root@localhost Redbooks#
```

We successfully retrieved our lost file using IBM N series SnapRestore technology.

Considerations:

- While SnapRestore is able to restore single files from a previously taken Snapshot, you cannot use SnapRestore for single file reversion on files with NT streams, or on directories.
- It means, for a Windows server with a CIFS-share, that ONLY single files in the *root* directory of the volume can be restored.
- For a UNIX-server with an NFS share, it means that single files can be restored from the *root* directory and *subdirectories* of the Snapshot, but not whole directories.
- Therefore it is more likely that server administrators will mount the .snapshot directory of a volume and retrieve the lost or damaged files from there.
- It is, however, not a SnapRestore function, but an integrated part of the Data ONTAP Snapshot technology.
- One case where SnapRestore single file restore does make up a very good tool, is when restoring LUNs, which, in terms of Data ONTAP, are single files and can be restored as single files.

Restoring a LUN with SnapRestore

Restoring an FCP-LUN is basically also just a SnapRestore single file restore. The resulting restore that takes place, is however, not just a single file, but an entire LUN with all its data and files in it. The filename we provide for LUNs being reverted with SnapRestore is the LUN-name, such as /vol/lun1/lun1.

Next, in Example 12-6, we show how to restore a LUN. The LUN must be offline before the SnapRestore action can take place, and the LUN must be manually taken online after the restore is finished.

Example 12-6 Reverting a LUN from a Snapshot

```
itsotuc1> snap restore -t file -s lun1-snap1 /vol/lun1/lun1

WARNING! This will restore a file from a snapshot into the active
filesystem.  If the file already exists in the active filesystem,
it will be overwritten with the contents from the snapshot.

Are you sure you want to do this? y

You have selected file /vol/lun1/lun1, snapshot lun1-snap1

Proceed with restore? y
Thu Mar 31 22:20:55 GMT [lun.snaprestore.notice:notice]: [/vol/lun1/lun1,
43928771, 96, 18539483] SnapRestore: started
itsotuc1> Thu Mar 31 22:20:55 GMT [lun.snaprestore.notice:notice]:
[/vol/lun1/lun1, 43928771, 96, 18539483] SnapRestore: completed

itsotuc1> lun online /vol/lun1/lun1

itsotuc1> lun show
        /vol/lun1/lun1              1.0g (1077511680)    (r/w, online, mapped)

itsotuc1>
```

12.5.6 Mounting the Snapshot directly from a server

In the event that a single file or a directory is lost and cannot be retrieved from a filesystem, Data ONTAP offers Snapshots from which data can be retrieved.

Mounting a Snapshot from a UNIX server with a NFS-share

In the following example, we demonstrate how to mount an existing Snapshot from a UNIX server with a NFS-share already mounted.

Example 12-7 shows the available Snapshots on our N series filer for the volume nfs_vol1.

Example 12-7 The available Snapshots

```
itsotuc1> snap list nfs_vol1
Volume nfs_vol1
working...

  %/used       %/total  date          name
----------  ----------  ------------  --------
  7% ( 7%)    0% ( 0%)  Mar 31 20:00  hourly.0
 16% (11%)    0% ( 0%)  Mar 31 18:36  nfs_vol1_snap2
 27% (15%)    0% ( 0%)  Mar 31 17:57  nfs_vol1_snap1
 50% (38%)    0% ( 0%)  Mar 31 16:00  hourly.1
 52% ( 7%)    0% ( 0%)  Mar 31 12:00  hourly.2
 53% ( 7%)    0% ( 0%)  Mar 31 08:00  hourly.3
 55% ( 7%)    0% ( 0%)  Mar 31 00:00  nightly.0

itsotuc1>
```

Example 12-8 shows a UNIX server where the mount command without parameters lists all mounted filesystems. We have an NFS-share to an N series NAS device mounted on /mnt/archive.

Example 12-8 Existing mounted filesystems on the server

```
[root@localhost ~]# mount
/dev/mapper/VolGroup00-LogVol00 on / type ext3 (rw)
proc on /proc type proc (rw)
sysfs on /sys type sysfs (rw)
devpts on /dev/pts type devpts (rw,gid=5,mode=620)
/dev/sda1 on /boot type ext3 (rw)
tmpfs on /dev/shm type tmpfs (rw)
none on /proc/sys/fs/binfmt_misc type binfmt_misc (rw)
sunrpc on /var/lib/nfs/rpc_pipefs type rpc_pipefs (rw)
9.11.218.114:/vol/nfs_vol1 on /mnt/archive type nfs (rw,addr=9.11.218.114)

[root@localhost ~]#
```

Example 12-9 shows how we mount the N series Snapshot to the filer and list files in the .snapshot directory.

Example 12-9 mounting the .snapshot and listing files (mount output modified for clarity)

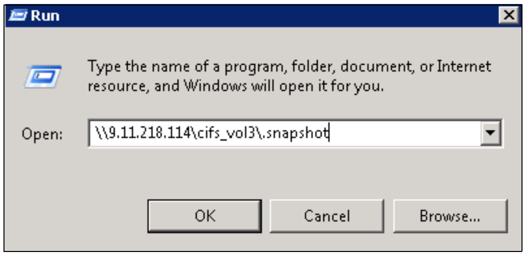
```
[root@localhost ~]# mount -t nfs 9.11.218.114:/vol/nfs_vol1/.snapshot
/mnt/archive/snapshot

[root@localhost ~]# mount
9.11.218.114:/vol/nfs_vol1 on /mnt/archive type nfs (rw,addr=9.11.218.114)
9.11.218.114:/vol/nfs_vol1/.snapshot on /mnt/archive/snapshot type nfs
(rw,addr=9.11.218.114)

[root@localhost Redbooks]# cd /mnt/archive/.snapshot

[root@localhost .snapshot]# ls -l
total 28
drwxr-xr-x 5 root root 4096 Mar 31 11:09 hourly.0
drwxr-xr-x 4 root root 4096 Mar 29 09:29 hourly.1
drwxr-xr-x 4 root root 4096 Mar 29 09:29 hourly.2
drwxr-xr-x 4 root root 4096 Mar 29 09:29 hourly.3
drwxr-xr-x 4 root root 4096 Mar 31 10:50 nfs_vol1_snap1
drwxr-xr-x 5 root root 4096 Mar 31 11:09 nfs_vol1_snap2
drwxr-xr-x 4 root root 4096 Mar 29 09:29 nightly.0

[root@localhost .snapshot]#
```

From this point, we can change directory down into all available Snapshots, and copy files or directories out of the Snapshots.

Mounting a Snapshot from a Windows server with a CIFS-share

In the following example, we demonstrate how to mount an existing Snapshot from a Windows server with a CIFS-share already mounted.

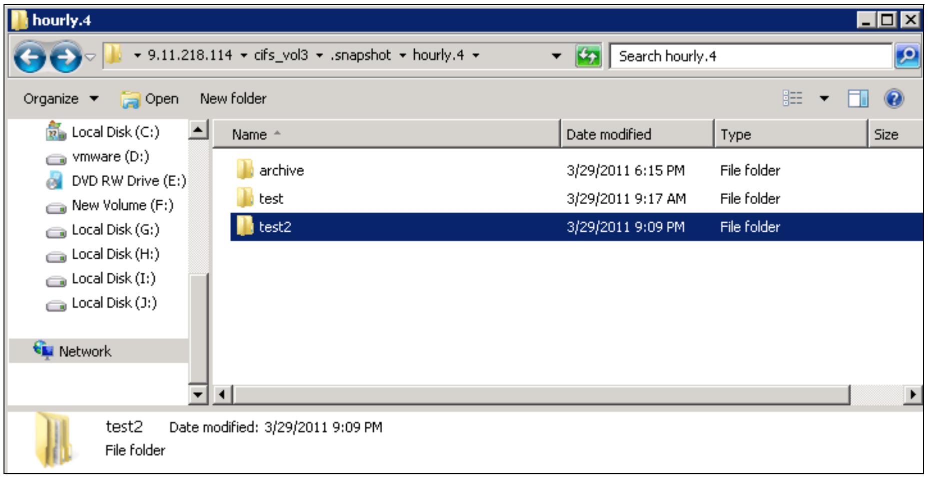
Figure 12-9 shows how a system currently connected to the CIFS-share \\9.11.218.114\cifs_vol3\ now also connects to the .snapshot directory of the cifs_vol3 volume. From the server, click **Start** → **Run** and type `\\9.11.218.114\cifs_vol3\.snapshot` then press Enter.


Run
Type the name of a program, folder, document, or Internet resource, and Windows will open it for you.
Open:
OK
Cancel
Browse...

\\9.11.218.114\cifs_vol3\.snapshot

Figure 12-9 mounting to the .snapshot directory

Figure 12-10 shows how Windows Explorer opens the .snapshot directory on the cifs_vol3 volume on the N series filer.



Name	Date modified	Type	Size
archive	3/29/2011 6:15 PM	File folder	
test	3/29/2011 9:17 AM	File folder	
test2	3/29/2011 9:09 PM	File folder	

test2 Date modified: 3/29/2011 9:09 PM
File folder

Figure 12-10 Windows Explorer opens the .snapshot directory

From here, all Snapshots and datafiles, or whole directories within them, can be copied out of the Snapshot.

12.6 SnapRestore for databases

SnapRestore for databases provides a unique solution to database recovery, rather than restoring large amounts of data from backup tape,

12.6.1 SnapRestore for databases overview

SnapRestore for databases uses this easy procedure:

1. It reverts the entire volume back in time to its state when the Snapshot was taken.
2. It allows the playing of change logs forward to complete the recovery.

This procedure effectively protects data without expensive mirroring or replication. Use SnapRestore where the time to copy data from either a Snapshot or tape into the active file system is prohibitive.

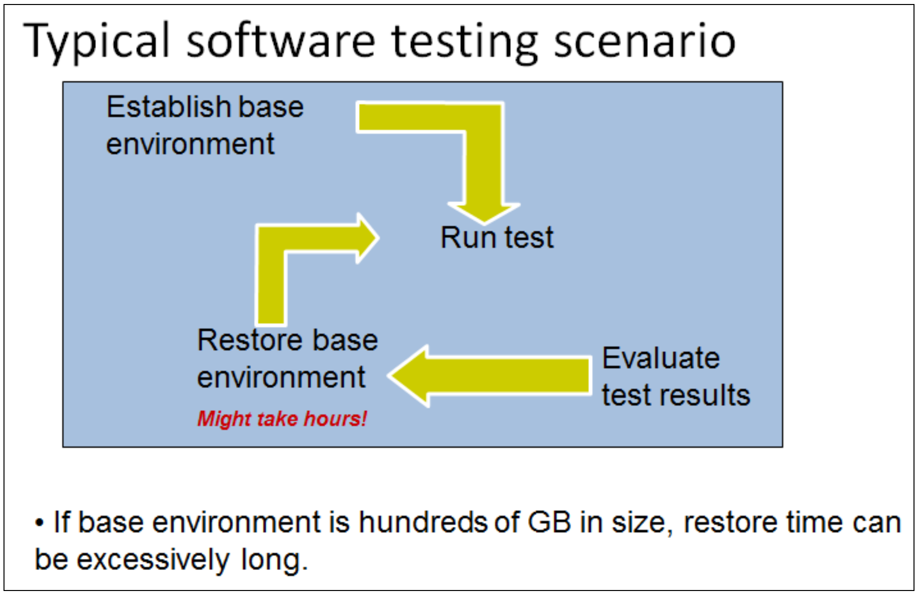
12.6.2 SnapRestore for databases scenario

In this scenario, an Oracle database is damaged and SnapRestore is used to restore it, as described here:

- There is a 550 GB Oracle database that requires recovery.
- An Ultrium tape drive can reach up to 140 MBps with 6 Gbps SAS interface connectivity.
- A normal recovery takes about 1.5 hours plus the log replay time.
- SnapRestore reverts the volume to the same state as when backup was taken, which takes 3 minutes.
- Total recovery using SnapRestore takes 3 minutes plus the log replay time.

Tip: A single use of SnapRestore can pay for the storage system in terms of cost of downtime for the enterprise.

Figure 12-11 shows a typical software testing scenario without using SnapRestore.



Typical software testing scenario

Establish base environment

Run test

Evaluate test results

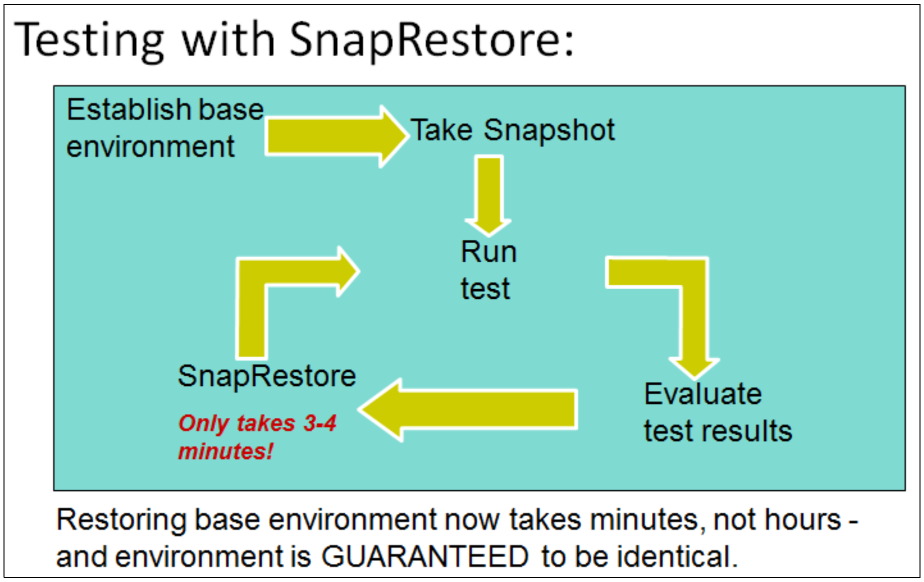
Restore base environment

Might take hours!

- If base environment is hundreds of GB in size, restore time can be excessively long.

Figure 12-11 Software testing scenario

Figure 12-12 shows a testing scenario using SnapRestore.



Testing with SnapRestore:

Establish base environment

Take Snapshot

Run test

Evaluate test results

SnapRestore

Only takes 3-4 minutes!

Restoring base environment now takes minutes, not hours - and environment is GUARANTEED to be identical.

Figure 12-12 Testing using SnapRestore

13

SnapMirror

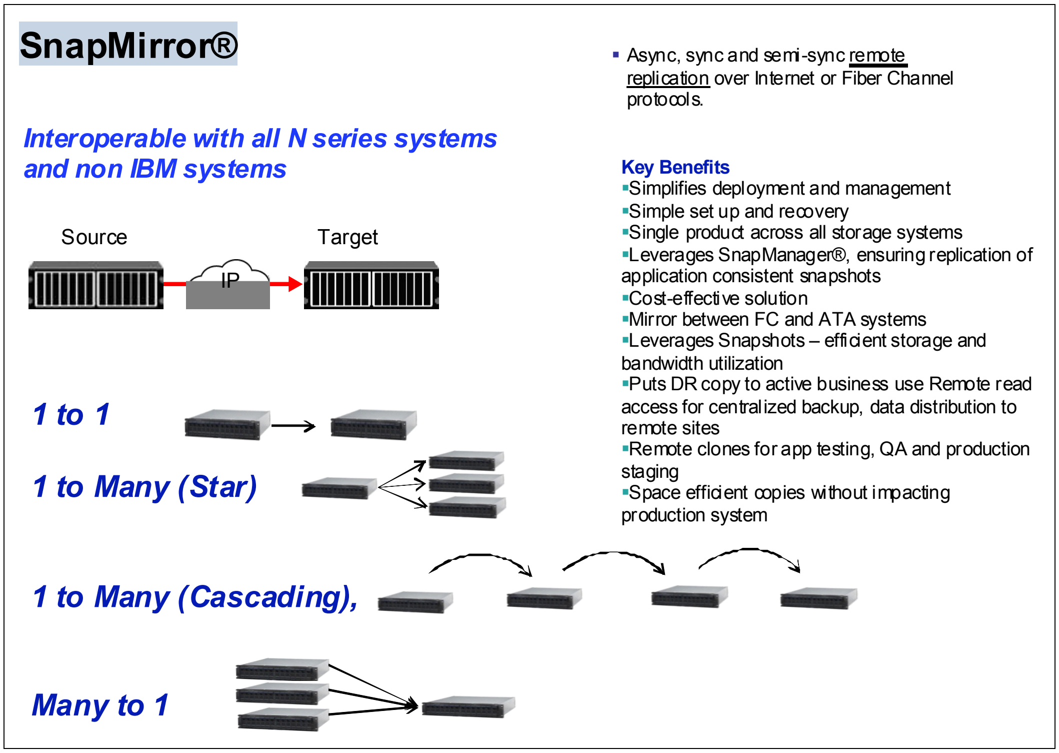
This chapter introduces IBM System Storage N series SnapMirror. It allows a volume (flexible or traditional) or qtree to be replicated between IBM System Storage N series storage systems over a network, typically for backup or disaster recovery purposes. However, it can be used also for application testing, load balancing, and remote access to data. SnapMirror is enhanced by the introduction of FlexVol and FlexClone technology, and by the introduction of synchronous and semi-synchronous modes.

The following topics are covered:

- SnapMirror at a glance
- Introduction to SnapMirror
- The three modes of SnapMirror
- SnapMirror applications
- Synchronous and asynchronous implications
- Volume capacity and SnapMirror
- Guarantees in a SnapMirror deployment
- SnapMirror architecture
- Isolating testing from production
- Cascading mirrors
- Performance impact of synchronous and semi-synchronous modes
- CPU impact of synchronous and semi-synchronous modes
- Network bandwidth considerations
- Replication considerations

13.1 SnapMirror at a glance

SnapMirror, as shown in Figure 13-1, is a feature of Data ONTAP that enables you to replicate data. SnapMirror enables you to replicate data from specified source volumes or qtrees to specified destination volumes or qtrees, respectively.



SnapMirror®

- Async, sync and semi-sync remote replication over Internet or Fiber Channel protocols.

Interoperable with all N series systems and non IBM systems

Key Benefits

- Simplifies deployment and management
- Simple set up and recovery
- Single product across all storage systems
- Leverages SnapManager®, ensuring replication of application consistent snapshots
- Cost-effective solution
- Mirror between FC and ATA systems
- Leverages Snapshots – efficient storage and bandwidth utilization
- Puts DR copy to active business use Remote read access for centralized backup, data distribution to remote sites
- Remote clones for app testing, QA and production staging
- Space efficient copies without impacting production system

Figure 13-1 SnapMirror overview

You need a separate license to use SnapMirror. After the data is replicated to the destination storage system, you can access the data on the destination to perform the following actions:

- Provide users immediate access to mirrored data in case the source goes down.
- Restore the data to the source to recover from disaster, data corruption (qtrees only), or user error.
- Archive the data to tape.
- Balance resource loads.
- Back up or distribute the data to remote sites.

You can configure SnapMirror to operate in one of the following modes:

- Asynchronous mode: SnapMirror replicates Snapshot copies to the destination at specified, regular intervals.
- Synchronous mode: SnapMirror replicates data to the destination as soon as the data is written to the source volume.
- Semi-synchronous mode: SnapMirror replication at the destination volume lags behind the source volume by 10 seconds. This mode is useful for balancing the need for synchronous mirroring with the performance benefit of asynchronous mirroring.

SnapMirror can be used with traditional volumes and FlexVol volumes.

13.2 Introduction to SnapMirror

SnapMirror is a chargeable feature of IBM System Storage N series storage systems (which requires a license code). It allows a volume or qtree to be replicated between IBM System Storage N series storage systems over a network for backup or disaster recovery purposes. But it can be used also for application testing, load balancing, and remote access to data.

After an initial baseline transfer of the entire volume or qtree, as shown in Figure 13-2, subsequent updates only transfer new and changed data from the source to the destination. It makes SnapMirror highly efficient in terms of network bandwidth utilization. The result is an online, read-only volume (mirror) that contains the same data as the source volume at the time of the most recent update.

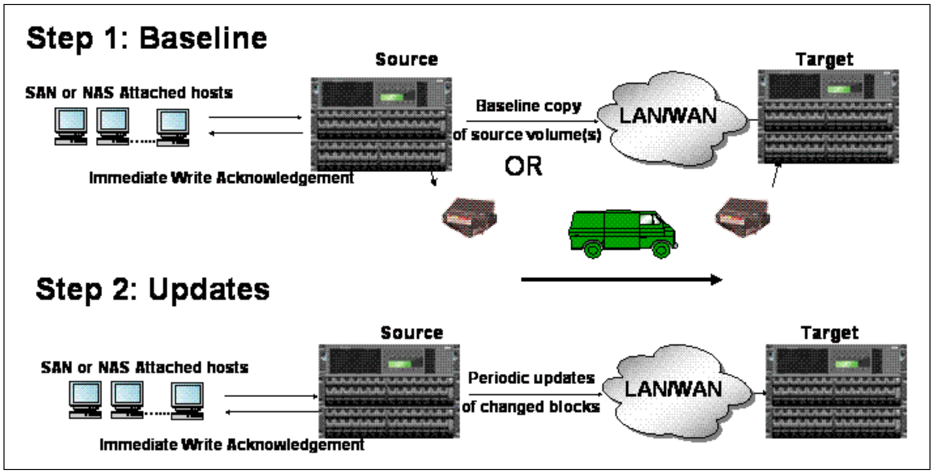

Step 1: Baseline
Source
Target
SAN or NAS Attached hosts
Baseline copy of source volume(s)
LAN/WAN
Immediate Write Acknowledgement
OR
Step 2: Updates
Source
Target
SAN or NAS Attached hosts
Periodic updates of changed blocks
LAN/WAN
Immediate Write Acknowledgement

Figure 13-2 Baseline creation

To replicate data for the first time, the storage system transfers the active file system and all Snapshots from the source volume to the mirror. After the storage system finishes transferring the data, it brings the mirror online. This version of the mirror is the baseline for future incremental changes. Also, like any other volume, after you finish, you can export the mirror for Network File System (NFS) mounting or add a share corresponding to this volume for Common Internet File System (CIFS) sharing.

To make incremental changes on the mirror, the storage system takes regular Snapshots on the source volume according to the schedule specified in the configuration file. By comparing the current Snapshot with the previous Snapshot, the storage system determines what changes it must make to synchronize the data in the source volume and the data in the mirror.

The destination volume is available for read-only access, or the mirror can be *broken* to enable writes to occur on the destination. After breaking the mirror, it can be re-established by synchronizing the changes made to the destination back onto the source file system.

A variation on the basic SnapMirror deployment involves a writable source volume replicated to multiple read-only destinations. The function of this deployment is to make a uniform set of data available on a read-only basis to users from various locations throughout a network to allow for updating that data uniformly at regular intervals.

13.2.1 The need for SnapMirror

SnapMirror software provides a fast, flexible enterprise solution for mirroring or replicating data over local or wide area networks. SnapMirror is used for these purposes:

- Disaster recovery
- Remote enterprise-wide online backup
- Data replication for local read-only access at a remote site
- Application testing on a dedicated read-only mirror
- Data migration between IBM System Storage storage systems

SnapMirror technology is a key component of enterprise data protection strategies. If a disaster occurs at a source site, businesses can access mission-critical data from a mirror on another IBM System Storage N series storage system, ensuring uninterrupted operation (Figure 13-3). Enterprise tape backups are made from SnapMirror, not a production system, reducing CPU load on the production system.

The IBM System Storage N series storage system can be located virtually any distance from the source. It can be in the same building, or on the other side of the world, as long as the interconnecting network has the necessary bandwidth to carry the replication traffic that is generated.

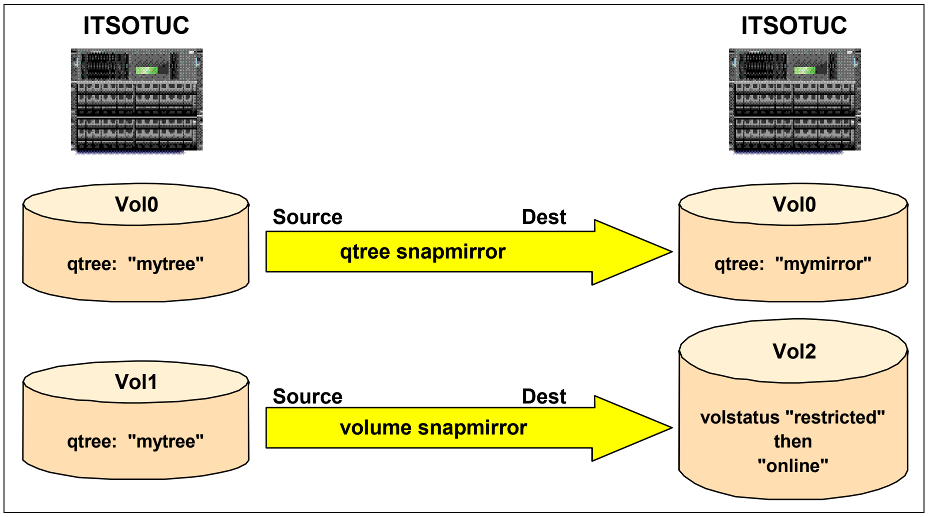


Figure 13-3 SnapMirror

The advantages of SnapMirror over copy vol (created with the `vol copy` command) is that SnapMirror supports these functions:

- Automated and scheduled updates of Snapshot
- Incremental Snapshot updates
- Qtree level replication between the source and the mirror

There are three modes of operation to replicate the data between the source and the mirror volume:

- Asynchronous mode: In the traditional asynchronous mode of operation, updates of new and changed data from the source to the mirror volume occur on a schedule defined by the storage administrator. These updates can be as frequent as once per minute or as infrequent as once per week, depending on user needs.

- Synchronous mode: This mode is also available, which sends updates from the source to the destination as they occur, rather than on a schedule. If configured correctly, it can guarantee that data written on the source system is protected on the mirror volume, even if the entire source system fails due to natural or human-caused disaster. In addition to a standard SnapMirror license, the synchronous feature requires a special license key.
- Semi-synchronous mode: This mode can minimize loss of data in a disaster while also minimizing the performance impact of replication on the source volume. In order to maintain consistency and ease of use, the asynchronous and synchronous interfaces are identical with the exception of a few additional parameters in the configuration file.

Important: Starting with Data ONTAP 8.1 N series systems support volume SnapMirror replication between 32-bit and 64-bit volumes.

13.2.2 Rules for using SnapMirror

The following rules apply when using SnapMirror:

- The source and the mirror volume must be of the same volume type, that is, both must be a traditional or a flexible volume.
- The mirror volume must be manually created because SnapMirror does not automatically create the mirror volume.
- SnapMirror can be used through a firewall. The port number that SnapMirror listens for connections is 10566. You have to allow a range of TCP ports from 10565 to 10569.
- The source volume must be online.
- The mirror cannot be the root volume.
- The capacity of the mirror must be greater than or equal to the capacity of the source volume. The configuration of the volumes, however, can be different.
- The mirror volume must run under a version of Data ONTAP equal to or later than that of the SnapMirror source volume. If the IBM System Storage N series storage systems must be upgraded, then the IBM System Storage N series storage system that hosts the mirror volume must be upgraded before the IBM System Storage N series storage systems that host the source volume. This requirement does not apply for qtree replication. This rule only applies for volume replication.
- Quotas cannot be enabled on a mirror.
- Qtrees cannot be created on a mirror. However, if one qtree exists in the source volume, the storage system mirrors the qtrees to the mirror.
- SnapMirror replicates a file system on one volume to a read-only copy on another volume.
- SnapMirror is based on Snapshot technology. Only changed blocks are copied after the initial mirror is established.
- It runs over IP or FC.
- Data is accessible read-only at remote sites.
- Replication is either volume based or qtree based.
- IP name resolution must be configured properly before the use of SnapMirror. If no DNS or host file is configured, then the IP address must be used and the Enable IP Checking must be enabled at SnapMirror.

Restriction: The maximum number of entries in /etc/snapmirror.conf is 1024 lines.

The maximum number of concurrent replication operations with the Nearstore feature enabled varies per N series model:

- For Volume SnapMirror:
  - N3000 models can handle up to 100 concurrent replication operations.
  - N6000 and N7000 models can handle up to 300 concurrent replication operations.
- For Qtree SnapMirror:
  - N3000 models can handle up to 160 concurrent replication operations.
  - N6000 and N7000 models can handle up to 512 concurrent replication operations.

13.3 The three modes of SnapMirror

SnapMirror can be used in three different modes:

- Asynchronous
- Synchronous
- Semi-synchronous

We explain these modes in more detail in the following sections.

13.3.1 Asynchronous mode

In asynchronous mode, shown in Figure 13-4, SnapMirror performs incremental, block-based replication as frequently as once per minute. Consult your technical team for the best plan for your environment or to determine whether synchronous SnapMirror is a better match. The performance impact on the source IBM System Storage N series storage system is minimal as long as the system is configured with sufficient CPU and disk I/O resources.

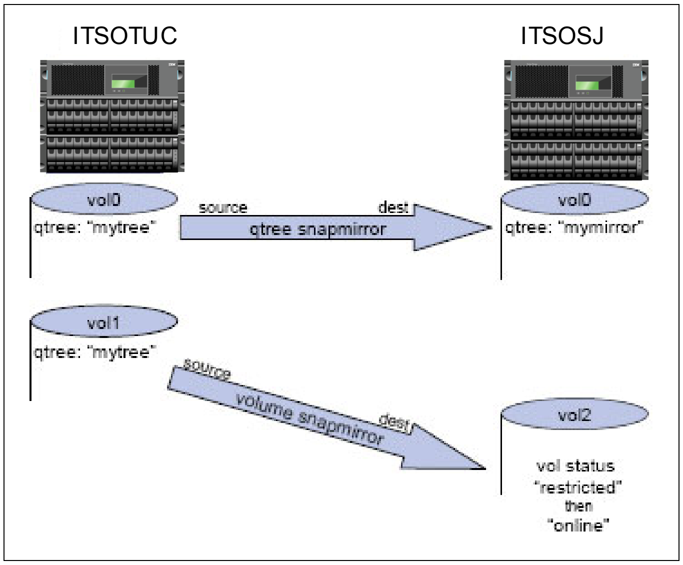


Figure 13-4 Asynchronous SnapMirror options

Asynchronous mode initialization

The first and most important step in asynchronous mode involves the creation of a one-time, baseline transfer of the entire data set. It is required before incremental updates can be performed.

This operation proceeds as follows:

1. The primary storage system takes a Snapshot (a read-only, point-in-time image of the file system).
2. Referring to Figure 13-2 on page 177, this Snapshot is called the *baseline* copy.
3. All data blocks referenced by this Snapshot and any previous Snapshot copies are transferred and written to the secondary file system.
4. After initialization is complete, the primary and secondary file systems will have at least one Snapshot in common.

Asynchronous mode updates

After initialization, both scheduled or manually triggered updates can occur. Each update transfers only the new and changed blocks from the primary to the secondary file system. This operation proceeds as follows:

1. The primary storage system takes a Snapshot.
2. The new Snapshot is compared with the baseline Snapshot to determine which blocks have changed.
3. The changed blocks are sent to the secondary and written to the file system.
4. After the update is complete, both file systems have the new Snapshot, which becomes the baseline Snapshot for the next update.

Because asynchronous replication is periodic, SnapMirror is able to consolidate writes on the source volume and conserve network bandwidth.

13.3.2 Synchronous mode

Synchronous SnapMirror is a SnapMirror feature that replicates data from a source volume to a partner destination volume at or near the same time that it is written to the source volume, rather than according to a predetermined schedule. This insures that data written on the source system is protected on the destination even if the entire source system fails. It guarantees zero data loss in the event of a failure, but can have a significant impact on performance. It is not necessary or appropriate for all applications.

Synchronous SnapMirror (Figure 13-5) replicates data between single storage systems or clustered storage systems located at remote sites using IP or FCP infrastructure with no special converters required. Synchronous SnapMirror is simply a mode of operation or feature that has recently been added to the SnapMirror software. This mode requires a special license key to function properly.

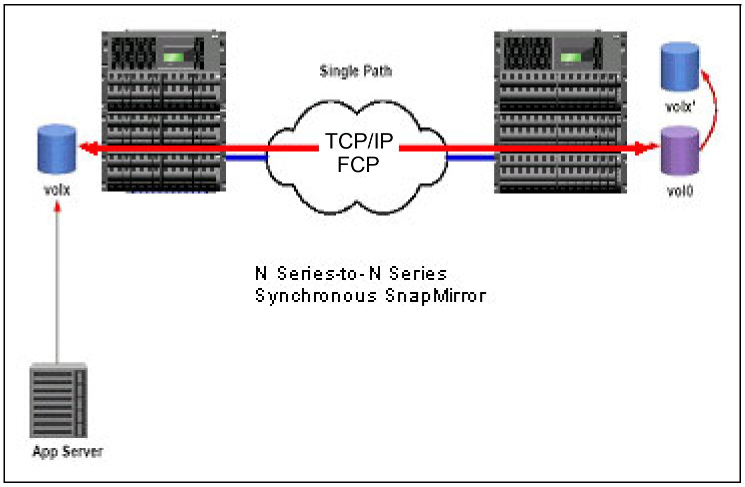


Figure 13-5 Single Path SnapMirror

Synchronous SnapMirror is supported only for configurations of which the source system and destination systems are the same type of system and have the same disk geometry. The type of system and disk geometry of the destination impacts the perceived performance of the source system. Therefore, the destination system must have the bandwidth for the increased traffic and for message logging. Log files are kept on the root volume. Therefore, you must ensure that the root volume spans enough disks to handle the increased traffic. The root volume must span four to six disks.

For the best performance, you need to have a dedicated high-bandwidth, low-latency network between the source and destination storage systems. Synchronous SnapMirror can support traffic over Fibre Channel and IP transports.

Disk configurations supported

The following configurations are supported for synchronous SnapMirror relationships:

- A source storage system with only ATA disks attached to a destination storage system with only ATA disks attached
- A source storage system with only Fibre Channel disks attached to a destination storage system with only Fibre Channel disks attached

Any other configuration of attached disks, such as a combination of ATA and Fibre Channel disks, is not supported.

Terminology

To avoid any potential confusion, it is appropriate to review exactly what is meant by the term *synchronous* in this context. The best way to do this task is to examine a scenario where the primary data storage device fails completely and then examine the disaster's impact on an application.

In a typical application environment, the following steps occur:

1. A user saves information in the application.
2. The client software communicates with a server and transmits the information.
3. The server software processes the information and transmits it to the operating system on the server.
4. The operating system software sends the information to the storage.
5. The storage acknowledges receipt of the data.
6. The operating system tells the application server that the write is complete.
7. The application server tells the client that the write is complete.
8. The client software tells the user that the write is complete.

In most cases, these steps take only tiny fractions of a second to complete. If the storage system fails in such a way that all data on it is lost (for example, as a result of a fire or flood that destroys all of the storage media), the impact to an individual transaction varies based on *when* the failure occurs, as explained here:

- If the failure occurs *before* step 5, the storage never acknowledges receipt of the data. It results in the user receiving an error message from the application, indicating that it failed to save the transaction.
- If the failure occurs *after* step 5, the user sees client behavior that indicates correct operation (at least until the following transaction is attempted). Despite the indication by the client software (in step 8) that the write was successful, the data is lost.

The first case is obviously preferable to the second, because it provides the user or application with knowledge of the failure and the opportunity to preserve the data until the transaction can be attempted again. In the second case, the data can be discarded based on the belief that it is already safely stored.

With traditional asynchronous SnapMirror, data is replicated from the primary storage to a secondary or destination storage device on a schedule. If this schedule were configured to cause updates once per hour, for example, it is possible for a full hour of transactions to be written to the primary storage, and acknowledged by the application, only to be lost when a failure occurs before the next update. For this reason, many customers attempt to minimize the time between transfers. Some customers replicate as frequently as once per minute, which significantly reduces the amount of data that can be lost in a disaster.

This level of flexibility is good enough for the vast majority of applications and users. In most real-world environments, loss of one minute or five minutes of data is of trivial concern compared with the downtime incurred during such an event. Any disaster that completely destroys the data on the IBM System Storage N series storage system will most likely also destroy the relevant application servers, critical network infrastructure, and so on.

However, there are some customers and applications that have a zero data loss requirement even in the event of a complete failure at the primary site, as shown in Figure 13-6.

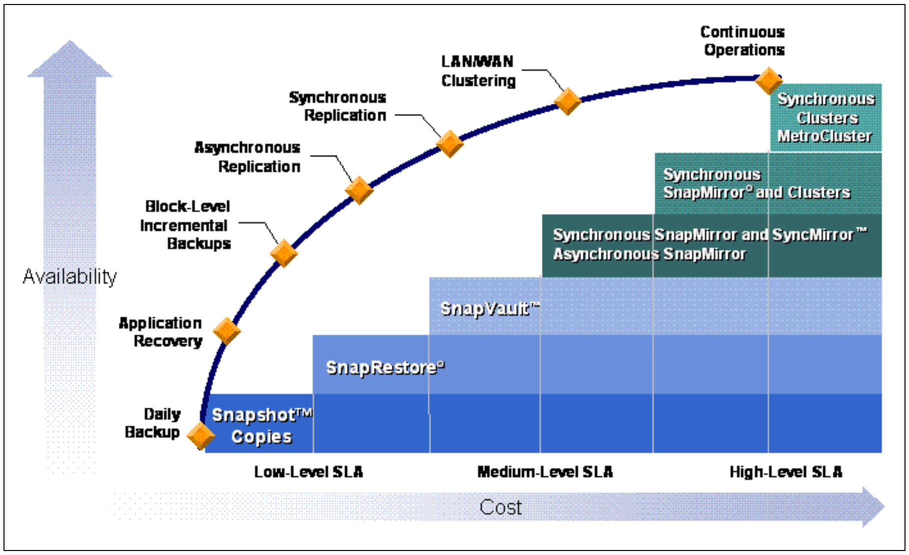


Figure 13-6 Availability

For these situations, synchronous mode is appropriate because it modifies the application environment described such that replication of data to the secondary storage occurs with *each* transaction, as explained here:

1. A user saves information in the application.
2. The client software communicates with a server and transmits the information.
3. The server software processes the information and transmits it to the operating system on the server.
4. The operating system software sends the information to the primary storage.
5. The primary storage sends the information to the secondary storage.
6. The secondary storage acknowledges receipt of the data.
7. The primary storage acknowledges receipt of the data.
8. The operating system tells the application server that the write is complete.
9. The application server tells the client that the write is complete.
10. The client software tells the user that the write is complete.

The key difference, from the application's point of view, is that the storage does not acknowledge the write until the data has been written to both the primary and the secondary storage. It has some performance impact, as described later, but modifies the failure scenario in beneficial ways:

- If the failure occurs *before* step 7, the storage never acknowledges receipt of the data. It results in the user receiving an error message from the application, indicating that it failed to save the transaction. It causes inconvenience, but no data loss.
- If the failure occurs *during or after* step 7, the data is safely preserved on the secondary storage system despite the failure of the primary.

Attention: Regardless of what technology is used, it is always possible to lose data. The key point here is that with synchronous mode, loss of data that has been acknowledged is prevented.

Operation

The first step involved in synchronous replication is a one-time, baseline transfer of the entire data set, just as in asynchronous mode, as described in 13.3.1, “Asynchronous mode” on page 180.

Tip: SnapMirror must be licensed before synchronous SnapMirror.

After the baseline transfer has completed, SnapMirror can change to synchronous mode, as follows:

1. Asynchronous updates occur, as described earlier, until the primary and secondary file systems are close to being synchronized.
2. NVLOG forwarding begins. It is a method for transferring updates as they occur.
3. Consistency point (CP) synchronization begins. It is a method for ensuring that writes of data from memory to disk storage are synchronized on the primary and secondary systems.
4. New writes from clients or hosts on the primary file system are blocked until acknowledgment of those writes has been received from the secondary system.
5. A final update occurs using the same method as asynchronous updates.

After SnapMirror has determined that all data acknowledged by the primary has been safely stored on the secondary, the system is in synchronous mode. At this point, the output of a SnapMirror status query shows that the relationship is *in sync*.

Attention: If the environment is unable to maintain synchronous mode (because of networking or destination issues), SnapMirror drops to asynchronous mode. When the connection is re-established, the source IBM System Storage N series asynchronously replicates data to the destination once each minute, until synchronous replication is re-established. After it occurs, a message will be logged of the change of status (*into* or *out of* synchronous status). This *safety net* is known as *fail-safe synchronous*.

Synchronous mode paths

More than one physical path might be required for a synchronous mirror. Synchronous SnapMirror supports up to two paths for a particular relationship. These paths can be Ethernet, Fibre Channel, or a combination of the two.

Multipath support allows synchronous and semi-synchronous traffic to be load-balanced between these paths and provides for failover in the event of a network outage. There are two modes of multipath operation:

- Multiplexing mode, as shown in Figure 13-7, in which both paths are used simultaneously and load balancing transfers across the two. When a failure occurs, the load from both transfers moves to the remaining path.

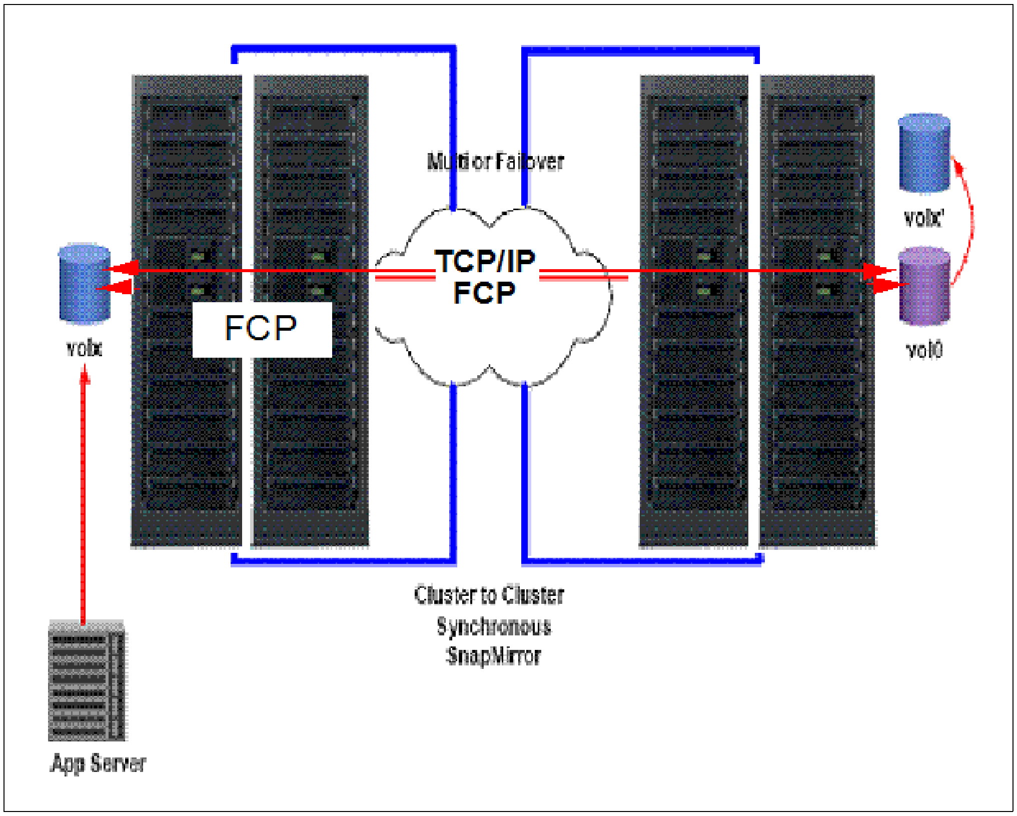


Figure 13-7 SnapMirror multipath

- Failover mode, in which one path is specified as the primary path in the configuration file. This path is the desired path and is used until a failure occurs. The second path is then used.

The role of NVLOG in synchronous SnapMirror

NVLOG forwarding is a critical component of synchronous mode operation. It is the method used for write operations submitted from clients against the primary file systems to be replicated to the destination.

When NVLOG forwarding is active in synchronous mode, some modifications are made as described here:

- The request is journaled in non-volatile RAM (NVRAM). It is also recorded in cache memory and forwarded over the network to the SnapMirror destination system, where it is journaled in NVRAM and cache memory.
- After the request is safely stored in NVRAM and cache memory on both the primary and secondary systems, Data ONTAP acknowledges the write to the client system, and the application that requested the write is free to continue processing.

As can be seen, NVLOG forwarding is the primary mechanism by which data is synchronously protected.

The synchronous SnapMirror replication mode synchronously replicates writes from the source NVLOG RAM to the destination NVLOG RAM. After the data transfer is completed, an acknowledgement is sent from the destination NVLOG RAM. It is known as NVLOG forwarding.

At this point, data is not yet written to the disk. After the destination NVRAM is half-full or 10 seconds to previous Consistency Point (CP) time, Data ONTAP creates a tetris that computes RAID parity information and data blocks to be written to the disk. It also forwards the same tetris data to a destination system until the CP data is written to the destination. It is known as *CP forwarding*.

If there is a delay in NVLOG or CP forwarding due to network or storage error synchronous replication, the synchronous mode falls back to the asynchronous mode. As there are two writes, one during NVLOG forwarding and another during CP forwarding, you must consider 2x of data written in synchronous replication mode.

13.3.3 Semi-synchronous mode

SnapMirror also provides a semi-synchronous mode, sometimes called *semi-sync*. Synchronous SnapMirror can be configured to lag behind the source volume by a user-defined number of write operations or milliseconds.

Semi-synchronous mode overview

This mode is like asynchronous mode in that the application does not need to wait for the secondary storage to acknowledge the write before continuing with the transaction. (Of course, for this reason, it is possible to lose acknowledged data.)

This mode is also like synchronous mode in that updates from the primary storage to the secondary storage occur right away, rather than waiting for scheduled transfers. This makes the potential amount of data lost in a disaster very small. Semi-synchronous mode minimizes data loss in a disaster, while also minimizing the extent to which replication impacts the performance of the source system.

Semi-synchronous mode provides a middle ground that keeps the primary and secondary file systems more closely synchronized than asynchronous mode. Configuration of semi-synchronous mode is identical to configuration of synchronous mode, with the addition of an option that specifies how many writes can be outstanding (unacknowledged by the secondary system) before the primary system delays acknowledging writes from the clients.

Internally, semi-synchronous mode works identically to synchronous mode in most cases. The only difference lies in how quickly client writes are acknowledged. The replication methods used are the same. However, it is possible to configure semi-synchronous mode in a way that changes the replication strategy. A CP is triggered when NVRAM is one-half full, or every 10 seconds, whichever occurs sooner. If semi-synchronous mode is configured to allow unacknowledged transactions greater than 10 seconds old, SnapMirror falls back to performing CP synchronization only. NVLOG forwarding is halted, because a CP synchronization is sufficiently frequent to meet the service level requested.

When a CP synchronization occurs under such circumstances, the tetris sent to the secondary IBM System Storage N series storage system includes not just the list of data blocks to be written, but also the content of those data blocks. It is because with NVLOG forwarding disabled, the secondary system does not have a copy of the data until the CP synchronization occurs.

For the vast majority of customer configurations, NVLOG forwarding is desirable. Thus, configuring SnapMirror to allow more than 10 seconds of outstanding data is not desirable for customers who want higher synchronicity levels.

However, if NVLOG forwarding is not required, specifying a large time value for outstanding data might reduce the overall CPU usage on the primary storage system. This configuration can allow for significant increases in overall throughput if CPU usage is a limiting factor.

Tip: Unlike asynchronous mode, which can replicate either volumes or quota trees, synchronous and semi-synchronous modes work only with volumes.

Semi-synchronous mode scenario

The semi-synchronous mode scenario consists of the following steps:

1. A user saves information in the application.
2. The client software communicates with a server and transmits the information.
3. The server software processes the information and transmits it to the operating system on the server.
4. The operating system software sends the information to the primary storage.
5. The primary storage sends the information to the secondary storage. The primary storage simultaneously acknowledges receipt of the data.
6. The operating system tells the application server that the write is complete.
7. The application server tells the client that the write is complete.
8. The client software tells the user that the write is complete.
9. At some point after step 5, the secondary acknowledges receipt of the data. (Note that step 9 can potentially occur before, or simultaneously with, step 6.)

If the secondary storage system is slow or unavailable, it is possible that a large number of transactions can be acknowledged by the primary storage system and yet not protected on the secondary. These transactions represent a window of vulnerability for the loss of acknowledged data.

For a window of zero size, customers can use fully synchronous mode rather than semi-sync. If using semi-sync, and the size of this window is customizable based on user and application needs. It can be specified as a number of operations, milliseconds, or seconds.

If the number of outstanding operations equals or exceeds the number of operations specified by the user, further write operations will not be acknowledged by the primary storage system until some have been acknowledged by the secondary.

Likewise, if the oldest outstanding transaction has not been acknowledged by the secondary within the amount of time specified by the user, further write operations will not be acknowledged by the primary storage system until all responses from the secondary are being received within that time frame.

13.4 SnapMirror applications

You can use SnapMirror for the following applications:

- For data replication for local read access at remote sites:
  - Slow access to corporate data is eliminated.
  - You can off-load tape backup CPU cycles to a mirror (Figure 13-8).

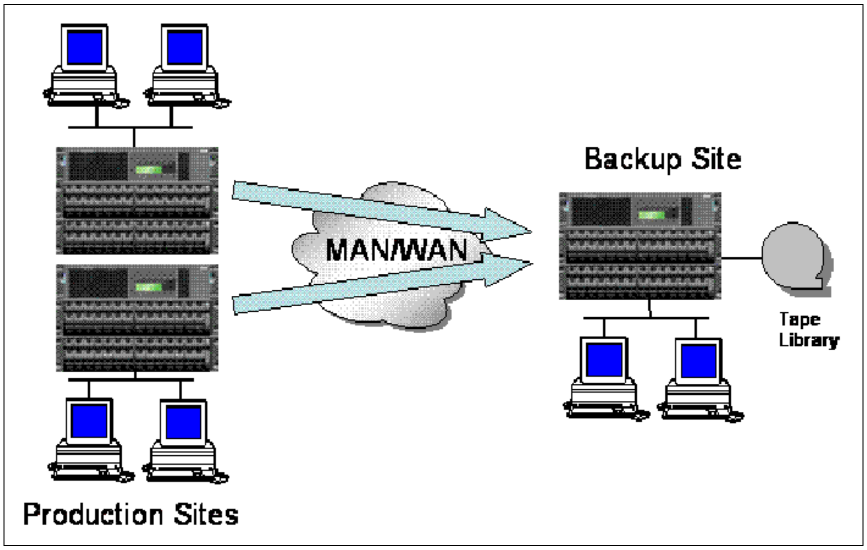


Figure 13-8 Data replication for warm backup/off-load

- To isolate testing from production volume:
  - ERP testing
  - Offline reporting
- For cascading mirrors:

  Replicated mirrors on a larger scale
- For disaster recovery.

  Replication to *hot site* for mirror failover and eventual recovery
- The Data ONTAP SnapMirror feature can be used in combination with FlexClone volumes to perform migration faster and more efficiently:
  - For enterprises with a warm backup site, or those that must off-load backups from production servers
  - For generating queries and reports on near-production data

13.5 Synchronous and asynchronous implications

With synchronous SnapMirror, a Snapshot is made on the destination volume every time that a write is done on the source. The Snapshot can be deleted from the clone, but not from the source volume, while the SnapMirror relationship is *in sync*. Synchronous SnapMirror has a *hard lock*. In contrast, asynchronous SnapMirror has a *soft lock*. If the process falls out of synchronous mode, it reverts to asynchronous mode and becomes a soft lock.

Synchronous SnapMirror keeps the source and destination in sync as much as possible:

- If the NVLOG channel requests (per op) time out
- If the CP on the source takes more than one minute
- If network errors persist even after three retransmissions
- If the source or destination fails to restart
- If the network connection fails

In such situations, synchronous SnapMirror completes an asynchronous update within one minute. It also turns on consistency point forwarding and NVLOG forwarding.

13.6 Volume capacity and SnapMirror

The source capacity must be less than or equal to the destination capacity when using flexible volumes. When the administrator performs a SnapMirror break and the destination capacity is greater than the source capacity, the destination volume shrinks to match the capacity of the smaller source volume. It is a much more efficient usage of disk space because it avoids consumption of unused space.

13.7 Guarantees in a SnapMirror deployment

Guarantees determine how the aggregate preallocates space to the flexible volume. SnapMirror never enforces guarantees, regardless of how the source volume is set. As long as the destination volume is a SnapMirror destination (replica), the guarantee is volume-disabled. Subsequently, the guarantee mode is the same as the volume mode when the volume is broken off using SnapMirror break.

13.8 SnapMirror architecture

A full Snapshot is created (named Snap A) and then a baseline transfer to the target volume is performed, as shown in Figure 13-9.

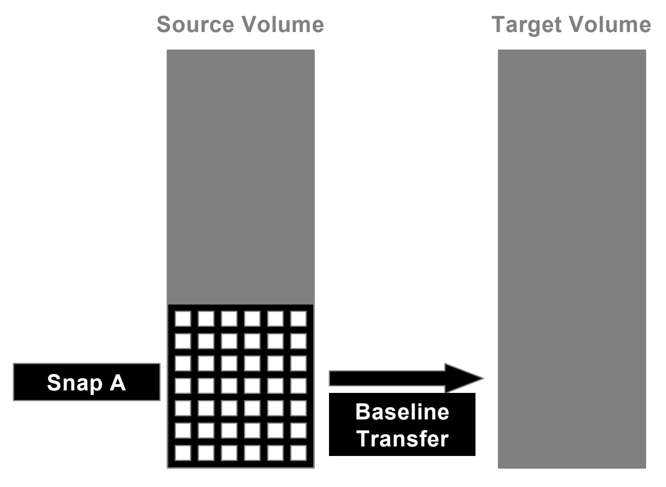


Figure 13-9 SnapMirror detail

Figure 13-10 displays the SnapMirror internal operation.

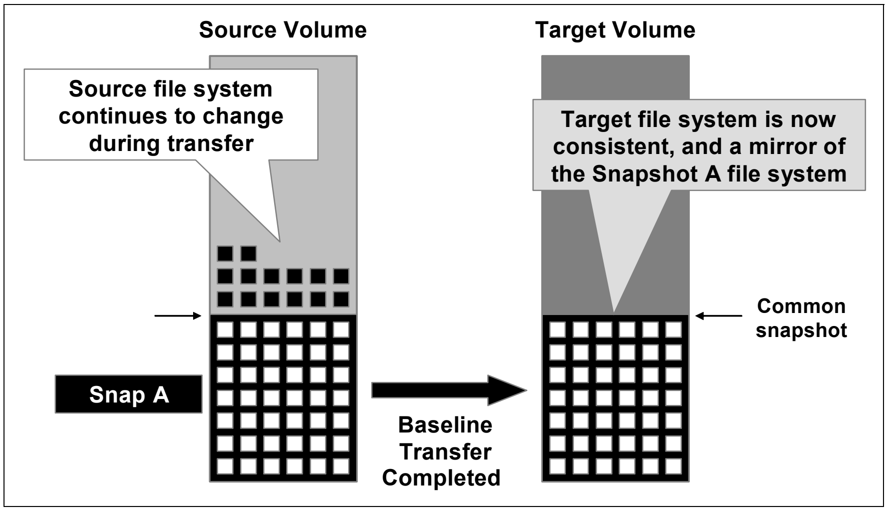

Source Volume
Target Volume
Source file system continues to change during transfer
Target file system is now consistent, and a mirror of the Snapshot A file system
Common snapshot
Snap A
Baseline Transfer Completed

Figure 13-10 SnapMirror internals

As you might expect in a 24x7 operation, updates to the source volume continue to occur while the baseline image is transferred, leading to the creation of Snap B. The integrity of Snap A is maintained with Snapshot and at a point in time. The baseline image of Snap A is transferred, as shown in Figure 13-11.

.

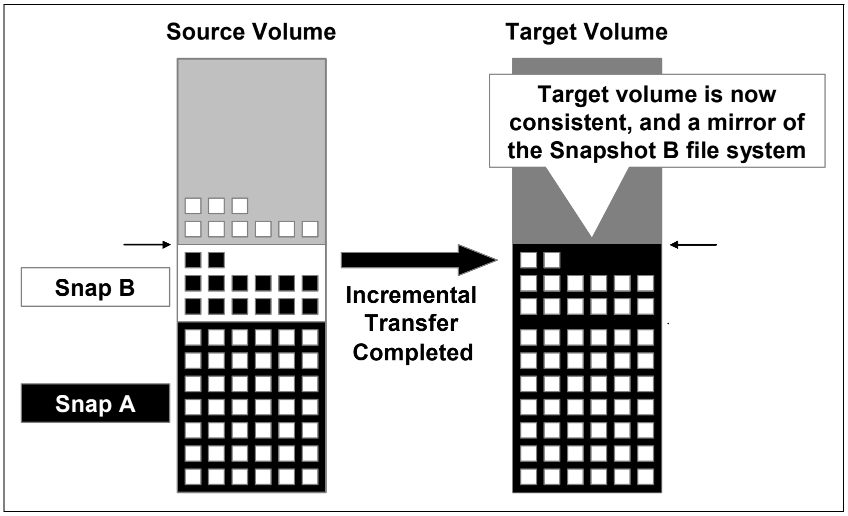

Source Volume
Target Volume
Target volume is now consistent, and a mirror of the Snapshot B file system
Snap B
Snap A
Incremental Transfer Completed

Figure 13-11 Consistent SnapMirror

For reference purposes, we refer to Snap A as T0 time and to Snap B as T1 time. At T1 time, a Snapshot is done again, capturing a image of the volume at that point in time. After completion of Snap B, an incremental transfer is initiated (it is incremental because only portions of the volume have changed since T0 time).

Updates continue to occur, but the Snapshot maintains the integrity of Snap B. After completion of the incremental transfer, there is now a consistent full image copy of the source volume as it looked at T1 time (Figure 13-12).

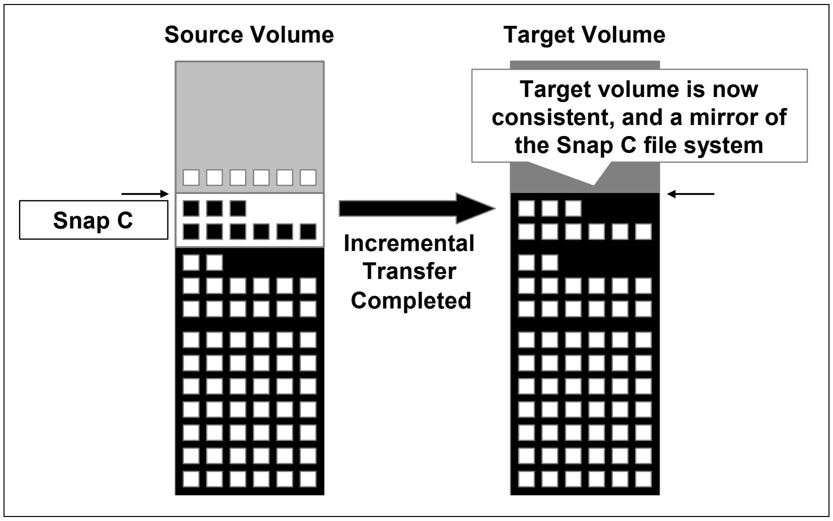

Source Volume
Target Volume
Target volume is now consistent, and a mirror of the Snap C file system
Snap C
Incremental Transfer Completed

Figure 13-12 Snap C consistency

Operations continue and now another Snapshot is done (Snap C or T2 time), capturing an image of the volume at that point in time. After completion of Snap C, an incremental transfer is initiated (it is incremental because only portions of the volume have changed since T1 time or snap B).

Updates continue to occur but the Snapshot maintains the integrity of Snap C. After completion of the incremental transfer, there is now a consistent full image copy of the source volume as it looked at T2 time.

13.9 Isolating testing from production

After a consistent image (that is, a baseline image and subsequent incremental transfers) is captured, the SnapMirror relationship is broken and the target is enabled for write operations, read for application testing, and so on (Figure 13-13).

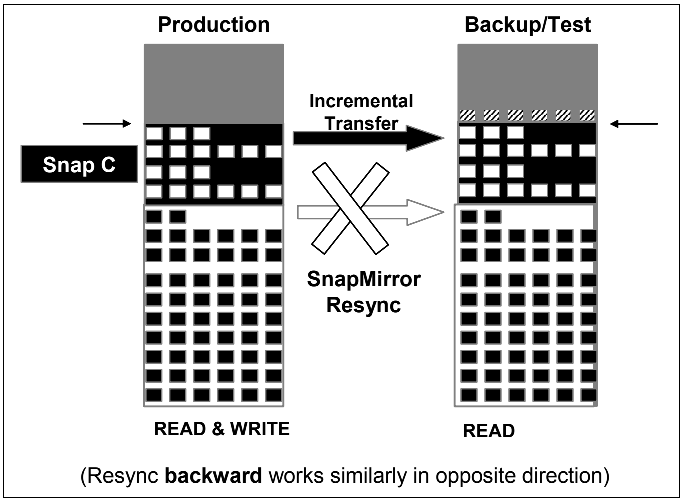


Figure 13-13 Isolate testing from production

During this time, the source volumes continue to be available online. Note that at any time, you can resync forward by re-establishing the mirror relationship.

13.10 Cascading mirrors

Cascading is a method of replicating from one destination system to another in a series, as shown in Figure 13-14. For example, you might want to perform synchronous replication from the primary site to a nearby secondary site, and asynchronous replication from the secondary site to a far-off tertiary site. Currently, only one synchronous SnapMirror relationship can exist in a cascade.

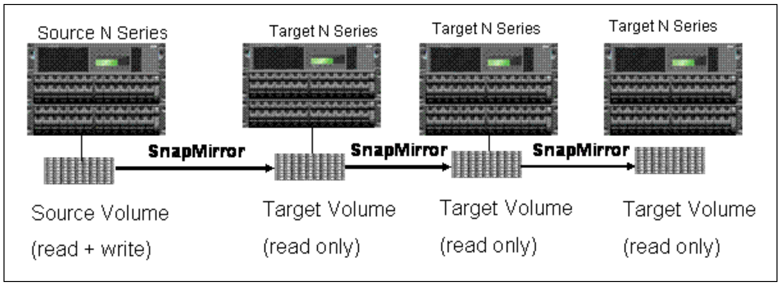


Figure 13-14 Cascading mirrors

13.10.1 Cascading replication

Figure 13-15 shows an example of cascading replication.

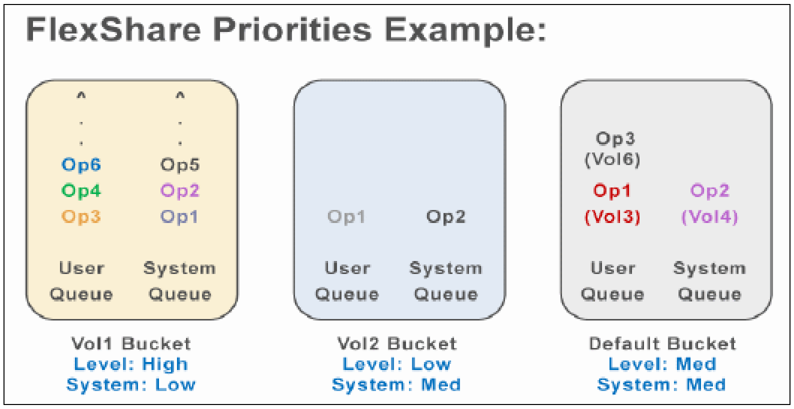


Figure 13-15 Cascading replication example

You can replicate to multiple (30) locations across the continent:

- You send data only once across the WAN.
- You reduce resource utilization on the source IBM System Storage N series storage system.

13.10.2 Disaster recovery

SnapMirror can become one of the methods to recover or continue operations in a disaster, as shown in Figure 13-16. Here are various requirements for business continuity that might require SnapMirror:

- Enterprises that cannot afford the downtime of a full restore from tape (days)
- Data-centric environments
- The mean time to recovery when a disaster occurs must be reduced

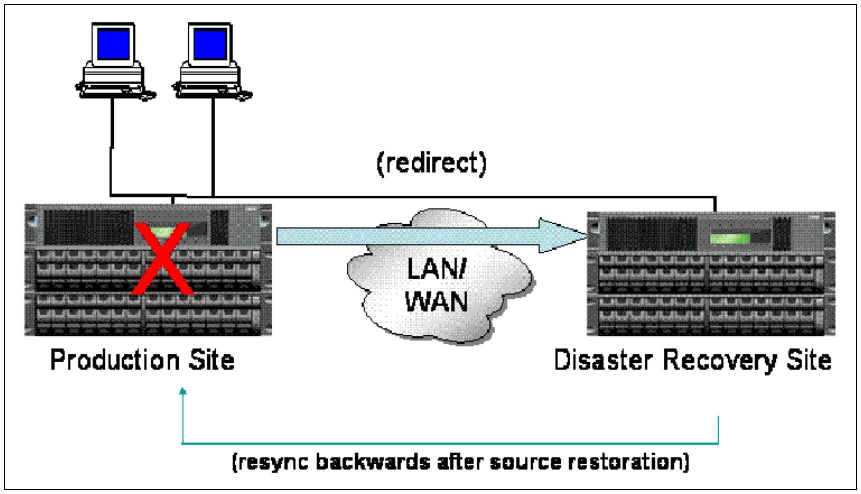


Figure 13-16 Disaster recovery example

13.11 Performance impact of synchronous and semi-synchronous modes

Performance is a complex and difficult area to quantify. It is beyond the scope of this book to describe IBM System Storage N series storage system performance. But we do examine what affects synchronous SnapMirror has on individual system performance and how synchronous SnapMirror affects overall performance.

It is important to note that the guidelines and preferred practices that follow are *not* exact measurements. Any synchronous replication method, regardless of the technology used, will have an impact on the performance of applications using the storage.

Understanding business requirements for application performance and data protection allows an organization to make informed choices between various data protection strategies. When examining the application performance impact of synchronous or semi-synchronous replication, there are two primary factors to consider:

- Overall system throughput might be reduced due to these factors:
  - CPU impact imposed by the replication process
  - Network bandwidth constraints between the primary and secondary storage
  - Slower system performance on the secondary storage than on the primary
  - Impact of workload on the secondary storage, thus reducing its ability to service replication traffic
  - Root volume performance on the secondary system
- Individual write operations take longer to complete due to the need for additional processing of each operation and network latency between the primary and secondary storage.

Our description of these factors focuses on the primary storage system and its client applications. The preferred practice is to provide a dedicated secondary storage system for synchronous or semi-synchronous replication, and we assume that this preferred practice is being followed. Thus, performance impact on the secondary storage system is not considered an important issue except insofar as it creates an impact on the primary storage system.

13.12 CPU impact of synchronous and semi-synchronous modes

When a system running SnapMirror in synchronous or semi-synchronous mode receives a write request from a client, it must do all of the standard processing that will be required normally. It also must do additional processing, related to SnapMirror, to transfer the information to the secondary storage system. This adds significant CPU impact to every write operation.

Although it is beyond the scope of this book to describe the individual components of this CPU impact in detail, it is helpful to illustrate the concept using an example. Reading or writing information over network connections is one task performed by an IBM System Storage N series storage system. Higher volumes of data being passed across the network result in more CPU usage on the storage system. So if the network-related CPU impact is considered independently of other factors, a client writing data to the IBM System Storage N series storage system at 30 MBps will use about half of the CPU used by a client writing data at 60 MBps.

When replicating data in synchronous or semi-synchronous mode, all of the data written to the primary by clients must also be passed across a network to the secondary system. So in addition to processing the data coming in from clients, the IBM System Storage N series CPU must do additional work to send the same data back out to the secondary system.

The same basic mechanism is at work in other CPU-intensive parts of the software in addition to networking. So in general, you can expect about double the CPU usage on a system with synchronous or semi-synchronous SnapMirror as compared with the same workload on a system without SnapMirror. You can use the **stats** command to display statistics on your CPU (Example 13-1).

Example 13-1 The stats command

```
itsotuc2*> stats show processor
processor:processor0:processor_busy:1%
processor:processor1:processor_busy:1%

itsotuc2*> stats show system
system:system:nfs_ops:0/s
system:system:cifs_ops:0/s
system:system:http_ops:0/s
system:system:dafs_ops:0/s
system:system:fcp_ops:0/s
system:system:iscsi_ops:0/s
system:system:net_data_recv:1KB/s
system:system:net_data_sent:0KB/s
system:system:disk_data_read:0KB/s
system:system:disk_data_written:8KB/s
system:system:cpu_busy:1%
system:system:avg_processor_busy:0%
system:system:total_processor_busy:1%
system:system:num_processors:2
```

13.13 Network bandwidth considerations

Because all of the data written to the primary storage must be replicated to the secondary storage as it is written, write throughput to the primary storage cannot generally exceed the bandwidth available between the primary and secondary storage devices. Because SnapMirror transfers can be performed over standard Ethernet networks and over Fibre Channel networks, there is a choice for transport. This choice will most likely be determined by preference or existing infrastructure rather than by performance needs.

In general, the configuration guideline is to configure the network between the primary and secondary storage with at least as much bandwidth as the network between the clients and the primary storage.

13.14 Replication considerations

Table 13-1 outlines the different maximum concurrent replication operations for different flavors of SnapMirror operations. Values are provided for Data ONTAP 8.1 with the Nearstore feature enabled. Without the Nearstore feature enabled, values would only be 50% from the stated values. Starting with Data ONTAP 8.1, the Nearstore feature is enabled by default.

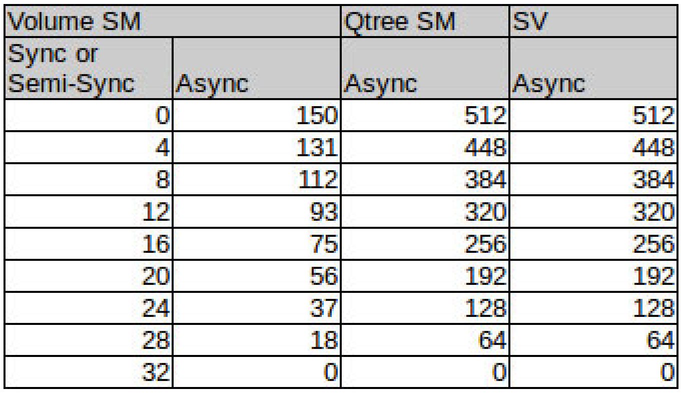
Table 13-1 Concurrent replication operations with Nearstore feature enabled

Storage system model	Async Volume SnapMirror	Sync or Semi-Sync Volume SnapMirror	Async Qtree SnapMirror	SnapVault	Open systems SnapVault
N3150	50/100	16	120/60	120	64
N3220					
N3240					
N6220	50/100	16	320/160	320	128
N6250					
N7550T	150/300[a]	32	512/256[b]	512	128
N7950T					

a. Source/Target
b. Single-path/Multi-path

Note that the system resources for replication all come from a shared pool. For example, an N7950T can support either 150/300 asynchronous Volume SnapMirror streams or 32 synchronous Volume SnapMirror streams; but not both at the same time.

Next we show the approximate number of replication streams available when used in combination on an N7950T. When maximum number of replication streams are in use, then there is no capacity for any more. A similar sharing of resources occurs on the other N series model. See Figure 13-17.



Volume SM		Qtree SM	SV
Sync or Semi-Sync	Async	Async	Async
0	150	512	512
4	131	448	448
8	112	384	384
12	93	320	320
16	75	256	256
20	56	192	192
24	37	128	128
28	18	64	64
32	0	0	0

Figure 13-17 Replication streams available

13.14.1 Maximum concurrent transfers for clustered configurations

In a cluster configuration, each node can have its own set of replication operations that are limited by the maximum number of concurrent transfers for that node.

In the event of a cluster takeover, replication operations from the node taken over are managed by the node that performed the takeover.

- If the combined number of concurrent transfers is less than the maximum allowed for the single node, all of the replication operations can run concurrently.
- If the combined number of concurrent transfers is greater than the maximum allowed for the single node, the replication operations for the node performing the takeover are run concurrently. Then, as a replication operation finishes, a replication operation from the taken over node replaces the finished operation.

Obviously this can present a problem if the total number of Synchronous replication streams in a cluster is greater than a single node can support in the event of a cluster takeover. We advise using a MetroCluster solution rather than using a large number of Synchronous SnapMirror relationships.

14

SnapLock

This chapter describes the N series SnapLock feature. It provides non-erasable and non-rewritable data protection that helps enable compliance with government and industry records retention regulations.

SnapLock configuration allows users to archive the data onto a permanent nonerasable, non-rewritable magnetic media while taking advantage of IBM technology for archival, backup, and disaster recovery.

There are two versions of SnapLock available:

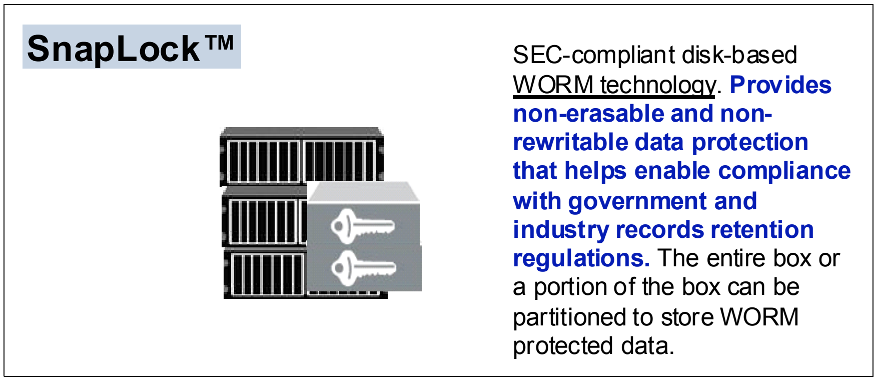
- SnapLock Compliance
- SnapLock Enterprise

The following topics are covered:

- SnapLock at a glance
- Introduction to SnapLock
- SnapLock setup

14.1 SnapLock at a glance

SnapLock, as shown in Figure 14-1, is a feature of Data ONTAP that implements high-performance, disk-based magnetic WORM storage. The primary objective of this Data ONTAP feature is to provide secure and storage-enforced data retention. SnapLock is a flexible and scalable solution that is supported on all IBM N series storage platforms. It is an open solution that uses standard protocols to enable seamless integration with ISV archiving applications as well as custom applications.



SnapLock™

SEC-compliant disk-based WORM technology. **Provides non-erasable and non-rewritable data protection that helps enable compliance with government and industry records retention regulations.** The entire box or a portion of the box can be partitioned to store WORM protected data.

Figure 14-1 SnapLock overview

SnapLock Compliance is designed to enable compliance in strictly regulated environments, such as those governed by SEC 240.17a-4. SnapLock Compliance provides an “untrusted storage administrator” model of operation in which the write once, read many (WORM) data, when committed to the storage system, is protected even from the storage administrator.

SnapLock Enterprise is designed for more flexible customer environments and operates under a “trusted storage administrator” model in which Data ONTAP permits some administrative actions while still enforcing WORM protection.

Attention: The SnapLock feature is only available in Data ONTAP 8.1. It is *not* available in Data ONTAP 8.0.x.

14.2 Introduction to SnapLock

SnapLock software products provide non-erasable, non-rewritable WORM functionality. SnapLock Compliance and SnapLock Enterprise are implemented as add-on licenses to Data ONTAP and the functionality of either is based on which license is active when the volume is created.

SnapLock is available in two versions:

- SnapLock Compliance
- SnapLock Enterprise

Both versions of SnapLock work exactly the same as each other, except with one noted exception explained in the next paragraph on SnapLock Enterprise. But, even though the products interoperate the same from an application point of view, the products are intended to accomplish different goals in terms of compliant archival versus long term protection of digital assets.

Restriction: SnapLock is not supported in all Data ONTAP versions. See the Interoperability matrixes and Release Notes of N series Data ONTAP. (For example, SnapLock was not supported with Data ONTAP 7.3.0 and 8.0, but in Data ONTAP 8.1.)

It is not possible to convert volumes directly (in place). If the data must be locked, it can be copied to a SnapLock volume. The simplest way to accomplish it is to use the **`vol copy`** command to an equivalently sized or larger SnapLock Enterprise or SnapLock Compliance volume. The destination compliance volume must be empty when **`vol copy`** is issued. Note that **`vol copy`** does not cause the files to be committed to WORM. When the data resides on the SnapLock volume:.

- A retention period must be set for the files.
- The files must be committed to WORM.

The retention period can either be set individually on each file by changing the *atime* (last access time) on the file or by using volume defaults. The commit to WORM can be accomplished either by committing individual files to WORM by removing the write permissions on the files or by using the *autocommit* feature of SnapLock, which automatically locks the files placed on the volume after the designated time period. A script or application can be used to accomplish these tasks.

Attention: SnapLock requires at least one file protocol such as CIFS or NFS in a SnapLock environment.

The unique properties and behavior of SnapLock volumes make them unsuitable for use as regular data storage volumes. In most ways, SnapLock volumes behave identically to regular volumes, but there are some very specific and critical differences in functionality and administration that make them unsuitable for use as regular volumes. For example:

- Renaming directories on SnapLock volumes is not allowed.
- Transition of file attributes from writable to read-only commits a file to WORM state.
- SnapLock volumes have the **`no_atime_update`** volume option set and explicit updates to atime are interpreted as changes to the retention period of the file.
- Administrative actions are restricted on SnapLock volumes.

Restriction: A SnapLock volume cannot be used as a regular volume. For example, the root volume cannot be the same as the one used for WORM purposes. Plan and size the environment wisely and provide enough drives and volumes for SnapLock, as well as regular volumes.

There are no restrictions on creation, deletion, or number of Snapshot copies on SnapLock volumes beyond the normal system limits because there is no risk of data loss with these operations. However, SnapRestore cannot be used to restore to a previous Snapshot copy on a SnapLock Compliance volume because it violates the immutability of compliance data. It is allowed on a SnapLock Enterprise volume because the storage administrator is trusted.

Attention: The aggregate that contains the SnapLock volume must be created with the attribute:

- To create a SnapLock aggregate, specify the `-L` flag with the **`aggr create`** command. This flag is only supported if either SnapLock Compliance or SnapLock Enterprise is licensed.
- The type of the SnapLock aggregate created, either Compliance or Enterprise, is determined by the installed SnapLock license. If both SnapLock Compliance and SnapLock Enterprise are licensed, use **`-L compliance`** or **`-L enterprise`** to specify the desired aggregate type.

14.2.1 SnapLock Compliance

SnapLock Compliance was designed to assist organizations in implementing a comprehensive archival solution for meeting strict regulatory requirements for data retention such as SEC 17 a-4. Records and files committed to WORM storage on a SnapLock Compliance volume cannot ever be altered or modified but can be deleted after the expiration of their retention periods. Moreover, a SnapLock Compliance volume cannot be deleted until all data stored on it has passed its retention period and been deleted by the archival application or some other process.

14.2.2 SnapLock Enterprise

SnapLock Enterprise is geared towards assisting organizations with meeting self-regulated and best-practice guidelines for protecting digital assets with WORM-type data storage. Data stored as WORM on a SnapLock Enterprise volume is permanently protected from alteration or modification but can be deleted after the expiration date. Functionality wise, SnapLock Enterprise matches SnapLock Compliance exactly with only one main difference: As the data being stored is not for the strictest regulatory applications, an administrator is trusted with the ability to delete a SnapLock Enterprise volume, including the data it contains.

Table 14-1 shows a comparison of SnapLock Compliance and SnapLock Enterprise.

Table 14-1 Comparison of Snaplock Compliance and SnapLock Enterprise

SnapLock Compliance	SnapLock Enterprise
“Strict” SnapLock ▸ Trust nobody	“Flexible” SnapLock ▸ Trust administrator
Permanently nonerasable, nonrewritable disk storage (WORM) ▸ Until file expiration ▸ Safe from any keyboard attack	Revision-safe, long-term storage solution ▸ Virus and application bug-proof ▸ Enables preferred practices business records retention
Complies w/ SEC regulations ▸ Meets SEC 17a-4 requirements ▸ Easy WORM-to-WORM replication	Partial storage admin control ▸ Admin can destroy volume to reclaim space ▸ Cannot modify/delete individual records

14.3 SnapLock setup

SnapLock is easy to integrate with because it allows the use of standard open protocols (NFS and CIFS) to set and manage the WORM data. It does this by using the atime (last access time stamp) file attribute to represent the retention period for the file. It also uses the removal of write access on the file to trigger the commit to WORM. Typically applications do the following:

- Select the files that must be retained for a certain time period (it is typically dictated by the governing regulations).
- Select the retention period (this too is typically dictated by regulations). The retention period can be set on a file basis (allowing file-level granularity); or volume-level defaults can be used to set the retention period on files that do not specify a retention period and that reside on the volume.
- Committing the files to WORM status. It can either be done at an individual file level (by removing the write permissions on the file) or by using the *autocommit* feature to automatically commit to WORM files that have not changed for a specified period of time.
- When the retention period has expired (that is, the value of *ComplianceClock* has surpassed the value of the atime), those files can then be deleted.

Many of these tasks can be automated.

Tip: SnapLock never automatically deletes any files from the volume. It is the responsibility of the applications to delete expired records.

SnapLock must be enabled on the N series as shown in Example 14-1:

Example 14-1 Enable the SnapLock license

```
tsosj-n01> license add <SnapLock license code
```

The ComplianceClock runs separately of the regular system clock. After being set, it cannot be changed. Be sure that the date and time settings are correct. See Example 14-2 regarding how to initialize the ComplianceClock.

Attention: Initializing the ComplianceClock has the following precautions:

- It is permanent; after being set, it cannot be changed or turned off.
- You CANNOT REVERT after the ComplianceClock has been initialized.
- The default and maximum retention period is 30 years.

Example 14-2 Initialize ComplianceClock:

```
itsosj-n01> date –c initialize

*** WARNING: YOU ARE INITIALIZING THE SECURE COMPLIANCE CLOCK ***
You are about to initialize the secure Compliance Clock of this system to the
current value of the system clock. This procedure can be performed ONLY ONCE on
this system so you should ensure that the system time is set correctly before
proceeding.

The current local system time is: Tue Jun 12 10:17:46 PST 2012
```

```
Is the current local system time correct? y
Are you REALLY sure you want initialize the Compliance Clock? y
Compliance Clock: Tue Jun 12 10:17:52 PST 2012
```

SnapLock volumes can only be created on SnapLock aggregates. Before a SnapLock Volume can be created, be sure you have created a Snaplock aggregate as shown in Example 14-3.

Example 14-3 Creating an SnapLock Aggregate

```
itsosj-n01> aggr create slc_aggr -L 3

WARNING: You have requested the creation of a new SnapLock Compliance aggregate.
SnapLock Compliance aggregates CANNOT be destroyed until all WORM content they
contain has expired.

Are you sure you want to create this SnapLock Compliance aggregate? y
Creation of an aggregate with 3 disks has completed.
```

After creating an aggregate, you can check the status, as shown in Example 14-4.

Example 14-4 Check status of created SnapLock aggregates

```
itsosj-n01> aggr status
Aggr State Status Options
slc_aggr online raid_dp, aggr snaplock_compliance
vol0 online raid4, root
```

Create a SnapLock volume with the **`vol create`** command, as shown in Example 14-5.

Example 14-5 Create SnapLock volume

```
itsosj-n01> vol create vol_slc_02 -L slc_aggr
```

Set up appropriate maximum, minimum, and default retention periods, as shown in Example 14-6.

Example 14-6 Set SnapLock retention periods

```
itsosj-n01> vol options vol_slc_02 snaplock_maximum_period 30y
itsosj-n01> vol options vol_slc_02 snaplock_minimum_period 30d
itsosj-n01> vol options vol_slc_02 snaplock_default_period 30d
```

Create a qtree in the SnapLock volume, as shown in Example 14-7.

Example 14-7 Create SnapLock qtree

```
itsosj-n01> qtree create /vol/vol_slc_02/locked
```

Then create a share. In our example it is a CIFS share, as you can see in Example 14-8.

Example 14-8 Create Share for SnapLock volume

```
itsosj-n01> cifs shares –add locked /vol/vol_slc_02/locked
```

If you want to check SnapLock volumes, use the **vol status** command, which can be used to identify SnapLock volumes. SnapLock volumes have an entry in the options field that identifies a volume as either **snaplock_compliance** or **snaplock_enterprise**.

To check the expiration date of a SnapLock volume, use the **-w** flag with the **vol status** command, which was introduced with Data ONTAP 7.3.1. The expiration date is the maximum retention time of WORM files on a SnapLock volume. SnapLock Compliance volumes cannot be deleted until the expiration date, as measured by the ComplianceClock, has passed.

To query the ComplianceClock, after being initialized, proceed as shown in Example 14-9.

Example 14-9 Query the ComplianceClock (initialized)

```
itsosj-n01> date –c
Compliance Clock: Tue Jun 12 10:17:52 PST 2012
```

If the ComplianceClock has never been initialized, the system will provide a different message, as shown in Example 14-10.

Example 14-10 Query the ComplianceClock (uninitialized)

```
itsosj-n01> date –c
Error: Compliance Clock has not been initialized.
```

15

SyncMirror

SyncMirror is designed to keep data available and up-to-date by maintaining two copies of data online. SyncMirror provides a synchronous replication, duplicate copy, or redundant copy of data within the same N series system. SyncMirror maintains a strict physical separation between the two copies. If an error occurs in one copy, the data is still accessible without any manual intervention at the RAID level.

The IBM N series System Storage also has the ability of implementing SyncMirror with Clustered Failover (now called HA pair in Data ONTAP 8.1 7-mode), SyncMirror provides a strict physical separation between two copies of your mirrored data, safeguarding your data in the event of an outage. The higher levels of data availability in SyncMirror's design allows users access to the mirrored data without operator intervention or disruption to applications.

This chapter provides an explanation and demonstration of SyncMirror. We describe SyncMirror in local and disaster recovery configurations, as well as explaining the difference between SnapMirror and SyncMirror.

The following topics are covered:

- Background
- Differences between SnapMirror and SyncMirror
- Implementing local SyncMirror
- How SyncMirror works with third-party storage
- Disaster recovery with SyncMirror

15.1 Background

By operating at the storage level instead of at the server or application level, IBM System Storage N series business continuance solutions ensure protection while off loading tasks from busy servers.

- SyncMirror maintains the following features:
  - Two copies of client's data online
  - Strict physical separation between copies of mirrored data
  - Ability to split the data copies with a simple command
  - Ability to operate with simple and consistent interfaces
  - Two plexes directly connected to same system
- SyncMirror can be used to mirror aggregates and traditional volumes (a traditional volume is an aggregate with a single volume that spans the entire aggregate).
- SyncMirror cannot be used to mirror FlexVol volumes, but FlexVol volumes can be mirrored as part of an aggregate.

IBM System Storage N series solutions rationalize business continuance strategies, simplify management, and greatly improve recovery times and reduce expensive downtime, protecting against lost revenue and damaged reputation. At the same time, their simplicity produces significant cost savings in the deployment and ongoing operation of a business continuance strategy. Business continuance is accomplished with a combination of SnapVault, SnapMirror, and SyncMirror (Figure 15-1).

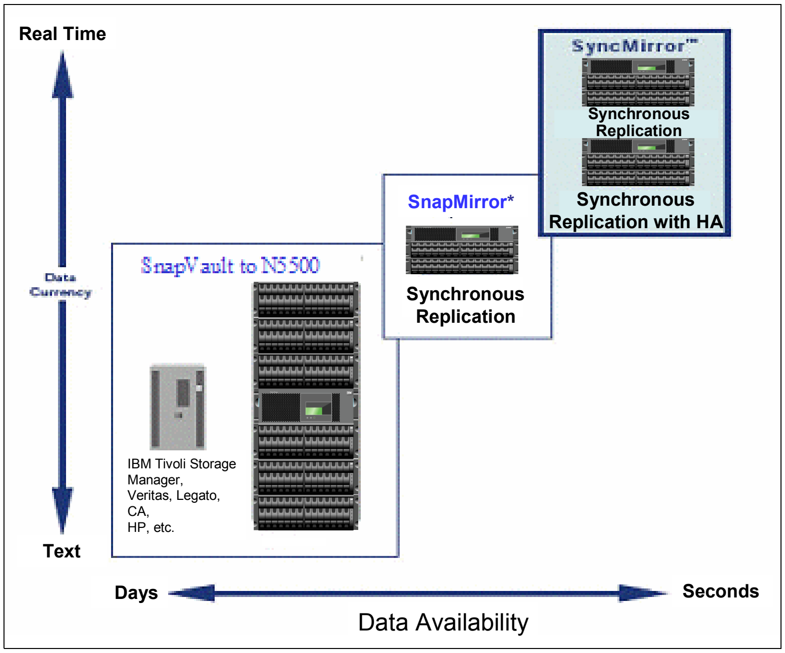
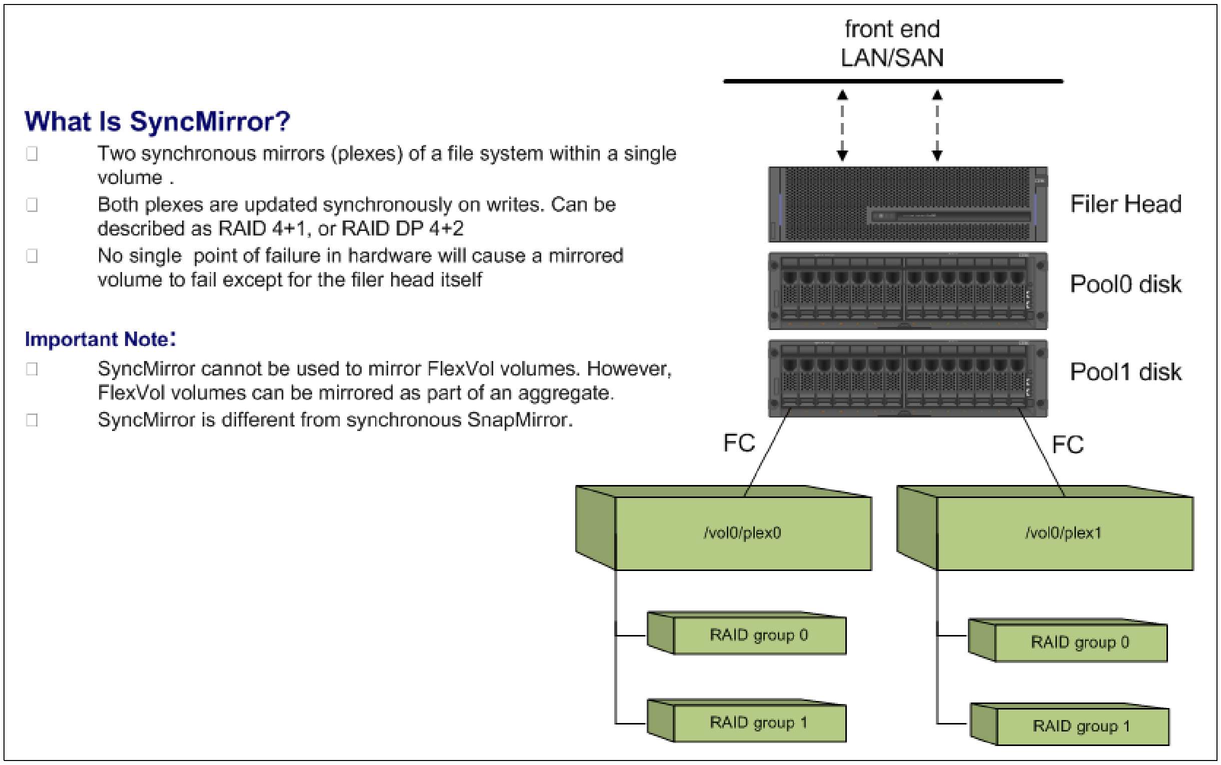


Figure 15-1 Business continuity with SyncMirror

The SyncMirror software creates aggregates or traditional volumes that consist of two copies of the same WAFL file system. The two copies, known as plexes, are synchronously updated (see Figure 15-2 on page 209 and Example 15-1 on page 210). As a result, the copies are always identical. Data ONTAP typically names the first plex *plex0*, and the second plex *plex1*.

Each plex is a physical copy of the same WAFL file system, and consists of one or more RAID groups. Because SyncMirror duplicates complete WAFL file systems, you cannot use the SyncMirror feature with a FlexVol volume; only aggregates are supported.



What Is SyncMirror?

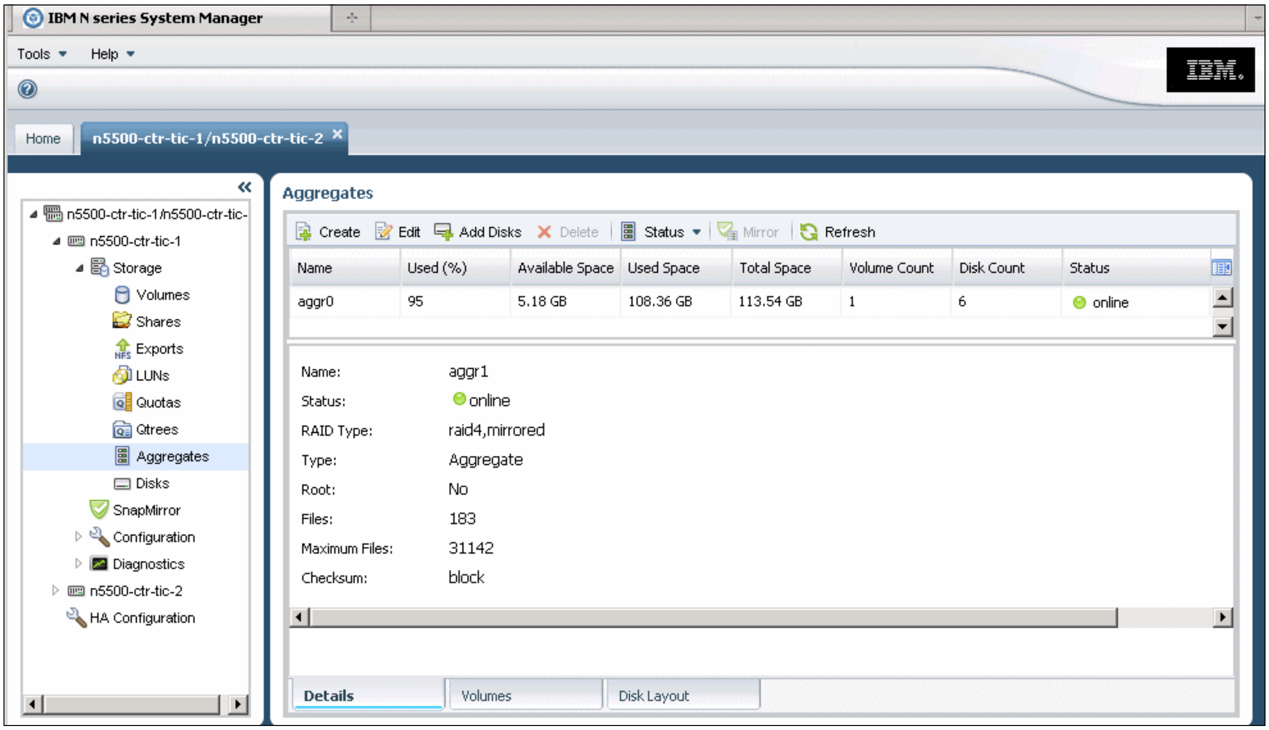
- Two synchronous mirrors (plexes) of a file system within a single volume .
- Both plexes are updated synchronously on writes. Can be described as RAID 4+1, or RAID DP 4+2
- No single point of failure in hardware will cause a mirrored volume to fail except for the filer head itself

Important Note:

- SyncMirror cannot be used to mirror FlexVol volumes. However, FlexVol volumes can be mirrored as part of an aggregate.
- SyncMirror is different from synchronous SnapMirror.

Figure 15-2 Synchronous mirroring

Figure 15-3 shows an mirrored aggregate using N series System Manager 2.



Name	Used (%)	Available Space	Used Space	Total Space	Volume Count	Disk Count	Status
aggr0	95	5.18 GB	108.36 GB	113.54 GB	1	6	online

Name: aggr1
Status: online
RAID Type: raid4,mirrored
Type: Aggregate
Root: No
Files: 183
Maximum Files: 31142
Checksum: block

Figure 15-3 System Manager aggregate view

Example 15-1 N series plexes status in local SyncMirror

```
itsotuc1> aggr status -v
           Aggr State           Status            Options
          aggr0 online          raid_dp, aggr     root, diskroot, nosnap=off,
                                mirrored          raidtype=raid_dp, raidsize=16,
                                32-bit            ignore_inconsistent=off,
                                                  snapmirrored=off,
                                                  resyncsnaptime=60,
                                                  fs_size_fixed=off,
                                                  snapshot_autodelete=on,
                                                  lost_write_protect=on,
                                                  ha_policy=cfo

                Volumes: vol0, Source_Volume

                Plex /aggr0/plex0: online, normal, active
                    RAID group /aggr0/plex0/rg0: normal

                Plex /aggr0/plex2: online, normal, active
                    RAID group /aggr0/plex2/rg0: normal
```

The additional resiliency features that SyncMirror offers over SnapMirror in synchronous mode provide protection against system downtime due to shelf failure, triple disk failure for RAID-DP RAID groups, or Fibre Channel loop failure, as explained here:

- When used in HA pair configurations, SyncMirror provides the highest resiliency levels in IBM System Storage N series storage for a local data center.
- The highest levels of storage resiliency ensure continuous data availability within a data center.
- IBM advises using SyncMirror and HA pair configurations for high levels of storage resiliency. See 15.5, "Disaster recovery with SyncMirror" on page 227.

Figure 15-4 shows the replication of data using SyncMirror with a single filer, which creates a redundant storage subsystem with an up-to-date mirror to support high data availability.

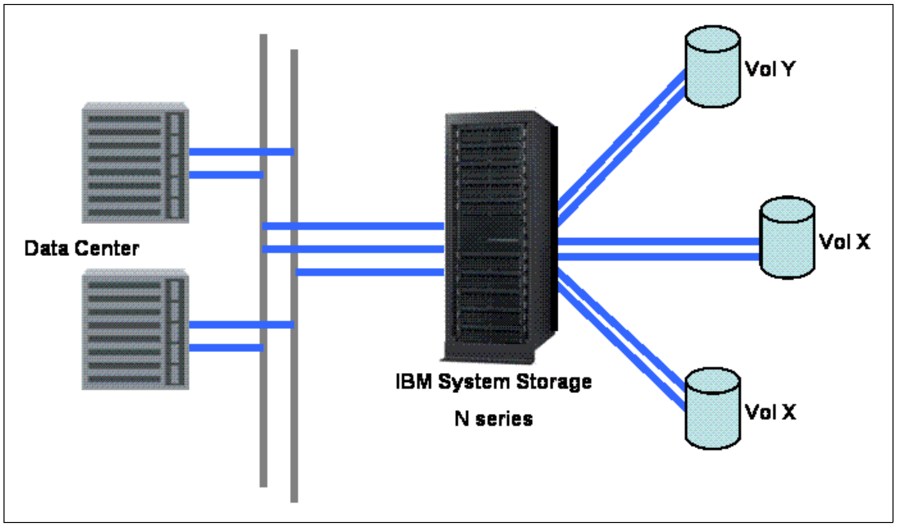


Figure 15-4 N series storage system with a SyncMirror as single system

Figure 15-5 shows the replication of data using SyncMirror with filers in an HA pair failover configuration, which supports high data availability, data redundancy, and automatic failover.

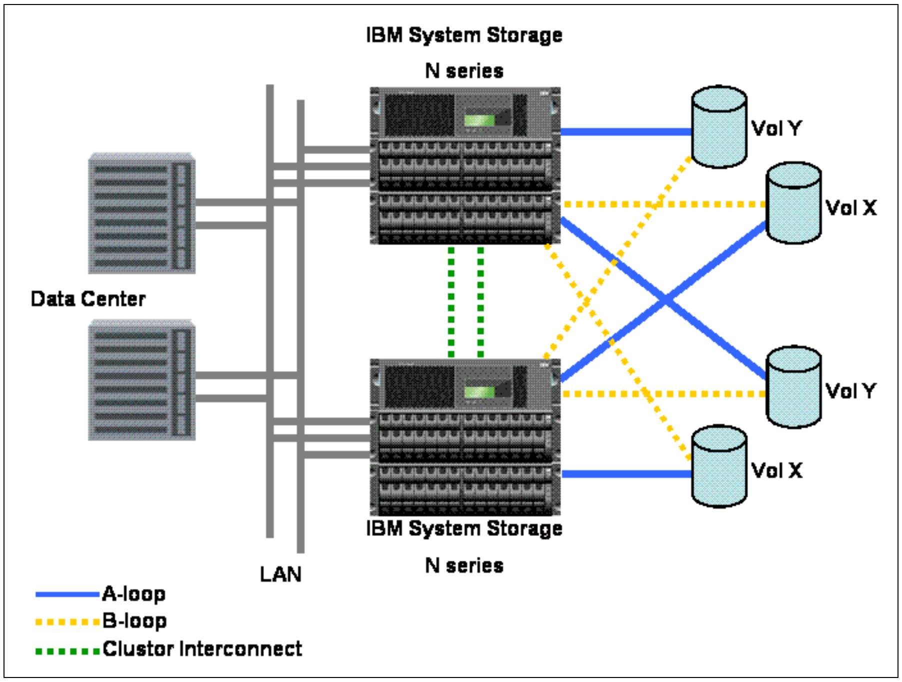


Figure 15-5 N series storage system with SyncMirror Clustered Failover HA pair setup

15.1.1 What is new in 8.2

Starting with Data ONTAP 8.2, you do not have to install the SyncMirror license, because it is enabled by default.

15.2 Differences between SnapMirror and SyncMirror

The differences between SyncMirror and synchronous SnapMirror involve ownership of the second copy of data:

- With SyncMirror, one host owns both plexes. It simply writes to both plexes from NVRAM. It also provides for instant failover in the event that one plex fails for any reason.
- SyncMirror using mirrored plexes results in quicker rebuild times.

When SyncMirror is enabled, disks are separated into two disk pools, and a copy of the plex is created. The physical separation of the plexes protects against data loss if one of the shelves or the storage array becomes unavailable. The unaffected plex continues to serve data while you fix the cause of the failure. After being fixed, the two plexes can be resynchronized.

Although SnapMirror in synchronous mode provides a similar capability, its generally intended use is replicating data between geographic locations to improve disaster recovery options in the event of a data center outage. At the local level, SyncMirror provides storage resiliency capabilities that SnapMirror or even H/A configurations by themselves do not.

With Synchronous SnapMirror, the secondary filer (IBM System Storage N series storage system) owns the replicated copy. It moves blocks similar to SnapMirror, where blocks move over IP to the other filer, rather than direct Fibre Channel writes to local disks. There is no automatic failover because the primary host does not own the second copy of the volume. SnapMirror is limited in functionality, not being able to fulfill automatic failover.

In contrast, if an aggregate using SnapMirror for replication becomes unavailable, you can use one of the following options to access the data on the SnapMirror destination (secondary):

- The SnapMirror destination cannot automatically take over the file serving functions. However, you can manually set the SnapMirror destination to allow read-write access to the data.
- You can restore the data from the SnapMirror destination to the primary (source) storage system.

15.3 Implementing local SyncMirror

Local SyncMirror provides synchronous mirroring between two different volumes or aggregates on the same storage controller so a duplicate copy of data exists, resulting in higher storage resiliency and data availability. Software disk ownership (SANOWN) or system configuration ownership is the mechanism to assign disks in the local SyncMirror configuration.

A SyncMirror aggregate consists of two plexes. The setup provides a high level of data availability, and consistency through RAID-level, block-level mirroring. SyncMirror enables the two plexes to simultaneously update insuring that the plexes are always identical.

Tip: A mirrored aggregate can have only two plexes.

15.3.1 Preliminary construction and considerations

It is important to evaluate preliminary actions and determine architecture considerations that will provide a successful implementation.

Plexes and aggregates

Here are some valuable considerations and ideas to help fulfill the objectives listed:

- Disks assigned to plexes:

  When assigning disks, you will need to understand how Data ONTAP assigns disks to plexes in order to configure your disk shelves and host adapters.
- Viewing plexes and spare pools:

  Ensure that you view all spare disk and LUN assignments, when adding additional disks or LUNs to an aggregate. It is also important to identify pools used by each plex.
- Creating a mirrored aggregate:

  In creating an aggregate, you can specify the aggregate to use SyncMirror. This ensures that the aggregate is a mirrored from disk initialization. The integrator has several options for how they want to specify the disks or LUNs in creation of the mirrored aggregate.
- Converting an aggregate to a mirrored aggregate:

  It is possible to convert an existing aggregate to a mirrored aggregate by adding a plex. You can use SyncMirror to mirror a previously unmirrored aggregate. The integrator has several options for how to specify the disks or LUNs when converting to a mirrored aggregate.

Disk selection policies when using mirrored aggregates

For N series systems using EXN expansion units, the two plexes must be on different shelves connected to the system with separate cables and adapters. Each plex has its own collection of spare disks. It is important to note that mixed speeds (RPMs) are supported in creating a successful mirrored aggregate selection.

An aggregate mirrored using SyncMirror requires twice as much storage as an unmirrored aggregate. Each of the two plexes require an independent set of disks or array LUNs. For example, you need 2,880 GB of disk space to mirror a 1,440 GB aggregate, that is, 1,440 GB for each plex of the mirrored aggregate.

Considerations:

- Data ONTAP names the plexes of the mirrored aggregate.
- The version of Data ONTAP requires a consistent relationship between filers and disk.
- You cannot set up SyncMirror with disks in one plex and array LUNs in the other plex.

Regardless of how the administrator initiated the creation of a mirrored aggregate, they will have an opportunity to choose *Automatic* (preferred), or *Manual* disk selections.

15.3.2 Implementation of SyncMirror

When implementing SyncMirror on an IBM N series filer, first enable the SyncMirror license and proceed with enabling the SyncMirror feature, prior to following the next examples.

Note: In Data ONTAP 8.2, the SyncMirror feature is enabled by default, and it is not necessary to enable it with a license key.

Assigning disks to plexes and pools

The rules for selection of disks or LUNs for using as a mirrored aggregate are as follows:

- Disks or array LUNs selected for each plex must be in different pools (Pool0 or Pool1).
- Equal amounts of disks or array LUNs are required in both plexes.
- When selecting disks, the first consideration needs to be on the basis of equivalent bytes per sector (bps) size, and secondly, on the size of the disk.
- If there is no equivalent-sized disk, Data ONTAP uses a larger-capacity disk, and limits the size to make it identical.
- Data ONTAP names the plexes of the mirrored aggregate.

Tip: When creating an aggregate, Data ONTAP selects disks from the plex which has the most available disks. You can override this selection policy by specifying the disks to use.

Creating pools

We recently added disks in Enclosure 0 slots 7, 8, 9, 11, and 12 (0c.23, 0a.24, 0c.25, 0c.27, and 0c.28) as listed in Example 15-2. We use a total of 10 disks between both pool0 and pool1. There are five 300 GB Fibre Channel drives in Pool0, and five 300 GB Fibre Channel drives unassigned as shown in Figure 15-6. The disks require presentation into Pool1. Thus it creates an equal amount of disks required in separate pools, and keeps them independent.

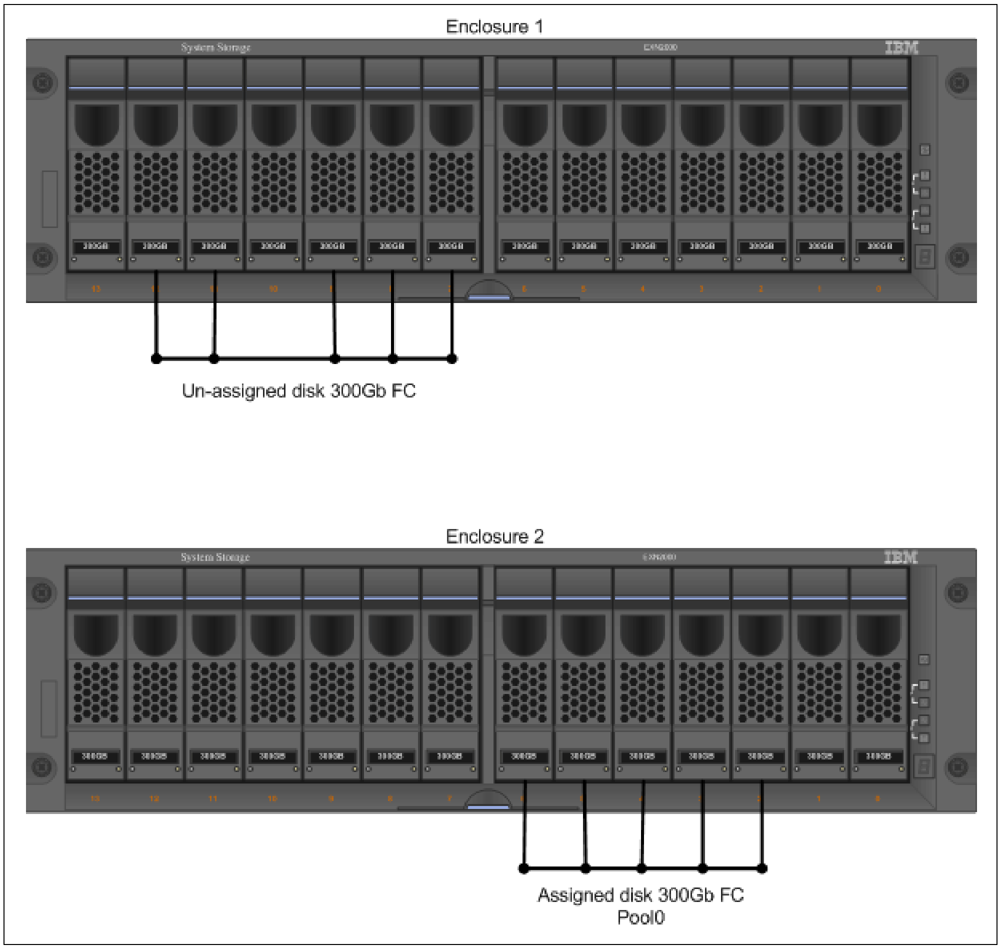


Figure 15-6   SyncMirror disk presentation and selection

Example 15-2   SyncMirror disk presentation and selection example

```
itsotuc1> disk show
  DISK       OWNER                      POOL   SERIAL NUMBER         HOME
------------ -------------              -----  -------------         -------------
0a.37        itsotuc1  (118052508)    Pool0  3KR14NZ3000076155TE0  itsotuc1  (118052508)
0c.32        itsotuc1  (118052508)    Pool0  3KR14Q1T00007615JGDE  itsotuc1  (118052508)
0c.33        itsotuc1  (118052508)    Pool0  3KR13HAP000076150YD3  itsotuc1  (118052508)
0c.35        itsotuc1  (118052508)    Pool0  3KR14PDT00007615XPCU  itsotuc1  (118052508)
0a.34        itsotuc1  (118052508)    Pool0  3KR149K700007615JL34  itsotuc1  (118052508)
0c.39        itsotuc1  (118052508)    Pool0  3KR14P6G00007615XN0D  itsotuc1  (118052508)
0c.38        itsotuc1  (118052508)    Pool0  3KR14QJZ000076114VYK  itsotuc1  (118052508)
0c.36        itsotuc1  (118052508)    Pool0  3KR158HQ00007615JE88  itsotuc1  (118052508)
0d.53        itsotuc1  (118052508)    Pool0  L59K3RKG              itsotuc1  (118052508)
0d.49        itsotuc1  (118052508)    Pool0  L50L3VYG              itsotuc1  (118052508)
0d.52        itsotuc1  (118052508)    Pool0  L59K6M1G              itsotuc1  (118052508)
0d.48        itsotuc1  (118052508)    Pool0  L50LJ2XG              itsotuc1  (118052508)
0d.50        itsotuc1  (118052508)    Pool0  L50QV25G              itsotuc1  (118052508)
0d.54        itsotuc1  (118052508)    Pool0  L59MP0XG              itsotuc1  (118052508)
0d.55        itsotuc1  (118052508)    Pool0  L59Q2H0G              itsotuc1  (118052508)

NOTE: Currently 13 disks are unowned. Use 'disk show -n' for additional
information.
```

```
itsotuc1> disk show -n
  DISK       OWNER                      POOL   SERIAL NUMBER         HOME
------------ -------------              -----  -------------         -------------
0a.26        Not Owned                  NONE   3KR14Q9P000076150Z1L
0c.20        Not Owned                  NONE   3KP2A73J00007629HUV5
0c.16        Not Owned                  NONE   3KP2BHVC000076296M05
0c.19        Not Owned                  NONE   3KP2BFPS00007630PWM6
0c.17        Not Owned                  NONE   3KP2BK5200007630PWL4
0a.18        Not Owned                  NONE   3KP2BHXZ000076296L73
0a.24        Not Owned                  NONE   3KR14PYK000076144JUJ
0c.25        Not Owned                  NONE   3KR14Q8A00007615JELN
0c.23        Not Owned                  NONE   3KR14NYL00007615DT0A
0c.27        Not Owned                  NONE   3KR158HN00007615VXBC
0c.21        Not Owned                  NONE   3KP30NY2000097363AYB
0c.28        Not Owned                  NONE   3KR14P3A000076150Z14
0c.22        Not Owned                  NONE   3KP2BGG600007630GMYA
```

Proceeding, we now assign disks 0c.23, 0a.24, 0c.25, 0c.27, 0c.28 for this example to Pool1 as depicted in Figure 15-7, and FilerView shown in Example 15-3. This ensures that we have an equal distribution of disks required in configuring a successful SyncMirror relationship.

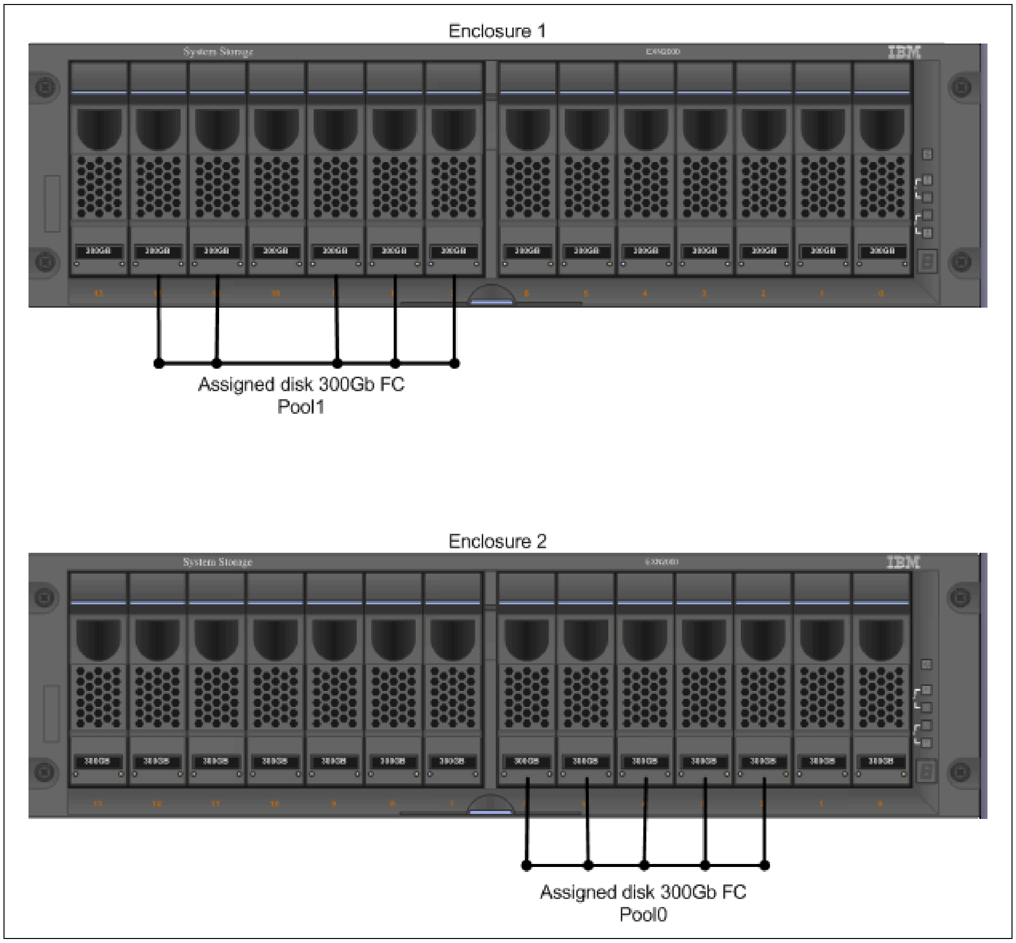
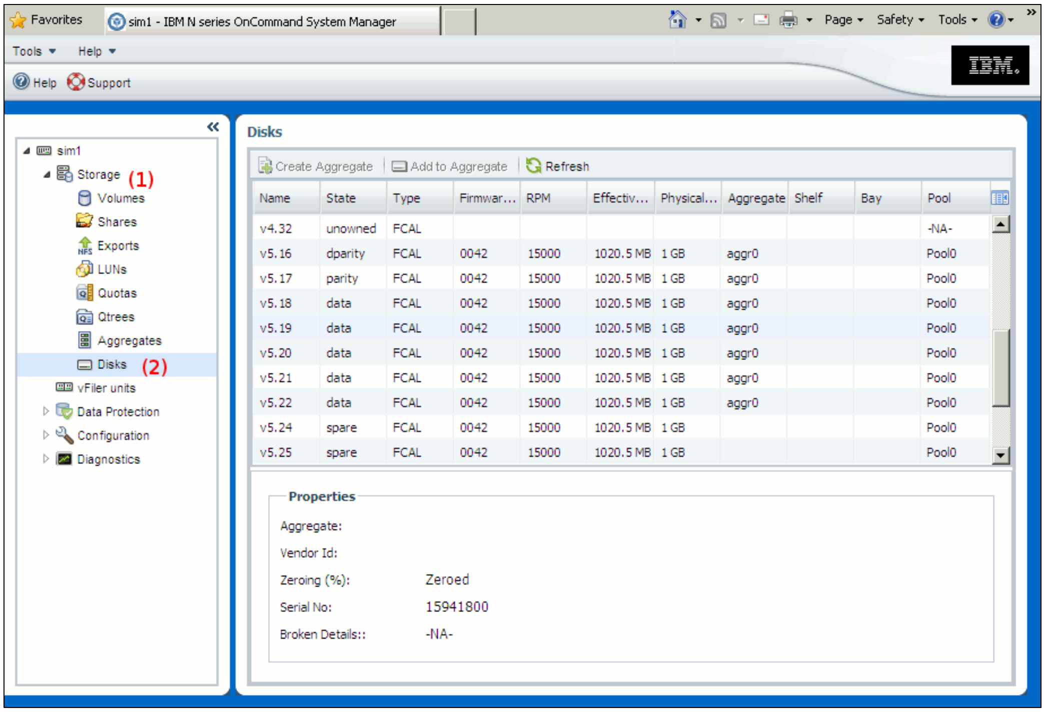


Figure 15-7 Pool1 creation for SyncMirror relationship

Example 15-3 Command disk assigned for Pool1 SyncMirror configuration

```
itsotuc1> disk assign 0c.28 0c.27 0c.25 0c.23 0a.24 -p 1 -f

Wed Apr 13 22:09:18 GMT [diskown.changingOwner:info]: changing ownership for disk 0c.28
(S/N 3KR14P3A000076150Z14) from unowned (ID 4294967295) to itsotuc1 (ID 118052508)
Wed Apr 13 22:09:18 GMT [diskown.changingOwner:info]: changing ownership for disk 0c.27
(S/N 3KR158HN00007615VXBC) from unowned (ID 4294967295) to itsotuc1 (ID 118052508)
Wed Apr 13 22:09:18 GMT [diskown.changingOwner:info]: changing ownership for disk 0c.25
(S/N 3KR14Q8A00007615JELN) from unowned (ID 4294967295) to itsotuc1 (ID 118052508)
Wed Apr 13 22:09:18 GMT [diskown.changingOwner:info]: changing ownership for disk 0c.23
(S/N 3KR14NYL00007615DT0A) from unowned (ID 4294967295) to itsotuc1 (ID 118052508)
Wed Apr 13 22:09:18 GMT [diskown.changingOwner:info]: changing ownership for disk 0a.24
(S/N 3KR14PYK000076144JUJ) from unowned (ID 4294967295) to itsotuc1 (ID 118052508)
Wed Apr 13 22:09:19 GMT [sfu.firmwareUpToDate:info]: Firmware is up-to-date on all disk
shelves.
disk show
  DISK        OWNER                         POOL   SERIAL NUMBER            HOME
------------ -------------                  -----  -------------            -------------
0c.28        itsotuc1  (118052508)    Pool1  3KR14P3A000076150Z14  itsotuc1  (118052508)
0a.37        itsotuc1  (118052508)    Pool0  3KR14NZ3000076155TE0  itsotuc1  (118052508)
0c.23        itsotuc1  (118052508)    Pool1  3KR14NYL00007615DT0A  itsotuc1  (118052508)
0c.32        itsotuc1  (118052508)    Pool0  3KR14Q1T00007615JGDE  itsotuc1  (118052508)
0c.33        itsotuc1  (118052508)    Pool0  3KR13HAP000076150YD3  itsotuc1  (118052508)
0c.35        itsotuc1  (118052508)    Pool0  3KR14PDT00007615XPCU  itsotuc1  (118052508)
0a.34        itsotuc1  (118052508)    Pool0  3KR149K700007615JL34  itsotuc1  (118052508)
0c.39        itsotuc1  (118052508)    Pool0  3KR14P6G00007615XN0D  itsotuc1  (118052508)
0c.38        itsotuc1  (118052508)    Pool0  3KR14QJZ000076114VYK  itsotuc1  (118052508)
0c.36        itsotuc1  (118052508)    Pool0  3KR158HQ00007615JE88  itsotuc1  (118052508)
0c.27        itsotuc1  (118052508)    Pool1  3KR158HN00007615VXBC  itsotuc1  (118052508)
0a.24        itsotuc1  (118052508)    Pool1  3KR14PYK000076144JUJ  itsotuc1  (118052508)
0c.25        itsotuc1  (118052508)    Pool1  3KR14Q8A00007615JELN  itsotuc1  (118052508)
0d.53        itsotuc1  (118052508)    Pool0  L59K3RKG              itsotuc1  (118052508)
0d.49        itsotuc1  (118052508)    Pool0  L50L3VYG              itsotuc1  (118052508)
0d.52        itsotuc1  (118052508)    Pool0  L59K6M1G              itsotuc1  (118052508)
0d.48        itsotuc1  (118052508)    Pool0  L50LJ2XG              itsotuc1  (118052508)
0d.50        itsotuc1  (118052508)    Pool0  L50QV25G              itsotuc1  (118052508)
0d.54        itsotuc1  (118052508)    Pool0  L59MP0XG              itsotuc1  (118052508)
0d.55        itsotuc1  (118052508)    Pool0  L59Q2H0G              itsotuc1  (118052508)
```

The System Manager tool can also be used to review the disk assignments, as shown in Figure 15-8. (Note that the disks shown do not match the previous examples and are included for information only.)



Name	State	Type	Firmwar...	RPM	Effectiv...	Physical...	Aggregate	Shelf	Bay	Pool
v4.32	unowned	FCAL								-NA-
v5.16	dparity	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.17	parity	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.18	data	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.19	data	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.20	data	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.21	data	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.22	data	FCAL	0042	15000	1020.5 MB	1 GB	aggr0			Pool0
v5.24	spare	FCAL	0042	15000	1020.5 MB	1 GB				Pool0
v5.25	spare	FCAL	0042	15000	1020.5 MB	1 GB				Pool0

Properties

Aggregate:

Vendor Id:

Zeroing (%): Zeroed

Serial No: 15941800

Broken Details:: -NA-

Figure 15-8 System Manager view of disk assignments

Creating a SyncMirrored aggregate

We will configure an aggregate called *rodgrad_med_flode* as shown in Figure 15-10. We will enable Synchronous Mirroring and RAID-DP, in one of the following ways:

- We can do it via the Command Line:

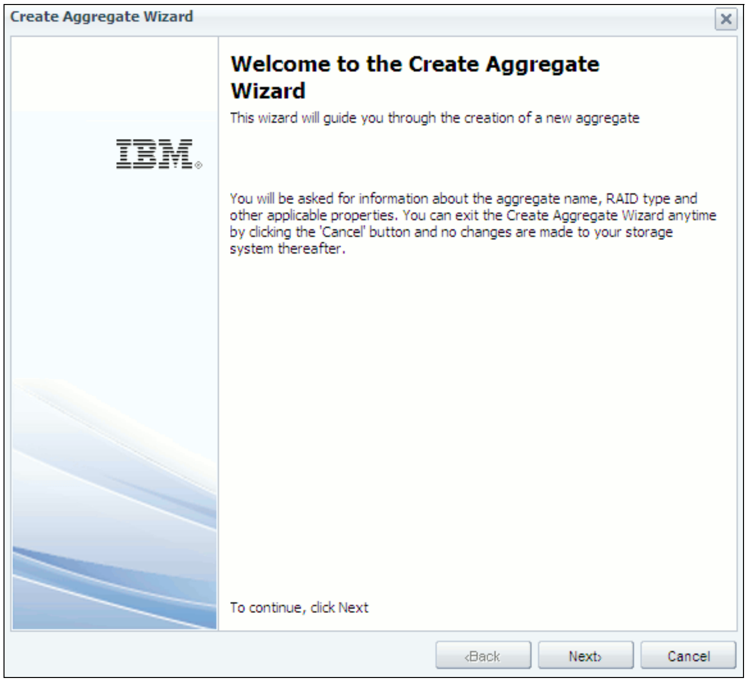
```
aggr create “aggr-name”

    [-f] [-L [compliance | enterprise]]
    [-B {32|64}]
    [-m] [-n] [-r <raid-group-size>] [-R <rpm>]
    [-T {ATA | BSAS | FCAL | LUN | SAS | SATA | SSD | XATA | XSAS}]
    [-t {raid4 | raid_dp}] [-v [-l <language-code>]] <disk-list>
```

If a mirrored aggregate is desired, make sure to specify an even number for <ndisks>, or to use two '-d' lists.

- Or, we can do it via System Manager:

  Expand **Storage** → **Aggregates**, and click **Create** to start the Create Aggregate Wizard, as shown in Figure 15-9.

  Click **Next**.



Create Aggregate Wizard

Welcome to the Create Aggregate Wizard

This wizard will guide you through the creation of a new aggregate

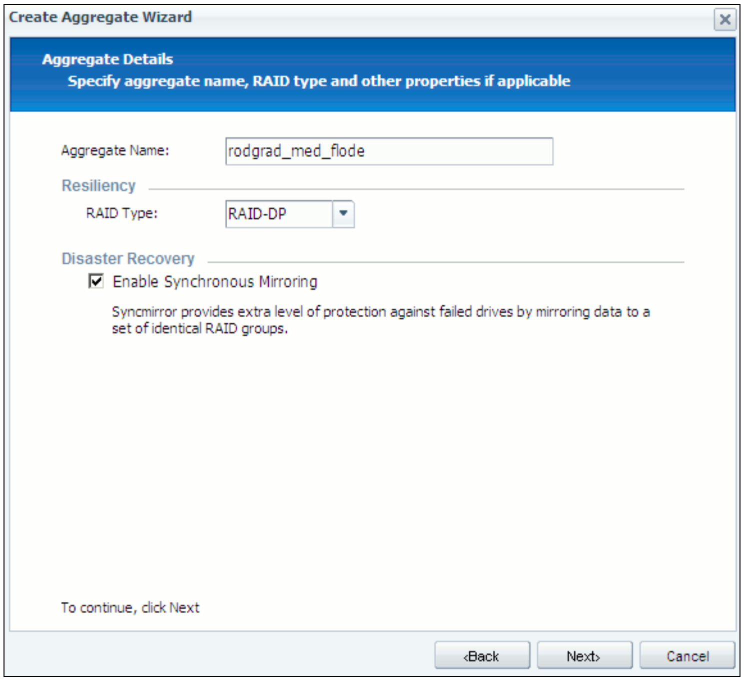
You will be asked for information about the aggregate name, RAID type and other applicable properties. You can exit the Create Aggregate Wizard anytime by clicking the 'Cancel' button and no changes are made to your storage system thereafter.

To continue, click Next

Figure 15-9 The Create Aggregate Wizard

Use the Wizard to enter the Aggregate name and RAID type, and enable Synchronous Mirroring, as shown in Figure 15-10.

Click **Next**.



Create Aggregate Wizard

Aggregate Details

Specify aggregate name, RAID type and other properties if applicable

Aggregate Name: rodgrad_med_flode

Resiliency

RAID Type: RAID-DP

Disaster Recovery

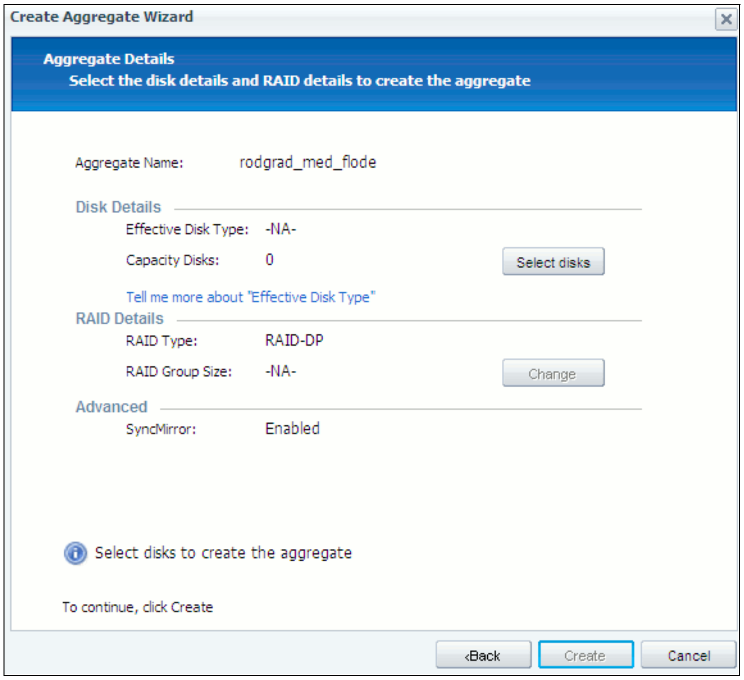
☑ Enable Synchronous Mirroring

Syncmirror provides extra level of protection against failed drives by mirroring data to a set of identical RAID groups.

To continue, click Next

Figure 15-10 Enable Synchronous Mirroring

In the *Disk Details* section, click **Select Disks** to select the number of disks to assign to the mirrored aggregate, as shown in Figure 15-11.


Create Aggregate Wizard
Aggregate Details
Select the disk details and RAID details to create the aggregate

Aggregate Name: rodgrad_med_flode

Disk Details

Effective Disk Type: -NA-

Capacity Disks: 0

Select disks
Tell me more about "Effective Disk Type"

RAID Details

RAID Type: RAID-DP

RAID Group Size: -NA-

SyncMirror: Enabled

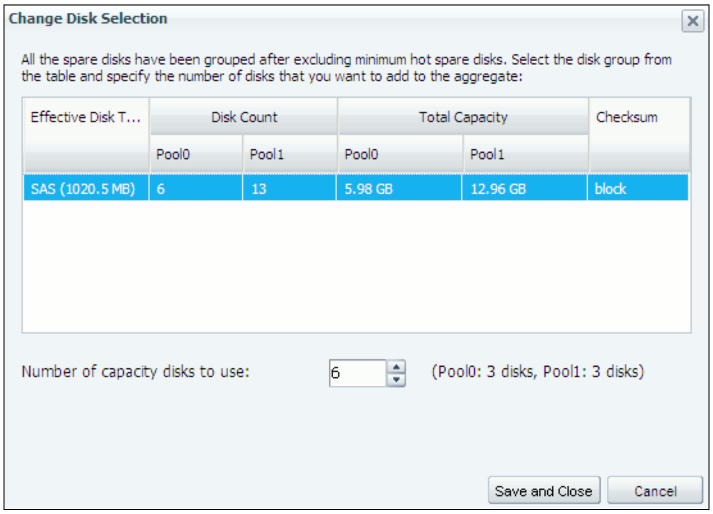
Select disks to create the aggregate

To continue, click Create

Figure 15-11 Select the number of disks to assign

Use the Change Disk Selection window to choose the number of disks to assign to the aggregate, as shown in Figure 15-12. Notice in the figure that selecting 6 disks will assign 3 disks from Pool0 and 3 disks from Pool1.

Click **Save and Close**.



All the spare disks have been grouped after excluding minimum hot spare disks. Select the disk group from the table and specify the number of disks that you want to add to the aggregate:

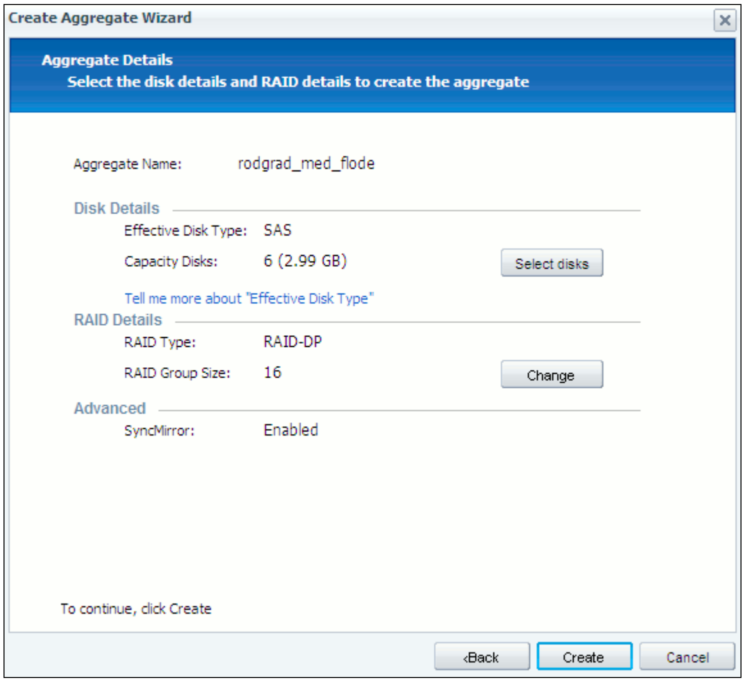
Effective Disk T...	Disk Count		Total Capacity		Checksum
	Pool0	Pool1	Pool0	Pool1	
SAS (1020.5 MB)	6	13	5.98 GB	12.96 GB	block

Number of capacity disks to use: 6 (Pool0: 3 disks, Pool1: 3 disks)

Figure 15-12 Choose the number of disks to assign

In the *RAID Details* section, click **Change** if you want to alter the maximum number of disks in each RAID group, as shown in Figure 15-13.

Then click **Create** to create the new mirrored Aggregate.



Create Aggregate Wizard

Aggregate Details

Select the disk details and RAID details to create the aggregate

Aggregate Name: rodgrad_med_flode

Disk Details

Effective Disk Type: SAS

Capacity Disks: 6 (2.99 GB)

Tell me more about "Effective Disk Type"

RAID Details

RAID Type: RAID-DP

RAID Group Size: 16

Advanced

SyncMirror: Enabled

To continue, click Create

Figure 15-13 Create the new mirrored Aggregate

After the Aggregate has been created successfully, click **Next** to exit the Wizard.

Disk type selection

While creating a new aggregate or traditional volume, you can select the available spare disks on a system connected to disks of different types. You can mix mutually compatible disks in one aggregate or traditional volume, when the option *raid.disktype.enable* is set to *off*.

If the option *raid.disktype.enable* is set to *on*, you will not be able to mix disk types. For example, if ATA, FCAL, SAS, and SATA are available as spare disks, then you can select one of them.

If the option *raid.disktype.enable* is set to *off*, you can mix disk types. For example, if ATA, FCAL, SAS, and SATA are available as spare disks, then you can select one of the following combinations:

- FCAL, SAS
- ATA, SATA

You can select one of the available spare disks, if they are not mutually compatible. For example, if you have only ATA and FCAL as available spare disks, which are not mutually compatible, you can choose one of the two.

When using mirrored aggregates, you can create a new aggregate with two mirrored plexes, or add a plex to an existing aggregate.

Considerations:

- To keep you from accidentally using the last spare, the maximum number of disks that you can select is the number of available disks minus one spare. To select all the available disks, use the command-line interface.
- On an IBM N series gateway, you can use the last spare.

When you create a new aggregate or traditional volume, you must have at least two disks available if you have selected RAID Level 4 protection. You must have at least three disks available if you are using double-parity RAID protection.

You can add additional disks later, but you cannot remove disks after they have been added.

You can also view the status of disk zeroing as shown in Example 15-4.

Example 15-4 SyncMirror initializing

```
Data ONTAP (itsotuc1.itso.com)
login: root
Password:
itsotuc1> aggr status
           Aggr State           Status            Options
rodgrad_med_flode creating        raid_dp, aggr     raidsize=3,
                                initializing      snapshot_autodelete=off,
                                mirrored          lost_write_protect=off
                                32-bit
          aggr0 online          raid_dp, aggr     root
                                32-bit
itsotuc1> aggr status -v
           Aggr State           Status            Options
rodgrad_med_flode creating        raid_dp, aggr     nosnap=off, raidtype=raid_dp,
                                initializing      raidsize=3,
                                mirrored          ignore_inconsistent=off,
                                32-bit            snapmirrored=off,
                                                  resyncsnaptime=60,
                                                  fs_size_fixed=off,
                                                  snapshot_autodelete=off,
                                                  lost_write_protect=off,
                                                  ha_policy=cfo
                Volumes: <none>

                Plex /rodgrad_med_flode/plex0: offline, empty, active

                Plex /rodgrad_med_flode/plex1: offline, empty, active

          aggr0 online          raid_dp, aggr     root, diskroot, nosnap=off,
                                32-bit            raidtype=raid_dp, raidsize=16,
                                                  ignore_inconsistent=off,
                                                  snapmirrored=off,
                                                  resyncsnaptime=60,
                                                  fs_size_fixed=off,
                                                  snapshot_autodelete=on,
                                                  lost_write_protect=on,
                                                  ha_policy=cfo

                Volumes: vol0, Source_Volume

                Plex /aggr0/plex0: online, normal, active
                    RAID group /aggr0/plex0/rg0: normal

itsotuc1>
```

15.3.3 Controlling plexes

A plex can either be in an online or offline state. In the online state, the plex is available for read or write access and the contents of the plex are current. In the offline state, the plex is not accessible for read or write.

Example 15-5 shows how to change the state of a plex from offline to online, and from online to offline.

Example 15-5 Changing the state of a plex

```
Data ONTAP (itsotuc1.itso.com)
login: root
Password:
itsotuc1> aggr status
           Aggr State           Status           Options
          aggr1 online          raid_dp, aggr    raidsize=3
                                mirrored
                                32-bit
          aggr0 online          raid_dp, aggr    root
                                32-bit
itsotuc1> aggr offline aggr1
Aggregate 'aggr1' is now offline.

itsotuc1> Wed Apr 13 20:09:26 GMT [volaggr.offline:CRITICAL]: Some aggregates are
offline. Volume creation could cause duplicate FSIDs.

itsotuc1> aggr status
           Aggr State           Status           Options
          aggr1 offline         raid_dp, aggr    raidsize=3,
                                mirrored         lost_write_protect=off
                                32-bit
          aggr0 online          raid_dp, aggr    root
                                32-bit
```

Viewing the status of a plex

To view the status of a plex (the plex must be online), use these commands:

- **`sysconfig -r`**
- **`aggr status -r`**
- **`vol status -r`**

An online plex can be in the following states.

- Active: The plex is available for use.
- Adding disks or array LUNs: Data ONTAP is adding disks or LUNs to a RAID group or groups of the plex.
- Empty: The plex is part of an aggregate that is being created and Data ONTAP needs to zero out one or more of the disks or array LUNs targeted to the aggregate before adding the disks to the plex.
- Failed: One or more of the RAID groups in the plex failed.
- Inactive:The plex is not available for use.
- Normal: All RAID groups in the plex are functional.

- Out-of-date: The plex contents are out of date and the other plex of the aggregate has failed.
- Resyncing: The plex contents are being resynchronized with the contents of the other plex of the aggregate.

Comparing plexes of a mirrored aggregate

The plexes of a mirrored aggregate are almost always synchronized. However, you might need to compare the plexes of a mirrored aggregate. You can also choose to correct any differences between the plexes:

- The mirrored aggregate must be online before you can compare the plexes.
- Comparing plexes can affect system performance.

When comparing the two plexes of a mirrored aggregate, you can choose one of the following options:

- Data ONTAP compares plexes without correcting differences. It is the default behavior.
- Data ONTAP compares plexes and corrects the differences it finds. To correct differences, you need to specify which plex to correct. The plex is specified as plex number (0, 1, and so on).

Attention: This process might use advanced Data ONTAP commands. Contact technical support before correcting differences using this option.

To compare the two plexes of a mirrored aggregate, choose one of the actions from Table 15-1.

Table 15-1 Comparing two plexes of a mirrored aggregate

If	Then
You do not want Data ONTAP to correct differences	Enter the following command: **`aggr verify start aggrname -n`** **`aggrname`** is the name of the mirrored aggregate whose plexes you are comparing.
You want Data ONTAP to correct differences	Enter the following command: **`aggr verify start aggrname -f plexnumber`** **`aggrname`** is the name of the mirrored aggregate whose plexes you are comparing.

Tip: If **`aggrname`** is not specified, Data ONTAP compares the plexes of all mirrored aggregates that are online.

Removing a plex from a mirrored aggregate

An administrator can remove a plex from a mirrored aggregate, and can do this if they want to stop mirroring the aggregate, or if there is a problem with the plex. Removing a plex results in an unmirrored aggregate.

About the task

In case of a failure that causes a plex to fail, you can remove the plex from the mirrored aggregate, fix the problem, and then re-create it. You can also re-create it using a different set of disks or array LUNs, if the problem cannot be fixed.

Procedure

Follow these steps:

1. Take the selected plex offline by entering the following command:
   - **`aggr offline plex-name`**
   - **`plex-name`** is the name of one of the mirrored plexes.

   > **Tip:** Only one plex at a time can be taken offline.

2. Destroy the plex you took offline by entering the following command:
   - **`aggr destroy plex-name`**

Result

Removing and destroying a plex from a mirrored aggregate results in an unmirrored aggregate, because the aggregate now has only one plex.

After removing the plex, Data ONTAP converts the disks or array LUNs used by the plex into hot spares.

Splitting a mirrored aggregate

Splitting a mirrored aggregate removes the relationship between its two plexes and creates two independent unmirrored aggregates. After splitting, both the aggregates come online.

- Before you begin:

  Ensure that both plexes of the mirrored aggregate you are splitting are online and operational.

You can split a mirrored aggregate for one of the following reasons:

- You want to stop mirroring an aggregate.
- You want to move a mirrored aggregate to another location.
- You want to modify the mirrored aggregate, and test the modification before applying it. You can apply and test the modifications on the split-off copy of the plex, then apply those changes to the untouched original plex.

Before splitting, a mirrored aggregate or traditional volume has two plexes, *plex0* and *plex1*. After splitting, the new unmirrored aggregate with the new name has one plex, *plex0*. The new unmirrored aggregate with the original name also has one plex, either *plex0* or *plex1*.

The plex name for an unmirrored aggregate is unimportant because the aggregate has only one plex. If you use SyncMirror to mirror one of the unmirrored aggregates presented, the resulting plex names will always be *plex0* and *plex1*.

> **Attention:** You do not need to stop applications that are using the aggregate, before splitting a mirrored aggregate.

- Enter the following command (see Example 15-6):
  - **`aggr split aggrname/plexname new_aggr`**
    - **`aggrname`** is the name of the mirrored aggregate.
    - **`plexname`** is the name of one of the plexes in the mirrored aggregate.
    - **`new_aggr`** is the name of the new aggregate that will be created.

Example 15-6 Splitting plex0 from mirrored aggregate aggr0

```
aggr split aggr0/plex0 aggrNew
```

After splitting, there are two unmirrored aggregates, aggr0 and aggrNew.

Rejoining split aggregates

You can rejoin split aggregates and might want to do this if you have set up an HA pair configuration in a MetroCluster, and a disaster occurs breaking the HA pair.

Attention: When you rejoin split aggregates, Data ONTAP mirrors the data from one aggregate to the other and destroys data that existed on that aggregate prior to the rejoin.

There are additional considerations when planning to rejoin split aggregates that previously used MetroCluster to mirror SnapLock volumes.

You can use MetroCluster to mirror SnapLock volumes from one site to another. With proper configuration, the SnapLock volumes retain their characteristics at the mirror site. In case of a failure at the primary site, and if necessary, you can use the **`cf forcetakeover –d`** command to break the mirror relationship and to bring the mirror site online. After the failure at the primary site is resolved, the MetroCluster mirror relationship can be reestablished. The mirrors can be resynchronized before resuming normal operation.

Attention: The primary node might have data that was not mirrored before using the **`cf forcetakeover –d`** command. For example, the data might have been written to the primary node while the link between the sites was inoperative. In such a case, you need to back up the SnapLock volumes in the aggregate on the primary site, before resynchronizing the two mirror aggregates. This step of creating an additional backup for the SnapLock volumes is required to ensure the availability of all data.

Follow these steps:

1. Determine the aggregate whose data you want to keep and the aggregate whose data you want to be overwritten.
2. If the aggregate whose data is to be overwritten is online, take it offline by entering the following command:

   ```
   aggr offline aggrname
   ```

   Here, **`aggrname`** is the name of the aggregate.

Tip: An error message appears if the aggregate is already offline.

3. Re-create the mirrored aggregate by entering the following command:

   ```
   aggr mirror aggrname1 -v aggrname2
   ```

   – **`aggrname1`** is the name of the aggregate whose data you want to keep.

   – **`aggrname2`** is the name of the aggregate whose data you want to be overwritten by **`aggrname1`**.

15.4 How SyncMirror works with third-party storage

For both aggregates composed of native disks and aggregates composed of array LUNs, SyncMirror creates two physically-separated copies of an aggregate.

These copies of the aggregate, called plexes, are simultaneously updated; therefore, the two copies of the data are always identical. Data continues to be served if one copy becomes unavailable.

For third-party storage, the physical separation of the plexes protects against data loss if the following events occur:

- An array LUN fails.

  For example, a LUN failure can occur because of a double disk failure on the storage array.
- A storage array becomes unavailable.
- In a MetroCluster configuration, an entire site fails.

  An entire site can fail because of a disaster or prolonged power failure. If this situation occurs, the site administrator enters a command to enable the surviving node to take over the functions of the partner. Data is accessed on the plex of the surviving node.

For third-party storage, each plex must be on a separate set of array LUNs. The plexes can be in two physically separate locations on the same storage array, or each of the two plexes can be on a different storage array. In a MetroCluster configuration with third-party storage, each plex must be on a separate set of LUNs on different storage arrays. (N series Gateway systems on which native disk shelves are installed cannot be deployed in a MetroCluster configuration.)

Data ONTAP needs to know whether a plex is local to the system on which the aggregate is configured or in a remote location. Local in the context of third-party storage means on the storage array connected to the V-Series system on which the aggregate is configured. The SyncMirror *pool* to which an array LUN is assigned provides the information that Data ONTAP needs to determine whether the plex is local or remote.

The illustration in Figure 15-14 shows the relationships of plexes and pools to an aggregate. One plex is associated with pool0 and one plex is associated with pool1. The number 0 is typically associated with the local pool and the number 1 is typically associated with the remote pool. The remote plex is the mirror of the aggregate.

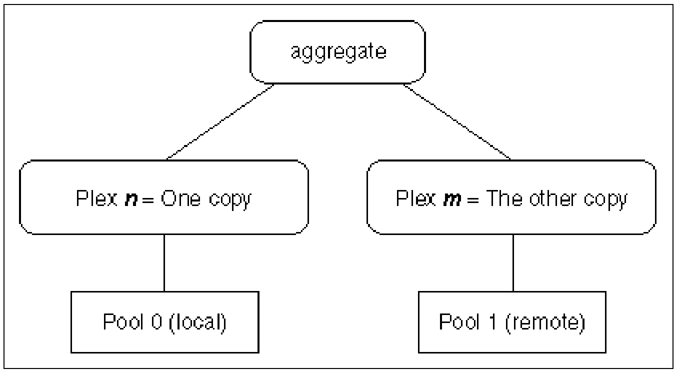


Figure 15-14 Relationship of plexes, and pools to an aggregate

15.5 Disaster recovery with SyncMirror

Another advantage of mirrored plexes is faster rebuild time. SyncMirror falls into the sixth tier of the disaster recovery hierarchy, as shown in Figure 15-15.

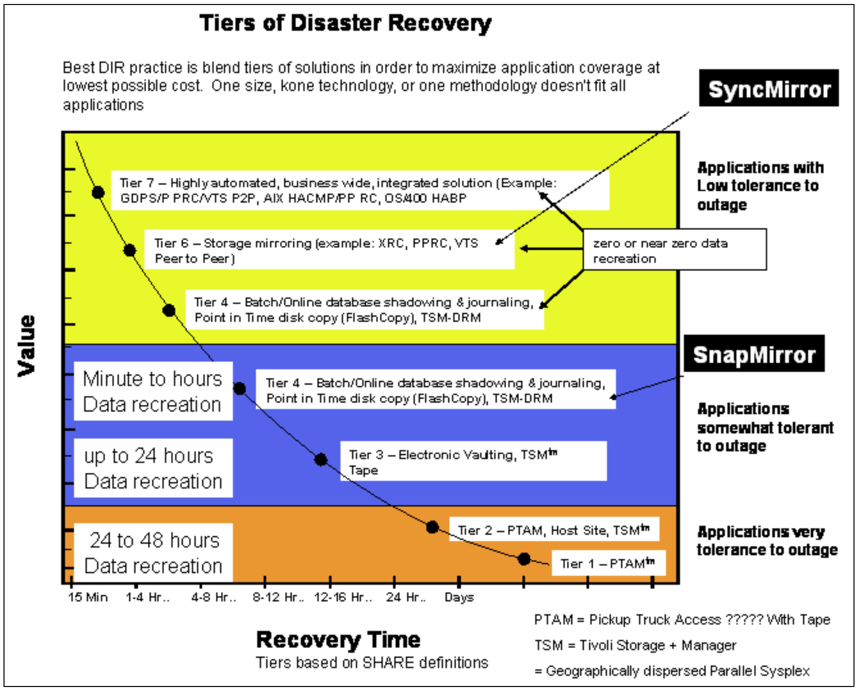


Figure 15-15 Tiers of disaster recovery

In contrast, if a SnapMirrored aggregate or traditional volume fails, its SnapMirror partner cannot automatically take over the file-serving functions and can only restore data to its condition at the time that the last Snapshot was created (you must issue commands to make the partner's data available).

With SyncMirror, IBM System Storage N series storage systems can tolerate multiple simultaneous disk failures across the RAID groups within the file system. This redundancy goes beyond typical mirrored (RAID 1) implementations seen in the market in that each SyncMirror RAID group is also RAID 4 or RAID-DP protected (Figure 15-16). A complete mirror can be lost, and an additional single drive loss within RAID-DP group can occur without data loss.

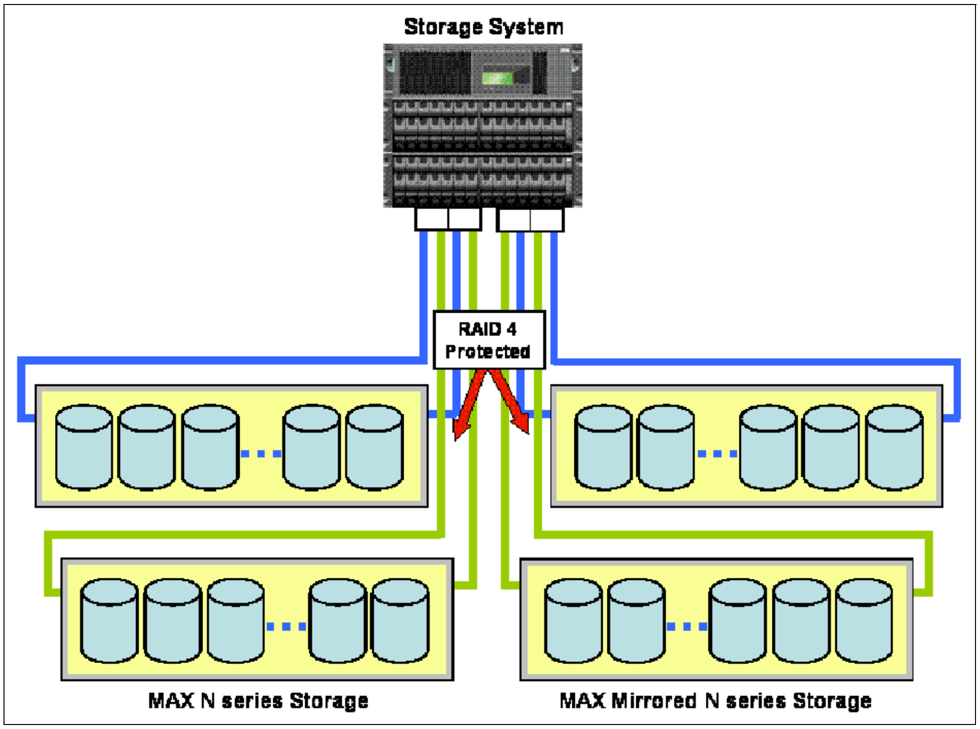


Figure 15-16 SyncMirror

Each RAID group is mirrored on storage connected to the server through completely independent data paths.

All mirrored storage is connected to separate host bus adapters (HBAs) with completely separate data paths for the greatest possible redundancy. Re-synchronization of a mirror volume occurs efficiently using Snapshots from the source volume. SyncMirror can be configured for use with stand-alone storage systems or clusters.

16

MetroCluster

This chapter describes the MetroCluster feature, which is an integrated, high-availability, business continuance solution that allows clustering of two N6000 or N7000 storage controllers at distances up to 100 kilometers.

The primary goal of MetroCluster is to provide mission-critical applications with redundant storage services in case of site-specific disasters. By synchronously mirroring data between two sites, it tolerates site-specific disasters with minimal interruption to applications and zero data loss.

The following topics are covered:

- Overview of MetroCluster
- Business continuity solutions
- Stretch MetroCluster
- Fabric Attached MetroCluster
- Synchronous mirroring with SyncMirror
- MetroCluster zoning and TI zones
- Failure scenarios

16.1 Overview of MetroCluster

IBM N series MetroCluster, as illustrated in Figure 16-1, is a solution that combines N series local clustering with synchronous mirroring to deliver continuous availability. MetroCluster expands the capabilities of the N series portfolio and works seamless with your host and storage environment to provide continuous data availability between two sites while eliminating the need to create and maintain complicated failover scripts. You will be able to serve data even if there is a complete site failure.

As a self-contained solution at the N series storage controller level, MetroCluster is able to transparently recover from failures, so business-critical applications continue uninterrupted.

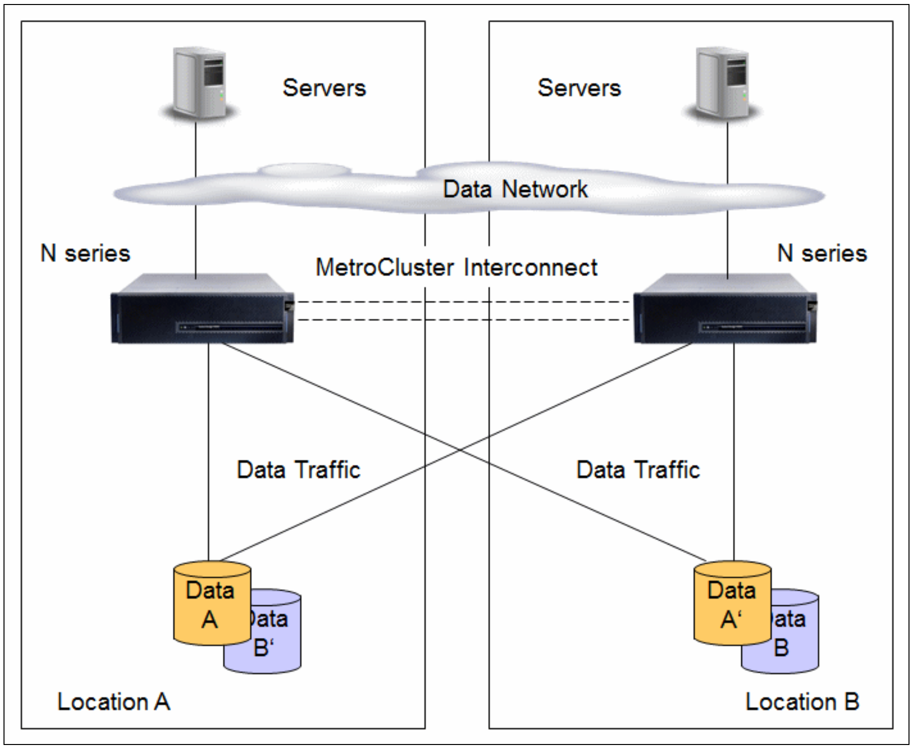


Figure 16-1 MetroCluster

MetroCluster is a fully integrated solution:

- Designed to be easy to administer
- Built on proven technology

It provides automatic failover to the remote data center and includes these benefits:

- Helps protect business continuity in the event of a failure in the primary data center
- Helps reduce dependency on IT staff for manual actions
- Provides synchronous mirroring up to 100 km

It has the following data replication capabilities:

- Designed to maintain a constantly up-to-date copy of data at a remote data center
- Supports replication of data from primary to remote site to maintain data currency

MetroCluster software provides an enterprise solution for high availability over wide area networks (WANs). MetroCluster deployments of N series storage systems are used for:

- Business continuance
- Disaster recovery
- Recovery point and recovery time:

  Achieving recovery point and recovery time objectives (instant failover), with more options regarding recovery point/time objectives in conjunction with other features

MetroCluster technology is an important component of enterprise data protection strategies. In case of a failure in one location (the local node or the disks are failing) then MetroCluster provides automatic failover to the remaining node and access to the data copy (because of SyncMirror) in the second location.

A MetroCluster system is made up of the following components:

- Two N series storage controllers, HA configuration: Provide the nodes for serving the data in case of a failure. N62x0 and N7950T systems are supported in MetroCluster configurations, whereas N3x00 is not supported.
- MetroCluster VI FC HBA, used for cluster interconnect.
- SyncMirror license: Provides an up-to-date copy of data at the remote site. Data is ready for access after failover without administrator intervention. (comes with Data ONTAP Essentials, and is not required in Data ONTAP 8.2).
- MetroCluster/Cluster remote and CFO license: Provides a mechanism for failover (automatically or administrator driven).
- FC switches: Provide storage system connectivity between sites/locations. (for fabric MetroClusters only).
- FibreBridges: Needed if you are going to use EXN3000 or EXXN3500 SAS Shelves.
- Cables: Multimode fiber optic cables (single-mode cables are not supported).

MetroCluster allows the Active/Active configuration to be spread across data centers up to 100 kilometers apart. In the event of an outage at one data center, the second data center can assume all affected storage operations lost with the original data center.

SyncMirror is required as part of MetroCluster to ensure that an identical copy of the data exists in the second data center in case the original data center is lost. If site A goes down, MetroCluster allows you to rapidly resume operations at a remote site minutes after a disaster, SyncMirror is used in MetroCluster environments to mirror data in two locations, as illustrated in Figure 16-2. Aggregate mirroring must be like-to-like disk types.

Licenses:

- Since the Data ONTAP 7.3 release, the cluster license and SyncMirror license are part of the base software bundle.
- Since the Data ONTAP 8.2 release, the SyncMirror feature is enabled by default.

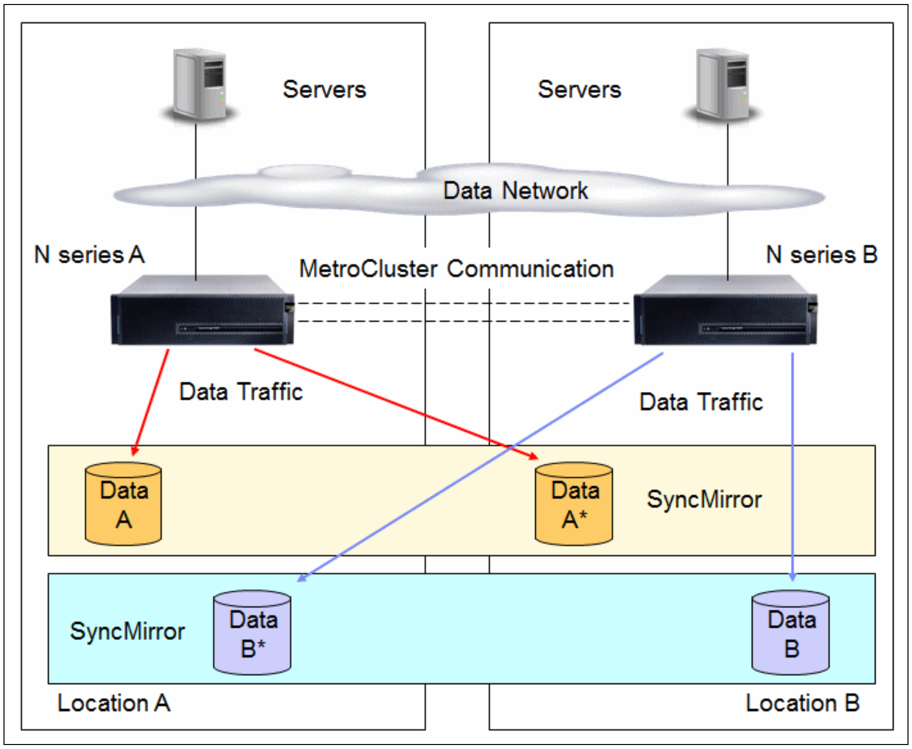


Figure 16-2 Logical view of MetroCluster utilizing SyncMirror

Geographical separation of N series nodes is implemented by physically separating controllers and storage, creating two MetroCluster halves. For distances under 500 m (campus distances), long cables are used to create Stretch MetroCluster configurations.

For distances more than 500 m but less than 100 km (metro distances), a fabric is implemented across the two geographies, creating a Fabric Attached MetroCluster configuration.

The Cluster_Remote license provides features that enable the administrator to declare a site disaster and initiate a site failover using a single command. The **`cf forcetakeover -d`** command initiates a takeover of the local partner even in the absence of a quorum of partner mailbox disks. This gives the administrator the ability to declare a site-specific disaster and have one node take over its partner's identity without a quorum of disks.

Several requirements must be in place to enable takeover in a site disaster:

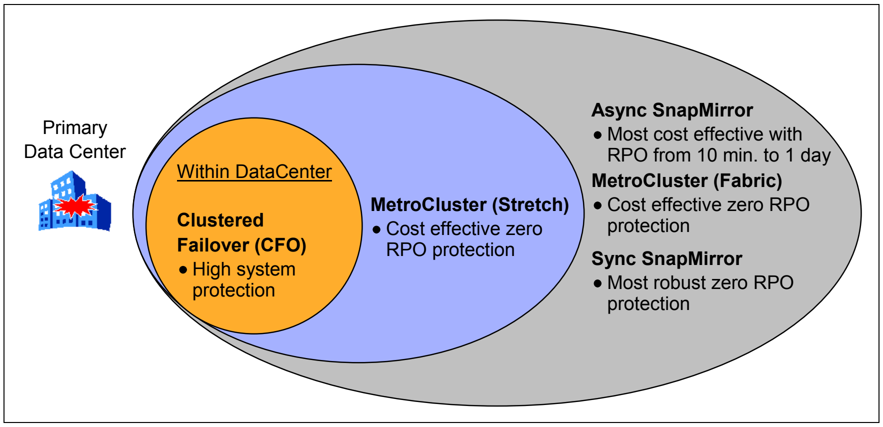
- Root volumes of both storage systems *must* be synchronously mirrored.
- Only synchronously mirrored aggregates are available during a site disaster.

Administrator intervention, that is, issuing the **`forcetakeover`** command, is required as a safety precaution against a *split brain* scenario.

Important: Site-specific disasters are not the same as a normal cluster failover.

16.2 Business continuity solutions

The N series offers several levels of protection with several different options. MetroCluster is just one of the options offered by the N series. MetroCluster fits into the campus-level distance requirement of business continuity. See Figure 16-3.



Primary
Data Center
Within DataCenter
Clustered
Failover (CFO)
• High system protection
MetroCluster (Stretch)
• Cost effective zero RPO protection
Async SnapMirror
• Most cost effective with RPO from 10 min. to 1 day
MetroCluster (Fabric)
• Cost effective zero RPO protection
Sync SnapMirror
• Most robust zero RPO protection

Figure 16-3 Business continuity with IBM System Storage N series

See Table 16-1 for differences between synchronous SnapMirror and MetroCluster with SyncMirror.

Table 16-1 Differences between Sync SnapMirror and MetroCluster SyncMirror

Feature	Synchronous SnapMirror	MetroCluster (SyncMirror)
Network for Replication	FC or IP	FC only
Concurrent transfer limited	Yes	No
Distance limitation	Up to 200 km (depending on latency)	100 km (Fabric MetroCluster)
Replication between HA pairs	Yes	No
Deduplication	Dedup volume and sync volume cannot be in same aggregate	Yes
Use of secondary node for an additional async mirroring	Yes	No, async replication occurs from primary plex

16.3 Stretch MetroCluster

The Stretch MetroCluster configuration uses two storage systems that are connected to provide high availability and data mirroring. You can place these two systems in separate locations. When the distance between the two systems is less than 500 meters, you can implement Stretch MetroCluster. The cabling is direct connected between nodes and shelves. FibreBridges are required when using SAS Shelves (EXN3000 and EXN3500).

16.3.1 Planning Stretch MetroCluster configurations

For planning and sizing Stretch MetroCluster environments, observe these considerations:

- Use multipath HA (MPHA) cabling.
- Use FibreBridges in conjunction with SAS Shelves (EXN3000 & EXN3500).
- N62x0 and N7950T systems require FC/VI cards for Fabric MetroClusters.
- Provide enough ports/loops to satisfy performance (plan additional adapters if appropriate).

  Stretch MetroCluster controllers connect directly to local shelves and remote shelves. The minimum is four FC ports per controller for a single stack (or loop) configuration. But keep in mind that you will mix the pools of the different two controllers in each stack (see Figure 16-4 for details).

  > **Tip:** A Stretch MetroCluster solution requires at least four disk shelves.

  It has a minimal impact on the environment, because in case of a disk failure, (+ replacement), you have to assign this disk manually to the correct controller/pool. Therefore, you must use disk pools on different stacks.

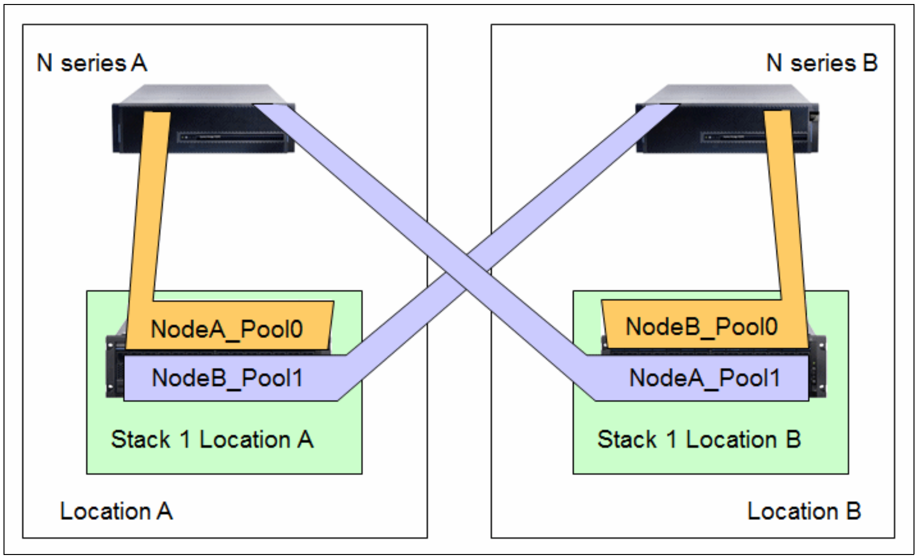


Figure 16-4 Stretch MetroCluster setup with only one stack per site

- Stretch MetroCluster has no imposed spindle limits, just the platform limit.
- Take care in planning N6210 MetroCluster configurations, because the N6210 has only two FC initiator onboard ports and two PCI expansion slots.

  Because you will use one slot for the FC/VI adapter, you have only one remaining slot for an FC initiator card. Due to the minimum of four FC ports, needed for Stretch MetroCluster, there are two configurations possible:

  – Two onboard FC ports + dual port FC initiator adapter

  – Quad port FC initiator HBA (frees up onboard FC ports)

  Remember that all slots are in use and the N6210 cannot be upgraded with other adapters.
- Mixed SATA and FC configurations are allowed, provided that the following requirements are met:

  – There is no intermixing of FC and SATA shelves on the same loop.

  – Mirrored shelves must be of the same type as their parents.

The Stretch MetroCluster heads can have a distance of up to 500 m (@2 Gbps). Greater distances can be available at lower speeds (check with RPQ/SCORE). Qualified distances are up to 500 m. If you have distances greater than 500 m, choose Fabric MetroCluster. Table 16-2 lists *theoretical* Stretch MetroCluster distances.

Table 16-2 T Theoretical MetroCluster distances

Data Rate in Gbps	OM-2 (50/125 um)	OM-3 (50/125 um)	OM-3+
1	500	860	1130
2	300	500	650
4	150	270	350

Attention: Maximum distance *supported* for Stretch MetroCluster is as follows:

- 2 Gpps: 500 meters
- 4 Gbps: 270 meters
- 8 Gbps: 150 meters

16.3.2 Cabling Stretch MetroClusters

Figure 16-5 shows an example of a Stretch MetroCluster with two EXN4000 FC shelves on each site.

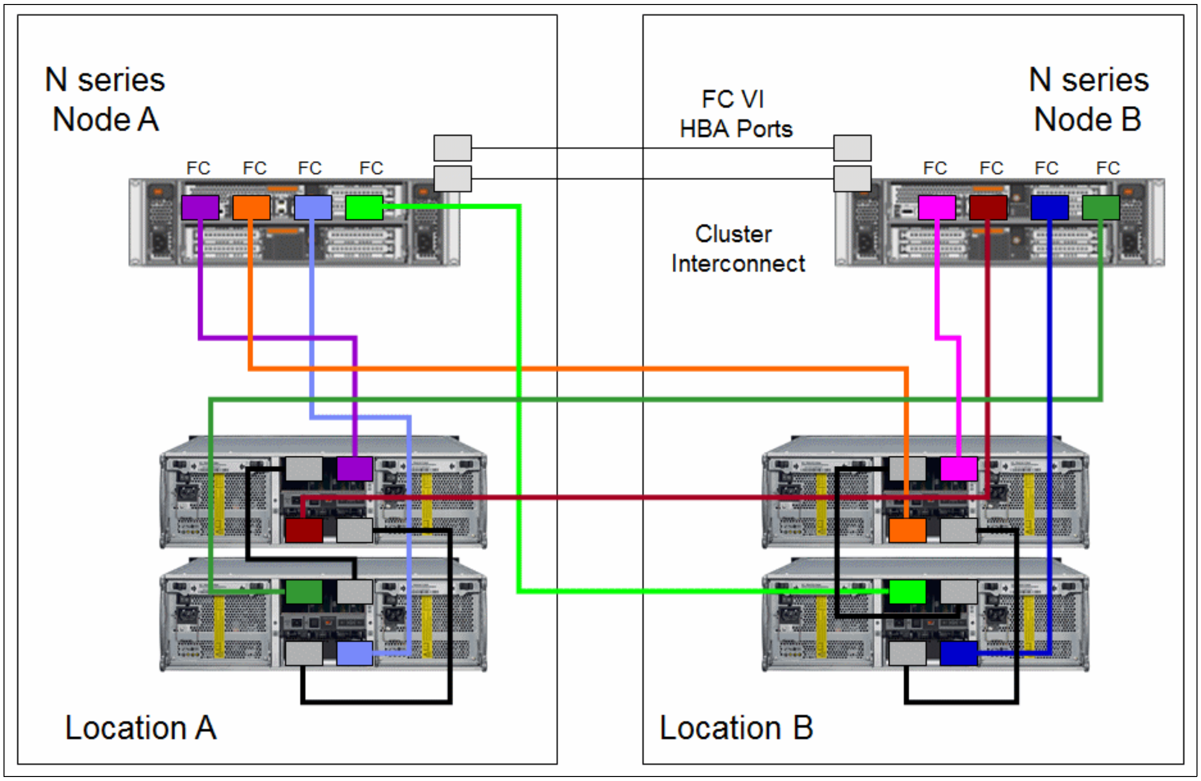


Figure 16-5 Stretch MetroCluster cabling with EXN4000

If you decide to use SAS Shelves (EXN3000 and EXN3500), then you need to use the FibreBridges.

Starting with Data ONTAP 8.1, EXN3000 (SAS or SATA) and EXN3500 are supported on Stretch MetroCluster (and Fabric MetroCluster as well) through SAS FC bridge (FibreBridge). The FibreBridge performs protocol conversion from SAS to FC and enables connectivity between Fibre Channel initiators and SAS storage enclosure devices, enabling SAS disks to appear as LUNs in a MetroCluster fabric. You need at minimum four FibreBridges (minimum is two per stack) in a MetroCluster environment. A sample is shown in Figure 16-6.

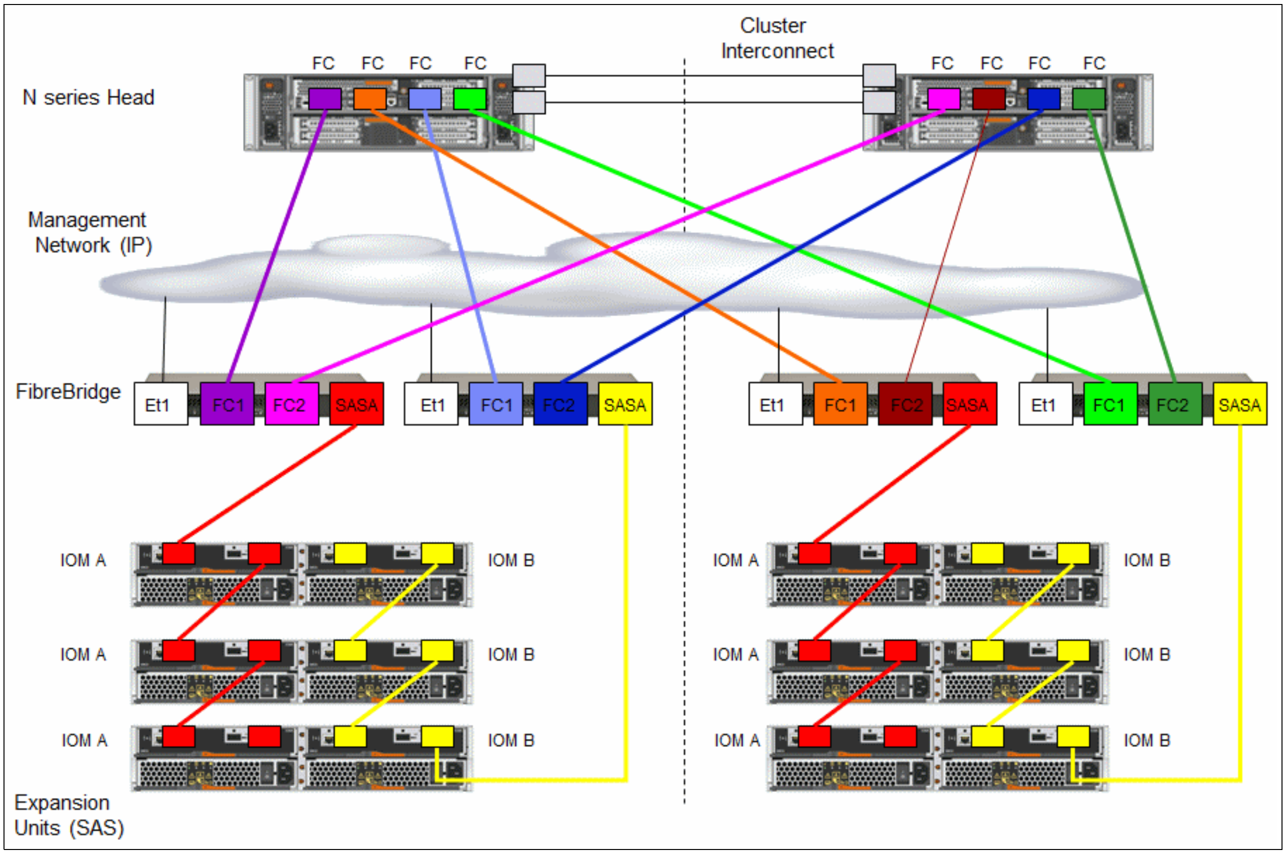


Figure 16-6 Cabling Stretch MetroCluster with FibreBridges and SAS Shelves

More details about the SAS Bridges can be found in the “SAS FibreBridges” chapter included in the companion book, *IBM System Storage N Series Hardware Guide*, SG24-7840, which is located at the following website:

`http://www.redbooks.ibm.com/abstracts/sg247840.html?Open`

16.4 Fabric Attached MetroCluster

Fabric Attached MetroCluster, sometimes referred to as Fabric MetroCluster, is based on the same concept as Stretch MetroCluster but provides greater distances (up to 100 km) throughout its SAN Fabrics. Both nodes in a Fabric MetroCluster are connected through four FC switches (two fabrics) for high availability and data mirroring. There is no direct connection as with Stretch MetroCluster. The nodes can be placed in different locations. Since Data ONTAP 8.0, Fabric Metro Clusters require dedicated fabrics for internal connectivity (back-end traffic and FC/VI communication). It is not supported to share this infrastructure with other systems.

A minimum of four FibreBridges are required when using SAS Shelves (EXN3000 and EXN3500) in a MetroCluster environment.

16.4.1 Planning Fabric Attached MetroCluster configurations

For planning and sizing Fabric MetroCluster environments, observe these considerations:

- Use FibreBridges in conjunction with SAS Shelves (EXN3000 & EXN3500).
- Provide enough ports/loops to satisfy performance (plan additional adapters if appropriate).
- Storage must be symmetric (such as the same storage on both sides). For storage that is not symmetric, but is similar, file an RPQ/SCORE.
- Keep in mind that N series native disk shelf disk drives are not supported with MetroClusters.
- Four Brocade/IBM B-Type FC Switches are needed. For supported Switches and firmware in Fabric MetroCluster environments, see the Interoperability Matrix:

  http://www-304.ibm.com/support/docview.wss?uid=ssg1S7003897

  One pair of FC switches is required at each location. The switches must be dedicated for the MetroCluster environment, and they cannot be shared with other systems. Remember that you might need licenses for FC switches; for example, the following licenses:

  - Extended distance license (if over 10 km)
  - Full-fabric license
  - Ports-on-demand (POD) licenses (for additional ports)
- Infrastructure / connectivity:
  - Dark fiber: Direct connections using long-wave SFPs (no standard offering available for these SFPs for large distances [> 30 km])) can be provided by the customer.
  - Leased metro-wide transport services from a service provider: Typically provisioned by dense wavelength division multiplexer/time division multiplexer/optical add drop multiplexer (DWDM/TDM/OADM) devices. Make sure that the device is supported by fabric switch vendor (IBM/Brocade).
  - Dedicated bandwidth between sites (mandatory): One interswitch link (ISL) per fabric, or two ISLs if using the traffic isolation (TI) feature and appropriate zoning. Do not use ISL trunking, because it is not supported,
- Take care in designing fabric MetroCluster infrastructure. Check ISL requirements and keep in mind that cluster interconnect needs proper planning and performance.
- Latency considerations:

  A dedicated fiber link has a round trip time (RTT) of appox1 ms for every 100 km (~ 60 miles). Additional nonsignificant latency might be introduced by devices (for example, multiplexers) en route. Generally speaking, as distance between sites increases (assuming 100 km = 1 ms link latency):

  - Storage response time increases by the link latency. For example, if storage has a response time of 1.5 ms for local access, then over 100 km, the response time increases by 1 ms to 2.5 ms.
  - Applications, in contrast, respond differently to the increase in storage response time. For some applications, the response time increases by approximately the link latency, while for other applications, the response time increases by greater than the link latency. For example, application A response time with local storage access is 5 ms and over 100 km is 6 ms, while application B response time with local storage access is 5 ms, and over 100 km is 10 ms.

- Take care in planning N6210 MetroCluster configurations, because the N6210 has only two FC initiator onboard ports and two PCI expansion slots.

  Because you will use one slot for the FC/VI adapter, you have only one remaining slot for an FC initiator card. Because of a minimum of four FC ports needed for Stretch MetroCluster, there are two configurations possible:

  – Two onboard FC ports + dual port FC initiator adapter

  – Quad port FC initiator HBA (frees up onboard FC ports)

  Remember that all slots are in use and the N6210 cannot be upgraded with other adapters.

- Currently, when using SAS Shelves, there is no spindle limit with *Fabric* MetroCluster and Data ONTAP 8.x. Only the platform spindle limit does apply (N62x0 and N7950T), as you can see in Table 16-3.

Table 16-3 Maximum number of spindles with DOT 8.x and Fabric MetroCluster

Platform	**Number of spindles SAS/SATA (requires FibreBridges)**	**Maximum number of FC disks**
N6210	480	480
N6240	600	600
N6270	960	840 (672 with DOT7.3.2 or 7.3.4)
N7950T	1176	840 (672 with DOT7.3.2 or 7.3.4)

Important: Fabric MetroClusters need four *dedicated* FC switches in two fabrics. Each fabric will be dedicated to the traffic for a single MetroCluster. No other devices can be connected to the MetroCluster fabric.

Beginning with Data ONTAP 8.1, MetroCluster supports shared-switches configuration with Brocade 5100 switches. Two MetroCluster configurations can be built with four Brocade 5100 switches. For more information about shared-switches configuration, see the *Data ONTAP High Availability Configuration Guide*.

Important: Always refer to the MetroCluster Interoperability Matrix on the IBM Support site for the latest information about components and compatibility.

16.4.2 Cabling Fabric Attached MetroClusters

Figure 16-7 shows an example of a Fabric MetroCluster with two EXN4000 FC shelves on each site.

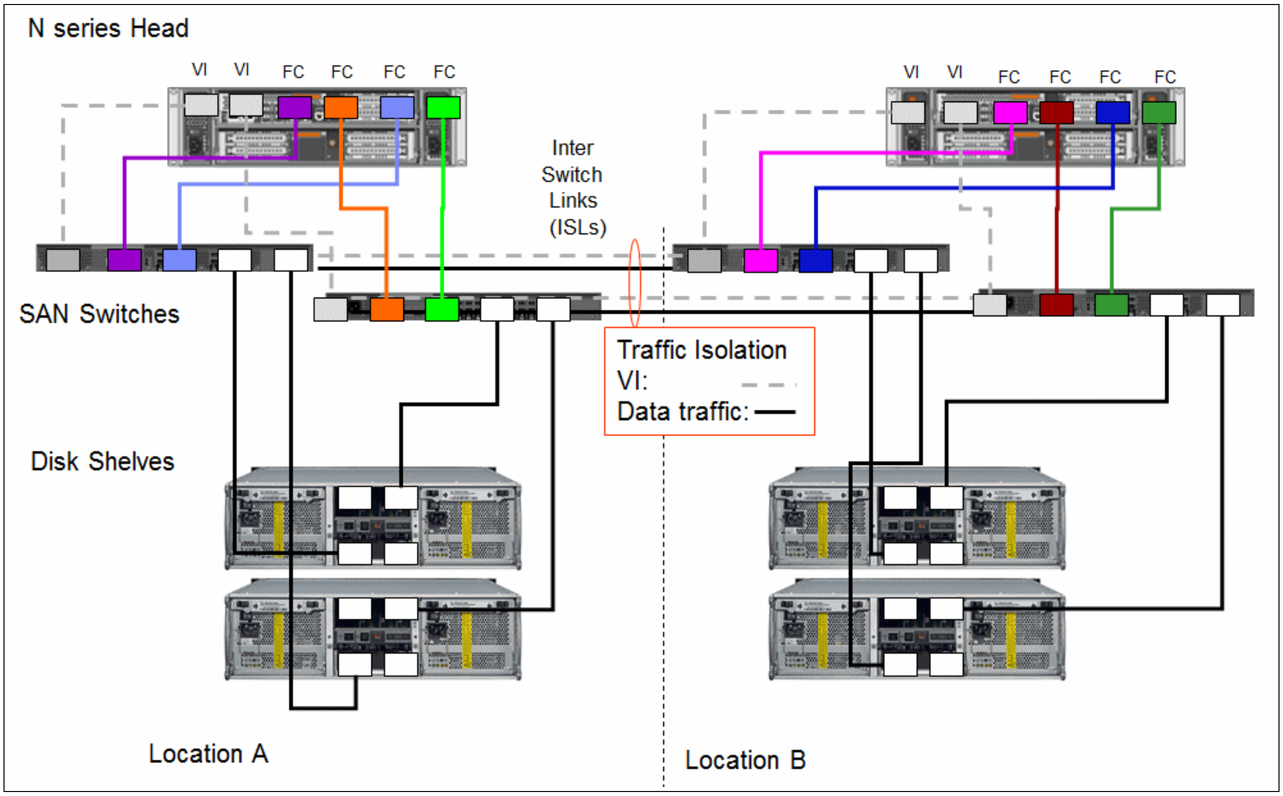


Figure 16-7 Fabric MetroCluster cabling with EXN4000

Fabric MetroCluster configurations use Fibre Channel switches as the means to separate the controllers by a greater distance. The switches are connected between the controller heads and the disk shelves, and to each other. Each disk drive or LUN individually logs into a Fibre Channel fabric.

The nature of this architecture requires, for performance reasons, that the two fabrics be completely dedicated to Fabric MetroCluster. Extensive testing was done to ensure adequate performance with switches included in a Fabric MetroCluster configuration. For this reason, Fabric MetroCluster requirements prohibit the use of any other model or vendor of Fibre Channel switch than the Brocade included with the Fabric MetroCluster.

If you decide to use SAS Shelves (EXN3000 and EXN3500), you have to use the FibreBridges).

Starting with Data ONTAP 8.1, EXN3000 (SAS or SATA) and EXN3500 are supported on Stretch MetroCluster (and Fabric MetroCluster as well) via SAS FC bridge (FibreBridge). The FibreBridge performs protocol conversion from SAS to FC and enables connectivity between Fibre Channel initiators and SAS storage enclosure devices to enabling SAS disks to appear as LUNs in a MetroCluster fabric. You need at minimum four FibreBridges (minimum is two per stack) in a MetroCluster environment (see Figure 16-8).

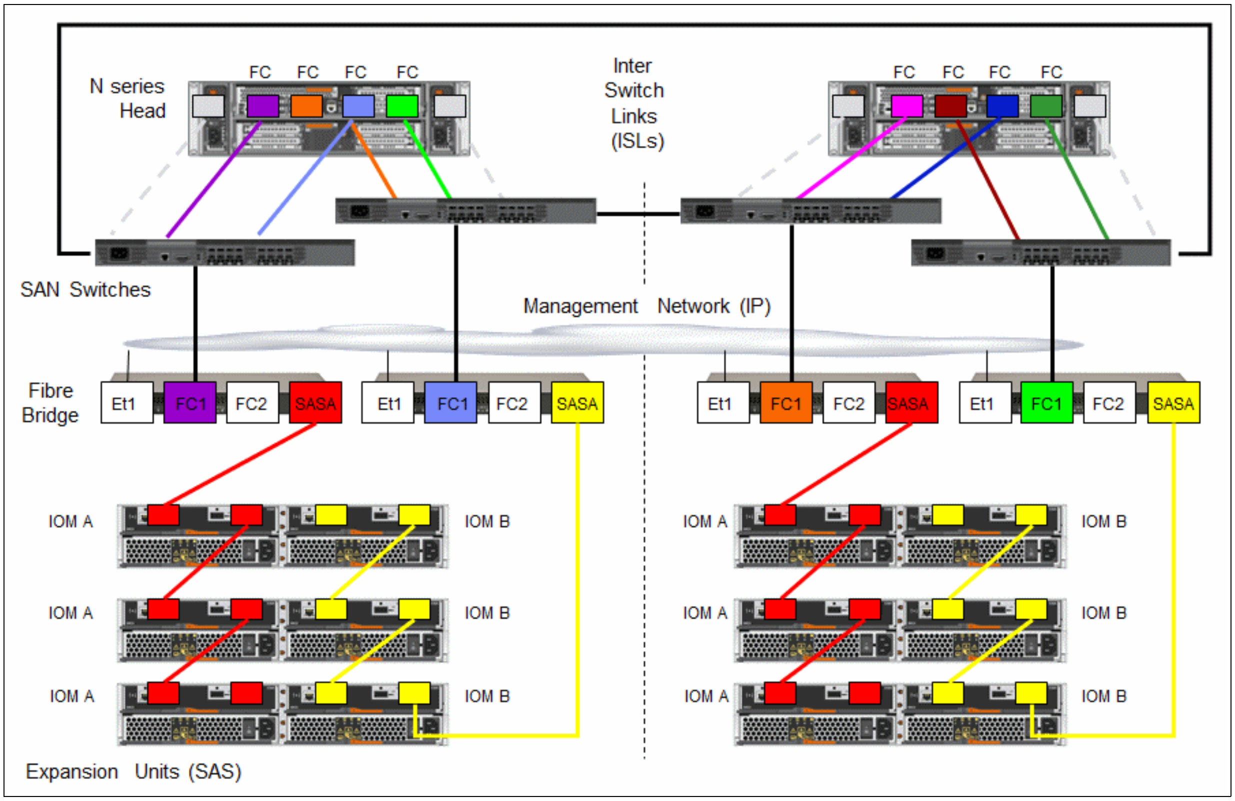


Figure 16-8 Cabling Fabric MetroCluster with FibreBridges and SAS Shelves

More details about the SAS Bridges can be found in the “SAS FibreBridges” chapter included in the companion book: *IBM System Storage N Series Hardware Guide*, SG24-7840, which is located at the following website:

http://www.redbooks.ibm.com/abstracts/sg247840.html?Open

16.5 Synchronous mirroring with SyncMirror

SyncMirror synchronously mirrors data across the two halves of the MetroCluster configuration by writing data to two plexes: the local plex (on the local shelf) actively serving data and the remote plex (on the remote shelf) normally not serving data. On local shelf failure, the remote shelf seamless takes over data-serving operations. Both copies or plexes are updated synchronously on writes, thus ensuring consistency.

16.5.1 SyncMirror overview

The design of IBM System Storage N series and MetroCluster provides data availability even in the event of an outage, whether it is due to a disk problem, cable break, or host bus adapter (HBA) failure. SyncMirror can instantly access the mirrored data without operator intervention or disruption to client applications. Read performance is optimized by performing application reads from both plexes (Figure 16-9).

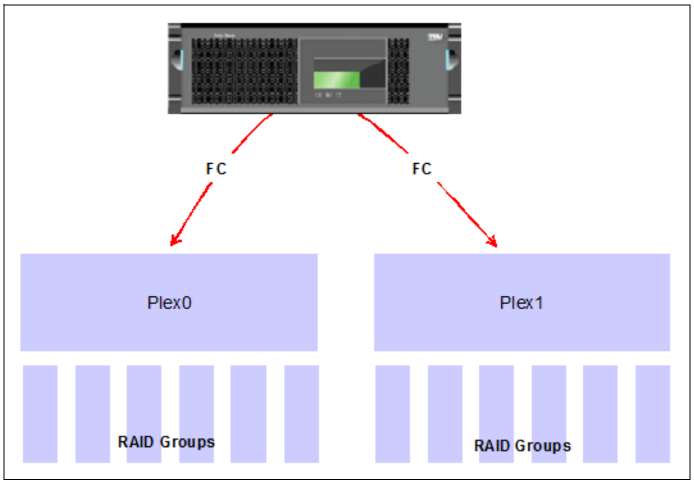


Figure 16-9 Synchronous mirroring

SyncMirror is used to create aggregate mirrors. When planning SyncMirror environments, keep in mind the following considerations:

- Aggregate mirrors must be on the remote site (geographically separated).
- In normal mode (no takeover), aggregate mirrors cannot be served out.
- Aggregate mirrors can exist only between like drive types.

When the SyncMirror license is installed, disks are divided into pools (pool0: local, pool1: remote/mirror). When a mirror is created, Data ONTAP pulls disks from pool0 for the local aggregate and from pool1 for the mirrored aggregate. Verify the correct number of disks in each pool before creating the aggregates. Any of the following commands can be used, as you can see in Example 16-1.

Example 16-1 Verification of SyncMirror

```
itsosj_n1>sysconfig -r
itsosj_n1>aggr status -r
itsosj_n1>vol status -r
```

To see the volume /plex/raidgroup relationship, use the **`sysconfig –r`** command, as shown in Example 16-2. Use the **`aggr mirror`** command to start mirroring the plexes.

Example 16-2 Viewing the aggregate status

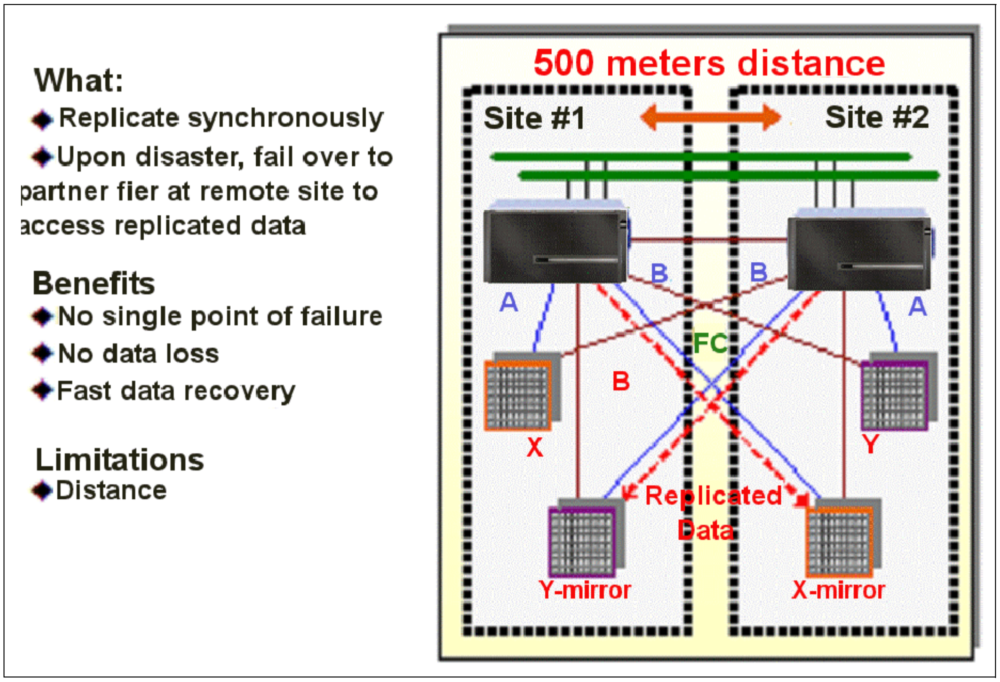
```
n5500-ctr-tic-1> sysconfig -r
Aggregate aggr0 (online, raid_dp, mirrored) (block checksums)
  Plex /aggr0/plex0 (online, normal, active, pool0)
    RAID group /aggr0/plex0/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------- ---- ---- ---- ----- --------------    --------------
      dparity   0a.16   0a      1   0   FC:A   0  FCAL 15000 136000/278528000  137104/280790184
      parity    0a.17   0a      1   1   FC:A   0  FCAL 15000 136000/278528000  137104/280790184
      data      0a.18   0a      1   2   FC:A   0  FCAL 15000 136000/278528000  137104/280790184

  Plex /aggr0/plex2 (online, normal, active, pool1)
    RAID group /aggr0/plex2/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------- ---- ---- ---- ----- --------------    --------------
      dparity   0c.25   0c      1   9   FC:B   1  FCAL 15000 136000/278528000  137104/280790184
      parity    0c.24   0c      1   8   FC:B   1  FCAL 15000 136000/278528000  137104/280790184
      data      0c.23   0c      1   7   FC:B   1  FCAL 15000 136000/278528000  137104/280790184

Aggregate aggr1 (online, raid4, mirrored) (block checksums)
  Plex /aggr1/plex0 (online, normal, active, pool0)
    RAID group /aggr1/plex0/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------- ---- ---- ---- ----- --------------    --------------
      parity    0a.19   0a      1   3   FC:A   0  FCAL 15000 136000/278528000  137104/280790184
      data      0a.21   0a      1   5   FC:A   0  FCAL 15000 136000/278528000  137104/280790184
      data      0a.20   0a      1   4   FC:A   0  FCAL 15000 136000/278528000  137104/280790184

  Plex /aggr1/plex1 (online, normal, active, pool1)
    RAID group /aggr1/plex1/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------- ---- ---- ---- ----- --------------    --------------
      parity    0c.26   0c      1   10  FC:B   1  FCAL 15000 272000/557056000  274845/562884296
      data      0c.20   0c      1   4   FC:B   1  FCAL 15000 136000/278528000  280104/573653840
      data      0c.29   0c      1   13  FC:B   1  FCAL 15000 136000/278528000  280104/573653840


Pool1 spare disks

RAID Disk       Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
---------       ------  ------------- ---- ---- ---- ----- --------------    --------------
Spare disks for block or zoned checksum traditional volumes or aggregates
spare           0c.28   0c      1   12  FC:B   1  FCAL 15000 272000/557056000  280104/573653840

Pool0 spare disks

RAID Disk       Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
---------       ------  ------------- ---- ---- ---- ----- --------------    --------------
Spare disks for block or zoned checksum traditional volumes or aggregates
spare           0a.22   0a      1   6   FC:A   0  FCAL 15000 136000/278528000  137104/280790184

Partner disks

RAID Disk       Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
---------       ------  ------------- ---- ---- ---- ----- --------------    --------------
partner         0a.25   0a      1   9   FC:A   1  FCAL 15000 0/0               137104/280790184
partner         0a.27   0a      1   11  FC:A   1  FCAL 15000 0/0               137104/280790184
partner         0a.26   0a      1   10  FC:A   1  FCAL 15000 0/0               137104/280790184
partner         0c.16   0c      1   0   FC:B   0  FCAL 15000 0/0               137104/280790184
partner         0c.21   0c     1   5   FC:B   0  FCAL 15000 0/0              137104/280790184
partner         0c.22   0c     1   6   FC:B   0  FCAL 15000 0/0              137104/280790184
partner         0a.29   0a     1   13  FC:A   1  FCAL 15000 0/0              137104/280790184
partner         0c.17   0c     1   1   FC:B   0  FCAL 15000 0/0              137104/280790184
partner         0c.27   0c     1   11  FC:B   0  FCAL 15000 0/0              137104/280790184
partner         0c.18   0c     1   2   FC:B   0  FCAL 15000 0/0              137104/280790184
partner         0a.23   0a     1   7   FC:A   1  FCAL 15000 0/0              137104/280790184
partner         0a.28   0a     1   12  FC:A   1  FCAL 15000 0/0              137104/280790184
partner         0a.24   0a     1   8   FC:A   1  FCAL 15000 0/0              137104/280790184
partner         0c.19   0c     1   3   FC:B   0  FCAL 15000 0/0              274845/562884296
```

16.5.2 SyncMirror without MetroCluster

SyncMirror local (without MetroCluster) is basically a standard cluster with one or both controllers mirroring their RAID to two separate shelves. The caveat in failover is that if you lose a controller and one of its RAID sets (plexes), the partner does not take over the other RAID set (plex). Therefore, without MetroCluster, all of the same rules apply as for a normal cluster:

- If controller A fails, partner B takes over.
- If loop A (Plex0) on controller A fails, controller A continues operation by running through loop B (Plex1).
- If controller A fails and either loop A or loop B (Plex0/Plex1) fails, you will not be able to continue.

MetroCluster protects against the following scenario: If controller A fails and its SyncMirrored shelves attached to loop A (Plex0) or loop B (Plex1) fail simultaneously, partner B takes over operation for partner A and its SyncMirrored plex on either loop A (Plex0) or loop B (Plex1). See Figure 16-10.



What:
- Replicate synchronously
- Upon disaster, fail over to partner fier at remote site to access replicated data

Benefits
- No single point of failure
- No data loss
- Fast data recovery

Limitations
- Distance

Figure 16-10 MetroCluster protection

16.6 MetroCluster zoning and TI zones

In a traditional SAN, there is great flexibility in connecting devices to ports as long as the ports are configured correctly and any zoning requirements are met. A MetroCluster, however, expects certain devices to be connected to specific ports or ranges of ports. It is therefore critical that cabling be exactly as described in the installation procedures. Also, no switch-specific functions such as trunking or zoning are currently used in a Fabric MetroCluster, making switch management minimal.

The Traffic Isolation (TI) zone feature of Brocade/IBM B type switches (FOS 6.0.0b or later) allows us to control the flow of interswitch traffic by creating a dedicated path for traffic flowing from a specific set of source ports. In the case of a fabric MetroCluster configuration, the traffic isolation feature can be used to dedicate an Inter Switch Link (ISL) to high-priority cluster interconnect traffic.

There is a need to isolate FCVI traffic. FCVI exchanges are high priority traffic that must not be subject to any interruption or congestions caused by storage traffic.

Fabric OS v6.0.0b introduces the concept of Traffic Isolation Zones:

- They can create a dedicated route.
- They do not modify the routing table.
- They are implemented across the entire data path from a single location.
- They do not require a license.
- TI Zones are called “zones”, but they are really about FSPF routing.
- TI zones need a standard zoning configuration being in effect.
- TI zones appear only in the defined zoning configuration (not in effective zoning configuration).
- You create TI Zones using Domain, Index (D, I) notation.
- E_Ports and F_ and FL_Ports must be included for an end-to-end route (initiator - target).
- Ports are only members of a single TI zone.

Without TI Zones, traffic is free to use either ISL, subject to the rules of Fabric Shortest Path First (FSPF) and Dynamic Path Selection (DPS), as you can see in Figure 16-11.

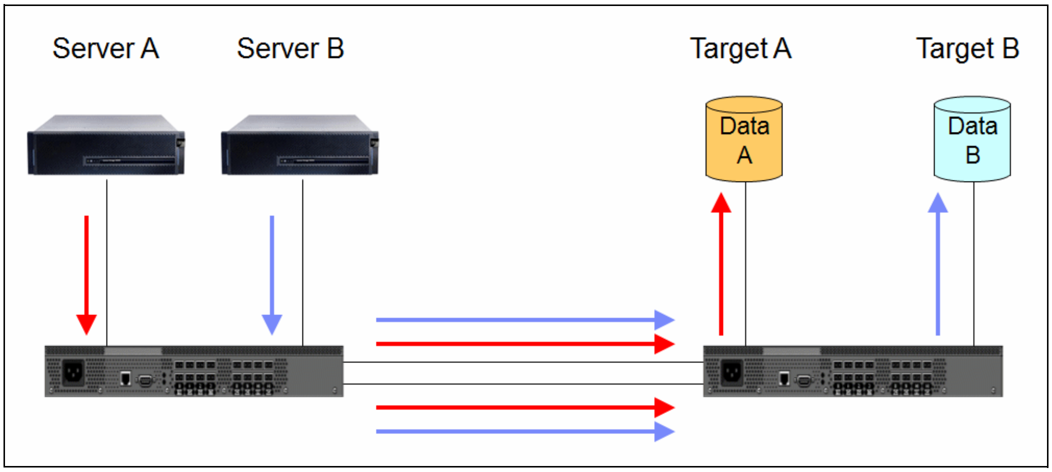


Figure 16-11 Traffic Flow without TI Zones

Customers can benefit from using two ISLs per fabric (instead of one ISL per fabric) to separate out high-priority cluster interconnect traffic from other traffic (to prevent contention) on the back-end fabric and for additional bandwidth in some cases. The TI feature is used to enable this separation. The TI feature provides better resiliency and performance but requires more fiber between sites.

Traffic isolation is implemented using a special zone, called a traffic isolation zone (TI zone). A TI zone indicates the set of ports and ISLs to be used for a specific traffic flow. When a TI zone is activated, the fabric attempts to isolate all interswitch traffic entering from a member of the zone to only those ISLs that have been included in the zone. The fabric also attempts to exclude traffic not in the TI zone from using ISLs within that TI zone.

TI Traffic Flow: TI Zones are a new feature of Fabric OS v6.0.0b:

- TI Zones only exist in the Defined Zoning Configuration.
- TI Zones must be created with Domain, Index notation only.
- TI Zones must include both E_Ports, and N_Ports in order to create a complete, dedicated, end-to-end route from Initiator to Target.

Each fabric will be configured to prohibit probing of the FCVI ports by the Fabric nameserver.

Figure 16-12 shows the dedicated traffic between Domain 1 and Domain 2. Data from system A would stay in the TI Zone “1-2-3-4” and would not pass TI Zone “5-6-7-8”“. So the traffic is routed on 2-3 for system A and 6-7 for system B.

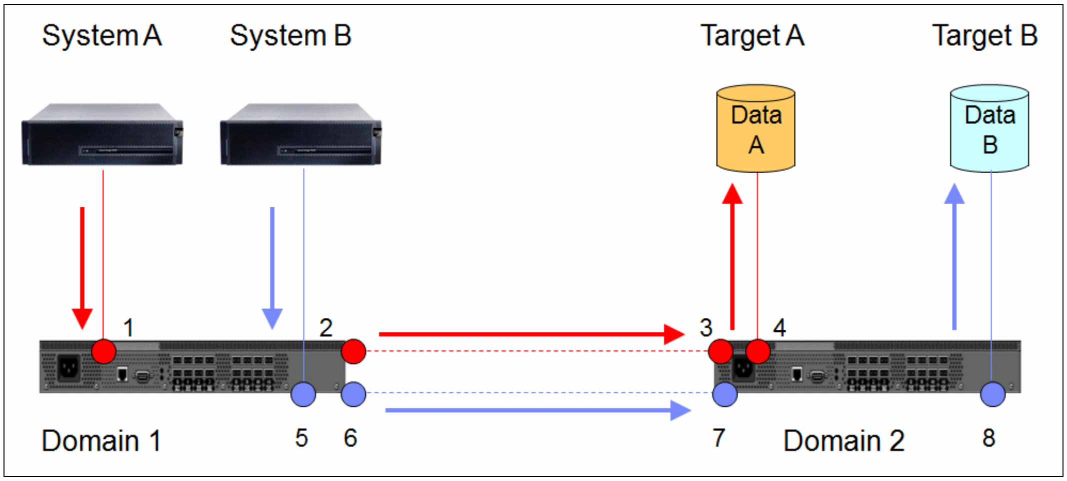


Figure 16-12 TI zones

Figure 16-13 shows an example of traffic isolation (TI) in a Fabric MetroCluster environment. VI traffic (orange) is separated from data/back-end traffic (black) by TI zones.

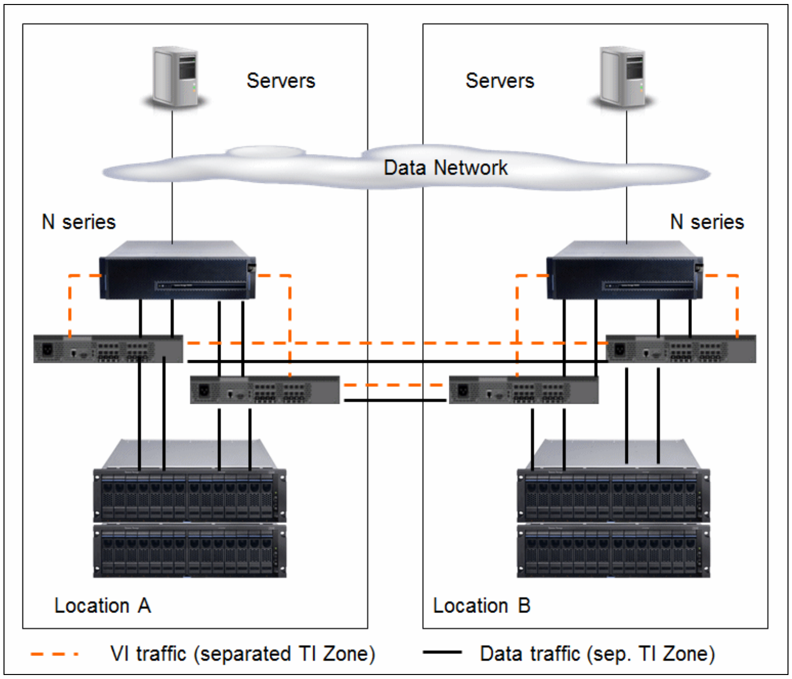


Figure 16-13 TI zones in MetroCluster environment

16.7 Failure scenarios

The following examples illustrate some possible failure scenarios and the resulting configurations when using MetroCluster.

16.7.1 MetroCluster host failure

In this scenario, N series N1 (Node 1) has failed. CFO/MetroCluster takes over the services and access to its disks (Figure 16-14). The fabric switches provide the connectivity for the N series N2, and the hosts to continue to access data without interruption.

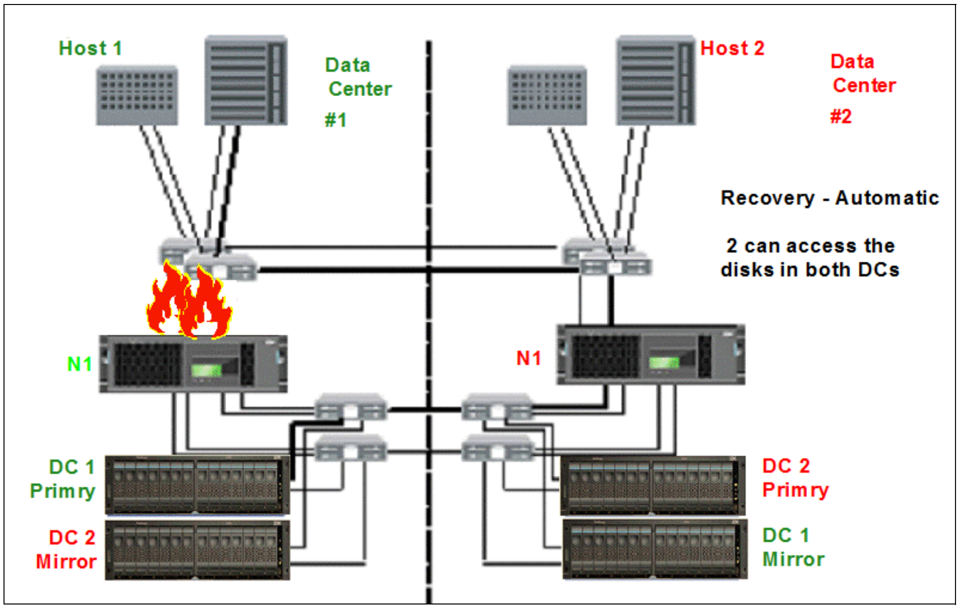


Figure 16-14 IBM System Storage N series failure

16.7.2 N series and expansion unit failure

This scenario (Figure 16-15) shows the loss of one site. resulting in failure of controller and shelves at the same time.

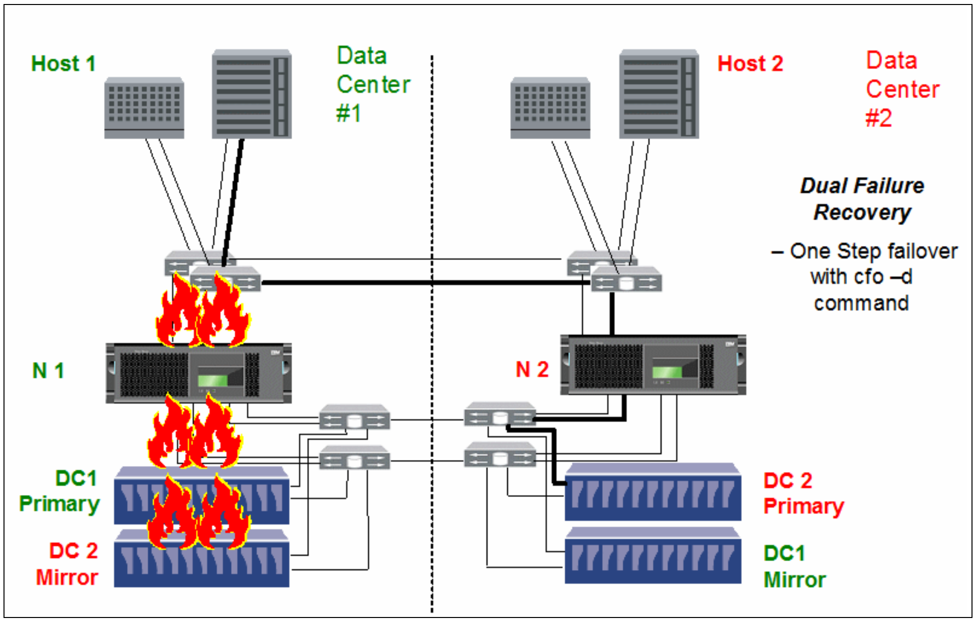


Figure 16-15 Controller and expansion unit failure

In order to continue access, a failover must be performed by the administrator issuing the `cfo -d` command. Data access is restored because DC1 mirror was in sync with DC1 primary. Through connectivity provided by the fabric switches, all hosts will again have access to required data.

16.7.3 MetroCluster interconnect failure

In this scenario, the fabric switch interconnects have failed (Figure 16-16). Although it is not a critical failure, resolution must occur promptly before a more critical failure occurs.

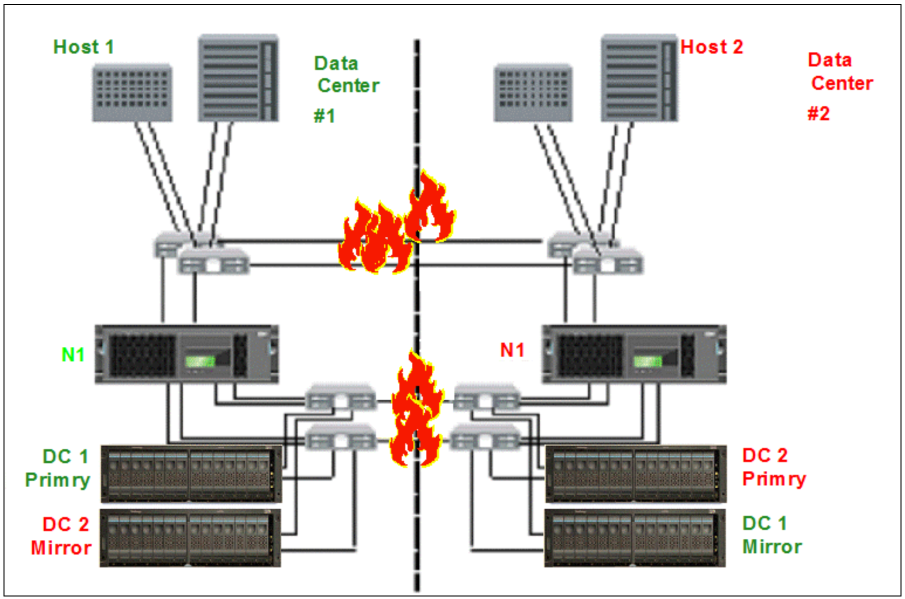


Figure 16-16 Interconnect failure

During this period, data access is uninterrupted to all hosts. No automated filer takeover occurs. Both filer heads will continue to run serving its LUNs/volumes. However, mirroring and failover are disabled, thus reducing data protection. When the interconnect failure is resolved, re-syncing of mirrors occurs.

16.7.4 MetroCluster site failure

In this scenario, a site disaster has occurred and all switches, storage systems, and hosts have been lost (Figure 16-17). To continue data access, a cluster failover must be initiated by using the **`cfo -d`** command. Both primaries now exist at data center 2, and hosting of Host1 is also done at data center 2.

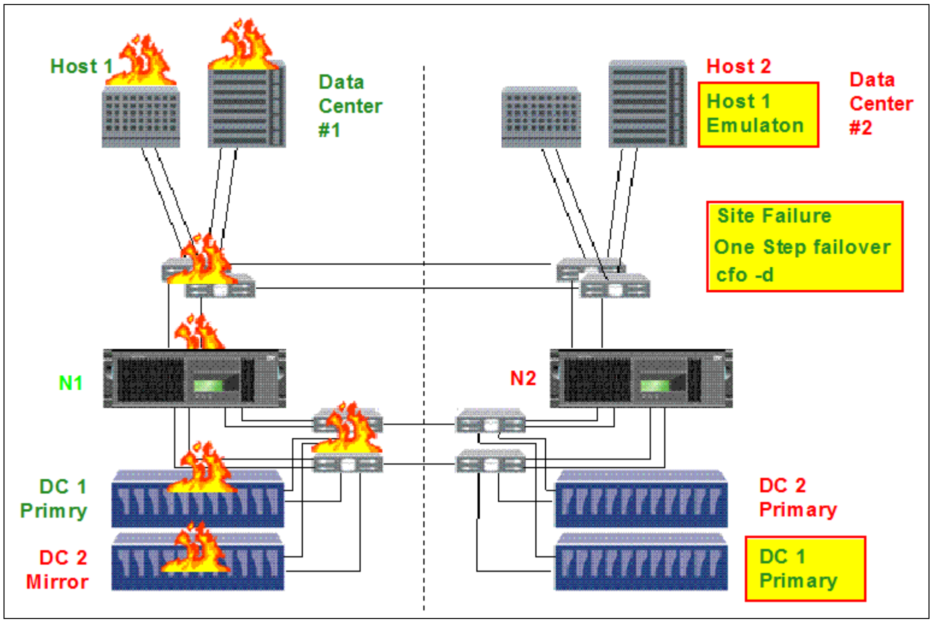


Figure 16-17 Site failure

Important: If the site failure is staggered in nature and the interconnect fails before the rest of the site is destroyed, there is a chance of data loss. It occurs because processing has continued after the interconnect has failed. Typically, site failures occur pervasively and at the same time.

16.7.5 MetroCluster site recovery

After the hosts, switches, and storage systems have been recovered at data center 1, a recovery can be performed. A **`cf giveback`** command is issued to resume normal operations (Figure 16-18). Mirrors are re-synchronized and primaries and mirrors are reversed to their previous status.

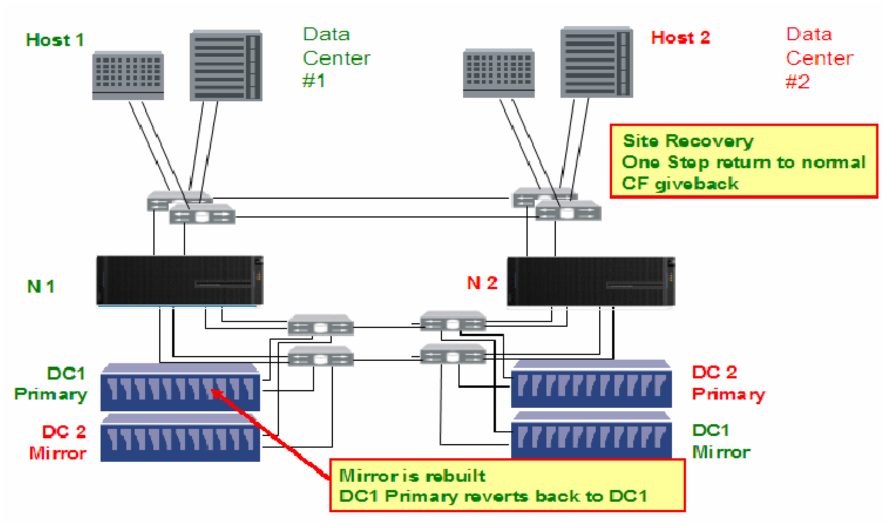


Figure 16-18 MetroCluster recovery

Part 3

Storage efficiency technology

The N series storage efficiency technology portfolio and monitoring tools can help you to reach never before seen levels of efficiency in your IT environment.

IBM N series customers can realize three key benefits through the use of our family of storage efficiency solutions:

- Store the maximum amount of data for the lowest possible cost
- Retain data on disk for longer periods of time
- Reduce data center power, cooling, and space requirements

Achieving these goals need not come at the cost of reduced performance or increased administration. We can reduce your storage burden while improving availability and manageability.

In this part of the book, you can learn more about our storage efficiency technology portfolio and products. The following topics are covered:

- SnapVault
- What storage efficiency is
- Deduplication
- Compression
- IBM Real-time Compression Appliance
- Thin replication using SnapVault and Volume SnapMirror

SnapVault

This chapter describes the SnapVault solution, which is a low-impact, disk-based, online backup of heterogeneous storage systems for fast and simple restores.

It introduces the SnapVault commands, which provide a means to create and manage Snapshot copies in each volume or aggregate.

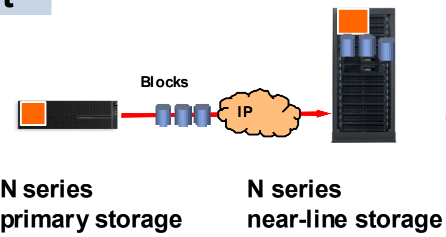
The following topics are covered:

- SnapVault at a glance
- Business applications of SnapVault
- Overview of SnapVault
- Benefits of using SnapVault
- SnapVault operation
- SnapVault backup
- SnapVault details
- Disaster recovery with SnapVault
- Remote solution using SnapVault
- Maximum number of concurrent SnapVault targets
- Preferred practices
- Summary

17.1 SnapVault at a glance

This section introduces the IBM System Storage N series SnapVault as shown in Figure 17-1.

SnapVault™



- Provides homogeneous super-efficient hourly disk-based online backup and restore by periodically backing up a Snapshot copy to another system
- **Full volume, native format copies** of primary N series systems reside on backup system and only incremental **block changes** are sent to backup system for efficient space management

Figure 17-1 SnapVault overview

SnapVault protects data on both IBM N series and non-IBM N series primary systems by maintaining a number of read-only versions of that data on a SnapVault secondary system and the SnapVault primary system.

SnapVault is a disk-based storage backup feature of Data ONTAP. SnapVault enables data stored on multiple systems to be backed up to a central, secondary system quickly and efficiently as read-only Snapshot copies.

In the event of data loss or corruption on a system, backed-up data can be restored from the SnapVault secondary system with less downtime and uncertainty than is associated with conventional tape backup and restore operations.

The following terms are used to describe the SnapVault feature:

- Primary system: A system whose data is to be backed up.
- Secondary system: A system to which data is backed up.
- Primary system qtree: A qtree on a primary system whose data is backed up to a secondary qtree on a secondary system.
- Secondary system qtree: A qtree on a secondary system to which data from a primary qtree on a primary system is backed up.
- Open systems platform: A server running IBM AIX®, Solaris, HP-UX, Red Hat Linux, SUSE Linux, or Windows, whose data can be backed up to a SnapVault secondary system.
- Open Systems SnapVault agent: A software agent that enables the system to back up its data to a SnapVault secondary system.
- SnapVault relationship: The backup relationship between a qtree on a primary system or a directory on an open systems primary platform and its corresponding secondary system qtree.
- SnapVault Snapshot copy: The backup images that SnapVault creates at intervals on its primary and secondary systems SnapVault Snapshot copies capture the state of primary qtree data on each primary system. This data is transferred to secondary qtrees on the SnapVault secondary system. The secondary system creates and maintains versions of Snapshot copies of the combined data for long-term storage and possible restore operations.

- SnapVault Snapshot basename: A name that you assign to a set of SnapVault Snapshot copies using the SnapVault `snap sched` command. As incremental Snapshot copies for a set are taken and stored on both the primary and secondary systems, the system appends a number (0, 1, 2, 3, and so on) to the basenames to track the most recent and earlier Snapshot updates.
- SnapVault baseline transfer: An initial complete backup of a primary storage qtree or an open systems platform directory to a corresponding qtree on the secondary system.
- SnapVault incremental transfer: A follow-up backup to the secondary system that contains only the changes to the primary storage data between the current and last transfer actions.

17.1.1 What is new in 8.2

Starting with Data ONTAP 8.2, you need not use separate licenses to enable SnapVault primary and SnapVault secondary. You must use the same SnapVault license to enable both the SnapVault primary and SnapVault secondary.

17.2 Business applications of SnapVault

SnapVault software is a reliable and economical way to protect enterprise data, and it offers many significant advantages over traditional backup methods. Although SnapVault can be deployed in configurations designed to emulate the legacy backup methods it replaces, the full value of the solution can be realized only by making a significant shift in the way you think about backup and recovery. SnapVault is so useful that it renders many common backup policies and schedules obsolete.

This chapter provides an overview of SnapVault, focusing on the differences between SnapVault and traditional backup applications. In particular, it covers some of the special benefits that are unique to SnapVault. Figure 17-2shows how SnapVault works.

17.3 Overview of SnapVault

SnapVault is a separately licensed feature in Data ONTAP that provides disk-based data protection for storage systems. The SnapVault server runs on the IBM System Storage N series platform. However, you can use an IBM System Storage N series storage system as a SnapVault client as well.

SnapVault replicates selected Snapshots from multiple client storage systems to a common Snapshot on the SnapVault server, which can store many Snapshots. These Snapshots on the server have the same function as regular tape backups (Figure 17-2). Periodically, data from the SnapVault server can be dumped to tape for extra security.

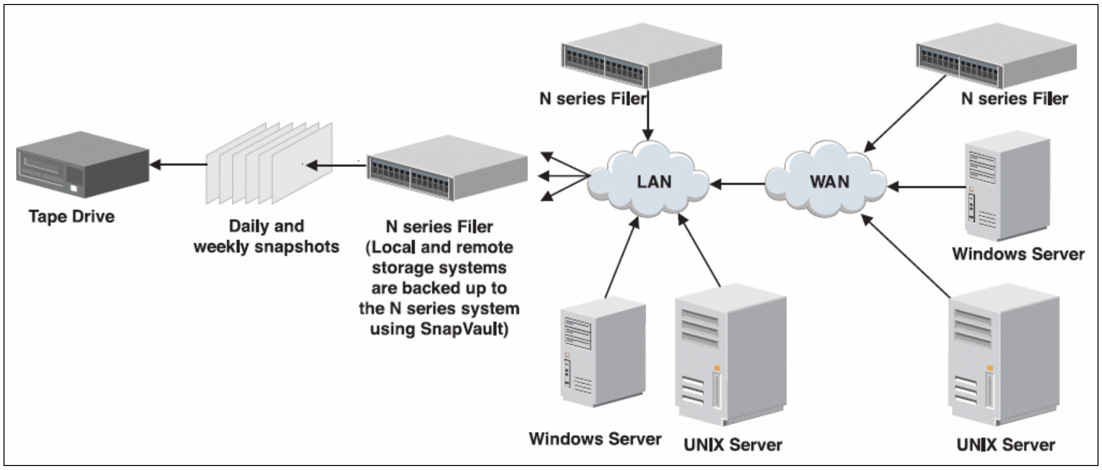

N series Filer
N series Filer
Tape Drive
Daily and weekly snapshots
N series Filer (Local and remote storage systems are backed up to the N series system using SnapVault)
LAN
WAN
Windows Server
Windows Server
UNIX Server
UNIX Server

Figure 17-2 SnapVault overview

IBM N series SnapVault is a heterogeneous disk-to-disk protection solution ideal for use with IBM N series filers and heterogeneous OS systems (Windows, Linux, Solaris, HPUX and AIX). SnapVault uses Snapshot technology to take point-in-time Snapshot and store them as online backups. In event of data loss or corruption on a filer, the backup data can be restored from the SnapVault filer with less downtime. It has significant advantages over traditional tape backups:

- It reduces backup windows versus traditional tape-based backup.
- You can expect media cost savings.
- There are no backup/recovery failures due to media errors.
- Recovery of corrupted or destroyed data is simple and fast.

SnapVault consists of major two entities, SnapVault clients and a SnapVault storage server. A SnapVault client (IBM N series filers and UNIX/Windows servers) is the system whose data must be backed-up. The SnapVault server is an IBM N series storage system which gets the data from clients and backs up data. SnapVault protects data on a client system by maintaining a number of read-only versions (Snapshots) of that data on a SnapVault filer. The replicated data on the SnapVault server system can be accessed via NFS or CIFS. The client systems can restore entire directories or single files directly from the SnapVault filer.

SnapVault requires a primary and secondary license. A SnapVault primary system corresponds to a backup client in the traditional backup architecture. The SnapVault secondary is always an IBM System Storage N series storage system running Data ONTAP. SnapVault software protects data residing on a SnapVault primary.

All of this heterogeneous data is protected by maintaining online backup copies (Snapshot) on a SnapVault secondary system. The replicated data on the secondary system can be accessed through Network File System (NFS) or Common Internet File System (CIFS), just as regular data can be. The primary systems can restore entire directories or single files directly from the secondary system. There is no corresponding equivalent to the SnapVault secondary in the traditional tape-based backup architecture.

17.4 Benefits of using SnapVault

The following section explains the benefits of utilizing SnapVault in a production environment for data protection.

17.4.1 Incremental backups forever

A *full backup* copies the entire data set to a backup medium, which is tape in traditional backup applications, or an N series near-line enabled system when using SnapVault. An *incremental backup* copies only the changes in a data set. Because incremental backups take less time and consume less network bandwidth and backup media, they are less expensive. Of course, because an incremental backup contains only the changes to a data set, at least one full backup is required in order for an incremental backup to be useful.

Traditional backup schedules involve a full backup once per week or once per month and incremental backups each day. There are two reasons why full backups are done so frequently:

- Reliability: Because a full backup is required to restore from an incremental backup, failure to restore the full backup due to media error or other causes renders all of the incremental backups useless when restoring the entire data set. Tapes used in traditional backup applications are offline storage; you cannot be sure that the data on the tape is readable without placing the tape in a drive and reading from it. Even if each piece of tape is individually read back and verified after being written, it can still fail after being verified, but before being restored.

  This problem is usually solved by taking full backups more frequently, and by duplicating backup tapes. Duplication of backup tapes serves several purposes, including providing an off-site copy of the backup and providing a second copy in case one copy is corrupted. However, for certain types of problems it is possible that the corrupted data will simply be copied to both sets of tapes.
- Speed of recovery: In order to restore a full data set, a full backup must be restored first, and possibly one or more incremental backups. If full backups are performed weekly and incremental backups daily, restores typically involve a level-zero restore and up to six incremental restores. If you perform fewer full backups and more incrementals, restoring a full data set will take considerably longer.

SnapVault addresses both of these issues. It ensures backup reliability by storing the backups on disk in a WAFL file system. Backups are protected by RAID, block checksums, and periodic disk scrubs, just like all other data on an IBM System Storage N series. Restores are simple because each incremental backup is represented by a Snapshot, which is a point-in-time copy of the entire data set, and is restored with a single operation.

For these reasons, only the incremental changes to a data set ever need to be backed up after the initial baseline copy is complete. It reduces load on the source, network bandwidth consumption, and overall media costs.

17.4.2 Self-service restores

One of the unique benefits of SnapVault is that users do not require special software or privileges to perform a restore of their own data. Users who want to restore their own data can do so without the intervention of a system administrator, saving time and money.

Important: When trying to restore from a SnapVault secondary, connectivity to the secondary must be in place and users must know where it is located to avoid data corruption and overwrite files in the wrong place.

Restoring a file from a SnapVault backup is simple. Just as the original file was accessed via an NFS mount or CIFS share, the SnapVault secondary can be configured with NFS exports and CIFS shares. As long as the destination qtrees are accessible to the users, restoring data from the SnapVault secondary is as simple as copying from a local Snapshot.

Users can restore an entire data set the same way, assuming that the appropriate access rights are in place. However, SnapVault provides a simple interface to restore an entire data set from a selected Snapshot using the **`snapvault restore`** command on the SnapVault primary (Example 17-1). This command must be used only by the filers' administrators because it can overwrite data.

Example 17-1 SnapVault restore syntax

```
itsotuc1> snapvault restore
usage:
snapvault restore [-f] [-s <snapname>] [-k <n>] -S <secondary_filer>:<secondary_
path> [<primary_filer>:]<primary_path>
```

Tip: When you use **`snapvault restore`**, the command prompt does not return until the restore has completed. If the restore needs to be cancelled, press Ctrl-c.

17.4.3 Consistent security

A common statement in the computer security community is that backups are "a reliable way to violate file permissions at a distance." With most common backup methods, the backup copy of the data is stored in a format that is usable by anyone with a copy of the appropriate backup software. Access controls can be implemented by the backup software, but they cannot be the same as the access controls on the original files.

SnapVault stores backup copies of the data in a WAFL file system, which replicates all of the file permissions and access control lists held by the original data. Users who are not authorized to access a file on the original file system are not authorized to access the backup copies of that file. It allows the self-service restores described earlier to be performed safely.

SnapVault usage provides the following benefits:

- It avoids the bandwidth limitations of tape drives, so restore can be faster.
- It does not require full dumps from the primary storage, so there is no need for a backup window.
- It is a data protection solution for heterogeneous storage environments.
- It performs disk-to-disk backup and recovery.
- It utilizes the incrementals forever model.
- It is designed to address *pain points* associated with tape:
  - Intelligent data movement reduces:
    - Network traffic
    - Impact on production systems
  - Frequent backups ensure superior data protection.
  - It uses Snapshot technology, which significantly reduces the amount of backup media.
- It provides reduced backup impact: Incrementals only, changed blocks only.
- It provides instant single file restore: The Snapshot directory displays SnapVault Snapshots.
- It can protect remote sites over a WAN.

17.5 SnapVault operation

SnapVault protects data on a SnapVault primary system (called a *SnapVault client* in earlier releases) by maintaining a number of read-only versions of data on a SnapVault secondary system (called a *SnapVault server* in earlier releases) and the SnapVault primary. The SnapVault secondary is always a data storage system running Data ONTAP.

First, a complete copy of the data set is sent across the network to the SnapVault secondary. This initial, or *baseline*, transfer can take a long time to complete because it is duplicating the entire source data set on the secondary, much like a level-0 backup to tape. Each subsequent backup transfers only the data blocks that have changed since the previous backup.

When the initial full backup are performed, the SnapVault secondary stores the data in a WAFL file system and creates a Snapshot image of the volume for the data being backed up. A Snapshot is a read-only, point-in-time version of a data set. SnapVault creates a new Snapshot with every transfer, and allows retention of a large number of copies according to a schedule configured by the backup administrator. Each copy consumes an amount of disk space proportional to the differences between it, and the previous copy.

For example, if SnapVault backed up a 100 GB data set for the first time, it consumed 100 GB of disk space on the SnapVault secondary. Over the course of several hours, users change 10 GB of data on the primary file system. When the next SnapVault backup occurs, SnapVault writes the 10 GB of changes to the SnapVault secondary, and creates a new Snapshot. Now, the SnapVault secondary contains two Snapshot copies: One contains an image of the file system as it appeared when the baseline backup occurred, and the other contains an image of the file system as it appeared when the incremental backup occurred. The copies consume a combined total of 110 GB of space on the SnapVault secondary.

Figure 17-3 shows how this process works. Blue arrows indicate the first copy, and red arrows indicate the updated data copy in a second moment.

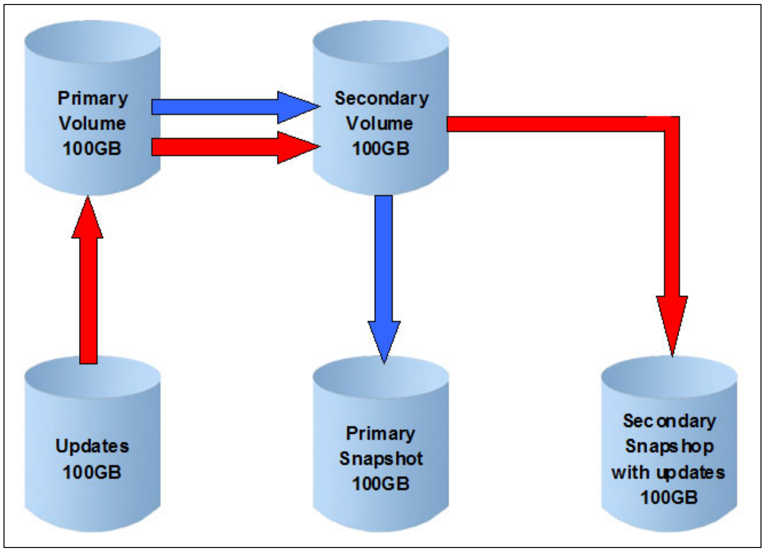


Figure 17-3 SnapVault to secondary

17.5.1 Snapshots, volumes, and qtrees

A *quota tree*, or *qtree*, is a logical unit used to allocate storage (Figure 17-4). The system administrator sets the size of a qtree and the amount of data that can be stored in it, but it can never exceed the size of the volume that contains it.

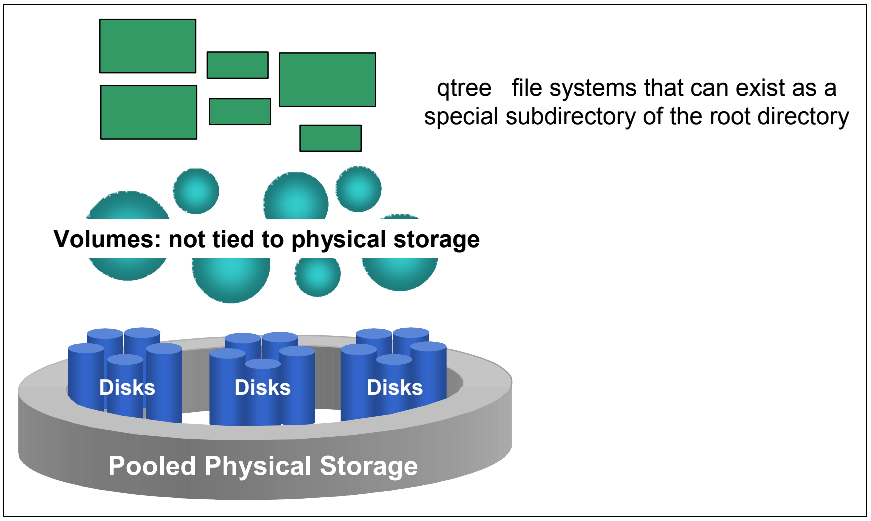

qtree file systems that can exist as a special subdirectory of the root directory
Volumes: not tied to physical storage
Disks
Disks
Disks
Pooled Physical Storage

Figure 17-4 The qtree concept

The smallest granularity for SnapVault is a qtree; each qtree can contain different application data, have different users, and have different scheduling needs. However, the SnapVault Snapshot creations and schedules of a SnapVault transfer per volume. Because the scheduling is on a volume level, when you create volumes on the secondary, be sure to group like qtrees (qtrees that have similar change rates and identical transfer schedules) into the same destination volume.

> **Storage:** Qtrees represent the third level at which filer storage can be partitioned. Disks are organized into aggregates, which provides pools of storage. In each aggregate, one or more flexible volumes can be created. Traditional volumes can also be created directly without the previous creation of an aggregate. Each volume contains a file system. Finally, the volume can be divided into qtrees.

A *volume* is a logical storage unit composed of a number of RAID groups. The space available within a volume is limited by the size and number of disks used to build the volume. A Snapshot is a read-only, point-in-time version of an entire volume. It contains images of all the qtrees within the volume.

When you start protecting a qtree using the **`snapvault start`** command, a Snapshot is created on the volume that contains the qtree you want to back up. The SnapVault primary reads the image of the qtree from this copy and transfers it to the SnapVault secondary.

Each time a SnapVault incremental backup occurs, the SnapVault primary compares the previous copy with the current copy and determines which data blocks changed and need to be sent to the SnapVault secondary. The SnapVault secondary writes these data blocks to its version of the qtree. When all qtrees in the secondary volume have been updated, a Snapshot is taken to capture and retain the current state of all the qtrees. After this copy has been created, it is visible for restoring data.

This mechanism effectively combines data from multiple Snapshots on multiple primaries into a single copy on the SnapVault secondary. But, it is important to remember that SnapVault does not transfer Snapshot copies; it only transfers selected data from within copies.

17.5.2 SnapVault example

With an IBM System Storage N series, SnapVault can be configured using the command-line interface.

The following example shows how to set up SnapVault between two separate IBM System Storage N series nodes. In the example, we consider two IBM System Storage N series: *itsotuc1* and *itsotuc2*.

The home directories are in a qtree on *itsotuc1,* called */vol/vol1/users,* and the database is on *itsotuc1,* in the volume called */vol/oracle*.

Perform the following steps:

1. Telnet to *itsotuc1* and *itsotuc2.*
2. License SnapVault.
3. Enable SnapVault on both systems (Example 17-2).

Example 17-2 Install SnapVault license

```
itsotuc1>license add ABCDEFG
itsotuc1>options snapvault.enable on
itsotuc1>options snapvault.access host=itsotuc2

itsotuc2>license add HIJKLMN
itsotuc2>options snapvault.enable on
itsotuc2>options snapvault.access host=itsotuc1
```

> **Tip:** Licenses on primary and secondary filers must be activated. Use the **`sv_ontap_pri`** license for the primary system and the **`sv_ontap_sec`** license for the secondary system.

4. Schedule Snapshot copies on the SnapVault primary, *itsotuc1* (Example 17-3).

Example 17-3 Schedule Snapshot copies on the SnapVault primary

```
a.
itsotuc1>snap sched vol1 0 0 0
itsotuc1>snap sched oracle 0 0 0
b.
itsotuc1>snapvault snap sched vol1 sv_hourly 22@0-22
c.
itsotuc1>snapvault snap sched oracle sv_daily 7@23
```

   a. Turn off the normal Snapshot schedules, which are replaced by SnapVault Snapshot schedules.

   b. Set up schedules for the home directory hourly Snapshots.

      This schedule takes a Snapshot every hour, except for 11 p.m. It keeps nearly a full day of hourly copies, and combined with the daily, or weekly backups at 11 p.m., ensures that copies from the most recent 23 hours are always available.

   c. Set up schedules for the oracle directory daily Snapshots.

      This schedule takes a Snapshot once each night at 11 p.m. and retains the seven most recent copies.

5. Schedule Snapshots on the SnapVault secondary, *itsotuc2* (Example 17-4).

Example 17-4 Schedule Snapshot copies on SnapVault secondary

```
a.
itsotuc2> aggr create sv_flex 10
itsotuc2> vol create vault sv_flex 10g
b.
itsotuc2>snap sched vault 0 0 0
c.
itsotuc2>snapvault snap sched -x vault sv_hourly 4@0-22
d.
itsotuc2>snapvault snap sched -x vault sv_daily 12@23@sun-fri
e.
itsotuc2>snapvault snap sched vault sv_weekly 13@23@sat
```

a. Create a FlexVol for use as a SnapVault destination.

b. Turn off the normal Snapshot schedules, which are replaced by SnapVault Snapshot schedules.

c. Set up schedules for the hourly backups.

   This schedule checks all primary qtrees backed up to the vault volume once per hour for a new Snapshot called sv_hourly.0. If it finds a copy, it updates the SnapVault qtrees with new data from the primary, and then takes a Snapshot on the destination volume, called sv_hourly.0.

d. Set up schedules for the daily backups.

   This schedule checks all primary qtrees backed up to the vault volume once each day at 11 p.m. (except on Saturdays) for a new Snapshot called sv_daily.0. If it finds a copy, it updates the SnapVault qtrees with new data from the primary, and then takes a Snapshot on the destination volume, called sv_daily.0.

e. Set up schedules for the weekly backups.

   This schedule creates a Snapshot of the vault volume at 11 p.m. each Saturday for a new Snapshot called sv_weekly.0. There is no need to create the weekly schedule on the primary. Because you have all the data on the secondary for this Snapshot, create and retain the weekly copies on the secondary only.

6. Perform the initial baseline transfer.

   At this point, you have configured schedules on both the primary and secondary systems, and SnapVault is enabled and running. However, SnapVault does not know which qtrees to back up, or where to store them on the secondary. Snapshots are taken on the primary, but no data is transferred to the secondary.

   To provide SnapVault with this information, use the **snapvault start** command on the secondary (Example 17-5).

Example 17-5 Perform baseline transfer

```
itsotuc2> snapvault start -S itsotuc1:/vol/vol1/users /vol/vault/itsotuc1_users
Snapvault configuration for the qtree has been set.
Transfer started.
Monitor progress with 'snapvault status' or the snapmirror log.

itsotuc2> snapvault start -S itsotuc1:/vol/oracle/ /vol/vault/oracle
Snapvault configuration for the qtree has been set.
Transfer started.
Monitor progress with 'snapvault status' or the snapmirror log.
```

17.5.3 Special case: Database and application server backups

Simply scheduling a Snapshot on a database volume might not create a safe, consistent image of the database. Most databases, such as Oracle and DB2, can be backed up while they continue to run and provide service, but they must first be put into a special hot backup mode. Other databases need to be quiesced (which means that they momentarily stop providing service), and some need to be shut down completely, enabling a cold backup.

In any of these cases, you must take certain actions before and after the Snapshot is created on the database volume. These are the same steps that you need to take for any other backup method, so your database administrators probably already have scripts that perform these functions.

Although you can set up SnapVault Snapshot schedules on such a volume and simply coordinate the appropriate database actions by synchronizing the clocks on the storage systems and database server, it is easier to detect potential problems if the database backup script creates the Snapshots using the **`snapvault snap create`** command.

In this example, you want to take a consistent image of the database every four hours, keeping the most recent day's worth of Snapshots (six copies), and you want to retain one version per day for a week. On the SnapVault secondary, you will keep even more versions.

Perform the following steps:

1. Provide SnapVault with the names of the Snapshot copies to use and how many copies to keep. No schedule needs to be specified, because all Snapshot creations will be controlled by the database backup script.

   This schedule takes a Snapshot called sv_hourly, and retains the most recent five copies, but does not specify when to take the copies (Example 17-6).

   Example 17-6 Hourly Snapshot

   ```
   itsotuc-pri> snapvault snap sched oracle sv_hourly 5@-
   ```

   This schedule takes a Snapshot called sv_daily, and retains only the most recent copy. It does not specify when to take the copy (Example 17-7).

   Example 17-7 Snapshot with no time specification

   ```
   itsotuc-pri> snapvault snap sched oracle sv_daily 1@-
   ```

2. Write the database backup script. In most cases, the script has the structure shown in Example 17-8.

   Example 17-8 Database backup script

   ```
   [ first commands to put the database into backup mode ]
   rsh itsotuc-pri snapvault snap create oracle sv_hourly
   [ end with commands to take the database out of backup mode ]
   ```

3. Use a scheduling application (such as cron on UNIX systems or the Windows Task Scheduler program) to take an sv_hourly Snapshot each day at every hour other than at 11 p.m. A single sv_daily copy will be taken each day at 11 p.m., except on Saturday evenings, when a sv_weekly copy will be taken instead.

   In most cases, it is entirely practical to run such a database backup script every hour because the database needs to be in backup mode for only a few seconds while the script creates the Snapshot.

17.5.4 Special case: Backup of FCP or iSCSI LUNs

Backing up logical units (LUNs) used by Fibre Channel Protocol (FCP) or iSCSI hosts presents the same issues as backing up databases. You must take steps to ensure that the Snapshots taken represent consistent versions of the user data.

If the LUN is being used as raw storage for a database system, then the steps to be taken are *exactly* the same as described in Example 17-8 on page 265.

If the LUN is being used as storage for a file system, such as UFS, NTFS, or VxFS, the steps to take depend on the file system. Some file systems have commands or APIs to synchronize and quiesce the file system, while others might require that the file system be unmounted or disconnected prior to taking the Snapshot. In some cases, certain logging file systems might not require any action at all, but it is rare.

In addition to the backup steps for the file system, it is important to take any steps required by applications that use the file system as well.

Finally, if you are backing up LUNs via SnapVault, consider turning space reservations on for the SnapVault secondary volume. Enabling space reservation allows writes to the LUN in case of an event where the amount of the data needed to be retained is greater than the available space in the LUN. Example 17-9 shows how to enable space reservation depending on the used data structure. For example, if you have a 10 GB LUN on the primary IBM System Storage N series storage system and rewrite all 10 GB, the next SnapVault transfer sends all 10 GB. The SnapVault transfer does not fail because it utilizes the 10 GB space reservation to complete those writes. SnapVault cannot delete the 10 GB that was overwritten because it is still required for the previous Snapshot.

Example 17-9 Space reservation syntax

```
qtree level: qtree reservation qtree_path [enable|disable]
file level: file reservation file_name [enable|disable]
LUN level: lun set reservation lun_path [enable|disable]
```

Here is a summary of how SnapVault works:

1. Administrators set up the backup relationship, backup schedule, and retention policy.

   Multiple qtrees or open system directories can be backed up to the same volume if they have the same schedule and retention policy.

2. The backup schedule begins the backup job.

   The backup job can involve backing up multiple SnapVault primaries.

3. Data moves from the SnapVault primary to the SnapVault secondary:

   Incremental forever backup after the initial level 0 transfer.

   – Storage systems transfer changed blocks to the SnapVault secondary.

   – Open systems transfer changed files to the SnapVault secondary.

4. Upon successful completion of a backup job, the IBM System Storage N series storage system takes a Snapshot on the secondary IBM System Storage N series storage system.

   SnapVault only saves changed blocks on the SnapVault secondary.

5. SnapVault maintains Snapshots on the SnapVault secondary based on retention policy.

6. A third level of protection is provided by a SnapMirror SnapVault secondary to a remote location for disaster recovery (optional).

7. A traditional backup application can be used to back up the SnapVault secondary to tape. Figure 17-5 shows the SnapVault backup flow.

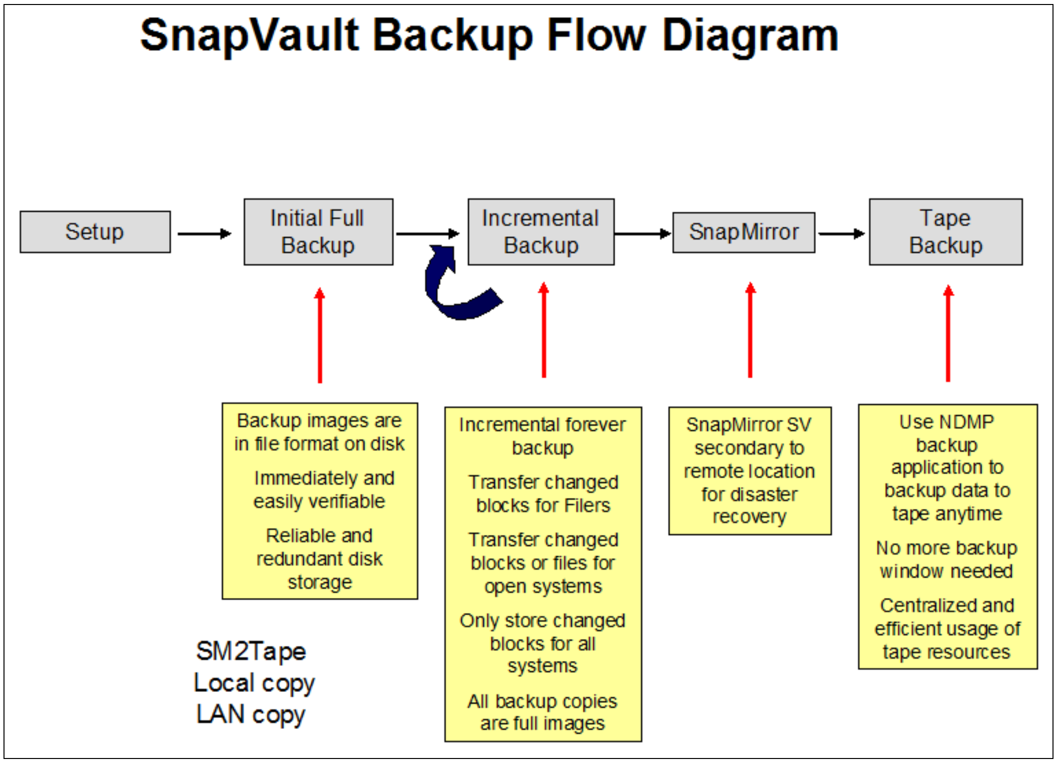

SnapVault Backup Flow Diagram
Setup
Initial Full Backup
Incremental Backup
SnapMirror
Tape Backup
Backup images are in file format on disk
Immediately and easily verifiable
Reliable and redundant disk storage
Incremental forever backup
Transfer changed blocks for Filers
Transfer changed blocks or files for open systems
Only store changed blocks for all systems
All backup copies are full images
SnapMirror SV secondary to remote location for disaster recovery
Use NDMP backup application to backup data to tape anytime
No more backup window needed
Centralized and efficient usage of tape resources
SM2Tape
Local copy
LAN copy

Figure 17-5 Backup flow diagram

After the simple installation of the SnapVault agent on the desired primary file and application servers, the SnapVault secondary system requests initial baseline image transfers from the primary storage system. This initial (or baseline) transfer can take some time to complete, because it is duplicating the entire source data set on the secondary, much like a level-0 backup to tape.

SnapVault protects data on a SnapVault primary system by maintaining a number of read-only versions of that data on a SnapVault secondary system. These transfers establish SnapVault relationships between the primary qtrees or directories and the SnapVault secondary qtrees. The baseline transfer is typically the most time-consuming process of the SnapVault implementation because it is duplicating the entire source data set on the secondary, much like a full backup to tape. You can use the **snapvault status** command to monitor the transfer progress, as shown in Example 17-10.

Example 17-10 SnapVault status

```
itsotuc2> snapvault status -l
Snapvault secondary is ON.

Source:                 itsotuc1:/vol/vol1/users
Destination:            itsotuc2:/vol/vault/itsotuc
Status:                 Idle
Progress:               -
State:                  Snapvaulted
Lag:                    00:16:28
Mirror Timestamp:       Wed Apr  6 15:03:10 MST 2011
Base Snapshot:          itsotuc2(0135019083)_vault-base.5
```

```
Current Transfer Type:  -
Current Transfer Error: -
Contents:               Replica
Last Transfer Type:     Resync
Last Transfer Size:     8 KB
Last Transfer Duration: 00:00:12
Last Transfer From:     itsotuc1:/vol/vol1/users

Source:                 itsotuc1:/vol/oracle
Destination:            itsotuc2:/vol/vault/oracle
Status:                 Idle
Progress:               -
State:                  Snapvaulted
Lag:                    00:15:48
Mirror Timestamp:       Wed Apr  6 15:03:50 MST 2011
Base Snapshot:          itsotuc2(0135019083)_vault-base.5
Current Transfer Type:  -
Current Transfer Error: -
Contents:               Replica
Last Transfer Type:     Resync
Last Transfer Size:     20 KB
Last Transfer Duration: -
Last Transfer From:     itsotuc1:/vol/oracle
```

Tip: The `snapvault status -l` command shows the state of each volume or qtree that has replication enabled. Use the command output for monitoring the first replication schedules so you can ensure that they are all working properly.

With SnapVault, one baseline transfer is required before any subsequent backups or restores can be performed, but unlike traditional tape backup environments, this initial baseline transfer is a one-time occurrence, not a weekly event. First, a complete copy of the data set is pulled across the network to the SnapVault secondary. Each subsequent backup transfers only the data blocks that have changed since the previous backup (incremental backup or incrementals forever).

When the initial full backup is performed, the SnapVault secondary stores the data in a WAFL file system and creates a Snapshot image of that data. A Snapshot is a read-only, point-in-time version of a data set. Each of these Snapshots can be thought of as full backups (although they are only consuming a fraction of the space). A new Snapshot is created each time that a backup is performed, and a large number of Snapshots can be maintained according to a schedule configured by the backup administrator. Each Snapshot consumes an amount of disk space equal to the differences between it and the previous Snapshot.

A very common scenario is data protection of the secondary system. A SnapVault secondary system can be protected by either backup to tape or backup to another disk-based system. The method used to back up to a tertiary disk-based system is volume-based SnapMirror. All Snapshots are transferred to the tertiary system and the SnapVault primaries can be directed to this tertiary system, if necessary. In order to back up an IBM System Storage N series secondary to a tape library, the SnapMirror to tape option or an NDMP backup to a tape library can be used.

17.6 SnapVault backup

SnapVault provides you with great flexibility in deciding which data, and at what granularity, to protect. The data structures that are backed up and restored through SnapVault depend on the primary storage system.

On IBM System Storage N series primary systems, the qtree is the basic unit of SnapVault backup and restore. SnapVault backs up specified qtrees on the primary system to associated qtrees on the SnapVault secondary storage system. If necessary, data is restored from the secondary qtrees back to their associated primary qtrees. Figure 17-6 shows qtree units.

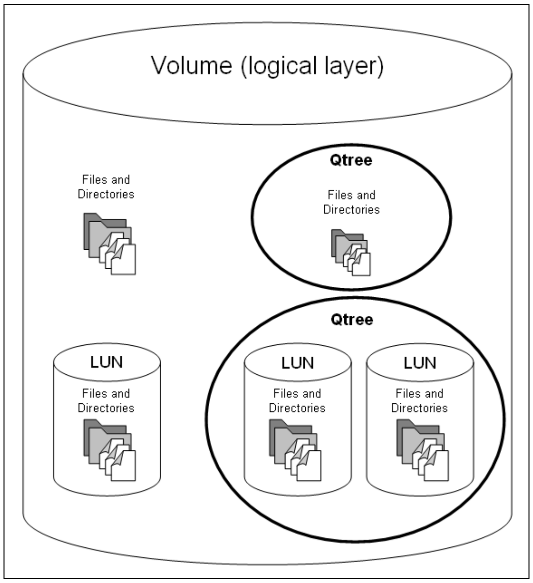


Figure 17-6 Illustration of qtrees

On open system storage platforms, the directory is the basic unit of SnapVault backup. SnapVault backs up specified directories from the native system to specified qtrees in the SnapVault secondary storage system. If necessary, SnapVault can restore an entire directory or a specified file to the primary storage platform.

17.6.1 Incremental backups (updates)

Incremental backups are updates to an existing baseline copy of the primary system's data set. The concept of obtaining and transferring primary system data to the secondary system after the baseline transfer has completed is typically referred to as *incremental backups forever*. With SnapVault, there is only one full backup (the baseline transfer) followed by incremental backups for the remainder of the qtree/directory relationships. An incremental backup occurs when the SnapVault secondary contacts the primary to update its qtrees with the latest data from the primary.

No additional full backups need to be performed after the first baseline copy has been completed. Each incremental backup and subsequent Snapshot of the data set on the secondary system can be used as full backups of the original data set except that they only consume the amount of disk space that was actually changed. The incremental backups are replications of the primary data set with the replication versions being updated as often as every hour without consuming the media that traditional tape-based architectures will require.

It is important to note that incremental backups only consume space on the secondary system for the data that has actually changed. For example, if a 10 GB file has had 100 KB of changes since the last incremental backup, the secondary server consumes only 100 KB of storage space to record that change. In other words, only the changed blocks in an updated file are stored on the secondary server. It is dramatically different from the behavior of most incremental tape backups, where the entire changed file is recorded from the incremental backup. It is a dramatic advantage in resource conservation for those deploying SnapVault in place of traditional backup architectures.

17.6.2 Scheduling/retention policy

The schedule details the frequency, number of copies to retain, date, and time to perform incremental backups for a specific SnapVault relationship. The SnapVault secondary system creates and maintains copies, based on the specified schedule, for each primary data set that it is responsible for protecting. Incremental backups can be scheduled every hour, week, or month depending on the needs of the environment, providing backup/storage administrators considerable flexibility when defining policies for data protection (Example 17-11). Use the command **snapvault snap sched** to schedule the SnapVault copies.

Example 17-11 SnapVault scheduling

```
snapvault snap sched -x vol1 sv_weekly 1@sat@19

snapvault snap sched -x vol1 sv_nightly 2@mon-fri@19

snapvault snap sched -x vol1 sv_hourly 11@mon-fri@7-18
```

> **Tip:** The option **-x** in the **snapvault snap sched** command indicates that the secondary must transfer data from the primaries before creating the secondary Snapshot.

In normal operation, updates and Snapshot creation proceed automatically according to the Snapshot schedule. However, SnapVault also supports manual operation through basic command-line and management interface operations, thus allowing for customer on demand application-level integration for specific applications/servers that require an application/event-driven backup capability.

All of these scheduling options result in the capability to increase the frequency of backups, without an increased requirement of baseline transfers and media cost. In most cases, traditional backup windows can be reduced or even eliminated after the initial full backup has been performed.

17.6.3 Snapshot copies

The backup data is stored in Snapshots on the secondary system in the RAID-protected IBM System Storage N series WAFL file system. A Snapshot backup is a read-only, point-in-time version of a data set. Each time that a backup is performed, a new Snapshot is created. Up to 250 Snapshot copies can be maintained on a SnapVault secondary system.

When creating a new Snapshot backup, SnapVault deletes the oldest Snapshot, renames the remaining copies as appropriate, and then creates a new base Snapshot backup. The data is readily available and safely stored on disk.

Most organizations then make a tape copy from the Snapshot, or better yet, replicate the Snapshot (through SnapMirror) to an off-site facility where tape copies are created. If multiple systems are being backed up, the transfers are synchronized so that all transfers are completed at the same time, enabling multiple backups to share a Snapshot duplicate, thereby preserving Snapshots for archiving on the secondary system.

Each Snapshot backup consumes an amount of disk space equal to the amount of data changed during the period between its creation and the creation of the previous Snapshot backup. As stated earlier, only data that has changed (not entire files that have changed) is saved. Snapshot allows you to keep hundreds of full backup equivalent images of the source data set, in a minimum amount of space. This boils down to each Snapshot representing a full backup of the primary storage data set minus the space requirements of typical full backups. Full backup versioning is achieved and, best of all, the data is online for immediate access for restore and recovery without sorting through hundreds of tape cartridges.

17.7 SnapVault details

In this section, we take a detailed look at SnapVault characteristics and associated benefits:

- Speed of recovery:

  To recover a full data set, you must first recover a full backup, and then recover each incremental backup, in order. If you perform full backups weekly and incremental backups daily, restores will typically involve a full restore and up to six incremental restores.

  > **Tip:** If you perform fewer full backups and more incrementals, restoring a full data set takes considerably longer.

  SnapVault ensures backup reliability by storing the backups on disk in the WAFL file system. These backups are protected by RAID, block check-sums, and periodic disk scrubs, just like all other data on a N series storage system. Restores are simple because each incremental backup is represented by a Snapshot. It is shown in Figure 17-7, which is a point-in-time view of the entire data set that can be restored in a single operation, eliminating the need to manage large quantities of tape cartridges.

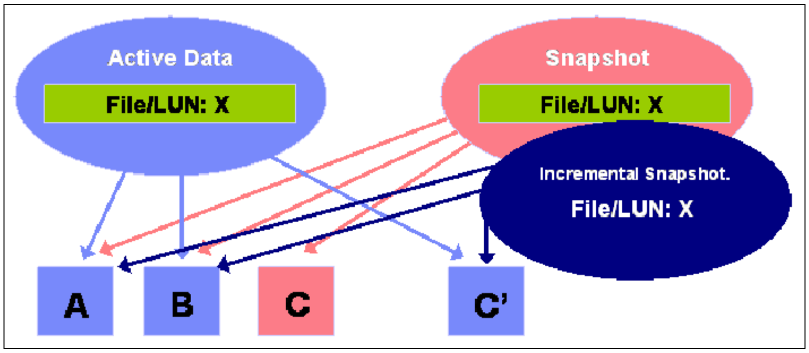


Figure 17-7 Incremental Snapshot

- Simplicity of restores:

  A unique benefit of SnapVault is that users do not require special software or privileges to perform a restore of their own data. Users who want to perform a restore of their own data can do so without the intervention of a system administrator, thus saving user time and resources and freeing up valuable administrator time. If required by policies, data recovery can be restricted to authorized individuals, as well.

  Recovering a file from a SnapVault backup is simple. Just as the original file was accessed through an NFS mount or CIFS share, the SnapVault secondary can be configured with NFS exports and CIFS shares. As long as the destination qtrees are accessible to the users, restoring data from the SnapVault secondary is as simple as copying from a local Snapshot image. Restores can be performed by drag-and-drop or a simple copy command, depending on the environment. If SnapVault has been deployed in an open systems environment, the restore process can be initiated directly from the primary system that was backed up through the command line.

  Recovery of an entire data set can be performed the same way if the user has appropriate access rights. SnapVault provides a simple interface to recover an entire data set from a selected Snapshot using the **snapvault restore** command (Example 17-12).

Example 17-12 SnapVault restore syntax

```
itsotuc1> snapvault restore
usage:
snapvault restore [-f] [-s <snapname>] [-k <n>] -S
<secondary_filer>:<secondary_
path> [<primary_filer>:]<primary_path>
```

A user can recover the complete contents of a secondary qtree/directory back to the primary with the **snapvault restore** command on the primary. The primary data set will be read-only until the transfer completes, at which time it becomes writable.

After a restore, the user can choose to resume backups from the recovered data set to the secondary qtree from which it was recovered. When used alone, SnapVault creates hourly, read-only Snapshots on the secondary. Thus, restores are done through a copy back. Because each Snapshot refers to a complete point-in-time image of the entire file system, the restore time is zero. There is no tape or incremental *unwind*.

> **Attention:** Restoring data with **snapvault** commands is easy, but users can find that confusing. It is best that system administrators own the restore commands and create CIFS or NFS shares to users so they can choose which file or folder they need to restore and copy over to the production location.

In comparison, recovery from tape can consume considerable resources. Single files can sometimes be recovered by users, but are typically recovered by an administrator. The tape that contains the backup file must be located (sometimes retrieved from an off-site storage location), and the backup application must transfer the file from the tape location to the requesting host.

The backup administrator starts the tape restore process and retrieves the tape from the appropriate location if necessary. The tape must be loaded (from seven seconds up to two minutes), positioned to the correct location on the tape (usually several seconds, sometimes more), and the data read.

If a full image must be restored, data must be recovered using the last full and subsequent incremental backups. If a restore requires recovery from one full backup and all incremental backups since that last full backup, then there is more of a chance that an error might occur with the media involved in a tape solution.

This process can be long and tedious, depending on the amount of data being recovered. Hours and days can be spent during the restore process. If there is a failure during the restore, the entire process must be reinitiated, thereby significantly adding to downtime. If the data to be restored is a large critical database, users are offline during the entirety of the restore process.

- Reliability:

  Because a full backup is required in order to recover a full data set from a traditional incremental backup, failure to recover the full backup due to a media error or other causes renders all of the incremental backups useless when recovering the entire data set. Tapes used in traditional backup applications are offline storage. You cannot be sure that the data on the tape is readable without placing the tape in a drive and reading from it. Even if each piece of tape media is individually read back and verified after being written, it can still fail after being verified, but before being recovered, due to improper handling, resulting in a tremendous amount of uncertainty when increasing the number of incremental backups per full backup.

  This problem is usually solved by taking full backups more frequently and by duplicating backup tapes. Duplication of backup tapes serves several purposes, including providing an off-site copy of the backup and providing a second copy of the media in case one copy is bad. However, it is possible that the corrupt data will simply be copied to both sets of tapes. SnapVault and SnapMirror prevent this from occurring.

- Reduced time spent in backup (incremental backups forever):

  The promise of incremental backups forever is delivered with SnapVault. An incremental backup copies only the changes in a data set to a backup media. Entire files are stored in traditional backup architectures. In contrast, only changed blocks are stored in a SnapVault configuration, which results in dramatic space savings. Because incremental backups take less time and consume less network bandwidth and backup media, they are less expensive. Traditional backup schedules involve a full backup once per week or once per month and incremental backups each day.

- Space/media savings:

  Because SnapVault only requires storage of changed blocks of data, the storage requirements are much less than that of traditional backup applications, which typically store the entire changed files.

17.8 Disaster recovery with SnapVault

In traditional tape-based solutions, it is common to duplicate tapes for off-site storage, ship the tape off-site, and store them remotely for disaster recovery purposes. Making duplicate copies of the backup data allows one copy to be kept locally for restore purposes, while the other is shipped off-site for remote recovery in the event of a disaster.

17.8.1 SnapVault options

SnapVault provides several superior disaster recovery and off-site options. One option is to back up to a remote SnapVault secondary or multiple SnapVault secondaries across a wide area network (WAN) for off-site storage (Figure 17-10 on page 276).

A second option is to back up the SnapVault secondary to tape for offline storage in the event that the SnapVault secondary system is unavailable. This deployment adds a tape backup of the SnapVault secondary storage system and can serve two purposes: It enables the storage of an unlimited number of network backups offline while keeping the more recent backups available online in secondary disk storage for quick recovery if necessary (Figure 17-8).

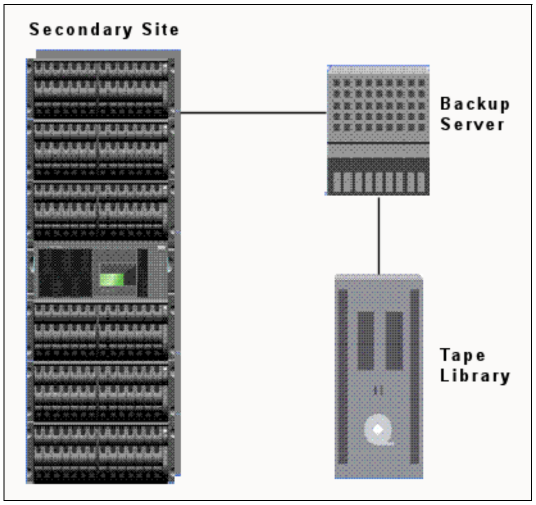


Figure 17-8 Tape backup

If a single tape backup is generated off the SnapVault secondary storage system, the IBM System Storage N series and open systems storage platforms are not subject to the performance degradation, system unavailability, or the complexity of direct tape backup of multiple systems. In this instance, tape augments the backup experience.

These options provide multiple lines of defense in the event of a disaster:

- Local Snapshot backups
- Secondary SnapVault storage
- Offline tape-archived data

Another variation to the basic SnapVault deployment protects replications stored on SnapVault secondary storage against interruption of the secondary storage system itself. The data backed up to SnapVault secondary storage is mirrored to a system configured as a SnapMirror partner, or destination. If the SnapVault secondary storage system fails, the SnapMirror destination can be converted to a secondary storage system and used to continue the SnapVault backup operation with minimum disruption to the environment.

Figure 17-9 shows the various SnapVault solutions.

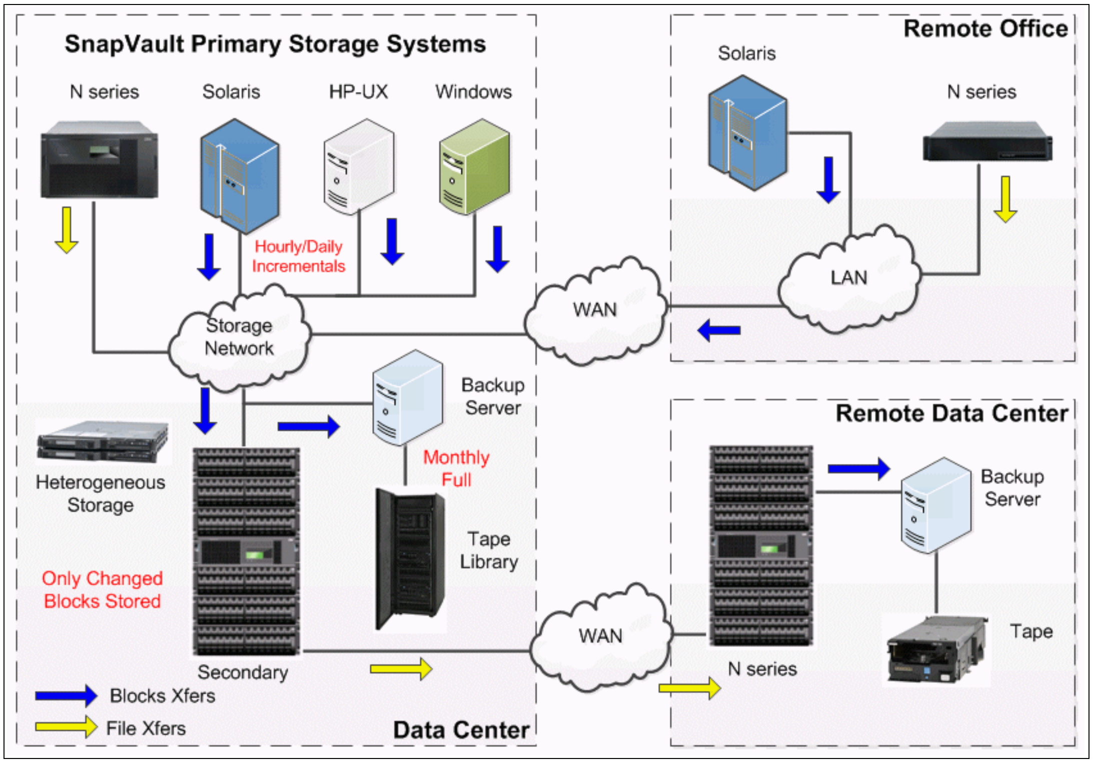


Figure 17-9 SnapVault solutions

Attention: In the SnapMirror protection architecture, both SnapMirror and SnapVault must be licensed on the secondary.

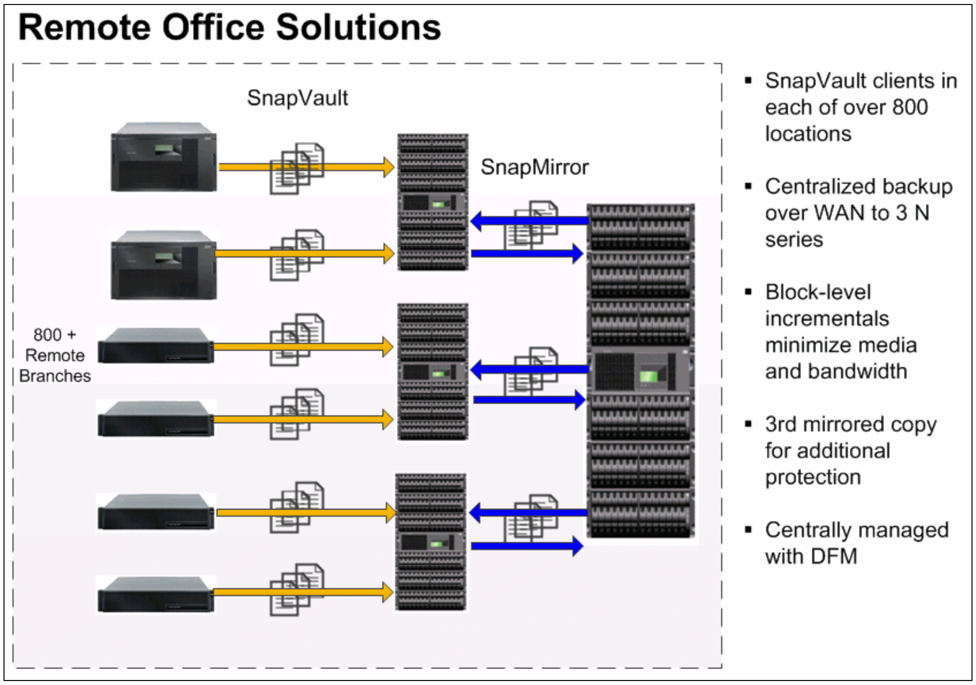
17.8.2 Comparing SnapMirror and SnapVault

In this section, we compare SnapMirror and SnapVault:

- Both SnapVault and SnapMirror use data replication.
- SnapMirror copies all Snapshots from a read/write source into a read-only destination.
- SnapVault copies the active file system data from a read/write source into a read-only destination, but protects and versions the data by creating destination Snapshots.
- SnapMirror provides up to per-minute updates.
- SnapVault provides up to per-hour updates.
- SnapMirror is homogeneous. It works on IBM System Storage N series storage systems only.
- SnapVault is heterogeneous.

17.9 Remote solution using SnapVault

SnapVault can be used to replicate data over to remote locations. Figure 17-10 shows an example of using SnapVault for local and remote backup. In this example, SnapVaults are occurring from a primary IBM System Storage N series to a secondary IBM System Storage N series. In addition, clients with agents using open systems SnapVault can also back up to the secondary system.



- SnapVault clients in each of over 800 locations
- Centralized backup over WAN to 3 N series
- Block-level incrementals minimize media and bandwidth
- 3rd mirrored copy for additional protection
- Centrally managed with DFM

Figure 17-10 Remote office solutions

Figure 17-11 shows the cost-efficiency storage of IBM System Storage N series storage systems, SnapVault can be used as a tapeless backup system. This configuration allows failover to the secondary N series system storage for access and high availability.

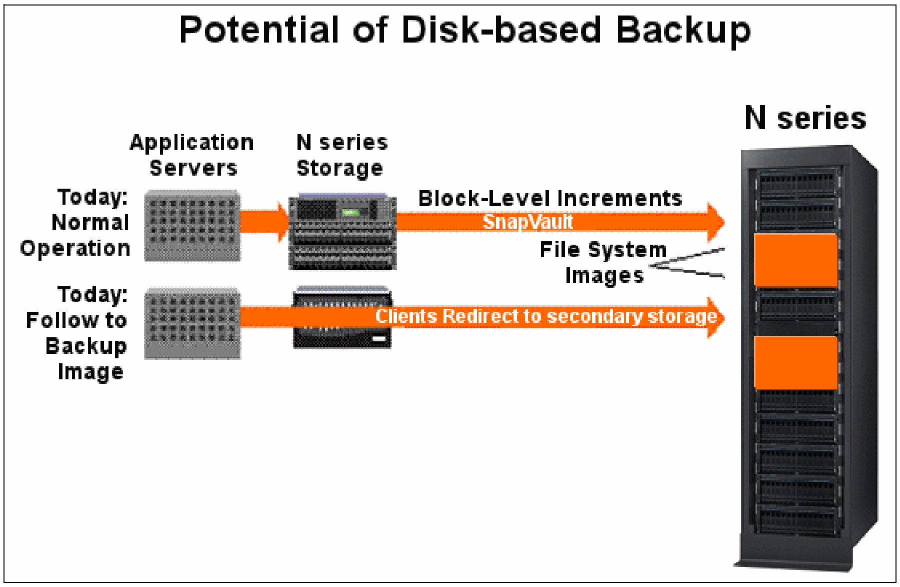


Figure 17-11 Disk-based backup

17.10 Maximum number of concurrent SnapVault targets

Starting with Data ONTAP 7.3, the maximum possible number of concurrent SnapVault transfers of individual qtrees on a storage system is greater than the maximum number in Data ONTAP 7.2. However, the maximum number of concurrent SnapVault targets has not changed between the release families.

Before Data ONTAP 7.3, the maximum number of concurrent SnapVault targets supported by a system was equal to the maximum number of concurrent SnapVault transfers possible for the system.

A SnapVault target is a process that controls the creation of a scheduled SnapVault Snapshot copy on a SnapVault destination volume. For each SnapVault destination volume that has qtrees being updated, there is a SnapVault target.

The maximum number of concurrent SnapVault targets for each platform is described in Table 17-1. At any point in time, no more than the listed number of volumes can have their qtrees updated concurrently. If the number of SnapVault targets exceeds the limit mentioned in the table, the excess SnapVault targets are queued and executed after the active SnapVault targets complete their backups.

Table 17-1 Maximum number of concurrent SnapVault targets Data ONTAP 8.1

Model	Suggested maximum number of concurrent SnapVault targets
N3240	128
N6210	128
N6240	256
N6270	256
N7950T	256

Tip: These maximum numbers apply only to SnapVault targets, and therefore to SnapVault qtrees. There is no restriction on the number of volumes that can be updated concurrently for SnapMirror qtrees.

17.11 Preferred practices

The following sections describe preferred practices and advice for implementing SnapVault. This information can be useful when planning the SnapVault deployment.

17.11.1 General preferred practices

There are many preferred practices to be aware of in order to ensure a successful SnapVault deployment. Some of the general preferred practices are explained here.

Monitoring logs

In most cases when a SnapVault transfer fails, the problem can be determined by reviewing the log file. All operations (both primary and secondary) are logged to the /etc/log/snapmirror log. This log contains various messages that can affect the scheduled transfers. In it, you can see the amount of time a SnapVault session might have been in the quiescing state, or whether it tried to roll back to the last good Snapshot in the event of a failed transfer.

Scheduling guidelines

When setting up the SnapVault schedule, first gather the following information:

1. What is the maximum size to which this qtree is expected to grow?
2. What is the estimated rate of change for this qtree in megabytes or gigabytes per day?
3. How many days of Snapshot copies must be maintained on the destination volume?

A primary consideration for grouping qtrees within destination volumes is the number of days that Snapshot copies will be retained on the destination volume. The available space on the destination volume is the secondary criterion.

Another thing to remember when setting up SnapVault schedules is that you need to disable all scheduled Snapshots that are invoked by `snap sched` on both the primary and the secondary. Also, a preferred practice is to keep all the Snapshot names the same on all primary systems, regardless of the volume, as shown in Table 17-2.

Table 17-2 Snapshot names and frequency

Snapshot name	Snapshot frequency
sv_hourly	Hourly
sv_daily	Daily
sv_weekly	Weekly

In addition to knowing which Snapshot must be used, this practice helps to determine the transfer schedule of the specific Snapshot.

When scheduling the Snapshots, make sure that you add up *all* qtrees in every volume. When adding new schedules for volumes, be sure to take into account the existing schedule. Also, when scheduling, be aware of how many Snapshots are going to be retained for the volume, including copies for SnapVault and SnapMirror.

Primary Snapshot retention

When planning your SnapVault transfer schedule, keep in mind that you can also retain SnapVault Snapshots at the primary. It might not be a requirement to keep hourly copies on the secondary, but it can be an ideal situation for primary copy retention.

In the schedule created in Example 17-11 on page 270, it was decided to keep more hourly Snapshot copies on the primary than on the secondary. The reasoning is that if you need to go back just 1 or even 10 hours, that copy must be maintained locally. This helps keep the amount of restore time lower than restoring from the secondary. It also reduces the number of transfers from the primary to the secondary and makes it easier to maintain a complex schedule. Given this scenario, you might have hourly copies on the primary but perhaps transfer only four hourly copies (one every six hours).

Changing the “tries” count

The `-t` option (tries) of the `snapvault` command sets the number of times that updates for the qtree must be tried before giving up; the default is 2. When the secondary starts creating a Snapshot, it first updates the qtrees in the volume (provided that the `-x` option was set on the Snapshot schedule). If the update fails for some reason (such as a temporary network outage), then the secondary tries the update again one minute later.

The `–t` option specifies how many times the secondary must try before giving up and creating a new Snapshot with data from all the other qtrees. When set to 0, the secondary does not update the qtree at all. It is one way to temporarily disable updates to a qtree.

If you leave this option at the default, the first attempt to update the secondary counts as the first try. In that case, SnapVault attempts only once more to update the destination before failing. If there are potential network issues, increase the number of tries for the transfer. If the tries count needs to be modified after the relationship has been set up, use the **snapvault modify** command (Example 17-13). It is useful when there is a planned network outage.

Example 17-13 SnapVault modify syntax

```
itsotuc*> snapvault modify
usage:
snapvault modify [-k <kbs>] [-t <n>] [-o <options>] [-S
[<primary_system>:]<primary_path>] [<secondary_filer>:]<secondary_path>
        where <options> is <opt_name>=<opt_value>[[,<opt_name>=<opt_value>]...]
```

Primary data layout

With FlexVols, there are alternative ways to lay out data on the SnapVault primary system, which can handle small files and millions of files. If the SnapVault primary system contains millions of files, then using a single FlexVol in place of each qtree, or even creating just one qtree in each volume, is advantageous for SnapVault performance. This minimizes the amount of scan time prior to data being sent during the SnapVault transfer.

When performing the baseline, the **snapvault start** command is still used, but a dash (**-**) is used in place of the source qtree name. The **-** signifies that SnapVault backs up all data in the volume that does not reside in a qtree. If qtrees also exist in that volume, a separate SnapVault relationship must be created for those qtrees.

Important: The non-qtree part of the primary storage system volume can be replicated only to the SnapVault secondary storage system. The data can be restored to a qtree on the primary storage system, but it *cannot* be restored as non-qtree data.

If the Data ONTAP CLI is used to perform restores, use one qtree inside the volume. Using this configuration allows restores to function like any other SnapVault restore.

17.11.2 Common misconfigurations

The following sections describe common misconfigurations that a user might encounter with SnapVault. These are issues that you need to consider prior to the deployment in order to achieve a successful SnapVault deployment.

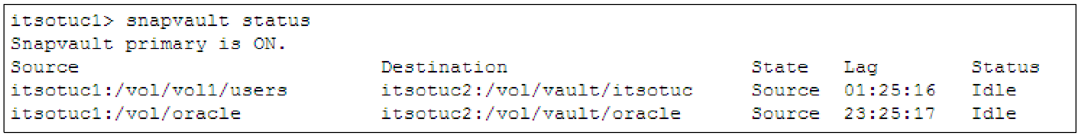
Time zones, clocks, and lag time

A scheduling consideration is that the SnapVault operations are initiated by the clock on the storage system. For example, on the primary, the Snapshots are scheduled by using the **snapvault snap sched** command. When the time for the copy to be created is reached, the primary storage system creates its copy.

On the secondary storage system, you use the **snapvault snap sched -x** command (**-x** tells the secondary to contact the primary for the Snapshot data) to schedule the SnapVault transfer. It can pose a huge problem with lag times if the clocks are not synchronized.

The following examples show the output from the **snapvault status** command.

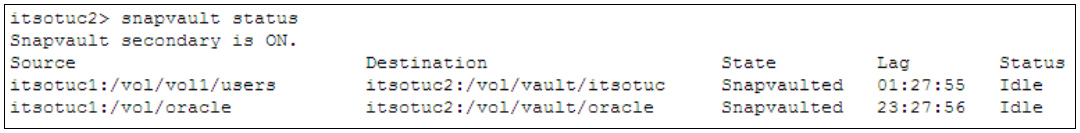
Figure 17-12 shows the output from the primary storage system.



```
itsotuc1> snapvault status
Snapvault primary is ON.
Source                        Destination                   State    Lag        Status
itsotuc1:/vol/vol1/users      itsotuc2:/vol/vault/itsotuc   Source   01:25:16   Idle
itsotuc1:/vol/oracle          itsotuc2:/vol/vault/oracle    Source   23:25:17   Idle
```

Figure 17-12 SnapVault status on primary

Figure 17-13 shows the output from the secondary storage system.



```
itsotuc2> snapvault status
Snapvault secondary is ON.
Source                        Destination                   State         Lag        Status
itsotuc1:/vol/vol1/users      itsotuc2:/vol/vault/itsotuc   Snapvaulted   01:27:55   Idle
itsotuc1:/vol/oracle          itsotuc2:/vol/vault/oracle    Snapvaulted   23:27:56   Idle
```

Figure 17-13 SnapVault status on secondary

Clocks and scheduling also come into play when the primary and secondary are in different time zones. In this case, it is important to remember that schedules are based on the local clock. Given this scenario, assume that there are two storage systems, one on the U.S. East Coast and one on the U.S. West Coast. You must ensure that schedules account for the three-hour difference, otherwise you will either have negative lag times or lag times greater than what is expected based on the schedule.

Managing the number of Snapshot copies

Each volume on the SnapVault secondary system can have up to 255 Snapshots. SnapVault software requires the use of 4 Snapshots (regardless of the number of qtrees or data sets being backed up), leaving 251 copies for scheduled or manual Snapshot creation. In most cases, fewer than 251 copies are maintained due to limitations on available disk space.

With improper scheduling, this limit can quickly be reached on the secondary storage system because SnapVault takes a Snapshot of the volume after every transfer. Again, it is important to make sure that the qtrees within a SnapVault destination have the same characteristics in order to avoid reaching the 250 copy limit.

Volume to Qtree SnapVault

When issuing the **`snapvault start`** command, you are not required to specify a qtree name for the source; however, this situation is to be avoided. This type of relationship increases the performance of the SnapVault transfer, and increases the amount of time it takes to perform a backup.

Because you must specify a qtree for the SnapVault destination, an entire volume now resides in a qtree on the destination. In the event of a restore via Data ONTAP CLI, the entire contents of the qtree, which contains all the data from the source volume, is restored to a qtree on the SnapVault primary system. Subsequently, you must copy manually the data back to the appropriate location.

Conclusion

SnapVault software can be configured and deployed with a minimum amount of time and planning to duplicate the capabilities of legacy backup solutions, while still providing several unique advantages. With some advance preparation and investigation of user needs, SnapVault can deliver data protection, backup, and recovery capabilities with orders of magnitude beyond those available with traditional solutions.

17.12 Summary

Traditional backup and recovery solutions based only on tape are difficult to maintain. Recovering data is labor-intensive, and scaling is difficult because it results in additional complexity. Utilizing disks as a tape cache is a step forward in that some of the backup data is now online. But without Snapshots, recovering data is still laborious.

SnapVault can significantly reduce the impact and bottlenecks associated with traditional backup and recovery solutions. Only the changes to the data are backed up, resulting in a very small backup window and less stress on the compute environment. The backed-up data is kept online, resulting in extremely fast recoveries.

18

What storage efficiency is

This chapter introduces storage efficiency, which includes technologies such as FlexVol volume, Snapshot copy, deduplication, SnapVault, SnapMirror, and FlexClone. These technologies help to increase storage utilization and decrease storage costs.

Storage efficiency enables you to store the maximum amount of data for the lowest cost and accommodates rapid data growth while consuming less space. IBM N series strategy for storage efficiency is based on the built-in foundation of storage virtualization and unified storage provided by its core Data ONTAP operating system and Write Anywhere File Layout (WAFL) file system.

The unified storage architecture allows you to efficiently consolidate a storage area network (SAN), network-attached storage (NAS), and secondary storage on a single platform.

High-density disk drives, such as serial advanced technology attachment (SATA) drives mitigated with RAID-DP technology, provide increased efficiency and read performance.

Technologies such as FlexVol volume, Snapshot copy, deduplication, SnapVault, SnapMirror, and FlexClone offer dramatic cost savings and performance improvements. You can use these technologies together to improve storage utilization.

The following topics are covered:

- The IBM N series advantage
- SATA storage disks and Flash Cache
- Protection against double disk failure with RAID-DP
- How space management works
- Thin provisioning using FlexVol volumes

18.1 The IBM N series advantage

IBM N series has a rich set of features such as SATA disks, Flash Cache, RAID-DP, FlexVol, Snapshot copies, deduplication, SnapVault, SnapMirror, and FlexClone, which help to achieve significant improvements in storage utilization. When used together, these technologies help to achieve increased performance.

IBM N series offers the following technologies to implement storage efficiency:

- SATA disks provide an attractive price point to reduce storage costs and optimize system performance. You can incorporate key features and products, which increase data reliability and availability, in the SATA deployment to implement a reliable storage solution.
- The Flash Cache improves performance for workloads that are random read-intensive without adding additional disk drives. This read cache helps to reduce latency and improve I/O throughput.
- RAID-DP is a double-parity RAID 6 implementation that protects against dual disk drive failures.
- Thin provisioning enables you to maintain a common unallocated storage space that is readily available to other applications as needed. It is based on the FlexVol technology.
- Snapshot copies are a point-in-time, read-only view of a data volume, which consumes minimal storage space. Two Snapshot copies created in sequence differ only by the blocks added or changed in the time interval between the two. This block incremental behavior limits the associated consumption of storage capacity.
- Deduplication saves storage space by eliminating redundant data blocks within a FlexVol volume.
- SnapVault is a centralized and cost-effective solution for replicating Snapshot copies to inexpensive secondary storage. This technology copies only the data blocks that changed since the last backup, and not the entire files, so backups run more frequently and use less storage capacity.
- SnapMirror is a flexible solution for replicating data over local area, wide area, and Fibre Channel networks. It can serve as a critical component in implementing enterprise data protection strategies. You can replicate your data to one or more storage systems to minimize downtime costs in case of a production site failure. You can also use SnapMirror to centralize the backup of data to disks from multiple data centers.
- FlexClone technology copies data volumes, files, and LUNs as instant virtual copies. A FlexClone volume is a writable point-in-time image of the FlexVol volume or another FlexClone volume. This technology enables you to use space efficiently, storing only data that changes between the parent and the clone. Use of this feature results in savings in space, power, and dollars. Additionally, clones have the same high performance as their parent.
- The unified architecture integrates multiprotocol support to enable both file-based and block-based storage on a single platform. With N series Gateways, you can virtualize your entire storage infrastructure under one interface, and you can apply all the preceding efficiencies to your non-IBM N series systems.

18.2 SATA storage disks and Flash Cache

SATA disks are low-cost, high-capacity storage solutions best suited for secondary storage applications, such as online backup and archiving. You can combine the Flash Cache with SATA disks to provide a low-cost alternative to Fibre Channel disks without sacrificing read performance.

The Flash Cache enables you to optimize the performance of mainstream storage platforms with workloads that are random read intensive, such as the file services. This intelligent read-cache technology helps to reduce latency and improve I/O throughput without adding more disk drives.

The Flash Cache module (formerly Performance Acceleration Module II) (Figure 18-1) works with deduplication to reduce the amount of redundant data it keeps in cache.

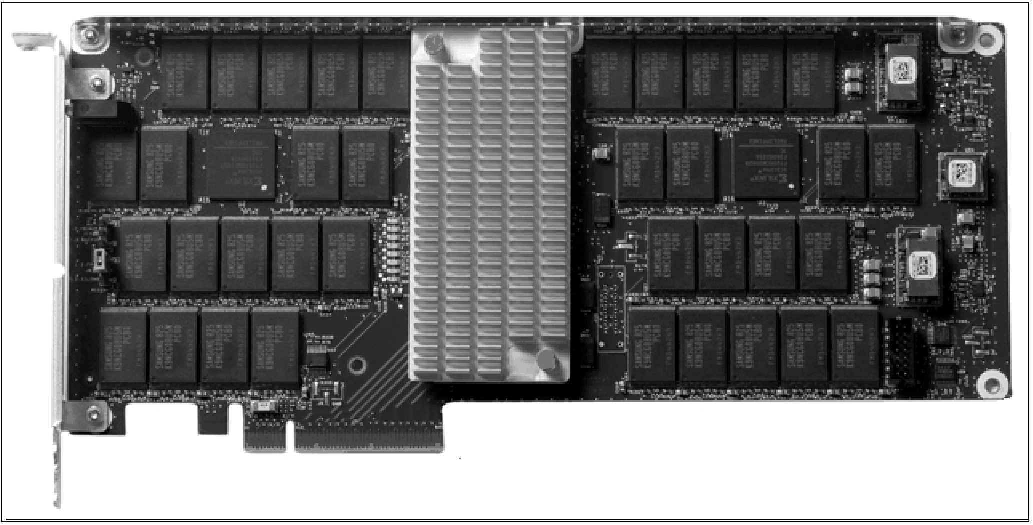


Figure 18-1 Flash Cache (PAM II)

18.3 Protection against double disk failure with RAID-DP

RAID-DP is a standard Data ONTAP feature that protects your data from double-disk failures during reconstruction. Understanding how RAID-DP provides this protection can help you administer your storage systems more effectively.

RAID-DP offers protection against either two failed disks within the same RAID group or against a single-disk failure followed by a bad block or bit error before reconstruction is complete. This protection is equivalent to traditional hardware disk mirroring but requires 43 percent fewer spindles.

RAID-DP protection makes a SATA disk a viable option for your enterprise storage. You can use the less-expensive SATA disks without worrying about data loss and also lower your storage acquisition costs. For more information about RAID-DP, see the Data ONTAP 8.0 7-Mode Storage Management Guide, which can be found at this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S7003173

Failed disks:

- The time to reconstruct data from two failed disks is slightly less than the time to reconstruct data from a single-disk failure.
- It is highly likely that one disk failed before the other and at least some information has already been recreated with traditional row parity. RAID-DP automatically adjusts for this occurrence by starting recovery where two elements are missing from the second disk failure.

18.3.1 What RAID-DP protection is

If an aggregate is configured for RAID-DP protection, Data ONTAP reconstructs the data from one or two failed disks within a RAID group and transfers that reconstructed data to one or two spare disks as necessary.

RAID-DP provides double-parity disk protection when the following conditions occur:

- There is a single-disk or double-disk failure within a RAID group.
- There are media errors on a block when Data ONTAP is attempting to reconstruct a failed disk.

The minimum number of disks in a RAID-DP group is three: at least one data disk, one regular parity disk, and one double-parity (or dParity) disk.

If there is a data-disk or parity-disk failure in a RAID-DP group, Data ONTAP replaces the failed disk in the RAID group with a spare disk and uses the parity data to reconstruct the data of the failed disk on the replacement disk. If there is a double-disk failure, Data ONTAP replaces the failed disks in the RAID group with two spare disks and uses the double-parity data to reconstruct the data of the failed disks on the replacement disks.

RAID-DP is the default RAID type for all aggregates.

18.3.2 How RAID-DP protection works

RAID-DP employs both the traditional RAID4 horizontal parity and diagonal parity structures to significantly increase the data fault tolerance from failed disks.

RAID-DP protection (Figure 18-2) adds a secondary parity disk to each RAID group in an aggregate or traditional volume. A RAID group is an underlying construct on which aggregates and traditional volumes are built. Each traditional RAID 4 group has a minimum of four data disks and one parity disk; aggregates and volumes contain one or more RAID4 groups.

Whereas the parity disk in a RAID4 volume stores row parity across the disks in a RAID4 group, the additional RAID-DP parity disk stores diagonal parity across the disks in a RAID-DP group. With these two parity stripes in RAID-DP, one the traditional horizontal and the other diagonal, data protection is obtained even if two disk failures occur in the same RAID group.

RAID-DP protection is depicted in Figure 18-2.

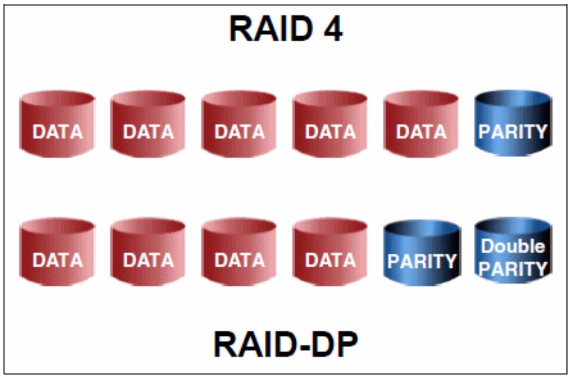


Figure 18-2 RAID-DP

18.4 How space management works

The space management capabilities of Data ONTAP allow you to configure your storage systems to provide the storage availability required by the users and applications accessing the system, while using your available storage as effectively as possible.

Data ONTAP enables space management using the following capabilities:

- Space guarantees
- Space reservations
- Fractional reserve
- Automatic free space preservation

18.4.1 What kind of space management to use

The type of space management to use depends on many factors, including your tolerance for out-of-space errors, whether you plan to overcommit your aggregates, and your rate of data overwrite. Table 18-1 can to help you determine which space management capabilities best suit your requirements.

Tip: LUNs in this context (Table 18-1) refer to the LUNs that Data ONTAP serves to clients, not to the array LUNs used for storage on a storage array.

Table 18-1 Space management capabilities to use

If...	Then use...	Typical usage	Notes
You want management simplicity.	FlexVol volumes with a space guarantee of volume OR Traditional volumes	NAS file systems	It is the easiest option to administer. As long as you have sufficient free space in the volume, writes to any file in this volume will always succeed.
Writes to certain files must always succeed. You want to overcommit your aggregate.	FlexVol volumes with a space guarantee (See “What space guarantees are” on page 289) of file OR Traditional volume AND space reservation enabled for files that require writes to succeed	LUNS Databases	This option enables you to guarantee writes to specific files.
You need even more effective storage usage than file space reservation provides. You actively monitor available space on your volume and can take corrective action when needed. Snapshot copies are short-lived. Your rate of data overwrite is relatively predictable and low.	FlexVol volumes with a space guarantee of volume OR Traditional volume AND Space reservation on for files that require writes to succeed AND Fractional reserve < 100%	LUNs (with active space monitoring) Databases (with active space monitoring)	With fractional reserve <100%, it is possible to use up all available space, even with space reservations on. Before enabling this option, be sure either that you can accept failed writes or that you have correctly calculated and anticipated storage and Snapshot copy usage.
You want to overcommit your aggregate. You actively monitor available space on your aggregate and can take corrective action when needed.	FlexVol volumes with a space guarantee of none	Storage providers who need to provide storage that they know will not immediately be used Storage providers who need to allow available space to be dynamically shared between volumes	With an overcommitted aggregate, writes can fail due to insufficient space.

18.4.2 What space guarantees are

Space guarantees on a FlexVol volume ensure that writes to a specified FlexVol volume or writes to files with space reservations enabled do not fail because of lack of available space in the containing aggregate.

A space guarantee is an attribute of the volume. It is persistent across storage system reboots, takeovers, and givebacks. Space guarantee values can be volume (the default value), file, or none, as follows:

- A space guarantee of *volume* reserves space in the aggregate for the volume. The reserved space cannot be allocated to any other volume in that aggregate. The space management for a FlexVol volume that has a space guarantee of volume is equivalent to a traditional volume.
- A space guarantee of *file* reserves space in the aggregate so that any file in the volume with space reservation enabled can be completely rewritten, even if its blocks are being retained on disk by a Snapshot copy.
- A FlexVol volume that has a space guarantee of *none* reserves no extra space for user data; writes to LUNs or files contained by that volume can fail if the containing aggregate does not have enough available space to accommodate the write.

Important: Because out-of-space errors are unexpected in a CIFS environment, do not set space guarantee to none for volumes accessed using CIFS.

When space in the aggregate is reserved for space guarantee for an existing volume, that space is no longer considered free space. Some operations consume free space in the aggregate, such as creation of Snapshot copies or creation of new volumes in the containing aggregate. They can occur only if there is enough available free space in that aggregate; these operations are prevented from using space already committed to another volume.

When the uncommitted space in an aggregate is exhausted, only writes to volumes or files in that aggregate with space guarantees are guaranteed to succeed.

Attention: Space guarantees are honored only for online volumes. If you take a volume offline, any committed but unused space for that volume becomes available for other volumes in that aggregate. When you bring that volume back online, if there is not sufficient available space in the aggregate to fulfill its space guarantees, you must use the **`force (-f)`** option, and the volume’s space guarantees are disabled. When a volume's space guarantee is disabled, the word (disabled) appears next to its space guarantees in the output of the **`vol status`** command.

18.4.3 What space reservation is

When space reservation is enabled for one or more files or LUNs, Data ONTAP reserves enough space in the volume (traditional or FlexVol) so that writes to those files or LUNs do not fail because of a lack of disk space.

For example, if you create a 100 GB space reserved LUN in a 500 GB volume, that 100 GB of space is immediately allocated, leaving 400 GB remaining in the volume. In contrast, if space reservation is disabled on the LUN, all 500 GB in the volume remain available until writes are made to the LUN.

Space reservation is an attribute of the file or LUN; it is persistent across storage system reboots, takeovers, and givebacks. Space reservation is enabled for new LUNs by default, but you can create a LUN with space reservations disabled or enabled. After you create the LUN, you can change the space reservation attribute by using the **`lun set reservation`** command.

When a volume contains one or more files or LUNs with space reservation enabled, operations that require free space, such as the creation of Snapshot copies, are prevented from using the reserved space. If these operations do not have sufficient unreserved free space, they fail. However, writes to the files or LUNs with space reservation enabled will continue to succeed.

18.4.4 What fractional reserve is

Fractional reserve is a volume option that enables you to determine how much space Data ONTAP reserves for Snapshot copy overwrites for LUNs, as well as for space-reserved files to be used after all other space in the volume is used. The fractional reserve setting defaults to 100%, but you can use the **`vol options`** command to set fractional reserve to any percentage from zero to 100.

Data ONTAP removes or reserves this space from the volume as soon as the first Snapshot copy is created and will only use this space when the rest of the volume has been filled:

- For example, as shown in Figure 18-3 on the left side, a 100 GB volume is shown with two 20 GB LUNs.
- Assuming the LUNs are full Data ONTAP will with the default fractional_reserve=100 reserve 40 GB (2 x 20 GB) of space in the volume to assure that there is always enough space for both the LUNs and all the Snapshot data, even if the LUNs will be completely overwritten. It is depicted in the middle of Figure 18-3.
- As depicted on the right side, if fractional_reserve is set to 60% when the Snapshot copy is created, instead of reserving 40 GB in the volume, Data ONTAP will reserve 24 GB (60% * [2 x 20 GB]).

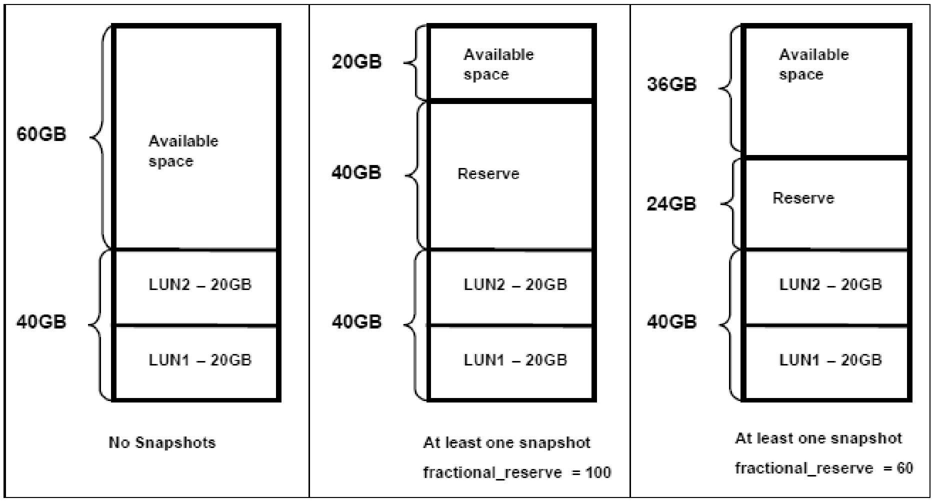


Figure 18-3 Fractional Reserve

It is usually best to set fractional_reserve to 0 for more control over space utilization (see “Reasons to set fractional reserve to zero” on page 292) and to use the **`autodelete`** function.

However, occasionally there are circumstances under which fractional reserve can be used:

- When Snapshot copies cannot be deleted
- When preserving existing Snapshot copies is more important than creating new ones
- On the following types of volumes:
  - Traditional volumes
  - FlexVol volumes with a space guarantee of volume

Attention: If the *guarantee* option for a FlexVol volume is set to *file,* then fractional reserve for that volume is set to 100 percent and is not adjustable. If the *guarantee* option for a FlexVol volume is set to *none*, then fractional reserve for that volume can be set to the desired value. For the vast majority of configurations, set fractional reserve to *zero* when the *guarantee* option is set to none because it greatly simplifies space management.

If fractional reserve is set to 100%, when you create space-reserved LUNs, you can be sure that writes to those LUNs will always succeed without deleting Snapshot copies, even if all of the space reserved LUNs is completely overwritten.

Setting fractional reserve to less than 100 percent causes the space reservation held for all space-reserved LUNs in that volume to be reduced to that percentage. Writes to the space-reserved LUNs in that volume are no longer guaranteed, which is why you need to use **`snap autodelete`** or **`vol autogrow`** for these volumes. Snap autodelete allows a flexible volume to automatically delete the Snapshots in the volume. It is useful when a volume is about to run out of available space and deleting Snapshots can recover space for current writes to the volume.

Fractional reserve is generally used for volumes that hold LUNs with a small percentage of data overwrite.

Considerations:

- If you are using fractional reserve in environments in which write errors due to lack of available space are unexpected, you must monitor your free space and take corrective action to avoid write errors. Data ONTAP provides tools for monitoring available space in your volumes.
- Reducing the space reserved for overwrites (by using fractional reserve) does not affect the size of the space-reserved LUN. You can write data to the entire size of the LUN. The space reserved for overwrites is used only when the original data is overwritten.

If you create a 500-GB space-reserved LUN, then Data ONTAP ensures that 500 GB of free space always remains available for that LUN to handle writes to the LUN.

If you then set fractional reserve to 50 for the LUN's containing volume, then Data ONTAP reserves 250 GB, or half of the space it was previously reserving for overwrites with fractional reserve set to 100. If more than half of the LUN is overwritten, then subsequent writes to the LUN can fail due to insufficient free space in the volume.

Tip: When more than one LUN in the same volume has space reservations enabled, and fractional reserve for that volume is set to less than 100 percent, Data ONTAP does not limit any space-reserved LUN to its percentage of the reserved space. In other words, if you have two 100-GB LUNs in the same volume with fractional reserve set to 30, one of the LUNs can use up the entire 60 GB of reserved space for that volume.

18.4.5 Reasons to set fractional reserve to zero

Setting fractional reserve on a volume to zero and managing your free space manually, rather than having Data ONTAP manage it for you, gives you more control over space utilization.

You can maximize your space utilization by exercising more control over how much space is allocated for overwrites. The cost is extra management time and the risk of encountering out-of-space errors on a space-reserved file or LUN.

For example, setting fractional reserve to zero might be helpful for a qtree SnapMirror destination of a space-reserved LUN. (For Volume SnapMirror, the fractional reserve of the target is always the same as the source.) Depending on the rate of change (ROC) of your data, how much free space is available in the volume, and whether you are taking Snapshots of the volume containing the target, it might make sense to reduce the fractional reserve to zero for the target.

18.4.6 Automatic space provisioning for full volumes

Data ONTAP can automatically make more free space available for a FlexVol volume when that volume is nearly full. You can choose to make the space available by first allowing the volume size to increase, or by first deleting Snapshot copies.

You enable this capability for a FlexVol volume by using the **`vol options`** command with the **`try_first`** option.

Data ONTAP can automatically provide more free space for the volume by using one of the following methods:

- Increase the size of the volume when it is nearly full. This method is useful if the volume's containing aggregate has enough space to support a larger volume. You can increase the size in increments and set a maximum size for the volume.
- Delete Snapshot copies when the volume is nearly full. For example, you can automatically delete Snapshot copies that are not linked to Snapshot copies in cloned volumes or LUNs, or you can define which Snapshot copies you want to delete first; usually your oldest or newest Snapshot copies. You can also determine when to begin deleting Snapshot copies; for example, when the volume is nearly full or when the volume’s Snapshot reserve is nearly full.

You can choose which method (increasing the size of the volume or deleting Snapshot copies) you want Data ONTAP to try first. If the first method does not provide sufficient extra free space to the volume, Data ONTAP will try the other method next.

18.5 Thin provisioning using FlexVol volumes

With thin provisioning, when you create volumes for different purposes in a given aggregate, you do not actually allocate any space for those volumes in advance. The space is allocated only when the application host needs it.

The unused aggregate space is available for the thinly provisioned volumes to expand or for creating new volumes. By allowing as-needed provisioning and space reclamation, thin provisioning can improve storage utilization and decrease storage costs.

A FlexVol volume can share its containing aggregate with other FlexVol volumes. Therefore, a single aggregate is the shared source of all the storage used by the FlexVol volumes it contains. Flexible volumes are no longer bound by the limitations of the disks on which they reside. A FlexVol volume is a pool of storage that can be sized based on how much data you want to store in it, rather than on the size of your disk. This flexibility enables you to maximize the performance and capacity utilization of the storage systems. Because FlexVol volumes can access all available physical storage in the system, dramatic improvements in storage utilization are possible.

The following exemplifies how using FlexVol volumes can help maximize the capacity utilization of storage systems:

A 500 GB volume is allocated with only 100 GB of actual data; the remaining 400 GB allocated has no data stored in it. This unused capacity is assigned to a business application, even though the application might not need all 500 GB until later. The allocated but unused 400 GB of excess capacity is temporarily wasted.

With thin provisioning, the storage administrator provisions 500 GB to the business application but uses only 100 GB for the data. The difference is that with thin provisioning, the unused 400 GB is still available to other applications. This approach allows the application to grow transparently, and the physical storage is fully allocated only when the application truly needs it. The rest of the storage remains in the free pool to be used as needed. Storage administrators can set thresholds, so they are alerted when more disks need to be added to the pool.

See Figure 18-4 for a comparison of thin provisioning with traditional provisioning.

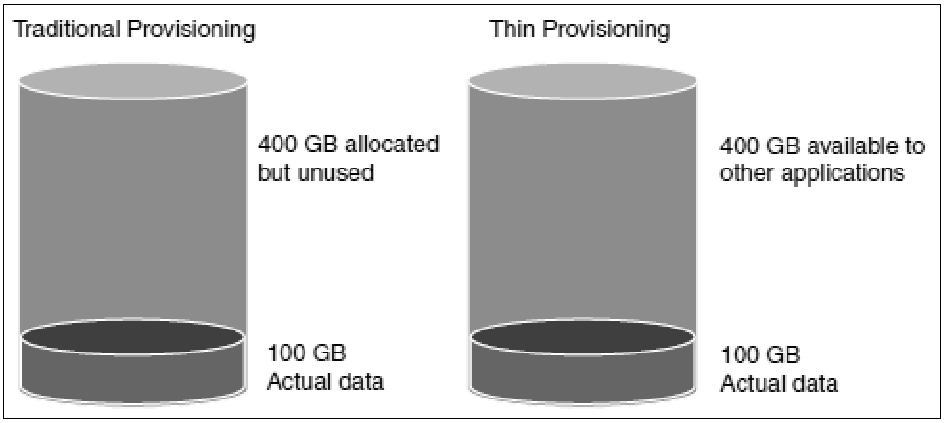


Figure 18-4 Thin provisioning compared to traditional provisioning

The FlexVol technology enables you to oversubscribe the free space to adapt rapidly to the changing business needs.

The benefits of using thin provisioning are as follows:

- It allows storage to be provisioned just like traditional storage, but it is not consumed until data is written.
- Storage-provisioning time is greatly reduced, because you can create the storage for an application quickly without depending on the actual physical space available.
- Through notifications and configurable threshold values, you can plan your procurement strategies well in advance and have enough storage for thin provisioned volumes to grow.

- You can set aggregate overcommitment thresholds by using Protection Manager. Using Provisioning Manager, you can also set policies for provisioning, exporting, and managing your space requirements. For more information about aggregate overcommitment threshold values and provisioning policies.

18.5.1 Storage space management using OnCommand

Provisioning Manager and Operations Manager, now part of the OnCommand management software suite, are part of an add-on management suite that integrates thin provisioning for cross-system configurations.

The Provisioning Manager provides the following capabilities:

- Provisioning policies that manage provisioning and exporting of storage
- Automatic provisioning of a dataset when you assign a provisioning policy to it
- Periodic checking of provisioned storage for conformance to the policy
- Manual controls for resizing volumes and for deleting old Snapshot copies on existing storage and newly provisioned storage
- Migration of datasets and vFiler units to new storage systems
- Deduplication to eliminate duplicate data blocks to reduce storage space

You can use Operations Manager for the following day-to-day activities on storage systems:

- Monitor the device or the object health, the capacity utilization, and the performance characteristics of a storage system
- View or export reports
- Configure alerts and thresholds for event management
- Group devices, vFiler units, host agents, volumes, qtrees, and LUNs

18.5.2 Automating thin provisioning using Provisioning Manager

With Provisioning Manager, you can automate thin provisioning and eliminate the need to provision extra storage space. You can use resource pools for more efficient aggregation of resources.

Provisioning Manager enables you to take advantage of thin provisioning, and resource pooling to get the highest possible level of storage efficiency from your storage resources. You can pool your resources based on attributes such as performance, cost, physical location, or availability.

By grouping related resources into a pool, you can treat the pool as a single unit for monitoring, provisioning, reporting, and role-based access control (RBAC). This approach simplifies the management of these resources and allows for a more flexible and efficient use of the storage.

19

Deduplication

This chapter describes the benefits and functions of deduplication on space saving.

Deduplication is an optional feature of Data ONTAP that significantly improves physical storage space by eliminating duplicate data blocks within a FlexVol volume.

Deduplication works at the block level on the active file system, and uses the WAFL block-sharing mechanism. Each block of data has a digital signature that is compared with all other signatures in a data volume. If an exact block match exists, the duplicate block is discarded and its disk space is reclaimed.

Deduplication removes data redundancies, as shown in Figure 19-1.

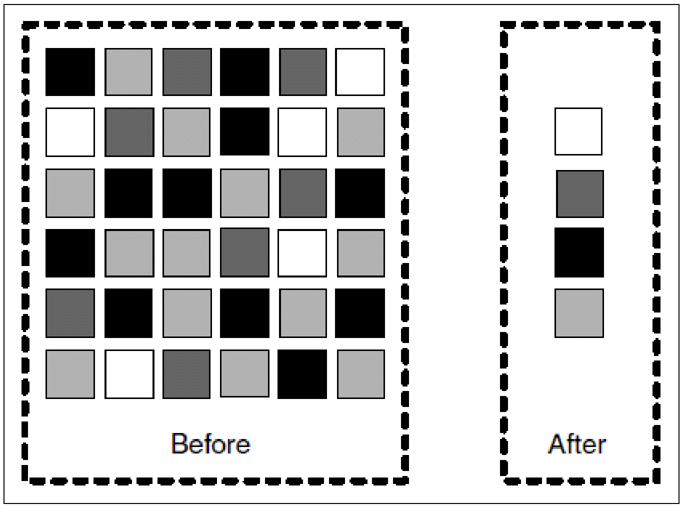


Figure 19-1 Deduplication results

You can configure deduplication operations to run automatically or on a schedule. You can deduplicate new and existing data, or only new data, on a FlexVol volume.

Important: Starting with Data ONTAP 8.1, you can enable the deduplication feature without adding a license. For deduplication, no limit is imposed on the supported maximum volume size. The maximum volume size limit is determined by the type of storage system regardless of whether deduplication is enabled.

The following topics are covered:

- How deduplication works
- What deduplication metadata is
- Guidelines for using deduplication
- Deduplication commands
- Performance considerations for deduplication
- How deduplication works with other features and products

19.1 How deduplication works

Deduplication operates at the block level within the entire FlexVol volume, eliminating duplicate data blocks and storing only unique data blocks.

Data ONTAP writes all data to a storage system in 4-KB blocks. When deduplication runs for the first time on a FlexVol volume with existing data, it scans all the blocks in the FlexVol volume and creates a digital fingerprint for each of the blocks. Each of the fingerprints is compared to all other fingerprints within the FlexVol volume. If two fingerprints are found to be identical, a byte-for-byte comparison is done for all data within the block. If the byte-for-byte comparison detects identical fingerprints, the pointer to the data block is updated, and the duplicate block is freed.

Figure 19-2 shows how the process works.

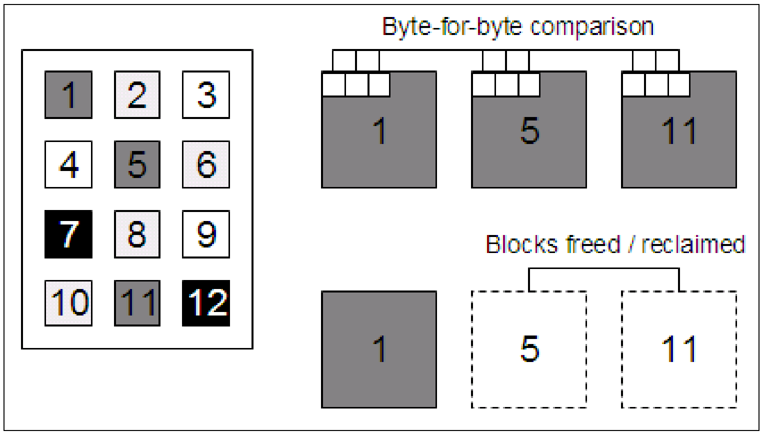


Figure 19-2 Fingerprints and byte-for-byte comparison

Deduplication runs on the active file system. Therefore, as additional data is written to the deduplicated volume, fingerprints are created for each new block and written to a change log file. For subsequent deduplication operations, the change log is sorted and merged with the fingerprint file, and the deduplication operation continues with fingerprint comparisons as previously described.

19.2 What deduplication metadata is

Deduplication uses fingerprints, which are digital signatures for every 4-KB data block in a FlexVol volume. The fingerprint database and the change logs form the deduplication metadata.

The fingerprint database and the change logs used by the deduplication operation are located outside the volume and in the aggregate. Therefore, the deduplication metadata is not included in the FlexVol volume Snapshot copies.

This approach enables deduplication to achieve higher space savings. However, some of the temporary metadata files created during the deduplication operation are still placed inside the volume and are deleted only after the deduplication operation is complete. The temporary metadata files, which are created during a deduplication operation, can be locked in the Snapshot copies. These temporary metadata files remain locked until the Snapshot copies are deleted.

While deduplication can provide substantial space savings, a percentage of storage overhead is associated with it, which you need to consider when sizing a FlexVol volume.

The deduplication metadata can occupy up to 6 percent of the total logical data of the volume, as follows:

- Up to 2 percent of the total logical data of the volume is placed inside the volume.
- Up to 4 percent of the total logical data of the volume is placed in the aggregate.

19.3 Guidelines for using deduplication

When using deduplication, remember the following guidelines about system resources and free space:

- Deduplication is a background process that consumes system resources during the operation. If the data does not change very often in a FlexVol volume, it is best to run deduplication less frequently. Multiple concurrent deduplication operations running on a storage system lead to a higher consumption of system resources.
- Ensure that sufficient free space exists for deduplication metadata in the volumes and aggregates. Before running deduplication for the first time, you must ensure that the aggregate has free space that is at least 4 percent of the total data usage for all volumes in the aggregate, in addition to 2 percent free space for FlexVol volumes. It enables additional storage savings by deduplicating any new blocks with those that existed before the upgrade. If there is not sufficient space available in the aggregate, the deduplication operation fails with an error message. During a deduplication failure, there is no loss of data and the volume is still available for read/write operations. However, depending upon the space availability in the aggregate, fingerprints of the newly added data might be lost.

  Tip: Use the **`df`** command to check free space on aggregates and volumes.

- You cannot increase the size of a volume that contains deduplicated data beyond the maximum supported size limit, either manually or by using the **`autogrow`** option.
- You cannot enable deduplication on a volume if it is larger than the maximum volume size. However, you can enable deduplication on a volume after reducing its size within the supported size limits.

Starting with Data ONTAP 8.0, FlexVol volumes can be either 32 bit or 64 bit. All FlexVol volumes created using releases earlier than Data ONTAP 8.0 are 32-bit volumes. A 32-bit volume, like its containing 32-bit aggregate, has a maximum size of 16 TB. A 64-bit volume has a maximum size as large as its containing 64-bit aggregate (up to 100 TB, depending on the storage system model).

Considerations:

- Even in 64-bit volumes, the maximum size for LUNs and files is still 16 TB.
- For best performance, if you want to create a large number of small files in a volume, it is best to use 32-bit volumes.

  64-bit aggregates have a larger address space and need more memory for their metadata, compared to 32-bit aggregates. This extra memory usage reduces the amount of memory available for user data. Therefore, for workloads that are memory intensive, you might experience a slight performance impact when running the workload from a FlexVol volume contained in a 64-bit aggregate compared to running the workload from a volume contained in a 32-bit aggregate.

  Workloads that are highly random in nature typically access more metadata over a given period of time compared to sequential workloads. Random read workloads with a very large active data set size might experience a performance impact when run on a FlexVol volume in a 64-bit aggregate, compared to when run on a FlexVol volume in a 32-bit aggregate. It is because the data set size combined with the increased metadata size can increase memory pressure on the storage system and result in an increased amount of on disk I/O. In such scenarios, if you want to run the random workload from a volume contained in a 64-bit aggregate, using PAM (or PAM II) improves the performance delivered by the storage system and helps alleviate any performance impact seen with 64-bit aggregates.

  Note that just having a 64-bit aggregate on the storage system does not result in any sort of performance degradation. The effects on performance, if any, are seen when data in any FlexVol volume in the 64-bit aggregate starts to be accessed.

19.4 Deduplication commands

This section describes the main deduplication commands on Data ONTAP.

19.4.1 Activating the deduplication license

You need to activate the deduplication license before enabling deduplication.

You can activate the deduplication license by using the **`license add`** command after installing Data ONTAP.

Enter the following command:

```
license add <license_key>
```

Here, **`license_key`** is the code for the deduplication license.

You also need to add the NearStore option license. Run the following command:

```
license add <nearstore_option license key>
```

Important: The deduplication license is only supported with Data ONTAP 7.2.2 or later releases up to Data ONTAP 8.x.

Starting with Data ONTAP 8.1, you can enable the deduplication feature without adding a license. Also with DOT 8.1, for deduplication, no limit is imposed on the supported maximum volume size.

The maximum volume size limit is determined by the type of storage system regardless of whether deduplication is enabled.

19.4.2 Common deduplication operations

Here we show how to perform various deduplication operations:

- Enable deduplication operations with the following command:

  ```
  sis on <path>
  ```

  Here, **`path`** is the complete path to the FlexVol volume.

  Example 19-1 Enabling deduplication in a volume

  ```
  itsotuc*> sis on /vol/flexvol
  SIS for "/vol/flexvol" is enabled.
  ```

- Start deduplication operations with the following command:

  ```
  sis start [-s] [-f] [-d] [-sp] /vol/volname
  ```

  The **`-s`** option scans the volume completely and you are prompted to confirm if deduplication must started on the volume.

  The **`-f`** option starts deduplication on the volume without any prompts.

  The **`-d`** option starts a new deduplication operation after deleting the existing checkpoint information.

  The **`-sp`** option initiates a deduplication operation by using the previous checkpoint regardless of how old the checkpoint is.

- View the deduplication status of a volume with the following command:

  ```
  sis status -l path
  ```

  Here, **`path`** is the complete path to the FlexVol; for example, /vol/flexvol.

  The **`sis status`** command is the basic command to view the status of deduplication operations on a volume. For more information about the sis status command, see the **`sis`** man page. Table 19-1 lists and describes status and progress messages that you might see after running the **`sis status -l`** command.

Table 19-1 Status and progress messages after sis commands

Message	Message type	Description
Idle	Status and progress	No active deduplication operation is in progress.
Pending	Status	The limit of maximum concurrent deduplication operations allowed for a storage system or a vFiler unit is reached. Any deduplication operation requested beyond this limit is queued.
Active	Status	Deduplication operations are running.

Message	Message type	Description
size Scanned	Progress	A scan of the entire volume is running, of which **size** is already scanned.
size Searched	Progress	A search of duplicated data is running, of which **size** is already searched.
size (pct) Done	Progress	Deduplication operations have saved **size** amounts of data. **pct** is the percentage saved of the total duplicated data that was discovered in the search stage.
size Verified	Progress	A verification of the metadata of processed data blocks is running, of which **size** is already verified.
pct% Merged	Progress	Deduplication operations have merged **pct%** (percentage) of all the verified metadata of processed data blocks to an internal format that supports fast deduplication operations.

- View deduplication space savings as follows:

  The **df -s** command displays the space savings in the active file system only. Space savings in Snapshot copies are not included in the calculation.

  Enter the following command to view space savings with deduplication as shown in Example 19-2:

  ```
  df -s volname
  ```

  Here, **volname** is the name of the FlexVol volume. For example, vol2.

  Example 19-2 Listing saved space in a volume

  ```
  itsotuc*> df -s vol2
  Filesystem                used      saved       %saved
  /vol/vol2/            82402564   17942796          18%
  itsotuc*>
  ```

  Tip: Using deduplication does not affect volume quotas. Quotas are reported at the logical level, and remain unchanged.

- Stop deduplication operations as follows:

  Enter the following command to stop the deduplication operation as shown in Example 19-3:

  ```
  sis stop path
  ```

  Here, **path** is the complete path to the FlexVol volume. For example, `/vol/flexvol`.

  Example 19-3 Stopping deduplication in a volume

  ```
  itsotuc*> sis stop /vol/flexvol
  Operation is currently idle: /vol/flexvol
  itsotuc*>
  ```

- Disable deduplication operations as follows:

  If deduplication on a specific volume has a performance impact greater than the space savings achieved, you might want to disable deduplication on that volume. If you want to remove deduplication license from your system, you must disable it before removing it.

  - If deduplication is in progress on the volume, enter the following command to abort the operation:

    ```
    sis stop path
    ```

    Here, **path** is the complete path to the FlexVol volume. For example, /vol/vol1.

  - Enter the following command to disable the deduplication operation:

    ```
    sis off path
    ```

    This command stops all future deduplication operations. See Example 19-4.

    Example 19-4 Disabling the deduplication in a volume

    ```
    itsotuc*> sis off /vol/flexvol
    SIS for "/vol/flexvol" is disabled.
    itsotuc*>
    ```

  Tip: Before removing the deduplication license, you must disable deduplication on all the FlexVol volumes, using the **sis off** command. Otherwise, you will receive a warning message asking you to disable this feature. Any deduplication operation that occurred before removing the license will remain unchanged.

- Deduplication checkpoint feature:

  The checkpoint is used to periodically log the execution process of a deduplication operation. When a deduplication operation is stopped for any reason (such as system halt, panic, reboot, or last deduplication operation failed or stopped) and checkpoint data exists, the deduplication process can resume from the latest checkpoint file.

  - You can restart from the checkpoint by using the following commands:

    ```
    sis start -s
    sis start (manually or automatically)
    ```

  - You can view the checkpoint by using the following command:

    ```
    sis status -l
    ```

  The checkpoint is created at the end of each stage or sub-stage of the deduplication process. For the **sis start -s** command, the checkpoint is created at every hour during the scanning phase.

  If a checkpoint corresponds to the scanning stage (the phase when the **sis start -s** command is run) and is older than 24 hours, the deduplication operation will not resume from the previous checkpoint automatically. In this case, the deduplication operation will start from the beginning. However, if you know that significant changes have not occurred in the volume since the last scan, you can force continuation from the previous checkpoint using the **-sp** option.

19.5 Performance considerations for deduplication

Certain factors affect the performance of deduplication. You need to check the performance impact of deduplication in a test setup, including sizing considerations, before deploying deduplication in performance-sensitive or production environments.

The following factors affect the performance of deduplication:

- Application and the type of data used
- The data access pattern (for example, sequential versus random access, the size and pattern of the input and output)
- The amount of duplicate data, the amount of total data, and the average file size.

> **Tip:** To avoid performance problems, run the first deduplication and monitor it. If you notice any performance degradation, abort the operation with command: `sis stop <vol-name>` . See Example 19-3 on page 301.

- The nature of data layout in the volume
- The amount of changed data between deduplication operations

> **Tip:** use the `df -s` command between deduplication operations to know the amount of changed data. See Example 19-2 on page 301.

- The number of concurrent deduplication operations

> **Tip:** You can run a maximum of eight concurrent deduplication operations on a system. If any more consecutive deduplication operations are scheduled, the operations are queued.

- Hardware platform (system memory and CPU module)
- Load on the system (for example, MBps)
- Disk types (for example, ATA/FC, and RPM of the disk)

19.6 How deduplication works with other features and products

When using deduplication with other features, be mindful of the following considerations.

19.6.1 Deduplication and Snapshot copies

You can run deduplication only on the active file system. However, this data can get locked in Snapshot copies created before you run deduplication, resulting in reduced space savings.

Data can get locked in Snapshot copies in two ways:

- One possibility is that the Snapshot copies were created before the deduplication operation is run. You can avoid this situation by always running deduplication before Snapshot copies are created.
- When the Snapshot copy is created, a part of the deduplication metadata resides in the volume and the rest of the metadata resides in the aggregate outside the volume. The fingerprint files and the change-log files that are created during the deduplication operation are placed in the aggregate and are not captured in Snapshot copies, which results in higher space savings. However, some temporary metadata files that are created during a deduplication operation are still placed inside the FlexVol; these files are deleted after the deduplication operation is complete.

These temporary metadata files can get locked in Snapshot copies if the copies are created during a deduplication operation. The metadata remains locked until the Snapshot copies are deleted. If a Snapshot copy is locked, the **snap delete** operation fails until you execute a **snapmirror release** or **snapvault release** command to unlock the Snapshot copy. Snapshot copies are locked because SnapMirror or SnapVault is maintaining these copies for the next update. Deleting a locked Snapshot copy will prevent SnapMirror or SnapVault from correctly replicating a file or volume as specified in the schedule you set up. Example 19-5 shows how to delete a locked SnapMirror Snapshot copy, and Example 19-6 shows how to delete a locked SnapVault Snapshot copy.

Example 19-5   Deleting a locked SnapMirror Snapshot copy

```
itsotuc*> snap delete vol0 oldsnap
Can't delete oldsnap: snapshot is in use by snapmirror.
Use 'snapmirror destinations -s' to find out why.
itsotuc*> snapmirror destinations -s vol0
Path Destination
/vol/vol0 itsotuc0*:vol0
itsotuc*> snapmirror release vol0 itsotuc0*:vol0
itsotuc*> snap delete vol0 oldsnap
```

Example 19-6   Deleting a locked SnapVault Snapshot copy

```
itsotuc*> snap delete vol0 oldsnap
Can't delete oldsnap: snapshot is in use by snapvault.
Use 'snapvault status -l' to find out why.
itsotuc*> snapvault status -l
SnapVault client is ON.
Source: itsotuc*:/vol/vol0/qt3
Destination itsotuc0*:/vol/sv_vol/qt3...
itsotuc*> snapvault release /vol/vol0/qt3
itsotuc0*:/vol/sv_vol/qt3
itsotuc*> snap delete vol0 oldsnap
```

To avoid conflicts between deduplication and Snapshot copies, follow these guidelines:

- Run deduplication before creating new Snapshot copies.
- Remove unnecessary Snapshot copies stored in deduplicated volumes.
- Reduce the retention time of Snapshot copies stored in deduplicated volumes.
- Schedule deduplication only after significant new data has been written to the volume.
- Configure appropriate reserve space for the Snapshot copies.
- If snap reserve is 0, turn off the schedule for automatic creation of Snapshot copies (which is the case in most LUN deployments).

Tips:

- To check the snap reserve value set, run the command:
  snap reserve <vol-name>
- To disable the automatic Snapshot copy, run the command:
  vol options <vol-name> nosnap on
- To enable the automatic Snapshot copy, run the command:
  vol options <vol-name> nosnap off

19.6.2 Deduplication and volume SnapMirror

You can use volume SnapMirror to replicate a deduplicated volume.

When using volume SnapMirror with deduplication, consider the following information:

- You need to enable both the deduplication and SnapMirror licenses.

> **Tips:**
>
> - To enable the deduplication license, see 19.4.1, "Activating the deduplication license" on page 299
> - To enable the SnapMirror license, do the same procedure, entering the command: **`license add`** `xxxxxx`, where `xxxxxx` is the license code you purchased.

- You can enable deduplication on the source system, the destination system, or both systems.

> **Attention:** A deduplication license is not required on the destination storage system. However, if the primary storage system is not available and the secondary storage system becomes the new primary, deduplication must be licensed on the secondary storage system for deduplication to continue. Therefore, you might want to license deduplication on both storage systems.

You can enable, run, and manage deduplication only from the primary storage system. However, the FlexVol volume in the secondary storage system inherits all the deduplication attributes and storage savings through SnapMirror:

- The shared blocks are transferred only once. Therefore, deduplication also reduces the use of network bandwidth. The fingerprint database and the change logs that the deduplication process uses are located outside a volume, in the aggregate. Therefore, volume SnapMirror does not transfer the fingerprint database and change logs to the destination. This change provides additional network bandwidth savings.
- If the source and destination volumes are on different storage system models, they might have different maximum volume sizes. The lower maximum applies. When creating a SnapMirror relationship between two different storage system models, ensure that the maximum volume size with deduplication is set to the lower maximum volume size limit of the two models.
- The volume SnapMirror update schedule does not depend on the deduplication schedule. When configuring volume SnapMirror and deduplication, you need to coordinate the deduplication schedule and the volume SnapMirror schedule. Start the volume SnapMirror transfers of a deduplicated volume after the deduplication operation is complete. This schedule prevents the sending of undeduplicated data and additional temporary metadata files over the network. If the temporary metadata files in the source volume are locked in Snapshot copies, these files consume extra space in the source and destination volumes. Volumes whose size has been reduced to within the limit supported by deduplication can be part of the SnapMirror primary storage system and the secondary storage system.

19.6.3 Deduplication and qtree SnapMirror

You can use deduplication for volumes that use qtree SnapMirror.

Deduplication operations are supported with qtree SnapMirror. Qtree SnapMirror does not automatically initiate a deduplication operation at the completion of every individual qtree SnapMirror transfer. You can set up a deduplication schedule independent of your qtree SnapMirror transfer schedule.

Reference: The `sis config` command is used to configure and view deduplication schedules for flexible volumes. For more details about the `sis` command, see the *IBM System Storage N series Software Guide*, SG24-7129.

When using qtree SnapMirror with deduplication, consider the following information:

- You need to enable both the deduplication and SnapMirror licenses.

  Tip: You can enable deduplication on the source system, the destination system, or both systems.

- Even when deduplication is enabled on the source system, duplicate blocks are sent to the destination system. Therefore, no network bandwidth savings is achieved.
- To recognize space savings on the destination system, run deduplication on the destination after the qtree SnapMirror transfer is complete.
- You can set up a deduplication schedule independently of the qtree SnapMirror schedule. For example, on the destination system, the deduplication process does not start automatically after qtree SnapMirror transfers are finished.
- Qtree SnapMirror recognizes deduplicated blocks as changed blocks. Therefore, when you run deduplication on an existing qtree SnapMirror source system for the first time, all the deduplicated blocks are transferred to the destination system. This process might result in a transfer several times larger than the regular transfers.

When using qtree SnapMirror with deduplication, ensure that qtree SnapMirror uses only the minimum number of Snapshot copies that it requires. To ensure this minimum, retain only the latest Snapshot copies.

19.6.4 Deduplication and SnapVault

The deduplication feature is integrated with the SnapVault secondary license. This feature increases the efficiency of data backup and improves the use of secondary storage.

The behavior of deduplication with SnapVault is similar to the behavior of deduplication with qtree SnapMirror, with the following exceptions:

- Deduplication is also supported on the SnapVault destination volume.
- The deduplication schedule depends on the SnapVault update schedule on the destination system. However, the deduplication schedule on the source system does not depend on the SnapVault update schedule, and it can be configured independently on a volume.
- Every SnapVault update (baseline or incremental) starts a deduplication process on the destination system after the archival Snapshot copy is taken.

- A new Snapshot copy replaces the archival Snapshot copy after deduplication finishes running on the destination system. (The name of this new Snapshot copy is the same as that of the archival copy, but the Snapshot copy uses a new timestamp, which is the creation time.)
- You cannot configure the deduplication schedule on the destination system manually or run the **`sis start`** command. However, you can run the **`sis start -s`** command on the destination system as shown in Example 19-7.

Example 19-7 Starting deduplication in a volume

```
itsotuc*> sis start -s /vol/flexvol01
The file system will be scanned to process existing data in /vol/flexvol01.
This operation may initialize related existing metafiles.
Are you sure you want to proceed (y/n)? y
The SIS operation for "/vol/flexvol01" is started.
```

- The SnapVault update does not depend on the deduplication operation. A subsequent incremental update is allowed to continue while the deduplication operation on the destination volume from the previous backup is still in progress. In this case, the deduplication operation continues; however, the archival Snapshot copy is not replaced after the deduplication operation is complete.
- The SnapVault update recognizes the deduplicated blocks as changed blocks. Thus, when deduplication is run on an existing SnapVault source for the first time, all saved space is transferred to the destination system. The size of the transfer might be several times larger than the regular transfers. Running deduplication on the source system periodically will help prevent this issue for future qtree SnapMirror transfers. Run deduplication before the SnapVault baseline transfer.

Tip: You can run a maximum of eight concurrent deduplication operations on a system. This number includes the deduplication operations linked to SnapVault volumes and those that are not linked to SnapVault volumes.

19.6.5 Deduplication and SnapRestore

The metadata created during a deduplication operation is located in the aggregate. Therefore, when you initiate a SnapRestore operation on a volume, the metadata is not restored to the active file system. The restored data, however, retains the original space savings.

After a SnapRestore operation, if deduplication is enabled on the volume, any new data written to the volume continues to be deduplicated. However, space savings is obtained for only the new data.

To run deduplication for all the data on the volume, use the **`sis start -s`** command.

This command builds the fingerprint database for all the data in the volume. The amount of time this process takes depends on the size of the logical data in the volume. Before using the **`sis start -s`** command, you must ensure that the volume and the aggregate containing the volume have sufficient free space for the deduplication metadata. See 19.3, “Guidelines for using deduplication” on page 298.

19.6.6 Deduplication and volume copy

Volume copy is a method of copying both data in the active file system and data in storage systems from one volume to another. The source and destination volumes must both be FlexVol volumes.

When deduplicated data is copied by using the **`vol copy`** command, the copy of the data at the destination inherits all the deduplication attributes and storage savings of the source data.

> **Tip:** When using the **`vol copy`** command, the destination volume must be in restrict mode. Use the **`vol restrict`** command to restrict it and allow the **`vol copy`** command.

The metadata created during a deduplication operation (fingerprint files and changelog files) are located outside the volume in the aggregate. Therefore, when you run the volume copy operation on a volume, the fingerprint files and change-log files are not restored to the active file system. After a volume copy operation, if deduplication is enabled on the volume, any new data written to the volume continues to be deduplicated. However, space savings are only obtained for the new data.

To run deduplication for all the data on the volume, use the **`sis start -s`** command.

This command builds the fingerprint database for all the data in the volume. The amount of time this process takes depends on the size of the logical data in the volume. Before using the **`sis start -s`** command, you must ensure that the volume and the aggregate containing the volume have sufficient free space for deduplication metadata. See 19.3, “Guidelines for using deduplication” on page 298.

19.6.7 Deduplication and FlexClone volumes

Deduplication is supported on FlexClone volumes. FlexClone volumes are writable clones of a parent FlexVol volume.

The FlexClone volume of a deduplicated volume is a deduplicated volume. The cloned volume inherits the deduplication configuration of the parent volume (for example, deduplication schedules).

The FlexClone volume of a non-deduplicated volume is a non-deduplicated volume. If you run deduplication on a clone volume, the clone is deduplicated, but the original volume remains nondeduplicated.

The metadata created during a deduplication operation (fingerprint files and change-log files) are located outside the volume in the aggregate; therefore, they are not cloned. However, the data retains the space savings of the original data.

Any new data written to the destination volume continues to be deduplicated and fingerprint files for the new data are created. Space savings is only obtained for the new data.

To run deduplication for all the data on the cloned volume, use the **`sis start -s`** command. The time the process takes to finish depends on the size of the logical data in the volume.

When a cloned volume is split from the parent volume, deduplication of all data in the clone that was part of the parent volume is undone after the volume-split operation. However, if deduplication is running on the clone volume, the data is deduplicated in the subsequent deduplication operation.

19.6.8 Deduplication in a High Availability pair

You can activate deduplication in a High Availability (HA) pair.

The maximum number of concurrent deduplication operations allowed on each node of an HA pair is eight. If one of the nodes fails, the other node takes over the operations of the failed node. In takeover mode, the working node continues with its deduplication operations as usual. However, the working node does not start any deduplication operations on the failed node.

Attention: Change logging for volumes with deduplication continues for the failed node in takeover mode. Therefore, you can perform deduplication operations on data written during takeover mode after the failed node is active, and there is no loss in space savings. To disable change logging for volumes that belong to a failed node, you can turn off deduplication on those volumes. You can also view the status of volumes with deduplication for a failed node in takeover mode.

19.6.9 Deduplication and VMware

You can run deduplication in VMware environments for efficient space savings.

While planning the Virtual Machine Disk (VMDK) and data store layouts, follow these guidelines:

- Operating system VMDKs deduplicate efficiently because the binary files, patches, and drivers are highly redundant between virtual machines. You can achieve maximum savings by keeping these VMDKs in the same volume.
- Application binary VMDKs deduplicate to varying degrees. Applications from the same vendor commonly have similar libraries installed; therefore, you can achieve moderate deduplication savings. Applications written by different vendors do not deduplicate at all.
- Application datasets when deduplicated have varying levels of space savings and performance impact based on the application and intended use. Carefully consider what application data needs to be deduplicated.
- Transient and temporary data, such as VM swap files, pagefiles, and user and system temp directories, does not deduplicate well and potentially adds significant performance impact when deduplicated. Therefore, it is best to keep this data on a separate VMDK and volume that are not deduplicated.

Application data has a major effect on the percentage of storage savings achieved with deduplication.

New installations typically achieve large deduplication savings.

Important: In VMware environments, proper partitioning and alignment of the VMDKs is important. Applications whose performance is impacted by deduplication operations are likely to have the same performance impact when you run deduplication in a VMware environment.

19.6.10 Deduplication and MultiStore

Deduplication commands are available in all the vFiler contexts. Deduplication support on vFiler units allows users to reduce redundant data blocks within vFiler units.

You can enable deduplication only on FlexVol volumes in a vFiler unit. Deduplication support on vFiler units ensures that volumes owned by a vFiler unit are not accessible to another vFiler unit.

Deduplication also supports disaster recovery and migration of vFiler units. If you enable deduplication on the volume in the source vFiler unit, the destination vFiler unit inherits all deduplication attributes.

You must license deduplication on the primary storage system. It is best that you also license deduplication on the secondary storage system. These licenses ensure that deduplication operations can continue without any disruption in case a failure causes the secondary vFiler unit to become the primary storage system.

To use the deduplication feature, activate the following licenses on the storage system:

- `multistore`
- `a_sis`

> **Licenses:** See 19.4.1, “Activating the deduplication license” on page 299 to activate licenses for deduplication operations.

You can run deduplication commands using the RSH or SSH protocol. Any request is routed to the IP address and IP space of the destination vFiler unit.

20

Compression

This chapter describes the N series data compression feature. N series data compression is a software-based solution that provides transparent data compression. It can be run inline or post-process and also includes the ability to perform compression of existing data. No application changes are required to use N series data compression.

The following topics are covered:

- Introduction to data compression
- Potential space savings
- Performance
- Compression examples

20.1 Introduction to data compression

Starting with Data ONTAP 8.1, you can compress data within a FlexVol volume using the data compression technology. You can use data compression only on FlexVol volumes that are created on 64-bit aggregates on primary, secondary, and tertiary storage tiers.

After you enable data compression on a FlexVol volume, all subsequent writes to the volume are compressed inline. However, existing data remains uncompressed. You can use the data compression scanner to compress the existing data.

Starting with Data ONTAP 8.1, data compression is supported on vFiler units. You can perform data compression operations from the CLI of all vFiler units, in addition to the CLI of vfiler0.

Important:

- Starting with Data ONTAP 8.1, data compression is supported on SnapLock volumes and in stretch and fabric-attached MetroCluster configurations. For compression, no limit is imposed on the supported maximum volume size.
- The maximum volume size limit is determined by the type of storage system regardless of whether compression is enabled.

For more information, see the *Data ONTAP 8.1 7-Mode Storage Efficiency Management Guide*, available at the following website:

- http://www.ibm.com/support/docview.wss?uid=ssg1S7003898

20.1.1 How N series data compression works

N series data compression does not compress the entire file as a single contiguous stream of bytes. This would be prohibitively expensive when it comes to servicing small reads or overwrites from part of a file because it requires the entire file to be read from disk and uncompressed before the request can be served. It would be especially difficult on large files. To avoid this situation, N series data compression works by compressing a small group of consecutive blocks, known as a compression group. In this way, when a read or overwrite request comes in, we only need to read a small group of blocks, not the entire file. It optimizes read and overwrite performance and allows greater scalability in the size of the files being compressed.

Compression groups

The N series compression algorithm divides a file into compression groups. The file must be larger than 8k or it will be skipped for compression and written to disk uncompressed. Compression groups are a maximum of 32K. A compression group contains data from one file only. A single file can be contained within multiple compression groups. If a file is 60k, it is contained within two compression groups. The first group is 32k and the second group is 28k.

Compressed writes

The N series handles compression write requests at the compression group level. Each compression group is compressed separately. The compression group is left uncompressed unless a savings of at least 25% can be achieved on a per-compression-group basis; this optimizes the savings while minimizing the resource overhead (Figure 20-1).

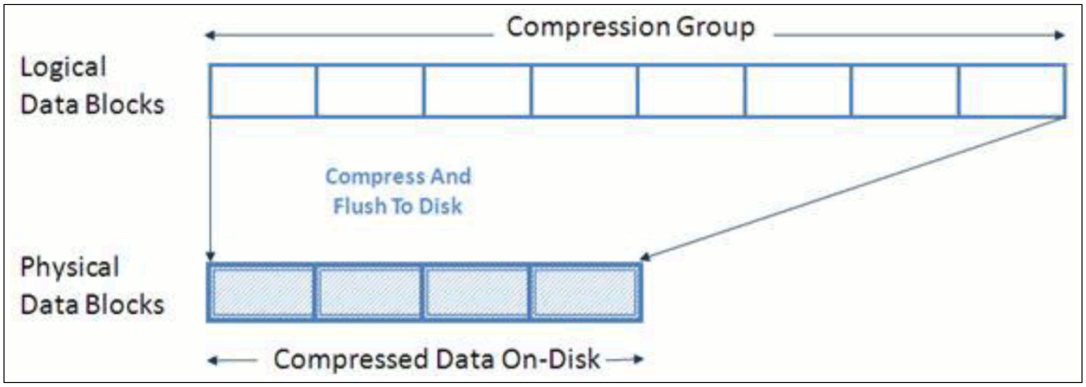


Figure 20-1 Compression write request handling

Because compressed blocks contain fewer blocks to be written to disk, compression reduces the amount of write I/Os required for each compressed write operation. Not only does it lower the data footprint on disk; it can also decrease the time to complete your backups.

Compressed reads

When a read request comes in, we read only the compression group(s) that contain the requested data, not the entire file. It optimizes the amount of I/O being used to service the request. When reading compressed data, only the required compression group data blocks will be transparently decompressed in memory. The data blocks on disk remain compressed. It has much less overhead on the system resources and read service times.

In summary, the N series compression algorithm is optimized to reduce overhead for both reads and writes.

20.1.2 When data compression runs

The N series data compression can be run either in-line or as a post-process operation.

In-line operations

N series data compression can be configured as an inline operation. In this way, as data is sent to the storage system it is compressed in memory before being written to the disk. The advantage of this implementation is that it can reduce the amount of write I/O. This implementation option can affect your write performance and thus must not be used for performance-sensitive environments that without proper testing to understand the impact.

In order to provide the fastest throughput inline compression will compress most new writes but will defer some more performance-intensive compression operations to compress when the next post-process compression process is run. An example of a performance-intensive compression operation includes partial compression group writes and overwrites.

Post-process operations

N series data compression includes the ability to run post-process compression. Post-process compression uses the same schedule as deduplication utilizes. If compression is enabled when the `sis` schedule initiates a post-process operation it runs compression first, followed by deduplication. It includes the ability to compress data that existed on disk prior to enabling compression.

When to use in-line or post-process operations

Inline compression provides immediate space savings; post-process compression first writes the blocks to disk as uncompressed and then at a scheduled time compresses the data.

Post-process compression is useful for environments that want compression savings but do not want to incur any performance penalty associated with new writes.

Inline compression is useful for customers who are not as performance sensitive and can handle some impact on new write performance as well as CPU during peak hours.

If both in-line and post-process compression are enabled, then post-process compression will try to compress only blocks that are not already compressed. This includes blocks that were bypassed by inline compression such as small partial compression group overwrites.

20.2 Potential space savings

This section describes the potential storage savings for three scenarios: deduplication only, inline compression only (disabling the post-process schedule), and the combination of compression and deduplication.

Comprehensive testing with various datasets was performed to determine typical space savings in different environments. These results (Table 20-1) were obtained from various customer deployments and lab testing, and are dependent upon the customer specific configuration.

Important: Compression results can vary, based on your individual data.

Table 20-1 Typical deduplication and compression space savings

Data type	Application type	In-line compression only	Deduplication only	Deduplication and compression
File Services/IT Infrastructure		50%	30%	65%
Virtual Servers and Desktops (Boot Volumes)		55%	70%	70%
Database	OLTP	65%	0%	65%
	Data warehouse	70%	15%	70%
E-mail, Collaborative	Exchange 2003/2007	35%	3%	35%
	Exchange 2010	35%	15%	40%
Engineering Data		55%	30%	75%
Geoseismic		40%	3%	40%
Archival Data		Application Dependent	25%	Application Dependent
Backup Data		Application Dependent	95%	Application Dependent

In the N series implementation, compression is run before deduplication. It provides us with the ability to use inline compression to get immediate space savings from compression followed by additional savings from deduplication. In our testing of other solutions we found that better savings were achieved by running compression prior to deduplication.

20.3 Performance

Because compression is part of Data ONTAP, it is tightly integrated with the N series WAFL (Write Anywhere File Layout) file structure. As a result, N series compression is optimized to perform with high efficiency.

Both compression and deduplication can be scheduled to run during non-peak hours. This minim uses the bulk of the overhead on the system during peak hours. When there is a lot of activity on the system, compression/deduplication runs as a background process and limits its resource usage. When there is not a lot of activity on the system, compression/deduplication speed will increase, and it will utilize available system resources. The potential performance impact must be fully tested prior to implementation.

Because compression/deduplication is run on a per-volume basis, the more volumes you have enabled, the greater the impact on system resources. We advise that you stagger the compression/deduplication schedule for volumes to help control the overhead.

When considering adding compression or deduplication, remember to use standard sizing and testing methods, as would be used when considering the addition of applications to the storage system. It is important to understand how inline compression will affect your system, how long post-process operations will take in your environment, and whether you have the bandwidth to run these with acceptable impact on the applications running on your storage system.

While N series with compression is optimized to minimize impact on throughput, there might still be an impact even if only using post process compression, because the system still has to uncompress some data in memory when servicing reads. This impact will continue so long as the data is compressed on disk regardless of whether compression is disabled on the volume at a future point.

The more compressible the data, the faster compression occurs. In other words it will be faster to compress data that has 75% savings from compression compared to compressing data that has only 25% savings from compression.

Because of these factors, we advise that performance with compression/deduplication be carefully measured in a test setup and taken into sizing consideration before deploying compression/deduplication in performance-sensitive solutions.

20.3.1 Performance impact of in-line and post-process compression

The N series compression feature can be run as either an in-line or post-process operation.

In-line compression performance

Inline compression will consume extra CPU resources whenever data is read or written to the volume; this includes peak hours. The more volumes that are enabled with compression, the more the resource demand and overhead will be. The impact will be shown by longer latencies on the volume that has compression enabled. Given the possible impact to peak time performance, we advise limiting typical use cases to those not as performance sensitive, such as file services, backup, and archive solutions.

On workloads such as file services, systems with less than 50% CPU utilization have shown an increased CPU usage of ~20% for datasets that were 50% compressible. For systems with more than 50% CPU utilization, the impact might be more significant.

Post-process compression performance

To get an idea of how long it takes for a single compression process to complete, suppose that the compression process is running on a flexible volume that contains data that is 50% compressible and at a conservative rate of 70 MB/sec on a N7900. If 1 TB of new data was added to the volume since the last compression process ran, this compression operation takes about 4 hours to complete. Remember that other factors such as different amounts of compressible data, different types of systems, or other applications running on the system can affect the compression performance.

20.3.2 I/O performance on compressed volumes

Compression also has an impact on I/O performance. File services-type benchmark testing with compression savings of 50% has shown a decrease in throughput of ~5%.

Write performance

The impact of compression on the write performance of a system is different depending on whether you are using inline or post-process compression.

If you use inline compression, the write performance is a function of the hardware platform that is being used, the type of write (that is, partial or full), the compressibility of the data, the number of volumes with compression enabled, as well as the amount of load that is placed on the system.

For post-process compression, the write performance will only be impacted for partial overwrites of previously compressed data; all other data will be written uncompressed. It will be compressed the next time post-process compression is run.

For physical backup environments such as volume, SnapMirror with datasets that provide good space savings, there is no CPU impact and there is reduced I/O on the destination system, faster replications, as well as network bandwidth savings during the transfer.

For logical backup environments such as qtree SnapMirror, the effect of enabling inline compression depends on a number of factors. For example, with four parallel qtrees, SnapMirror transfers to a N6270 with four separate compression-enabled volumes.

We saw that the backup window remained constant, given the following factors:

- CPU utilization increased ~35% when compression was enabled on all four volumes on the destination system.
- Dataset was 70% compressible.

The backup window will be affected the most if CPU becomes a bottleneck. We advise testing in your environment with various amounts of concurrency to understand the ideal configuration for your environment.

Read performance

When data is read from a compressed volume, the impact on the read performance varies depending on the access patterns, the amount of compression savings on disk, and how busy the system resources are (CPU and disk).

In a sample test with a 50% CPU load on the system, read throughput from a dataset with 50% compressibility showed decreased throughput of 25%. On a typical system, the impact might be higher because of the additional load on the system. Typically, the most impact is seen on small random reads of highly compressible data, and on a system that is more than 50% CPU busy. Impact on performance will vary and must be tested before implementing in production.

20.4 Compression examples

This section describes the steps necessary to enable and configure compression and deduplication on a FlexVol volume using Data ONTAP 8.1, operating in 7-mode.

The first example describes the process of creating a new flexible volume and then configuring, running, and monitoring compression and deduplication on it.

The second example describes the process of adding compression and deduplication to an already existing flexible volume that already contains data. We then run compression to get savings on the already existing data on disk.

> **Tip:** The steps are spelled out in detail, so the process appears much longer than it would be in the real world.

20.4.1 Creating a new volume and enabling compression and deduplication

This example creates a place to archive several large data files.

Step 1

Create a flexible volume (no larger than the maximum volume size limit for your system) (Example 20-1).

Example 20-1 Create a flexvol

```
nas2> vol create volArchive aggrTest 200g
Creation of volume 'volArchive' with size 200g on containing aggregate
'aggrTest' has completed.
```

Step 2

Enable deduplication on the flexible volume (sis on), followed by compression (**`sis config –C true –I true`**) (**`-I`** is only required if you want to use inline compression), and verify that it is turned on. The **`sis config`** command shows the compression and deduplication configuration for flexible volumes.

After you turn deduplication on, Data ONTAP lets you know that if it was an existing flexible volume that already contained data before deduplication was enabled, you need to run **`sis start –s`**. In this example, there is a brand-new flexible volume, so it is not necessary (Example 20-2).

Example 20-2 Enable deduplication

```
nas2> sis on /vol/volArchive

SIS for "/vol/volArchive" is enabled.
```

```
Already existing data could be processed by running "sis start -s
/vol/volArchive".

nas2> sis config -C true -I true /vol/volArchive

nas2> sis config /vol/volArchive

Inline
Path Schedule Compression Compression
-------------------- ------------ ----------- -----------
/vol/volArchive sun-sat@0 Enabled Enabled
```

Step 3

Another way to verify that deduplication is enabled on the flexible volume is to check the output from running **sis status** on the flexible volume (Example 20-3).

Example 20-3 Check the sis status

```
nas2> sis status /vol/volArchive

Path State Status Progress
/vol/volArchive Enabled Idle Idle for 00:03:19
```

Step 4

Turn off the default deduplication schedule (Example 20-4).

Example 20-4 Disable the deduplication schedule

```
nas2> sis config /vol/volArchive

Inline
Path Schedule Compression Compression
-------------------- -------------- ------------------ -------------------
/vol/volArchive sun-sat@0 Enabled Enabled

nas2> sis config -s - /vol/volArchive

nas2> sis config /vol/volArchive

Inline
Path Schedule Compression Compression
-------------------- ------------ ------------------ ------------------
/vol/volArchive - Enabled Enabled
```

Step 5

Mount the flexible volume and copy data into the new archive directory flexible volume.

Step 6

Examine the flexible volume. Use the **df –S** command to examine the storage consumed and the space saved. Note that only compression savings have been achieved so far by simply copying data to the flexible volume, even though deduplication is also turned on. What has happened is that the inline compression compressed the new data as it was written to the volume. Since deduplication was enabled, all the new blocks have had their fingerprints written to the change log file. Until deduplication is actually run, the duplicate blocks will not be removed (Example 20-5).

Example 20-5 Examine the flexible volume

```
nas2*> df -S

Filesystem used total-saved %total-saved deduplicated %deduplicated compressed
%compressed
/vol/volArchive/ 139178264 36126316 21% 0 0% 36126316 21%
```

Step 7

Manually run compression and deduplication on the flexible volume. It causes the compression engine to compress any blocks that were skipped by inline compression followed by processing of the change log, which includes fingerprints to be sorted and merged, and duplicate blocks to be found (Example 20-6).

Example 20-6 Run the sis process

```
nas2> sis start /vol/volArchive

The SIS operation for "/vol/volArchive" is started.
```

Step 8

Use **sis status** to monitor the progress of compression and deduplication operations (Example 20-7).

Example 20-7 Check the sis status

```
nas2> sis status /vol/volArchive

Path State Status Progress
/vol/volArchive Enabled Active 23 GB (77%) Compressed

nas2> sis status /vol/volArchive

Path State Status Progress
/vol/volArchive Enabled Active 164 GB Searched

nas2> sis status /vol/volArchive

Path State Status Progress
/vol/volArchive Enabled Active 39 GB (43%) Done

nas2> sis status /vol/volArchive
Path State Status Progress
/vol/volArchive Enabled Idle Idle for 00:01:03
```

Step 9

When **sis status** indicates that the flexible volume is once again in the Idle state, compression and deduplication have finished running, and you can check the space savings they provided in the flexible volume (Example 20-8).

Example 20-8 View disk space

```
nas2> df –S

Filesystem used total-saved %total-saved deduplicated %deduplicated compressed
%compressed
/vol/volArchive/ 72001016 103348640 59% 64598736 47% 38749904 35%
```

Step 10

Adjust the compression and deduplication schedule as required in your environment.

20.4.2 Enabling compression and deduplication on an existing volume

This example adds compression and deduplication savings to an existing flexible volume with data already on the volume. While it is not necessary, this example includes the steps involved if you wanted to compress and deduplicate the data already on disk, in addition to the new writes to disk. The destination N series storage system is called fas6070c-ppe02, and the volume is /vol/volExisting.

Step 1

Enable deduplication on the flexible volume (sis on), followed by compression (**sis config –C true –I true**) (**-I** is only required if you want to use inline compression), and verify that it is turned on. The **sis config** command shows the compression and deduplication configuration for flexible volumes (Example 20-9).

Example 20-9 Enable deduplication

```
nas2> sis on /vol/volExisting

SIS for "/vol/volExisting" is enabled.
Already existing data could be processed by running "sis start -s
/vol/volExisting".

nas2> sis config -C true -I true /vol/volArchive

nas2> sis config /vol/volArchive

Inline
Path Schedule Compression Compression
-------------------- --------------- ------------------- -------------------
/vol/volArchive      sun-sat@0       Enabled             Enabled
```

Step 2

Examine the flexible volume. Use the **df –S** command to examine the storage consumed and the space saved (Example 20-13).

Example 20-10 View disk space

```
nas2> df -S

Filesystem used total-saved %total-saved deduplicated %deduplicated compressed
%compressed
/vol/volExisting/ 173952092 0 0% 0 0% 0 0%
```

At this time, only new data will be compressed and have fingerprints created. From here, if you want to compress/deduplicate only new writes, you can skip to step 9.

If you want to compress/deduplicate existing data on disk, the following additional steps are required.

Step 3

Disable the post-process compression and deduplication schedule (Example 20-11).

Example 20-11 Disable the post-process schedule

```
nas2> sis config /vol/volExisting

Inline
Path Schedule Compression Compression
-------------------- --------------- ------------------- -------------------
/vol/volExisting sun-sat@0 Enabled Enabled

nas2> sis config -s - /vol/volExisting

nas2> sis config /vol/volExisting

Inline
Path Schedule Compression Compression
-------------------- --------------- ------------------- -------------------
/vol/volExisting - Enabled Enabled
```

Step 4

Record the current Snapshot schedule for the volume. Disable the snap schedule (Example 20-12).

Example 20-12 View Snapshot schedule

```
nas2> snap sched volExisting

Volume volExisting: 0 2 6@8,12,16,20
fas6070-ppe02> snap sched volExisting 0 0 0
```

Step 5

Delete as many Snapshot copies as possible (Example 20-13).

Example 20-13 Delete Snapshots

```
nas2> snap list volExisting

Volume volExisting
working...
```

```
%/used %/total date name
---------- ---------- ------------ --------
23% ( 0%) 20% ( 0%) Jul 18 11:59 snap.1
24% ( 1%) 20% ( 0%) Jul 18 23:59 snap.2
26% ( 3%) 23% ( 2%) Jul 19 11:59 snap.3
26% ( 0%) 23% ( 0%) Jul 19 23:59 snap.4
27% ( 2%) 24% ( 1%) Jul 20 11:59 snap.5

nas2> snap delete volExisting snap.1

nas2> Tue Jul 20 16:20:06 EDT [wafl.snap.delete:info]: Snapshot copy sn ap.1 on
volume volExisting NetApp was deleted by the Data ONTAP function snapcmd _delete.
The unique ID for this Snapshot copy is (3, 6768).

nas2> snap delete volExisting snap.2

nas2> Tue Jul 20 16:20:10 EDT [wafl.snap.delete:info]: Snapshot copy sn ap.2 on
volume volExisting NetApp was deleted by the Data ONTAP function snapcmd _delete.
The unique ID for this Snapshot copy is (2, 6760).

nas2> snap delete volExisting snap.3

nas2> Tue Jul 20 16:20:15 EDT [wafl.snap.delete:info]: Snapshot copy sn ap.3 on
volume volExisting NetApp was deleted by the Data ONTAP function snapcmd _delete.
The unique ID for this Snapshot copy is (4, 6769).

nas2> snap list volExisting

Volume volExisting
working...
%/used %/total date name
---------- ---------- ------------ --------
12% ( 0%) 10% ( 0%) Jul 19 23:59 snap.4
13% ( 1%) 10% ( 1%) Jul 20 11:59 snap.5
```

Step 6

Start compression and deduplication of the existing data on the volume (run this task during low system usage times). You can run with the **–D** option if you only want to run deduplication on the existing data. You can run with the **–C** option if you only want to compress the existing data. However, the next time the deduplication process runs, it will deduplicate the existing data (Example 20-14).

Example 20-14 Start compression and deduplication

```
nas2> sis start -s /vol/volExisting

The file system will be scanned to process existing data in /vol/volExisting.
This operation may initialize related existing metafiles.
Are you sure you want to proceed (y/n)? y

The SIS operation for "/vol/volExisting" is started.
[fas6070-ppe02:wafl.scan.start:info]: Starting SIS volume scan on volume
volExisting.
```

Step 7

Use **sis status** to monitor the progress of compression and deduplication (Example 20-15).

Example 20-15 Monitor sis status

```
nas2> sis status /vol/volExisting

Path State Status Progress
/vol/volExisting Enabled Active 122 GB Scanned, 25 GB Compressed

nas2> sis status /vol/volExisting

Path State Status Progress
/vol/volExisting Enabled Active 164 GB Searched

nas2> sis status /vol/volExisting

Path State Status Progress
/vol/volExisting Enabled Active 41 GB (45%) Done

nas2> sis status /vol/volExisting

Path State Status Progress
/vol/volExisting Enabled Idle Idle for 00:02:43
```

Step 8

When **sis status** indicates that the flexible volume is once again in the Idle state, compression and deduplication have finished running, and you can check the additional space savings they provided in the flexible volume (Example 20-16).

Example 20-16 View disk space

```
nas2> df –S

Filesystem used total-saved %total-saved deduplicated %deduplicated compressed
%compressed
/vol/volExisting/ 72005148 103364200 59% 64621288 47% 38742912 35%
```

Step 9

Reconfigure the Snapshot schedule for the volume (Example 20-17).

Example 20-17 Configure the Snapshot schedule

```
nas2> snap sched volExisting 5 7 10
```

Step 10

Adjust the compression/deduplication schedule as required in your environment.

21

IBM Real-time Compression Appliance

This chapter briefly describes the features and functions of the IBM Real-time Compression™ Appliance (RTCA).

For more information about IBM Real-time Compression, see the Introduction to the Redbooks publication, *IBM Real-time Compression Appliance*, which can be found at the following website:

http://www.redbooks.ibm.com/abstracts/sg247953.html?Open

The following topics are covered:

- Introduction to data compression
- IBM Real-time Compression
- Benefits
- IBM RTCA RACE technology

21.1 Introduction to data compression

The industry need for data compression is clearly for it to be fast, reliable, and scalable. The compression algorithm used must assure data consistency and a very good compression rate in order to be implemented. In addition, the data compression solution must also be easy to implement. The compression must occur without impacting the production use of the data at any time. A generic overview of the RTCA solution is presented in Figure 21-1.

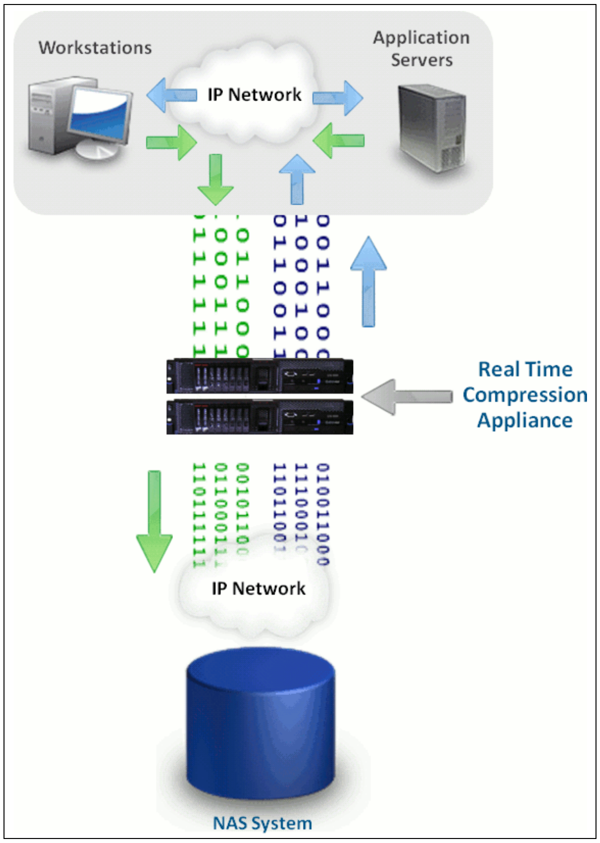


Figure 21-1 Real-time Compression Appliance overview

21.2 IBM Real-time Compression

To understand the basic design of the IBM Real-time Compression technology, we need to review in detail the basics of modern compression techniques.

The IBM Real-time Compression Appliance (IBM RTCA) is based on a reversible data compression algorithm that operates in a real-time method.

The IBM RTCA product compresses data on initial write in order to assure that less data is stored on primary storage. As a result, the storage system has to process less data, using less CPU overhead and lower disk spindles utilization. The storage system can therefore serve more requests from its read/write cache, while some reads can be served from the RTCA product’s read-ahead cache.

In addition to compressing data in real-time, the IBM RTCA product also enables its customers to non-disruptively compress existing data that is already saved to disk with the Compression Accelerator utility. Compression Accelerator is a high-performance and intelligent software application running on the IBM RTCA product which, by policy, allows users to compress data that has already been saved to disk while that data remains online and accessible by applications and end users.

The policies allow users to throttle how decompressed data gets compressed so as not to have an impact on existing storage performance. The ability to compress already stored data significantly enhances and accelerates the benefit to end users, allowing them to see a tremendous return on their IBM RTCA investment. On initial purchase of an IBM RTCA product, users can defer their purchase of new storage. As new storage needs to be acquired, IT purchases less than half of the “required” storage before compression. The IBM RTCA product enables IT to save on their overall storage investment.

21.3 Benefits

IBM Real-time Compression Appliance solutions enable five key benefits:

- Real-time efficient operation: Supports the performance and accessibility requirements of business-critical applications because data is compressed up to 80% in real time, without performance degradation.
- Transparency: 100% transparent to systems, storage, and applications. Provides compatibility with downstream processes such as backups, Snapshots, cloning, mirroring and archiving. And it complements deduplicated environments.
- Non-disruptive: Requires no changes to applications, servers or storage systems. All IT processes remain the same.
- Less Cost, Greener: Cost reduction benefits carry throughout the storage lifecycle: less storage to power, cool, and manage means less cost and a greener data center.
- Performance: By offloading the compression to an appliance, the storage controller is not handling the compression itself and has more processor cycles free for serving storage.

21.4 IBM RTCA RACE technology

The IBM Random Access Compression Engine (RACE) technology is the core of IBM RTCA products for NAS. RACE technology is based on 35 patents that are not about compression. Rather, they define how to make industry standard LZ compression of primary storage operate in real-time and allow random access. The primary intellectual property behind it is our RACE engine. The IBM RACE engine sits on an appliance in front of any NFS or CIFS deployment, acting as an “intelligent cable” between the IP switch and the storage. No software agents or drivers are required on clients or servers.

The RACE technology (see Figure 21-2) is made up of three components:

- Random Access Compression Engine (RACE): Enables random-access data compression without compromising performance.
- Unified Protocol Manager (UPM): Enables transparent support of multiple storage and network protocols, including CIFS and NFS.
- Monitoring and Reporting Manager (MRM): Enables online storage compression trending, analysis, and reporting.

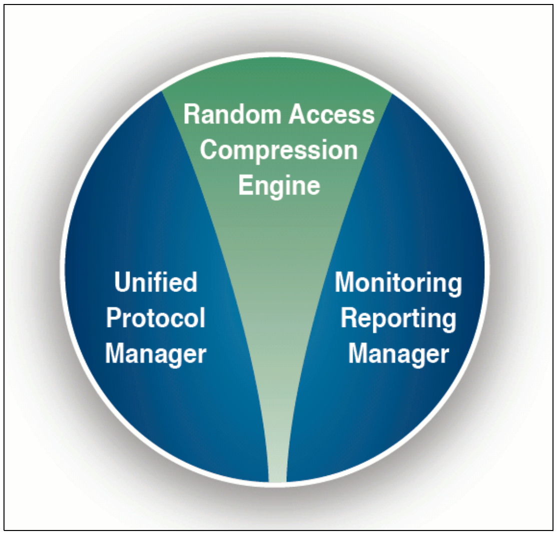


Figure 21-2 RACE technology overview

The traditional compression technologies start from a constant file size and, after compression, the result is a variable file in terms of capacity. As a result, when using large data chunks, the performance impact is high. However, when using small data chunks, although the performance impact is small, the compression ratio is also very small. Over time, there are many disadvantages that can occur. They include the need for garbage collection, poor performance while the volume of the data increases, or losing parts of metadata, such as the date of creation, date accessed, user rights, or modification dates. Another issue can be fragmentation in the target storage space. It occurs because after the file is stored in its original size, the result of compression is stored in a new zone and then the input is deleted.

The Random Access Compression Engine (RACE) starts from an unknown data stream and compresses data coming from the host. The resulting compressed file keeps all attributes from the original; metadata is not changed. Also, because of the in-line approach, there is no need at the storage level to write original data, read it, write the result of compression, and finally, delete the initial file. At the end, there is not any garbage or fragmentation on the storage system. The performance needed at the storage level is decreased because the writes and reads are made only in compressed format instead of a complete one.

A logical overview of the Random Access Compression Engine is presented in Figure 21-3.

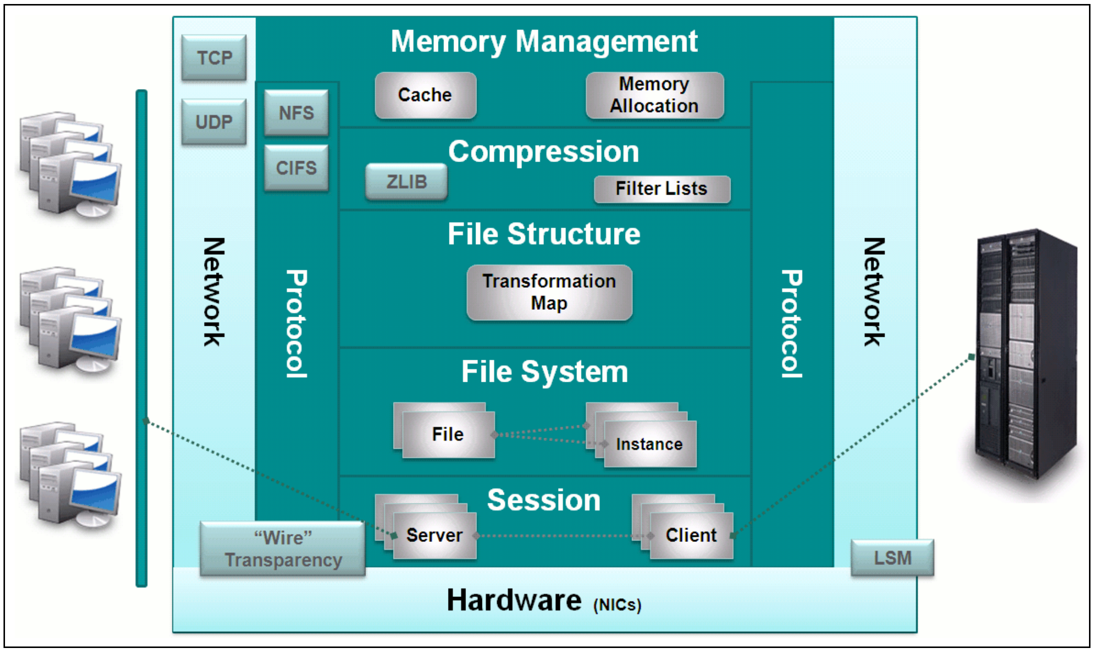


Figure 21-3 Random Access Compression Engine

RACE takes incoming data streams and compresses the data within these data requests, leaving the metadata intact, to the storage array. The data is stored in the array and the acknowledgement that the write has been committed is sent directly back from the array to the end user or application. This process flow is important because from a data availability perspective, it is imperative that it is the storage array that acknowledges the write commitment, not the IBM RTCA product. It is for this reason that the IBM RTCA product has no write cache. All storage commits come from the array, preserving the integrity of the data between the storage and the application.

As stated before, the IBM RTCA technology uses industry standard LZ compression algorithms. The “secret sauce” is not the compression algorithms that the RTCA product uses to do its compression, but rather the manner in which that compression is accomplished. One of the key ways that the RTCA product is able to achieve its high compression ratios and performance is by compressing data utilizing random access techniques.

The benefits of compressing data using random access techniques are twofold. First, the ability to read or write only the blocks of the compressed file that require read or modification means faster access performance for these operations. If you only need to write a small piece of data in order to update a whole file, your storage performance is maximized. Second, because the RTCA product has this capability, updates to the file are accomplished in a way that does not disrupt the other blocks in the compressed file.

We are operating under the assumption that upstream data compression significantly reduces downstream data deduplication ratios. Therefore it is preferable to apply data deduplication technologies over decompressed data prior to performing a deduplicated backup. It can be true for data compressed with traditional techniques. But the unique, random access nature of the IBM RTCA product's compression preserves data deduplication ratios. It allows end users to experience maximum data optimization in both primary and downstream tiers.

22

Thin replication using SnapVault and Volume SnapMirror

This chapter introduces the topic of thin replication using SnapVault and Volume SnapMirror. Thin replication refers to the copying of data to another facility during backup and disaster recovery. Snapshot copies can be backed up and replicated to another facility using SnapVault and Volume SnapMirror. These two technologies increase storage efficiency by transferring only changed data blocks after the baseline copy is created.

This chapter also explains how you can combine deduplication with these thin replication technologies to achieve greater savings by eliminating redundant data. Furthermore, you can use compression with these technologies to reduce the size of the replicated data, saving network bandwidth. The data on a SnapVault or volume SnapMirror source is compressed, transferred over the network, and then uncompressed on the destination before being written to the disk.

Starting with Data ONTAP 8.1, you can replicate volumes by using SnapMirror technology between 32-bit and 64-bit volumes. For both synchronous and asynchronous volume replication, the SnapMirror source and destination volumes can be either 32-bit or 64-bit.

The following topics are covered:

- Disk-to-Disk backups using SnapVault
- Efficient data protection using volume SnapMirror

Figure 22-1 illustrates how SnapVault and SnapMirror store data using thin transfers.

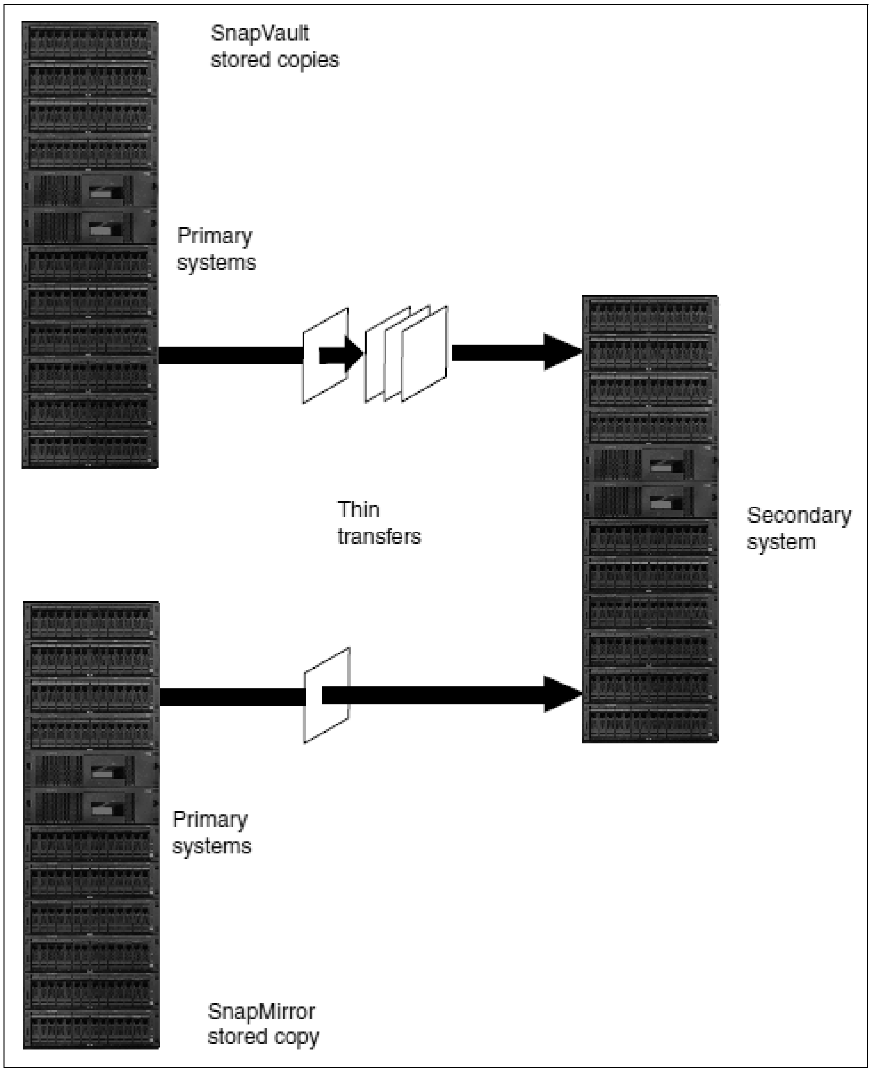


Figure 22-1 How SnapVault and SnapMirror store data using thin transfers

22.1 Disk-to-Disk backups using SnapVault

SnapVault uses network bandwidth efficiently, because it transfers only the blocks that changed since the last Snapshot copy. It automatically eliminates the duplication that results from other backup technologies, such as tape.

Furthermore, deduplication facilitates space reduction at the source and destination systems and during the data transfer between the two systems.

22.1.1 What data gets backed up and restored through SnapVault

The data structures that are backed up and restored through SnapVault depend on the primary system:

- On systems running Data ONTAP, the qtree is the basic unit of SnapVault backup and restore. SnapVault backs up specified qtrees on the primary system to associated qtrees on the SnapVault secondary system. If necessary, data is restored from the secondary qtrees back to their associated primary qtrees. See Figure 22-2.

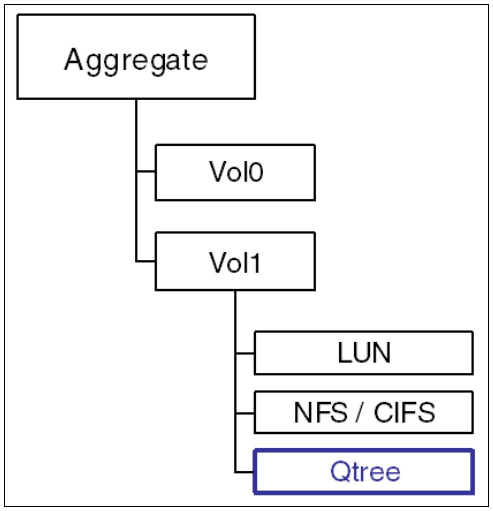


Figure 22-2 Qtree representation

- On open systems storage platforms, the directory is the basic unit of SnapVault backup. SnapVault backs up specified directories from the native system to specified qtrees in the SnapVault secondary system.

  If necessary, SnapVault can restore an entire directory or a specified file to the open systems platform.

> **Tip:** You can back up the qtrees from multiple primary systems, or directories from multiple open systems storage platforms, to associated qtrees on a single SnapVault secondary volume.

Figure 22-3 shows the backup of qtrees and directories on different systems to a single secondary volume.

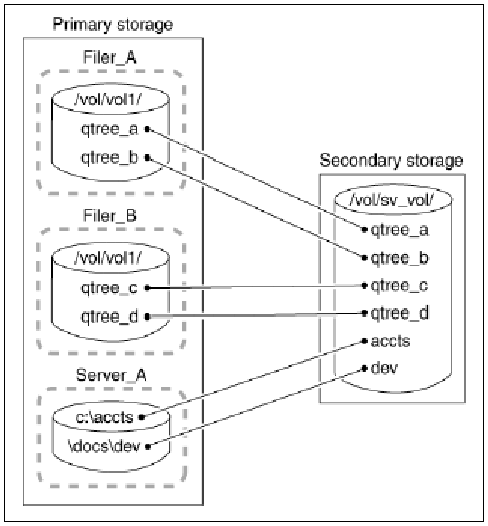


Figure 22-3 Backup of qtree and directories from primary storage to secondary storage

22.1.2 Types of SnapVault deployment

You can deploy SnapVault in three ways for your business requirements:

- Basic SnapVault deployment
- Primary to secondary to tape backup variation
- Primary to secondary to SnapMirror variation

Basic SnapVault deployment

The basic SnapVault backup system deployment consists of a primary system and a secondary system.

Primary storage system: Primary systems are the platforms that run Data ONTAP and open systems storage platforms to be backed up:

- On primary systems, SnapVault backs up primary qtree data, non-qtree data, and entire volumes, to qtree locations on the SnapVault secondary systems.
- Supported open systems storage platforms include Windows servers, Solaris servers, AIX servers, Red Hat Linux servers, SUSE Linux servers, and HP-UX servers. On open systems storage platforms, SnapVault can back up directories to qtree locations on the secondary system.

Secondary storage system: The SnapVault secondary system is the central disk-based unit that receives and stores backup data from the system as Snapshot copies. Any system can be configured as a SnapVault secondary system. However, the preferable hardware platform is an IBM Near-line system, which is a solution that offers IBM System Storage N series populated with SATA disk drives. The NearStore feature is designed to address those storage challenges by providing near-primary storage performance at a significantly lower cost.

Figure 22-4 shows a basic SnapVault deployment.

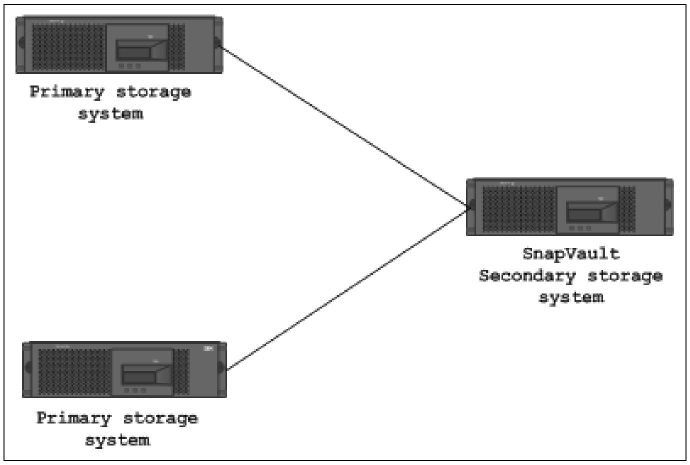


Figure 22-4 Basic SnapVault deployment

Primary to secondary to tape backup variation

A common variation to the basic SnapVault backup deployment adds a tape backup of the SnapVault secondary system.

This deployment can serve two purposes:

- It enables you to store an unlimited number of network backups offline while keeping the most recent backups available online in secondary storage. It can help in the quick restoration of data. If you run a single tape backup off the SnapVault secondary storage system, the storage platforms are not subject to the performance degradation, system unavailability, and complexity of direct tape backup of multiple systems.
- It can be used to restore data to a SnapVault secondary system in case of data loss or corruption on that system.

> **Tip:** Some UNIX attributes are not preserved using this method; notably, UNIX access control lists (ACLs).

Figure 22-5 shows a basic SnapVault deployment with tape backup.

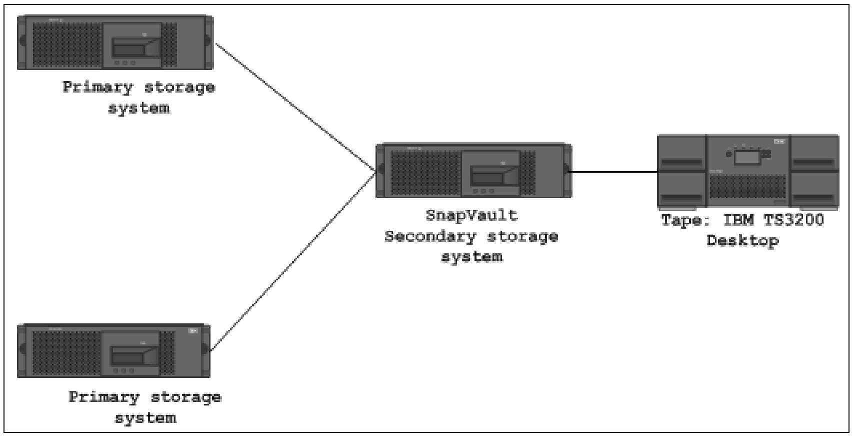


Figure 22-5 Basic SnapVault deployment with tape backup

Primary to secondary to SnapMirror variation

In addition to the basic SnapVault deployment, you can replicate the SnapVault secondary using SnapMirror. This protects the data stored on the SnapVault secondary against problems with the secondary system itself.

The data backed up to SnapVault secondary storage is replicated to a SnapMirror destination.

If the secondary system fails, the data mirrored to the SnapMirror destination can be converted to a secondary system and used to continue the SnapVault backup operation with minimum disruption. Figure 22-6 shows an example of SnapVault with SnapMirror.

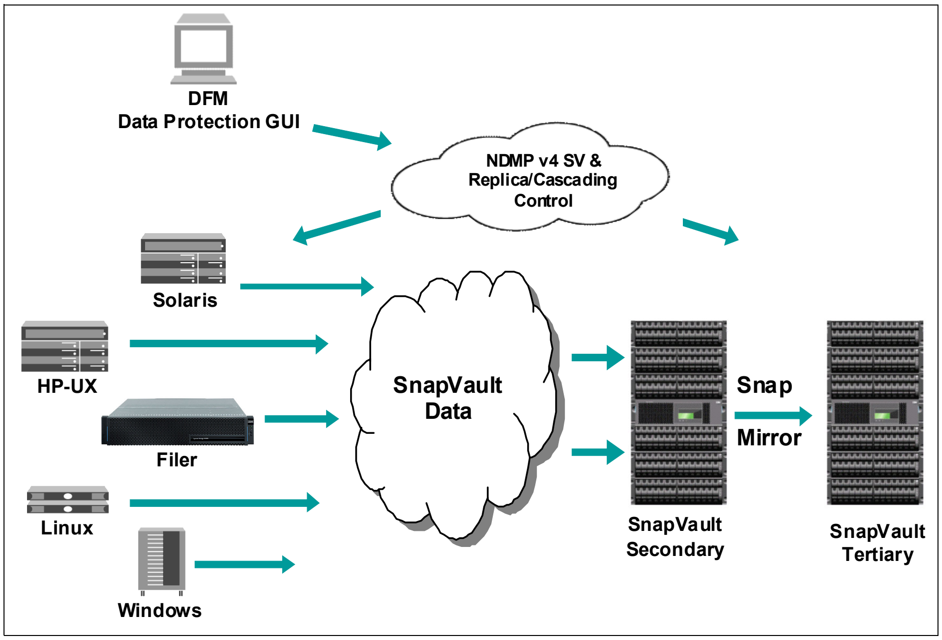


Figure 22-6 SnapVault with SnapMirror

22.1.3 How SnapVault backup works

Backing up qtrees using SnapVault involves five main steps:

1. Add license to primary and secondary systems.
2. Disable normal Snapshot schedules.
3. Schedule incremental transfers.
4. Start the baseline transfers (first initial full backup).
5. Restore data upon request.

Tip: License on primary and secondary filers must be activated. Use the **`sv_ontap_pri`** license for primary system and **`sv_ontap_se`**c license for secondary system.

Disabling normal Snapshot schedules

SnapVault Snapshots will replace the normal Snapshots, so it is required to disable the normal Snapshots schedules as they will be managed by SnapVault.

Scheduling incremental transfers

Each primary system, in response to command-line input, creates sets of scheduled SnapVault Snapshot copies of the volumes containing the qtrees to be backed up. For tracking purposes, you might name according to frequency, for example, sv_hourly, sv_nightly, and so on.

For each Snapshot set, SnapVault saves the number of primary storage Snapshot copies you specify and assigns each Snapshot a version number (0 for most current, 1 for second most recent, and so on).

The SnapVault secondary system, in response to command-line input, carries out a specified set of scheduled data transfer and Snapshot actions. For each of its secondary qtrees on a given volume, SnapVault retrieves, from the Snapshot data of each corresponding primary qtree, the incremental changes to the primary qtrees made since the last data transfer.

Then SnapVault creates a volume Snapshot copy of the changes in the secondary qtrees. For each transfer and Snapshot set, SnapVault saves the number of secondary storage Snapshot copies that you specify and assigns each Snapshot copy a version number (0 for most current, 1 for second most recent, and so on).

Example 22-1 shows how to schedule SnapVault incremental transfers.

Example 22-1 Scheduling SnapVault incremental transfers

```
itsotuc1>snapvault snap sched vol1 sv_hourly 22@0-22
itsotuc1>snapvault snap sched oracle sv_daily 7@23
```

Starting initial baseline transfers

In response to command-line input, the SnapVault secondary system requests initial base transfers of qtrees specified for backup from a primary storage volume to a secondary storage volume. These transfers establish SnapVault relationships between the primary and secondary qtrees.

Each primary system, when requested by the secondary system, transfers initial base images of specified primary qtrees to qtree locations on the secondary system.

At this point, you have configured schedules on both the primary and secondary systems, and SnapVault is enabled and running. However, SnapVault does not know which qtrees to back up, or where to store them on the secondary. Snapshots are taken on the primary, but no data is transferred to the secondary.

To provide SnapVault with this information, use the **`snapvault start`** command on the secondary.

Restoring data upon request

One of the unique benefits of SnapVault is that users do not require special software or privileges to perform a restore of their own data. Users who want to restore their own data can do so without the intervention of a system administrator, saving time and money.

Important: When trying to restore from a SnapVault secondary, connectivity to the secondary must be in place. Users must know where it is located in order to avoid data corruption and overwriting files in the wrong place.

Restoring a file from a SnapVault backup is simple. Just as the original file was accessed by an NFS mount or CIFS share, the SnapVault secondary can be configured with NFS exports and CIFS shares. As long as the destination qtrees are accessible to the users, restoring data from the SnapVault secondary is as simple as copying from a local Snapshot.

Users can restore an entire data set the same way, assuming that the appropriate access rights are in place. However, SnapVault provides a simple interface to restore an entire data set from a selected Snapshot using the **snapvault restore** command on the SnapVault primary (Example 22-2). This command must used only by filers administrators because it can overwrite data.

Example 22-2 SnapVault restore syntax

```
itsotuc1> snapvault restore
usage:
snapvault restore [-f] [-s <snapname>] [-k <n>] -S <secondary_filer>:<secondary_
path> [<primary_filer>:]<primary_path>
```

> **Tip:** When you use **`snapvault restore`**, the command prompt does not return until the restore has completed. If the restore needs to be cancelled, press Ctrl-c.

Figure 22-7 illustrates SnapVault functionality:

- Protects multiple qtrees/volumes
- On multiple primary storage systems
- On a specific secondary storage system

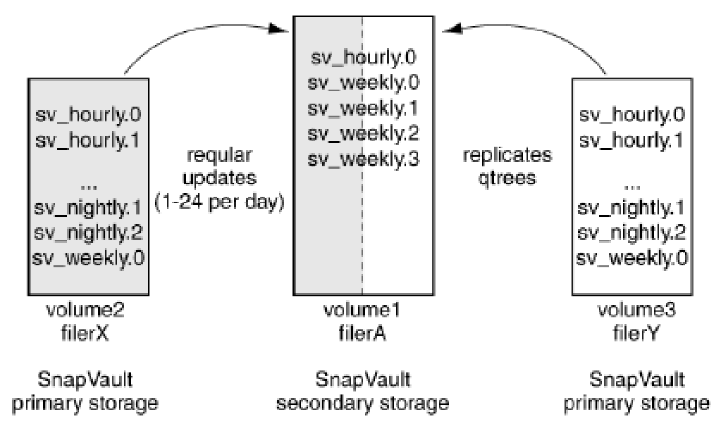


Figure 22-7 SnapVault functionality

22.1.4 Guidelines for creating a SnapVault relationship

When creating a SnapVault relationship, follow these guidelines for volumes and qtrees.

- Establish a SnapVault relationship between volumes that have the same vol language code settings. Use the **`vol lang [vol_name]`** command to check the settings.
- After you establish a SnapVault relationship, do not change the language assigned to the destination volume.
- Avoid white space (spaces and tab characters) in names of source and destination qtrees.
- Do not rename volumes or qtrees after establishing a SnapVault relationship.
- The qtree cannot exist on the secondary system before the baseline transfer.

22.1.5 About LUN clones and SnapVault

A LUN clone is a space-efficient copy of another LUN. Initially, the LUN clone and its parent share the same storage space. More storage space is consumed only when one LUN or the other changes.

> **Tip:** LUNs in this context refer to the LUNs that Data ONTAP serves to clients, not to the array LUNs used for storage on a storage array.

SnapDrive for Windows

Starting with Data ONTAP 7.3, SnapVault can transfer LUN clones in an optimized way by using SnapDrive for Windows. To manage this process, SnapDrive for Windows creates two Snapshot copies:

- Backing Snapshot copy, which contains the LUN to be cloned
- Backup Snapshot copy, which contains both the LUN and the clone

Modes of transfer

Starting with Data ONTAP 7.3, a SnapVault transfer with LUN clones can run in two modes:

- In non-optimized mode, a LUN clone is replicated as a LUN. Therefore, a LUN clone and its backing LUN get replicated as two separate LUNs on the destination. SnapVault does not preserve space savings that come from LUN clones.
- In optimized mode, a LUN clone is replicated as a LUN clone on the destination. Transfers of LUN clones to the secondary system in optimized mode are possible only with SnapDrive for Windows.

These modes apply to newly created LUN clones. On successive update transfers, only the incremental changes are transferred to the destination in both modes.

> **Attention:** A single relationship must either be optimized or non-optimized. Switching between the two modes is not allowed.

22.2 Efficient data protection using volume SnapMirror

Volume SnapMirror provides an easy-to-administer replication solution that makes efficient use of available network bandwidth by transferring only changed blocks. If a disaster occurs, businesses can access data from a replica on a remote storage system for uninterrupted operation.

When mirroring asynchronously, SnapMirror replicates Snapshot copies from a source system to a destination system. When an update occurs, a new Snapshot copy is created and is compared against the previous Snapshot copy to determine the changes since the last update. Only the new and changed blocks are sent to the destination system. At the destination system, the changed blocks are merged with the existing data blocks resulting in a full mirror copy of the source system.

Because SnapMirror is based on the Snapshot copy technology and also seamlessly integrates with deduplication, it consumes minimal storage space and saves on network bandwidth. When volume SnapMirror is combined with deduplication any savings on the SnapMirror source volume are inherited at the destination volume.

22.2.1 How SnapMirror works

SnapMirror replicates data from a source volume or qtree to a partner destination volume or qtree, respectively, by using Snapshot copies. Before using SnapMirror to copy data, you need to establish a relationship between the source and the destination.

You can specify a SnapMirror source and destination relationship between volumes or qtrees by using one of the following options.

- The /etc/snapmirror.conf file
- The /etc/snapmirror.allow file
- The snapmirror.access option

Example 22-3 shows how to check and manage the three options just listed.

Example 22-3 SnapMirror source and destination relationship

```
itsotuc*> rdfile /etc/snapmirror.conf
#Regenerated by registry Mon Apr 11 21:22:38 GMT 2011
9.11.218.110:Source_Volume itsotuc2:Mirror_Volume - * * * *

itsotuc*> rdfile /etc/snapmirror.allow
9.11.218.238
9.11.218.110

itsotuc*> options snapmirror
snapmirror.access            host=9.11.218.110
snapmirror.checkip.enable    off
snapmirror.delayed_acks.enable on
snapmirror.enable            on
snapmirror.log.enable        on
snapmirror.vbn_log_enable    off        (value might be overwritten in takeover)
```

The SnapMirror feature does the following tasks:

1. It creates a Snapshot copy of the data on the source volume.
2. It copies it to the destination, a read-only volume or qtree.
3. It updates the destination to reflect incremental changes on the source, according to the schedule you specify.

The result of this process is an online, read-only volume or qtree that contains the same data as the source at the time of the most recent update.

Each volume SnapMirror replication, qtree SnapMirror replication, or SnapVault replication consists of a pair of operations. There is one operation each at these locations:

- The source storage system
- The destination storage system

Therefore, if a storage system is the source for one replication and the destination for another replication, it uses two replication operations. Similarly, if a storage system is the source as well as the destination for the same replication, it uses two replication operations.

22.2.2 SnapMirror use cases

SnapMirror is used to replicate data. Its qualities make SnapMirror useful in several scenarios, including disaster recovery, data backup, and data restoration.

You can copy or use the data stored on a SnapMirror destination. The additional advantages of SnapMirror make it useful in data retrieval situations such as those described in Table 22-1.

Table 22-1 SnapMirror data retrieval situations and usage

Situation	How to use SnapMirror
Disaster recovery: You want to provide immediate access to data after a disaster has made a qtree, volume, or system unavailable.	You can make the destination writable so clients can use the same data that was on the source volume the last time data was copied.
Disaster recovery testing: You want to test the recovery of data and restoration of services in the event of a disaster.	You can use FlexClone technology on the SnapMirror destination, and test for disaster recovery, without stopping or pausing other replication operations.
Data restoration: You want to restore lost data on a qtree or volume source from its mirrored qtree or volume SnapMirror partner.	You can temporarily reverse the roles for the source and destination qtrees or volumes and copy the mirrored information back to its source.
Application testing: You want to use an application on a database, but you want to test it on a copy of the database in case the application damages the data.	You can make a copy of the database to be used in the application testing to ensure that the data on the source cannot be lost.
Load balancing: A large number of users need read-only access to a qtree or volume.	You can copy the data in a qtree or volume to multiple volumes or systems to distribute the load.
Off-loading tape backups: You need to reserve all processing and networking resources on a system for serving NFS and CIFS requests.	After copying data on the source system, you can back up the data in the destination to tape. This means that the source system does not have to allocate resources for performing backups.
Access to remote data: Users who need read access to a volume are distributed over a large geographical area.	You can copy the source volume to other systems that are geographically closer to the users. Users accessing a local system can read the data using less resource time than if they connected to a distant system.

22.2.3 Preferred practices while using SnapMirror

While using SnapMirror, you can increase the efficiency of data copying by performing certain actions. This includes the staggering of Snapshot copy schedules and SnapMirror update schedules:

- To optimize performance, stagger your Snapshot copy update schedules so that SnapMirror activity does not begin or end at the exact minute that a **`snap sched`** command operation attempts to create a Snapshot copy.
- If the SnapMirror feature is scheduled to perform Snapshot copy management at the same time as a **`snap sched`** activity, then the Snapshot copy management operations scheduled using the **`snap sched`** command might fail with syslog messages: `"Skipping creation of hourly snapshot"` and `"Snapshot already exists."`

- For optimum SnapMirror volume replication performance, ensure that the SnapMirror source volume and destination volume contain disks of the same size, organized in the same RAID configuration:
  - If the SnapMirror source and destination are FlexVol volumes, the RAID configurations do not make a difference.
  - If the SnapMirror source and destination are qtrees, volume size and configuration do not make any difference.

22.2.4 SnapMirror deployment variations

There are several variations possible while deploying SnapMirror. These variations allow you to customize the solution to suit your requirements.

Source to destination to tape variation: A common variation to the basic SnapMirror backup deployment adds a tape backup of the destination volume. By running a tape backup off the SnapMirror destination volume (as shown in Figure 22-8), you do not subject the heavily-accessed source volume to the performance degradation and complexity of a direct tape backup.

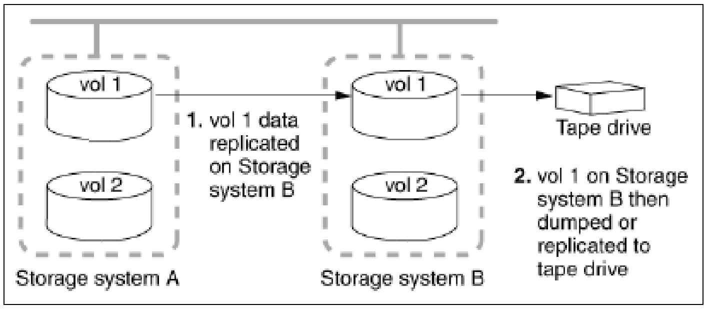

vol 1
vol 2
1. vol 1 data replicated on Storage system B
vol 1
vol 2
Tape drive
2. vol 1 on Storage system B then dumped or replicated to tape drive
Storage system A
Storage system B

Figure 22-8 SnapMirror deployment: Source to destination to tape

Source to tape to destination variation: A SnapMirror deployment that supports SnapMirror replication over low-bandwidth connections accommodates an initial mirroring between a source and destination volume using physically-transported tape (as shown in Figure 22-9). After the large base Snapshot copy has been replicated, smaller, incremental Snapshot copy updates can be carried out over a low-bandwidth connection.

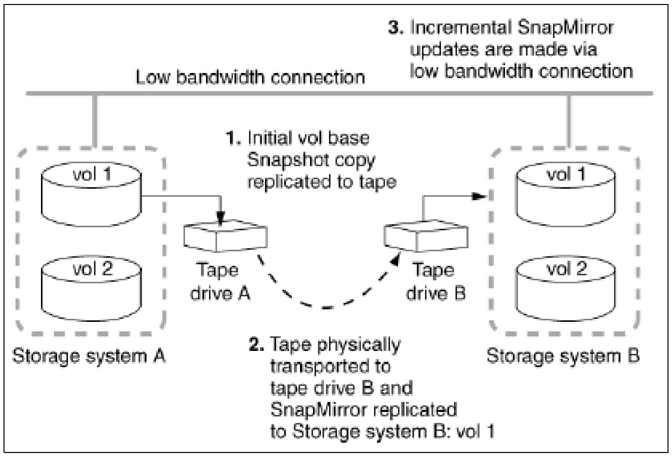

3. Incremental SnapMirror updates are made via low bandwidth connection
Low bandwidth connection
1. Initial vol base Snapshot copy replicated to tape
vol 1
vol 2
Tape drive A
Tape drive B
vol 1
vol 2
Storage system A
2. Tape physically transported to tape drive B and SnapMirror replicated to Storage system B: vol 1
Storage system B

Figure 22-9 SnapMirror deployment: Source to tape to destination

Cascading destinations variation: A variation on the basic SnapMirror deployment and function involves a writable source volume replicated to multiple read-only destinations. The function of this deployment is to make a uniform set of data available on a read-only basis to users from various locations throughout a network and to allow for updating that data uniformly at regular intervals. Figure 22-10 shows a cascade deployment.

Support: The cascade deployment is supported for volume SnapMirror only. It is not supported for qtree SnapMirror.

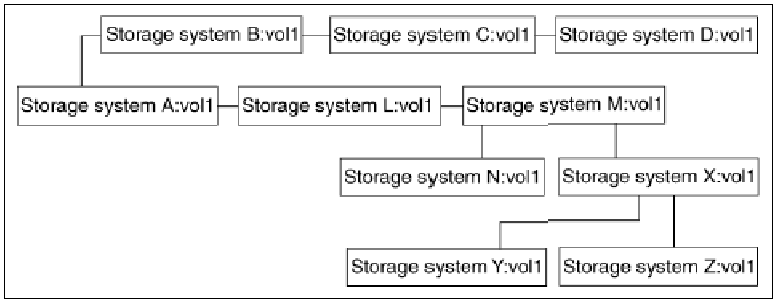


Figure 22-10 SnapMirror deployment: Cascade

22.2.5 Cascading data replication

You can replicate data from a SnapMirror destination to another system using SnapMirror (cascading). Therefore, a system that is a destination for one SnapMirror relationship can act as the source for another SnapMirror relationship. It is useful when you need to copy data from one site to many sites.

Instead of replicating data from a single source to each of the destinations, you can replicate data from one destination to another destination, in a series. It is referred to as cascading.

Tip: You can replicate data from a destination volume in the same way you replicate from a writable source volume.

Part 4

Storage access protocols

With a unified, multiprotocol architecture, the N series network storage solutions take advantage of the benefits of Ethernet storage and work as a "unification engine" supporting network file system (NFS), Common Internet File System (CIFS), iSCSI, and FCoE in the same system, while also providing support for traditional Fibre Channel network storage. The N series offers a wide range of solutions to deliver the consolidation and unification today's businesses need as storage foundation.

In this part of the book, we describe the file and block based protocols available on the N series. The following topics are covered:

- CIFS and Active Directory
- NFS
- Multiprotocol data access
- Fibre Channel
- FCoE
- iSCSI
- Other protocols

23

CIFS and Active Directory

This chapter explains how to set up Common Internet File System (CIFS) shares for use with Microsoft SQL (MS SQL) or other databases, and explains the use of the Active Directory service with an IBM System Storage N series storage system.

The following topics are covered:

- Supported CIFS versions
- Joining the N series CIFS service to Active Directory
- Prerequisite steps for Active Directory integration
- Selecting a user account
- Precreating a computer object
- Running the CIFS setup wizard
- Troubleshooting the domain joining process
- Device discovery
- Automatic home shares

23.1 Supported CIFS versions

Beginning in Data ONTAP 8.1, clients can use the SMB 2.0 protocol to access files on the storage system. With Data ONTAP 7.x and 8.x, the SMB 1.x version is also supported on all N series systems.

23.2 Joining the N series CIFS service to Active Directory

In order for resources on a network to be locatable, a mechanism must exist so that the resources can easily be found. A directory service, in this case, Active Directory, keeps track of all known resources and responds to requests with a list of currently available devices and services. But before you can be trusted to query for resources, you must be granted membership in the Active Directory domain.

Active Directory works on a container basis. A *container* can be a domain, organization unit (OU), or computer.

These are the key benefits that an IBM System Storage N series storage system receives by joining Active Directory:

- Controlled security and management through group management (that is, group policy objects (GPOs) and access control lists (ACLs) placed on objects and organization units (OUs))
- Single sign-on and pass-through authentication for users
- Interoperability by extending control beyond the native Windows environment through the Microsoft management interface by providing a read-only computer management view of the following items:
  - Shared folders, shares, sessions, and open files
  - Local users and groups to the IBM System Storage N series storage system

23.2.1 Data ONTAP

Data ONTAP is a proprietary operating system. It is not based on the Windows operating system. Consequently, the current Data ONTAP operating system requires that additional rights be assigned to the user or to the precreated device object when an administrator or administrator equivalent account is not used.

After the computer object has successfully joined the Active Directory domain, the user account credentials will no longer be used and are not stored in any way in the operating system. They are used only to allow the IBM System Storage N series storage system to become an active member of Active Directory and to write standard properties to the object during the join process.

23.2.2 Machine accounts

Every computer running a Windows workstation (Windows NT 4.0 or later), Windows server operating system, or IBM System Storage N series storage system must have a computer account. System users also require a valid account before being allowed to access a networked resource. Workstations, servers, and other devices participating in an Active Directory domain must have an account too. This account provides a means for authenticating and auditing computer access to the network, and access control, security, and management to domain resources.

> **Reference:** For more information about how to set up and configure CIFS, see the *IBM System Storage N series Data ONTAP 8.0 7-Mode File Access and Protocols Management Guide*, available at this website:
>
> `http://www.ibm.com/storage/support/nas`

23.3 Prerequisite steps for Active Directory integration

Perform the following steps:

1. Determine the host name of the IBM System Storage N series storage system. You can accomplish this task by issuing the **hostname** command (Example 23-1).

   Example 23-1 The hostname command

   ```
   itsotuc1> hostname
   ```

2. Determine the IP address of the IBM System Storage N series by running the **ifconfig -a** command (Example 23-2).

   Example 23-2 The ifconfig -a command

   ```
   itsotuc1> ifconfig -a
   e0a: flags=0x2f48867<UP,BROADCAST,RUNNING,MULTICAST,TCPCKSUM> mtu 1500
           inet 9.11.218.114 netmask 0xffffff00 broadcast 9.11.218.255
           ether 00:a0:98:09:8a:07 (auto-1000t-fd-up) flowcontrol full
   ...
   ```

Our system is available at IP address 9.11.218.114.

3. Verify that the IBM System Storage N series is licensed for CIFS (Example 23-3).

Example 23-3 The license command

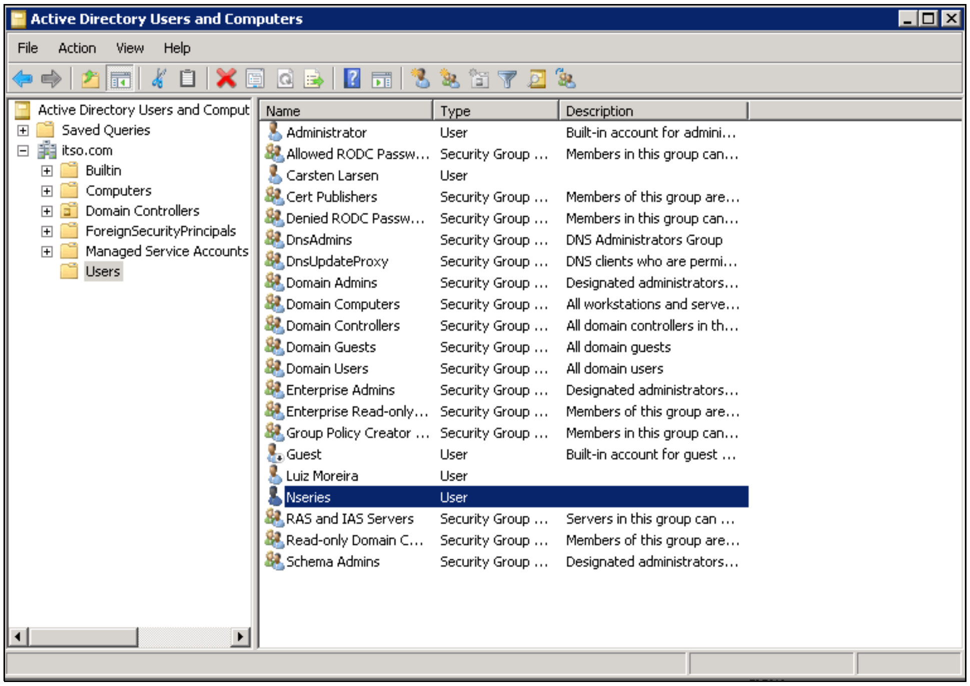
```
itsotuc1> license
                 a_sis not licensed
                  cifs site ABCDEFG
               cluster not licensed
        cluster_remote not licensed
        ...
```

23.4 Selecting a user account

This section describes how to specify the desired user account.

23.4.1 Preparation

You must create or select a user account that will be used for the precreation of the IBM System Storage N series computer object (Figure 23-1).



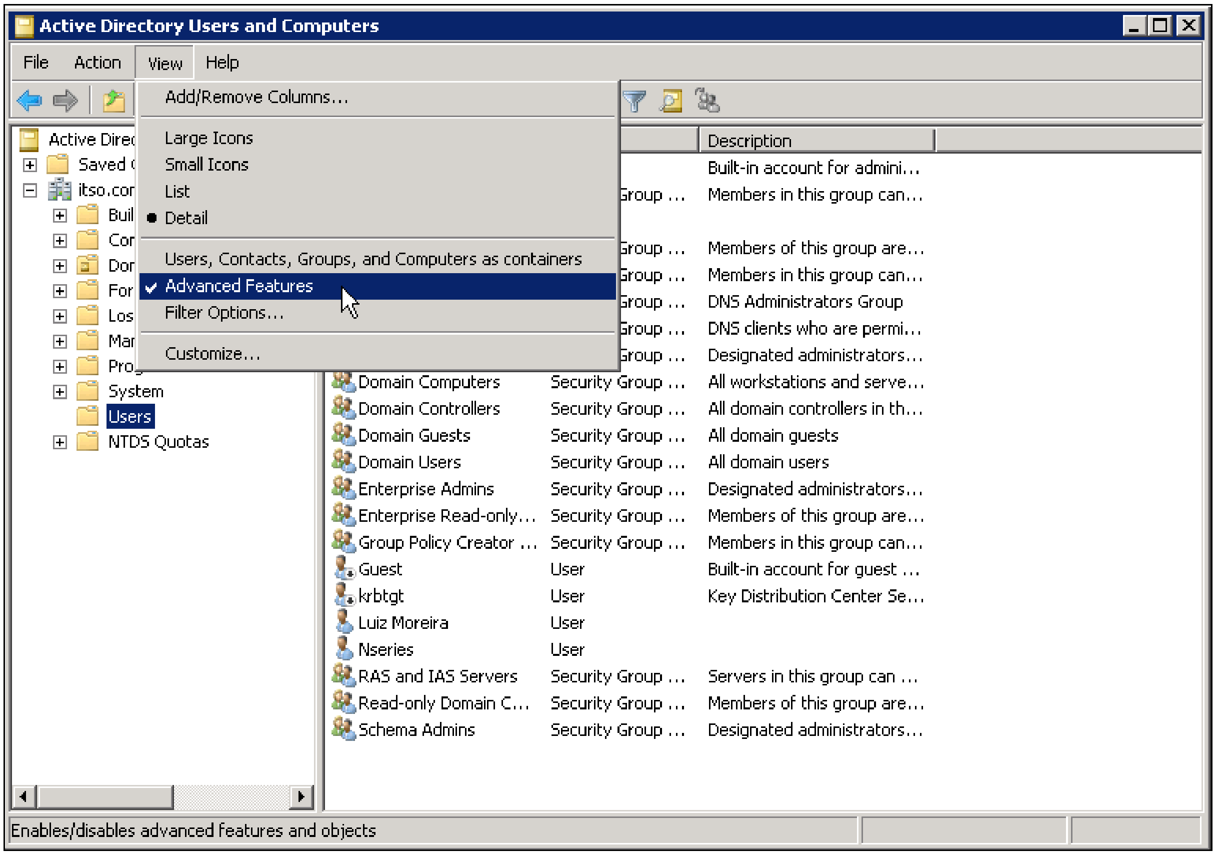
Name	Type	Description
Administrator	User	Built-in account for admini...
Allowed RODC Passw...	Security Group ...	Members in this group can...
Carsten Larsen	User	
Cert Publishers	Security Group ...	Members of this group are...
Denied RODC Passw...	Security Group ...	Members in this group can...
DnsAdmins	Security Group ...	DNS Administrators Group
DnsUpdateProxy	Security Group ...	DNS clients who are permi...
Domain Admins	Security Group ...	Designated administrators...
Domain Computers	Security Group ...	All workstations and serve...
Domain Controllers	Security Group ...	All domain controllers in th...
Domain Guests	Security Group ...	All domain guests
Domain Users	Security Group ...	All domain users
Enterprise Admins	Security Group ...	Designated administrators...
Enterprise Read-only...	Security Group ...	Members of this group are...
Group Policy Creator ...	Security Group ...	Members in this group can...
Guest	User	Built-in account for guest ...
Luiz Moreira	User	
Nseries	User	
RAS and IAS Servers	Security Group ...	Servers in this group can ...
Read-only Domain C...	Security Group ...	Members of this group are...
Schema Admins	Security Group ...	Designated administrators...

Figure 23-1 User to be used to create the IBM System Storage N series computer object

23.4.2 Acquiring rights

Ensure that the N series user *"Nseries"* has the minimum rights for operation by performing these steps:

1. Select **View** from the top of the window and make sure that **Advanced View** is selected (Figure 23-2).

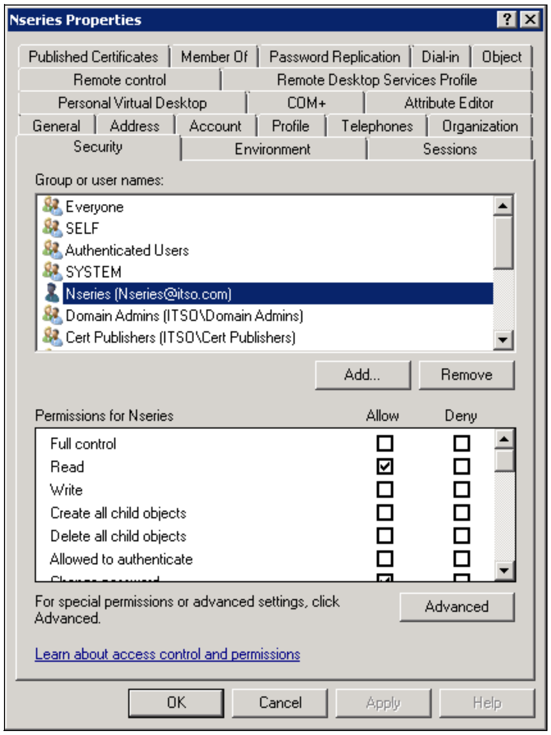


		Description
		Built-in account for admini…
	roup …	Members in this group can…
	roup …	Members of this group are…
	roup …	Members in this group can…
	roup …	DNS Administrators Group
	roup …	DNS clients who are permi…
	roup …	Designated administrators…
Domain Computers	Security Group …	All workstations and serve…
Domain Controllers	Security Group …	All domain controllers in th…
Domain Guests	Security Group …	All domain guests
Domain Users	Security Group …	All domain users
Enterprise Admins	Security Group …	Designated administrators…
Enterprise Read-only…	Security Group …	Members of this group are…
Group Policy Creator …	Security Group …	Members in this group can…
Guest	User	Built-in account for guest …
krbtgt	User	Key Distribution Center Se…
Luiz Moreira	User	
Nseries	User	
RAS and IAS Servers	Security Group …	Servers in this group can …
Read-only Domain C…	Security Group …	Members of this group are…
Schema Admins	Security Group …	Designated administrators…

Enables/disables advanced features and objects

Figure 23-2 Enabling advanced features for users

2. Open the *Nseries* user and select **Security** (Figure 23-3).



Group or user names:

- Everyone
- SELF
- Authenticated Users
- SYSTEM
- Nseries (Nseries@itso.com)
- Domain Admins (ITSO\Domain Admins)
- Cert Publishers (ITSO\Cert Publishers)

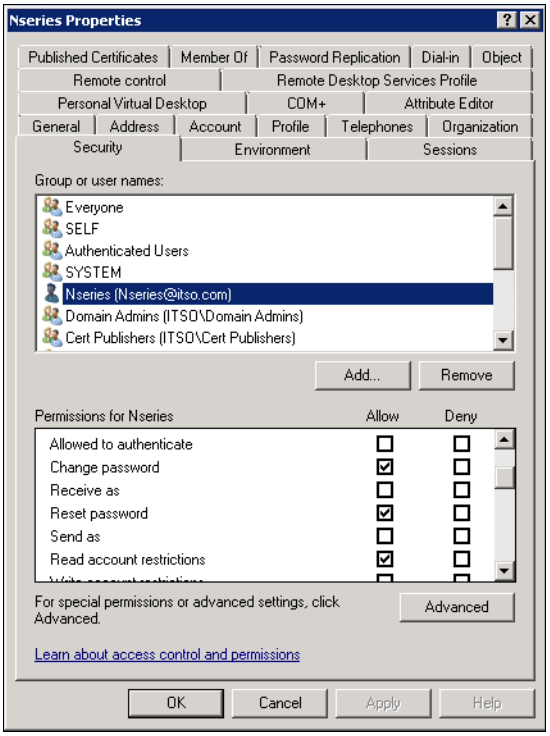
Permissions for Nseries	Allow	Deny
Full control	□	□
Read	☑	□
Write	□	□
Create all child objects	□	□
Delete all child objects	□	□
Allowed to authenticate	□	□

For special permissions or advanced settings, click Advanced.

Learn about access control and permissions

Figure 23-3 Selecting security permissions

3. In the Permissions pane, make sure that **Allow Change Password** and **Reset Password** are selected (Figure 23-4).



Group or user names:

- Everyone
- SELF
- Authenticated Users
- SYSTEM
- Nseries (Nseries@itso.com)
- Domain Admins (ITSO\Domain Admins)
- Cert Publishers (ITSO\Cert Publishers)

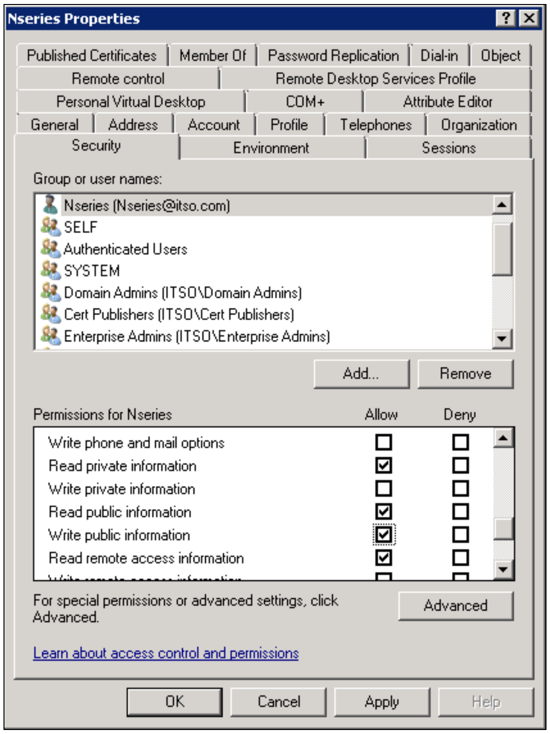
Permissions for Nseries	Allow	Deny
Allowed to authenticate	□	□
Change password	☑	□
Receive as	□	□
Reset password	☑	□
Send as	□	□
Read account restrictions	☑	□

For special permissions or advanced settings, click Advanced.

Learn about access control and permissions

Figure 23-4 Selecting password permissions

4. On the same Security tab in the Permissions scroll-down area, make sure that **Write Public Information** is checked (Figure 23-5).



Group or user names:

- Nseries (Nseries@itso.com)
- SELF
- Authenticated Users
- SYSTEM
- Domain Admins (ITSO\Domain Admins)
- Cert Publishers (ITSO\Cert Publishers)
- Enterprise Admins (ITSO\Enterprise Admins)

Permissions for Nseries	Allow	Deny
Write phone and mail options	□	□
Read private information	☑	□
Write private information	□	□
Read public information	☑	□
Write public information	☑	□
Read remote access information	☑	□

For special permissions or advanced settings, click Advanced.

Learn about access control and permissions

Figure 23-5 Selecting Write Public Information permission

23.5 Precreating a computer object

Many Active Directory administrators employ a set of preferred practices that place strict controls on who can create computer objects. If the join is performed in the manner described in "Creating the computer object" on page 354, then security risks are minimized because the need for Active Directory administrator rights at the device during the setup process is eliminated.

Precreate the computer object using an account with the required privileges, and then use an account with fewer privileges to log on to the computer and issue the appropriate commands to complete the join process. Precreating a computer object is the best method for joining an IBM System Storage N series to Active Directory.

23.5.1 Creating the computer object

> **Important:** Ensure that the IBM System Storage N series date and time are within 5 minutes of the domain controller date and time (that is, synchronized).

In order for the IBM System Storage N series to join Active Directory, you have to create a computer object reference by performing the following steps:

1. Open the Active Directory MMC and select the computer objects in your domain. Right-click **Computer** and create a **New** → **Computer** object (Figure 23-6).

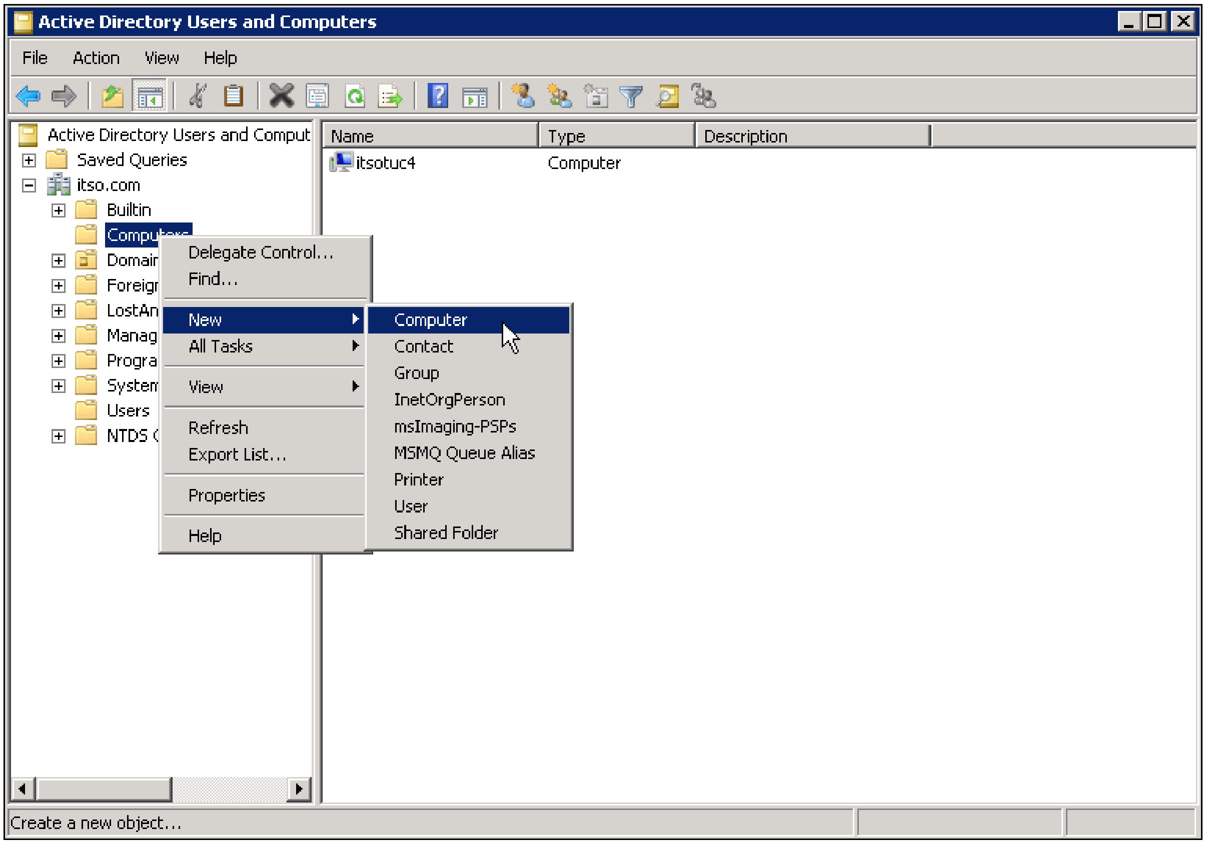
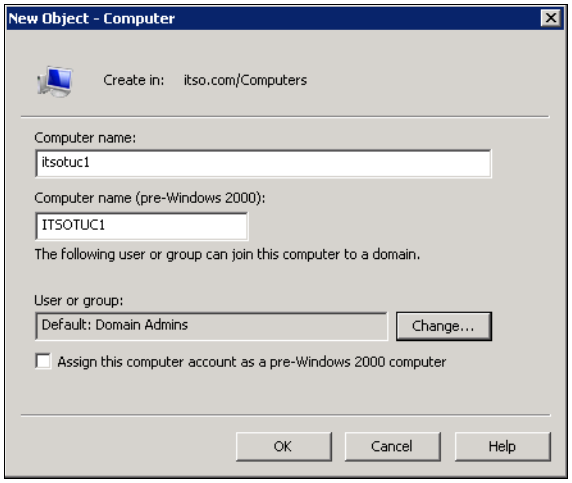


Figure 23-6 Creating a computer in Active Directory

2. Add a new computer object that references your IBM System Storage N series by using the pre-selected account, as shown in described in Figure 23-1 on page 350. Next, click **Change** to select a non-default user as shown in Figure 23-7.


New Object - Computer

Create in: itso.com/Computers

Computer name:
itsotuc1

Computer name (pre-Windows 2000):
ITSOTUC1

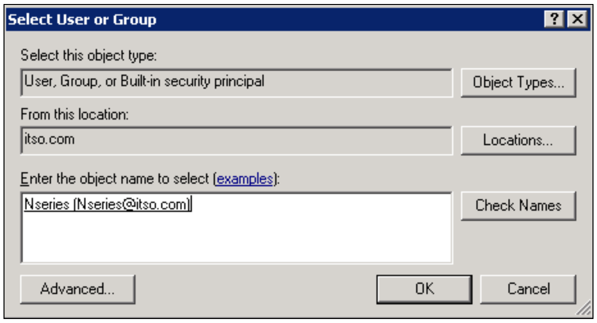
The following user or group can join this computer to a domain.

User or group:
Default: Domain Admins

□ Assign this computer account as a pre-Windows 2000 computer

Figure 23-7 Adding a new computer object

Figure 23-8 shows the selection of a user for the new computer object. We typed in our pre-defined N series account and clicked **Check Names.** Now click **OK**.


Select User or Group

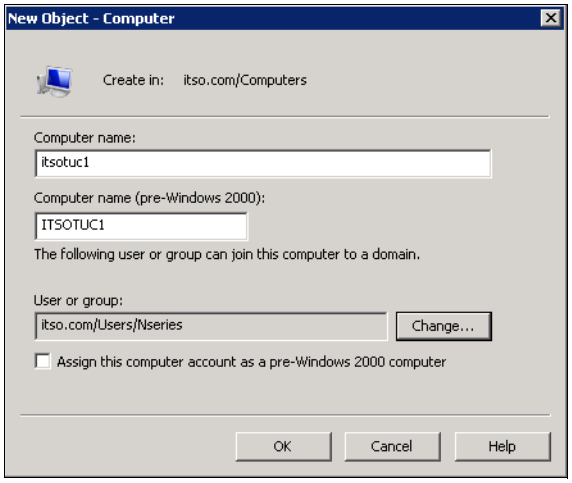
Select this object type:
User, Group, or Built-in security principal

From this location:
itso.com

Enter the object name to select (examples):
Nseries (Nseries@itso.com)

Figure 23-8 Selecting a user for an IBM System Storage N series computer object

Figure 23-9 shows the final step in the new computer object. Click **OK** to finish.



New Object - Computer

Create in: itso.com/Computers

Computer name:
itsotuc1

Computer name (pre-Windows 2000):
ITSOTUC1

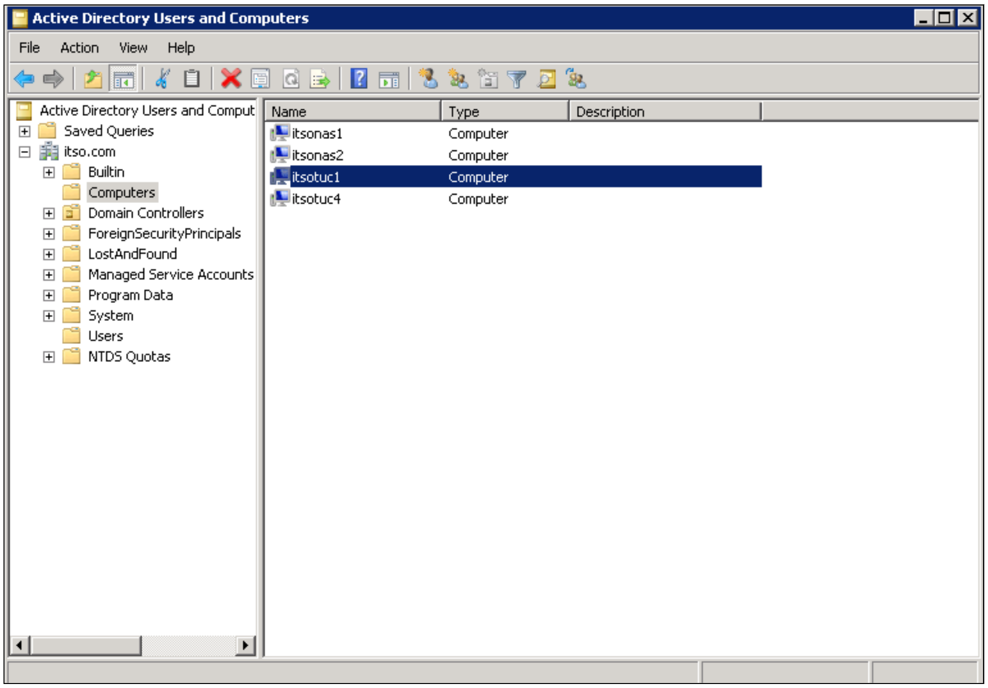
The following user or group can join this computer to a domain.

User or group:
itso.com/Users/Nseries

☐ Assign this computer account as a pre-Windows 2000 computer

Figure 23-9 Results of specifying a user

At this point, you can expect to see the computer object that was created for the IBM System Storage N series itsutuc1 in the computer object container of your Active Directory (Figure 23-10).



Name	Type	Description
itsonas1	Computer	
itsonas2	Computer	
itsotuc1	Computer	
itsotuc4	Computer	

Figure 23-10 Verification of computer object creation

23.5.2 Completion of the Active Directory integration

At the completion of the Active Directory join process, a number of properties are written to the computer account:

- DNS host name
- Several service principal names
- Object classes
- Operating system name and version
- A randomly generated password set for this account through KPASSWD.

> **Attention:** This is the only instance in which the IBM System Storage N series join process differs from the Microsoft join process. Microsoft uses proprietary RPC calls to change the password. In contrast, IBM uses the published KPASSWD APIs to change the password.

23.6 Running the CIFS setup wizard

In addition to performing initial CIFS configuration, the **cifs setup** command enables you to perform several tasks. With the **cifs setup** command, you can perform the following tasks:

- Create and name a CIFS server that your CIFS clients can access
- Join the CIFS server to a domain or workgroup, or move between them
- Create a default set of local CIFS users and groups

In this section, we describe how to run the CIFS setup wizard through a command-line interface (CLI) and a GUI.

23.6.1 Running the CIFS setup wizard using Data ONTAP CLI

Perform the following steps to join our N series itsotuc1 storage system to be an Active Directory member:

1. Run the **cifs setup** command to access the CIFS setup wizard (Example 23-4).
2. If CIFS is running at this point, then we have to type **cifs terminate** or the **cifs setup** command will fail.

   Example 23-4 lists the entire **cifs setup** dialogue and shows how we attach our NAS device to Active Directory.

Example 23-4 the entire CIFS setup dialogue

```
itsotuc1> cifs setup
This process will enable CIFS access to the filer from a Windows(R) system.
Use "?" for help at any prompt and Ctrl-C to exit without committing changes.

        This filer is currently a member of the /etc/passwd-style workgroup
        'WORKGROUP'.
Do you want to continue and change the current filer account information? [n]: y
        Your filer is currently visible to all systems using WINS. The WINS
        name server currently configured is: [ 9.11.218.102 ].

(1) Keep the current WINS configuration
(2) Change the current WINS name server address(es)
(3) Disable WINS
```

```
Selection (1-3)? [1]: 1
        This filer is currently configured as a multiprotocol filer.
Would you like to reconfigure this filer to be an NTFS-only filer? [n]: n
        The default name for this CIFS server is 'ITSOTUC1'.
Would you like to change this name? [n]: n
        Data ONTAP CIFS services support four styles of user authentication.
        Choose the one from the list below that best suits your situation.

(1) Active Directory domain authentication (Active Directory domains only)
(2) Windows NT 4 domain authentication (Windows NT or Active Directory domains)
(3) Windows Workgroup authentication using the filer's local user accounts
(4) /etc/passwd and/or NIS/LDAP authentication

Selection (1-4)? [1]: 1
What is the name of the Active Directory domain? [ITSO.COM]: itso.com
        In Active Directory-based domains, it is essential that the filer's
        time match the domain's internal time so that the Kerberos-based
        authentication system works correctly. If the time difference between
        the filer and the domain controllers is more than 5 minutes,
        authentication will fail. Time services are currently not configured
        on this filer.
Would you like to configure time services? [n]: n
        In order to create an Active Directory machine account for the filer,
        you must supply the name and password of a Windows account with
        sufficient privileges to add computers to the ITSO.COM domain.
Enter the name of the Windows user [Administrator@ITSO.COM]: nseries@itso.com
Password for nseries@itso.com:xxxxxxxx
CIFS - Logged in as nseries@itso.com.
        An account that matches the name 'ITSOTUC1' already exists in Active
        Directory: 'cn=itsotuc1,cn=computers,dc=itso,dc=com'. This is normal
        if you are re-running CIFS Setup. You can continue by using this
        account or changing the name of this CIFS server.
Do you want to re-use this machine account? [y]: y
CIFS - Starting SMB protocol...
Welcome to the ITSO.COM (ITSO) Active Directory(R) domain.

CIFS local server is running.
itsotuc1>
```

In the foregoing example, we chose to enable WINS, which were already enabled from the last time **`cifs setup`** was run. WINS is not required for CIFS, but NFS must be activated for CIFS to work.

Tip: **`cifs setup`** remembers the settings from last time it was run on the system, so re-running **`cifs setup`** might cause the prompts to vary slightly.

Example 23-5 shows how, by using the command **cifs domaininfo**, we can verify that our N series system is now a member of the Active Directory domain itso.com.

Example 23-5 Verification of connectivity to AD

```
itsotuc1> cifs domaininfo
NetBios Domain:           ITSO
Windows 2003 Domain Name: itso.com
Type:                     Windows 2003
Filer AD Site:            Default-First-Site-Name

Current Connected DCs:    \\WIN-OZI5BOZ5IEJ
Total DC addresses found: 2
Preferred Addresses:
                          None
Favored Addresses:
                          9.11.218.102    WIN-OZI5BOZ5IEJ  PDC
Other Addresses:
                          169.254.119.68                   PDC

Connected AD LDAP Server: \\win-ozi5boz5iej.itso.com
Preferred Addresses:
                          None
Favored Addresses:
                          9.11.218.102
                           win-ozi5boz5iej.itso.com
Other Addresses:
                          None
itsotuc1>
```

References: For additional information about how to activate CIFS, see the following documentation:

- IBM System Storage N series Data ONTAP 8.1 7-Mode Software Setup Guide
- IBM System Storage N series Data ONTAP 8.1 7-Mode File Access and Protocols Management Guide

Both guides can be found at this website:

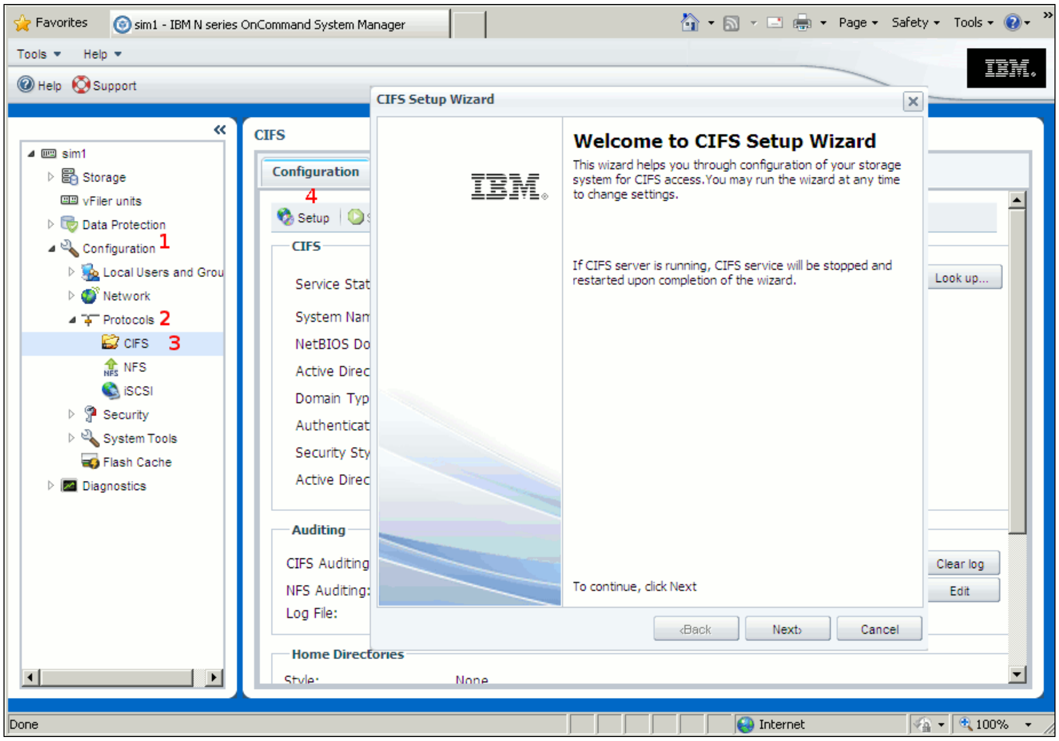
http://www.ibm.com/storage/support/nas

23.6.2 Running the CIFS setup wizard using System Manager

To use System Manager to setup the CIFS service on a Filer or vFiler, perform the following steps:

1. Start the System Manager tool her **Start Menu** → **Programs** → **IBM** → **N series OnCommand System Manager** → **IBM N series OnCommand System Manager 3.0**, then connect to your storage system.
2. Open the CIFS setup wizard by clicking on **Configuration** → **Protocols** → **CIFS** and then clicking on the **Setup** button, as shown in Figure 23-11.

   Click **Next.**



Welcome to CIFS Setup Wizard

This wizard helps you through configuration of your storage system for CIFS access.You may run the wizard at any time to change settings.

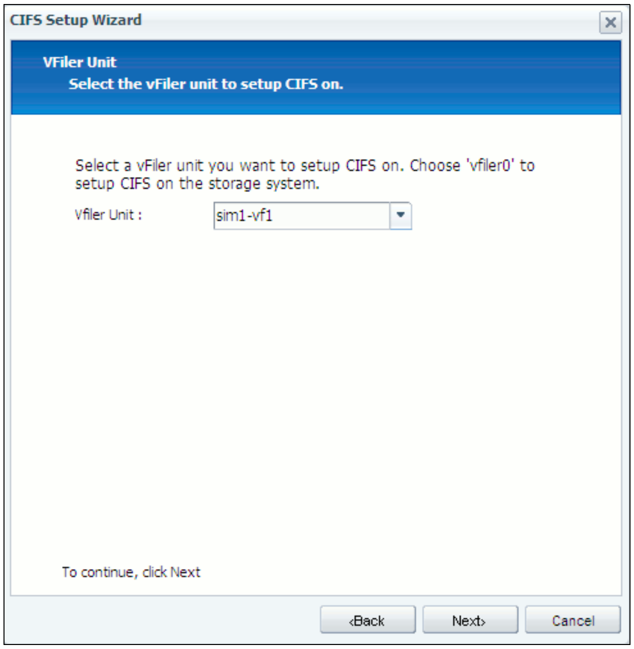
If CIFS server is running, CIFS service will be stopped and restarted upon completion of the wizard.

To continue, click Next

Figure 23-11 System Manager CIFS setup Wizard - start

3. Select the Vfiler unit to work with (Figure 23-12). Remember, each Vfiler has its own CIFS setup.

   Click **Next**.



VFiler Unit

Select the vFiler unit to setup CIFS on.

Select a vFiler unit you want to setup CIFS on. Choose 'vfiler0' to setup CIFS on the storage system.

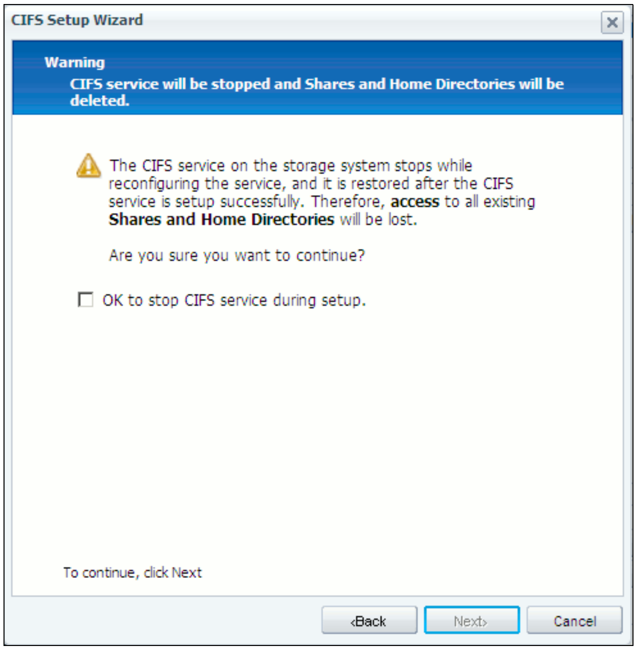
Vfiler Unit : sim1-vf1

To continue, click Next

Figure 23-12 FilerView CIFS setup Wizard - vFiler Unit

4. Select **OK** to stop the CIFS service during setup (Figure 23-13).

   Click **Next**.



CIFS Setup Wizard

Warning

CIFS service will be stopped and Shares and Home Directories will be deleted.

The CIFS service on the storage system stops while reconfiguring the service, and it is restored after the CIFS service is setup successfully. Therefore, **access** to all existing **Shares and Home Directories** will be lost.

Are you sure you want to continue?

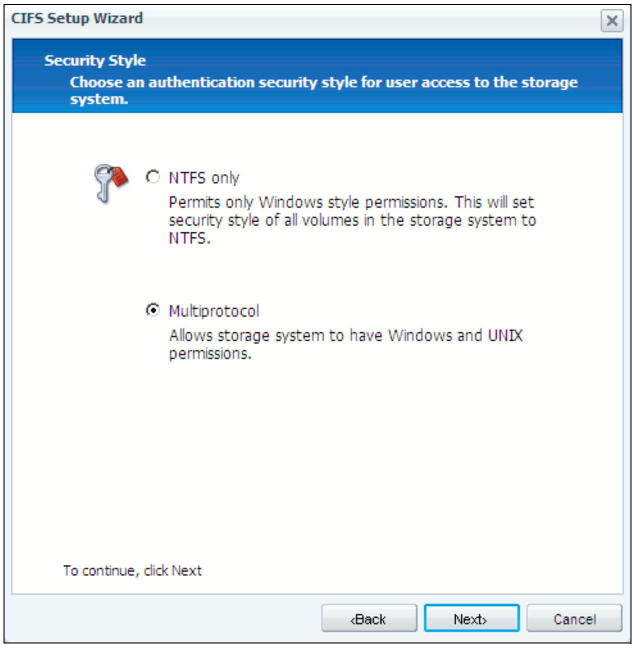
□ OK to stop CIFS service during setup.

To continue, click Next

Figure 23-13 FilerView CIFS Setup Wizard - stop CIFS during setup

5. Select the security style (Figure 23-14). The answer will depend on your environment, whether this Vfiler will only be serving Windows clients, UNIX clients, or a mixture of the two.

   Click **Next**.



CIFS Setup Wizard

Security Style

Choose an authentication security style for user access to the storage system.

○ NTFS only

Permits only Windows style permissions. This will set security style of all volumes in the storage system to NTFS.

● Multiprotocol

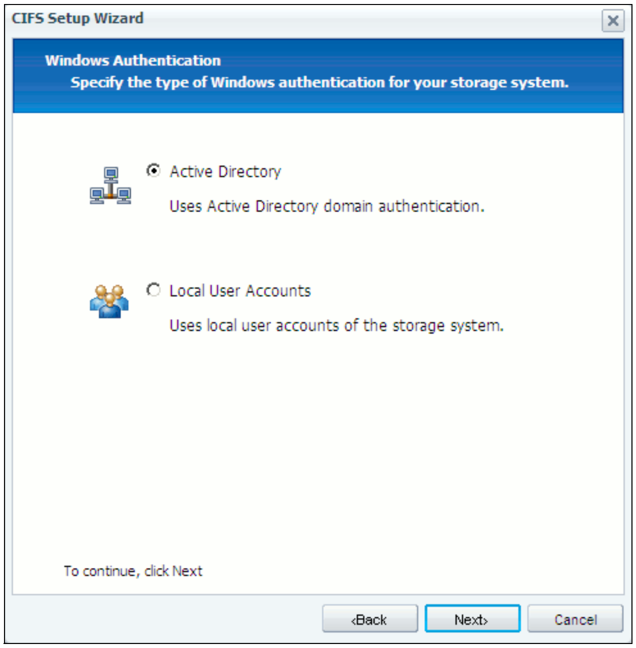
Allows storage system to have Windows and UNIX permissions.

To continue, click Next

Figure 23-14 FilerView CIFS setup Wizard - security style

6. Select the authentication method (Figure 23-15). In most environments, this will be Active Directory.

   Click **Next**.



CIFS Setup Wizard

Windows Authentication

Specify the type of Windows authentication for your storage system.

● Active Directory

Uses Active Directory domain authentication.

○ Local User Accounts

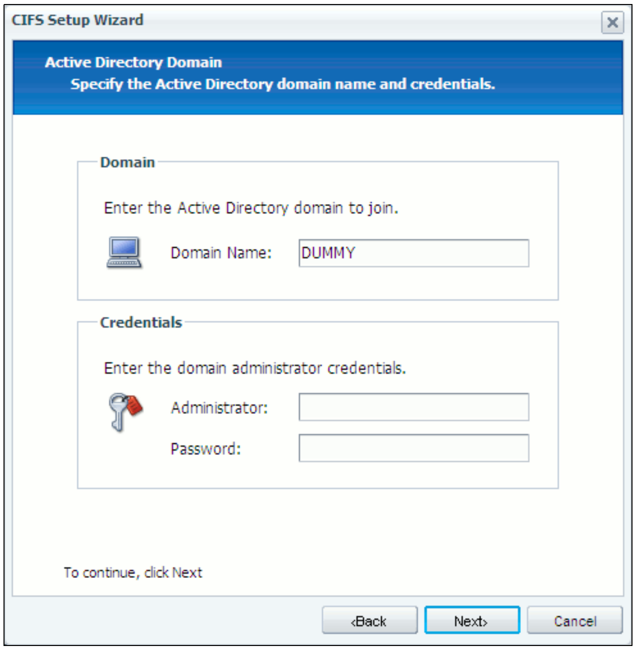
Uses local user accounts of the storage system.

To continue, click Next

Figure 23-15 FilerView CIFS setup Wizard - authentication method

7. If you selected Active Directory, then you will need to enter an AD domain name, and credentials to join that domain (Figure 23-16).

   Click **Next**.



CIFS Setup Wizard

Active Directory Domain

Specify the Active Directory domain name and credentials.

Domain

Enter the Active Directory domain to join.

Domain Name: DUMMY

Credentials

Enter the domain administrator credentials.

Administrator:

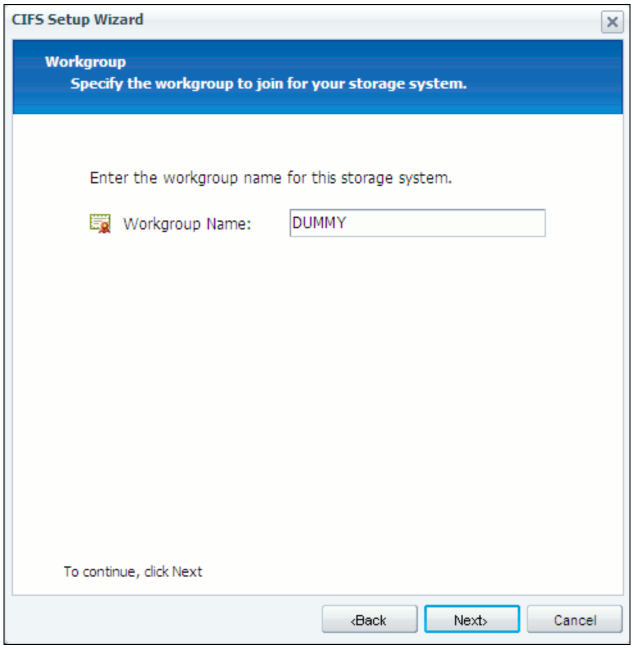
Password:

To continue, click Next

Figure 23-16 FilerView CIFS setup Wizard - Active Directory

8. If you selected Local User Accounts, then you will need to enter the Workgroup name to associate with (Figure 23-17).

   Click **Next**.



CIFS Setup Wizard

Workgroup

Specify the workgroup to join for your storage system.

Enter the workgroup name for this storage system.

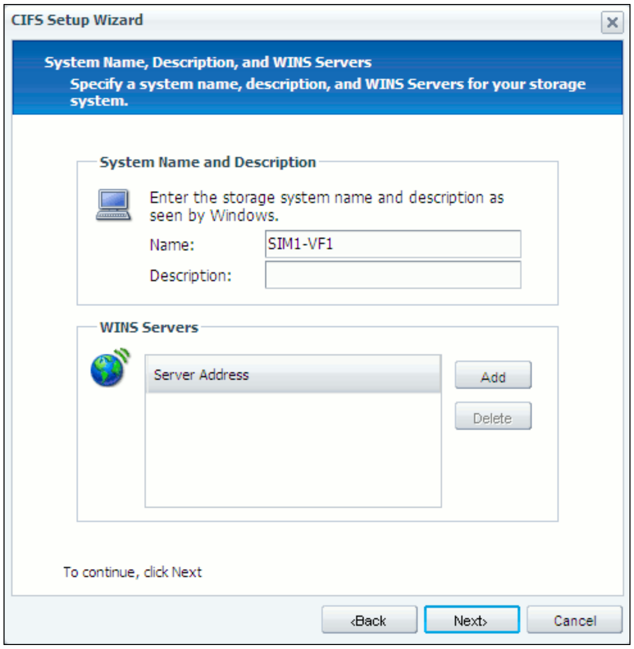
Workgroup Name: DUMMY

To continue, click Next

Figure 23-17 FilerView CIFS setup Wizard - Local User Accounts

9. Enter a system name and description, and WINS servers if required (Figure 23-18).

   Click **Next**.



CIFS Setup Wizard

System Name, Description, and WINS Servers

Specify a system name, description, and WINS Servers for your storage system.

System Name and Description

Enter the storage system name and description as seen by Windows.

Name: SIM1-VF1

Description:

WINS Servers

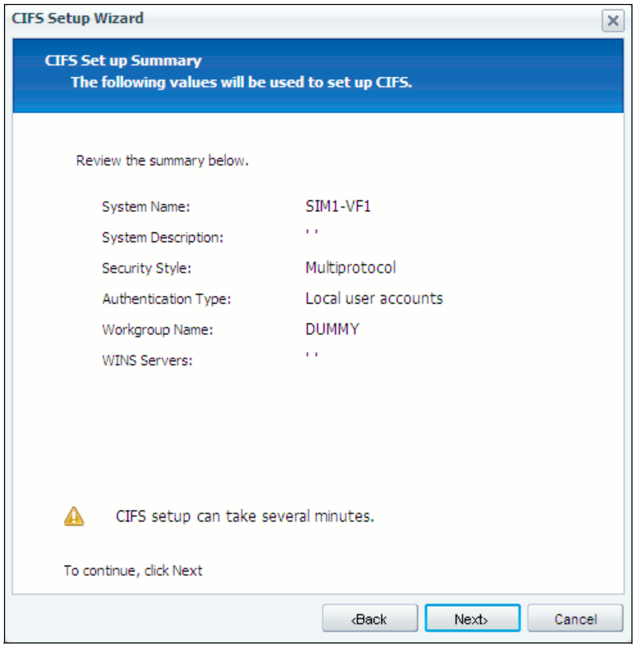
Server Address

To continue, click Next

Figure 23-18 FilerView CIFS setup Wizard - system name

10. Review the CIFS setup summary (Figure 23-19).

Click **Next**.



CIFS Setup Wizard

CIFS Set up Summary

The following values will be used to set up CIFS.

Review the summary below.

System Name:	SIM1-VF1
System Description:	' '
Security Style:	Multiprotocol
Authentication Type:	Local user accounts
Workgroup Name:	DUMMY
WINS Servers:	' '

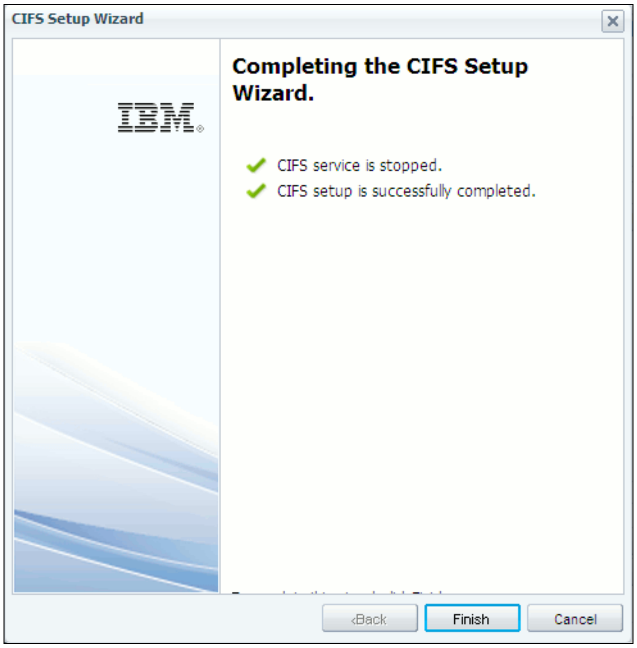
CIFS setup can take several minutes.

To continue, click Next

Figure 23-19 FilerView CIFS setup Wizard - summary

11. The Wizard will now set up the CIFS service (Figure 23-20).

Click **Finish**.



CIFS Setup Wizard

Completing the CIFS Setup Wizard.

✓ CIFS service is stopped.

✓ CIFS setup is successfully completed.

Figure 23-20 FilerView CIFS setup Wizard - complete

23.6.3 Active Directory: Mixed mode or native mode

The terms *mixed mode* and *native mode* refer to functional levels in a Windows 2000 server. In a Windows 2003 server, the terms mixed and native have been superseded by the *raise function level*.

23.6.4 Domain function levels (mixed and native)

There are now four domain levels in which a Windows 2003 server can operate:

- Windows 2003 server:

  All the servers are Windows 2003 servers, and there are no other domain controllers. However, even at this level, the entire range of clients (including IBM System Storage N series storage systems) and member servers can still join the domain.

- Windows 2003 server interim:

  There are Windows NT 4.0 servers and Window 2003 servers (but no Windows 2000 servers). This level applies when you upgrade a Windows NT 4.0 PDC to a Windows 2003 server.

  Interim mode is important when you have Windows NT 4.0 groups with more than 5,000 members. Windows 2000 does not allow you to create groups with more than 5,000 members.

- Windows 2000 native:

  There are Windows 2000 servers and Windows 2003 servers (but no Windows NT 4.0 servers).

- Windows 2000 mixed:

  There are Windows NT 4.0 BDCs and Windows 2000 servers. Windows 2000 mixed is the default function level, because it supports all types of domain controllers.

Joining an IBM Storage System N series to AD: An IBM System Storage N series storage system can be joined to the Active Directory in mixed, native, interim, or pure Windows 2003 server modes.

23.7 Troubleshooting the domain joining process

This section describes the various methods of troubleshooting the domain joining process.

23.7.1 DNS

To determine whether the IBM device is joining a Windows NT 4.0 domain or Active Directory, and to locate the domain controllers, a key distribution center (KDC) (used for Kerberos), and other services are necessary, as CIFS relies on DNS.

If DNS is not enabled or is configured incorrectly, the domain joining phase either fails or, if a Microsoft Windows Internet-Naming Server (WINS) is running, assumes that the domain being joined is a Windows NT 4.0 domain.

23.7.2 Time synchronization

If time synchronization is not enabled, and the IBM System Storage N series storage system’s time diverges more than five minutes from the domain's time, client authentication attempts to the IBM System Storage N series storage system will fail.

23.7.3 Active Directory replication

Based on the size of the Active Directory domain, when propagating a change for a small organization with one site, replication usually takes less than 15 minutes. For a global company with many sites, replication can take up to several hours to complete.

23.8 Device discovery

The IBM System Storage N series storage system performs an intelligent discovery process to locate the most appropriate domain controller (DC) in the network with which to communicate. For its first connection, Data ONTAP attempts to use servers that appear in the CIFS **prefdc** list (in list order), if configured.

Example 23-6 shows a **cifs prefdc print** command with **prefdc** configured.

Example 23-6 The cifs prefdc print command with prefdc configured

```
itsotuc1> cifs prefdc print
Preferred DC ordering per domain:

ITSO:
        1. 9.11.218.102
itsotuc1>
```

If none of these preferred servers are available, or if none are configured, all server addresses are discovered at once, and then categorized, prioritized, and cached.

Example 23-7 shows a **prefdc** list with nothing configured.

Example 23-7 The prefdc list with nothing configured

```
itsotuc1> cifs prefdc print
No preferred Domain Controllers configured.
DCs will be automatically discovered.
itsotuc1>
```

Preferred addresses are ordered as specified by using the **cifs prefdc** command. *Favored* categories and *other* categories are sorted according to the fastest response. Data ONTAP simultaneously pings all addresses listed in both categories and waits one second for responses.

The **cifs prefdc** command gives you control over the order in which Data ONTAP attempts to contact a server. The list is consulted for all Windows service connections, not just domain controllers.

When configuring CIFS on an IBM System Storage N series device in a Windows 2000 or Windows 2003 or 2008 domain, an LDAP query to Active Directory checks to ensure that a computer object with the same name does not already exist. If the name does exist, the setup process makes sure that it is not a domain controller. These are precautionary measures used to guarantee that no computer object names are duplicated in error.

In order to locate resources on a network, a mechanism must exist by which the resources can easily be found. A directory service such as Active Directory keeps track of all known resources and responds to requests with a list of currently available devices and services.

But before you can be trusted to query for resources, you must be granted membership in the domain. Joining a domain accomplishes two tasks:

- For an IBM System Storage N series storage system, it grants the required rights to query Active Directory, if it needs to find other resources.
- It provides a single management interface through MMC for administration of security and user access levels to the IBM System Storage N series storage system.

23.9 Automatic home shares

The IBM N-Series storage controller has the ability to dynamically create a CIFS share on demand when a user access their home directory. When the user disconnects, the share is automatically removed, thus reducing the load on the NAS controller, because it now needs only to maintain those CIFS home shares that are actively in use at any particular time. It also reduces the administrative burden of needing to define individual user home shares, which might be many thousands of users in a large environment.

23.9.1 Visibility of shares

Another interesting feature of automatic home shares is that they are only visible to the individual account owner, even the NAS administrator cannot list this type of share. It is done by design, because listing all of the CIFS shares on a large NAS with tens of thousands of users can soon become unwieldy. Visibility is a form of *access based share enumeration*.

23.9.2 Configuring the NAS controller

This section describes how to configure the auto home share feature on the IBM N-Series storage controller. (Each major step is divided into procedural steps.)

Step 1: Create a parent directory for the home shares

Follow these steps to configure the parent directory and CIFS share:

1. Log in to the NAS administrative interface (CLI examples are shown here).
2. Create a new volume to contain the user home shares, as shown in Example 23-8:
   a. Set the size and other parameters to suit your requirements.
   b. The CIFS home shares will be created as sub-directories under this volume.

Example 23-8 Create a NAS volume to contain the home shares

```
NAS> vol create cifs_home -s none aggr0 10g
Creation of volume 'cifs_home' with size 10g on containing aggregate
'aggr0' has completed.
```

3. Configure the NAS file system security mode for CIFS access, as shown in Example 23-9.

   The security might already be correct, depending on the NAS default security style.

Example 23-9 Configure the NAS security style for the volume

```
NAS> qtree status cifs_home
Volume   Tree     Style Oplocks  Status
-------- -------- ----- -------- ---------
cifs_home          unix  enabled  normal

NAS> qtree security /vol/cifs_home ntfs
Thu Oct 13 15:09:53 GMT [N5600-B: wafl.quota.sec.change:notice]: security style
for /vol/cifs_home/ changed from unix to ntfs
```

4. Create a CIFS share for the containing volume, as shown in Example 23-10:
   a. Remove access permissions for “everyone.”
   b. Create access permissions for “administrator.”

Example 23-10 Create a CIFS share and set the access permissions

```
NAS> cifs shares -add cifs_home /vol/cifs_home
The share name 'cifs_home' will not be accessible by some MS-DOS workstations

NAS> cifs shares cifs_home
Name         Mount Point                       Description
----         -----------                       -----------
cifs_home    /vol/cifs_home
         everyone / Full Control

NAS> cifs access -delete cifs_home everyone
1 share(s) have been successfully modified

NAS> cifs access cifs_home administrator "Full Control"
1 share(s) have been successfully modified

NAS> cifs shares cifs_home
Name         Mount Point                       Description
----         -----------                       -----------
cifs_home    /vol/cifs_home
         N5600-B\administrator / Full Control
```

The CIFS share is now ready for administrative access.

Tip: This CIFS share will not be accessed directly by the users. Instead it will be used by the RTCA to access the contents of the auto home shares that are in sub-directories of the containing volume.

Step 2: Set the parent directory location

Follow these steps to configure the parent directory location:

1. Check whether a CIFS home directory has already been configured, as shown in Example 23-11.

   If this parameter is already configured, you need to investigate before continuing.

Example 23-11 Check the CIFS home directory location

```
NAS> cifs homedir
No CIFS home directory paths.

NAS> rdfile /etc/cifs_homedir.cfg
#
# This file contains the path(s) used by the filer to determine if a
# CIFS user has a home directory. See the System Administrator's Guide
# for a full description of this file and a full description of the
# CIFS homedir feature.
#
# There is a limit to the number of paths that may be specified.
# Currently that limit is 1000.
# Paths must be entered one per line.
#
# After editing this file, use the console command "cifs homedir load"
# to make the filer process the entries in this file.
#
# Note that the "#" character is valid in a CIFS directory name.
# Therefore the "#" character is only treated as a comment in this
# file if it is in the first column.
#
# Two example path entries are given below.
# /vol/vol0/users1
# /vol/vol1/users2
#
```

2. Configure the CIFS home directory parameter, as shown in Example 23-12.

   Be careful with the **wrfile** command, because it is easy to make a mistake and overwrite the configuration file with blank data.

Example 23-12 Set the CIFS home directory location

```
NAS> wrfile -a /etc/cifs_homedir.cfg /vol/cifs_home

NAs> cifs homedir load

NAS> cifs homedir
/vol/cifs_home
```

The CIFS service now knows where to look for the users home directories when automatically creating the user’s home shares.

Step 3: Set the auto home directory name style

Follow these steps to configure the naming standard for the auto home shares:

1. Configure the naming style for the auto home share, as shown in Example 23-13.

   In our lab, we chose the default naming style. You must set it to match the requirements of your environment.

Example 23-13 Set the CIFS home directory name style

```
NAS> options cifs.home_dir_namestyle ntname
```

See the NAS product manual for a description of the other directory naming styles.

Step 4: Create the user’s home directories

Follow these steps to create the user’s home directory on the NAS controller:

1. Log in to the NAS client (for example, a Windows PC).
2. Create the home directory on the NAS controller, as shown in Example 23-14:
   a. Detach any existing CIFS connections.
   b. Authenticate to the NAS controller as the “administrator” user account.
   c. Create the CIFS user’s home directory (under the “homedir” location that we configured in “Step 2: Set the parent directory location” on page 368).

Example 23-14 Make a home directory for the CIFS user/s

```
C:\>net use * /d
You have these remote connections:

                   \\nas2\IPC$
Continuing will cancel the connections.

Do you want to continue this operation? (Y/N) [N]: y
The command completed successfully.

C:\>net use \\nas2 /user:administrator
The password or user name is invalid for \\nas2.

Enter the password for 'administrator' to connect to 'nas2':********
The command completed successfully.

C:\>mkdir \\nas2\cifs_home\cifsuser1
```

At this point, the CIFS auto home share is ready to access from the NAS client.

Step 5: Access the auto home share

Follow these steps to connect the user to their new home directory on the NAS controller:

1. Log in to the NAS client (for example, a Windows PC).
2. Connect to the new auto home share, as shown in Example 23-15.
   a. Detach any existing CIFS connections.
   b. Make a connection to the new CIFS share:

      Authenticate to the NAS controller as the “cifsuser1” user account.
      (Of course, you need to change it to suit your user account name.)

Example 23-15 Connect to the new CIFS auto home share

```
C:\>net use * /d
You have these remote connections:

                   \\nas2\IPC$
Continuing will cancel the connections.

Do you want to continue this operation? (Y/N) [N]: y
The command completed successfully.

C:\>net use h: \\nas2\cifsuser1 /user:cifsuser1
The password or user name is invalid for \\nas2\cifsuser1.

Enter the password for 'cifsuser1' to connect to 'nas2':
The command completed successfully.
```

Assuming that you have configured the share permissions and security style correctly, the user must now be able to read and write to their new CIFS home share.

When the user disconnects their CIFS session, the share will automatically be removed. Of course, the underlying directory and its contents remain on the NAS controller, ready to be accessed on demand the next time the user connects to the system.

24

NFS

Most UNIX clients use NFS for remote file access. Sun Microsystems introduced NFS in1985. Since then, it has become a de facto standard protocol, used by 10 million systems worldwide.

In this chapter, we describe NFS. It is particularly common on UNIX-based systems, but NFS implementations are available for virtually every modern computing platform in current use, from desktops to supercomputers. Only when used by UNIX-based systems, however, does NFS closely resemble the behavior of a client's local file system.

The following topics are covered:

- What NFS is
- NFS versions
- File access using NFS
- NFS shares
- NFS Data ONTAP 8.1 commands
- Enabling Kerberos v5 security services for NFS
- Interoperability

24.1 What NFS is

The filer supports NFS versions V2, V3, and V4 of the NFS protocol as shown in Figure 24-1.

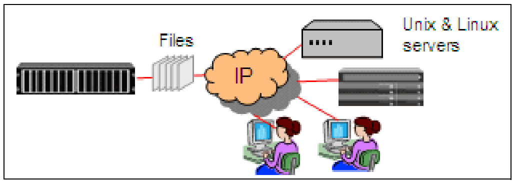


Figure 24-1 NFS

NFS is a widely used file sharing protocol supported on a broad range of platforms. The protocol is designed to be stateless, allowing easy recovery in the event of server failure. Associated with the NFS protocol are two ancillary protocols, the MOUNT protocol and the NLM protocol. The MOUNT protocol provides a means of translating an initial path name on a server to an NFS file-handle which provides the initial reference for subsequent NFS protocol operations. The NLM protocol provides file locking services, which are stateful by nature, outside of the stateless NFS protocol.

NFS is supported on both TCP and UDP transports. Support for TCP and UDP is enabled by default. Either one can be disabled by setting the nfs.tcp.enable or nfs.udp.enable options using the options command.

24.2 NFS versions

This section describes the supported NFS versions.

24.2.1 NFS V1

NFS was first introduced by Sun Microsystems in the early 1980s. NFS V1 was Sun’s prototype version and was never released for public use.

24.2.2 NFS V2

NFS V2 was released in 1985 with the SunOS V2 operating system. Many UNIX vendors licensed this version of NFS from Sun. NFS V2 suffered many undocumented and subtle changes throughout its 10-year life. Some vendors allowed NFS V2 to read or write more than 4 K bytes at a time, while others increased the number of groups provided as part of the RPC authentication from 8 to 16. These minor changes created occasional incompatibilities between different NFS implementations. However, the protocol continued to provide an exceptional degree of compatibility between systems made by different vendors NFS V2 is a default and cannot be disabled.

24.2.3 NFS V3

The NFS V3 specification was developed during July 1992. Working code for NFS V3 was introduced by some vendors in 1995 and was made widely available in 1996. Version 3 incorporated many performance improvements over Version 2. But did not significantly change the way that NFS worked or the security mode used by the network file system. It is backwards compatible with Version 2 and it supports 64-bit file size. It has asynchronous writes, which eliminates the synchronous write bottleneck of Version 2.

Since the initial NFS protocol specification defined file sizes as being 32 bits long, supporting 64-bit file sizes required the NFS protocol revision to be updated.

Protocol revisions are rare, so it is not sensible to make just one change. As a result, NFSv3 includes several other changes along with the large file size support. The most interesting ones are a collection of performance improvements.

Large block transfers

The NFSv2 protocol specification restricts read and write operations to 8 KB (kilobytes). In NFSv3, the client and server can negotiate any size they like for reads and writes. Current NFSv3 implementations are indicate a consensus for using 32-KB transfer sizes for 10- and 100-Mbps (Megabit per second) networks, and 48-KB in HiPPI environments which run at 100 MBps (MegaByte per second) or higher.

Allowing the client and server to negotiate the optimal transfer size provides flexibility that will allow NFSv3 implementations to evolve in the future, if necessary, in case new networking technology makes even larger block sizes desirable.

Safe asynchronous writes

This feature allows the server to reply to writes immediately, instead of waiting for the data to be put safely on disk or in NVRAM. A new operation, called Commit, lets clients check with the server at some point after the WRITE operation, to verify that the server actually has written the data. The client is required to keep its own copy of the written data until the Commit succeeds, and if the Commit fails, the client is required to resend its copy of the written data.

For systems without NVRAM, this feature improves write performance for large files. On servers that do use NVRAM, it can reduce the CPU time spent copying data into NVRAM, thereby increasing the total throughput capability of the server.

NFSv3 support for asynchronous writing does not enhance by much the speed with which small files can be written. (Writing a small file might only require one or two async Write requests followed by a Commit.) And it does not help operations such as Create, Remove, and Rename at all. Therefore, NVRAM will continue to be critical to fast NFS service, even with NFSv3.

Improved attribute returns

In NFSv2, some operations return less information than they ought to. For instance, the Symlink operation creates a new link, but it does not return the file handle or attributes of the link. As a result, an NFSv2 client must send a Lookup request immediately after the Symlink.

In NFSv3, operations return additional information as appropriate, thus reducing the total number of operations that need to be sent.

The readdirplus operation

In NFSv2, the Readdir operation returns the names of the files in a directory, but not the attributes. So to handle a command like `ls -l`, the Readdir must be followed by a Lookup operation for each file in the directory. An `ls -l` on a directory with 100 entries would require 101 NFS operations.

NFSv3 supports a Readdirplus operation that returns both directory names and file attributes. As a result, `ls -l` could be handled with just one Readdirplus operation. It is especially useful in speeding up recursive tree-walking commands such as **`find`** and **`ls -R`**.

24.2.4 NFS V4

In 1998, Sun initiated an effort to design NFSv4 (RFC 3530). This design resulted in significant changes for NFS. NFSv4 incorporates the following new features:

- File system name-space
- Access control lists (ACLs)
- Improved client caching efficiency
- Stronger security
- Stateful design
- Improved ease of use in respect to the Internet

In that same year, Sun relinquished control over NFS development to the Internet Engineering Task Force (IETF), which then assumed responsibility for further development of the standard

Data ONTAP support of NFSv4

Supporting NFSv4 clients involves enabling or disabling the NFSv4 protocol, specifying an NFSv4 user ID domain, managing NFSv4 ACLS and file delegation, and configuring file and record locking. Data ONTAP supports all of the mandatory functionality in NFSv4 except the SPKM3 and LIPKEY security mechanisms.

This functionality consists of the following features.

COMPOUND

Allows a client to request multiple file operations in a single Remote Procedure Call (RPC) request.

Open delegation

Allows the server to delegate file control to some types of clients for read and write access.

Pseudo-fs

Used by NFSv4 servers to determine mount points on the storage system. There is no mount protocol in NFSv4.

Locking

Lease-based. There are no separate Network Lock Manager (NLM) or Network Status Monitor (NSM) protocols in NFSv4.

Named attributes

Similar to Windows NT streams.

24.3 File access using NFS

To be able to set up and use NFS functionality, your storage system must have an NFS license installed. You can use the license command to configure licenses on the storage system.

You need to be sure that you have obtained a valid NFS license. You can display the features that are currently licensed on the storage system by entering the following command:

```
license
```

If the NFS feature displays a licence code, then your storage system is already licensed. Otherwise, install an NFS license by entering the following command:

```
license add license_code.
```

24.3.1 Exporting or unexporting file system paths

You can export or unexport a file system path, making it available or unavailable to NFS clients, by editing the **/etc/exports** file or running the **exportfs** command.

To support secure NFS access (through using the **sec=krb*** export option), you must first enable Kerberos v5 security services.

If you need to make permanent changes to several export entries at once, it is usually easiest to edit the **/etc/exports** file directly. However, if you need to make changes to a single export entry or you need to make temporary changes, it is usually easiest to run the **exportfs** command.

24.3.2 Editing the /etc/exports file

To specify which file system paths Data ONTAP exports automatically when NFS starts, you can edit the **/etc/exports** file.

If the **nfs.export.auto-update** option is on, which it is by default, Data ONTAP automatically updates the **/etc/exports** file when you create, rename, or delete volumes.

> **Tip:** The maximum number of lines in the **/etc/exports** file is 10,240. This includes commented lines. The maximum number of characters in each export entry, including the end of line character, is 4,096.

An export entry has the following syntax:

```
path -option[, option...]
```

In the export entry syntax, **path** is a file system path (for example, a path to a volume, directory, or file) and **option** is an export option that specifies the following information:

- Which NFS clients have which access privileges (read-only, read-write, or root)
- The user ID (or name) of all anonymous or root NFS client users that access the file system path
- Whether NFS client users can create **setuid** and **setgid** executables and use the **mknod** command when accessing the file system path
- The security types that an NFS client must support to access the file system path
- The actual file system path corresponding to the exported file system path

Use these steps to edit the /etc/exports file:

1. Open the /etc/exports file in a text editor on an NFS client that has root access to the storage system.
2. Make your changes.
3. Save the file.

If you edit the `/etc/exports` file using a text editor, your changes will not take effect until you export all file system paths in the `/etc/exports` file or synchronize the currently exported file system paths with those specified in the /etc/exports file.

> **Tip:** Running the `exportfs` command with the `-b, -p, or -z` option also changes the `/etc/exports` file.

24.4 NFS shares

File shares, as shown in Figure 24-2, are exports to the user or application.

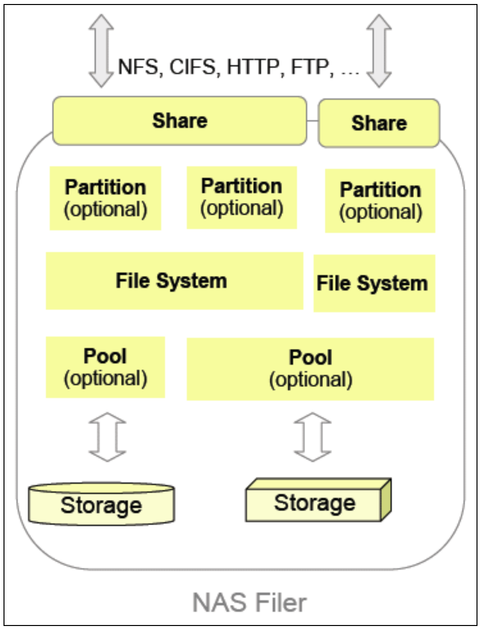
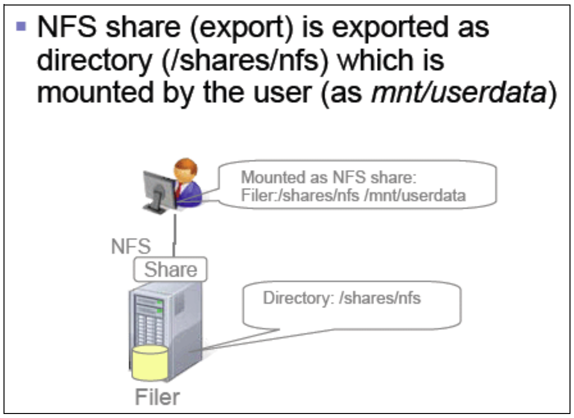


Figure 24-2 NFS shares

Here are some common characteristics for NFS exports:

- NFS share (export) is exported as a directory (/shares/nfs) which is mounted by the user (as mnt/userdata). See Figure 24-3.
- NFS v3 is stateless.
- Upon TCPIP address failover, the NFS client experiences a short interruption.
- NFS v4 introduces stateful protocol.



- NFS share (export) is exported as directory (/shares/nfs) which is mounted by the user (as *mnt/userdata*)

Figure 24-3 NFS share export

24.5 NFS Data ONTAP 8.1 commands

The NFS command reference includes only admin level commands. Advanced commands are not included.

The **nfs** command is used to manage the Network File System service. It can be used to turn the NFS service on or off, do Kerberos setup, manage the NSDB cache, or check nfs status. With no arguments, **nfs** shows the current state of the NFS service. This behavior is now deprecated and might be removed in a future release. The NFS command has the following functions:

nfs

na_nfs manages Network File System service.

nfs [help | on | off | setup | stat | status | nsdb]

nfs help [subcommand] displays the available subcommands with no further arguments. Otherwise, it displays a short help message for the subcommand.

nfs off turns off the NFS subsystem.

nfs on turns on the NFS subsystem if it is licensed.

nfs setup enters a setup dialog that is used to set system parameters needed for Kerberos V5 support in NFS.

nfs stat [options] is an alias for **nfsstat**; consult the man page for **nfsstat** for usage.

nfs status displays the current state of the NFS service.

nfs nsdb is used to manage the name server database cache (NSDB).

24.6 Enabling Kerberos v5 security services for NFS

To enable Kerberos v5 security services for NFS, you can use the **`nfs setup`** command.

Data ONTAP provides secure NFS access using the Kerberos v5 authentication protocol to ensure the security of data and the identity of users within a controlled domain.

The Data ONTAP Kerberos v5 implementation for NFS supports two Kerberos Key Distribution Center (KDC) types: Active Directory-based and UNIX-based, as described in Table 24-1.

Table 24-1 Key distribution

KDC type	**Description**
Actice Directory based	The Kerberos realm for NFS is an Active Directory-based KDC. You must configure CIFS with Microsoft Active Directory authentication (which is Kerberos-based); then NFS will use the CIFS domain controller as the KDC.
UNIX based	The Kerberos realm for NFS is an MIT or Heimdal KDC.
Multirealm	Uses a UNIX-based KDC for NFS and an Active Directory-based KDC for CIFS. Available in Data ONTAP 7.3.1 and later releases.

To support Kerberos multirealm configurations, Data ONTAP uses two sets of principal and keytab files. For Active Directory-based KDCs, the principal and keytab files are **`/etc/krb5auto.conf and /etc/krb5.keytab,`** respectively, just as in releases prior to Data ONTAP 7.3.1. For UNIX-based KDCs, however, the principal and keytab files are **`/etc/krb5.conf and /etc/UNIX_krb5.keytab`**, respectively.

Starting with Data ONTAP 7.3.1, the keytab file for UNIX-based KDCs has changed from **`/etc/krb5.keytab to /etc/UNIX_krb5.keytab`**. Data ONTAP continues to use the old keytab file /etc/krb5.keytab, however, if you upgrade from a release prior to Data ONTAP 7.3.1 in which Data ONTAP was configured to use a UNIX based KDC for NFS. You need only use the new keytab file **`/etc/UNIX_krb5.keytab`** for UNIX-based KDCs if you are reconfiguring CIFS after upgrading from such a release or if you are configuring NFS for the first time after configuring an Active-Directory-based KDC for CIFS.

24.7 Interoperability

Usage of NFS V2, V3, and V4 over UDP or TCP is supported on any platform that supports the industry standard NFS protocol.

The latest version of the NFS interoperability matrix can be found at the following website:

http://www-304.ibm.com/support/docview.wss?uid=ssg1S7003770

25

Multiprotocol data access

This chapter describes multiprotocol data access for the IBM System Storage N series storage systems, which provide fast, simple, and reliable network data access to the following clients: Network File System (NFS), Common Internet File System (CIFS) for Microsoft Windows networking), and Hypertext Transfer Protocol (HTTP) primarily for web browsers.

Support for all three protocols is woven into the Data ONTAP microkernel and file system. It provides multiprotocol data access that transcends the enclosed perspective of general purpose operating systems, as explained in this chapter.

The following topics are covered:

- Introduction to multiprotocol access
- File system permissions
- File service for environments with NFS and CIFS
- Altering qtree security at the qtree level or volume level
- NFS
- CIFS
- NFS compared to CIFS
- Mixing NFS and CIFS
- Multiprotocol file service

25.1 Introduction to multiprotocol access

As mentioned, support for the NFS, CIFS, and HTTP protocols has been designed into the Data ONTAP microkernel and file system, and it provides you with multiprotocol data access that transcends the enclosed perspective of general-purpose operating systems.

In the context of file service to Windows clients, Data ONTAP software for Windows is virtually indistinguishable from other Microsoft Windows servers in a Windows domain. For example, in addition to many other Windows-compatible features, note the following points:

- Access control lists (ACLs) can be set on shares, files, and directories.
- IBM System Storage N series storage systems can be administered through Windows Server Manager and User Manager.
- UNIX users are mapped to Windows users.
- Multi-language support is available through UNICODE.
- File access logging can be tracked for Windows and UNIX users.
- IBM System Storage N series storage systems interoperate with NTFS and Active Directory.

Windows-style ACLs and UNIX-style file access permissions are fully integrated on the IBM System Storage N series storage system. Furthermore, Windows users are automatically mapped dynamically to their respective UNIX accounts (to assess file permissions), thus simplifying the unification of the two separate namespaces.

This setup is especially powerful in conjunction with the IIBM System Storage N series autohome feature, which provides all Windows users with share-level access to their own home directories without the painstaking administrative efforts typically required on other Windows file servers. (Users automatically see their own home directory as a share in the Network Neighborhood, but not other users' home directories, unless those others have been deliberately and explicitly exported as publicly visible shares.)

With multiprotocol access, Windows clients can store and access data side-by-side with UNIX-based clients without compromising their respective file attributes, security models, or performance. Users with Windows desktops can work within the single instances of their home or project directories, with Windows-based applications executing locally, or with UNIX-based applications running on a server. Whether written to the storage system through NFS or CIFS, documents can be accessed directly by a wide variety of web browsers through HTTP.

Multiprotocol access, as illustrated in Figure 25-1, liberates the data infrastructure, largely freeing it from the constraints of operating system preference or existing investments.

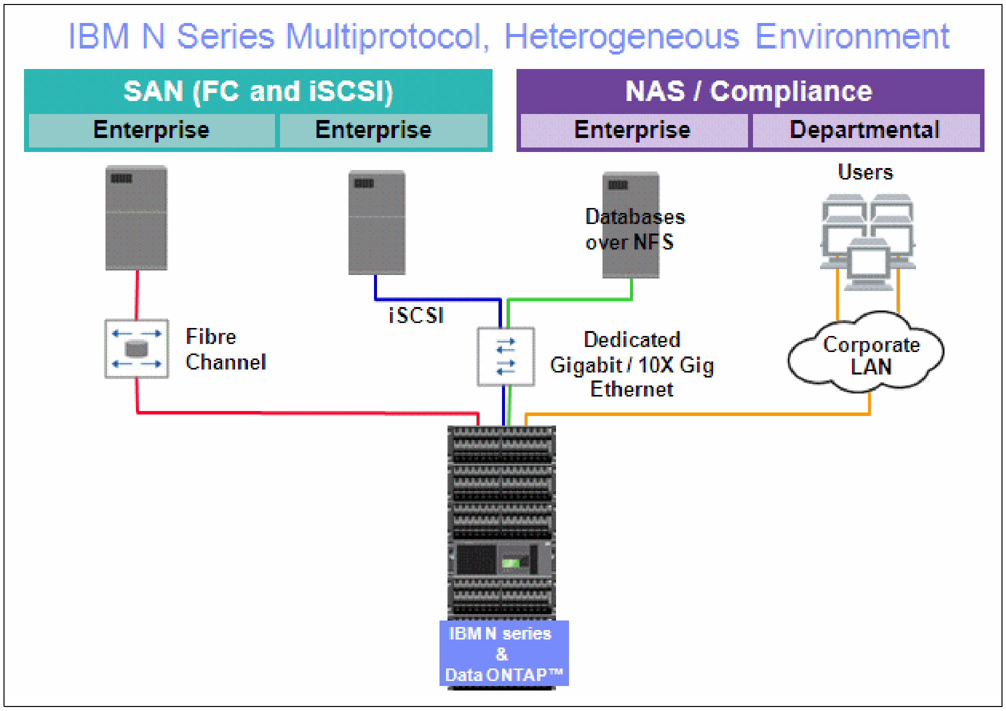


Figure 25-1 Network-attached storage and SAN protocols

In the following sections, we describe these topics:

- The IBM System Storage N series multiprotocol storage system architecture
- The implications of multiprotocol access for system administrators and users
- The evolution of Data ONTAP software for Windows

25.2 File system permissions

IBM System Storage N series storage systems support both UNIX-style and NTFS-style file permissions. Because the ACL security model in NTFS is more complex than the file security model used in UNIX, no one-to-one mapping can be made between them.

The fundamental problem occurs when a Windows or similar type of client (which expects an ACL) accesses a UNIX file, or when a UNIX client (which expects UNIX file permissions) accesses a Windows file. In these cases, the file server must sometimes authorize the request using a user identity that has been mapped from one system to the other, or in some cases using even a set of permissions that has been synthesized for one system based on the actual permissions for the file in the other system.

IBM System Storage N series storage systems ensures that these synthesized file permissions are at least as restrictive as the true file permissions. In other words, if user XYZ cannot access a file using the true file permissions, the same is true when using the synthesized file permissions. Data ONTAP has a mechanism called *UID-to-SID mapping* to address this issue.

25.2.1 UNIX file permissions

UNIX file permissions are usually represented as three sets of concatenated rwx triplets. Example 25-1 illustrates a directory listing in a UNIX File System.

Example 25-1 UFS permissions

```
lrwxrwxrwx 1 agy eng 10 Sep 2 14:42 perms.doc->perms.html
-rw-r--r-- 1 agy eng 1662 Sep 2 14:32 perms.html
-rw-rw---- 1 agy eng 2399 Feb 19 1998 privileges.nt.txt
drwxr-xr-x 2 agy eng 4096 Sep 2 14:42 work
```

The first 10 characters on each line indicate the file type and permissions for the listed file:

- The first character is as follows:

  d A letter (d) indicates that the file is a *directory*.
  l A letter (l) indicates that the file is a symbolic *link*.
  - A dash (-) character indicates that it is a regular *file*.

- The next three characters specify whether the user (agy, in this example) can read (r), write (w), or execute (x) the file.
- The following three characters specify the permissions for the group associated with the file (eng, in this example).
- The last three characters specify the permissions for users who are not the owner or members of the file's group.

Referencing Example 25-1, perms.doc is a symbolic link that anyone can traverse (obtaining the file perms.html). `perms.html` is a regular file that anyone can read, but only the user agy can write. `privileges.nt.txt` is a file that agy (or anyone from the group eng) can both read and write. Finally, `work` is a directory that anyone can search and read files in, but only agy can insert files into or delete files from it.

When user agy attempts to access a UNIX file named `nfsfile`, the behavior of the file system depends on what kind of file it is. First, though, the request is checked against the permissions associated with the file. For example, if it is a read request, the following processing occurs:

- If agy is the `nfsfile` owner and the owner has read permission on `nfsfile`, then the request can be honored.
- Otherwise, if agy is a member of the file's group and the group has read permission, the request can be honored.
- Otherwise, if all others have read permission, the request can be honored.
- However, if none of these tests succeed, the request is denied.

25.2.2 NTFS file permissions

NTFS uses a different system for denoting file permissions. On the FAT and FAT32 file systems, which were designed as single-user file systems, there are no permissions. Anyone who can gain access to the machine has unlimited privilege on every file in the system.

The NTFS file system, however, has a sophisticated security model. This same security model is also available for use by the CIFS network file system protocol on IBM System Storage N series storage systems, so Windows clients accessing files on a storage system, whether or not they are running NTFS, can also use this security model.

In NTFS and CIFS, each file has a data structure associated with it known as a *security descriptor* (SD). This contains (among other things) the file owner's security ID (SID), as shown in Example 25-2, and another data structure known as an *access control list* (ACL).

Example 25-2 An owner's SID

```
User Name          SID
================== =============================================
itso\administrator S-1-5-21-2057036396-1631034848-1296613683-500
```

An ACL consists of one or more access control entries (ACEs), each of which explicitly allows or denies access to a single user or group. Suppose that user agy attempts to open file `pcfile` for reading. The algorithm used to determine whether to grant agy permission to do this task is conceptualized as follows:

- First, search all the ACEs that deny access to anyone. If any of them deny read access to agy specifically, or to any of the groups of which agy is a member, stop searching the ACL and reject the request.
- If no denials of access are found, continue searching the rest of the ACEs in the ACL. If one is found that grants read access to agy or to any of the groups that agy is in, stop searching the ACL and allow the request.
- If the entire ACL has been searched and no ACEs were found that allow agy to read the file specified in the request, reject the request.

It ought to be clear by now why it is not always possible to make a one-to-one mapping from the ACL model to the UNIX security model. For example, using the ACL security model, you can allow access to all the members of a group, except to some specified user. However, It cannot be done using the UNIX model.

25.2.3 NFS access of data

Figure 25-2 portrays the process when a user on a UNIX host accesses data with UNIX security style using NFS. You will see that the IBM System Storage N series behaves with little differentiation in the UNIX environment.

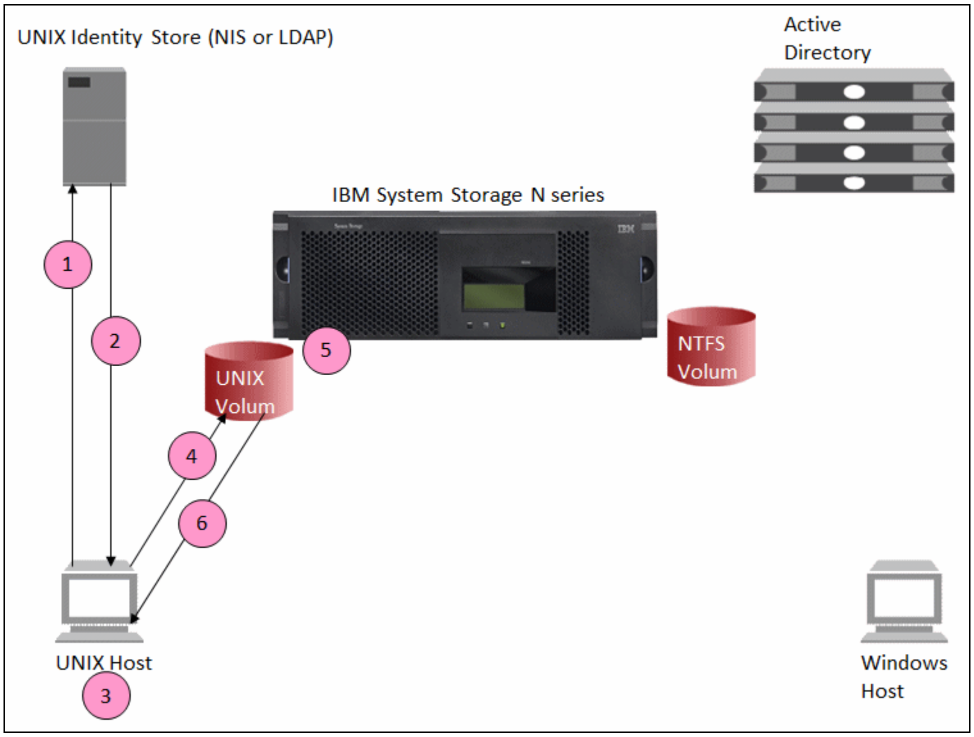


Figure 25-2 NFS access of UNIX security style data

Steps for NFS access of UNIX security style data

To provide NFS access to UNIX data, perform the following steps:

1. The user begins a login from the UNIX host.

   As part of the login process, the host requests user and group information for the user from the name services configured in the `/etc/nsswitch.conf` file. The data can be retrieved from local files, an Network Information Service (NIS) server, or an LDAP server.

2. The configured name service returns user and group information to the UNIX host.

   All user information needed to log in, including UID, GID, and the user shell and home directory, is returned. Additionally, the user's secondary group GIDs are returned.

3. Using the user information retrieved from the identity store, the user is authenticated and allowed access to the UNIX host.

   The user and group information retrieved in Step 1 is cached and used to determine access rights to local and remote resources.

4. The user requests access to a mounted IBM System Storage N series file system.

   The storage system checks export options at the time the file system is mounted. Export options can affect the user's ability to perform a requested action. For example, if the file system is mounted read only, the user cannot write to the file system even if file permissions allow it.

   The file or folder access request contains the user's UID and GIDs, including secondary GIDs.

5. The storage system uses the UID and GIDs sent in the access request for the access check.

   Because the file system use the UNIX security style, with UNIX `rwxrwxrwx` type of file permissions, the storage system uses the UID and GIDs with the access check. The system determines if the UID is the owner of the file. If not, the system determines if any of the user's groups are the objects group owner. If not, then `other` permissions apply. The system then determines if the desired action is allowed or not allowed.

6. The system replies to the access request, either permitting the requested action if the file permissions allow it or denying the action if the permissions do not allow it.

> **Scope:** NFS access of UNIX security style data does not include a user mapping step. This section does not describe how the process handles access requests by root.

NFS access of NTFS security style data

The System Storage N series offers the unique capability of providing mixed access to both UNIX and Windows users/hosts. Figure 25-3 describes the process when a user on a UNIX host accesses data with NTFS security style using NFS.

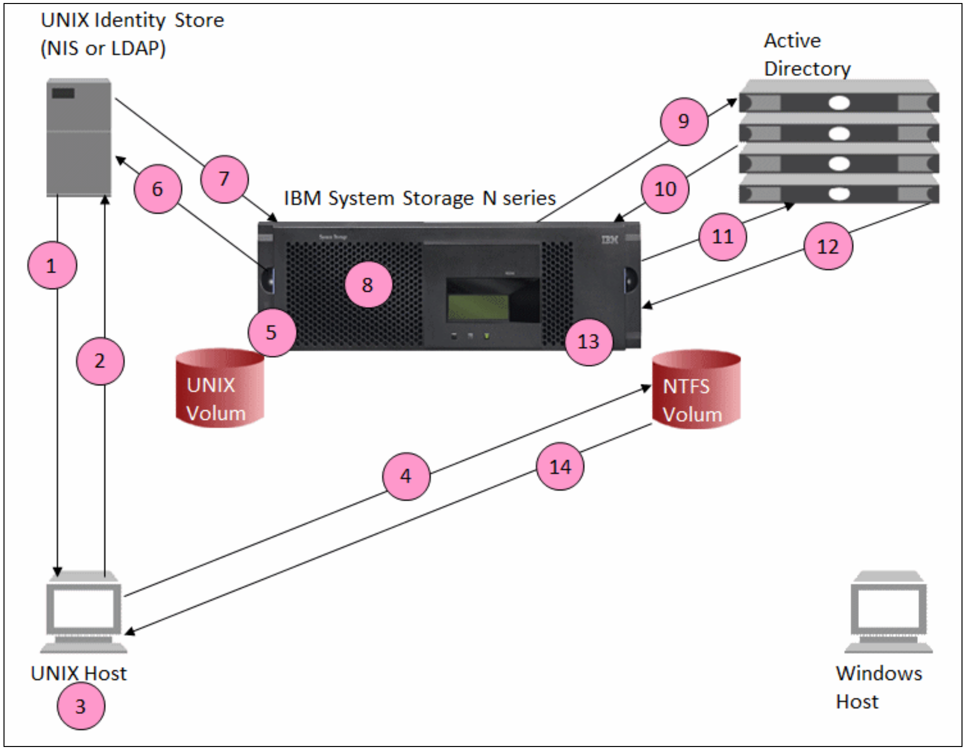


Figure 25-3   NFS access of NTFS security style data

Steps for NFS access of NTFS security style data

To provide NFS access of NTFS data, perform the following steps:

1. The user begins a login from the UNIX host.

   As part of the login process, the host requests user and group information for the user from the name services configured in the `/etc/nsswitch.conf` file. The data can be retrieved from local files, an NIS server, or an LDAP server.

2. The configured name service returns user and group information to the UNIX host.

   All user information needed to log in, including UID, GID, and the user shell and home directory, is returned. Additionally, the user's secondary group GIDs are returned.

3. Using the user information retrieved from the identity store, the user is authenticated and allowed access to the UNIX host.

   The user and group information retrieved in step 1 is cached and used to determine access rights to local and remote resources.

4. The user requests access to a mounted IBM System Storage N series file system.

   The storage system checks export options at the time the file system is mounted. Export options can affect the user's ability to perform a requested action. For example, if the file system is mounted read only, the user cannot write to the file system even if file permissions allow it.

   The UNIX client sends an access request for a storage system mount that contains the user's UID and GIDs, including secondary GIDs.

5. The user requests access to data on the storage system that uses NTFS-style permissions, but the request contains UNIX-style UID and GIDs.

   The storage system cannot use the UID and GIDs sent in the access request for the access check. Instead, the storage system must compare Windows user and group information to the file's Window ACL to determine access for the user. Therefore, the storage system maps the UNIX user to the corresponding Windows user.

6. The system queries the configured UNIX identity store, supplying the user's UID, and requests the user name.

   When the user requests access from a UNIX host, the request contains the user's UID, but does not contain the user name. Because mapping is done by user name, the system must determine the UNIX user name before the mapping process can proceed.

7. The UNIX name service returns the name of the UNIX user to the storage system.

8. The storage system uses the UNIX user name to begin the user mapping process.

   The storage system checks `/etc/usermap.cfg` and the LDAP user mapping entries (if configured) to see if there is a specific mapping entry for the UNIX user.

   If there is a specific mapping entry, the storage system uses this entry for the user mapping process.

   If there is not a specific mapping entry, the storage system assumes that the Windows user name is the same as the UNIX user name and uses this name automatically during the mapping process.

Important: If there is not a specific mapping entry in /etc/usermap.cfg or in the LDAP store (if LDAP user mapping is configured), the Windows user name is assumed to be the same as the UNIX user name, and this name is automatically used in the mapping process. In addition to correlating a Windows user name with a UNIX user name, the mapping process is important in determining if the mapped user is a valid user or not. If the mapped name is not a valid user, access can be allowed as the default NT user or completely disallowed, depending on how the WAFL option wafl.default_nt_user is configured.

9. The storage system queries Active Directory to determine if the mapped user name is a valid Windows user.

   If the user name is a valid Windows user, the mapping process continues.

   If the user name is not a valid Windows user, the user is either mapped to the generic NT user or access is denied, depending on how the WAFL option `wafl.default_nt_user` is configured.
10. Active Directory replies to the query, either supplying information about the user or returning an error indicating that the user is not found.
11. The storage system queries Active Directory for the mapped Windows user's SIDs (user SID and all group SIDs).
12. Active Directory replies to the query, supplying information about the user, including the user's SIDs.
13. The storage system uses the user's SIDs and compares them to the ACLs of files and directories for which access has been requested. The system then determines if the desired action is allowed or not allowed.
14. The system replies to the access request, either permitting the requested action if the file permissions allow it or denying the action if the permissions do not allow it.

Steps: The complete credential retrieved in steps 5 through 12 is stored in the WAFL credential cache. With subsequent access requests, the wcc is checked instead of repeating Steps 5 through 12. The wcc entries expire after 20 minutes (default). After the entry expires, the user mapping step is required again, with the results again being cached.

25.2.4 NTFS access modes

The Windows file permissions model defines more access modes than UNIX does (read, write, and execute). Table 25-1 explains what each of the basic file access modes means.

Table 25-1 Basic file access modes

Request type	The object is a folder	The object is a file
Read (r)	Displays the file's data, attributes, owner, and permissions.	Displays the file's data, attributes, owner, and permissions.
Write (w)	Writes the file, appends the file, and reads or changes its attributes.	Writes the file, appends the file, and reads or changes its attributes.
Read and execute (x)	Displays the folder's contents. Displays the data, attributes, owner, and permissions for files within the folder. Runs files within the folder.	Displays the file's data, attributes, owner, and permissions, and runs the file.

Request type	The object is a folder	The object is a file
Modify	Reads, writes, modifies, and executes files in the folder. Changes attributes and permissions. Takes ownership of the folder or files within.	Reads, writes, modifies, executes, and changes the file's attributes.
Full control	Reads, writes, modifies, and executes files in the folder. Changes attributes and permissions. Takes ownership of the folder of files within.	Reads, writes, modifies, executes, and changes the file's attributes and permissions, and takes ownership of the file.
List folder contents	Displays the folder's contents. Displays the data, attributes, owner, and permissions for files within the folder. Runs files within the folder.	

Windows XP, 2000, 2003, and 2008 also support special access permissions, which are made by combining the permissions described in Example 25-1 on page 384.

Table 25-2 lists these special access permissions and their combinations.

Table 25-2 Special access permissions and combinations

File special permissions	Full control	Modify	Read and execute	Read	Write
Traverse folder/execute file.	X	X	X		
List folder/read data.	X	X	X	X	
Read attributes.	X	X	X	X	
Read extended attributes.	X	X	X	X	
Create files/write data.	X	X			X
Create folders/append data.	X	X			X
Write attributes.	X	X			X
Write extended attributes.	X	X			X
Delete Subfolders and files.	X				
Delete.	X	X			
Read permissions.	X	X	X	X	X
Change permissions.	X				
Take ownership.	X				
Synchronize.	X	X	X	X	X

25.3 File service for environments with NFS and CIFS

IBM System Storage N series storage systems support both NFS-style and CIFS-style file permissions. NFS-style file permissions are widely used in most UNIX systems. CIFS-style file permissions are used in Windows when communicating over networks.

Because the ACL security model in CIFS is more complex than the NFS file security model used in UNIX, no one-to-one mapping can be done between them. This mathematical fact has forced all vendors of multiprotocol file storage products to develop non-mathematical strategies to blend the two systems and make them as compatible as possible. This section explains the IBM System Storage N series approach to this problem.

File service for heterogeneous environments (UNIX workstations plus Windows servers) is challenging. Windows NFS software can be installed on Windows clients or SAMBA can be installed on a UNIX server, but these approaches are either costly or time-consuming, or they introduce an extra layer of file system emulation.

To reduce complexity, making changes must be done at the file server rather than altering a large (and growing) number of Windows clients or adding a file system emulation layer that reduces performance. In other words, in a heterogeneous environment, the file server must support remote file access protocols for both UNIX-based clients and Windows clients.

The alternative, that is, using separate file servers for each protocol, can increase costs due to administrative complexity and redundant investments in storage. Routine administrative functions such as backup and restore are duplicated, and it is still difficult to implement applications that must facilitate sharing of data between UNIX and Windows users.

Perhaps worst of all, perpetuating an arrangement of separate servers for distinct sets of UNIX and Windows clients creates an awkward situation for users that need to access the same files (in their home directories, for example) with locally executing applications on their server, and by means of an X Window System session on a UNIX host. See Figure 25-4.

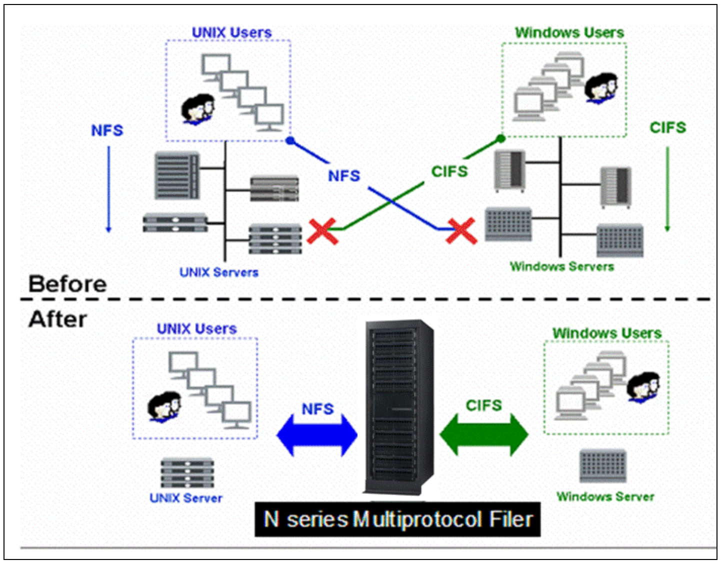


Figure 25-4 Multiprotocol IBM System Storage N series storage system

25.3.1 CIFS access of UNIX security data

In a multiprotocol environment enabled by the IBM System Storage N series, requests for heterogeneous access to data are common. A method of access for CIFS to UNIX data and its security structure must be provided. This section explains how to provide that access and structure on the IBM System Storage N series.

Steps for CIFS access of UNIX security style data

Figure 25-5 describes the process when a user on a Windows host accesses data with UNIX-style security using CIFS.

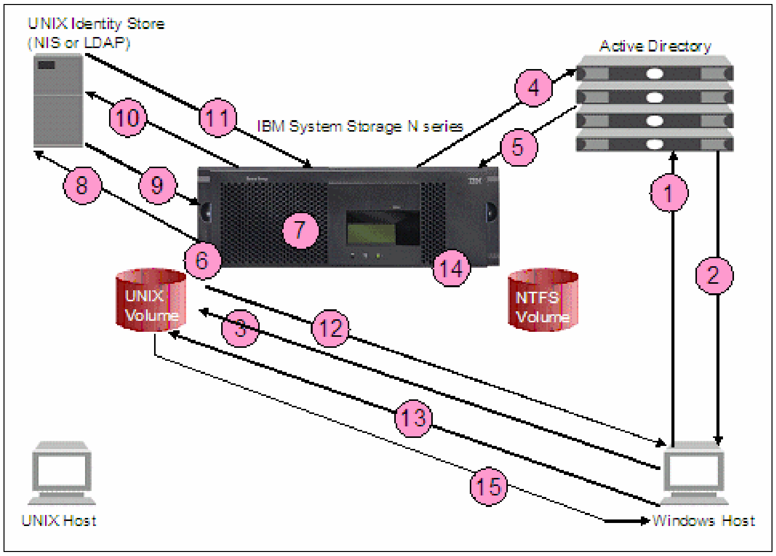


Figure 25-5 CIFS access of UNIX security style data

Detailed procedure

The following steps outline the steps illustrated in Figure 25-5 in detail:

1. The user begins a login from the Windows host.

   As part of the login process, the host communicates with the user's domain controller or the local security database (if the user is logging in as a local user). The process can vary depending on whether the user is part of an Active Directory domain or an NT domain, or is a local user.

2. The user's credential is returned to the Windows host.

   The result is a credential that contains the user SID and the user's group SIDs.

   If the user is from a non-trusted domain and the domain guest account is enabled, the domain controller authenticates the user with the guest credentials. If a user is logging in locally to a Windows workstation where the local guest account is enabled and does not have a local account, the user is authenticated with the local guest credentials.

3. The user requests access to data stored on a IBM System Storage N series file system through the session setup and connect requests (either through mapping a drive or accessing through a UNC path).

   When a Windows user requests access to data on a IBM System Storage N series file system, an authenticated session must first be set up and then the connection to the share must be made. The steps for session setup depend on the authentication protocol negotiated between the storage system and the client.

The Windows client determines which protocol is used. The storage system supports NTLMv1, NTLMv2, Kerberos, and clear text password authentication:

- Windows NT LAN Manage (NTLM) authentication:

  Used when the storage system is a member of an NT 4 domain, when operating in local workgroup mode, or if the Windows client requests that NTLM authentication be used.

  After the NTLM authentication protocol is negotiated, the client sends a session setup request. The request contains the user name along with NTLM authentication information and other information used in session setup.

- Kerberos authentication:

  Used when the storage system is a member of an Active Directory domain and Kerberos authentication is negotiated between the client and the storage system.

  After the Kerberos authentication protocol is negotiated, the client sends a session setup request. Kerberos uses tickets for authenticating user requests to network services. The session setup request contains a Ticket Granting Service (TGS) ticket for the storage system's CIFS service. This ticket is embedded in the session setup request and contains a security blob with all of the information necessary for the storage system to authenticate the user. The TGS ticket also contains the user SID and the SIDs for all of the user's domain groups.

4. The storage system processes the session setup request:

   - Windows NT LAN Manage (NTLM) authentication:

     In a domain environment, the storage system does not store information about the user's Windows password; therefore, with NTLM authentication, the storage system must send the user information and NTLM authentication information to the domain controller, which performs the actual user authentication.

   - Kerberos authentication:

     Even though the request contains all of the information needed to authenticate the user session, the user name is not in the security blob; therefore, the storage system first consults the system's SID cache to see if the user name for that SID is in cache. If it is not, the storage system queries the user's domain controller and does a SID lookup.

5. The domain controller processes the request sent by the storage system and provides the information that the storage system needs to complete the session setup:

   - Windows NT LAN Manage (NTLM) authentication:

     The domain controller returns success or failure for the user authentication. If the user is authenticated, the domain controller also returns SIDs for the user and the user's groups. The storage system then includes the SIDs of any local groups to which the user belongs and stores this information in the user's CIFS session cache.

     If the user who is requesting session setup is a local storage system user, the session authentication is handled locally by the storage system, and the SIDs of the local groups to which the user belongs are added to the user's session cache.

   - Kerberos authentication:

     The domain controller returns the user name to the storage system, which already has the user SID and the user's domain group SIDs. The storage system then includes the SIDs of any local groups to which the user belongs, and all SID information is stored in the user's session cache.

Requirements: In a domain environment, before the storage system can query the domain controllers, the storage system must have a machine account in the domain and must authenticate to the domain and establish a Netlogon pipe. In a Kerberos environment, the storage system must obtain a valid Granting Ticket for the machine account. There can be additional traffic between storage system and domain controllers during the session setup; however, the basic process is as outlined here.

If the user is connecting as a guest, the storage system denies access unless the option to allow guest account connection is set, as shown in Example 25-3.

Example 25-3 Setting the option to allow a guest connection

```
itsotuc2*> options cifs.guest_account unix_name
No confirmation will be given that this has been successful. Run the following
to display current settings
itsotuc2*> options cifs.guest_account
cifs.guest_account         ‘value of unix_name’
```

In this example, `unix_name` is a user created in the UNIX identity store, that is, `/etc/passwd`, NIS, or LDAP.

If guest access to the storage system is not desired, set this option to a null value, as shown in Example 25-4. All guest access requests will be denied. The default for this option is NULL. Guest access is denied by default.

Example 25-4 Setting CIFS guest account to null

```
itsotuc2*> options cifs.guest_account ""
No confirmation will be given that this has been successful. Run the following
to display current settings
itsotuc2*> options cifs.guest_account
cifs.guest_account
```

If the user is connecting to the storage system as a local user, the `cifs.guest_account` option determines whether access as a guest user is allowed or denied.

6. The user requests access to data on the storage system that uses UNIX-style file permissions, but the request contains Windows-style SIDs.

   The storage system cannot use the SIDs sent in the file access request for the access check. Instead, the storage system must compare UNIX user and group information to the file UNIX permission to determine access for the user. Therefore, the storage system maps the Windows user to the corresponding UNIX user.

   Additionally, current Data ONTAP implementations store the UNIX credentials with the user’s session cache. Therefore, the Windows-to-UNIX user mapping occurs prior to completion of the session setup and connection to the share.

7. The storage system uses the Windows user name to begin the user mapping process.

   The storage system checks `/etc/usermap.cfg` or the LDAP user mapping entries (if configured) to see if there is a specific mapping entry for the Windows user.

   If there is a specific mapping entry, the storage system uses this entry for the user mapping process.

   If there is not a specific mapping entry, the storage system assumes that the UNIX user name is the same as the Windows user name and uses this name during the mapping process automatically.

Important: If there is not a specific mapping entry in /etc/usermap.cfg or in the LDAP store (if LDAP user mapping is configured), and the Windows user name is used automatically, the mapping process is still performed. It is important to determine if the UNIX user is a valid user or not. If the mapped name is not a valid user, access can be as the default user or it can be completely disallowed, depending on how the WAFL option wafl.default_unix_user is configured.

8. The storage system queries the configured UNIX name services to determine if the mapped user name is a valid UNIX user.

   If the user name is a valid UNIX user, the mapping process continues.

   If the user name is not a valid UNIX user, the user is either mapped to the generic UNIX account or access is denied, depending on how the WAFL option `wafl.default_unix_user` is configured.

9. The configured UNIX name service replies to the query, either supplying information about the user or returning an error indicating that the user is not found.

10. The storage system queries the UNIX name services for the mapped UNIX user's user and group information.

11. The UNIX name service replies to the query, supplying information about the UNIX user's credential, including the user's UID and GIDs (including secondary group GIDs).

    The UNIX credential is stored with the user's cached authenticated session information.

    The user mapping information in this process is not stored in the wcc. If the same user establishes a new CIFS connection, the process is re-executed.

12. Now that the UNIX credentials have been added to the authenticated session cache, the storage system can reply to the session setup request, with either a success or a failure.

    If the session request was successful, the connection request to the share can be processed, and is either allowed or denied based on evaluation of user and group share permissions.

13. The Windows user requests access through the mapped drive to a file or folder in the UNIX security style volume.

14. The storage system uses the UNIX user information stored in the user's session cache and compares it to the UNIX permissions on files and directories for which access has been requested.

    The system then determines if the desired action is allowed or not allowed.

    Because the file system has UNIX security style, with UNIX `rwxrwxrwx` type of file permissions, the storage system uses the UID and GIDs with the access check. The system determines if the UID is the owner of the file. If not, the system determines if any of the user's groups are the objects group owner. If not, then `other` permissions apply (if the default user is configured). The system then determines if the desired action is allowed or not allowed.

15. The system replies to the access request, either permitting the requested action if the file permissions allow it or denying the action if the permissions do not allow it.

Scope: This section does not describe how the process handles access requests by Windows administrators.

25.3.2 CIFS access of NTFS data

This section describes native CIFS access to NTFS data. It provides high-level steps for CIFS access.

Steps for CIFS access of NTFS security style data

Figure 25-6 describes the process when a user on a Windows host accesses data with NTFS security style using CIFS.

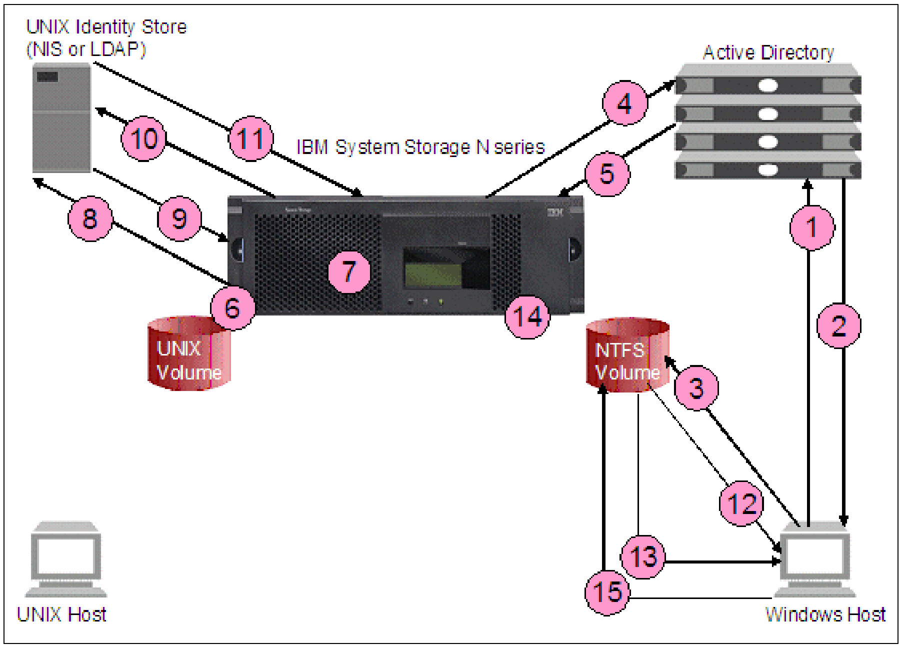


Figure 25-6 CIFS access of NTFS security style data

Detailed procedure

The following steps outline the steps illustrated in Figure 25-6 in detail:

1. The user begins a login from the Windows host.

   As part of the login process, the host communicates with the user’s domain controller or the local system’s local security database (if the user is logging in as a local user). The process can vary, depending on whether the user is part of an Active Directory domain or an NT domain, or is a local user.

2. The user’s credential is returned to the Windows host.

   The result is a credential that contains the user SID and the user’s group SIDs.

   If a user is from a nontrusted domain and the domain guest account is enabled, the domain controller authenticates the user with the guest credentials. If a user is logging in locally to a Windows workstation where the local guest account is enabled and does not have a local account, the user is authenticated with the local guest credentials.

3. The user requests access to data stored on an IBM System Storage N series file system through the session setup and connect requests, either through mapping a drive or accessing through a Uniform Naming Convention (UNC) path.

Path: The format for a UNC path is \\server name\shared volume\shared directory\name of file and is not case-sensitive.

Typically, the location of a file is described by the drive letter and folder in which it is located. It is likely to be unique to the host that this location is mapped to, whereas specifying a UNC path is more specific and is common across all operating systems.

When a user requests access to data on a IBM System Storage N series file system, an authenticated session must first be set up and then the connection to the share must be made. The steps for session setup depend on the authentication protocol negotiated between the storage system and the client. The Windows client determines which protocol is used. The storage system supports NTLMv1, NTLMv2, Kerberos, and clear text password authentication:

– NTLM authentication:

  Used when the storage system is a member of an NT 4 domain, when operating in local workgroup mode, or if the Windows client requests that NTLM authentication be used.

  After the NTLM authentication protocol is negotiated, the client sends a session setup request. The request contains the user name, along with NTLM authentication information and other information used in session setup.

– Kerberos authentication:

  Used when the storage system is a member of an Active Directory domain and Kerberos authentication is negotiated between the client and the storage system.

After the Kerberos authentication protocol is negotiated, the client sends a session setup request. Kerberos uses tickets for authenticating user requests to network services. The session setup request contains a Ticket Granting Service (TGS) ticket for the storage system's CIFS service. This ticket is embedded in the session setup request and contains a security blob with all of the information necessary for the storage system to authenticate the user. The TGS ticket also contains the user SID and the SIDs for all of the user's domain groups.

4. The storage system processes the session setup request:

   – Windows NT LAN Manage (NTLM) authentication:

     In a domain environment, the storage system does not store information about the user's Windows password; therefore, with NTLM authentication, the storage system must send the user information and NTLM authentication information to the domain controller, which performs the actual user authentication.

   – Kerberos authentication:

     Even though the request contains all of the information needed to authenticate the user session, the user name is not in the security blob; therefore, the storage system consults the system's SID cache to see if the user name for that SID is in the cache. If it is not, the storage system queries the user's domain controller and does a SID lookup.

5. The domain controller processes the request sent by the storage system and provides the information that the storage system needs to complete the session setup:

   – Windows NTLM authentication:

     The domain controller returns success or failure for the user authentication. If the user is authenticated, the domain controller also returns SIDs for the user and the user's groups. The storage system then includes the SIDs of any local groups to which the user belongs and stores this information in the user's CIFS session cache.

     If the user who is requesting session setup is a local storage system user, the session authentication is handled locally by the storage system, and the SIDs of the local groups to which the user belongs are added to the user's session cache.

   – Kerberos authentication:

     The domain controller returns the user name to the storage system, which already has the user SID and all the user's domain group SIDs. The storage system then includes the SIDs of any local groups to which the user belongs, and all SID information is stored in the user's session cache.

> **Requirements:** In a domain environment, before the storage system can query the domain controllers, the storage system must have a machine account in the domain and must authenticate to the domain and establish a Netlogon pipe. In a Kerberos environment, the storage system must obtain a valid Ticket Granting Ticket for the machine account. There can be additional traffic between storage system and domain controllers during the session setup; however, the basic process is as outlined here.

If the user is connecting as a guest, the storage system denies access unless the option to allow guest account connection is set, as shown in Example 25-5.

Example 25-5 Setting the option to allow guest connection

```
itsotu2*> options cifs.guest_account unix_name
No confirmation will be given that this has been successful. Run the following
to display current settings
itsotuc2*> options cifs.guest_account
cifs.guest_account         'value of unix_name'
```

In this example, `unix_name` is a user created in the UNIX identity store, that is, `/etc/passwd`, NIS, or LDAP.

If guest access to the storage system is not desired, set this option to a null value, as shown in Example 25-6. All guest access requests will then be denied. The default for this option is NULL. Guest access is denied by default.

Example 25-6 Setting CIFS guest account to null

```
itsotuc2*> options cifs.guest_account ""
No confirmation will be given that this has been successful. Run the following
to display current settings
itsotuc2*> options cifs.guest_account
cifs.guest_account
```

If the user is connecting to the storage system as a local user, the option `cifs.guest_account` determines if access as a guest user is allowed or denied.

6. The user requests access to data on the storage system that uses NTFS-style file permissions. The storage system uses the Windows-style SIDs when evaluating file access rights; however, current implementations of Data ONTAP always perform a user mapping when data is requested via CIFS.

   The UNIX credentials, containing all of the mapped UNIX user's UID and GIDs, are stored with the Windows user's session authentication cache; therefore, a Windows-to-UNIX user mapping must be done before the session setup is complete.

   Therefore, the storage system maps the Windows user to the corresponding UNIX user, even in the case where a Windows user is accessing data stored in an NTFS volume.

7. The storage system uses the Windows user name to begin the user mapping process.

   The storage system checks `/etc/usermap.cfg` or the LDAP user mapping entries (if configured) to see if there is a specific mapping entry for the Windows user.

   If there is a specific mapping entry, the storage system uses this entry for the user mapping process.

   If there is not a specific mapping entry, the storage system assumes that the UNIX user name is the same as the Windows user name and uses this name during the mapping process automatically.

   Important: If there is not a specific mapping entry in /etc/usermap.cfg or in the LDAP store (if LDAP user mapping is configured) and the Windows user name is automatically used, the mapping process is still performed. It is important to determine if the UNIX user is a valid user or not. If the mapped name is not a valid user, access can be as the default user, or it can be completely disallowed, depending on how the WAFL option wafl.default_unix_user is configured.

8. The storage system queries the configured UNIX name services to determine if the mapped user name is a valid UNIX user.

   If the user name is a valid UNIX user, the mapping process continues.

   If the user name is not a valid UNIX user, the user is either mapped to the generic UNIX account or access is denied, depending on how the WAFL option `wafl.default_unix_user` is configured.

9. The configured UNIX name service replies to the query, either supplying information about the user or returning an error indicating that the user is not found.

10. The storage system queries the UNIX name services for the mapped UNIX user's user and group information.

11. The UNIX name service replies to the query, supplying information about the user, including the user's UID and GIDs (including secondary group GIDs).

    The UNIX credential is stored with the user's cached authentication session information.

    The user mapping information in this process is not stored in the wcc. If the same user establishes a new CIFS connection, the process is executed again.

12. Now that the UNIX credentials have been added to the authenticated session cache along with the Windows user SID information, the storage system can reply to the session setup request, with either a success or a failure.

    If the session request was successful, the connection request to the share can be processed, and is either allowed or denied, based on evaluation of user and group share permissions.

13. The Windows user requests access through the mapped drive to a file or folder in the NTFS security style volume.

14. The storage system uses the Windows user information stored in the user’s session cache and compares it to the NTFS permissions on files and directories for which access has been requested. The system uses the Windows user and group SIDs to determine if the desired action is allowed or not allowed. The system then determines if the desired action is allowed or not allowed.

15. The system replies to the access request, either permitting the requested action if the file permissions allow it or denying the action if the permissions do not allow it.

Scope: This section does not describe how the process handles access requests by Windows administrators.

25.3.3 CIFS access of NTFS data

A IBM System Storage N series storage system supports four CIFS access methods. The method is chosen during CIFS setup.

Joining a workgroup using `/etc/password`, NIS, or LDAP for authentication offers a substantially different method for managing file and folder access. Example 25-7 illustrates this method.

Example 25-7 CIFS setup and access method designation

```
1) Active Directory domain authentication (Active Directory domains only)
(2) Windows NT 4 domain authentication (Windows NT or Active Directory domains)
(3) Windows Workgroup authentication using the filer's local user accounts
(4) /etc/passwd and/or NIS/LDAP authentication

Selection (1-4)? [1]:
```

In this example, we describe CIFS access in workgroup mode.

Steps for CIFS access in workgroup mode

When CIFS access is based on workgroup using `/etc/password`, NIS, or LDAP, session authentication is done on the basis of user names and passwords that are stored in the UNIX directory stores. Even if local Windows users are created on the storage system using the **`useradmin`** command, they are not used for session authentication. All authentication is done based on UNIX user information that is stored in the UNIX identity stores.

If the volume that is accessed uses NTFS-style security, ACLs are not used during the access check, even if the file or folders has valid ACLs. File access is determined by share level permissions. From the user's point of view, an NTFS-style security volume is a FAT volume.

If the volume being accessed uses UNIX-style security, UNIX file permissions in conjunction with share permissions are used to determine access.

Detailed procedure

Here are the steps for CIFS access in workgroup mode:

1. The Windows client requests access to the data.

   The user who is logged in to the Windows client can be a domain user or a local user. However, when mapping the drive, the user must use one of the following methods:

   – Map the drive using a different user name, where the user name and password are the name and password of a UNIX user stored in the UNIX identity store.

   – Use pass-through authentication, where the logged-in user is using the same user name and password as a user stored in the UNIX identity store of the storage device.

2. The storage device maps the Windows account name to the UNIX account name.
3. The storage device checks `/etc/password` and `/etc/group`, NIS, or LDAP to retrieve the UNIX UID and GIDs.
4. The storage system compares account information with share level permissions.
5. The storage system compares account information with UNIX permissions or DOS attribute bit.
6. If the user has both share and file level access, then access is granted.

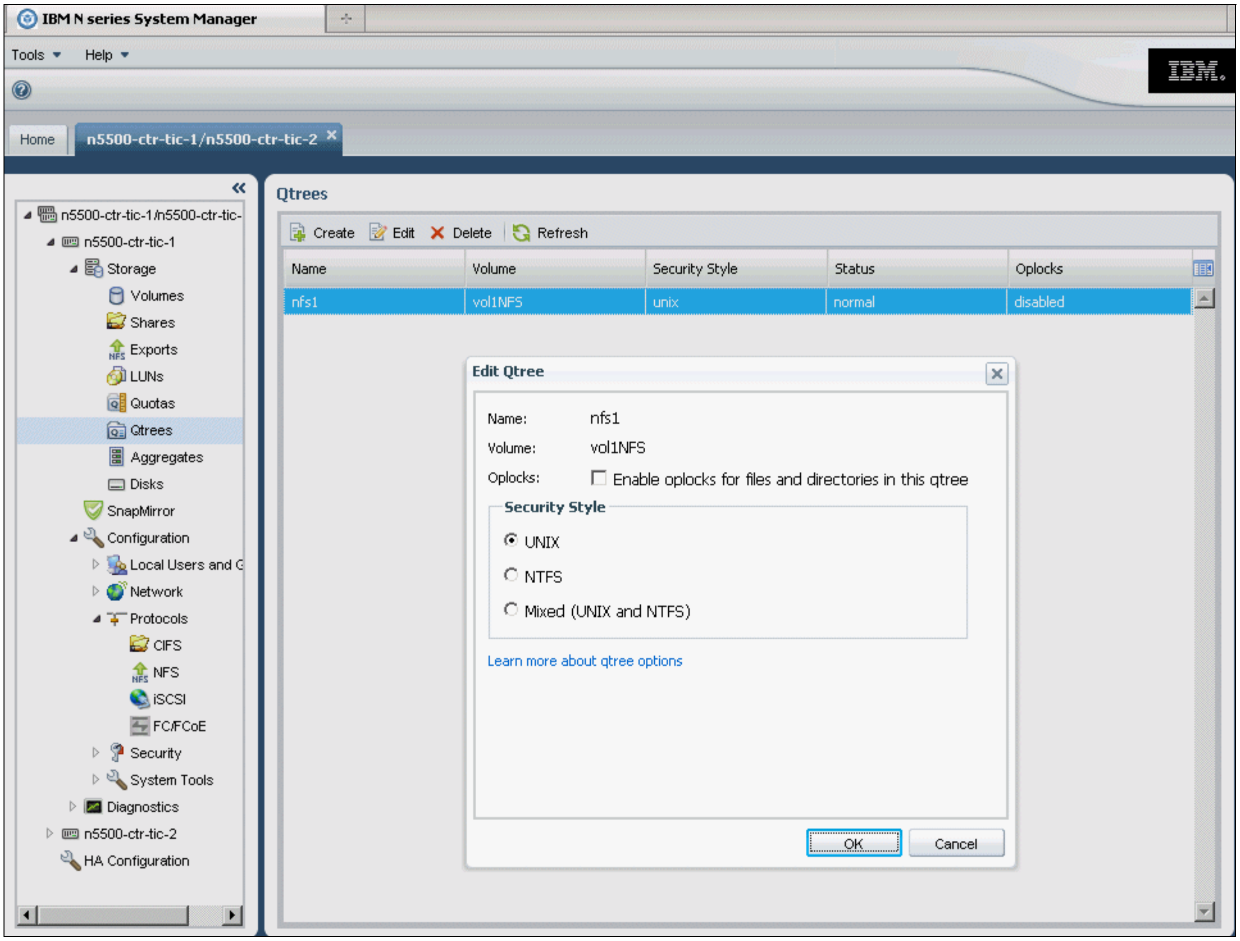
25.4 Altering qtree security at the qtree level or volume level

This section gives an illustration of how to alter qtree security at the qtree level or volume level.

To alter qtree security at the qtree level or volume level, perform the following steps:

1. Start the N series System Manager.
2. Select your storage system.
3. Select **storage**.
4. Select **qtrees**.

5. Select your qtree and perform a right mouse click on the qtree to edit the security options as shown in Figure 25-7.



Name	Volume	Security Style	Status	Oplocks
nfs1	vol1NFS	unix	normal	disabled

Edit Qtree

Name: nfs1

Volume: vol1NFS

Oplocks: □ Enable oplocks for files and directories in this qtree

Security Style

● UNIX

○ NTFS

○ Mixed (UNIX and NTFS)

Figure 25-7 qtree security style change

25.5 NFS

Most UNIX clients use NFS for remote file access. Sun Microsystems introduced NFS in 1985. Since then, it has become a *de facto* standard protocol, used by 10 million systems worldwide. NFS is particularly common on UNIX-based systems, but NFS implementations are available for virtually every modern computing platform in current use, from desktops to supercomputers. Only when used by UNIX-based systems, however, does NFS closely resemble the behavior of a client's local file system.

25.6 CIFS

The operating systems running on Windows clients do not include NFS. Instead, the protocol for remote file access is CIFS, formerly known as Server Message Block (SMB). SMB was first introduced by Microsoft and Intel in the early 1980s, and is the protocol used in several diverse PC network environments.

25.7 NFS compared to CIFS

The NFS protocol versions 1 to 3 are stateless protocols. NFS operations are *idempotent* (can be repeatedly applied harmlessly), or, if non-idempotent (file deletion, for example), are managed safely by the server. Clients are oblivious to server restarts (if service is restored promptly), with a few exceptions. The NFS protocol emphasizes that error recovery over file locking error recovery is simple if no state must preserved. To the contrary, NFS version 4 is a stateful protocol.

A CIFS file server is stateful (not stateless). The CIFS protocol emphasizes locking over error recovery, because Windows applications rely on strict locking. Strict locking requires a sustained connection. It is imperative that an active session not be interrupted.

Applications executing on Windows clients react to a CIFS server in exactly the same manner as they do to local disk drives. A down server is no different from an unresponsive disk drive. Therefore, Windows clients must be warned and allowed time to gracefully disengage (that is, save files, exit applications, and so on) before a server shuts down or restarts.

25.8 Mixing NFS and CIFS

Software solutions exist that allow UNIX-based servers to provide remote file access functionality to Windows clients without requiring NFS. Running in user mode (not in the UNIX kernel), these applications support Windows clients through CIFS. Of these, the most widely used are Samba, Hummingbird NFS Maestro, and Windows Service for UNIX (SFU) by Microsoft. Samba is a server-side installation. NFS Maestro and SFU are NFS emulators installed on the clients running NTFS.

For users with a casual need for CIFS access, or who are new to PCs and are trying to get a feel for what Windows service is like, Samba offers several advantages:

- It is available at no cost.
- It is easily available.
- It runs on most popular UNIX systems.
- It is relatively reliable for simple uses.

For more serious requirements (for example, providing primary file service for a large organization), however, Samba falls short in several important areas:

- Shallow integration with the underlying UNIX-centric file system (particularly with respect to locking mechanisms)
- Difficulty of installation, configuration, and administration
- Lack of reliable support (It is public domain software.)

25.9 Multiprotocol file service

A file server cannot easily deliver functionality to clients beyond what its local file system provides. For example, the UNIX File System (UFS) does not store the *creation* time stamp for a file. A Windows client cannot retrieve that information through the CIFS protocol if the server does not have that information to provide.

The NFS protocol does not offer mechanisms for data access beyond the capabilities of UFS. NFS was developed to extend UFS across networks. Similarly, the CIFS protocol extends a Windows-oriented file system to remote clients. In both of these cases, the remote file access protocol implementation is described as *native* to the operating system context in which it originated.

25.9.1 Emulated multiprotocol file service

When application software such as Samba provides remote data access to the files on a UNIX file server through the CIFS protocol, it must do more than simply provide semantically correct responses in its communication with clients over the wire. It must make up the difference between the personal computer's requirements and the intrinsic facilities of the server's local file system (UFS).

On a file-by-file basis, it must store additional information (in supplementary files, if the file system itself has no provision for it). For example, the server must offer case-insensitive file name lookup for personal computer clients. For older personal computer clients, the server must also generate DOS-style *8.3* file names (consisting of up to eight characters, plus up to three characters for an optional suffix).

An 8.3 file name is not inherently included in UFS, so the CIFS emulation application must store it elsewhere. Similarly, UFS does not provide case-insensitive file name lookup. The CIFS server emulation application must do that for itself.

The mapping from the Windows clients' expectations to the server's UNIX context is awkward and incomplete. The reverse situation (that is, a Windows server serving UNIX clients through NFS application software) is similarly mismatched. This dissonance is characteristic of *emulated* remote file access protocol implementations.

25.9.2 Native multiprotocol file service

IBM System Storage N series storage systems are not UNIX-based, nor are they Windows based. The microkernel operating system and WAFL file system are designed specifically for extensible file service.

Therefore NFS, CIFS, and HTTP (and in the future, additional protocols) can be implemented natively. There is no functionality mismatch (as with the emulated approaches), and kernel-based security and file locking enforcement are inherently stronger than user-space application software methods, as illustrated in Figure 25-8.

Note that in Figure 25-8, all IBM System Storage N series processing of a client's request is executed within the kernel as a series of function calls, thus eliminating the data copies and impact of the interprocess communication (IPC) between the separate processes in the emulated approach.

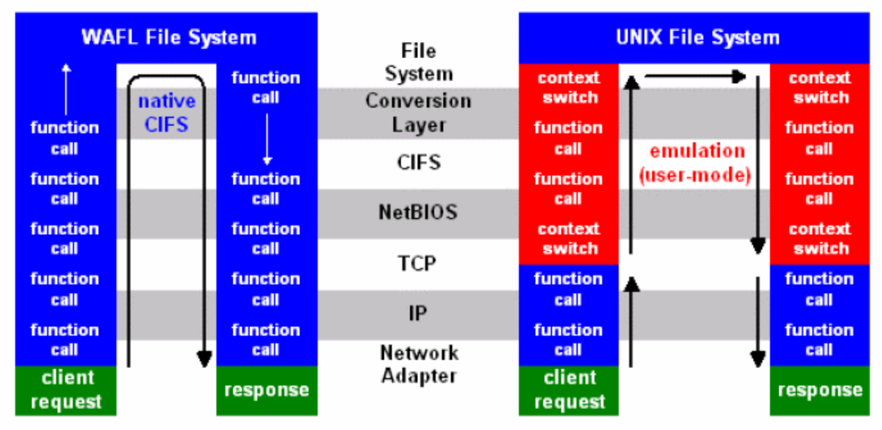


Figure 25-8 Processing of client requests

These are the three primary elements in the Data ONTAP microkernel:

- A real-time mechanism for process execution
- The WAFL file system
- The RAID manager

Of these, only the RAID manager is insensitive to whether protocol blocks are read and written in the same way, whether originally triggered by NFS or CIFS.

Figure 25-9 provides a block diagram view of Data ONTAP software. Note that NFS can operate with either TCP or UDP transport mechanisms, but IBM System Storage N series support for the CIFS file-sharing protocol uses TCP exclusively.

Attention: CIFS uses NetBIOS over TCP/IP (NBT) in the IBM System Storage N series implementation, but can also use NETBEUI or IPX/SPX in other environments.

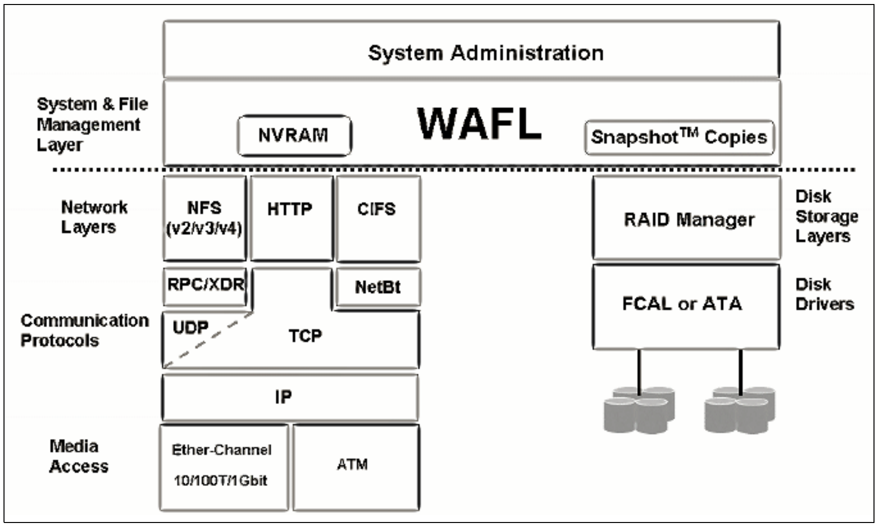


Figure 25-9 Data ONTAP software

Within the microkernel, all incoming NFS, CIFS, and HTTP requests are received by the network interface driver. After being initiated, the processing of a request is uninterrupted and continuous, as far as possible, utilizing a series of function calls. (It is entirely different from traditional file servers, which employ separate processes for the network protocol stack, the remote file system semantics, the local file system, and the disk subsystem.)

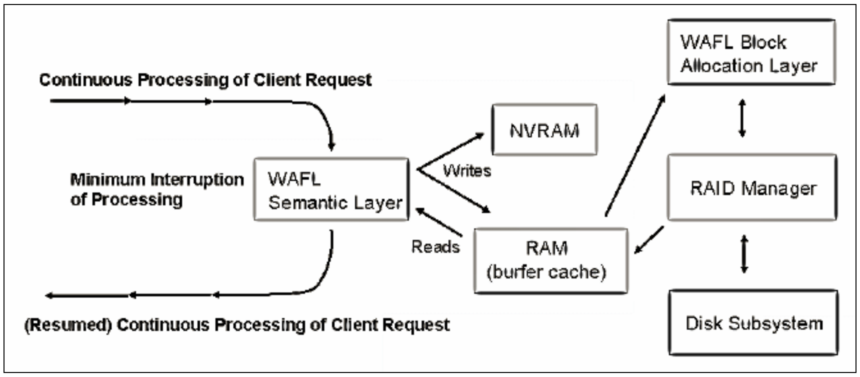
The advantages of this approach are performance and simplicity, and the simplicity enables extraordinary reliability. The process (which begins with the network interface driver) executes continuously and blocks only when waiting is unavoidable.

Ordinarily, this will occur only when the request has reached the WAFL file system. By this stage, the request has been interpreted with respect to the applicable protocol (NFS, CIFS, or HTTP) and tested for correctness and legality. If the request cannot be serviced (for example, because it is illegal or otherwise incorrect) an error code is immediately returned.

Assuming that the request can and ought to be serviced, a reply to the client is generated, *unless* one of the following conditions causes the process to block:

- A request requires data not already in memory.
- An administrative event preempts other processing.

Figure 25-10 illustrates the process flow for all cases except administrative events. (Administrative commands take effect immediately, possibly causing other operations to block.)



Continuous Processing of Client Request
Minimum Interruption of Processing
WAFL Semantic Layer
Writes
Reads
NVRAM
RAM (burfer cache)
WAFL Block Allocation Layer
RAID Manager
Disk Subsystem
(Resumed) Continuous Processing of Client Request

Figure 25-10 Client request processing

If a read request cannot be satisfied with data already in memory (from a previous read request or because of read-ahead), the process will block, while WAFL requests the RAID manager to retrieve the requested data (and the next several blocks deeper in the file. It occurs unless the minra option has been enabled, which limits read-ahead to a single block).

IBM System Storage N series storage systems serve all incoming requests with a continuous series of function calls (not IPCs) that block only when necessary, as described. This architecture is exceptionally well-suited to a multiprotocol context because it simply does not matter which protocol is used. All requests are handled expeditiously as single processes running in the kernel, regardless of protocol. Circumstances where a process might block are similar across protocols. The WAFL file system is extensible, allowing for native accommodation of both UNIX-style and Windows-style file and directory attributes.

26

Fibre Channel

This chapter provides a high level overview of the Fibre Channel (FC) protocol.

The following topics are covered:

- Fibre Channel defined
- What FC nodes are
- How FC target nodes connect to the network
- How FC nodes are identified
- Further information

26.1 Fibre Channel defined

First started in 1988 and got ANSI standard approval in 1994, Fibre Channel (FC) is now the most common connection type for storage area network (SAN). Nowadays FC SAN is already an indispensable infrastructure component in any current complex IT environment. With the proliferation of various in-house developed and packaged applications supported by various implementations from different infrastructure components, managing modern IT environment is becoming more complicated because everybody is competing with resources and expecting the best service level with minimal unscheduled downtime.

FC is a licensed service on the storage system that enables you to export LUNs and transfer block data to hosts using the SCSI protocol over a Fibre Channel fabric.

26.2 What FC nodes are

In an FC network, nodes include targets, initiators, and switches.

Targets are storage systems, and initiators are hosts. Nodes register with the Fabric Name Server when they are connected to an FC switch.

26.3 How FC target nodes connect to the network

Storage systems and hosts have adapters, so they can be directly connected to each other or to FC switches with optical cables. For switch or storage system management, they might be connected to each other or to TCP/IP switches with Ethernet cable.

When a node is connected to the FC SAN, it registers each of its ports with the switch's Fabric Name Server service, using a unique identifier.

26.4 How FC nodes are identified

Each FC node is identified by a worldwide node name (WWNN) and a worldwide port name (WWPN).

26.4.1 How WWPNs are used

WWPNs identify each port on an adapter. They are used for creating an initiator group and for uniquely identifying a storage system's HBA target ports.

- **Creating an initiator group:** The WWPNs of the host's HBAs are used to create an initiator group (igroup). An igroup is used to control host access to specific LUNs. You can create an igroup by specifying a collection of WWPNs of initiators in an FC network. When you map a LUN on a storage system to an igroup, you can grant all the initiators in that group access to that LUN. If a host's WWPN is not in an igroup that is mapped to a LUN, that host does not have access to the LUN. This means that the LUNs do not appear as disks on that host. You can also create port sets to make a LUN visible only on specific target ports. A port set consists of a group of FC target ports. You can bind an igroup to a port set. Any host in the igroup can access the LUNs only by connecting to the target ports in the port set.
- **Uniquely identifying a storage system's HBA target ports:** The storage system's WWPNs uniquely identify each target port on the system. The host operating system uses the combination of the WWNN and WWPN to identify storage system adapters and host target IDs. Some operating systems require persistent binding to ensure that the LUN appears at the same target ID on the host.

26.4.2 How storage systems are identified

When the FC protocol service is first initialized, it assigns a WWNN to a storage system based on the serial number of its NVRAM adapter. The WWNN is stored on disk.

Each target port on the HBAs installed in the storage system has a unique WWPN. Both the WWNN and the WWPN are a 64-bit address represented in the following format: nn:nn:nn:nn:nn:nn:nn:nn, where n represents a hexadecimal value.

You can use commands such as **`fcp show adapter`**, **`fcp config`**, **`sysconfig -v`**, or **`fcp nodename`** to see the system's WWNN as FC Nodename or nodename, or the system's WWPN as FC portname or portname.

26.4.3 How hosts are identified

You can use the **`fcp show initiator`** command to see all of the WWPNs, and any associated aliases, of the FC initiators that have logged on to the storage system. Data ONTAP displays the WWPN as Portname.

To know which WWPNs are associated with a specific host, see the FC Host Utilities documentation for your host. These documents describe commands supplied by the Host Utilities or the vendor of the initiator, or methods that show the mapping between the host and its WWPN. For example, for Windows hosts, you must use the *LightPulse* utility (**`lputilnt`**), *HBAnyware*, or *SANsurfer* applications, and for UNIX hosts, you must use the **`sanlun`** command.

26.4.4 How switches are identified

Fibre Channel switches have one worldwide node name (WWNN) for the device itself, and one worldwide port name (WWPN) for each of its ports.

For example, Figure 26-1 shows how the WWPNs are assigned to each of the ports on a 16-port Brocade switch. For details about how the ports are numbered for a particular switch, see the vendor-supplied documentation for that switch.

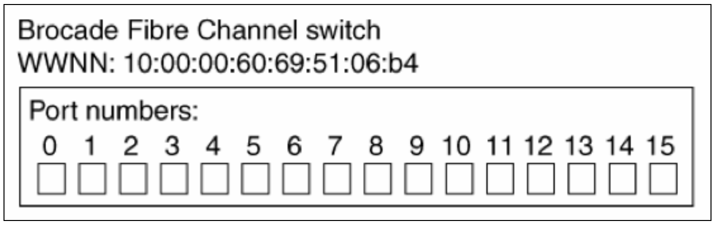


Figure 26-1 Sample switch WWNN

Port **0**, WWPN 20:**00**:00:60:69:51:06:b4
Port **1**, WWPN 20:**01**:00:60:69:51:06:b4
...
Port **14**, WWPN 20:**0e**:00:60:69:51:06:b4
Port **15**, WWPN 20:**0f**:00:60:69:51:06:b4

26.5 Further information

More details on SAN and the Fibre Channel protocol can be found in the following Redbooks publications:

- *Designing an IBM Storage Area Network*, SG24-5758, located at this website:

  http://www.redbooks.ibm.com/abstracts/sg245758.html?Open

- *Introduction to Storage Area Networks and System Networking*, SG24-5470, located at this website:

  http://www.redbooks.ibm.com/abstracts/sg245470.html?Open

27

FCoE

This chapter provides a high level overview of the Fibre Channel over Ethernet (FCoE) protocol.

The following topics are covered:

- Benefits of a unified infrastructure
- Fibre Channel over Ethernet (FCoE)
- Data center bridging
- Further information

27.1 Benefits of a unified infrastructure

Data centers run multiple parallel networks to accommodate both data and storage traffic. To support these different networks in the data center, administrators deploy separate network infrastructures, including different types of host adapters, connectors and cables, and fabric switches. Use of separate infrastructures increases both capital and operational costs for IT executives. The deployment of a parallel storage network, for example, adds to the overall capital expense in the data center, while the incremental hardware components require additional power and cooling, management, and rack space that negatively impact the operational expense.

Consolidating SAN and LAN in the data center into a unified, integrated infrastructure is referred to as network convergence. A converged network reduces both the overall capital expenditure required for network deployment and the operational expenditure for maintaining the infrastructure.

With recent enhancements to the Ethernet standards, including increased bandwidth (10 GbE) and support for congestion management, bandwidth management across different traffic types, and priority- based flow control, convergence of data center traffic over Ethernet is now a reality. The Ethernet enhancements are collectively referred to as Data Center Bridging (DCB).

27.2 Fibre Channel over Ethernet (FCoE)

Fibre Channel over Ethernet (FCoE) is a protocol designed to seamlessly replace the Fibre Channel physical interface with Ethernet. FCoE protocol specification is designed to fully exploit the enhancements in DCB to support the lossless transport requirement of storage traffic.

FCoE encapsulates the Fibre Channel (FC) frame in an Ethernet packet to enable transporting storage traffic over an Ethernet interface. By transporting the entire FC frame in Ethernet packets, FCoE makes sure that no changes are required to FC protocol mappings, information units, session management, exchange management, services, and so on.

With FCoE technology, servers hosting both host bus adapters (HBAs) and network adapters reduce their adapter count to a smaller number of Converged Network Adapters (CNAs) that support both TCP/IP networking traffic and FC storage area network (SAN) traffic. Combined with native FCoE storage arrays and switches, an end-to-end FCoE solution can now be deployed to exploit all the benefits of a converged network in the data center.

FCoE provides the following compelling benefits to data center administrators and IT executives:

- Compatibility with existing FC deployments protects existing investment and provides a smooth transition path.
- 100% application transparency for both storage and networking applications eliminates the need to recertify applications.
- High performance comparable to the existing Ethernet and FC networks with a road map to increase the bandwidth up to 100Gbps and more is provided.
- Compatibility with existing management frameworks including FC zoning, network access control lists, and virtual SAN and LAN concepts minimizes training of IT staff.

Figure 27-1 shows a converged network enabled by the FCoE technology. Servers use a single CNA for both storage and networking traffic instead of a separate network interface card (NIC) and an FC HBA. The CNA provides connectivity over a single fabric to native FCoE storage and other servers in the network domain. The converged network deployment using FCoE reduces the required components, including host adapters and network switches.

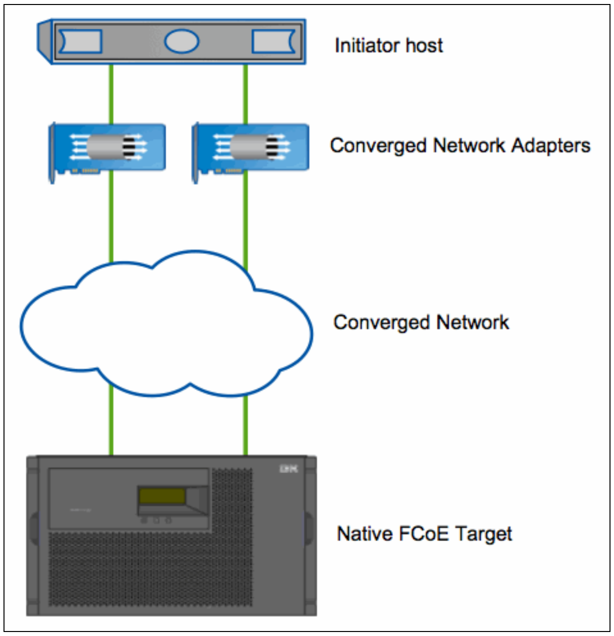


Figure 27-1 Implemented converged network

27.3 Data center bridging

FCoE and converged Ethernet are possible due to enhancements made to the Ethernet protocol, collectively referred to as Data Center Bridging (DCB). DCB enhancements include bandwidth allocation and flow control based on traffic classification and end-to-end congestion notification. Discovery and configuration of DCB capabilities are performed using Data Center Bridging Exchange (DCBX) over LLDP.

Bandwidth allocation is performed with enhanced transmission selection (ETS), which is defined in the IEEE 802.1Qaz standard. Traffic is classified into one of eight groups (0-7) using a field in the Ethernet frame header. Each class is assigned a minimum available bandwidth. If there is competition or oversubscription on a link, each traffic class will get at least its configured amount of bandwidth. If there is no contention on the link, any class can use more or less than it is assigned.

Priority-based flow control (PFC) provides link-level flow control that operates on a per-priority basis. It is similar to 802.3x PAUSE, except that it can pause an individual traffic class. It provides a network with no loss due to congestion for those traffic classes that use PFC. Not all traffic needs PFC. Normal TCP traffic provides its own flow control mechanisms based on window sizes. Because the Fibre Channel protocol expects a lossless medium, FCoE has no built-in flow control and requires PFC to give it a lossless link layer. PFC is defined in the 802.1Qbb standard.

ETS and PFC values are generally configured on the DCB-capable switch and pushed out to the end nodes. For ETS, the sending port controls the bandwidth allocation for that segment of the link (initiator to switch, switch to switch, or switch to target). With PFC, the receiving port sends the per-priority pause, and the sending port reacts by not sending traffic for that traffic class out of the port that received the pause.

Congestion notification (CN) will work with PFC to provide a method for identifying congestion and notifying the source of the traffic flow (not just the sending port). The source of the traffic could then scale back sending traffic going over the congested links. This was developed under 802.1Qau, but is not yet implemented in production hardware.

Fibre Channel over Ethernet (FCoE) is a SAN transport protocol that allows FC frames to be encapsulated and sent over a DCB capable Ethernet network. For this to be possible, the Ethernet network must meet certain criteria; specifically, it must support DCB.

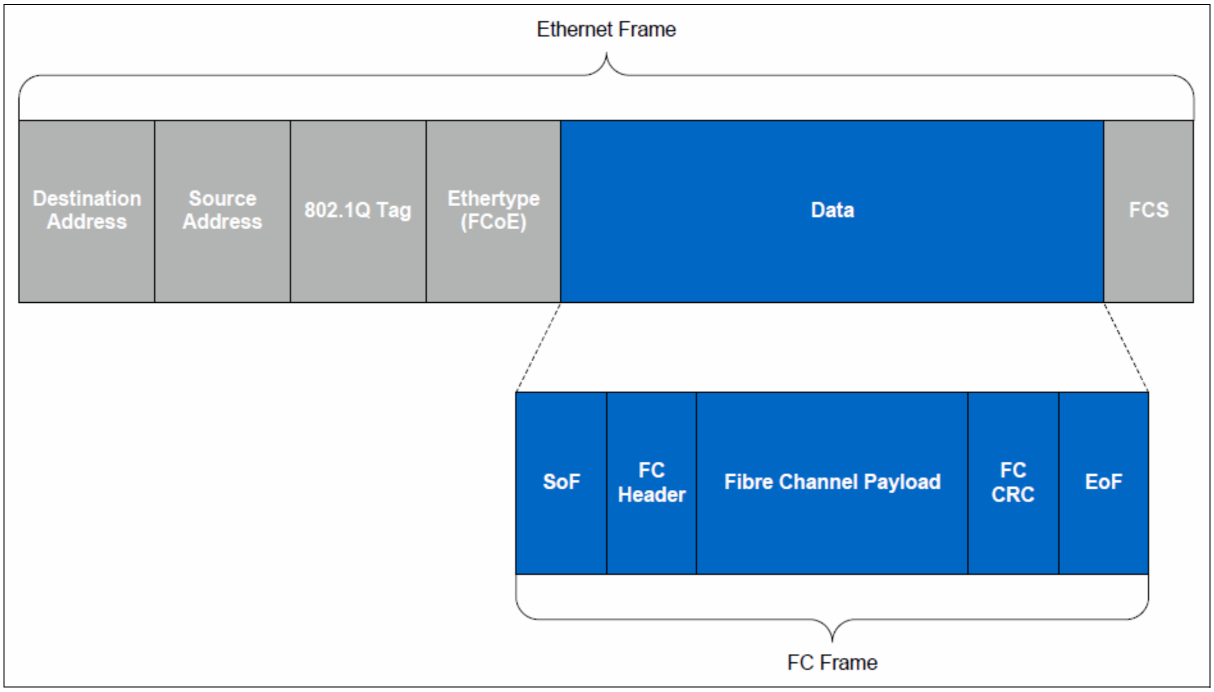


Figure 27-2 FCoE sample frame

Because the FC frames are transported with the FC header all encapsulated in the Ethernet frame (see Figure 27-2), movement of data between an Ethernet network and traditional Fibre Channel fabric is simple. Also, because the FC frames are being transported over Ethernet, the nodes and switches do not have to be directly connected. In fact, the FCoE standard was written to account for one or more DCB-capable switches to be in place between a node and an FCoE switch. Both of these points provide a great amount of flexibility in designing an FCoE storage solution.

A Fibre Channel frame can be up to 2,148 bytes. including the header. Consider that a standard Ethernet frame has only 1,500 bytes available for data, and it is obvious that a larger frame is needed. Luckily Ethernet frame sizes greater than 1,500 bytes have been available on many networking devices for some time now to improve performance of high-bandwidth links. For FCoE, jumbo frames are required, and all FCoE devices must support *baby jumbo* frames of 2,240 bytes. That is the maximum FC frame size plus related Ethernet overhead.

Because traditional Fibre Channel expects a highly reliable transport, the protocol does not have any built-in flow control mechanisms. In traditional FC, the transport layer with buffer-to-buffer credits handles flow control. TCP/IP traffic assumes an unreliable transport and utilizes TCP's adjustable window size and allows retransmits to make sure that all data is transferred. Therefore, a means of making sure of the reliable transport of all FCoE frames had to be established.

Ethernet does have 802.3X PAUSE flow control (defined in 802.3 Annex 31B), but it acts on all traffic coming in on the link. The lack of granularity prevents it from being suitable for a converged network of FCoE and other traffic. The DCB working group addressed this gap with the enhancements described in the DCB section.

The general process by which FCoE is initialized is called FCoE Initialization Protocol (FIP). Before going into the process, we first go over FCoE-specific terms:

- Converged network adapter (CNA): A unified adapter that acts as both an FCoE initiator and a standard network adapter.
- Node: A Fibre Channel initiator or target that is able to transmit FCoE frames.
- Node MAC address: The Ethernet MAC address used by the ENode for FIP.
- FCoE forwarder (FCF): A Fibre Channel switch that is able to process FCoE frames.
- FCoE: Fibre Channel over Ethernet.
- FIP: FCoE Initialization Protocol.
- Fabric-provided MAC address (FPMA): FPMA or SPMA is the FIP MAC address of the ENode.
- Unified target adapter (UTA): An adapter used in a N series storage array that provides FCoE target ports and standard network ports.
- Virtual E_Port (VE_Port): Used to connect two FCFs using FCoE.
- Virtual F_Port (VF_Port): The port on an FCF to which a VN_Port connects.
- Virtual N_Port (VN_Port): The port on an end node used for FCoE communication.

When a node (target or initiator) first connects to an FCoE network, it does so using its ENode MAC address. It is the MAC address associated with its physical, lossless Ethernet port. The first step is DCB negotiation. After the ETS, PFC, and other parameters are configured, the ENode sends a FIP VLAN request to a special MAC address that goes to all FCFs. Available FCFs respond indicating the VLANs on which FCoE services are provided.

Now that the ENode knows which VLAN to use, it sends a discovery solicitation to the same ALL-FCF-MACS address to obtain a list of available FCFs and whether those FCFs support FPMA. FCFs respond to discovery solicitations, and they also send out discovery advertisements periodically.

The final stage of FIP is for the ENode to log into an FCF (FLOGI). During this process, the ENode is assigned a FIP MAC address. It is the MAC address that will be used for all traffic carrying Fibre Channel payloads. The address is assigned by the FCF (FPMA).

27.4 Further information

More details on converged networking and the FCoE protocol can be found in the Redbooks publication, *Storage and Network Convergence Using FCoE and iSCSI,* SG24-7986, which is located at the following website:

http://www.redbooks.ibm.com/abstracts/sg247986.html?Open

28

iSCSI

This chapter provides a high level overview of the iSCSI protocol.

The following topics are covered:

- What iSCSI is
- How iSCSI nodes are identified
- How the storage system checks initiator node names
- Default port for iSCSI
- What target portal groups are
- What iSNS is
- What CHAP authentication is
- How iSCSI communication sessions work
- How iSCSI works with HA pairs
- Further information

28.1 What iSCSI is

The iSCSI protocol is a licensed service on the storage system that enables you to transfer block data to hosts using the SCSI protocol over TCP/IP. The iSCSI protocol standard is defined by RFC 3720.

In an iSCSI network, storage systems are targets that have storage target devices, which are referred to as LUNs (logical units). A host with an iSCSI host bus adapter (HBA), or running iSCSI initiator software, uses the iSCSI protocol to access LUNs on a storage system. The iSCSI protocol is implemented over the storage system's standard gigabit Ethernet interfaces using a software driver.

The connection between the initiator and target uses a standard TCP/IP network. No special network configuration is needed to support iSCSI traffic. The network can be a dedicated TCP/IP network, or it can be your regular public network. The storage system listens for iSCSI connections on TCP port 3260.

In an iSCSI network, there are two types of nodes: targets and initiators. Targets are storage systems, and initiators are hosts. Switches, routers, and ports are TCP/IP devices only, and are not iSCSI nodes.

Storage systems and hosts can be direct-attached through FC or connected through a TCP/IP network.

iSCSI can be implemented on the host using hardware or software. You can implement iSCSI in one of the following ways:

- Initiator software that uses the host's standard Ethernet interfaces.
- An iSCSI host bus adapter (HBA): An iSCSI HBA appears to the host operating system as a SCSI disk adapter with local disks.
- TCP Offload Engine (TOE) adapter that offloads TCP/IP processing. The iSCSI protocol processing is still performed by host software.

You can implement iSCSI on the storage system using software solutions.

Target nodes can connect to the network in the following ways:

- Over the system's Ethernet interfaces using software that is integrated into Data ONTAP. iSCSI can be implemented over multiple system interfaces, and an interface used for iSCSI can also transmit traffic for other protocols, such as CIFS and NFS.
- On the 20xx, 30xx, and 60xx systems, using an iSCSI target expansion adapter, to which some of the iSCSI protocol processing is offloaded. You can implement both hardware-based and software-based methods on the same system.
- Using a unified target adapter (UTA).

28.2 How iSCSI nodes are identified

Every iSCSI node must have a node name.

The two formats, or type designators, for iSCSI node names are *iqn* and *eui.* The storage system always uses the iqn-type designator. The initiator can use either the iqn-type or eui-type designator.

28.2.1 The iqn-type designator

The iqn-type designator is a logical name that is not linked to an IP address. It is based on the following components:

- The type designator, such as iqn
- A node name, which can contain alphabetic characters (a to z), numbers (0 to 9), and three special characters:
  - Period (“.”)
  - Hyphen (“-”)
  - Colon (“:”)
- The date when the naming authority acquired the domain name, followed by a period
- The name of the naming authority, optionally followed by a colon (:)
- A unique device name

> **Tip:** Some initiators might provide variations on the preceding format. Also, even though some hosts do support underscores in the host name, they are not supported on N series systems. For detailed information about the default initiator-supplied node name, see the documentation provided with your iSCSI Host Utilities.

An example format is given in Example 28-1.

Example 28-1 The iSCSI format

```
iqn.yyyymm.backward naming authority:unique device name

yyyy-mm is the month and year in which the naming authority acquired the domain
name.
backward naming authority is the reverse domain name of the entity responsible for
naming this device. An example reverse domain name is com.microsoft.
unique-device-name is a free-format unique name for this device assigned by the
naming authority.

The following example shows the iSCSI node name for an initiator that is an
application server: iqn.1991-05.com.microsoft:example
```

28.2.2 Storage system node name

Each storage system has a default node name based on a reverse domain name and the serial number of the storage system's non-volatile RAM (NVRAM) card.

The node name is displayed in the following format:

```
iqn.1992-08.com.ibm:sn.serial-number
```

The following example shows the default node name for a storage system with the serial number 12345678:

```
iqn.1992-08.com.ibm:sn.12345678
```

28.2.3 The eui-type designator

The eui-type designator is based on the type designator, eui, followed by a period, followed by sixteen hexadecimal digits.

A format example is as follows: eui.0123456789abcdef

28.3 How the storage system checks initiator node names

The storage system checks the format of the initiator node name at session login time. If the initiator node name does not comply with storage system node name requirements, the storage system rejects the session.

28.4 Default port for iSCSI

The iSCSI protocol is configured in Data ONTAP to use TCP port number 3260.

Data ONTAP does not support changing the port number for iSCSI. Port number 3260 is registered as part of the iSCSI specification and cannot be used by any other application or service.

28.5 What target portal groups are

A target portal group is a set of network portals within an iSCSI node over which an iSCSI session is conducted.

In a target, a network portal is identified by its IP address and listening TCP port. For storage systems, each network interface can have one or more IP addresses and therefore one or more network portals. A network interface can be an Ethernet port, virtual local area network (VLAN), or interface group.

The assignment of target portals to portal groups is important for two reasons:

- The iSCSI protocol allows only one session between a specific iSCSI initiator port and a single portal group on the target.
- All connections within an iSCSI session must use target portals that belong to the same portal group.

By default, Data ONTAP maps each Ethernet interface on the storage system to its own default portal group. You can create new portal groups that contain multiple interfaces.

You can have only one session between an initiator and target using a given portal group. To support some multipath I/O (MPIO) solutions, you need to have separate portal groups for each path. Other initiators, including the Microsoft iSCSI initiator version 2.0, support MPIO to a single target portal group by using different initiator session IDs (ISIDs) with a single initiator node name.

Tip: Although this configuration is supported, it is not advised for N series storage systems. For more information, see the technical report on iSCSI multipathing.

28.6 What iSNS is

The Internet Storage Name Service (iSNS) is a protocol that enables automated discovery and management of iSCSI devices on a TCP/IP storage network. An iSNS server maintains information about active iSCSI devices on the network, including their IP addresses, iSCSI node names, and portal groups.

You can obtain an iSNS server from a third-party vendor. If you have an iSNS server on your network, and it is configured and enabled for use by both the initiator and the storage system, the storage system automatically registers its IP address, node name, and portal groups with the iSNS server when the iSNS service is started. The iSCSI initiator can query the iSNS server to discover the storage system as a target device.

If you do not have an iSNS server on your network, you must manually configure each target to be visible to the host.

Currently available iSNS servers support different versions of the iSNS specification. Depending on which iSNS server you are using, you might have to set a configuration parameter in the storage system.

28.7 What CHAP authentication is

The Challenge Handshake Authentication Protocol (CHAP) enables authenticated communication between iSCSI initiators and targets. When you use CHAP authentication, you define CHAP user names and passwords on both the initiator and the storage system.

During the initial stage of an iSCSI session, the initiator sends a login request to the storage system to begin the session. The login request includes the initiator's CHAP user name and CHAP algorithm. The storage system responds with a CHAP challenge. The initiator provides a CHAP response. The storage system verifies the response and authenticates the initiator. The CHAP password is used to compute the response.

28.8 How iSCSI communication sessions work

During an iSCSI session, the initiator and the target communicate over their standard Ethernet interfaces, unless the host has an iSCSI HBA or a CNA.

The storage system appears as a single iSCSI target node with one iSCSI node name. For storage systems with a MultiStore license enabled, each vFiler unit is a target with a different iSCSI node name.

On the storage system, the interface can be an Ethernet port, interface group, UTA, or a virtual LAN (VLAN) interface.

Each interface on the target belongs to its own portal group by default. It enables an initiator port to conduct simultaneous iSCSI sessions on the target, with one session for each portal group. The storage system supports up to 1,024 simultaneous sessions, depending on its memory capacity. To determine whether your host's initiator software or HBA can have multiple sessions with one storage system, see your host OS or initiator documentation.

You can change the assignment of target portals to portal groups as needed to support multi-connection sessions, multiple sessions, and multipath I/O.

Each session has an Initiator Session ID (ISID), a number that is determined by the initiator.

28.9 How iSCSI works with HA pairs

HA pairs provide high availability because one system in the HA pair can take over if its partner fails. During failover, the working system assumes the IP addresses of the failed partner and can continue to support iSCSI LUNs.

The two systems in the HA pair must have identical networking hardware with equivalent network configurations. The target portal group tags associated with each networking interface must be the same on both systems in the configuration. This ensures that the hosts see the same IP addresses and target portal group tags whether connected to the original storage system or connected to the partner during failover.

28.10 Further information

More details on the iSCSI protocol can be found in the Redbooks publication, *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240, which is located at the following website:

`http://www.redbooks.ibm.com/abstracts/sg246240.html?Open`

29

Other protocols

This chapter describes other protocols that can be used with N series systems. Being a Unifies Storage solution, the N series provides more by far than CIFS, NFS, FCP, and iSCSI access.

We cover the following protocols:

- File Transfer Protocol (FTP)
- Secure File Transfer Protocol (SFTP)
- File Transfer Protocol over SSL (FTPS)
- Hypertext Transfer Protocol (HTTP)
- WebDAV

29.1 File Transfer Protocol (FTP)

N series systems do support File Transfer Protocol (FTP) and Trivial File Transfer Protocol (TFTP) protocols to access data on the system. Transport Layer Security (TLS), File Transfer Protocol over SSH (SFTP), and File Transfer Protocol over SSL (FTPS) were added to the security feature set in Data ONTAP 8.1. Note that these features are not available in Data ONTAP 8.0.x.

In Data ONTAP 8.1, the maximum number of concurrent FTP connections are 5,000. The maximum FTP Command Log size and FTP Transfer Log size is 4 gigabytes and maximum FTP Logs is 100.

To specify the maximum number of FTP connections that the FTP server allows, you can use the **ftpd.max_connections** option. By default, the maximum number of FTP connections is 500. Example 29-1 shows how you can set it.

Example 29-1 Maximum number of FTP connections

```
options ftpd.max_connections n
```

Here, *n* is the maximum number of FTP connections that the FTP server allows.

If you set the **ftpd.max_connections** option to a value that is less than the current number of FTP connections, the FTP server refuses new connections until the number falls below the new maximum. The FTP server does not interrupt existing FTP connections. In an HA configuration, the maximum number of FTP connections doubles automatically when the storage system is in takeover mode.

29.1.1 Enabling and disabling FTP

On storage systems shipped with Data ONTAP 8.0 or later, secure protocols are enabled and non-secure protocols are disabled by default. SecureAdmin is set up automatically on storage systems shipped with Data ONTAP 8.0 or later. The default security settings for these systems are as follows:

- Secure protocols (including SSH, SSL, and HTTPS) are enabled by default.
- Non-secure protocols (including RSH, Telnet, FTP, and HTTP) are disabled by default.

Therefore, the options shown in Table 29-1 are set by default on your N series system.

Table 29-1 FTP options

Service	Setting
File Transfer Protocol (FTP)	off
File Transfer Protocol over SSH (SFTP)	off
File Transfer Protocol over SSL (FTPS)	off
Trivial File Transfer Protocol (TFTP	off

You can enable or disable the FTP server by modifying the ftpd.enable option (Example 29-2). It allows clients to access files using FTP.

Example 29-2 Enable FTP

```
options ftpd.enable on
```

The FTP server begins listening for FTP requests on standard FTP port 21.

To disable the FTP server, see Example 29-3.

Example 29-3 Disable FTP

```
options ftpd.enable off
```

You can access the messages log files using your NFS or CIFS client, or using FilerView /OnCommand or N series System Manager.

The FTP command audit log is located in /etc/log/ftp.cmd, the FTP transfer log is located in /etc/log/ftp.xfer.

29.1.2 Blocking and protecting data and access

This section describes options available to manage FTP access.

FTP file locking

To prevent users from modifying files while the FTP server is transferring them, you can enable FTP file locking. Otherwise, you can disable FTP file locking. By default, FTP file locking is disabled.

Example 29-4 shows how to enable FTP file locking for deleting or renaming.

Example 29-4 Enable FTP file locking for deleting or renaming

```
options ftpd.locking delete
```

Example 29-5 shows how to enable FTP for deleting, renaming, and writing.

Example 29-5 Enable FTP file locking for deleting, renaming, or writing

```
options ftpd.locking write
```

Example 29-6 shows how to disable FTP file locking.

Example 29-6 Disable FTP file locking

```
options ftpd.locking none
```

Restricting FTP access

You can restrict FTP access by blocking FTP users and restricting FTP users to a specific directory (either their home directories or a default directory).

Blocking specific FTP users

To prevent specific FTP users from accessing the storage system, you can add them to the /etc/ftpusers file.

Blocking data

The interface.blocked.ftpd option allows you to block or unblock a protocol on an interface. See Example 29-7 for details.

Example 29-7 Blocking an interface for FTP

```
ftpd e0f options interface.blocked.ftpd e0f
```

Data ONTAP provides a firewall for protocols (CIFS, NFS, ftp, ndmp, iSCSI, and SnapMirror). Protocol blocking can be seen as a protocol firewall. Some examples for protocol blocking are shown in (Example 29-8).

Example 29-8 Various examples to block access

```
itsosj_n1> options interface.blocked.cifs e4c
itsosj_n1> options interface.blocked.nfs e1a,e1b
itsosj_n1> options interface.blocked.iscsi e5b
itsos_n1i> options interface.blocked.ftpd e4c,e1a
```

To configure the FTP server to use UNIX, Windows, or both authentication styles, you can set the **ftpd.auth_style option** to **unix, ntlm**, or **mixed**, respectively. By default, this option is **mixed**.

29.2 Secure File Transfer Protocol (SFTP)

The Secure File Transfer Protocol (SFTP) is a secure replacement for the File Transfer Protocol (FTP). SFTP is based on the Secure Shell protocol. Similar to FTP, SFTP is an interactive file transfer program that performs all operations over an encrypted SSH transport. Unlike FTP, SFTP encrypts both commands and data, providing effective protection against common network security risks.

The SSH client and server provide both command-line SFTP tools and a graphical user interface for Windows users. SFTP encrypts the session, preventing the casual detection of your user name, password, or anything you have transmitted. This protocol assumes that it runs over a secure channel, that the server has already authenticated the user at the client end, and that the identity of the client user is externally available to the server implementation. SFTP runs from the SSH Connection Protocol as a subsystem.

29.2.1 Limitations of Data ONTAP support for SFTP

There are some limitations of Data ONTAP support for SFTP. Because WAFL can write a maximum of 64 KB of data in one I/O operation, SFTP packet size is limited to 64 KB as well. If Data ONTAP receives a packet larger than 64 KB over SFTP, it closes the session and resets the connection. If Data ONTAP receives a read request for more than 64 KB of data, it reads only 64 KB of data.

29.2.2 Enabling or disabling SFTP

To enable or disable SFTP, you can set the **`sftp.enable`** option to on or off, respectively. It allows clients to access files using SFTP. By default, this option is off.

Before you can enable SFTP, you must set up and start SSH using the **`secureadmin`** command.

Example 29-9 shows how to enable SFTP.

Example 29-9 Enable SFTP

```
options sftp.enable on
```

Example 29-10 shows how to disable SFTP.

Example 29-10 Disable SFTP

```
options sftp.enable off
```

29.2.3 Enabling or disabling SFTP file locking

To prevent users from modifying files while the SFTP server is transferring them, you can enable SFTP file locking. By default, SFTP file locking is disabled.

If you want SFTP file locking to be enabled so that files being retrieved cannot be deleted or renamed, see Example 29-11 for details.

Example 29-11 SFTP locking enabled for retrieving but no deletion or renaming

```
options sftp.locking delete
```

In order to enable so that files being retrieved cannot be opened for writing or deletion, see Example 29-12.

Example 29-12 SFTP locking set to retrieve but no writing or deletion

```
options sftp.locking write
```

To disable SFTP file locking, see Example 29-13.

Example 29-13 Disable SFTP file locking

```
options sftp.locking none
```

29.3 File Transfer Protocol over SSL (FTPS)

You can manage FTP over SSL (FTPS) by enabling or disabling implicit FTPS, enabling or disabling explicit FTPS, and specifying whether explicit FTPS connections can be opened in secure mode.

29.3.1 Differences between implicit and explicit FTPS

FTP over SSL (FTPS) allows FTP software to perform secure file transfers.

Typically, data sent over an FTP connection, whether over a control connection or a data connection, is sent in clear text and without any freshness or integrity guarantees. FTPS provides an extension to the FTP protocol that allows FTP software to perform secure file transfers over an implicit FTPS connection or an explicit FTPS connection.

29.3.2 Implicit FTPS

Data ONTAP provides an industry-standard implementation of implicit FTPS; there is no corresponding RFC. With implicit FTPS, security is achieved by encrypting and decrypting data in the transport layer by SSL. In particular, implicit FTPS works as follows:

- Data ONTAP listens on port 990.
- The FTPS client connects to port 990.

  An SSL handshake is initiated on connection. If the handshake fails, no further communication is allowed.
- After the completion of a successful SSL handshake, all further FTP communication goes through SSL and is secure.
- The command channel cannot be restored back to clear text.

  New packets that some clients might require are PBSZ and PROT. The only argument that Data ONTAP supports for the PROT command is P, meaning private encrypted communications.
- The default port for the data channel is 989.

29.3.3 Explicit FTPS

Data ONTAP implements explicit FTPS in accordance with RFC 2228 and RFC 4217. In particular, explicit FTPS works as follows:

- Data ONTAP listens on port 21 (the standard FTP port).
- The FTP client connects to *port 21* over a normal TCP connection.

  Any communication over the connection is clear text to begin with. The connection can be made secure by issuing the **AUTH** command.
- After receiving the **AUTH** command, Data ONTAP initiates an SSL handshake.

  You can use the **CCC** command to restore the command channel back to clear text.
- Before starting a data connection, the client must issue **PBSZ** and **PROT** commands.

  Without these commands, the data connection would be clear text. The only arguments that Data ONTAP supports for the **PROT** command are **C** and **P**, meaning clear text or private data channels.

- As specified in the **RFC**, the **PBSZ** command must be preceded by a successful authentication data exchange, and the **PROT** command must be preceded by a successful **PBSZ** command.
- The default port for the data channel is *20*.

29.4 Hypertext Transfer Protocol (HTTP)

To let HTTP clients (web browsers) access the files on your storage system, you can enable and configure Data ONTAP's built-in HyperText Transfer Protocol (HTTP) server.

To let HTTP clients (web browsers) access the files on your N series storage system, you can enable and configure Data ONTAP's built-in Hypertext Transfer Protocol (HTTP) server. Alternatively, you can purchase and connect a third-party HTTP server to your storage system.

Depending on the enterprise security structure, the state of any service depends on where the service is deployed and how deep it is in the infrastructure. The following settings can be configured through the options command (see Table 29-2).

Table 29-2 Httpd.admin options

HTTP service	Default setting
FilerView https://<filer_IP>/na_admin (httpd.admin.ssl.enable)	off
FilerView http://<filer_IP>/na_admin (httpd.admin.enable)	off

The default HTTP timeout setting is 300 (**options httpd.timeout 300**). the HTTP log can be found in the following directory:

```
/etc/log/http.log
```

29.4.1 Managing the Data ONTAP HTTP server

Managing the HTTP server that is built into Data ONTAP involves several tasks.

Enabling or disabling the Data ONTAP HTTP server: You can use the **httpd.enable option** to enable or disable the HTTP server that is built into Data ONTAP. By default, this option is **off**. When this option is enabled, web browsers can access all of the files in the HTTP server's root directory.

To enable HTTP on an N series system, use the setting **on** as shown in Example 29-14.

Example 29-14 Enable HTTP

```
options httpd.enable on
```

To disable HTTP on your system, use the setting **off** as shown in Example 29-15.

Example 29-15 Disable HTTP

```
options httpd.enable off
```

29.4.2 Enabling or disabling the bypassing of HTTP traverse checking

You can **enable** or **disable** the bypassing of HTTP traverse checking by setting the **httpd.bypass_traverse_checking** option to **on** or **off**, respectively. By default, this option is set to off.

If the **httpd.bypass_traverse_checking** option is set to **off**, when a user attempts to access a file using the HTTP protocol, Data ONTAP checks the traverse (execute) permission for all directories in the path to the file. If any of the intermediate directories does not have the "**X**" (*traverse permission*), Data ONTAP denies access to the file.

If the **http.bypass_traverse_checking** option is set to **on**, when a user attempts to access a file, Data ONTAP does not check the traverse permission for the intermediate directories when determining whether to grant or deny access to the file.

29.4.3 Specifying the root directory for the HTTP server

The **httpd.rootdir** option can be set to specify the root directory (full path) for the HTTP server that is built into Data ONTAP. It is the directory that contains all of the files that an HTTP client can access. See Example 29-16) for how the root directory can be set.

Example 29-16 Set HTTP root directory

```
options httpd.rootdir /vol0/home/HTTP_root_n1
```

29.4.4 Testing the HTTP server

In order to confirm that the HTTP server is working, you can copy an HTML file into the HTTP root directory and then access the file from a web browser. You can also access the HTTP server's root directory (or a subdirectory of the HTTP server's root directory) directly from a web browser.

First copy an HTML file into the HTTP server's root directory (*httpd.rootdir).* Then access this file from a web browser (running on a separate system).

The URL is *http://www.hostname.com/myfile.html,* where *hostname* is the host name of the storage system and *myfile.html* is the name of the file you copied into the HTTP server's root directory. As a result, you must be able to see the contents of the file in the web browser.

If you do not specify a certain filename, such as **http://www.hostname.com**, without specifying the filename, such as *myfile.html*, the HTTP server looks for the following files in the following order in the directory that you specify:

1. index.html
2. default.htm
3. index.htm
4. default.html

If none of these files exists, the storage system automatically generates an HTML version of the directory listing for that directory (if the **httpd.autoindex.enable option** is **on**) or responds with the "403" (forbidden) error code (if the **httpd.autoindex.enable option** is **off**).

29.5 WebDAV

To let users use WebDAV interoperable, collaborative applications, you can add the WebDAV Web-based Distributed Authoring and Versioning) protocol to your existing HTTP service. Alternatively, you can purchase and connect a third-party WebDAV server to your storage system.

> **Tip:** You can use the WebDAV protocol on your storage system as an extension of HTTP only if you purchased the license for HTTP. Future versions of Data ONTAP might require the use of a WebDAV license key in order to use WebDAV with HTTP.

The WebDAV protocol defines the HTTP extensions that enable distributed Web authoring tools to be broadly interoperable, while supporting user needs. WebDAV allows you to create HTTP directories.

The WebDAV protocol provides support for remote software development teams though a wide range of collaborative applications. WebDAV benefits from the success of HTTP and acts as a standard access layer for a wide range of storage repositories. HTTP gives read access, WebDAV gives write access.

This protocol includes the following major features:

- Locking: Long-duration exclusive and shared write locks prevent two or more collaborators from writing to the same resource without first merging changes. To achieve robust Internet-scale collaboration, where network connections might be disconnected arbitrarily, and for scalability, because each open connection consumes server resources, the duration of DAV locks is independent of any individual network connection.
- Properties: XML properties provide storage for arbitrary metadata, such as a list of authors on Web resources. These properties can be efficiently set, deleted, and retrieved using the DAV protocol. DASL (DAV Searching and Locating) protocol provides searches of Web resources based on the values in XML properties.
- Namespace manipulation: Because resources sometimes need to be copied or moved as the Web evolves, DAV supports copy and move operations. Collections, similar to file system directories, can be created and listed.
- HTTP feature support: Data ONTAP WebDAV implementation supports your HTTP configuration settings, such as redirect rules, authentication, and access restrictions. To use WebDAV, you need to have HTTP service enabled and configured.
- CIFS feature support: Data ONTAP WebDAV implementation supports CIFS home directories when you have valid CIFS and HTTP licenses, and you have enabled WebDAV.

Managing the WebDAV server that is built into Data ONTAP includes tasks of enabling or disabling the WebDAV protocol and pointing a WebDAV client to a home directory.

You can set the **`webdav.enable option`** to **`on`** or **`off`**, respectively, to enable or disable the WebDAV server that is built into Data ONTAP. By default, this option is on.

Part 5

Application and host OS integration

N series storage systems provide all the capabilities to protect your critical data and are closely integrated with N series hardware and Data ONTAP.

In this part of the book, we describe storage solutions that outline how the N series is the heart of your datacenter with closely integrated applications and host platforms. The following topics are covered:

- SnapDrive
- SnapManager
- Snap Creator
- VMware vSphere
- Virtual Storage Console 4.1
- Consistency groups

30

SnapDrive

This chapter provides an overview of SnapDrive, which provides storage virtualization of IBM System Storage N series volumes and Snapshot backup and restore operations over iSCSI or Fibre Channel (FCP) transport protocols.

The following topics are covered:

- Challenges
- SnapDrive overview
- SnapDrive integration with the host operating system
- Snapshots using SnapDrive
- SnapDrive for Windows
- SnapDrive for UNIX
- Flexible networked storage
- Summary

30.1 Challenges

Today's enterprise IT administrator is challenged to provision storage quickly to support applications for new business initiatives, all on a minimal budget. For most IT administrators, it does not stop at this point. They also must protect the application against data corruption, disasters, and attacks with the help of well-planned backup mechanisms. Even more importantly, the backup process must not disrupt the service or the performance of the service.

To address these issues, administrators maintain their own set of scripts that are executed periodically or manually to speed up and automate the tasks of storage provisioning and backing up. But this situation comes with an additional burden of maintaining the scripts. For example, if the host operating system is updated to a newer version, the output of the command used in the script or command syntax might change, resulting in a complete rewrite of the scripts.

Therefore, IT and host administrators require a solution that enables them to do these tasks:

- Quickly provision storage and bring it online from the host.
- Adapt to different operating systems and other environmental changes without requiring maintenance of scripts.
- Take backups without any performance degradation to the application.
- Schedule and maintain their own backup policies, depending on the type of application.
- Modify space allocations without taking applications offline.

30.2 SnapDrive overview

SnapDrive software provides a host administrator efficient storage and data management, including the ability to implement quick backup and restore policies for application data.

An IBM System Storage N series storage system provides SnapDrive for Windows and SnapDrive for UNIX solutions to address the issues described in 30.1, "Challenges" on page 438. SnapDrive helps host administrators provision storage and manage it directly from the host. SnapDrive gives flexibility to application administrators by enabling them to define their backup policies and, more importantly, it allows administrators to resize storage on the fly without any disruption of application service.

Both versions of SnapDrive simplify storage and data management by using the host operating system and IBM System Storage N series technologies, hiding the complexity of steps that must be executed on both the storage system and the host system, and removing the dependency on the storage administrator.

Key SnapDrive functionality includes storage provisioning on the host, consistent data Snapshots, and rapid application data recovery from Snapshots. SnapDrive complements the native file system and volume manager technology, and integrates seamlessly with the clustering technology supported by the host operating system to provide high availability of the service to its users (Figure 30-1).

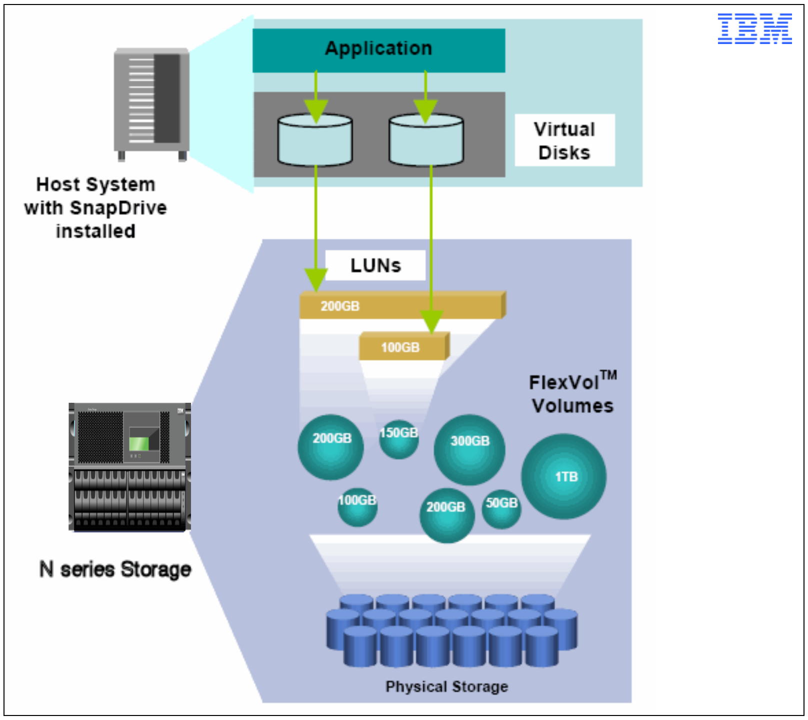


Figure 30-1 Typical SnapDrive deployment

SnapDrive provides a layer of abstraction between an application running on the host operating system and the underlying IBM System Storage N series storage systems. Applications that are running on a server with SnapDrive use virtual disks (or LUNs) on IBM System Storage N series storage systems as though they were locally connected drives or mount points. It allows applications that require locally attached storage, such as DB2, Oracle database, Microsoft Exchange, and Microsoft SQL, to use IBM System Storage N series technologies, including Snapshot, flexible volumes, cloning, and space management technologies.

30.2.1 Components of SnapDrive

SnapDrive includes all the necessary drivers and software to manage interfaces, protocols, storage, and Snapshots. Snapshots are nondisruptive to applications and functions on execution. Snapshot backups can also be mirrored across LAN or WAN links for centralized archiving and disaster recovery.

SnapManager solutions are designed to invisibly use SnapDrive to trigger backups, restores, and mirroring of specific data sets.

30.2.2 Benefits of SnapDrive

Most of today's enterprises use business-critical applications, and their IT and storage management teams face a number of challenges such as these:

- Support new business initiatives with a minimal increase in operating budget.
- Protect data from corruption, disasters, and attacks.
- Back up data without any performance degradation, quickly and consistently, without any errors.

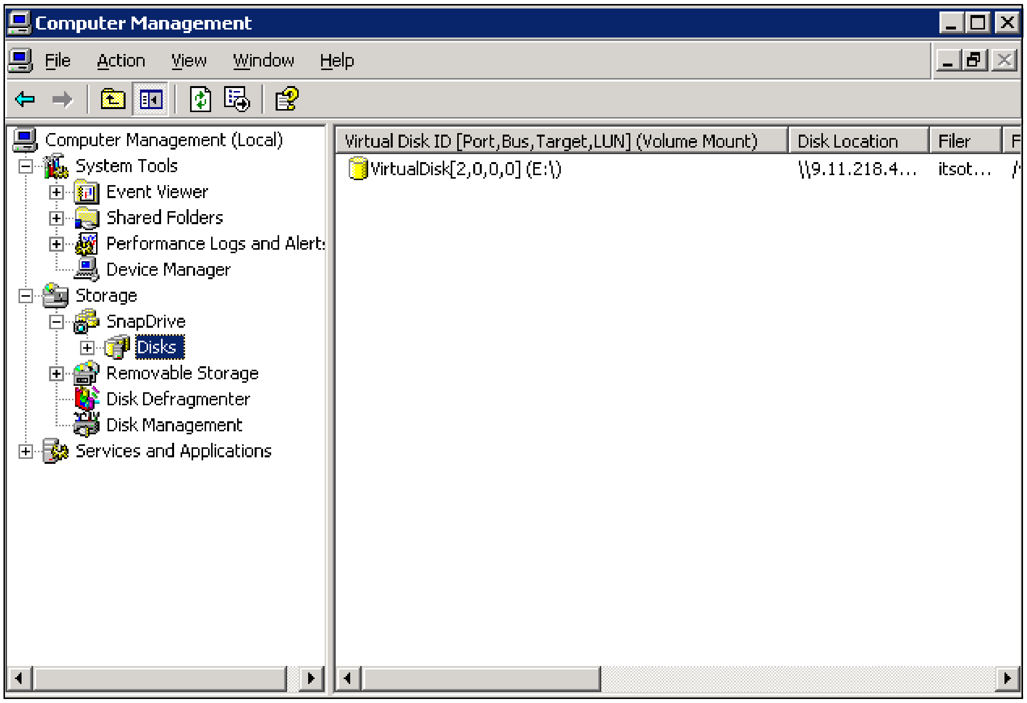
SnapDrive addresses these problems by providing simplified and intuitive storage management and data protection from a host/server perspective. Here we list some important benefits of SnapDrive:

- It allows host and application administrators to quickly create virtual disks within a dynamic pool of storage that can be reallocated, scaled, and enlarged in real time, even while systems are accessing data.
- It allows dynamic on-the-fly file system expansion; new disks are usable within seconds.
- Snapshots provide rapid backup and restore capability with minimal resource and capacity requirements.
- It allows mirroring, data replication, and clustering for high availability.
- It supports multipath technology for high performance.
- It enables connecting to existing Snapshots from the original host or a different host.
- Independent of the underlying storage access media and protocol. SnapDrive supports FCP, iSCSI, and Network File System (NFS) as the transport protocols (NFS supports only Snapshot management).
- It provides a patented, high-performance, and low-latency file system with industry-leading reliability.
- It includes robust yet easy-to-use data and storage management features and software.
- It has industry-leading availability, exceeding 99.99% availability on nonclustered systems.
- It provides robust data integrity features, such as advanced RAID functionality and built-in file system checksums, help protect against disk drive failures and disk errors.

30.3 SnapDrive integration with the host operating system

After being installed, SnapDrive can be used to create and manage virtual disks on IBM System Storage N series storage systems from the host operating system. For Windows-based hosts, these appear as basic disks to Windows Server 2003, Windows Server 2008, and their applications. The virtual disks that reside on the IBM System Storage N series storage system can be expanded, unlike Windows-native basic disks.

SnapDrive is also used to create, delete, and manage all aspects of the application Snapshot backups (Figure 30-2). It allows applications that require locally attached storage to use IBM System Storage N series functionality. It has been estimated that SnapDrive saves 100 steps in the creation of virtual disks.



Virtual Disk ID [Port,Bus,Target,LUN] (Volume Mount)	Disk Location	Filer	F
VirtualDisk[2,0,0,0] (E:\)	\\9.11.218.4...	itsot...	/‹

Figure 30-2 SnapDrive

30.4 Snapshots using SnapDrive

A Snapshot is a frozen, read-only image of a traditional volume, a FlexVol, or an aggregate that captures the state of the file system at a point in time. You use SnapDrive to ensure that you create consistent Snapshots in the event that you need to restore a LUN from that copy.

30.4.1 Consistent Snapshots

Snapshot operations on a single LUN actually make a Snapshot of all the LUNs on the volume. Because a storage system volume can contain LUNs from multiple hosts, the only consistent Snapshots are those of LUNs connected to the host that created the SnapDrive Snapshot. In other words, within a Snapshot, a LUN is not consistent if it is connected to any host other than the one that initiated the Snapshot. (This is why you are advised to dedicate your storage system volumes to individual hosts.) Therefore, it is important to back up a LUN using a SnapDrive Snapshot rather than using other means, such as creating Snapshots from the storage system console.

Snapshots in SAN and NAS environments

Making Snapshots in a SAN environment differs from doing so in an NAS environment in one fundamental way: In a SAN environment, the storage system does not control the state of the file system.

Snapshots are useful only when they can be successfully restored. Snapshots of a single storage system volume that contains all the LUNs on the host file system are always consistent, provided that the file system supports the freeze operation. But if the LUNs on the host file system span different storage system volumes or storage systems, then the copies might not be consistent, unless they are made at exactly the same time across different storage system volumes or storage systems and they can be restored successfully. For UNIX, consistent Snapshots can be created using the Data ONTAP consistency group feature, which is supported beginning with Data ONTAP V7.2.

SnapDrive and SnapManager for Oracle

One of the major deployments for SnapDrive is deploying along with SnapManager for Oracle. SnapManager for Oracle has become one of the most popular tools for making and managing backups for Oracle in an IBM System Storage N series environment. SnapManager for Oracle relies on SnapDrive to execute all backup and restore commands on the storage system.

For more information about the SnapManager for Oracle product, see the following website:

http://www.ibm.com/support/docview.wss?rs=1302&uid=ssg1S7002523

Attention: If you use the SnapManager product to manage your database instead of SnapDrive, you must use SnapManager to create Snapshots.

30.4.2 Preferred practices for Snapshots

For space management when using SnapDrive, use the following preferred practices:

- Disable automatic Snapshot creation on the storage system for the volume on which the LUNs are created.
- Periodically, use the **`snapdrive snap list`** command and delete old Snapshots that can unnecessarily occupy space.
- Set the snap reserve value to 0%, because the automatic Snapshots that are made by the storage system might not capture the LUN in a consistent state and therefore allow use of all the volume space dedicated for the LUNs.
- Use SnapDrive to create and manage all the LUNs on your storage system.
- Place all LUNs connected to the same host on a dedicated volume accessible by just that host.
- To avoid space contentions, do not have LUNs on the same storage system volume as other data (for example, NFS share).
- Unless you can be sure that name resolution publishes only the storage system interface that you intend, configure each network interface by IP address, rather than by name.
- If you use Snapshots, you cannot use the entire space on a storage system volume to store your LUN. The storage system volume hosting the LUN must also include sufficient space to store the Snapshot delta (incremental changes over time). This overhead can vary significantly depending on the volume settings, Snapshot retention schedule, and rate of change, and needs to be carefully planned.
- Do not create any LUNs in `/vol/vol0`. It is a storage system limitation. This volume is used by Data ONTAP to administer the storage system and must not be used to contain any LUNs.

30.4.3 Volume-size rules

SnapDrive uses space on a storage system volume for LUNs and their data, and also for the data that changes between Snapshots, the LUN's active file system, and for metadata.

Storage system volumes that will hold LUNs must be large enough to hold all the LUNs in the volume, as well any Snapshots if Snapshots are created.

The following factors govern the appropriate minimum size for a volume that holds a LUN:

- The volume and LUN settings must be configured as per the N series preferred practices. For example: Fractional Reserve, Volume Autosize, Snapshot Autodelete, and LUN and/or volume thin provisioning. The specific combination of settings will vary depending on your environment and intention.
- The volume must be at least as large as the LUN that it is to contain.
- The volume must also provide enough additional space to hold the number of Snapshots that you intend to keep online.

  The amount of space consumed by a Snapshot depends on the amount of data that changes after the Snapshot is taken. The maximum number of Snapshots is 255 per storage system volume.
- If you are not implementing the automatic capacity and Snapshot management features (that is, not following preferred practice) then we advise that you set the volume size to be twice the LUN size, plus some additional overhead to store the Snapshot delta (such as 20%).

Tip: Although this configuration is common in legacy systems, we advise that you instead use the modern preferred practices, as they significantly reduce capacity overhead.

30.5 SnapDrive for Windows

SnapDrive enables Windows and UNIX applications to access storage resources on IBM System Storage N series storage systems, which are presented to the Windows 2003 (or later) operating system as locally attached disks. IBM System Storage N series storage systems and SnapDrive software represent a complete data management solution for Windows applications.

SnapDrive includes Windows 2003 and later device drivers and software that is used to manage application Snapshot backups. Snapshot backups, shown in Figure 30-3, are nondisruptive to applications and occur quickly. Restoring data from a Snapshot is nearly instantaneous. Snapshot backups can also be mirrored across LAN or WAN links for centralized archiving and disaster recovery purposes. This section provides a brief outline of the architecture of SnapDrive.

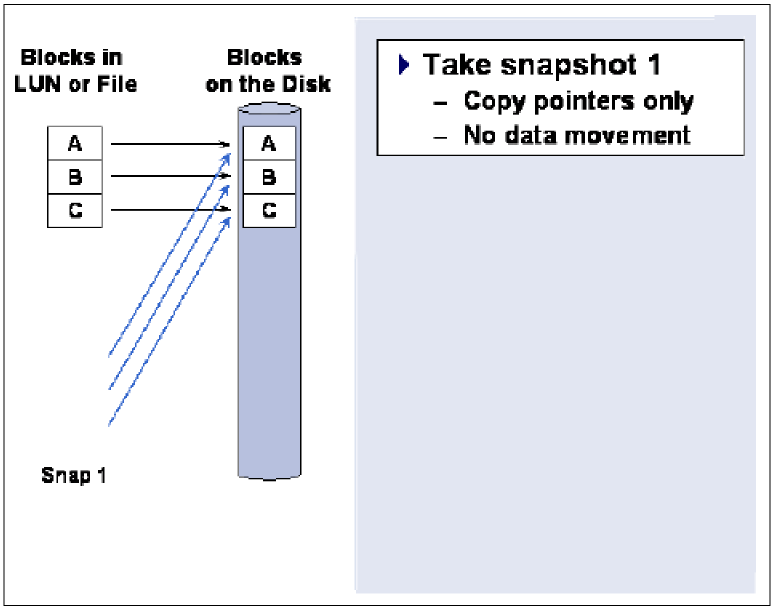


Figure 30-3 Snapshot

30.5.1 SnapDrive software components

SnapDrive software combines IBM System Storage N series functionality, Windows 2003 and later IBM System Storage N series device drivers, and a Microsoft Management Console (MMC) application into a complete data management solution.

The SnapDrive software installed on a Windows 2003 server has the following components:

- SnapDrive Win32 device drivers
- SnapDrive Win32 service
- SnapDrive Microsoft Management Console application

30.5.2 Windows Device Manager

In the SCSI and RAID controllers section of the Windows Device Manager, the Emulex LightPulse PCI Fibre Channel HBA and Microsoft iSCSI Initiator are both visible, as shown in Figure 30-4. SnapDrive is capable of accessing virtual disks over iSCSI and FCP on one or more storage systems simultaneously.

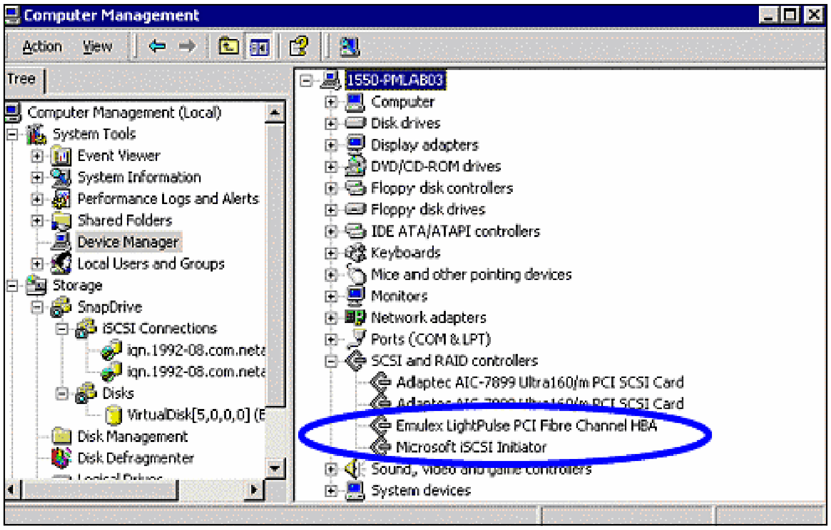
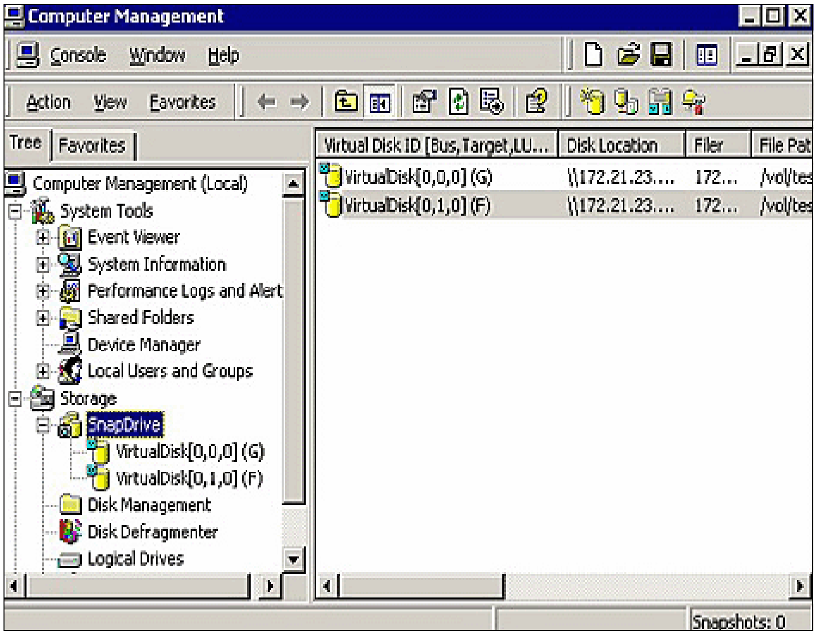


Figure 30-4 Computer management

Figure 30-5 shows the SnapDrive MMC management interface, which is used to manage virtual disks. In this example, there are two LUNs in use by this server.

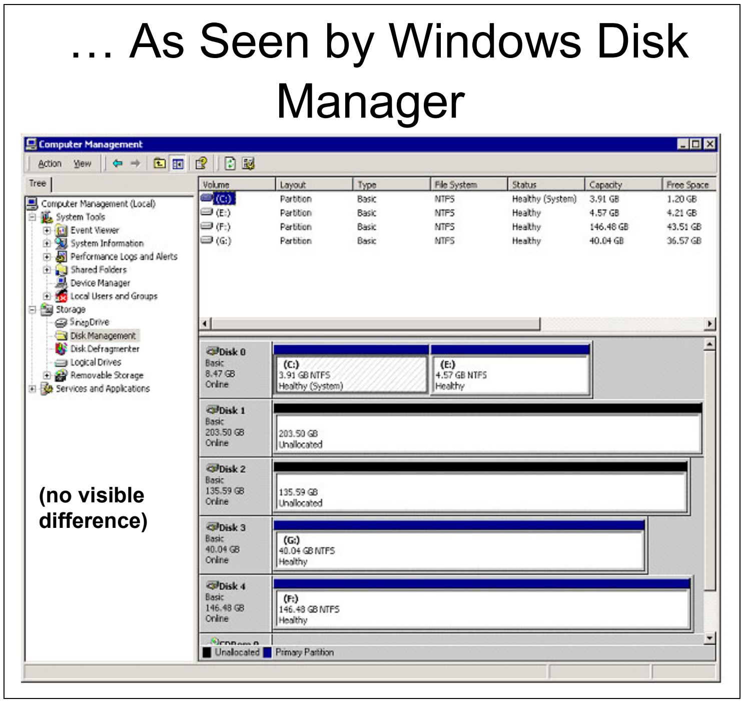


Virtual Disk ID [Bus,Target,LU...	Disk Location	Filer	File Pat
VirtualDisk[0,0,0] (G)	\\172.21.23....	172...	/vol/tes
VirtualDisk[0,1,0] (F)	\\172.21.23....	172...	/vol/tes

Figure 30-5 SnapDrive

After it is installed, SnapDrive can be used to create and manage virtual disks on IBM System Storage N series storage systems, which appear as basic disks to the Windows 2003 server and its applications. Virtual disks that reside on an IBM System Storage N series storage system can be expanded, unlike Windows-native basic disks. SnapDrive is also used to create, delete, and manage all aspects of the application Snapshot backups.

After a SnapDrive virtual disk is created, it appears in the Microsoft Disk manager, as shown in Figure 30-6.



… As Seen by Windows Disk Manager

(no visible difference)

Volume	Layout	Type	File System	Status	Capacity	Free Space
(C:)	Partition	Basic	NTFS	Healthy (System)	3.91 GB	1.20 GB
(E:)	Partition	Basic	NTFS	Healthy	4.57 GB	4.21 GB
(F:)	Partition	Basic	NTFS	Healthy	146.48 GB	43.51 GB
(G:)	Partition	Basic	NTFS	Healthy	40.04 GB	36.57 GB

Disk 0 Basic 8.47 GB Online	(C:) 3.91 GB NTFS Healthy (System)	(E:) 4.57 GB NTFS Healthy
Disk 1 Basic 203.50 GB Online	203.50 GB Unallocated	
Disk 2 Basic 135.59 GB Online	135.59 GB Unallocated	
Disk 3 Basic 40.04 GB Online	(G:) 40.04 GB NTFS Healthy	
Disk 4 Basic 146.48 GB Online	(F:) 146.48 GB NTFS Healthy	

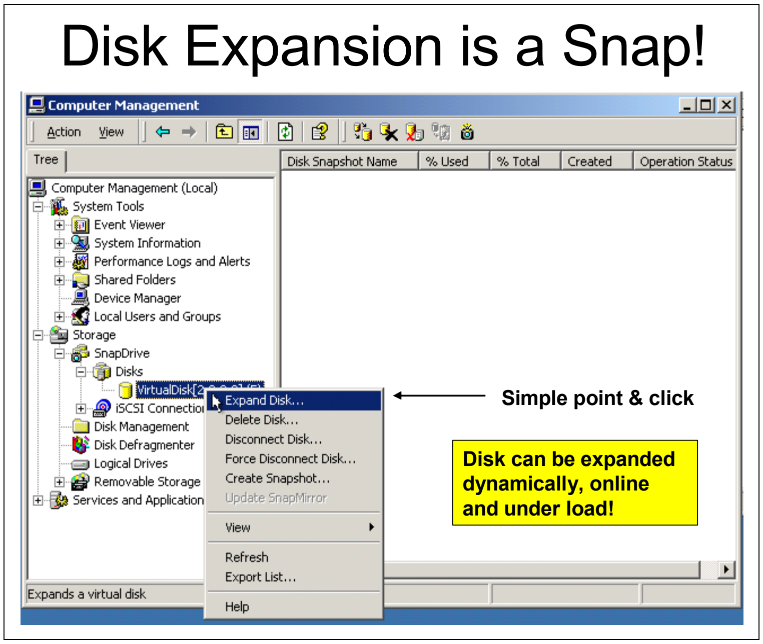
Figure 30-6 Microsoft Disk Manager

30.5.3 Dynamic file system expansion

As your storage needs increase, you might need to expand a virtual disk to hold more data. A good opportunity for doing this task is right after you have expanded your IBM System Storage N series volumes.

Planned downtime is also minimized with online disk expansion. Scheduled maintenance is minimized. Dynamic file system expansion satisfies those unique occurrences when unplanned growth or data movement is required and capacity must be increased on the fly. The following figures show how easily It can be done.

Figure 30-7 shows disk expansion.



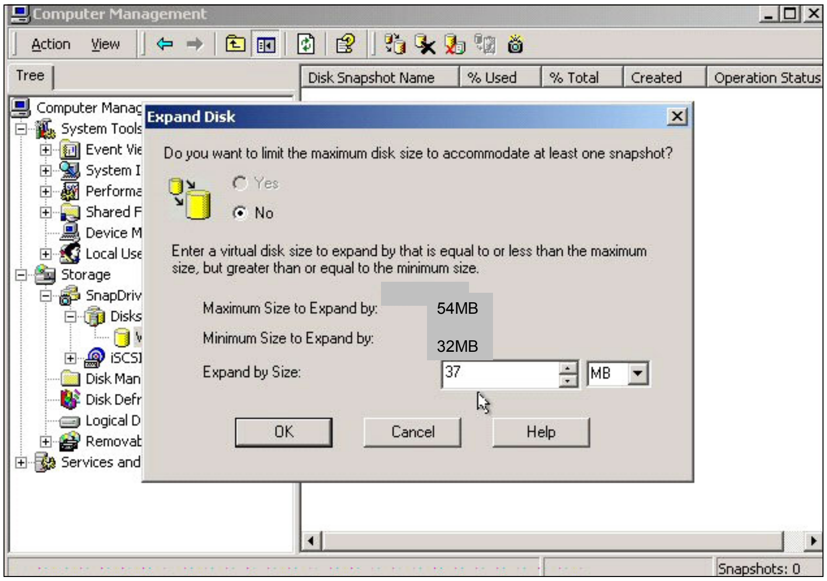
Disk Expansion is a Snap!

Simple point & click

Disk can be expanded dynamically, online and under load!

Figure 30-7 Disk expansion

Figure 30-8 shows dynamic expansion.



Expand Disk

Do you want to limit the maximum disk size to accommodate at least one snapshot?

○ Yes
● No

Enter a virtual disk size to expand by that is equal to or less than the maximum size, but greater than or equal to the minimum size.

Maximum Size to Expand by: 54MB

Minimum Size to Expand by: 32MB

Expand by Size: 37 MB

Figure 30-8 Dynamic expansion

30.5.4 Volumes, RAID groups, and virtual disks

Physical disks are grouped together in the form of volumes on an IBM System Storage N series storage system and can consist of one or more RAID groups.

Virtual disks are created and managed within the limits of an IBM System Storage N series storage system available storage capacity on a per-volume basis. The size of a volume is determined by the number of disks multiplied by the capacity of the disks.

The disks that make up a single volume can be divided into multiple RAID groups. Each RAID group calculates parity information for the drives within the RAID group. The use of multiple RAID groups is transparent to data access and only important at the RAID layer. Volumes function as a whole regardless of the number of RAID groups.

Application programs that run on a Windows server (such as a database application) access virtual disks as though they were locally attached physical disks. Virtual disks are units of storage that are designated for use by one or more host servers.

Virtual disks can also be used with Microsoft Cluster Server (MSCS). MSCS can use virtual disks for data storage and as quorum disks. Virtual disks and their attributes (such as file system format (NTFS) and size) are defined by the system administrator.

Consider the following points regarding virtual disks:

- Virtual disks are created on the IBM System Storage N series storage system and mounted as disks on Windows servers.
- Virtual disks are accessed and function as physical disks to the Windows server.
- Virtual disks appear within Windows as basic disks, not dynamic disks.
- Virtual disks can be expanded, unlike the native basic disks within Windows.
- Virtual disks are formatted with the NTFS file system.
- Virtual disks reside on physical N series volumes that are RAID protected and distributed among multiple disks for maximum data integrity and performance. See Figure 30-9.

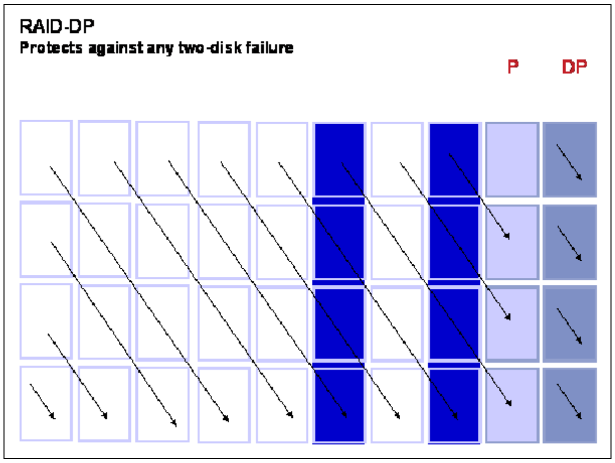


Figure 30-9 RAID-DP

- Dynamic-disk features and functionality are provided by the IBM System Storage N series storage system.
- Multiple virtual disks can be created on a single IBM System Storage N series storage system volume.
- Virtual disks can be accessed using the iSCSI over Gigabit Ethernet or FCP over Fibre Channel access methods.

30.6 SnapDrive for UNIX

SnapDrive for UNIX is a tool that simplifies the backup of data so that you can recover it if it is accidentally deleted or modified. SnapDrive for UNIX uses IBM System Storage N series Snapshot technology to create an image (that is, a Snapshot) of the data on a storage system attached to a UNIX host at a specific point in time. If the need arises later, you can restore the data to the storage system. When you restore a Snapshot, it replaces the current data on the storage system with the image of the data in the Snapshot.

In addition, SnapDrive for UNIX lets you provision storage on the storage system. SnapDrive for UNIX provides a number of storage features that enable you to manage the entire storage hierarchy, from the host-side application-visible file through the volume manager to the storage system-side LUNs providing the actual repository.

SnapDrive for UNIX is supported on the following host platforms:

- IBM AIX
- HP-UX
- Linux: Includes Red Hat Enterprise Linux, SUSE Linux, and Oracle Enterprise Linux.
- Solaris

SnapDrive for UNIX has the following key features:

- Advanced storage virtualization: Virtualizes IBM System Storage N series storage systems and integrates with native disk and volume management
- Automated mapping and management of new storage resources: Provisions new storage resources cleanly and efficiently, without application or server downtime
- Dynamic storage allocation: Quickly and easily reallocates storage systems in response to shifts in application or server demand
- Host file system consistent Snapshots: Implements a mechanism by which Snapshots taken are file system consistent
- Backup and restore: Provides a quicker way to create and restore storage from backups using IBM System Storage N series Snapshot technology

SnapDrive for UNIX allows you to create and delete a storage LUN or connect to a LUN that has already been created on the storage controller. If you must perform these tasks without SnapDrive for UNIX, you must log in to the storage system to create and map a LUN. Then, from the host and using the host commands, identify the LUN on the host, create a file system, and mount it. SnapDrive for UNIX achieves all these tasks with one command, reducing the time and probable errors during this process (Figure 30-11).

Although SnapDrive for UNIX can create storage using minimal options, you still need to understand the default values and use them appropriately.

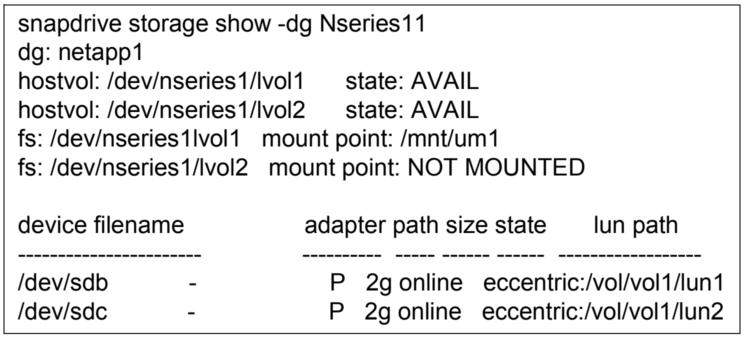
> **Attention:** SnapDrive for UNIX works only with Snapshots that it creates. It cannot restore Snapshots that it did not create.

30.6.1 How SnapDrive for UNIX works

The SnapDrive for UNIX software interacts with the host operating system and volume manager. It lets you easily and quickly back up and restore data about host volume groups that you stored on an IBM System Storage N series storage system. You can use it to manage the Snapshots that you create using it.

SnapDrive for UNIX coordinates the host Logical Volume Manager (LVM) volume groups and file systems to ensure that the host file systems stored on IBM System Storage N series LUNs have consistent images in the Snapshot. This action enables you to restore data from the backup Snapshots without requiring significant data recovery steps on the host.

You can also use SnapDrive for UNIX to create and manage storage. The SnapDrive storage commands, shown in Figure 30-10, work with LVM to let you create LVM objects and file systems that use the storage. They also let you remove the mappings between the storage and the host and delete the storage.



```
snapdrive storage show -dg Nseries11
dg: netapp1
hostvol: /dev/nseries1/lvol1      state: AVAIL
hostvol: /dev/nseries1/lvol2      state: AVAIL
fs: /dev/nseries1lvol1   mount point: /mnt/um1
fs: /dev/nseries1/lvol2   mount point: NOT MOUNTED

device filename               adapter path size state       lun path
-----------------------            ----------  ----- ------ ------  ------------------
/dev/sdb          -                   P   2g online   eccentric:/vol/vol1/lun1
/dev/sdc          -                   P   2g online   eccentric:/vol/vol1/lun2
```

Figure 30-10 UNIX SnapDrive

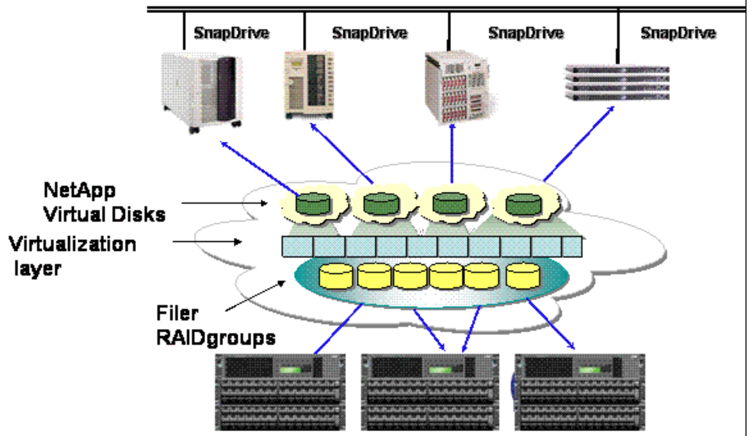
SnapDrive for UNIX communicates with the storage system using the host IP interface that you specified when you set up the storage system.

30.6.2 SnapDrive for UNIX and logical volumes

The host LVM combines LUNs from a storage system into disk or volume groups. This storage is then divided into logical volumes, which are used as though they were raw disk devices to hold file systems or raw data.

> **Volumes:** This book refers to logical volumes as *host volumes* to distinguish them from IBM System Storage N series storage system volumes.

SnapDrive for UNIX integrates with the host LVM to determine which IBM System Storage N series LUNs, as shown in Figure 30-11, make up each disk group, host volume, and file system requested for Snapshot. Because data from any given host volume can be distributed across all disks in the disk group, Snapshots can be taken and restored only for entire disk groups.



- Instant point-in-time Snapshot™ for non-disruptive backup, Data Mining, App testing
- Greater application performance: run other processes off of independent snapshots
- Near-instantaneous restoration with SnapRestore
- Replication using SnapMirror and rolling Snapshots

Figure 30-11 IBM System Storage N series LUNS

30.7 Flexible networked storage

SnapDrive is independent of the underlying storage access media and protocol. The iSCSI protocol provides storage access when the IBM System Storage N series storage system and host server are joined using Gigabit Ethernet. The FCP protocol facilitates storage access through a Fibre Channel host bus adapter (HBA) and storage area network (SAN).

The functionality and features intrinsic to SnapDrive are identical regardless of the underlying storage access protocol. It is because SnapDrive software utilizes either of the two access methods to access virtual disks, which are created and stored on storage systems. Thus, a virtual disk can be created and accessed using the iSCSI or FCP access protocols.

Virtual disks are referred to as logical unit numbers (LUNs) when accessed over the iSCSI and FCP protocols. Within the IBM System Storage N series storage system, LUNs are just special files.

30.8 Summary

SnapDrive is a complete storage management solution. It helps administrators execute nearly instantaneous Snapshot backups and restorations of application data. It also provisions storage from the host system as required by each application.

31

SnapManager

SnapManager empowers storage administrators with the tools necessary to do these tasks:

- Perform policy-driven data management
- Schedule and create regular database and application backups with minimal impact
- Restore data from these backups in the event of data loss or disaster
- Create clones for non-disruptive testing

With SnapManager, you can create backups on primary storage and create protected backups on secondary storage.

In this chapter, we describe several of the environments supported by SnapManager. As an installation example, we pay special attention to the installation on Microsoft Hyper-V.
The following topics are covered:

- Introduction to SnapManager
- Supported databases and applications
- SnapManager for Hyper-V
- SnapManager for Microsoft Exchange
- SnapManager for Oracle on UNIX
- SnapManager for SAP on Windows

31.1 Introduction to SnapManager

SnapManager uses N series technologies while integrating with the latest database and application releases. SnapManager is integrated with the following N series applications and technologies:

- Protection Manager uses resource pools, datasets, and protection policies to provide policy-based automation for SnapVault and SnapMirror capabilities.
- Operations Manager provides role-based access control of storage features for enhanced security.
- SnapDrive automates storage provisioning tasks and simplifies the process of creating error-free, host-consistent Snapshot copies of the storage.
- Snapshot (a feature of Data ONTAP) creates point-in-time copies of the database.
- SnapVault (a licensed feature of Data ONTAP) uses disk-based backups for reliable, low-overhead backup and recovery of databases.
- SnapMirror (a licensed feature of Data ONTAP) replicates database data across a global network at high speeds in a simple, reliable, and cost-effective manner.
- SnapRestore (a licensed feature of Data ONTAP) recovers an entire database in seconds, regardless of capacity or number of files.
- FlexClone (a licensed feature of Data ONTAP) helps to create fast, space-efficient clones
- of databases from the Snapshot backups.

31.2 Supported databases and applications

SnapManager is compatible with the following database and applications:

- Microsoft Hyper-V
- Microsoft Exchange Server
- Microsoft Exchange Server 2000
- Microsoft Office SharePoint Server
- Microsoft SQL Server
- Oracle
- SAP

31.3 SnapManager for Hyper-V

SnapManager for Hyper-V provides a solution for data protection and recovery for Microsoft Hyper-V virtual machines (VMs) running on Data ONTAP on cluster shared volumes and Windows shared and dedicated volumes.

You can perform application-consistent and crash-consistent dataset backups according to protection policies set by your backup administrator. You can also restore VMs from these backups. Reporting features enable you to monitor the status of and get detailed information about your backup and restore jobs.

31.3.1 What you can do with SnapManager for Hyper-V

SnapManager for Hyper-V enables you to back up and restore multiple virtual machines across multiple hosts. You can create datasets and apply policies to them to automate backup tasks such as scheduling, retention, and replication.

You can perform the following tasks with SnapManager for Hyper-V:

- Group virtual machines into datasets that have the same protection requirements and apply policies to those datasets
- Back up and restore dedicated and clustered virtual machines on storage systems running Data ONTAP software
- Back up and restore virtual machines running on cluster shared volumes and using Windows Failover clustering
- Automate dataset backups using scheduling policies
- Perform on-demand backups of datasets
- Retain dataset backups for as long as you need them, using retention policies
- Update the SnapMirror destination location after a backup successfully finishes
- Specify custom scripts to run before or after a backup
- Restore virtual machines from backups
- Monitor the status of all scheduled and running jobs
- Manage hosts remotely from a management console
- Provide consolidated reports for dataset backup, restore, and configuration operations
- Perform a combination of crash-consistent and application-consistent backups
- Perform disaster recovery operations using PowerShell cmdlet

31.3.2 Installing and uninstalling SnapManager for Hyper-V

Before you install the SnapManager for Hyper-V software, be aware of the following requirements:

- Data ONTAP requirements
- Hyper-V parent host requirements
- Hotfix requirements
- License requirements

Hyper-V parent host requirements

Hyper-V parent hosts are physical servers on which the Hyper-V role is enabled. Hosts that contain virtual machines are added to SnapManager for Hyper-V for protection and recovery. To install and run all of the SnapManager for Hyper-V software components, the Hyper-V parent hosts must meet certain technical requirements. All Hyper-V parent hosts must meet minimum operating system and Hyper-V requirements.

SnapManager for Hyper-V runs on any of the following operating system versions:

- Windows Server 2008 R2 x64 or later
- Windows Server 2008 R2 SP1 x64

Management consoles must be running one of the following operating systems:

- Windows Server 2003, SP2
- Windows XP Professional, SP3
- Windows Server 2008
- Windows Vista Business and Ultimate, SP1
- Windows Server 2008 R2
- Windows Server 2008 R2, SP1
- Windows 7

Installing or upgrading SnapManager for Hyper-V

You can install or upgrade SnapManager for Hyper-V so that you are able to back up and restore your data:

1. Download SnapManager for Hyper-V from the N series support site.
2. Launch the SnapManager for Hyper-V executable file.
3. Complete the steps in the SnapManager for Hyper-V InstallShield wizard.

Installation order

You need to install SnapDrive for Windows on all hosts before installing SnapManager for Hyper-V. If the hosts are members of a cluster, all nodes in the cluster require the installation of SnapDrive for Windows.

When SnapManager for Hyper-V starts it communicates with SnapDrive for Windows to get the list of all virtual machines running on a host. If SnapDrive for Windows is not installed on the host, this API fails and the SnapManager for Hyper-V internal cache does not update with the virtual machine information.

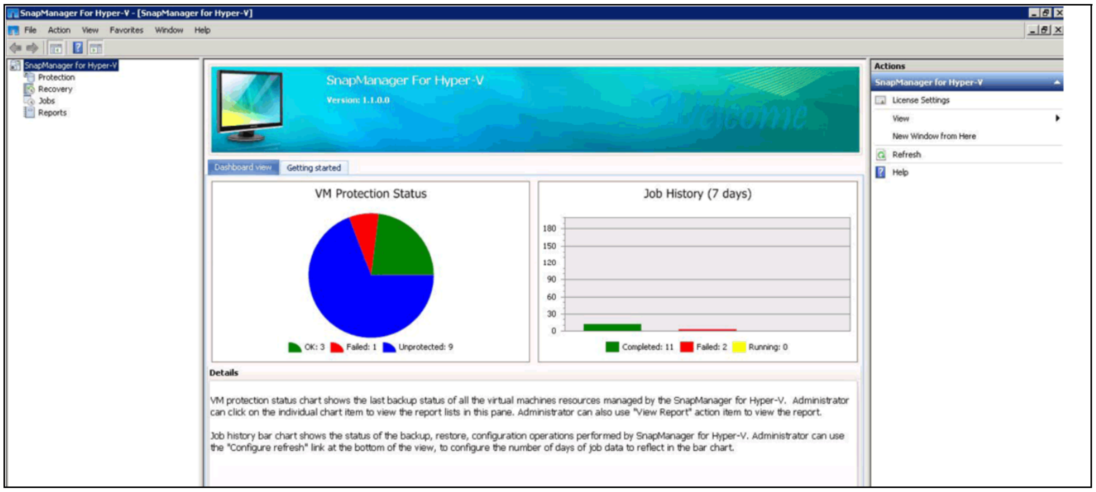
31.3.3 Configuring SnapManager for Hyper-V

You can configure and manage your hosts and virtual machine resources with policies to protect and restore your data.

Dashboard

The dashboard displays an overview of resources that are currently being protected, as well as those that are not protected. You can select different segments of either the VM Protection Status pie chart or the Job History bar graph to view general information about the status of your jobs, resources, and history.

The dashboard (Figure 31-1) displays an overview of resources.



VM protection status chart shows the last backup status of all the virtual machines resources managed by the SnapManager for Hyper-V. Administrator can click on the individual chart item to view the report lists in this pane. Administrator can also use "View Report" action item to view the report.

Job history bar chart shows the status of the backup, restore, configuration operations performed by SnapManager for Hyper-V. Administrator can use the "Configure refresh" link at the bottom of the view, to configure the number of days of job data to reflect in the bar chart.

Figure 31-1 SnapManager for Hyper-V Windows console

When you select a segment in the VM Protection Status pie chart, you can view information about the protection status of the virtual machines in the Details pane.

Adding a Hyper-V parent host or host cluster

You can add a Hyper-V parent host or host cluster to back up and restore your virtual machines.

1. From the navigation pane, click **Protection**.
2. From the Actions pane, click **Add host**.
3. Either type the name of the host or click **Browse** to select, and click **Add**.

> **Tip:** When you add a host to a cluster, the information about the new host is not automatically displayed in the GUI. Manually add the host information to the xml file in the installation directory.

Exporting the VM from a non-Data ONTAP host

You must first export a virtual machine (VM) from a non-Data ONTAP host before you can import it to a Data ONTAP host.

The VM that you want to migrate must be powered off.

1. Open the application Server Manager.
2. In the left pane, click **Role** → **Hyper-V** → **Hyper-V Manager**.
3. Select the name of the non-Data ONTAP host on which the VM that you want to migrate currently resides.
4. From the Virtual Machines pane, right-click the name of the VM that you want to migrate.
5. Click **Export**.
6. Click **Browse**.
7. A window displaying available hard disk drives opens.
8. Click the Data ONTAP host destination for the VM.
9. Click **Select Folder**.
10. The VM is exported to the Data ONTAP destination that you chose in Step 7.

31.3.4 Managing backup jobs

You can manage scheduled backups using the Jobs Management window and you can also create and monitor on-demand backups in SnapManager for Hyper-V.

You can create scheduled backup jobs using policies attached to datasets. You can create, modify, view, and delete the policies that make up scheduled backup jobs.

You can create on-demand backup jobs when you want them. An on-demand backup job can include retention and replication policies as well as scripts to run before and after the backup has taken place.

Types of backup jobs that SnapManager for Hyper-V can perform

SnapManager for Hyper-V enables you to use two types of backup jobs: application consistent and crash consistent. A combination of the two provides an ideal backup strategy.

Application-consistent backup

These backups are thorough, reliable, and resource intensive. Application-consistent backups are made in coordination with Microsoft Volume Shadow Copy Service (VSS) to ensure that each application running on the VM is quiesced before making a Snapshot copy. This backup method guarantees application data consistency. It can be used to restore VMs and the applications running on them. However, application-consistent backups are time consuming and can be complex.

Crash-consistent backup

These backups are quick Snapshot copies of all the LUNs used by VMs involved in a dataset. The resulting backup copy is similar to the data capture of a VM that crashes or is otherwise abruptly powered off. Crash-consistent backups are a quick way to capture data, but the VMs must be present in order to be restored from a crash consistent backup.Crash-consistent backups are not intended to replace application-consistent backups.

Crash-consistent backups take only one Snapshot copy, always. They do not provide VSS integration.

Multiple crash-consistent backups can execute in parallel. A crash-consistent backup can run in parallel with an application-consistent backup.

Manually backing up a dataset

You can create an on-demand backup of a dataset. You must have the following information available:

- Backup name and description
- Policy name, if necessary
- Policy override information
  (if you plan to change any of the previously specified policy options)
- Backup type
- Backup options information

Follow these steps:

1. From the navigation pane, click **Protection** → **Datasets**.
2. Select the dataset for which you want to create a manual backup and click Backup.The Backup wizard appears.
3. Complete the steps in the wizard to create your on-demand backup. Closing the wizard does not cancel the on-demand backup.

Monitoring backup jobs

You can view the scheduled backup jobs for a particular dataset by using the Jobs Management window Scheduled tab. You can also view the backup and restore jobs that are currently running by using the Jobs Management window Running tab.

1. From the navigation pane, click Jobs.
2. Click either the Scheduled tab or the Running tab.
3. Select the scheduled or running backup job, or the restore job, that you want to monitor. Information about the job appears in the Details pane.
4. Use the Running Job report in Reports view, if you want to view a live report of a running job.

> **Tip:** You can also monitor backup jobs with Microsoft's SCOM console. See the Microsoft web site for more information.

31.3.5 Restoring a virtual machine

You can restore a virtual machine (VM) from a backup by using SnapManager for Hyper-V. You can also restore a VM that is part of a cluster. SnapManager for Hyper-V determines the appropriate node in the cluster to restore the VM.

To restore a VM, SnapManager for Hyper-V uses the file-level restore feature in SnapDrive for Windows. You can spread the associated files of a VM, including the configuration file, Snapshot copies, and any VHDs, across multiple Data ONTAP LUNs. A LUN can contain files belonging to multiple VMs.

If a LUN contains only files associated with the VM that you want to restore, SnapManager for Hyper-V restores the LUN by using LCSR (LUN clone split restore). If a LUN contains additional files not associated with the virtual machine that you want to restore, SnapManager for Hyper-V restores the virtual machine by using the file copy restore operation. You can follow these steps:

1. From the navigation pane, click **Recovery**.
2. Select the virtual machine that you want to restore.
3. In the Backups pane, select the backup name that you want to restore and click **Restore**. The Restore wizard appears.

   If you start a restore operation of a Hyper-V virtual machine, and another backup or restoration of the same virtual machine is in process, it fails.
4. Complete the steps in the wizard to restore the virtual machine backup. Closing the wizard does not cancel the restore operation. SnapManager for Hyper-V validates the virtual machine configuration before beginning the restore operation. If there have been any changes in the virtual machine configuration, a warning appears and you can choose to continue or cancel the operation.

31.4 SnapManager for Microsoft Exchange

SnapManager provides you with an integrated data management solution for Microsoft Exchange that enhances the availability, scalability, and reliability of Exchange databases. SnapManager provides rapid online backup and restoration of databases, along with local or remote backup set mirroring for disaster recovery.

31.4.1 What SnapManager for Microsoft Exchange does

SnapManager uses online Snapshot technology that is part of Data ONTAP and integrates Exchange backup and restore APIs and Volume Shadow Copy Service (VSS). SnapManager uses SnapMirror to support disaster recovery.

SnapManager provides the following data management capabilities:

- Migrating Exchange databases and transaction logs to LUNs on storage systems
- Backing up Exchange databases and transaction logs from LUNs on storage systems
- Verifying the backed-up Exchange databases and transaction logs
- Managing backup sets
- Archiving backup sets
- Restoring Exchange databases and transaction logs from previously created backup sets

You can use SnapManager with configurations having multiple servers. You can perform local administration, remote administration, and remote verification.

SnapManager provides the following capabilities:

- Local administration:

  You install SnapManager on the same Windows host system as your Exchange server.
- Remote administration:

  If you install SnapManager on a remote computer, you can run SnapManager remotely to perform any task that you can perform on a locally installed SnapManager system.

- Remote verification:

  You can also perform remote database verification from a remote administration server that is configured with SnapDrive and Exchange server. Remote verification offloads the CPU-intensive database verification operations that can affect the performance of your production Exchange server.

31.4.2 Installation on a stand-alone Windows host system

You can install SnapManager on a stand-alone Windows host system that is used for a production Exchange server, a remote administration server, or a remote verification server.

You can run the software installation utility for SnapManager in either the interactive mode or the unattended mode. SnapManager guides you through the interactive mode; the unattended mode requires that you type certain commands and then installation takes place on its own.

31.4.3 Installing SnapManager in interactive mode

You can install SnapManager using the software installation utility in the interactive mode. The InstallShield wizard guides you through the installation.

The SnapManager Server Identity account must meet the following requirements:

- The account must have administrator privileges on the Exchange server.
- The account must also have system administrator server privileges.

The CD-ROM installation is the same as the network installation. The only difference is the name of the installation executable and the distribution media.

Do not use Terminal Services for any type of SnapManager administration, because you might miss critical information that is displayed only in pop-up boxes at the system console.

You do not need to stop Exchange services while you install SnapManager. Exchange can continue to run while you install SnapManager and afterward.

The SnapManager software installation program does not allow you to continue with the installation process if Microsoft .NET 3.5 is not installed. If .NET is not installed on a 32-bit system, the installation setup automatically installs it; on a 64-bit system, SnapManager prompts you to install it.

Follow these steps:

1. Network:

   Download the SnapManager package from the network, save it on the Windows host system, and then launch the SnapManager installation package by double-clicking it in your Windows Explorer.

   CD-ROM:

   Browse to the SnapManager installation package and double-click **`setup.exe`**.
2. In the *License Agreement* window, accept the license agreement.
3. In the *Customer Information* window, specify the user name, the organization name, and the SnapManager license type.
4. Note the full path of the folder in which you want to install SnapManager.
   You can change this default directory by clicking **Change**.

5. In the **S***napManager Server Identity* window, specify the user account you want to use to run SnapManager.
6. Type and confirm the password.
7. Click **Install**.
8. Wait until the *InstallShield Wizard Completed* window appears; then click **Finish** to exit the software installation utility.

31.5 SnapManager for Oracle on UNIX

SnapManager for Oracle simplifies and automates database backup, recovery, and cloning by using the Snapshot, SnapRestore, and FlexClone technologies.

31.5.1 What SnapManager for Oracle on UNIX does

SnapManager for Oracle on UNIX provides the following capabilities:

- Create space-efficient backups to primary or secondary storage and schedule backups to occur on a regular basis.
- Restore full or partial databases using file-based or volume-based restore operations and preview restore operations before they occur from primary or secondary storage.
- Automatic restore and recovery of database backups.
- Prune archive log files from the archive log destinations while creating the archivelog-only backups.
- Automatically retain minimum number of archive log backups by retaining only the backups with unique archive log files.
- Track operation details and produce reports by host, profile, backup, or clone.
- Verify backup status.
- Maintain history of the SnapManager operations associated with a profile.
- Create space-efficient clones of backups on primary or secondary storage.

31.5.2 Installing or upgrading SnapManager on a UNIX host

You can install or upgrade SnapManager software on any approved UNIX host. Install only one SnapManager server instance per host.

For HP-UX PA-RISC platform, ensure that you installed the 32-bit Oracle client software.

The software is installed in the following paths:

- Solaris/opt/NTAPsmo
- For all other platforms: /opt/Nseries/smo

Follow these steps:

1. Log in as root.
2. On the UNIX database host, change to the directory where you saved the software.

3. If the file is not an executable file, use the following command to change permissions on the downloaded file so it is executable:

   ```
   chmod 544 Nseries.smo*
   ```

4. For upgrades, stop the server:

   ```
   smo_server stop
   ```

5. Use the appropriate command to start the installation of the software on the host:

   ```
   Solaris (SPARC-x86_64)
   # ./Nseries.smo.sunos-sparc64-3.2.bin
   Solaris (*x86_64)
   # ./Nseries.smo.sunos-x64-3.2.bin
   AIX
   # ./Nseries.smo.aix-ppc-3.2.bin
   AIX (ppc64)
   # ./Nseries.smo.aix-ppc64-3.2.bin
   HP-UX (Itanium)
   # ./Nseries.smo.hpux-ia64-3.2.bin
   HP-UX (PA-RISC)
   # ./Nseries.smo.hpux-hppa-3.2.bin
   Linux
   # ./Nseries.smo.linux-x86-3.2.bin
   # ./Nseries.smo.linux-x64-3.2.bin
   ```

6. After the Introduction text, complete the following steps:
   a. Press Enter to continue.
   b. If it is an upgrade, press Enter when you receive the following message:
   Existing SnapManager For Oracle Detected.

7. For SnapManager for Oracle, at the prompt for the operating system user, press Enter to accept the default value. For a new installation, the default value for the user is oracle.

8. At the prompt for operating system group, press Enter to accept the default value. For a new installation, the default value is dba.

9. At the prompt for the **Server Startup Type**, press Enter to accept the default value.

10. At the Configuration Summary page, press Enter to continue.

11. or HP-UX PA-RISC platform that run Oracle 11gR2, you need to set the Oracle client home path in the SnapManager configuration file (smo.config). The Oracle client home path is the location in which you have installed the 32-bit Oracle client software. Enter the path of the 32-bit Oracle client software in the Oracle client home variable of the configuration file (smo.config): oracle.client.home=<>.

12. Start the SnapManager server by entering this command:

    ```
    smo_server start
    ```

    SnapManager displays a message stating that the SnapManager server is running.

13. Verify that the SnapManager system is running correctly by entering this command:

    ```
    smo system verify
    ```

    SnapManager displays a message stating that the operation succeeded.

14. For upgrades, upgrade each SnapManager repository by entering this command:

    ```
    smo repository update -repository -dbname repo_service_name -host repo_host
    -login -username repo_username -port repo_port
    ```

    SnapManager displays a message stating that the operation succeeded.

31.6 SnapManager for SAP on Windows

SnapManager for SAP simplifies and automates database backup, recovery, and cloning by using the Snapshot, SnapRestore, and FlexClone technologies.

31.6.1 What SnapManager for SAP on Windows does

SnapManager for SAP has several advantages for managing data and databases over other products:

- Create space-efficient backups to primary or secondary storage and schedule backups to occur on a regular basis.
- Restore full or partial databases using file-based or volume-based restore operations and preview restore operations before they occur from primary or secondary storage.
- Perform automatic restore and recovery of database backups.
- Prune archive log files from the archive log destinations while creating the archivelog-only backups.
- Automatically retain minimum number of archive log backups by retaining only the backups with unique archive log files.
- Track operation details and produce reports by host, profile, backup, or clone.
- Verify backup status.
- Maintain history of the SnapManager operations associated with a profile.
- Create space-efficient clones of backups on primary or secondary storage. For example, you can use the clone for testing updates in non-production environments.

31.6.2 Installing SnapManager for SAP on Windows

You can install or upgrade SnapManager software on any approved Windows host. Install only one SnapManager server instance per host.

Follow these steps:

1. For upgrades, stop the server by completing the following actions:
   a. In the Windows Services window, select N series SnapManager 3.1 for SAP.
   b. In the left panel, click **Stop**.
2. Double-click the downloaded executable file.

   For Windows x86, use Nseries.smsap.windows-x86-3.1.exe. For Windows x64, use Nseries.smsap.windows-x64-3.1.exe.
3. A dialog box might appear with this message:

   ```
   The publisher could not be verified. Are you sure you want to run this software?
   ```
   Click **OK**.
4. In the Introduction window, click **Next**.
5. If it is an upgrade, a pop-up window appears. To continue, click **OK**.
6. For new installations (not upgrades) at the next Choose Install Folder window, either click **Next** to accept the default location for the installation folder, or choose a new location.
7. On the Menu Availability window, click **Next**.

8. On the Specify Service Properties window, enter the account and password for the Windows service for SnapManager.
9. On the Pre-Installation Summary window, click **Install**.
10. At the Install Complete window, click **Next**.
11. At the Important Information window, click **Done** to exit the installer.
12. Start the SnapManager server by completing the following steps:
    a. In the Windows Services window, select N series SnapManager 3.1 for SAP.
    b. In the left panel, click **Start**.
13. Verify that the SnapManager system is running correctly by following these steps:
    a. Open the SnapManager Command Line Interface (CLI) command prompt window by selecting **Start** → **Programs** → **N series** → **SnapManager for SAP** → **Start SMSAP Command Line Interface (CLI)**
    b. In the CLI window, enter the following command:

    ```
    smsap system verify
    ```

    SnapManager displays a message stating that the operation succeeded.
14. For upgrades, upgrade each SnapManager repository by following these steps:
    a. Open the SnapManager Command Line Interface (CLI) command prompt window by selecting **Start** → **Programs** → **N series** → **SnapManager for SAP** → **Start SMSAP Command Line Interface (CLI)**
    b. In the CLI window, enter the following command:

    ```
    smsap repository update -repository -dbname repo_service_name  -host
    repo_host
    -login -username repo_username   -port repo_port
    ```

32

Snap Creator

Snap Creator is a backup and recovery software solution that enables you to integrate Snapshot technology with any application that is not supported by SnapManager products.

Snap Creator is platform and operating system independent. It provides application integration through plug-ins that enable it to support any application on a storage system. Snap Creator uses the plug-ins to handle quiesce and unquiesce actions for a given application or database.

Snap Creator supports application plug-ins for Oracle, DB2, MySQL, Sybase ASE (Sybase), IBM Lotus® Domino® (Domino), SnapManager for Microsoft SQL Server, SnapManager for Microsoft Exchange, MaxDB, and VMware (vSphere and vCloud Director). Additional application plug-ins are available through the Snap Creator Community.

Snap Creator provides a management interface for Snapshot technology, SnapVault, Open Systems SnapVault, SnapMirror, Protection Manager, Operations Manager, and FlexClone technology.

In this chapter, we describe Snap Creator architecture, features, and installation.

The following topics are covered:

- Snap Creator architecture
- Installing Snap Creator on UNIX
- Configuring Snap Creator Server

32.1 Snap Creator architecture

Snap Creator consists of a server and agent layer. The GUI, configuration, and CLI reside in the server layer. The agent runs remotely or locally and allows the Snap Creator Server to send quiesce or unquiesce operations to a given database.

The communication layer from the agent to the server is Simple Object Access Protocol (SOAP) over HTTP.

The illustration in Figure 32-1 shows the Snap Creator architecture.

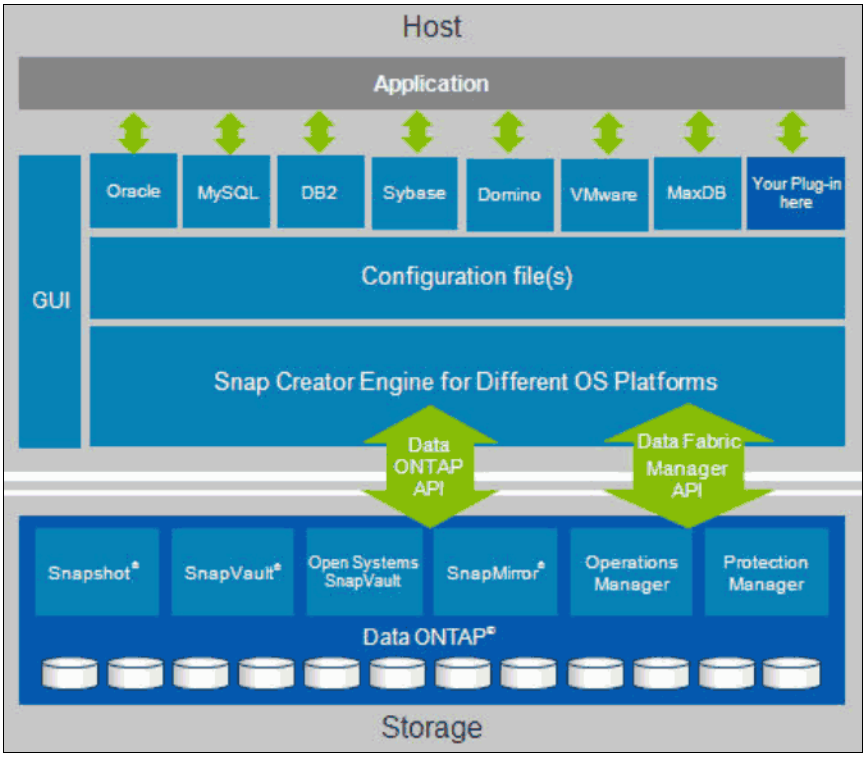


Figure 32-1 Snap Creator architecture

32.1.1 Security features of Snap Creator

Snap Creator provides security features such as RBAC for Storage controller, Host security for Snap Creator Agent, and RBAC for Snap Creator users through GUI.

RBAC for Storage controller

If you are not using DataFabric Manager server proxy, you need a user name and password to communicate with storage controllers. Passwords can be encrypted so that they are not saved in clear text.

Network communications are done through HTTP (80) or HTTPS (443), so you must have one or both of these ports open between the host where Snap Creator runs and the storage controllers. A user must be created on the storage controllers for authentication purpose. In the case of HTTPS, ensure that the user is enabled and configured on the storage controllers.

Snap Creator Agent security

Snap Creator uses host security to allow only authorized hosts to access the agent. Additionally, it checks the user name or password if you are not using a Snap Creator Server to communicate with the agent. This feature allows you to specify multiple host lines. To restrict access for third party applications through SOAP, the agent offers a user/password authentication.

If a command contains the path of the Snap Creator installation directory, it is blocked.

RBAC for Snap Creator users through GUI

You can create and manage multiple user accounts within Snap Creator GUI. The existing user in the GUI (which is created during installation or profile setup) acts as a super user and has access to the complete system. The ability to assign a set of profiles restricts a user to operate in a defined area. It is useful in a multi-tenant environment.

There are four types of users:

- Super user: Has access to everyone's work space and is the only user that can create users and profiles.
- Admin user: Has access only to a set of profiles and can perform all actions without any restrictions only on those profiles.
- Read-only user: Has access to a set of profiles but can only perform a set of read-only operations. This user cannot perform write or execute operations like creating a Snapshot copy or creating a configuration.
- Custom user: Has access to a set of profiles and a set of actions.

32.1.2 Snap Creator integration

Snap Creator integrates either fully or optionally with other software products and technologies:

- Optionally integrates with both SnapDrive for UNIX (SDU) and SnapDrive for Windows.
- If SnapDrive is used instead of Manage ONTAP Solution, which sends a call to the storage controller for the Snapshot copy, Snap Creator runs SnapDrive.
- Optionally uses SnapVault directly instead of Protection Manager to transfer Snapshot copies to secondary.
- Snapshot, SnapVault, SnapMirror, LUN cloning, volume cloning, and igroup mapping using Data ONTAP API.
- Any application or database that runs in an open systems environment (you can write the application backup script or plug-in if one does not exist).
- NetBackup, CommVault, or any backup software with CLI commands.
- Optionally integrates with Operations Manager for monitoring (the ability to create events in Operations Manager).
- Optionally integrates with Protection Manager to perform secondary backup (Snap Creator backup copies can be registered in Protection Manager).
- Optionally integrates with Open Systems SnapVault.

32.1.3 What a Snap Creator agent is

The Snap Creator Agent is a lightweight daemon that runs remotely or locally and allows the Snap Creator Server to send quiesce or unquiesce operations to a given database.

The Snap Creator Agent remotely handles operations on application through the plug-ins. All Snap Creator configurations are stored centrally on the Snap Creator Server and all backup jobs can be scheduled from the same host. It provides a single pane of glass (SPOG) for backup and restore.

Snap Creator uses the Snap Creator Agent, which runs as a daemon, to quiesce the application. The default port used is 9090, but any other port can also be used.

SOAP is used over the HTTP for communication. Based on a WSDL, any SOAP client can interact with the agent. Currently, Apache CXF (for Java) and PowerShell (for Windows) can be used. The supported application plug-ins are built into the agent.

32.2 Installing Snap Creator on UNIX

Snap Creator installation for UNIX differs from Windows in that the software package is an executable that when extracted contains both the Snap Creator Server and the Snap Creator Agent.

UNIX Services (agent/server) feature offers a start script for the Snap Creator Agent and Snap Creator Server. The start scripts are written in UNIX shell script (bourne shell) and are designed to run on all UNIX environments supported by Snap Creator.

32.2.1 Installing the Snap Creator Server

You can install Snap Creator Server on UNIX. The Snap Creator Server is designed to run on any open systems platform.

Follow these steps:

1. Extract the .tgz file to /usr/local. Change directory to the Snap Creator Server root directory /path/to/scServer_v<#>.
2. Run Snap Creator setup by entering the following command:

   ```
   ./snapcreator --profile setup
   ```

   > **Tip:** The Snap Creator executable must already be configured upon extraction with the proper permissions to be executed. If the **`profile setup`** command does not work, the permissions must be added by running the command: **`chmod 755 snapcreator`**.

3. Accept the EULA license agreement.
4. Optional: Enter the serial number of the storage system that will be used with Snap Creator.
5. To enable GUI job monitoring, enter: **`y`**
6. Enter the job monitor size.
7. Enter the user name and password for the administrative user for the GUI.
8. Start the Snap Creator GUI by following the instructions provided on the window.

9. To start Snap Creator GUI, type the following URL in the web browser:

   ```
   http://<HostName>:<port>
   ```

   Where:

   – HostName is the host name or IP address of the Snap Creator Server.

   – Port is the port number where the Snap Creator Server is running. By default, it is port 8080.

32.2.2 Installing Snap Creator agent

The Snap Creator Agent is designed to run on any open systems platform. If the agent is not required, you can choose to run the Snap Creator Server on the application server locally.

Follow these steps:

1. Extract the .tgz file to /usr/local.
2. Change directory to the Snap Creator Server root directory /path/to/scAgent_v<#>.
3. Run the Snap Creator setup by entering the following command:

   ```
   ./snapcreator --profile setup
   ```

   The Snap Creator Agent setup on UNIX configures the /path/to/scAgent_v<#>/xscript and prints usage information.

4. Install the agent:

   The Snap Creator Agent has the ability to run as a daemon under UNIX. The agent uses either the default port 9090 or a user-specified port. To set a non-default port number, configure the following environment variable: SC_AGENT_PORT.

32.3 Configuring Snap Creator Server

The Snap Creator configuration file is located in the following path:

```
/path/to/scServer_v<#>/configs/<profile>/<config>.conf
```

You can create multiple configurations, but Snap Creator Server runs only one configuration at a time.

You can edit this file by using Visual Interactive (VI) (UNIX) or any text editor in Windows. Additionally, you can use the Snap Creator GUI to edit and manage configuration files.

32.3.1 Creating a configuration file using CLI

You can create a new directory or profile for your configuration under /path/to/scServer_v<#>/configs. It is a preferred practice to name it after the host or application that is backed up.

Follow these steps:

1. Create the following directory:

   ```
   mkdir /path/to/scServer_v<#>/configs/oraprod01
   ```

2. Copy or rename the following default template to your new configuration directory:

   ```
   cp /path/to/scServer_v<#>/configs/default/default.conf
   /path/to/scServer_v<#>/configs/oraprod01/oraprod01.conf
   ```

3. Edit your configuration file by using VI (UNIX) or any text editor in Windows.

32.3.2 Creating a configuration file using the GUI

You can create a configuration file using the GUI.

Follow these steps:

1. Open the web browser to the following URL:

   ```
   http://myserver.mydomain.com:8080
   ```

   Then log in.

2. In the *Management Configurations* window, click **+ Add backup profile** and enter the new profile name.

   The profile name must relate to the application being backed up. Adding a profile creates a directory under the/path/to/scServer_v<#>/configs directory. The new backup profile is created.

3. Right-click the backup profile and select + **New Configuration**.
4. Proceed through the configuration wizard.
5. Review the summary and click **Finish**.

33

VMware vSphere

This chapter covers the integration of VMware vSphere and N series storage systems. N series software features and vSphere complement each other very well. We can provide only an overview of possible solution architecture and mutual benefits. For further details, see 33.6, “Further reading” on page 482 at the end of this chapter.

The following topics are covered:

- Server virtualization
- Benefits of N series with VMware vSphere 5
- Virtual Storage Console
- Using N series deduplication with VMware
- Coupling deduplication and compression
- Further reading

33.1 Server virtualization

With virtualization, one computer does the job of multiple computers, by sharing the resources of a single computer across multiple environments (Figure 33-1). By using virtual servers and virtual desktops, you can host multiple operating systems and multiple applications locally and in remote locations, freeing you from physical and geographical limitations. Server virtualization also offers energy savings and lower capital expenses because of more efficient use of your hardware resources. You also get high availability of resources, better desktop management, increased security, and improved disaster recovery processes when you build a virtual infrastructure.

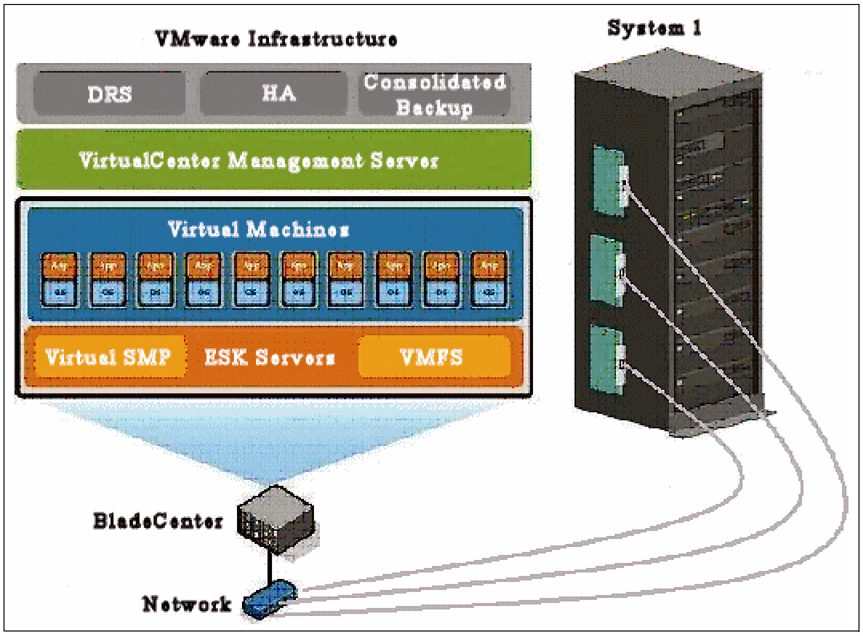


Figure 33-1 Server virtualization

The virtualization concept became more popular with the introduction of hypervisors (software responsible for the virtualization layer) in the x86 platform. However, server virtualization is not a new technology. Virtualization was first implemented more than 30 years ago by IBM as a way to logically partition mainframe computers into separate virtual machines. These partitions allowed mainframes to *multitask* (run multiple applications and processes at the same time), but because of the high cost of the mainframes, the virtualization technology did not become popular.

The broad adoption of Microsoft Windows and the emergence of Linux as server operating systems in the 1990s established x86 servers as the industry standard. The growth in x86 server and desktop deployments has introduced new IT infrastructure and operational challenges. Virtualization in the x86 platform allowed an option for companies that needed to centralize the management of servers and desktops together with a reduction in cost of management.

33.1.1 VMware Virtual Infrastructure

The VMware approach to virtualization inserts a thin layer of software directly on the computer hardware or on a host operating system. This software layer creates virtual machines. It also contains a virtual machine monitor or “hypervisor” that allocates hardware resources dynamically and transparently. With the hypervisor, multiple operating systems can run unaware concurrently on a single physical computer.

The VMware Virtual Infrastructure, with IBM System Storage N series storage and its storage virtualization capabilities, brings several benefits to data center management:

- Server consolidation and infrastructure optimization:

  Virtualization makes it possible to achieve higher resource utilization by pooling common infrastructure resources and breaking the “one application to one server” model.

- Physical infrastructure cost reduction:

  With virtualization, you can reduce the number of servers and related IT hardware in the data center. The benefit is reductions in real estate and power and cooling requirements, resulting in lower IT costs.

- Improved operational flexibility and responsiveness:

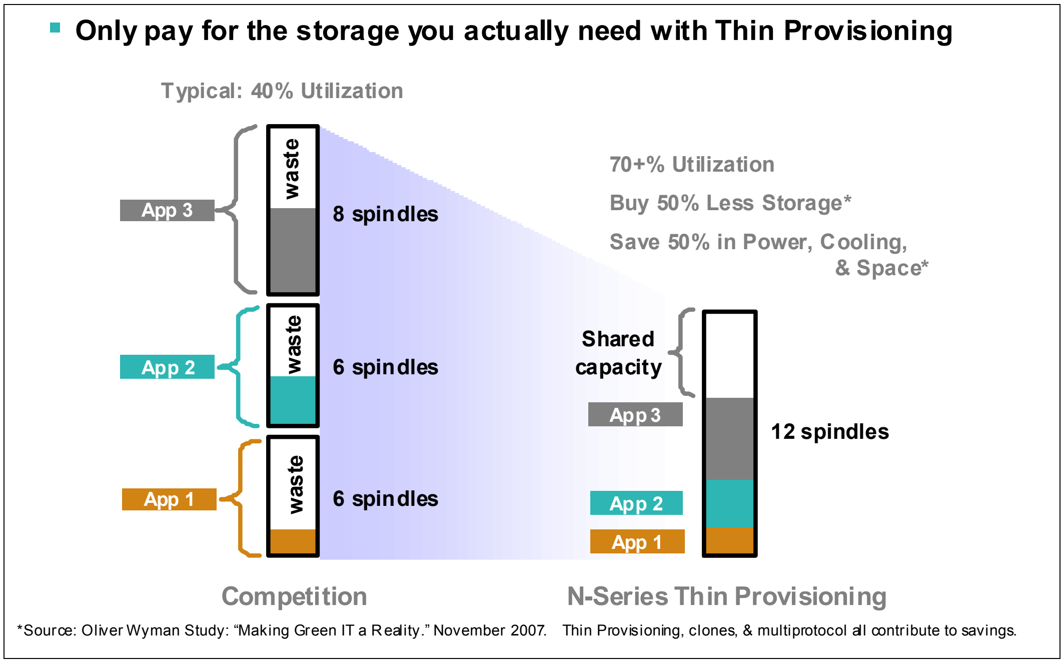
  Virtualization offers a new way to manage IT infrastructure and can help IT administrators spend less time on repetitive tasks, such as provisioning, configuration, monitoring, and maintenance.

- Increased application availability and improved business continuity:

  You can eliminate planned downtime and recover quickly from unplanned outages with the ability to securely back up and migrate entire virtual environments with no interruption in service.

- Storage saving:

  By taking advantage of the N series thin provisioning capability, you can allocate the space of the used files only (see Figure 33-2).



Only pay for the storage you actually need with Thin Provisioning

*Source: Oliver Wyman Study: “Making Green IT a Reality.” November 2007. Thin Provisioning, clones, & multiprotocol all contribute to savings.

Figure 33-2 Thin provisioning savings

- Rapid data center deployment:

  With the LUN clone capability of N series system, you can quickly deploy multiple VMware hosts in the data center.

33.1.2 Implementation example

This section provides an example of one of the several configurations that was used and implemented in the development of this Redbooks publication.

The environment has the following setup:

- Server: IBM System x3850 system
- Storage: IBM System Storage N series N6240
- Storage protocol: FCP used for the connection between the storage system and the server; boot from SAN
- SAN switch
- Network:
  - One Gigabit NIC for VMware Service Console if using VMware vSphere 5
  - One Gigabit NIC for vMotion
  - One Gigabit NIC for guest operating systems
- Virtualization software:
  - VMware Virtual Infrastructure (ESXi V5)
  - VMware vSphere vCenter 5

Network and storage redundancy: This example does not consider redundancy for network and storage.

33.2 Benefits of N series with VMware vSphere 5

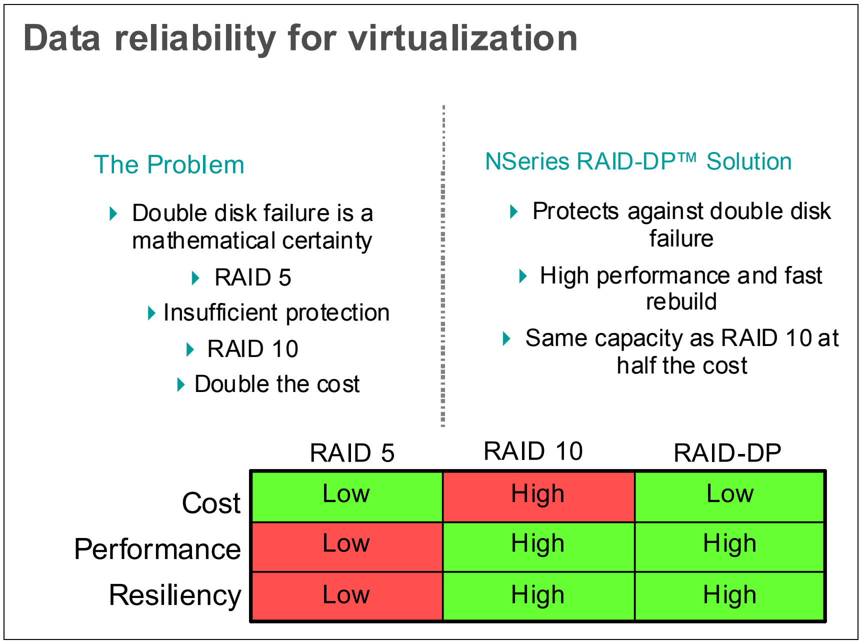
This section outlines the benefits that the IBM System Storage N series provides to VMware vSphere 5 environments. It includes the following topics:

- Increased protection with RAID-DP
- Cloning virtual machines
- N series LUNs for VMWare host boot
- N series LUNs for VMFS datastores
- Using N series LUNs for Raw Device Mappings
- Growing VMFS datastores
- Backup and recovery of the virtual infrastructure (SnapVault, Snapshot, SnapMirror)
- Using N series deduplication with VMware

33.2.1 Increased protection with RAID-DP

In a VWware vSphere 5 environment, the performance and availability of the storage system are important. Generally many different server systems are consolidated onto each VMware ESX host, and a failure can cause all of the machines to have an outage or data loss.

RAID-DP (Figure 33-3) provides the benefit of both performance and availability without the requirement to double the physical disks. This benefit is achieved by using two dedicated parity disks. Each disk has separate parity calculations, which allows the loss of any two disks in the Redundant Array of Independent Disks (RAID) set while still providing excellent performance.



Data reliability for virtualization

The Problem

- Double disk failure is a mathematical certainty
- RAID 5
- Insufficient protection
- RAID 10
- Double the cost

NSeries RAID-DP™ Solution

- Protects against double disk failure
- High performance and fast rebuild
- Same capacity as RAID 10 at half the cost

	RAID 5	RAID 10	RAID-DP
Cost	Low	High	Low
Performance	Low	High	High
Resiliency	Low	High	High

Figure 33-3 RAID-DP

33.2.2 Cloning virtual machines

Although you can clone guests natively with VMware, cloning from the N series provides significant storage space savings. This type of cloning is helpful when you need to test existing VMware guests. Guests can be cloned at the N series level and use little additional disk capacity due to the deduplication.

33.2.3 Multiprotocol capability for storing files on iSCSI, SAN, or NFS volumes

The N series storage system provides flexibility in the method and protocol used to connect to storage. Each has advantages and disadvantages, depending on the existing solution and the VMware environment requirements.

Traditionally, most VMware scenarios use standard Fibre Channel SAN connectivity. With N series, you can keep using this method if it is already in the environment. However, fiber connectivity can be expensive if new purchases are required. As a result, more environments are now implementing network connectivity methods to storage. Such methods include iSCSI, Network File System (NFS), and Common Internet File System (CIFS), as illustrated in Figure 33-4.

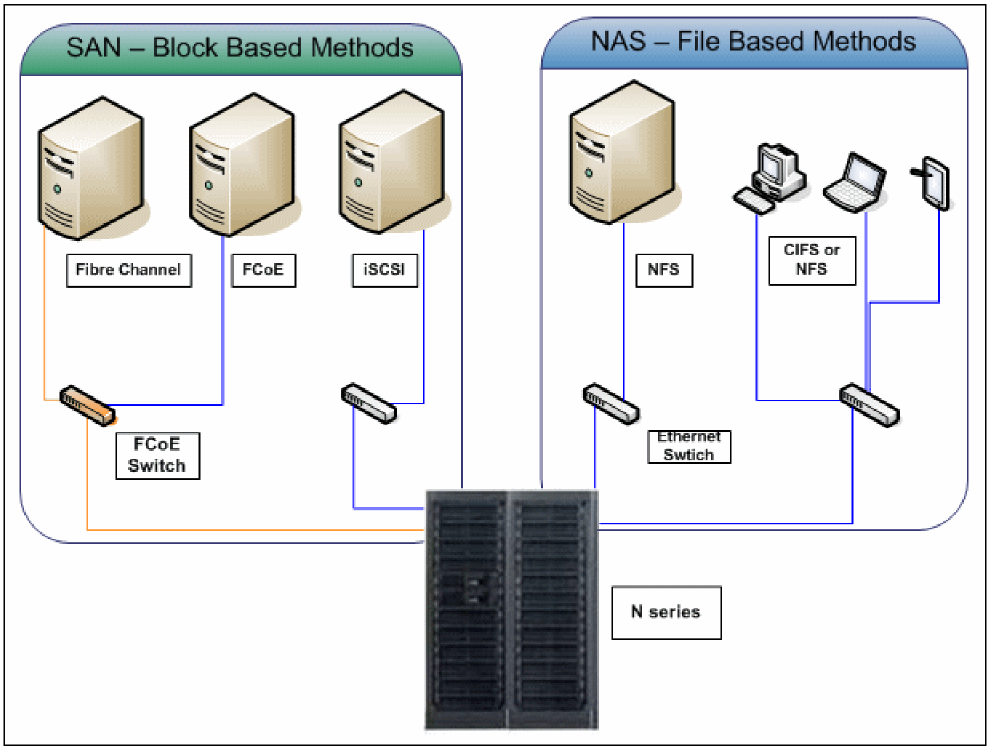


Figure 33-4 Storage protocols used by VMWare and available on N series family

33.2.4 N series LUNs for VMWare host boot

N series storage systems provide a set of features that make the boot from SAN reliable, secure, and cost effective. You can use these features as follows:

- With Snapshot, you can take Snapshots of a logical unit number (LUN) and restore it later. You can use Snapshot restores in a case of a storage failure or for corrupted file systems if necessary to recreate the entire LUN (Figure 33-5).
- With FlexClone, you can clone a LUN and make it available to other servers. This method can be used to deploy multiple ESXi hosts. For example, you can install the ESXi operating system on a single server, then use FlexClone to make a copy of that LUN to multiple servers. This N series feature is also helpful when you want to reproduce your production environment on a test area.

FlexClone functionality is shown in Figure 33-5.

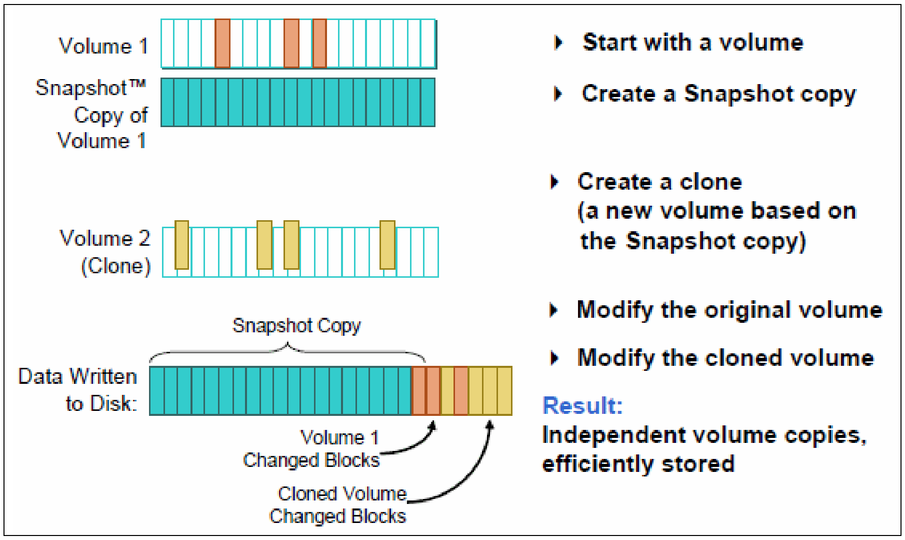

Volume 1
Snapshot™ Copy of Volume 1
Volume 2 (Clone)
Snapshot Copy
Data Written to Disk:
Volume 1 Changed Blocks
Cloned Volume Changed Blocks
Start with a volume
Create a Snapshot copy
Create a clone (a new volume based on the Snapshot copy)
Modify the original volume
Modify the cloned volume
Result:
Independent volume copies, efficiently stored

Figure 33-5 Flexclone cloning and space savings

> **Customizing the ESXi operating system:** After using FlexClone, the ESXi operating system must to be customized to avoid IP and name conflicts with the original server from which the FlexClone was taken.

33.2.5 N series LUNs for VMFS datastores

Including many hard drives in the aggregate provides improved performance for LUNs created over them. As a preferred practice, ensure that each LUN is used by a single datastore, thus making them easier to manage.

Similar backup and recovery requirements provide a good criteria when deciding which servers must share the same datastores. Consider having very important servers on their own datastore, so you can take full advantage of N series advanced functionalities, which are implemented on the volume level.

33.2.6 Using N series LUNs for Raw Device Mappings

Using Raw Device Mappings (RDM) with VMware ESXi offers the following benefits:

- Mapping file references to persistent names
- Unique ID for each mapped device
- Distributed locking for raw SCSI devices
- File permission enablement
- Redo log tracking for a mapped device
- Virtual machine migration with vMotion
- Use of file system utilities
- SAN management within a virtual machine

The N series can facilitate these benefits by providing virtual LUNs though flexible volumes (Figure 33-6).

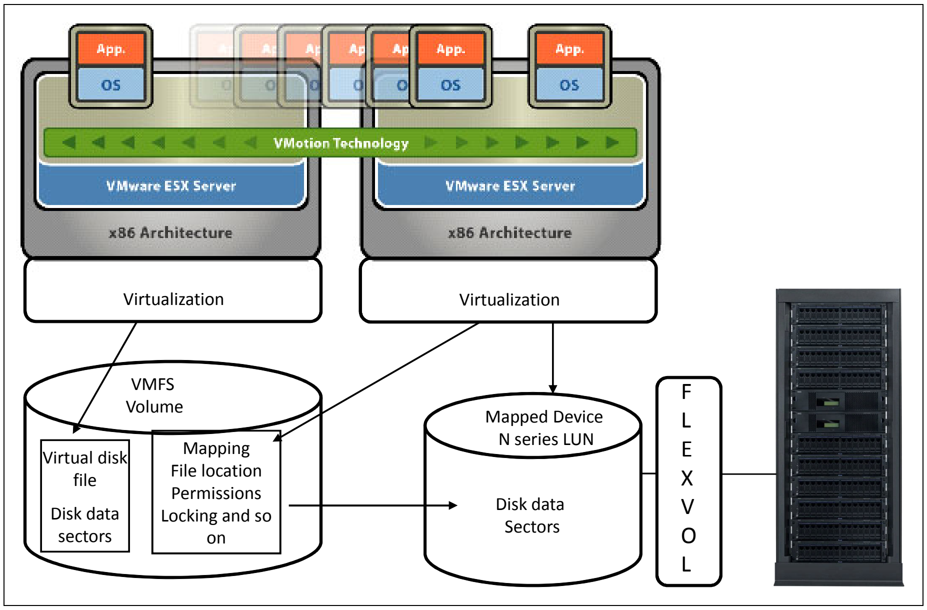


Figure 33-6 Mapping file data

33.2.7 Growing VMFS datastores

You can easily increase the storage for a Virtual Machine File System (VMFS) datastore by increasing the size of the N series LUN. Then you add an extent on the VMware ESX Server. However, you must complete this process only when all virtual machines stored on the datastore are shut down.

33.2.8 Backup and recovery of the virtual infrastructure (SnapVault, Snapshot, SnapMirror)

The use of N series functions, such as Snapshot, allow for fast backup of a whole disk volume without using much additional disk space. The backup can then be written to tape or mirrored to auxiliary storage at the same or different location.

Recovery of a disk volume from Snapshot is fast, because the volume is quickly replaced with the Snapshot. If less data is required for restoration, such as a single file or a guest virtual machine disk (files with .vmdk extension), then the restore depends on the backup strategy:

- If *Snapshot* is used, a clone of the Snapshot can be created and just the required files can be copied back manually. For a guest, the cloned volume can be mounted by VMware and the required guests can be registered and started.
- If backup was to *tape*, a restore of the required files is performed.
- If a *mirror* exists, the required files can also be copied back manually.

It is important to note that if no other tool is implemented and a volume backup is taken, only the entire volume can be restored. To overcome that limitation, IBM offers the IBM Tivoli Storage Manager product. This product interacts with VMWare vSphere APIs for Data Protection, formerly known as Virtual Consolidated Backup (VCB) on earlier VMWare versions. When used together, these products can restore on the image, volume, and file levels from a single backup.

For more information, see the following website:

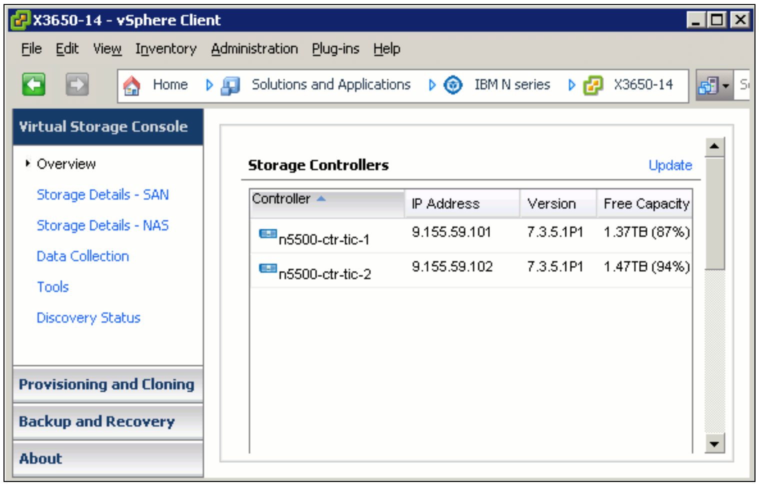
http://www.ibm.com/developerworks/wikis/display/tivolistoragemanager/IBM+Tivoli+Storage+Manager+for+Virtual+Environments

33.3 Virtual Storage Console

The Virtual Storage Console (VSC) software is a single vCenter Server plug-in. It provides end-to-end virtual machine lifecycle management for VMware environments running N series storage. The plug-in provides these features:

- Storage configuration and monitoring, using the Monitoring and Host Configuration capability (previously called the Virtual Storage Console capability)
- Datastore provisioning and virtual machine cloning, using the Provisioning and Cloning capability
- Backup and recovery of virtual machines and datastores, using the Backup and Recovery capability

As a vCenter Server plug-in, shown in Figure 33-7, the VSC is available to all vSphere Clients that connect to the vCenter Server. This availability is different from a client-side plug-in that must be installed on every vSphere Client. You can install the VSC software on a Windows server in your data center, but you must not install it on a client computer.



Storage Controllers

Controller	IP Address	Version	Free Capacity
n5500-ctr-tic-1	9.155.59.101	7.3.5.1P1	1.37TB (87%)
n5500-ctr-tic-2	9.155.59.102	7.3.5.1P1	1.47TB (94%)

Figure 33-7 Virtual Storage Console 2

Virtual Storage Console (VSC) integrates VSC storage discovery, health monitoring, capacity management, and preferred practice-based storage setting. It offers additional management capabilities with two capability options in a single vSphere client plug-in. Thus it enables centralized, end-to-end management of virtual server and desktop environments running on N series storage. VSC is composed of three main components:

- Virtual Storage Console Capability (base product): Provides a storage view of the VMware environment with a VM administrator perspective. It automatically optimizes the customer’s host and storage configurations, including HBA timeouts, NFS tunables, and multipath configurations. Using the Virtual Storage Console, a VM administrator can quickly and easily view controller status and capacity information. Also, the administrator can accurately report back utilization information in order to make more informed decisions about VM object placement.

- Provisioning and Cloning Capability: Provides end-to-end datastore management (provisioning, resizing, and deletion). Also offers rapid, space-efficient VM server and desktop cloning, patching, and updating by using FlexClone technology.
- Backup and Recovery capability (formerly SnapManager for Virtual Infrastructure): Automates data protection processes by enabling VMware administrators to centrally manage backup and recovery of datastores and VMs. It can be done without impacting guest performance. The administrator can also rapidly recover from these backup copies at any level of granularity: datastore, VM, VMDK, or guest file.

VSC is designed to simplify storage management operations, improve efficiencies, enhance availability, and reduce storage costs in both SAN- and NAS-based VMware infrastructures. It provides VMware administrators with a window into the storage domain. It also provides the tools to effectively and efficiently manage the lifecycle of virtual server and desktop environments running on N series storage.

33.4 Using N series deduplication with VMware

Deduplication is the concept of storing multiple instances of the same information into a single point. Then a pointer is used to refer to it on the next occurrence, so files that potentially might be stored in an environment many times are stored only once. Microsoft Exchange and Symantec Vault are commercial products known for the usage of deduplication.

N series deduplication provides Advanced Single Instance Storage (A-SIS) at the storage level, rather than the application level. Doing this significantly reduces the amount of storage that is used when the same files are stored multiple times. The deduplication process is shown in Figure 33-8.

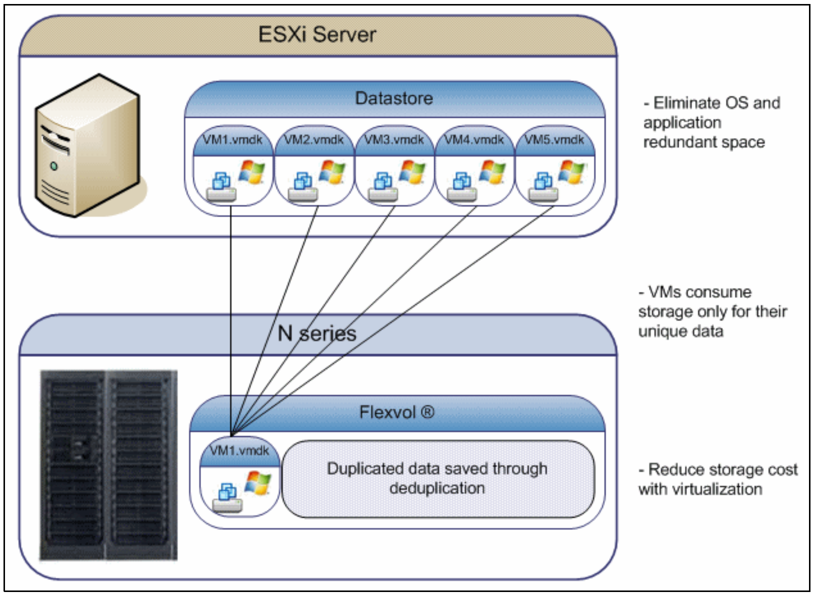

ESXi Server
Datastore
VM1.vmdk
VM2.vmdk
VM3.vmdk
VM4.vmdk
VM5.vmdk
- Eliminate OS and application redundant space
- VMs consume storage only for their unique data
N series
Flexvol ®
VM1.vmdk
Duplicated data saved through deduplication
- Reduce storage cost with virtualization

Figure 33-8 Storage Consumption with N series A-SIS

33.5 Coupling deduplication and compression

You can further increase savings by using N series deduplication and compression with the IBM Real-time Compression solution. Compression, which has been around for several years, has not met the strict IT demands for primary storage until now. To solve primary storage capacity optimization, vendors need to ensure data integrity and availability, without impacting performance or forcing IT to change their applications or process.

The IBM Real-time Compression technology meets these requirements with its Random Access Compression Engine (RACE), through an appliance called Real Time Compression Appliance (RTCA). It provides a tremendous reduction in capital and operational costs when it comes to storage management and the additional benefits of less to manage, power, and cool. Additionally, similar to server virtualization, IBM Real-time Compression fits seamlessly into your storage infrastructure. It is done without requiring changes to any processes and offering significant savings throughout the entire data life cycle.

IBM Real-time Compression provides data compression solutions for primary storage, enabling companies to dramatically increase storage efficiencies. IBM Real-time Compression provides the following benefits:

- Up to 80% of data footprint reduction.
- Resource savings: Compressing data at the origin triggers a cascading effect of multiple savings across the entire information life cycle. As less data is initially written to storage, it results in these improvements:
  - There is a reduction in storage CPU and disk utilization.
  - Effective storage cache size increases in proportion to the compression ratio and enables higher performance.
  - Snapshots, replications, and backup and restore-related operations all benefit from the data reduction and perform better.
- Transparency: No configuration changes are required on the storage, networks, or applications. The IBM Real-time Compression system is agnostic to both data types and storage systems.
- Simplicity: IBM Real-time Compression Plug and Play real-time data compression appliances are simple to deploy, with a typical installation taking no more than 30 minutes.

For more details on integration of vSphere 4.x environments with the IBM Real-time Compression Appliance (RTCA), see the IBM Redbooks publication: *Introduction to IBM Real-time Compression Appliances*, SG24-7953. It is located at the following website:

http://www.redbooks.ibm.com/abstracts/sg247953.html?Open

33.6 Further reading

For more information about the various possibilities and options available for the N series storage system with VMWare vSphere, see the following Redbooks publications:

- *IBM System Storage N series and VMware vSphere Storage Best Practices,* SG24-7871:

  http://www.redbooks.ibm.com/abstracts/sg247871.html?Open

- *IBM System Storage N series with VMware vSphere 4.1*, SG24-7636:

  http://www.redbooks.ibm.com/abstracts/sg247636.html?Open

- *IBM System Storage N series with VMware vSphere 4.1 using Virtual Storage Console 2*, REDP-4863:

  http://www.redbooks.ibm.com/abstracts/redp4863.html?Open

34

Virtual Storage Console 4.1

The ability to quickly back up tens of hundreds of virtual machines without affecting production operations can accelerate the adoption of VMware within an organization, as explained in this chapter.

The following topics are covered:

- Virtual Storage Console
- Installing the Virtual Storage Console 4.1
- Adding storage controllers to the VSC
- Optimal storage settings for ESXi host
- SnapMirror integration
- VSC in an N series MetroCluster environment
- Backup and recovery
- Provisioning and cloning
- Optimum VM availability
- VSC commands
- Scripting

34.1 Virtual Storage Console

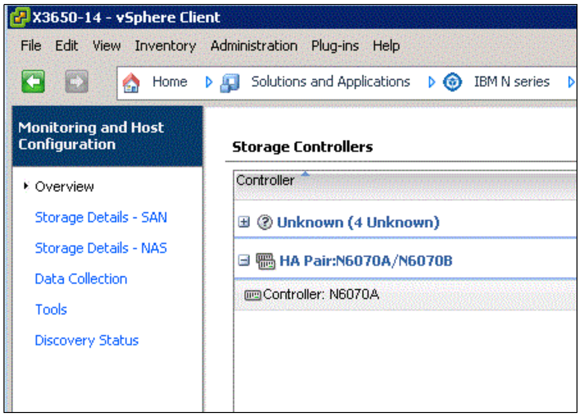
The Virtual Storage Console (VSC) feature was formerly provided in a separate interface and was called SnapManager for Virtual Infrastructure (SMVI). It builds on the N series SnapManager portfolio by providing array-based backups. They consume only block-level changes to each VM and can provide multiple recovery points throughout the day. The backups are an integrated component within the storage array. Therefore, VSC provides recovery times that are faster than times provided by any other means.

34.1.1 Introduction to the Virtual Storage Console

The Virtual Storage Console (VSC) software is a single vCenter Server plug-in. It provides end-to-end virtual machine lifecycle management for VMware environments running N series storage. The plug-in provides these features:

- Storage configuration and monitoring, using the Monitoring and Host Configuration capability (previously called the Virtual Storage Console capability)
- Datastore provisioning and virtual machine cloning, using the Provisioning and Cloning capability
- Backup and recovery of virtual machines and datastores, using the Backup and Recovery capability

As a vCenter Server plug-in, shown in Figure 34-1, the VSC is available to all vSphere Clients that connect to the vCenter Server. This availability is different from a client-side plug-in that must be installed on every vSphere Client. You can install the VSC software on a Windows server in your data center, but you must not install it on a client computer.



Storage Controllers

Controller
⊞ Unknown (4 Unknown)
⊟ HA Pair:N6070A/N6070B
Controller: N6070A

Figure 34-1 Virtual Storage Console

Virtual Storage Console (VSC) integrates VSC storage discovery, health monitoring, capacity management, and preferred practice-based storage setting. It offers additional management capabilities with two capability options in a single vSphere client plug-in. Thus it enables centralized, end-to-end management of virtual server and desktop environments running on N series storage. VSC is composed of three main components:

- Virtual Storage Console Capability (base product): Provides a storage view of the VMware environment with a VM administrator perspective. It automatically optimizes the customer's host and storage configurations, including HBA timeouts, NFS tunables, and multipath configurations. Using the Virtual Storage Console, a VM administrator can quickly and easily view controller status and capacity information. Also, the administrator can accurately report back utilization information in order to make more informed decisions about VM object placement.
- Provisioning and Cloning Capability: Provides end-to-end datastore management (provisioning, resizing, and deletion). Also offers rapid, space-efficient VM server and desktop cloning, patching, and updating by using FlexClone technology.
- Backup and Recovery capability (formerly SnapManager for Virtual Infrastructure): Automates data protection processes by enabling VMware administrators to centrally manage backup and recovery of datastores and VMs. This can be done without impacting guest performance. The administrator can also rapidly recover from these backup copies at any level of granularity: datastore, VM, VMDK, or guest file.

VSC is designed to simplify storage management operations, improve efficiencies, enhance availability, and reduce storage costs in both SAN- and NAS-based VMware infrastructures. It provides VMware administrators with a window into the storage domain. It also provides the tools to effectively and efficiently manage the lifecycle of virtual server and desktop environments running on N series storage.

34.1.2 License requirements

Table 34-1 summarizes the N series license requirements to perform different VSC functions.

Table 34-1 VSC license requirements

Task	License
Provision datastores	NFS, FCP, iSCSI
Restore datastores	SnapRestore
Use vFilers in Provisioning and Cloning operations	MultiStore
Clone virtual machines	FlexClone (NFS only)
Configure deduplication settings	A-SIS
Distribute templates to remote vCenters	SnapMirror

34.1.3 Architecture overview

Figure 34-2 illustrates the architecture for VSC. It also shows the components that work together to provide a comprehensive and powerful backup and recovery solution for VMware vSphere environments.

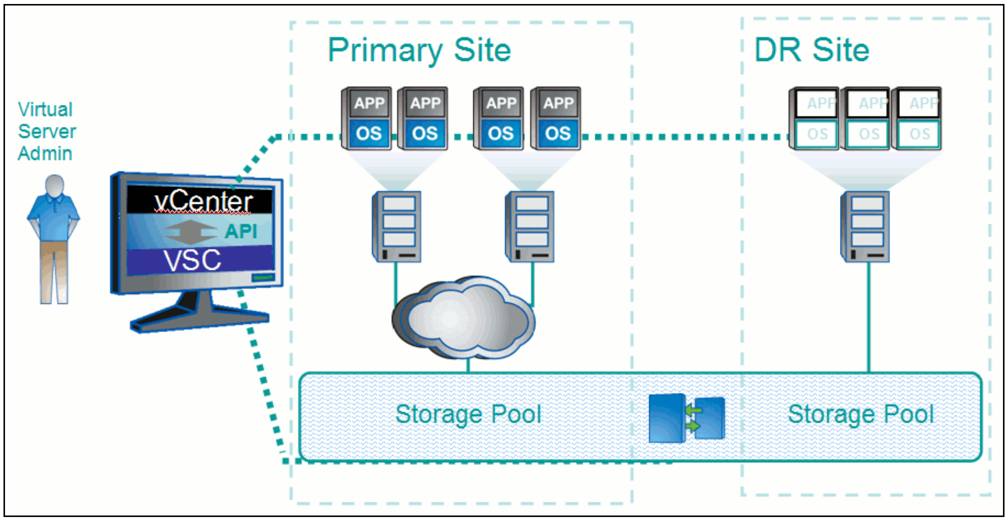


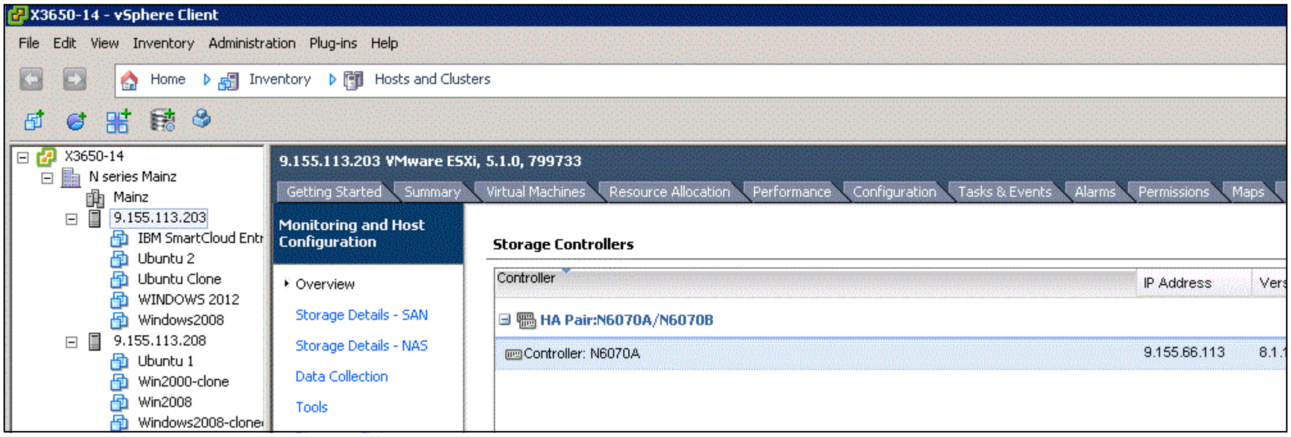
Figure 34-2 Architecture overview

34.1.4 Monitoring and host configuration

The Monitoring and Host Configuration capability enables you to manage ESXi servers connected to N series storage systems. You can set host timeout, NAS, and multipathing values, view storage details, and collect diagnostic data. You can use this capability to do the following tasks:

- View the status of storage controllers from a SAN (FC, FCoE, and iSCSI) perspective
- View the status of storage controllers from a NAS (NFS) perspective
- View SAN and NAS datastore capacity utilization
- View the status of VMware vStorage APIs for Array Integration (VAAI) support in the storage controller
- View the status of ESX hosts, including ESX version and overall status
- Check at a glance whether the following settings are configured correctly, and if not, automatically set the correct values:
  - Storage adapter timeouts
  - Multipathing settings
  - NFS settings
- Set credentials to access storage controllers
- Launch the vCenter GUI to create LUNs and manage storage controllers
- Collect diagnostic information from the ESXi hosts, storage controllers, and Fibre Channel switches
- Access tools to set guest operating system timeouts and to identify and correct misaligned disk partitions

When you click the N series tab in the vCenter Server and click Monitoring and Host Configuration in the navigation pane, the Overview panel displays. It is similar to Figure 34-3.



Controller	IP Address	Vers
HA Pair:N6070A/N6070B		
Controller: N6070A	9.155.66.113	8.1.1

Figure 34-3 VSC overview

Alternatively, you can find the VSC plug-in under Solutions and Applications (Figure 34-4).

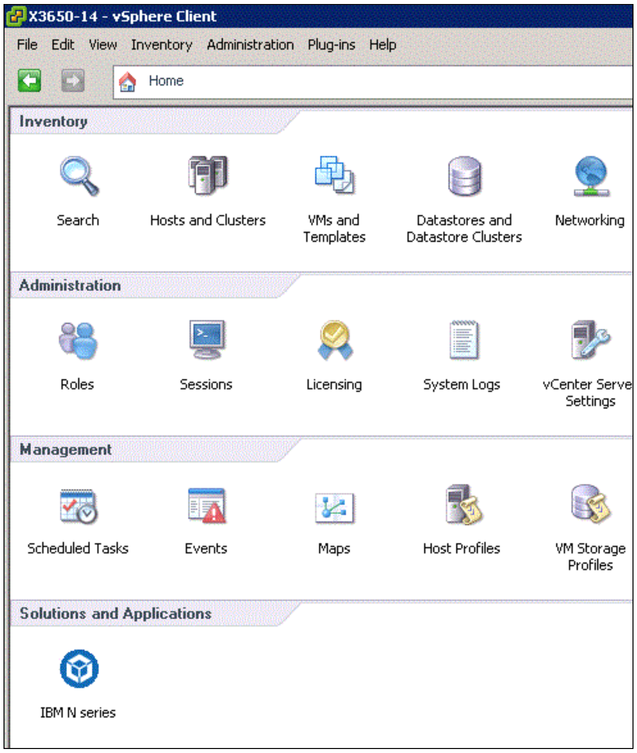


Figure 34-4 VSC location

34.1.5 Provisioning and cloning

The Provisioning and Cloning capability of Virtual Storage Console helps you to provision datastores and quickly create multiple clones of virtual machines in the VMware environment. Using FlexClone technology, the Provisioning and Cloning capability allows you to efficiently create, deploy, and manage the lifecycle of virtual machines. These tasks can be done from an easy-to-use interface integrated into the VMware environment. It is ideal for virtual server, desktop, and cloud environments.

You can use the provisioning and cloning capability for the following purposes:

- Clone individual virtual machines and place in new or existing datastores
- Create, resize, or delete datastores
- Apply guest customization specifications and power up new virtual machines
- Run deduplication operations
- Monitor storage savings
- Redeploy virtual machines from a baseline image
- Replicate NFS datastores across sites
- Import virtual machines into virtual desktop infrastructure connection brokers and management tools

Managing datastores and cloning virtual machines

To manage datastores and clone virtual machines, right-click an object in the Inventory panel of the vSphere Client and select **IBM N series** → **Provisioning and Cloning** (Figure 34-5):

- Right-click a powered-down virtual machine or template to create clones.
- Right-click a datacenter, cluster, or host to provision datastores.

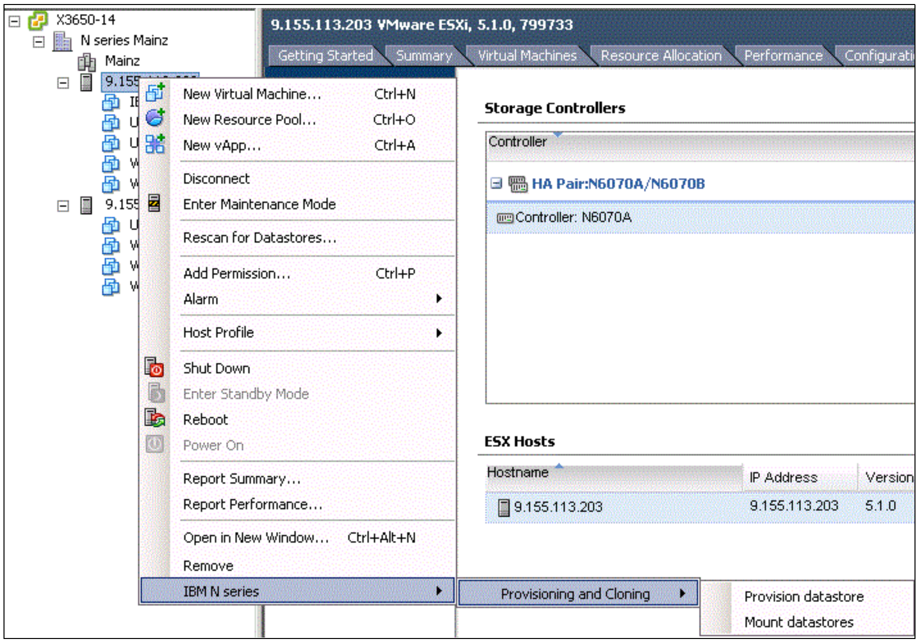


Figure 34-5 Accessing Provisioning and Cloning

Managing controllers, replicating datastores, and redeploying clones

Click the Inventory button in the navigation bar, and then select **Solutions and Applications** → **IBM N series** → **Provisioning and Cloning**. Use the following options:

- Select **Storage controllers** to add, remove, or modify properties of storage controllers.
- Select **Connection brokers** to add and remove connection broker definitions.
- Select **DS Remote Replication** to clone NFS datastore templates to multiple target sites.
- Select **Redeploy** to redeploy virtual machines.

34.2 Installing the Virtual Storage Console 4.1

The VSC provides full support for hosts running ESX/ESXi 4.0 and later. It provides limited reporting functionality with hosts running ESX/ESXi 3.5 and later.

34.2.1 Basic installation

Before downloading and installing the VSC, make sure that your deployment has the required components:

- You need a vCenter Server version 5.0 or later. The VSC can be installed on the vCenter Server or on another server or VM (see Figure 34-6).
- If installing on another server or VM, this system must run 32-bit or 64-bit Windows Server 2008, 2003 SP2 and later.
- A storage array is required to run Data ONTAP 7.3.1.1 or later.

Attention: Before installing, verify supported storage adapters and firmware.

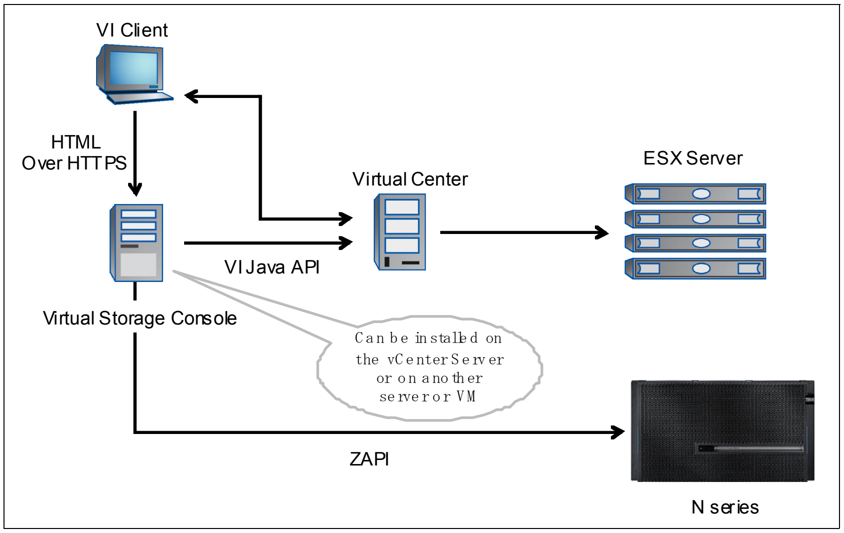


Figure 34-6 VSC possible deployments

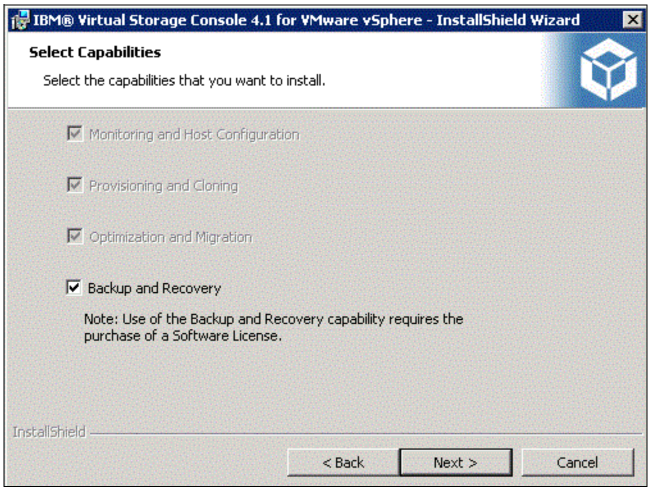
Tip: To keep it simple, we suggest installing the VSC on the vCenter server.

Complete the following steps to install the VSC 4.1:

1. Download the installation program to the Windows server.
2. Run the installation wizard and select the features you would like to install as shown in Figure 34-7.

3. Follow the on-screen instructions.

   During the installation process, a prompt displays to select the features of the VSC 4.1 () to be enabled in the environment. The core VSC must be selected. The Provisioning and Cloning and Backup and Recovery features are the former RCU and the SMVI interfaces. Certain subfeatures might require licensing, as described previously. See Figure 34-7.



IBM® Virtual Storage Console 4.1 for VMware vSphere - InstallShield Wizard

Select Capabilities

Select the capabilities that you want to install.

☑ Monitoring and Host Configuration

☑ Provisioning and Cloning

☑ Optimization and Migration

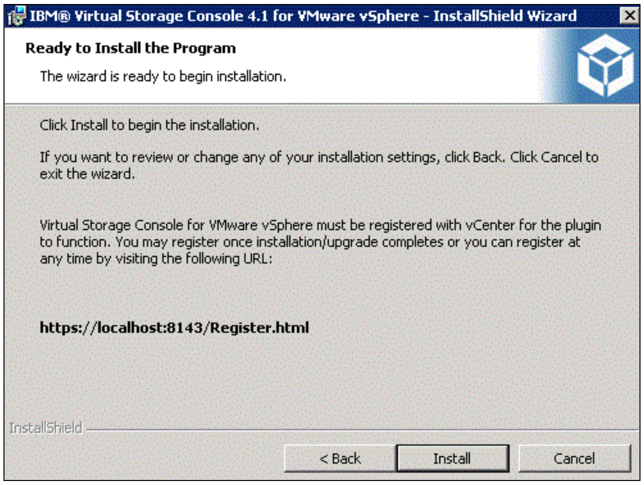
☑ Backup and Recovery

Note: Use of the Backup and Recovery capability requires the purchase of a Software License.

Figure 34-7 Select VSC features

4. Register the VSC as a plug-in, in the vCenter Server in the window that opens when the process is complete.

   The installation process launches the vCenter registration process as shown in Figure 34-8.



IBM® Virtual Storage Console 4.1 for VMware vSphere - InstallShield Wizard

Ready to Install the Program

The wizard is ready to begin installation.

Click Install to begin the installation.

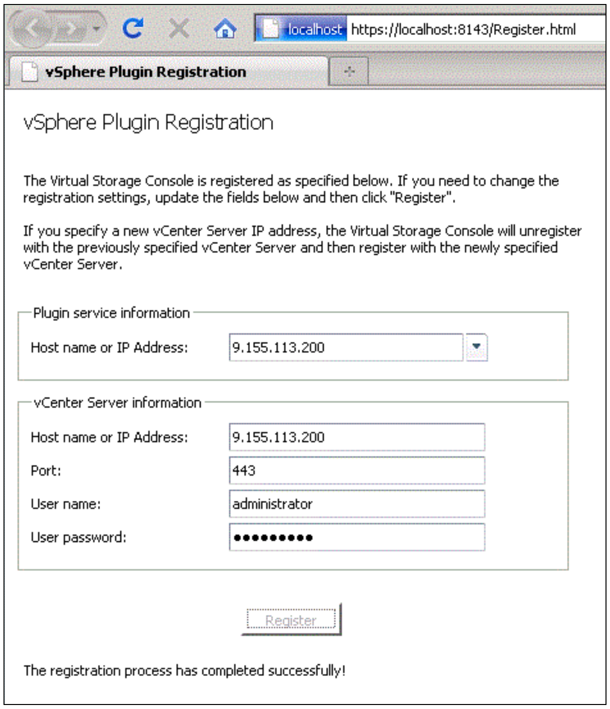
If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.

Virtual Storage Console for VMware vSphere must be registered with vCenter for the plugin to function. You may register once installation/upgrade completes or you can register at any time by visiting the following URL:

https://localhost:8143/Register.html

Figure 34-8 vCenter registration process

5. Finally, register the VSC plug-in with a vCenter server (Figure 34-9). This final step requires a user with vCenter administrator credentials to complete the registration process.


localhost
https://localhost:8143/Register.html
vSphere Plugin Registration

vSphere Plugin Registration

The Virtual Storage Console is registered as specified below. If you need to change the registration settings, update the fields below and then click "Register".

If you specify a new vCenter Server IP address, the Virtual Storage Console will unregister with the previously specified vCenter Server and then register with the newly specified vCenter Server.

Plugin service information

Host name or IP Address: 9.155.113.200

vCenter Server information

Host name or IP Address: 9.155.113.200

Port: 443

User name: administrator

User password: ●●●●●●●●●

The registration process has completed successfully!

Figure 34-9 VSC registration with vCenter server

34.2.2 Registration completion

Upon successful registration, the system confirms by issuing the following message on the web page: `The registration process has completed successfully!`

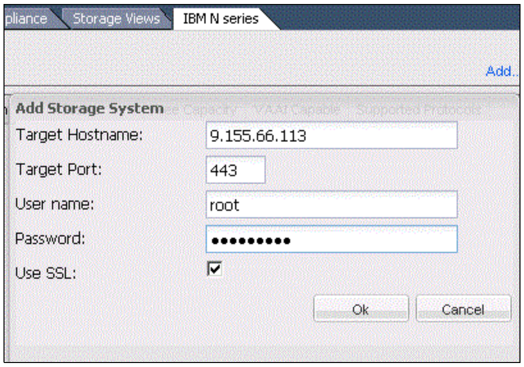
34.3 Adding storage controllers to the VSC

Adding the storage controllers that host the virtual infrastructure to the VSC is fairly simple:

1. Connect to vCenter by using the vSphere client.
2. Double-click the IBM N series tab on the home panel.
3. Select the Virtual Storage Console tab on the left.

After these steps are completed, the VSC launches and automatically identifies all storage controllers powered by Data ONTAP with the storage connected to the ESXi host in the environment. As an alternative to running discovery for the entire environment, you can select an ESXi host or cluster in the vSphere client and then select the IBM N series tab in the left panel. The VSC then begins discovery of all storage controllers with storage connected to the host or cluster that was selected.

The window pops up, as displayed in Figure 34-10, allowing you to enter the user or service account assigned for VSC management on the storage controller. This account can be the root account or one created specifically for the VSC core feature, as described previously.



Add Storage System

Target Hostname: 9.155.66.113

Target Port: 443

User name: root

Password: •••••••••

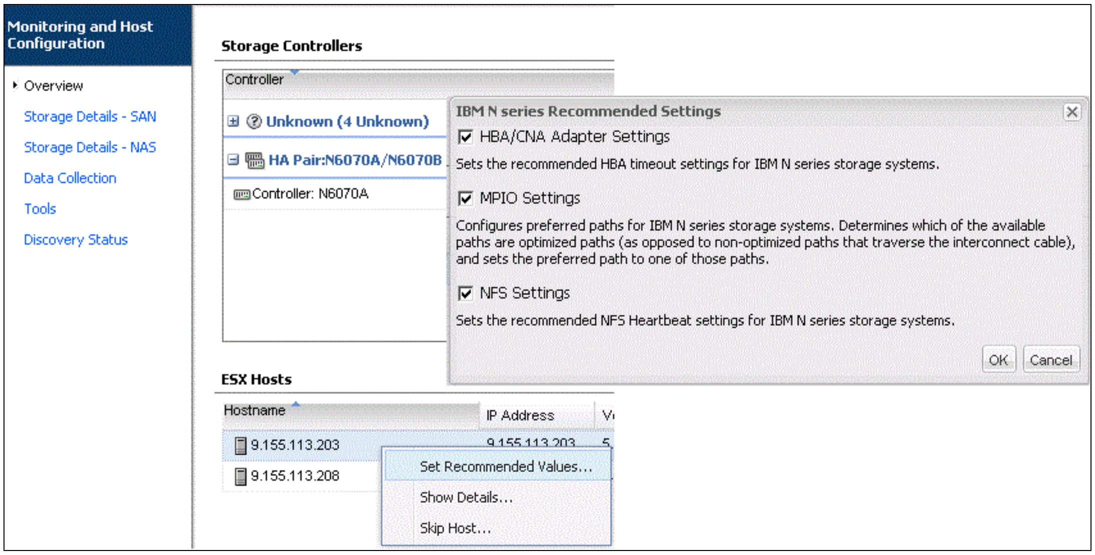
Use SSL: ☑

Figure 34-10 Adding storage controller access in VSC

34.4 Optimal storage settings for ESXi host

The VSC enables the automated configuration of storage-related settings for all ESXi 5.x hosts connected to N series storage controllers. VMware administrators can right-click individual or multiple ESXi host and set the preferred values for these hosts. This functionality sets values for HBAs and CNAs, sets appropriate paths and path selection plug-ins, and provides appropriate settings for software-based I/O (NFS and iSCSI).

To perform the setting, go to the VSC pane, right-click the designated ESXi server, and run the settings as shown in Figure 34-11.



IBM N series Recommended Settings

☑ HBA/CNA Adapter Settings

Sets the recommended HBA timeout settings for IBM N series storage systems.

☑ MPIO Settings

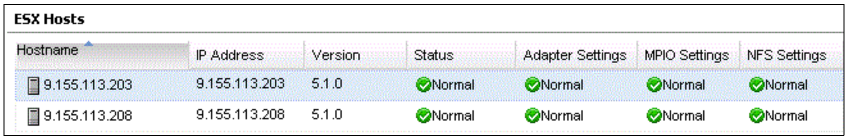
Configures preferred paths for IBM N series storage systems. Determines which of the available paths are optimized paths (as opposed to non-optimized paths that traverse the interconnect cable), and sets the preferred path to one of those paths.

☑ NFS Settings

Sets the recommended NFS Heartbeat settings for IBM N series storage systems.

Figure 34-11 Optimize ESX settings

After rebooting the ESX server, we can verify the improved settings. All status indicators are green (see Figure 34-12).



ESX Hosts

Hostname	IP Address	Version	Status	Adapter Settings	MPIO Settings	NFS Settings
9.155.113.203	9.155.113.203	5.1.0	✓Normal	✓Normal	✓Normal	✓Normal
9.155.113.208	9.155.113.208	5.1.0	✓Normal	✓Normal	✓Normal	✓Normal

Figure 34-12 Optimized ESX adapter settings

34.5 SnapMirror integration

SnapMirror relationships cannot be configured through VSC. However, VSC can update an existing SnapMirror relationship on the volume underlying the datastore or virtual machine. Preferably, test the SnapMirror relationship from the storage system command line before updating through VSC. This method aids in identifying where any potential issues might occur. If the SnapMirror update is successful from the CLI, but fails from within VSC, the administrator has a better understanding of where to concentrate troubleshooting efforts.

Also, identify the destination storage within VSC in the same manner that the relationship is configured on the storage system. For example, if a SnapMirror relationship is configured on the storage system using IP addresses rather than a DNS name, identify the auxiliary storage to VSC by the IP address and vice versa.

Because its support is for SnapMirror volume only, map one volume per datastore.

During backup creation, SnapManager provides the option of updating an existing SnapMirror relationship. That way, every time a Snapshot is created, the data is transferred to a remote storage system. Whenever the backup of a virtual machine or datastore is initiated with the SnapMirror option, the update starts as soon as the backup completes, after of the current SnapMirror schedule.

For example, by configuring regular SnapMirror updates on a filter after the VSC schedule, you can cut down the time required to update the mirror, because it is done in the interim. However, keep in mind that the updates must be scheduled in such a way that they do not conflict with the SnapManager backup.

34.5.1 SnapMirror destinations

A single SnapMirror destination is supported per volume. If a SnapMirror update is selected as part of a backup on a volume with multiple destinations, the backup fails.

If multiple SnapMirror destinations are required, use a tiered approach when configuring the SnapMirror relationships. For example, if the data must be transferred to four destinations, configure one destination from the primary storage system supported to one destination. Then configure three additional destinations from the auxiliary storage through the storage system CLI.

34.5.2 SnapMirror and deduplication

Preferably, do not use deduplication with Sync SnapMirror. Although technically it works, the integration and scheduling of deduplication with Sync SnapMirror are complicated to implement in the type of rigorous real-world scenarios that demand synchronous replication.

When configuring volume SnapMirror and deduplication, consider the deduplication schedule and the volume SnapMirror schedule. Start volume SnapMirror transfers of a deduplicated volume after deduplication completes (that is, not during the deduplication process). This technique avoids sending undeduplicated data and additional temporary metadata files over the network. If the temporary metadata files in the source volume are locked in Snapshot copies, they also consume extra space in the source and destination volumes. Volume SnapMirror performance degradation can increase with deduplicated volumes.

The scenario described previously has a direct impact on backups configured within VSC when the SnapMirror update option was selected. Avoid scheduling a backup with the SnapMirror update option until a a confirmation of the volume deduplication completeness. Although a few hours must be scheduled to ensure avoiding this issue, the actual scheduling configuration is data and customer dependent.

34.6 VSC in an N series MetroCluster environment

N series MetroCluster configurations consist of a pair of active-active storage controllers. They are configured with mirrored aggregates and extended distance capabilities to create a high-availability solution. This type of configuration has the following benefits:

- Higher availability with geographic protection
- Minimal risk of lost data, easier management and recovery, and reduced system downtime
- Quicker recovery when a disaster occurs
- Minimal disruption to users and client applications

A MetroCluster (either Stretch or Fabric) behaves in most ways similar to an active-active configuration. All of the protection provided by core N series technology (RAID-DP, Snapshot copies, automatic controller failover) also exists in a MetroCluster configuration. However, MetroCluster adds complete synchronous mirroring along with the ability to perform a complete site failover from a storage perspective with a single command.

The following N series MetroCluster types exist and work seamlessly with the complete VMware vSphere and ESX server portfolio:

- *Stretch MetroCluster* (sometimes called a *nonswitched cluster*) is an active-active configuration that can extend up to 500 m depending on speed and cable type. It includes synchronous mirroring (SyncMirror) and the ability to do a site failover with a single command.
- *Fabric MetroCluster* (also called a *switched cluster*) uses four Fibre Channel switches in a dual-fabric configuration. It uses a separate cluster interconnect card to achieve an even greater distance (up to 100 km depending on speed and cable type) between primary and secondary locations.

The integration of the MetroCluster and VMware vSphere is seamless and provides storage and application redundancy. In addition to connecting to the vSphere environment using FCP, iSCSI, or NFS, this solution can serve other network clients with CIFS, HTTP, and FTP at the same time.

The solution shown in Figure 34-13 provides a redundant VMware server, redundant N series heads, and redundant storage.

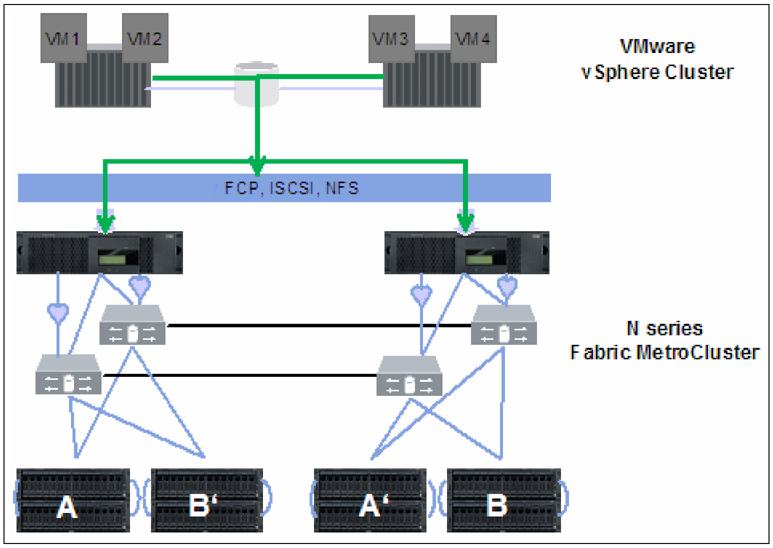


Figure 34-13 MetroCluster and VMware vSphere integrated solution

For more information about N series MetroCluster, see the "MetroCluster" chapter in the Redbooks publication, *IBM System Storage N series Software Guide*, SG24-7129.

34.7 Backup and recovery

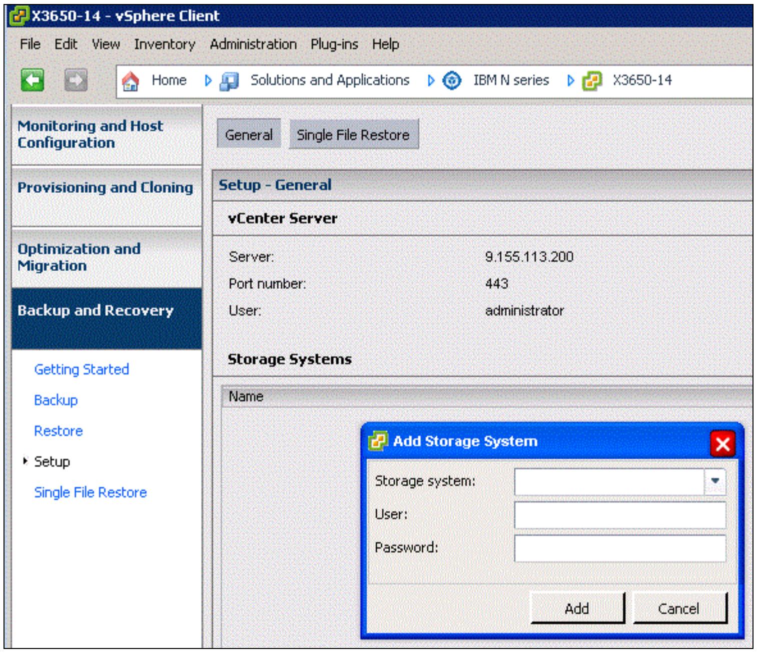
This section provides examples of backing up a single virtual machine or the entire DataCenter. The Backup and Recovery capability of the Virtual Storage Console provides rapid backup and recovery of multi-host configurations running on N series storage systems.

You can use this capability to do the following tasks:

- Perform on-demand backups of individual virtual machines, datastores, or a datacenter
- Schedule automated backups of individual virtual machines, datastores, or a datacenter
- Support virtual machines and datastores that are located on either NFS directories or VMFS file systems
- Mount a backup to verify its content prior to restoration
- Restore datastores or virtual machines to the original location
- Restore virtual machine disks (VMDKs) to the original or an alternate location
- Restore one or more files to a guest VMDK without having to restore the entire virtual machine or VMDK using single file restore feature

To configure your storage systems, click the N series icon in the vCenter Server and click **Setup** under Backup and Recovery in the navigation pane. The Setup panel displays. Click **Add** on the left side and register your N series system as shown in Figure 34-14.

> **Important:** You must register your N series system two times; first, for the VSC and second, for backup and recovery.



Server: 9.155.113.200
Port number: 443
User: administrator

Figure 34-14 N series registration for backup and restore

34.7.1 Data layout

Layout is indicated by N series preferred practices for vSphere environments. Move any transient and temporary data, such as the guest operating system swap file, temp files, and page files, to a separate virtual disk on another datastore. The reason is that snapshots of this data type can consume a large amount of storage in a short time because of the high rate of change.

When a backup is created for a virtual machine with VSC, VSC is aware of all VMDKs associated with the virtual machine. VSC initiates a Snapshot copy on all datastores upon which the VMDKs reside. For example, a virtual machine running Windows as the guest operating system has its C drive on datastore ds1, data on datastore ds2, and transient data on datastore td1. In this case, VSC creates a Snapshot copy against all three datastores at underlying volume level. It defeats the purpose of separating temporary and transient data.

Considerations for transient and temporary data

To exclude the datastore that contains the transient and temporary data from the VSC backup, configure the VMDKs residing in the datastore as “Independent Persistent” disks within the VMware Virtual Center (vCenter). After the transient and temporary data VMDKs are configured, they are excluded from both the VMware Virtual Center snapshot and the N series Snapshot copy initiated by VSC.

You must also create a datastore dedicated to transient and temporary data for all virtual machines with no other data types or virtual disks residing on it. This datastore avoids having a Snapshot copy taken against the underlying volume as part of the backup of another virtual machine. Do not deduplicate the data on this datastore.

SnapManager 2.0 for Virtual Infrastructure can include independent disks and exclude datastores from backup.

Including independent disks and excluding datastores

You can avoid having a Snapshot copy performed on the underlying volume as part of the backup of another virtual machine. In this case, preferably, create a datastore that is dedicated to transient and temporary data for all virtual machines. Exclude datastores that contain transient and temporary data from the backup. By excluding those datastores, snapshot space is not wasted on transient data with a high rate of change. In VSC 4.1, when selected entities in the backup span multiple datastores, one or more of the spanning datastores might be excluded from the backup.

After configuration, the transient and temporary data .vmdk are excluded from both the VMware vCenter Snapshot and the N series Snapshot copy initiated by VSC. In VSC 1.0, datastores with only independent disks were excluded from the backup. In VSC 4.1, an option is available to include them in the backup. Datastores with a mix of independent disks and normal disks or configuration files for a VM are included in the backup irrespective of this option.

If you have a normal disk and an independent disk for backup on the same datastore, it is always included for backup irrespective of the “include datastore with independent disk” option. Designate a separate datastore exclusively for swap data.

Restore from backup: If you exclude non-independent disks from the backup of a VM, that VM cannot be completely restored. You can perform only virtual disk restore and single file restore from such a backup.

34.7.2 Backup and recovery requirements

Your datastore and virtual machines must meet the following requirements before you can use the Backup and Recovery capability:

- In NFS environments, a FlexClone license is required to mount a datastore, restore guest files, and restore a VMDK to an alternate location.
- Snapshot protection is enabled in the volumes where those datastore and virtual machine images reside.
- SnapRestore is licensed for the storage systems where those datastore and virtual machine images reside.

34.7.3 Single wizard for creating backup jobs

With the wizard, you can create manual and scheduled backup jobs. In the right pane, you click **Backup**, name your new backup job, and select the per-backup job options:

- Initiate SnapMirror update.
- Perform VMware consistency snapshot.
- Include datastores with independent disks.

Virtual Machine backup

To back up individual VMs, follow these steps:

1. Right-click the **VM Backup** and drill down until you reach the selection to run or schedule a backup, as shown in Figure 34-15.

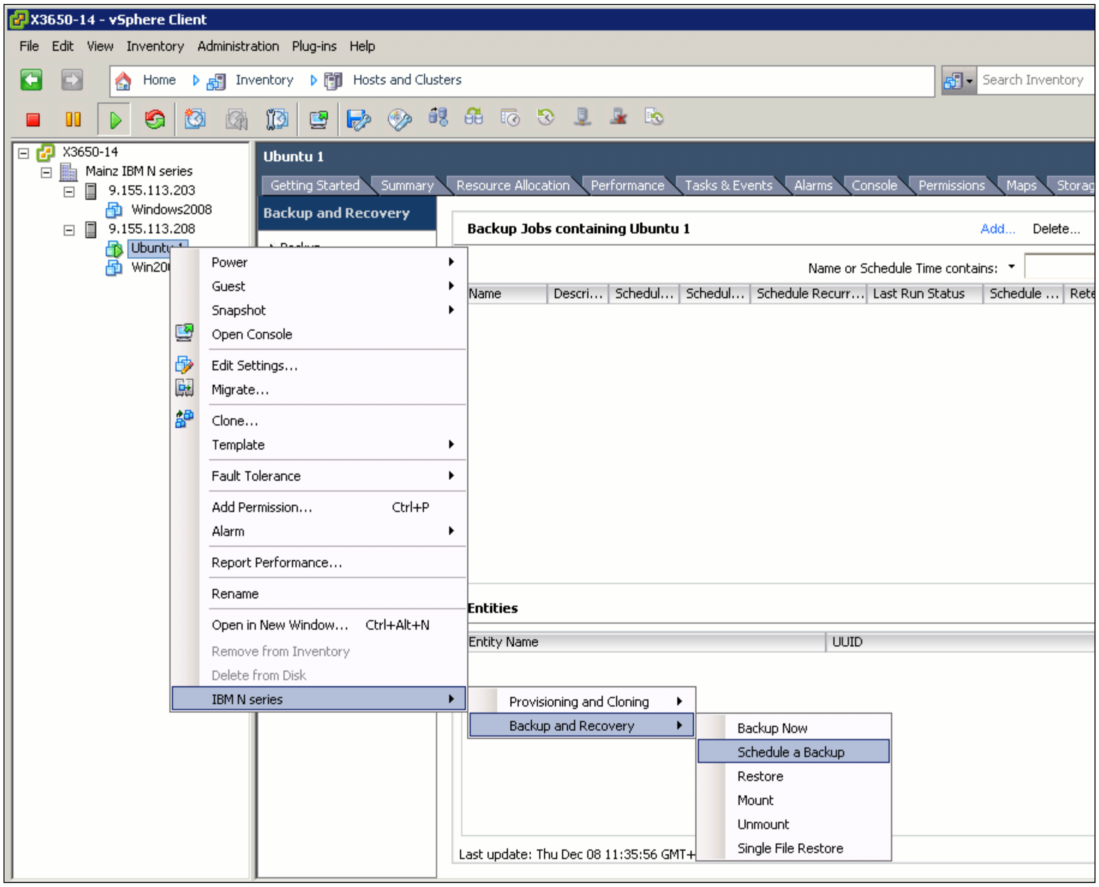
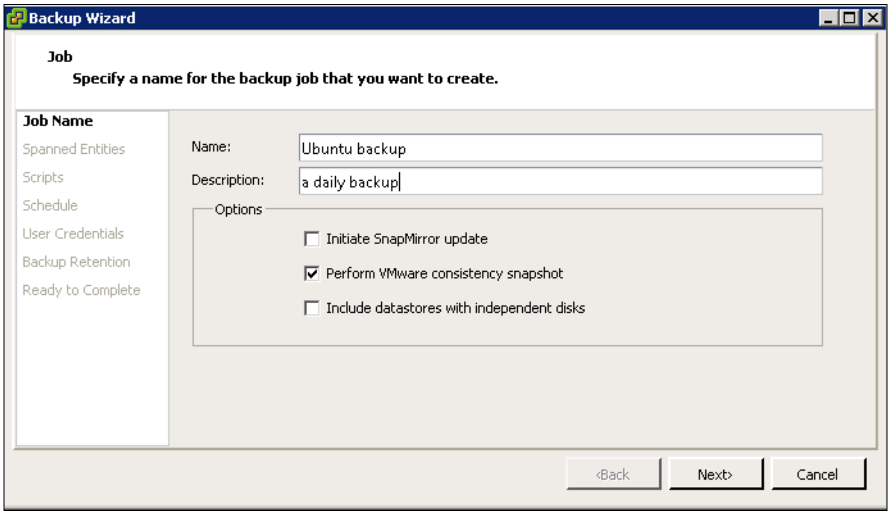


Figure 34-15 Adding a backup

2. Go to the Welcome panel, and then click **Next**.
3. Set a Name and Description, specify possible SnapMirror update, or include independent disks (see Figure 34-16), then click **Next**.



Job

Specify a name for the backup job that you want to create.

Name: Ubuntu backup

Description: a daily backup

Options

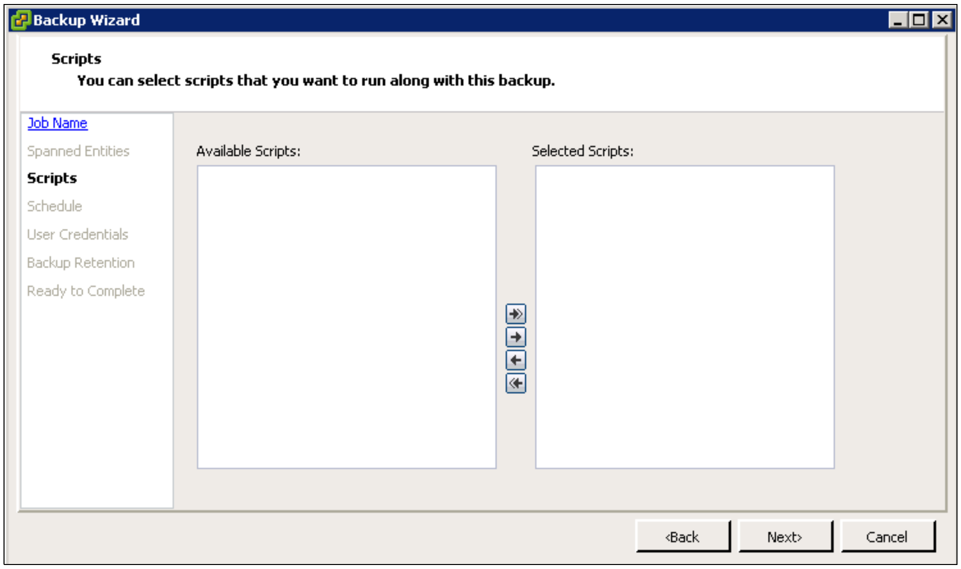
☐ Initiate SnapMirror update

☑ Perform VMware consistency snapshot

☐ Include datastores with independent disks

Figure 34-16 Backup options

4. In the following window, you can select scripts to be included in the backup job (see Figure 34-17).



Scripts

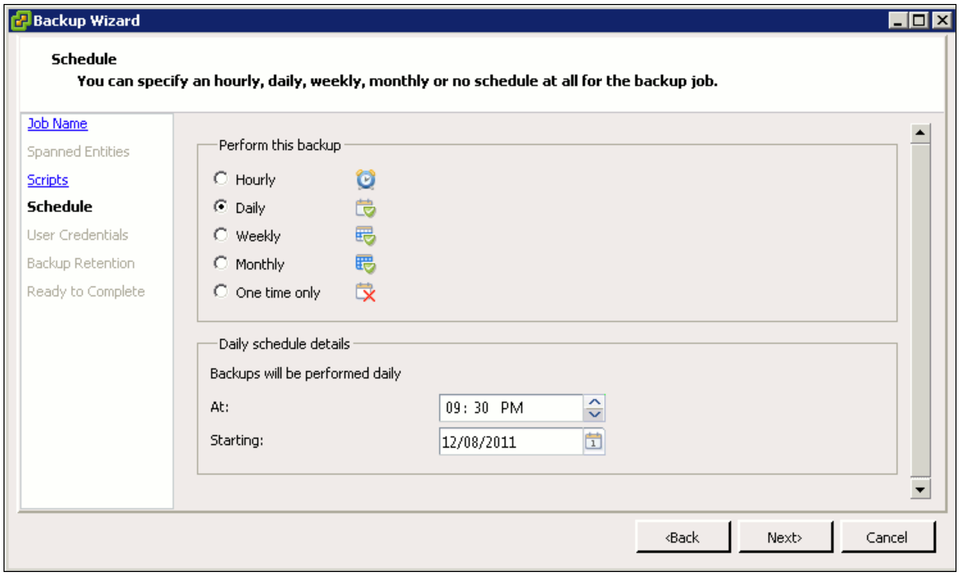
You can select scripts that you want to run along with this backup.

Available Scripts:

Selected Scripts:

Figure 34-17 Backup scripts

5. Now you can specify the schedule for the backup job as shown in Figure 34-18, and click **Next**.



Schedule

You can specify an hourly, daily, weekly, monthly or no schedule at all for the backup job.

Perform this backup

○ Hourly
● Daily
○ Weekly
○ Monthly
○ One time only

Daily schedule details

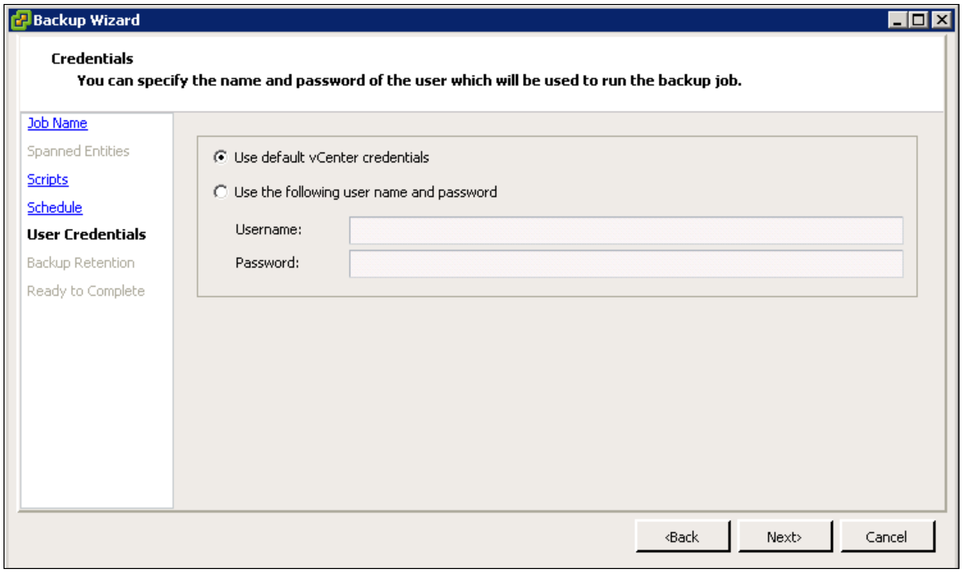
Backups will be performed daily

At: 09: 30 PM

Starting: 12/08/2011

Figure 34-18 Backup schedule

6. Confirm your credentials on the next panel as shown in Figure 34-19, and click **Next**.



Backup Wizard

Credentials

You can specify the name and password of the user which will be used to run the backup job.

● Use default vCenter credentials

○ Use the following user name and password

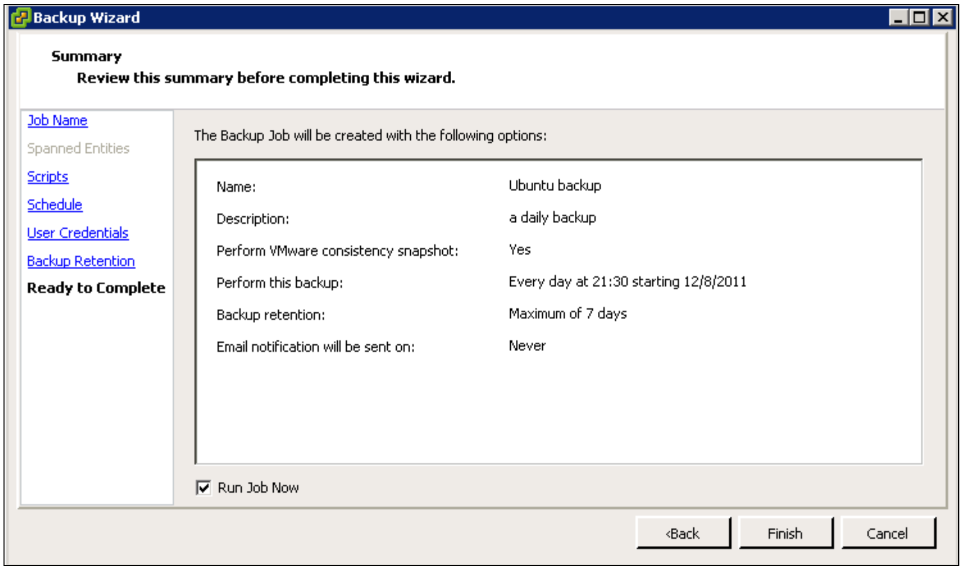
Username:

Password:

Figure 34-19 Backup job credentials

7. Revise the information entered and click **Finish** on the Schedule a Backup Wizard and click **Next**.

8. Select to run your new backup job immediately if you want, as shown in Figure 34-20.



Backup Wizard

Summary

Review this summary before completing this wizard.

The Backup Job will be created with the following options:

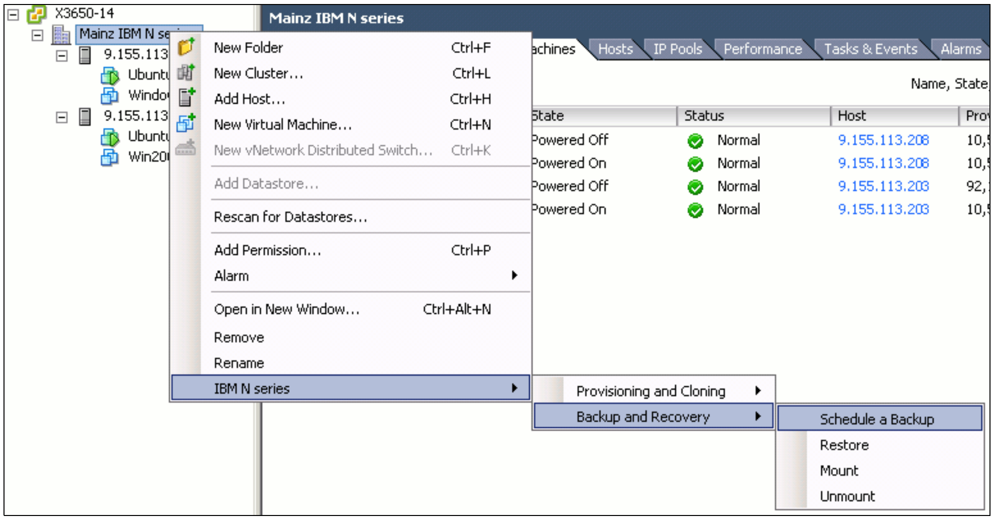
Name:	Ubuntu backup
Description:	a daily backup
Perform VMware consistency snapshot:	Yes
Perform this backup:	Every day at 21:30 starting 12/8/2011
Backup retention:	Maximum of 7 days
Email notification will be sent on:	Never

☑ Run Job Now

Figure 34-20 Revise scheduled backup job

Datacenter backup

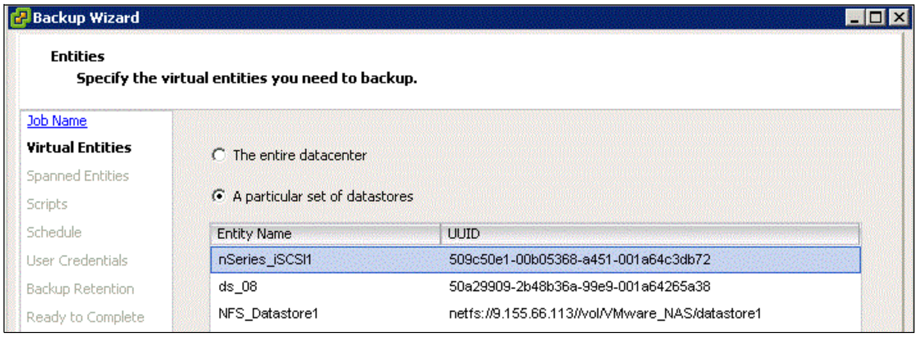
Alternatively, you can also select to back up the whole datacenter as shown in Figure 34-21. Some options are then added to the previously described process.



State	Status	Host	Pro
Powered Off	Normal	9.155.113.208	10,
Powered On	Normal	9.155.113.208	10,
Powered Off	Normal	9.155.113.203	92,
Powered On	Normal	9.155.113.203	10,

Figure 34-21 Datacenter backup

The backup wizard adds the option to select the whole datacenter of backup individual datastores as displayed in Figure 34-22.



Entities

Specify the virtual entities you need to backup.

○ The entire datacenter

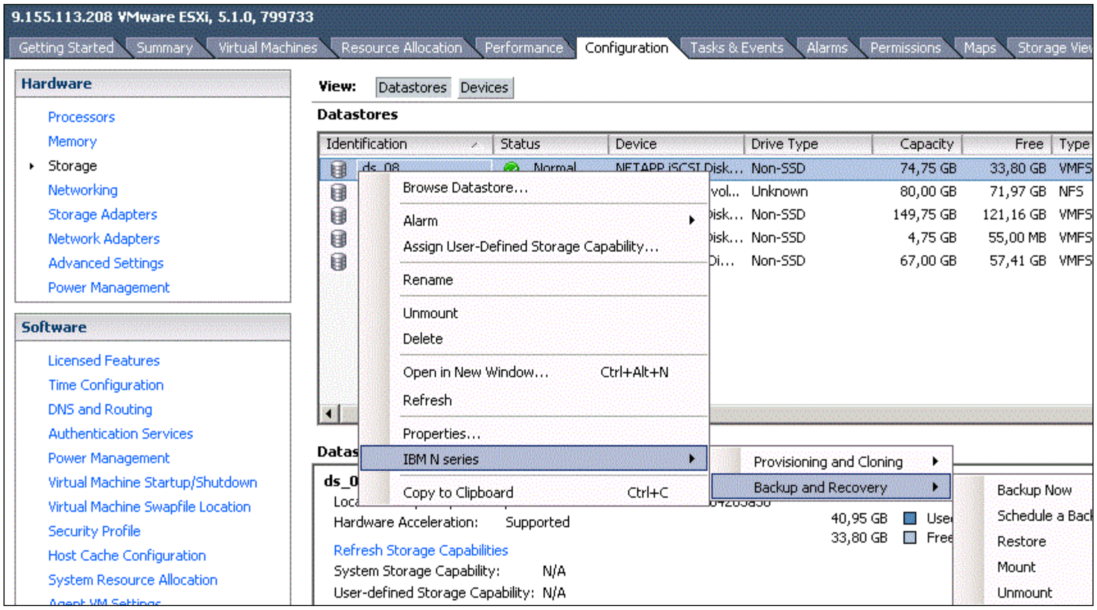
● A particular set of datastores

Entity Name	UUID
nSeries_iSCSI1	509c50e1-00b05368-a451-001a64c3db72
ds_08	50a29909-2b48b36a-99e9-001a64265a38
NFS_Datastore1	netfs://9.155.66.113//vol/VMware_NAS/datastore1

Figure 34-22 Datacenter backup options

Datastore backup

Alternatively, you can also select to back up an individual datastore as shown in Figure 34-23. Some options are then added to the previously described process.



Identification	Status	Device	Drive Type	Capacity	Free	Type
ds_08	Normal	NETAPP iSCSI Disk...	Non-SSD	74,75 GB	33,80 GB	VMFS
		vol...	Unknown	80,00 GB	71,97 GB	NFS
		isk...	Non-SSD	149,75 GB	121,16 GB	VMFS
		isk...	Non-SSD	4,75 GB	55,00 MB	VMFS
		Di...	Non-SSD	67,00 GB	57,41 GB	VMFS

Datas

ds_0

Loca

Hardware Acceleration: Supported

Refresh Storage Capabilities

System Storage Capability: N/A

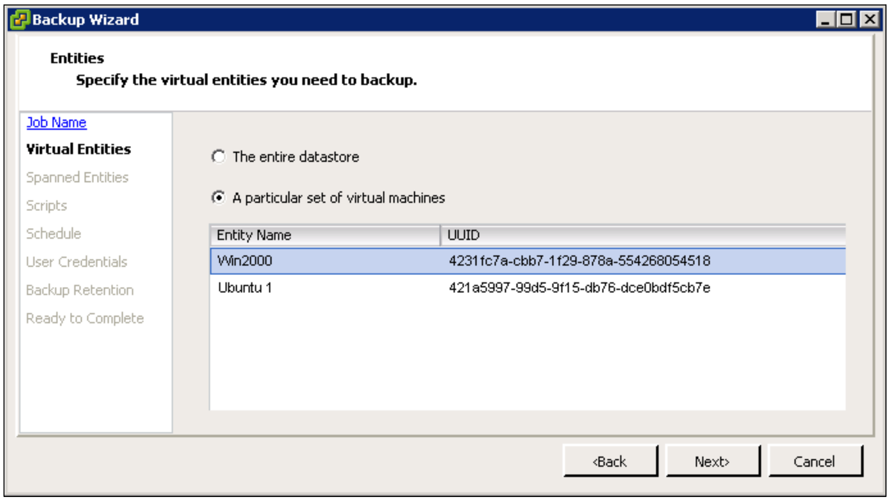
User-defined Storage Capability: N/A

40,95 GB ■ Use

33,80 GB □ Free

Figure 34-23 Datastore backup

The backup wizard adds the option to select the whole datastore of backup individual datastores as displayed in Figure 34-24.



Entities

Specify the virtual entities you need to backup.

○ The entire datastore

● A particular set of virtual machines

Entity Name	UUID
Win2000	4231fc7a-cbb7-1f29-878a-554268054518
Ubuntu 1	421a5997-99d5-9f15-db76-dce0bdf5cb7e

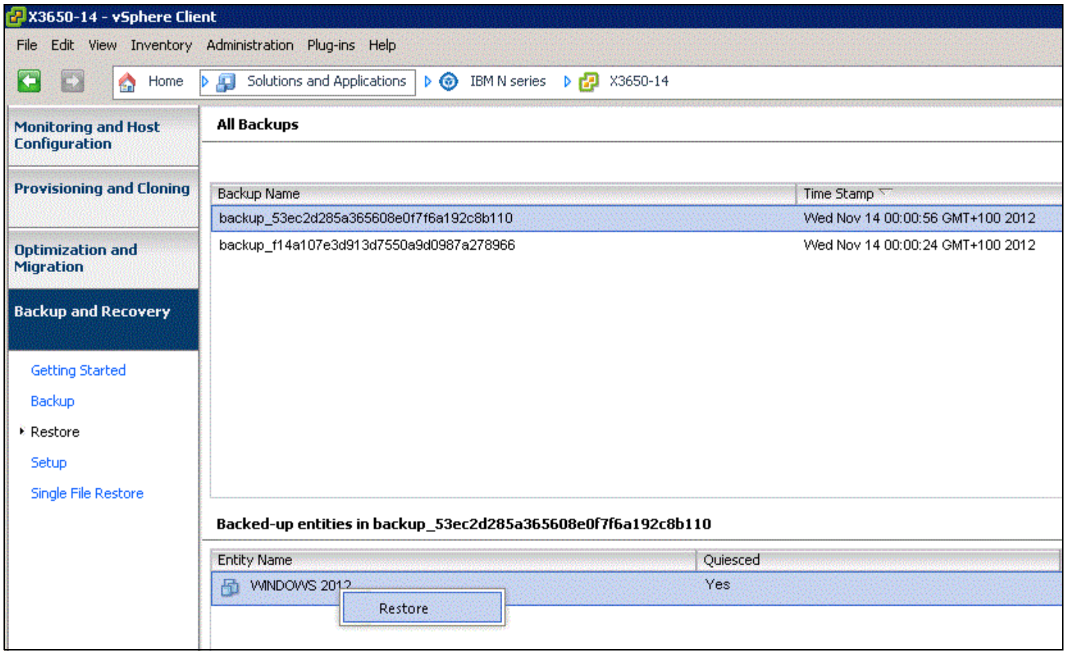
Figure 34-24 Datastore backup options

34.7.4 Granular restore options

The following granular restore options are available:

- Restore datastores or virtual machines to the original location.
- Restore virtual machine disks (VMDKs) to the original or an alternate location.
- Restore one or more files to a guest VMDK without having to restore the entire virtual machine or VMDK using single file restore feature.

You can access these options by the tabs as shown in Figure 34-25. Right-click the object that you want to restore.


X3650-14 - vSphere Client
File Edit View Inventory Administration Plug-ins Help
Home
Solutions and Applications
IBM N series
X3650-14
Monitoring and Host Configuration
Provisioning and Cloning
Optimization and Migration
Backup and Recovery
Getting Started
Backup
Restore
Setup
Single File Restore
Restore

All Backups

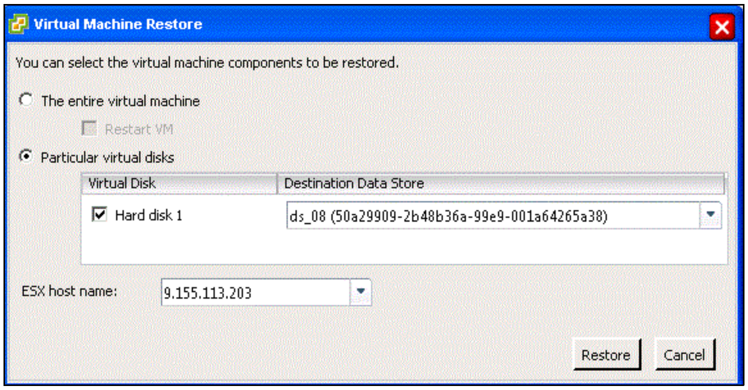
Backup Name	Time Stamp
backup_53ec2d285a365608e0f7f6a192c8b110	Wed Nov 14 00:00:56 GMT+100 2012
backup_f14a107e3d913d7550a9d0987a278966	Wed Nov 14 00:00:24 GMT+100 2012

Backed-up entities in backup_53ec2d285a365608e0f7f6a192c8b110

Entity Name	Quiesced
WINDOWS 2012	Yes

Figure 34-25 Restore options

You can also select whether you want to restore the entire virtual machine or individual virtual disks, as shown in Figure 34-26. Furthermore, you can select the original or a new location.



You can select the virtual machine components to be restored.

○ The entire virtual machine

□ Restart VM

● Particular virtual disks

Virtual Disk	Destination Data Store
☑ Hard disk 1	ds_08 (50a29909-2b48b36a-99e9-001a64265a38)

ESX host name: 9.155.113.203

Figure 34-26 VSC enhanced restore options

34.7.5 Other features

In addition, VSC offers these features:

- Consistent backup naming
- Serialization of VMware vSphere snapshots
- AutoSupport (ASUP) logging
- vFiler unit support for multiple IP addresses
- Advanced Find option to find specific backups

34.8 Provisioning and cloning

This section provides information and examples of the Provisioning and Cloning functions integrated in VSC.

34.8.1 Features and functions

The provisioning features require at least Data ONTAP 7.3.3 to accomplish the following tasks:

- Creation, resizing, and deletion of VMFS/NFS datastores
- Ability to provision, clone, and resize volumes on secure vFiler units
- Adding storage system using a domain account
- Automation of pathing for both LUNs and NFS datastores
- Running deduplication operations
- Monitoring storage savings and performance
- Protection against failover of NFS mounts to non-redundant VMkernel ports by limiting multiple TCP sessions to iSCSI only

The cloning features allow you to perform the following tasks:

- Creation of multiple virtual machine clones in new or existing datastores (using FlexClone technology)
- Application of guest customization specifications and powering up of new virtual machines
- Redeployment of virtual machines from a baseline image
- Importing virtual machines into virtual desktop infrastructure connection brokers and management tools
- Clone misalignment alert and prevention:
  - VM misalignment detection and user notification
  - Support for VMFS- and NFS-based VMs
- Ability to import virtual machine settings from a file:
  - Non-contiguous virtual machine names
  - Guest customization specifications
  - Computer name as virtual machine name
  - Power-on settings
- Support for these products:
  - VMware View 4.0, 4.5, 4.6 & 5.0 or later
  - Citrix XenDesktop 4.0 and 5.0 or later

Further features are included:

- Space reclamation management
- Addition of new datastores to new ESX Servers within a cluster
- Service catalog-based provisioning API with enhanced SOAP API to support creation, deletion, and resizing of NFS/VMFS datastores by Storage Services in Provisioning Manager
- Space Reclamation Management

- Mounting of existing datastores when new ESX hosts are added to a cluster or datacenter with support for both NFS and VMFS datastores
- Capability for the user to mount any existing datastore to newly added ESX hosts:
  - VDI One-click Golden Template distribution
  - This feature allows the user to copy a datastore from a source vCenter to one or more target vCenters
- VMware Virtual Desktop Infrastructure (VDI) enhancements:
  - XenDesktop/View import from API
  - VDI One-click Golden Template distribution
  - Saving of View credentials
  - Soap API support for importing newly created clones into Citrix XenDesktop and VMware View
  - Storing of View Server credentials
  - Elimination of the need to add VMware View Server credentials each time by the cloning wizard
  - Creation of multiple View Server pools

34.8.2 Provisioning datastores

With the Provisioning and Cloning feature of the VSC 4.1, you can create new datastores at the datacenter, cluster, or host level. The new datastore displays on every host in the datacenter or the cluster.

This process launches the N series Datastore Provisioning wizard, which allows you to select the following features:

- Storage controller
- Type of datastore (VMFS or NFS)
- Datastore details, including storage protocol and block size (if deploying a VMFS datastore)
- Specifying whether the LUN should be thin-provisioned

The provisioning process connects the datastore to all nodes within the selected group. For iSCSI, FC, and FCoE datastores, the VSC handles storage access control as follows:

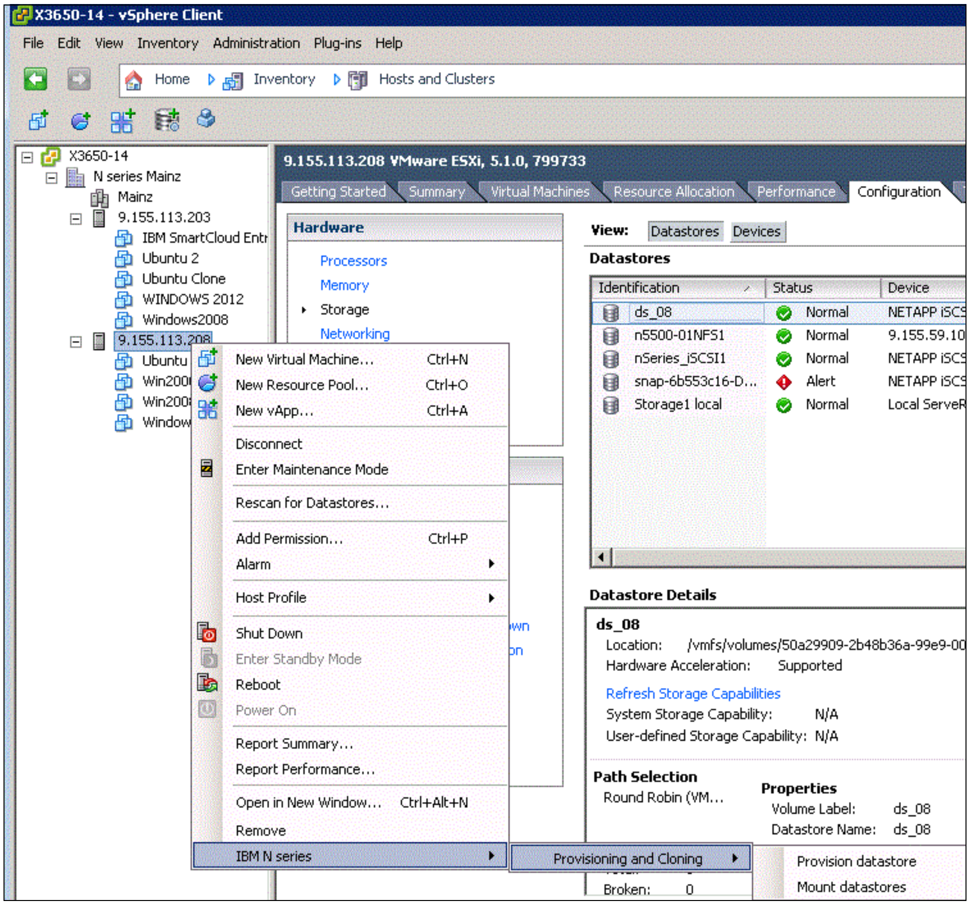
- Creating initiator groups
- Enabling ALUA
- Applying LUN masking
- Applying path selection policies
- Formatting the LUN with VMFS

For NFS datastores, the VSC handles storage access control by managing access rights in the exports file, and it balances the load across all available interfaces.

Tip: Remember, if you plan to enable data deduplication, then thin-provisioned LUNs are required to return storage to the free pool on the storage controller.

Follow these steps:

1. In the vSphere Client Inventory, right-click a datacenter, cluster, or host and select **N series** → **Provisioning and Cloning** → **Provision datastore** (see Figure 34-27).

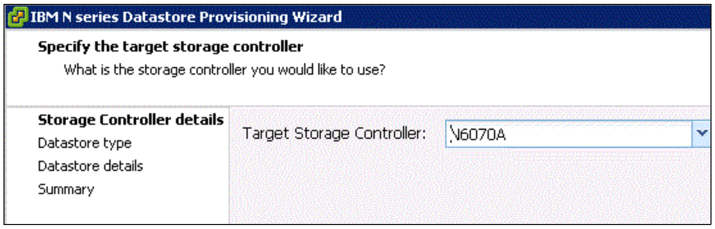


Identification	Status	Device
ds_08	Normal	NETAPP iSCS
n5500-01NFS1	Normal	9.155.59.10
nSeries_iSCSI1	Normal	NETAPP iSCS
snap-6b553c16-D...	Alert	NETAPP iSCS
Storage1 local	Normal	Local ServeR

ds_08
Location: /vmfs/volumes/50a29909-2b48b36a-99e9-00
Hardware Acceleration: Supported
System Storage Capability: N/A
User-defined Storage Capability: N/A
Round Robin (VM...
Volume Label: ds_08
Datastore Name: ds_08
Broken: 0

Figure 34-27 Provision a datastore

2. Next specify the N series system to use (see Figure 34-28).

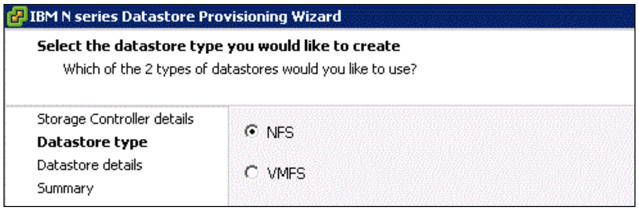

IBM N series Datastore Provisioning Wizard
Storage Controller details
Datastore type
Datastore details
Summary

Specify the target storage controller
What is the storage controller you would like to use?

Target Storage Controller: N6070A

Figure 34-28 Select storage controller for provisioning

3. In the following window, select the protocol to use. Here we only have NFS available, as shown in Figure 34-29.



IBM N series Datastore Provisioning Wizard

Select the datastore type you would like to create

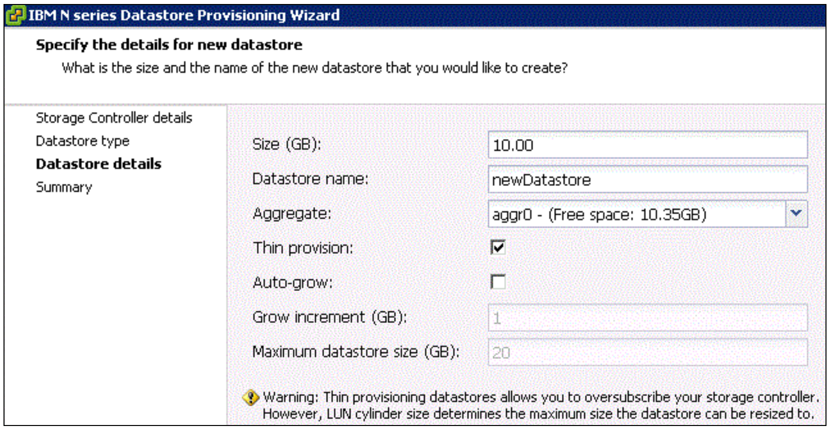
Which of the 2 types of datastores would you like to use?

● NFS

○ VMFS

Figure 34-29 Specify datastore type

4. Now specify the new datastore details (see Figure 34-30).



IBM N series Datastore Provisioning Wizard

Specify the details for new datastore

What is the size and the name of the new datastore that you would like to create?

Size (GB): 10.00

Datastore name: newDatastore

Aggregate: aggr0 - (Free space: 10.35GB)

Thin provision: ☑

Auto-grow: □

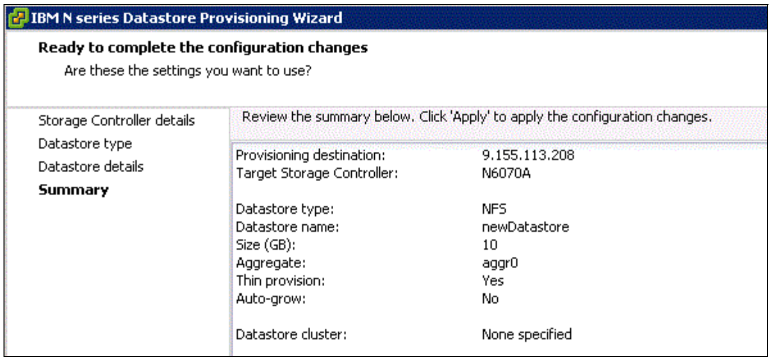
Grow increment (GB): 1

Maximum datastore size (GB): 20

Warning: Thin provisioning datastores allows you to oversubscribe your storage controller. However, LUN cylinder size determines the maximum size the datastore can be resized to.

Figure 34-30 New datastore details

5. Before applying your selection, verify the information as shown in Figure 34-31.



IBM N series Datastore Provisioning Wizard

Ready to complete the configuration changes

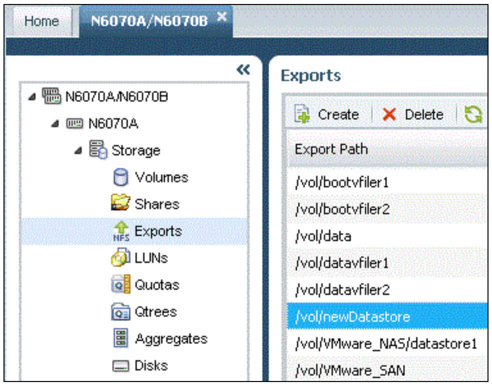
Are these the settings you want to use?

Review the summary below. Click 'Apply' to apply the configuration changes.

```
Provisioning destination:       9.155.113.208
Target Storage Controller:      N6070A

Datastore type:                 NFS
Datastore name:                 newDatastore
Size (GB):                      10
Aggregate:                      aggr0
Thin provision:                 Yes
Auto-grow:                      No

Datastore cluster:              None specified
```

Figure 34-31 Review new datastore settings

6. The new datastore named *newDatastore* was created on the N series. It can now be mounted to the host you want. Figure 34-32 shows System Manager access and the NFS exports.



Export Path
/vol/bootvfiler1
/vol/bootvfiler2
/vol/data
/vol/datavfiler1
/vol/datavfiler2
/vol/newDatastore
/vol/VMware_NAS/datastore1
/vol/VMware_SAN

Figure 34-32 Verify NFS exports

34.8.3 Managing deduplication

Deduplication eliminates redundant objects on a selected datastore and only references the original object. Figure 34-33 shows how VSC is able to manage deduplication for each individual datastore.

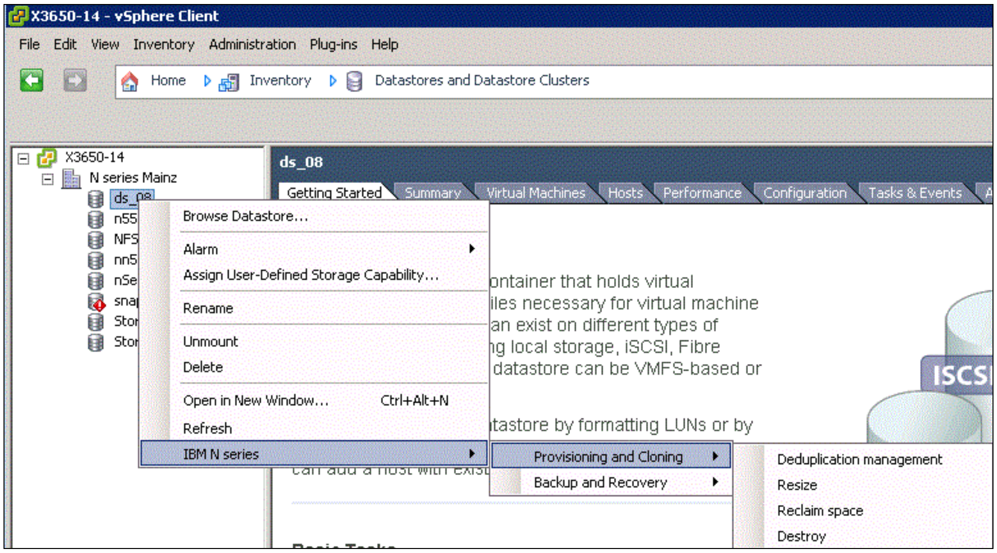
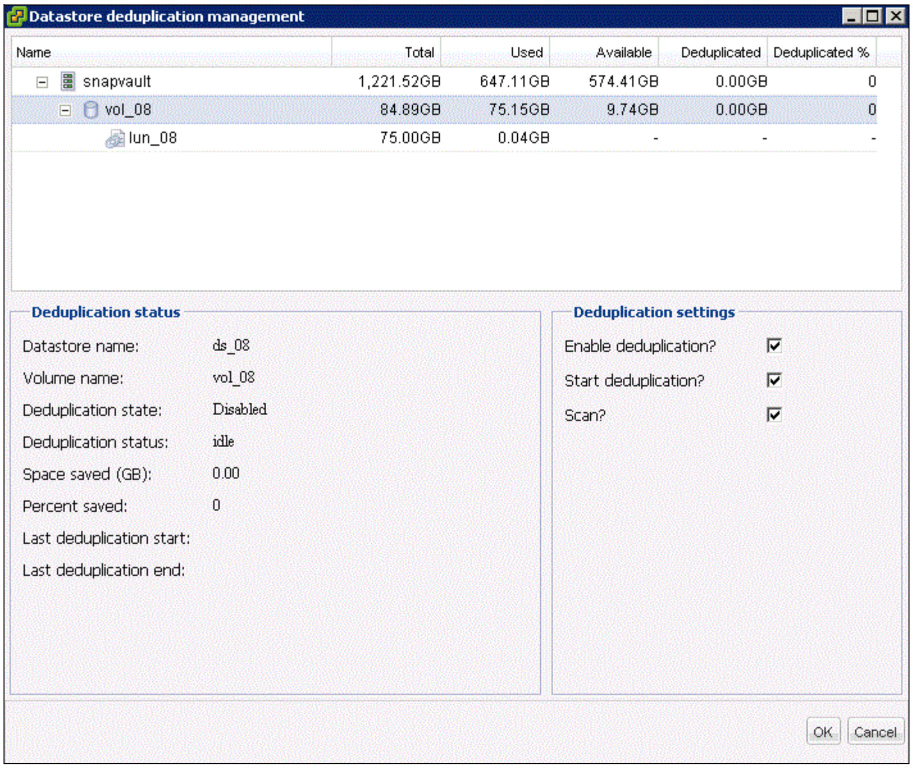

X3650-14 - vSphere Client
File Edit View Inventory Administration Plug-ins Help
Home
Inventory
Datastores and Datastore Clusters
X3650-14
N series Mainz
ds_08
n55
NFS
nn5
nSe
snap
Stor
Stor
ds_08
Getting Started
Summary
Virtual Machines
Hosts
Performance
Configuration
Tasks & Events
Browse Datastore...
Alarm
Assign User-Defined Storage Capability...
Rename
Unmount
Delete
Open in New Window... Ctrl+Alt+N
Refresh
IBM N series
Provisioning and Cloning
Backup and Recovery
Deduplication management
Resize
Reclaim space
Destroy
ontainer that holds virtual
iles necessary for virtual machine
an exist on different types of
ng local storage, iSCSI, Fibre
datastore can be VMFS-based or
tastore by formatting LUNs or by
can add a host with exist
ISCSI

Figure 34-33 Managing deduplication

Possible options to use N series advanced deduplication features are displayed in Figure 34-34. Click **OK** to apply your settings.



Name	Total	Used	Available	Deduplicated	Deduplicated %
snapvault	1,221.52GB	647.11GB	574.41GB	0.00GB	0
vol_08	84.89GB	75.15GB	9.74GB	0.00GB	0
lun_08	75.00GB	0.04GB	-	-	-

Deduplication status

Datastore name: ds_08
Volume name: vol_08
Deduplication state: Disabled
Deduplication status: idle
Space saved (GB): 0.00
Percent saved: 0
Last deduplication start:
Last deduplication end:

Deduplication settings

Enable deduplication? ☑
Start deduplication? ☑
Scan? ☑

Figure 34-34 Manage deduplication features

34.8.4 Cloning virtual machines

The Provisioning and Cloning capability can theoretically create thousands of virtual machine clones and hundreds of datastores at one time. In practice, however, multiple executions of fewer requests are preferred. The exact size of these requests depends on the size of the vSphere deployment and the hardware configuration of the vSphere Client managing the ESX hosts.

Follow these steps:

1. In the vSphere Client Inventory, right-click a powered-down virtual machine (Figure 34-35) or template and select **N series** → **Provisioning and Cloning** → **Create rapid clones**.

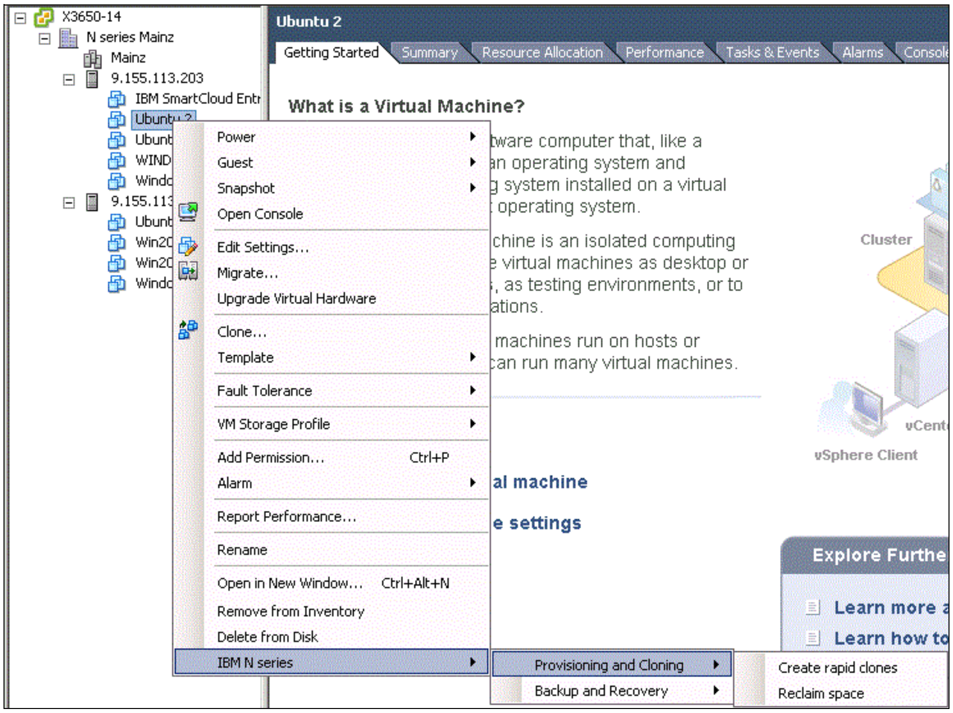
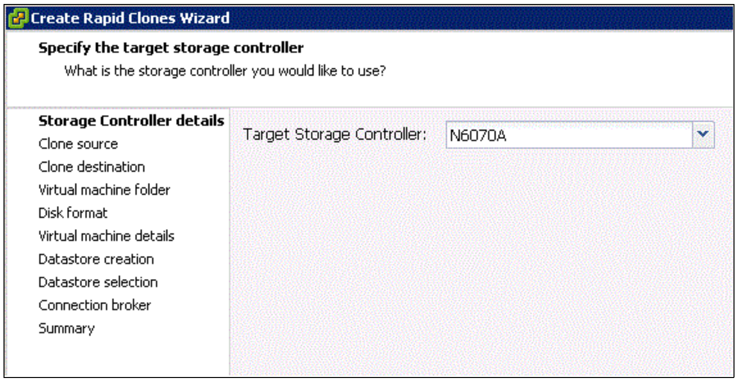


Figure 34-35 Select VM for cloning

2. Next select the controller you want to use for cloning (see Figure 34-36).



Create Rapid Clones Wizard

Specify the target storage controller

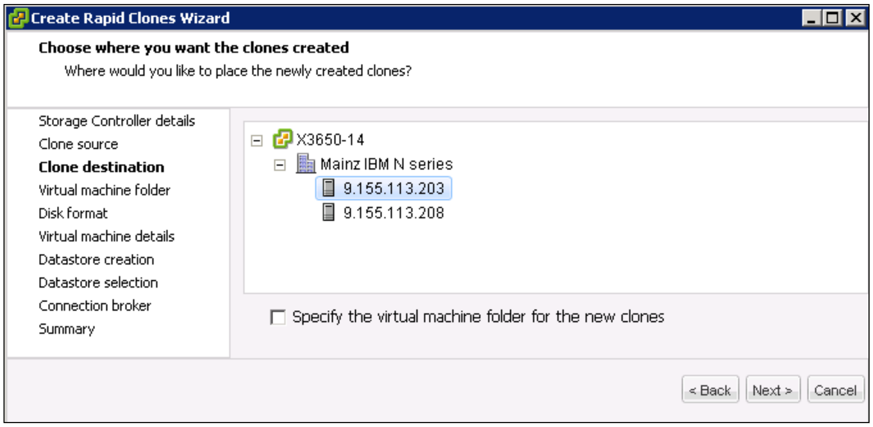
What is the storage controller you would like to use?

Storage Controller details
Clone source
Clone destination
Virtual machine folder
Disk format
Virtual machine details
Datastore creation
Datastore selection
Connection broker
Summary

Target Storage Controller: N6070A

Figure 34-36 Select controller for cloning

3. In the following window, select the destination N series system (see Figure 34-37).



Choose where you want the clones created

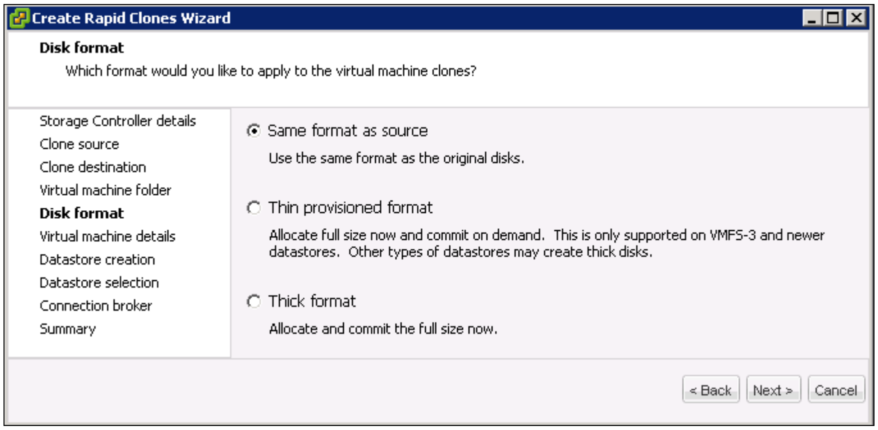
Where would you like to place the newly created clones?

- X3650-14
  - Mainz IBM N series
    - 9.155.113.203
    - 9.155.113.208

□ Specify the virtual machine folder for the new clones

Figure 34-37 Select clone target

4. Now specify the VM format for the clone as shown in Figure 34-38.



Disk format

Which format would you like to apply to the virtual machine clones?

● Same format as source

Use the same format as the original disks.

○ Thin provisioned format

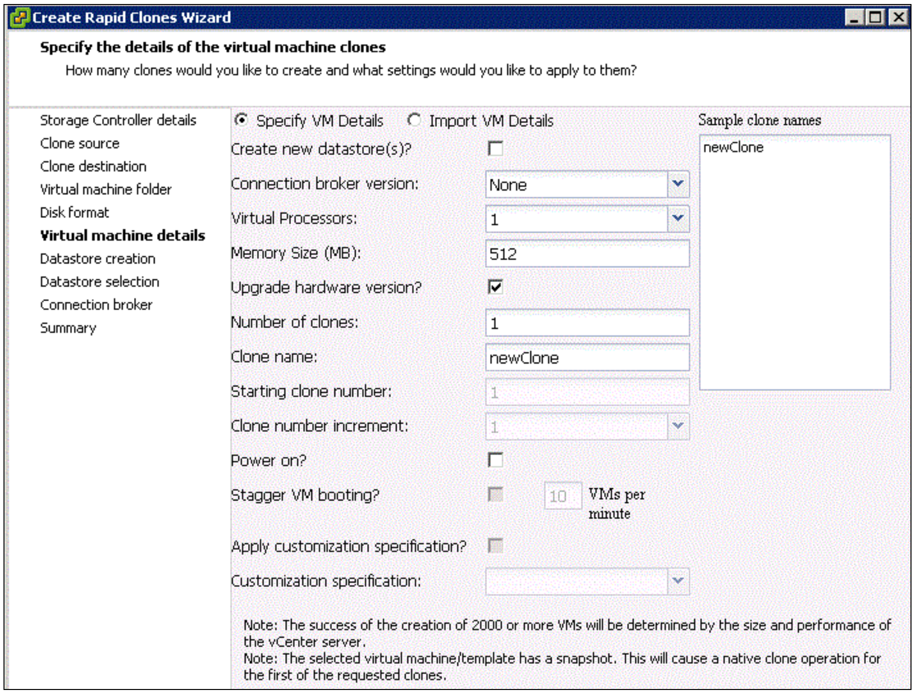
Allocate full size now and commit on demand. This is only supported on VMFS-3 and newer datastores. Other types of datastores may create thick disks.

○ Thick format

Allocate and commit the full size now.

Figure 34-38 Clone VM format

5. In the following window, specify details for the new datastores as displayed in Figure 34-39.



Specify the details of the virtual machine clones

How many clones would you like to create and what settings would you like to apply to them?

● Specify VM Details ○ Import VM Details

Create new datastore(s)?	□
Connection broker version:	None
Virtual Processors:	1
Memory Size (MB):	512
Upgrade hardware version?	☑
Number of clones:	1
Clone name:	newClone
Starting clone number:	1
Clone number increment:	1
Power on?	□
Stagger VM booting?	□ 10 VMs per minute
Apply customization specification?	□
Customization specification:	

Sample clone names

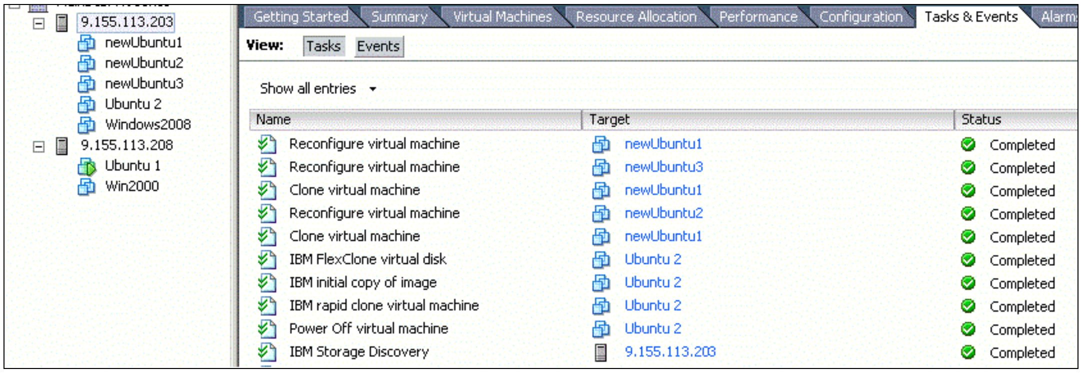
newClone

Note: The success of the creation of 2000 or more VMs will be determined by the size and performance of the vCenter server.
Note: The selected virtual machine/template has a snapshot. This will cause a native clone operation for the first of the requested clones.

Figure 34-39 Clone VM details

6. When a summary is provided, click **Apply** to execute your selection.

After successful completion of the cloning tasks, the new VMs are configured and ready for further use. Figure 34-40 shows the cloning results.



Name	Target	Status
Reconfigure virtual machine	newUbuntu1	Completed
Reconfigure virtual machine	newUbuntu3	Completed
Clone virtual machine	newUbuntu1	Completed
Reconfigure virtual machine	newUbuntu2	Completed
Clone virtual machine	newUbuntu1	Completed
IBM FlexClone virtual disk	Ubuntu 2	Completed
IBM initial copy of image	Ubuntu 2	Completed
IBM rapid clone virtual machine	Ubuntu 2	Completed
Power Off virtual machine	Ubuntu 2	Completed
IBM Storage Discovery	9.155.113.203	Completed

Figure 34-40 Clone results

34.8.5 Reclaiming space on virtual machines

You can use the Reclaim space feature to find free clusters on NTFS partitions and make them available to the operating system.

Before you begin

The Reclaim space feature allows Data ONTAP to use space freed when data is deleted in guest operating systems.

This feature has the following requirements:

- VMDKs attached to the virtual machine must be on NFS-backed datastores.

> **Tip:** The Reclaim space feature is not supported if the NFS datastore is backed by a qtree on a vFiler unit.

- VMDKs must have NTFS partitions.

> **Tip:** If the VMDK is unpartitioned or FAT, the Provisioning and Cloning capability incorrectly lists the disk as having an NTFS partition after the task completes and displays a “Yes” in the “Has NTFS partition(s)?” column. Even though the VMDK now appears to be partitioned, it is still unpartitioned or FAT, and you cannot reclaim space on it.

- ISOs mounted to the virtual machine must be contained in an NFS datastore.
- Storage systems must be running Data ONTAP 7.3.4 or later.
- You should have the VMware guest tools installed.
- When the Reclaim space feature is running, you must not power on the virtual machine.
- You cannot use the cloning feature when the target virtual machine is being used by either the Backup and Recovery capability or the Optimization and Migration capability.

Steps

Follow these steps:

1. Right-click a datastore or virtual machine and select **IBM N series** → **Provisioning and Cloning** → **Reclaim space** (Figure 34-41).

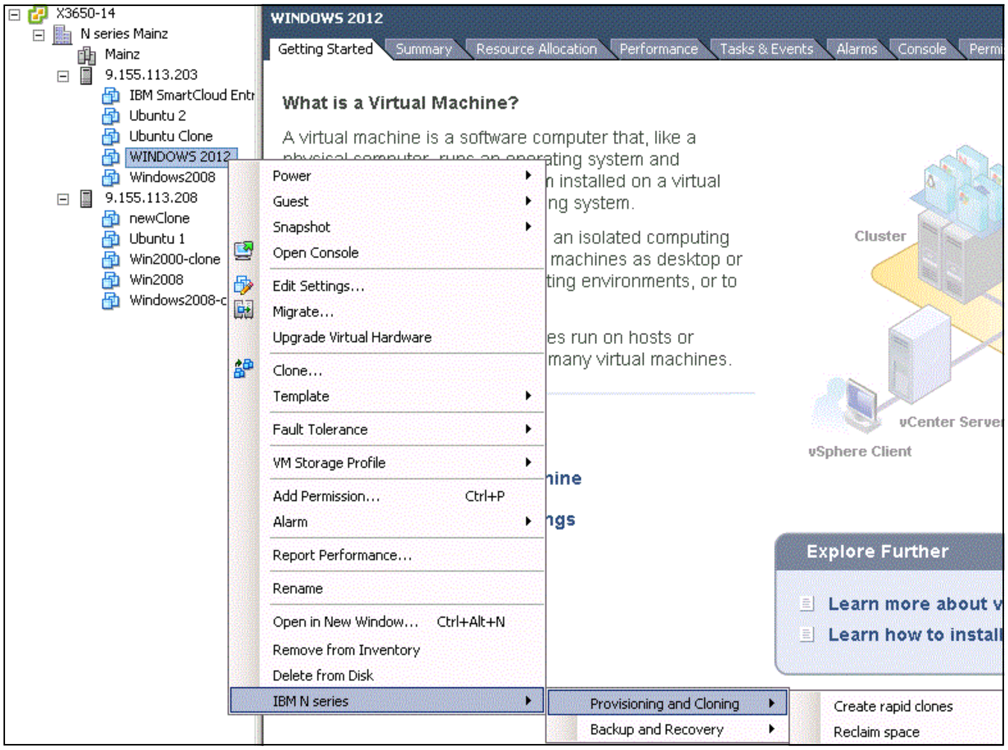


Figure 34-41 VM reclaim space

2. Click **OK**.

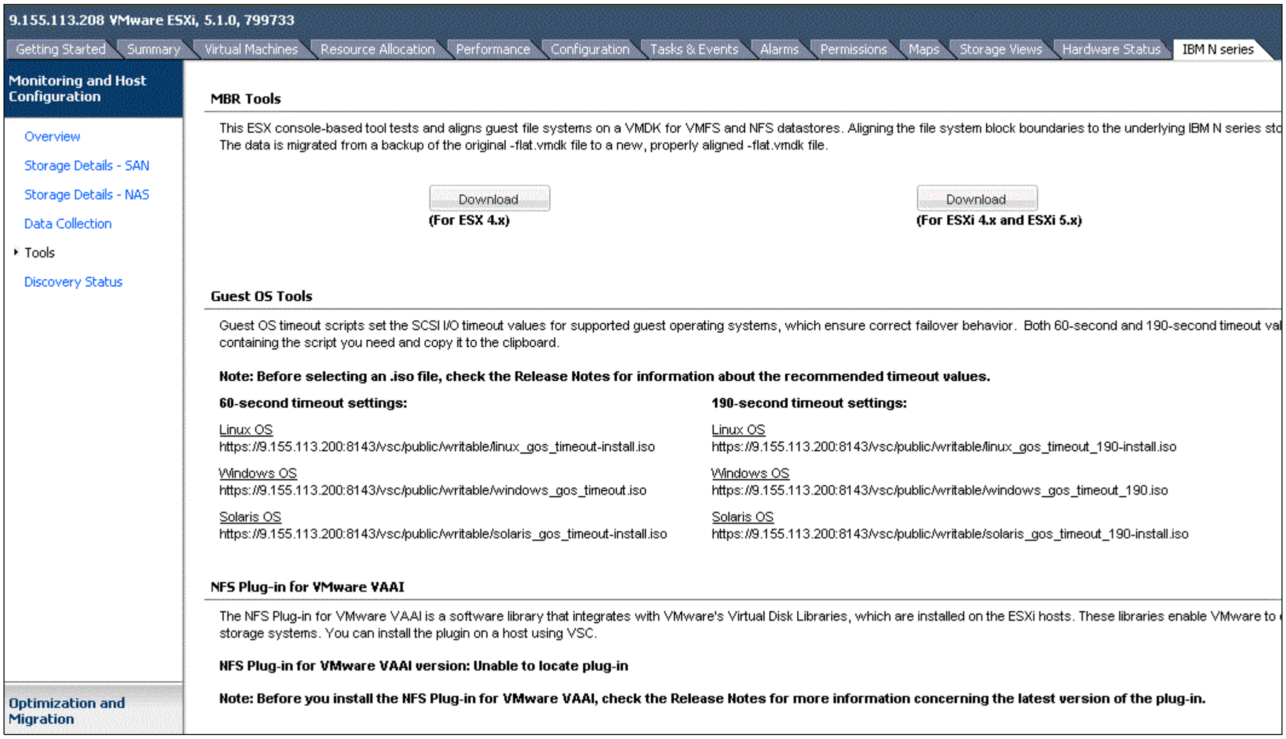
If the virtual machine is powered on, the Reclaim space feature powers it off. After the process completes, the Reclaim space feature returns the virtual machine to its previous state.

Tip: If you are using this feature when the virtual machine is powered on, make sure you have the guest operating system tools installed. Without these tools, the Reclaim space feature does not work when it has to power down the virtual machine.

If you do not want to install these tools, then you should power down the virtual machine before running the Reclaim space feature.

34.9 Optimum VM availability

The Monitoring and Host Configuration capability includes tools for detecting and correcting misaligned disk partitions and for setting virtual machine timeouts as shown in Figure 34-42.



MBR Tools

This ESX console-based tool tests and aligns guest file systems on a VMDK for VMFS and NFS datastores. Aligning the file system block boundaries to the underlying IBM N series sto
The data is migrated from a backup of the original -flat.vmdk file to a new, properly aligned -flat.vmdk file.

(For ESX 4.x)

(For ESXi 4.x and ESXi 5.x)

Guest OS Tools

Guest OS timeout scripts set the SCSI I/O timeout values for supported guest operating systems, which ensure correct failover behavior. Both 60-second and 190-second timeout val
containing the script you need and copy it to the clipboard.

Note: Before selecting an .iso file, check the Release Notes for information about the recommended timeout values.

60-second timeout settings:	190-second timeout settings:
Linux OS https://9.155.113.200:8143/vsc/public/writable/linux_gos_timeout-install.iso	Linux OS https://9.155.113.200:8143/vsc/public/writable/linux_gos_timeout_190-install.iso
Windows OS https://9.155.113.200:8143/vsc/public/writable/windows_gos_timeout.iso	Windows OS https://9.155.113.200:8143/vsc/public/writable/windows_gos_timeout_190.iso
Solaris OS https://9.155.113.200:8143/vsc/public/writable/solaris_gos_timeout-install.iso	Solaris OS https://9.155.113.200:8143/vsc/public/writable/solaris_gos_timeout_190-install.iso

NFS Plug-in for VMware VAAI

The NFS Plug-in for VMware VAAI is a software library that integrates with VMware's Virtual Disk Libraries, which are installed on the ESXi hosts. These libraries enable VMware to
storage systems. You can install the plugin on a host using VSC.

NFS Plug-in for VMware VAAI version: Unable to locate plug-in

Note: Before you install the NFS Plug-in for VMware VAAI, check the Release Notes for more information concerning the latest version of the plug-in.

Figure 34-42 VSC tools

Tip: The Optimization and Migration capability of VSC allows you to perform online alignments on VMFS-based datastores without having to take your VM down. This capability also lets you review the alignment status of VMs and migrate groups of VMs.

34.9.1 Optimizing VM SCSI BUS

One of the components of the VSC is the GOS timeout scripts. These scripts are a collection of ISO images that can be mounted by a VM to configure its local SCSI to values that are optimal for running in a virtual infrastructure.

Installing GOS scripts

The ISO images of the guest operating system (GOS) scripts are loaded on the VSC for VMware vSphere server. Mount and run them from the vSphere Client to set the storage timeouts for virtual machines.

Before you begin

Ensure the following prerequisites:

- The virtual machine must be running.
- The CD-ROM must already exist in the virtual machine or it must be added.
- The script must be installed from the copy of the VSC for VMware vSphere registered to the vCenter Server that manages the VM.

Steps

Follow these steps:

1. Open the vSphere Client and log into your vCenter Server.
2. Select a **Datacenter** in the Inventory panel, and then select the **IBM N series** tab.
3. In the Monitoring and Host Configuration capability, select the **Tools** panel.
4. Under **Guest OS Tools**, right-click the link to the ISO image for your guest operating system version and select **Copy to clipboard**.
5. In the vSphere Client, select the desired VM and click the **CD/DVD Connections** icon.
6. Select **CD/DVD Drive 1 > Connect to ISO image on local disk**.
7. Paste the link you copied into the **File Name** field and then click **Open**.

If you receive an authorization error, be sure you select the IBM N series tab and click **Yes** to proceed if a security certificate warning is displayed.

Also, be sure that the link you are using is from the copy of the VSC for VMware vSphere running on the vCenter Server that manages the VM.

After you finish

Log on to the VM and run the script to set the storage timeout values

34.9.2 Optimal storage performance

VMs store their data on virtual disks. Similar to physical disks, these virtual disks contain storage partitions and file systems, which are created by the guest operating system of the VM. To provide optimal disk I/O within the VM, you must align the partitions of the virtual disks to the block boundaries of VMFS and the block boundaries of the storage array. Failure to align all three of these items results in a dramatic increase of I/O load on a storage array and negatively affects the performance of all VMs being served on the array.

IBM, VMware, other storage vendors, and VMware partners advise aligning the partitions of VMs and the partitions of VMFS datastores to the blocks of the underlying storage array.

Datastore alignment

N series systems automate the alignment of VMFS with iSCSI, FC, and FCoE LUNs. This task is automated during the LUN provisioning phase of creating a datastore when you select the LUN type "VMware" for the LUN. Customers deploying VMware over NFS do not need to align the datastore. With any type of datastore, VMFS or NFS, the virtual disks contained within should have the partitions aligned to the blocks of the storage array.

VM partition alignment

When aligning the partitions of virtual disks for use with N series systems, the starting partition offset must be divisible by 4,096. For example, the starting partition offset for Microsoft Windows 2000, 2003, and XP operating systems is 32,256. This value does not align to a block size of 4,096.

Virtual machines running a clean installation of Microsoft Windows 2008, Windows 7, or Windows Vista operating systems automatically have their starting partitions set to 1,048,576. By default, this value does not require any adjustments.

> **Tip:** If your Windows 2008 or Windows Vista VMs were created by upgrading an earlier version of Microsoft Windows to one of these versions, then it is highly probable that these images require partition alignment.

34.9.3 VM partition alignment

Storage alignment is critical, so aligning the file system within the VMs to the storage array is very important. This process should not be considered optional. Misalignment at a high level results in decreased usage.

Issues with partition alignment

Failure to align the file systems results in a significant increase in storage array I/O to meet the I/O requirements of the hosted VMs. Customers might notice this impact in these situations:

- Running high-performance applications
- Achieving less than impressive storage savings with deduplication
- Perceiving a need to upgrade storage array hardware

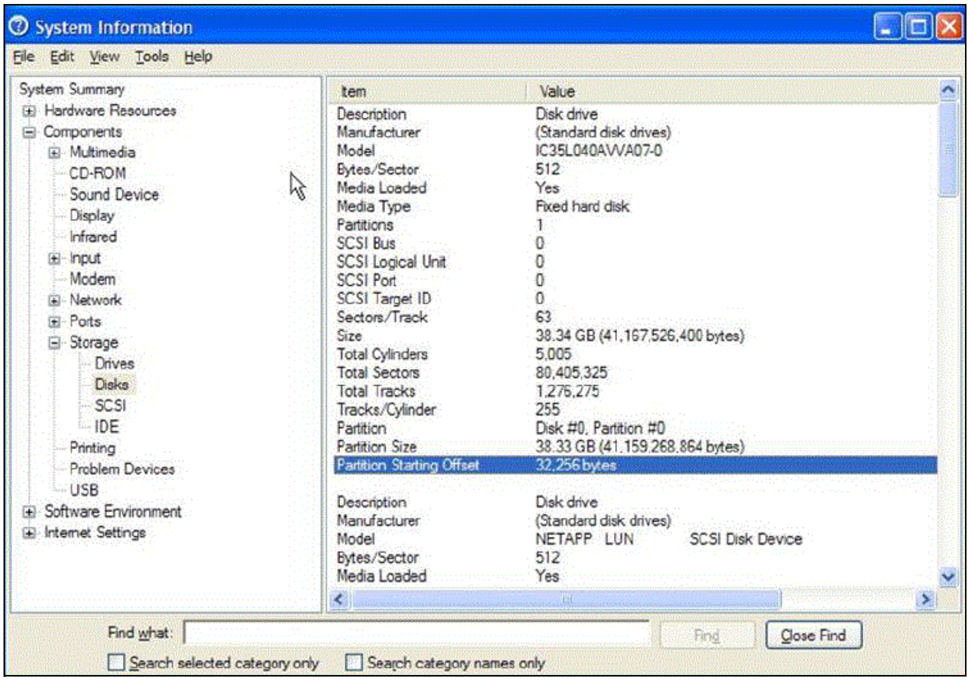
The reason for these types of issues is misalignment results in every I/O operation executed within the VM to require multiple I/O operations on the storage array.

Simply put, you can save your company a significant amount of capital expenditures by optimizing the I/O of your VMs.

Identifying partition alignment

To verify the starting partition offset for a VM based on Windows, complete the following steps:

1. Log in to the VM.
2. Run the system information utility (or `msinfo32`) to find the starting partition offset setting.
3. To run `msinfo32`, click **Start** > **All Programs** > **Accessories** > **System Tools** > **System Information** (Figure 34-43).



Item	Value
Description	Disk drive
Manufacturer	(Standard disk drives)
Model	IC35L040AVVA07-0
Bytes/Sector	512
Media Loaded	Yes
Media Type	Fixed hard disk
Partitions	1
SCSI Bus	0
SCSI Logical Unit	0
SCSI Port	0
SCSI Target ID	0
Sectors/Track	63
Size	38.34 GB (41,167,526,400 bytes)
Total Cylinders	5,005
Total Sectors	80,405,325
Total Tracks	1,276,275
Tracks/Cylinder	255
Partition	Disk #0, Partition #0
Partition Size	38.33 GB (41,159,268,864 bytes)
Partition Starting Offset	32,256 bytes
Description	Disk drive
Manufacturer	(Standard disk drives)
Model	NETAPP LUN SCSI Disk Device
Bytes/Sector	512
Media Loaded	Yes

☐ Search selected category only ☐ Search category names only

Figure 34-43 System information

34.9.4 N series MBR Tools: Identification of partition alignment status

IBM N series systems provides a tool, MBRScan, that runs on an ESX host and can identify if partitions are aligned with Windows and Linux VMs running within VMFS and NFS datastores. MBRScan runs against the virtual disk files that compose a VM. Although this process only requires a few seconds per VM to identify and report on the status of the partition alignment, each VM must be powered off. For this reason, it might be easier to identify the file system alignment from within each VM, because this action is nondisruptive.

MBRScan is an integrated component of the VSC.

Corrective actions for VMs with misaligned partitions

After you identify that your VMs have misaligned partitions, we advise correcting the partitions in your templates as the first corrective action. This step makes sure that any newly created VM is properly aligned and does not add to the I/O load on the storage array.

Correcting partition misalignment with N series MBR Tools

As part of the VSC tools, IBM N series systems provides a tool, MBRAlign, that runs on an ESX host and can correct misaligned primary and secondary master boot record-based partitions for guest operating systems. When using MBRAlign, the VM that is undergoing the corrective action must be powered off.

MBRAlign provides flexible repair options. For example, it can be used to migrate and align a virtual disk as well as change the format from a thin to thick vmdk. We highly advise creating a Snapshot copy before executing MBRAlign. This Snapshot copy can be safely discarded after a VM has been corrected, powered on, and the results have been verified.

You must download these tools before you can use them. There is a set of tools for ESX hosts and one for ESXi hosts. You must download the correct tool set for your hosts. MBRAlign can be obtained from the Tools Download link in the VSC.

Enabling the ESXi secure shell

When you are using ESXi, it is a good practice to enable the Secure Shell (SSH) protocol before you download the MBR tools. That way you can use the `scp` command if you need to copy the files.

Before you begin

ESXi does not enable this shell by default.

Steps

Follow these steps:

1. From an ESXi host, press the key combination **ALT F2** to access the Direct Console User Interface (DCUI) panel.
2. Press the **F2** function key to get to the Customize System panel.
3. Go to **Troubleshooting Options**.
4. Press **Enter** at the Enable SSH prompt.
5. Press **Enter** at the Modify ESX Shell timeout prompt.
6. Disable the timeout by setting the value to zero (0) and pressing **Enter**.
7. Go to **Restart Management Agents** and press **Enter**.
8. Press **F11**.

Downloading and installing MBR tools for ESXi hosts

If you have an ESXi host, you must download and install the version of the MBR (master boot record) tools for ESXi. The MBR tools enable you to detect and correct misaligned disk partitions for guest operating systems. These tools must be installed and run directly on the ESXi host. Before you install them, you must extract them from the .tar file into the root directory on the ESXi host.

Before you begin

You must be able to open a console connection to the ESXi host.

> **Tip:** The MBR tools can only be used when the virtual machine (VM) is powered off. If you want to perform online alignments on VMFS-based datastores without having to take your VM down, you can use the Optimization and Migration capability. In that case, you do not need to download the MBR tools.

Steps

Follow these steps:

1. Open the vSphere Client and log into your vCenter Server.
2. Select a Datacenter in the **Inventory** panel, and then select the **IBM N series** tab.
3. In the Monitoring and Host Configuration capability, select the **Tools** panel.
4. Under **MBR Tools**, click the **Download (For ESXi 4.x and ESXi 5.x)** button.

   Make sure you download the MBR Tools for ESXi. If you download the wrong MBR tools file, the tools will **not** work.
5. When the File Download dialog is displayed, click **Save**.
6. **(ESXi 4.x)** If you are using ESXi 4.x, manually enable the ESXi shell and SSH so that you can use the scp command to copy the files to the correct directories if needed.

ESXi 4.x does not enable the ESXi shell and SSH by default. You can enable these options from the physical host or from the vCenter. The following steps enable these options from the vCenter.

Tip: vCenter creates a configuration alert for each ESXi host that has the options enabled.

To enable the ESXi shell, perform the following steps:

1. From vCenter, highlight the appropriate ESXi host.
2. Go to the **Configuration** Tab.
3. In the left pane under **Software**, select **Security Profile**.
4. Select **Properties** from the **Services** pane.
5. Highlight the **ESXi Shell** service and select **Options**.
6. Select **Start and Stop with Host**.
7. Click **Start**.

To enable the ESXi SSH, perform the following steps:

1. From vCenter, highlight the appropriate ESXi host.
2. Go to the **Configuration** Tab.
3. In the left pane under **Software**, select **Security Profile**.
4. Select **Properties** from the **Services** pane.
5. Highlight the **SSH service** and select **Options**.
6. Select **Start and Stop with Host**.
7. Click **Start**.
8. Copy the MBR tools for ESXi file to the root (/) directory of the ESXi host. If you are using ESXi 4.x, use the Troubleshooting Console. If you are using ESXi 5.x, use the Technical Service Console.

   You might need to open ESXi firewall ports to enable copying the tools to the host.

   Tip: The MBR tools libraries must be located in specific directories on the host. Be sure to download the file to the root directory of the ESXi host.

9. Extract the files by entering the following command:

   ```
   tar -zxf mbrtools_esxi.tgz
   ```

   If you did not download the file to the root directory, you must manually move the files to that directory.

   Tip: ESXi does not support -P with the tar command.

After you finish

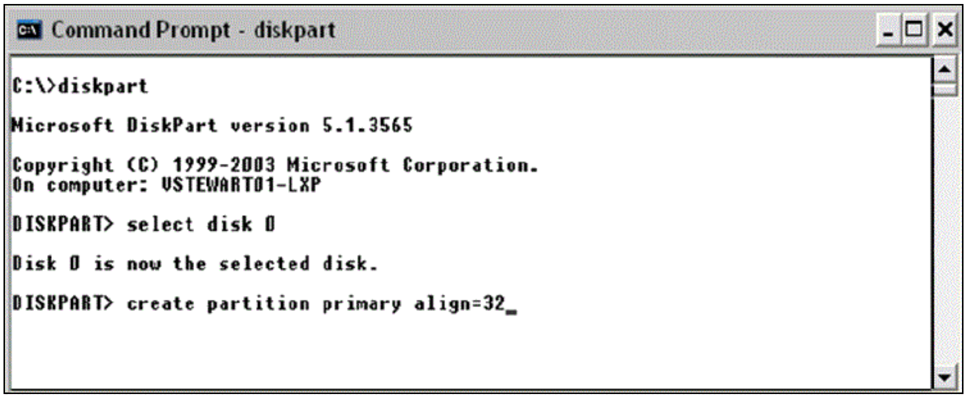
Run the `mbralign` tool to check and fix the partition alignment.

Linux VMs that boot using the GRUB boot loader require the following steps after MBRAlign has been run.

1. Connect a Linux CD or CDROM ISO image to the Linux VM.
2. Boot the VM.
3. Select to boot from the CD.
4. Execute GRUB setup to repair the boot loader, when appropriate.

Creating properly aligned partitions for new VMs

Virtual disks can be formatted with the correct offset at the time of creation by simply booting the VM before installing an operating system and manually setting the partition offset. For Windows guest operating systems, consider using the Windows Preinstall Environment boot CD or the alternative “live DVD” tools. See Figure 34-44 as an example.



```
C:\>diskpart

Microsoft DiskPart version 5.1.3565

Copyright (C) 1999-2003 Microsoft Corporation.
On computer: VSTEWART01-LXP

DISKPART> select disk 0

Disk 0 is now the selected disk.

DISKPART> create partition primary align=32_
```

Figure 34-44 Running diskpart to set a proper starting partition offset

To set up the starting offset, complete the following steps:

1. Boot the VM with the Microsoft WinPE CD.
2. Select **Start**, select **Run**, and enter `diskpart`
3. Enter `select disk 0`
4. Enter `create partition primary align=32`
5. Reboot the VM with the WinPE CD.
6. Install the operating system as normal.

You can also create properly aligned VMDKs with `fdisk` from an ESX console session.

34.9.5 Windows VM file system performance

If your VM is not acting as a file server, consider implementing the following change to your VMs, which disables the access time updates process in the Microsoft Windows NT File System (NTFS). This change reduces the amount of IOPS occurring within the file system.

Reducing IOPS in the file system

To make the proposed change, complete the following steps:

1. Log into a Windows VM.
2. Click **Start** → **Run**, and enter CMD.
3. Enter:

   ```
   fsutil behavior set disablelastaccess 1
   ```

Disk defragmentation utilities

VMs stored on N series storage arrays should not use disk defragmentation utilities because the WAFL file system is designed to optimally place and access data at a level below the guest operating system (GOS) file system.

34.10 VSC commands

You can use the Virtual Storage Console command-line interface to perform specific Backup and Recovery capability tasks.

All VSC commands can be performed by using either the GUI or the CLI, with some exceptions. For example, only the creation of scheduled jobs and their associated retention policies and single file restore can be performed through the GUI.

Remember the following general information about the commands:

- VSC commands are case-sensitive.
- There are no privilege levels; any user with a valid user name and password can run all commands.

You can launch the Virtual Storage Console CLI by using the desktop shortcut or the Windows Start menu. Double-click the VSC CLI desktop icon or navigate to **Start** → **All Programs** → **IBM** → **Virtual Storage Console** → **IBM N series VSC CLI**.

34.11 Scripting

VSC provides users the ability to run pre, post, and failure backup phase scripts as stated in the previous section. These scripts are any executable process on the operating system in which the VSC is running. When defining the backup to run, the pre, post, and failure backup scripts can be chosen by using either the VSC GUI or CLI. The scripts must be saved in the <VSC Installation>\smvi\server\scripts\ directory. Each chosen script runs as a pre, post, and failure backup script.

From the GUI, you can select multiple scripts by using the backup creation wizard or when editing an existing backup job as shown in Figure 34-17 on page 499. The UI lists all files found in the `<VSC Installation>\smvi\server\scripts\` directory. VSC runs the scripts before creating the VMware snapshots and after the cleanup of VMware snapshots.

When VSC starts each script, a progress message is logged indicating the start of the script. When the script completes, or is terminated by SAN volume controller because it was running too long, a progress message is logged. It indicates the completion of the script and states if the script was successful or failed. If a script is defined for a backup but is not found in the scripts directory, a message is logged stating that the script cannot be found.

The VSC maintains a global configuration value to indicate the amount of time that a script can execute. After a script runs for this length of time, the script is terminated by the VSC to prevent run-away processing by scripts. If VSC must terminate a script, it is implicitly recognized as a failed script and might force termination of the VSC backup in the pre-backup phase.

With the default settings, VSC waits for up to 30 minutes for each script to complete in each phase. This default setting can be configured by using the following entry in the `<VSC Installation>\smvi\server\etc\smvi.override` file:

```
smvi.script.timeout.seconds=1800
```

VSC backup scripts receive input from the environment variables. This way, the input can be sent in a manner that avoids CLI line length limits. The set of variables varies based on the backup phase.

35

Consistency groups

This chapter describes the N series support for Snapshot consistency groups.

Consistency group support is provided in N series Data ONTAP 7.2 and later and allows creation of consistent Snapshot backups across multiple volumes and controllers.

We also explain the need for application backup consistency, as well as the methods and software available for using N series consistency groups.

The following topics are covered:

- Snapshot backup and application consistency
- Consistency groups
- How to use consistency groups

35.1 Snapshot backup and application consistency

Enabling recovery from any type of array-based backup (Snapshot) solution requires some planning to ensure application-level data consistency and recoverability. If the Snapshot backup is created (or replicated) in an inconsistent state, then it might not be recoverable.

The methods of achieving application Snapshot consistency are described next.

35.1.1 Application-consistent

The following method is one possibility for achieving application Snapshot consistency:

- Consistent Snapshots are created after applications are gracefully shut down, quiesced, or put in hot backup mode:
  - It is generally the preferred backup mode as it captures the application in a “clean state” and provides for restart without any further recovery steps.

    Due to the overhead involved in application integration, this method is best suited to creating infrequent Snapshot backups (such as RPO=1 day, every X hours)
  - It provides a LOW RISK for recovery, because the Snapshot was created from a “known good” application state.
- Various tools can be used to drive application-consistent Snapshot backup, including the following tools:
  - N series SnapManager (for common applications such as Exchange, Oracle, or VMWare)
  - N series SnapCreator (for applications without SnapManager support)
  - N series SnapDrive (for OS support such as Windows, Solaris, or Linux)

    SnapDrive includes support for consistency groups, such that Snapshot copies that span multiple volumes are self consistent.
  - Custom backup script (such as to perform pre/post actions around an application backup)

35.1.2 Crash-consistent

The following method is another possibility for achieving application Snapshot consistency:

- It creates crash-consistent Snapshots WITHOUT coordinating with applications. A consistency group must be defined to ensure that write ordering is maintained for dependent writes in Snapshot copies across multiple volumes and/or controllers.
  - This method captures the application in a “dirty-but-consistent state”, which can then be recovered using the application’s own recovery mechanisms (such as automatic transaction log roll-forward or roll-back during application restart)

    By avoiding the overhead involved in application integration, this method is well suited to creating frequent Snapshot backups (such as RPO=5 min).
  - Assuming that the application can recover from what is effectively a crashed state; this also provides a LOW RISK for recovery. It is because the Snapshot was created from a “self/crash-consistent” application state.
- When an application spans multiple volumes or controllers, achieving a Snapshot that is crash-consistent requires coordination to make sure that the set of volume Snapshots are consistent with each other.

  - An example of this type of application would be a database that references documents on an external file system. A consistency group enabled backup ensures that the file system and database are captured at a mutually consistent point in time.
- The NAS controllers provide a consistency group (CG) feature that can be used to capture crash-consistent Snapshot copies that span multiple volumes and/or controllers.
  - Consistency group Snapshots briefly suspend application I/O while Snapshots are occurring to ensure consistency across multiple volumes and/or NAS controllers.
  - This feature can be called directly via script, integrated into Protection Manager Policies, and is also embedded in other N series solutions such as SnapDrive, SnapManager, and SnapCreator.

35.1.3 Non-consistent

The following method is another possibility for achieving application Snapshot consistency:

- This option is described here only as a counter point to the previous methods.
- It is NOT ADVISED and might result in an UNRECOVERABLE Snapshot backup:
  - It creates Snapshot copies WITHOUT coordinating with the application or preserving multi-volume write ordering.

    It results in a HIGH RISK for recovery, because the Snapshot was created from an unknown application state, possibly with no inter-volume consistency.
  - However, if an application has all of its data (control files, data files, online redo logs, and archived logs) contained within a single NAS volume, then a Snapshot copy of that single volume WILL provide a crash-consistent copy (see 35.1.2, “Crash-consistent” on page 524.

    In this specific case, host or storage coordination is not necessary to provide crash-consistency.

> **Tip:** It is not uncommon for a complex environment to use a mixture of application-consistent and crash-consistent Snapshot backups. An example might be daily application-consistent Snapshots of a database’s data files, complemented with frequent crash-consistent (also known as “blind”) Snapshots of the database transaction logs.

35.2 Consistency groups

The consistency group feature is available in Data ONTAP 7.2 and later. A consistency group is a grouping of a set of volumes that must be managed as a single logical entity. The functional objective of a consistency group is to provide storage-based, crash-consistent checkpoints from which an application can restart. These checkpoints are performed without interaction and coordination of the source application.

A checkpoint represents a collection of Snapshot copies, one Snapshot copy per volume, for all volumes defined in a consistency group. This collection of Snapshot copies is not the same as a regular group of Snapshot copies, where each copy is created independently of each volume. This special collection of Snapshot copies, the checkpoint, has some distinct characteristics:

- The copy of volumes occurs as an atomic operation.
- The resulting Snapshot copy preserves write ordering across all volumes for dependent writes.

Writes are dependent if an application issues a write based on the success or acknowledgement of the previous write. In case 1 of a dependent write (Figure 35-1), the application only issues W(d+1) after it has received the acknowledgement for W(d). In case 2, the application issues W(d+1) after it reads the newly written data (performed by W(d)) even if it has not received a success signal for W(d). In the case of an independent write, all writes are issued without waiting for individual acknowledgments. Hence, W(i), W(i+1), and W(i+2) might be issued all at the same time.

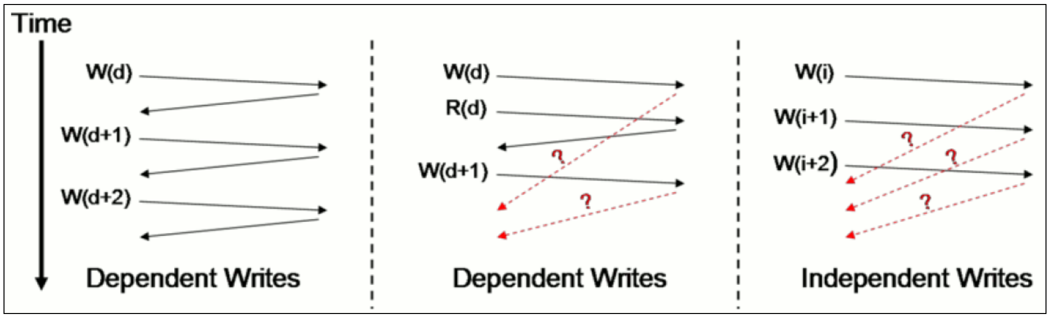


Figure 35-1 Dependent writes/implicit write ordering

35.2.1 Consistency group architecture overview

The overall framework of consistency group is based on a set of core Data ONTAP consistency group APIs. The CG APIs enable agents such as SnapManager for Oracle to create crash-consistent or restore-consistent checkpoints.

The internal operation of consistency group can be expressed by the following high-level sequence of actions:

1. The agent issues a start-checkpoint call to all participating controllers.
2. The controllers fence write access on volumes in a consistency group.
3. The controllers prepare a Snapshot copy of all volumes.
4. The agent receives fence-success from all participating controllers and issues a commit-checkpoint to all controllers.
5. Upon receiving a commit-checkpoint from the agent, controllers commit the Snapshot-creates in all volumes.
6. The controllers unfence all volumes in the consistency group.

The **`cg-start`** and **`cg-commit`** calls are closely related and operate as a pair. The **`cg-commit`** call must follow **`cg-start`** within a certain time interval. This time interval depends on the timeout argument specified in **`cg-start`**.

Figure 35-2 shows the high-level view of the layers involved in the CG architecture. Agent domain refers to any external application, such as SnapDrive, SnapManager, or a custom Perl script, that uses CG APIs to coordinate and create consistent Snapshot copies.

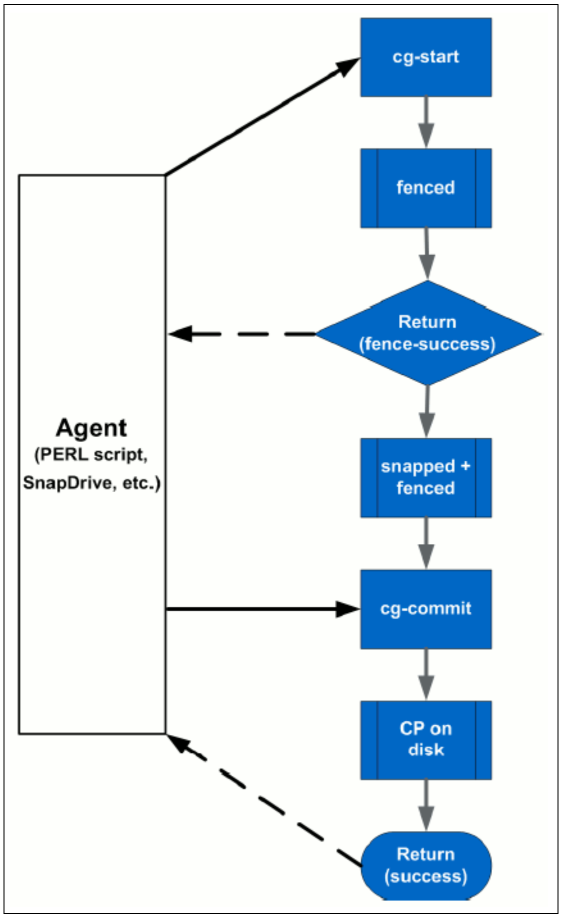


Figure 35-2 High-level flow diagram of consistency group

35.2.2 CG Primitives: APIs

A set of APIs is made available to support the core consistency group primitives. These APIs are used by the N series application integration tools, such as SnapDrive and SnapCreator, but can also be called in custom scripts.

There are three particular APIs (also known as ZAPIs in the N series context) that are used for managing consistency groups. Two of the three APIs, **`cg-start`** and **`cg-commit`**, are highly fundamental to the creation of a CG Snapshot copy.

1. **`cg-start`**

   Starts the checkpoint cycle for externally synchronized checkpoints in the controller. This operation fences the specified volumes and returns **`success`** if successful, then the call starts a snap create operation in these volumes. If the API returns success, this operation must be followed by a call to **`cg-commit`**.

   It is an asynchronous, nonblocking call.

2. **`cg-commit`**

   Commits the Snapshot copies that were started during the preceding **`cg-start`** call that returned the **`cg-id`** key, and unfences the volumes that were fenced.

   It is a synchronous, blocking call.

3. **cg-delete**

   Deletes the Snapshot copies associated with a CG checkpoint in this controller.

The **cg-start** and **cg-commit** calls are closely related and operate as a pair. The **cg-commit** call must follow **cg-start** within a certain time interval. This time interval depends on the timeout argument specified in **cg-start**.

The CG process is controlled by an external application, such as SnapDrive, SnapManager, or a custom Perl script, that uses CG APIs to coordinate and create consistent Snapshot copies.

35.3 How to use consistency groups

The CG APIs enable external applications to manage and create a crash-consistent Snapshot copy of multiple volumes residing on a single controller or spanning across two or more controllers. Applications such as SnapDrive, SnapManager, and SnapCreator use the CG APIs to provide crash-consistent Snapshot copies.

While it is beyond the context of this paper to describe these applications in full detail, it is necessary to understand how these applications make use of consistency groups.

35.3.1 SnapDrive

SnapDrive helps to automate storage provisioning and simplify storage management in N series storage environments. It offers the following key benefits:

- Simplified provisioning from the host with a protocol-agnostic approach
- Automated backup and restores
- Snapshot management with OS/application-consistent Snapshot copies

SnapDrive provides a layer of abstraction between an application running on the host operating system and the underlying N series storage systems. It has the unique advantage of understanding the host operating system, the file system, and N series storage. Thus, SnapDrive has the ability to coordinate operations between these core layers.

One notable feature of SnapDrive lies in the area of Snapshot copy management. SnapDrive can provide host/application-consistent Snapshot copies and crash-consistent Snapshot copies. When Snapshot copies of an application data set that spans multiple volumes or storage systems are required, SnapDrive provides consistency by synchronizing the data in the file system cache and freezing I/O operations to the requested LUNs.

What makes SnapDrive interesting is the mechanism with which it suspends I/O. SnapDrive will freeze I/O operations to the LUNs at the host OS/file-system layer or directly within the N series storage system. If a storage-based I/O fencing mechanism is elected, then SnapDrive relies on the native consistency group feature of Data ONTAP.

In an environment where all participating controllers support consistency groups, SnapDrive will use a Data ONTAP consistency group as the preferred (default) method to capture multicontroller/volume Snapshot copies.

If using a consistency group is not possible or is perhaps undesirable, such as when one of the controllers does not support CG or more than one application is sharing LUNs off a single volume where volume I/O fencing might have an effect on other applications, then SnapDrive resorts to the I/O suspension capability offered at the host level.

In such a scenario, SnapDrive attempts to use the lowest-level freezing mechanism available on the host to enforce consistency, which might require direct interaction with the file systems or volume managers. This approach suggests that consistency must always be enforced at the lowest possible level of the storage stack to minimize the duration of I/O suspension.

SnapDrive dramatically simplifies the use of consistency groups. When the file specs or file system dictates a Snapshot copy that spans multiple volumes and controllers, and all target controllers support consistency group (Data ONTAP 7.2 and higher), SnapDrive automatically recognizes this requirement and creates consistency groups to enable crash-consistent Snapshot copies. No change to the SnapDrive syntax is necessary to take advantage of consistency groups.

35.3.2 SnapCreator

SnapCreator is a backup plug-in framework for integrating applications with N series Snapshot technology. It currently supports application-consistent backups for Oracle, MySQL, MaxDB, Sybase, DB2, and IBM Lotus Notes®. Through Perl APIs, SnapCreator integrates seamlessly with other N series technologies such as Snapshot, SnapVault, SnapMirror, LUN cloning, volume cloning, SnapDrive, and so on.

SnapCreator supports consistency groups to create consistent Snapshot copies across multiple volumes. It uses the primary CG APIs, **`cg-start`** and **`cg-commit`**, to accomplish the task.

35.3.3 Custom scripting

Because native consistency group APIs are made available, any external third-party applications or custom scripts can take advantage of the consistency group feature of Data ONTAP.

The IBM N series Manageability SDK includes a number of example scripts (in various languages) to create a consistency group Snapshot. This script can be used “as is” or can be customized to suit your specific environment.

For a demonstration of using the sample script, see Example 35-1.

Example 35-1 Using the CG sample script

```
C:\temp> cg_operation.pl
cg_operation.pl <filer> <user> <password> <operation> <value1>[<value2>] [<volum
es>]
<filer>     -- Filer name
<user>      -- User name
<password>  -- Password
<operation> -- Operation to be performed: cg-start/cg-commit
<value1>    -- Depends on the operation
[<value2>]  -- Depends on the operation
[<volumes>] --List of volumes.Depends on the operation

C:\temp> cg_operation.pl myfiler myuser mypassword cg-start snapname urgent vol1
vol2 vol3
Consistency Group operation started successfully with cg-id=1

C:\temp> cg_operation.pl myfiler myuser mypassword cg-commit 1
Consistency Group operation commited successfully
```

On the N series controller, the consistent Snapshots can be seen on the referenced volumes (Example 35-2).

Example 35-2 List the CG Snapshots

```
nsim1> snap list
Volume vol0

[...trimmed output...]

Volume vol1
working....

  %/used       %/total  date          name
----------  ----------  ------------  --------
 27% (27%)    0% ( 0%)  Jun 04 16:43  snapname

Volume vol2
working....

  %/used       %/total  date          name
----------  ----------  ------------  --------
 27% (27%)    0% ( 0%)  Jun 04 16:43  snapname

Volume vol3
working....

  %/used       %/total  date          name
----------  ----------  ------------  --------
 27% (27%)    0% ( 0%)  Jun 04 16:43  snapname
```

For more details on the Manage ONTAP SDK, contact your IBM technical representative:

http://support.netapp.com/NOW/download/software/nmsdk/5.0/

Part 6

Storage management

In this part of the book, we explain how to manage your storage using N series storage management software such as OnCommand, System Manager, and the Command Line interface. The following topics are covered:

- Remote management
- Command line administration
- N series System Manager
- AutoSupport
- OnCommand

36

Remote management

This chapter introduces the remote management networks through Remote LAN Module (RLM) and Baseboard Management Controller (BMC). You can manage your storage system remotely by using a remote management device, which can be the Service Processor (SP), the Remote LAN Module (RLM), or the Baseboard Management Controller (BMC), depending on the storage system model.

The RLM is included in the following systems:

N5200, N5300, N5500, N5600
N6040, N6060, N6070
N7600, N7700, N7800, N7900

The BMC is included in the 20xx systems (N3300, N3400, N3600).

The SP is included in all other systems (N62xx, N7950T).

The Service Processor (SP) is a remote management device that is included in the N62xx and N7950T systems. It enables you to access, monitor, and troubleshoot the storage system remotely.

The Remote LAN Module (RLM) is a remote management card that is supported on the N6000 and N7000 systems. The RLM provides remote platform management capabilities, including remote access, monitoring, troubleshooting, logging, and alerting features.

The Baseboard Management Controller (BMC) is a remote management device that is built into the motherboard of the N3000 systems. It provides remote platform management capabilities, including remote access, monitoring, troubleshooting, logging, and alerting features.

The following topics are covered:

- Remote LAN Module (RLM)
- Baseboard Management Controller (BMC)
- Service Processor (SP)
- CLI administration

36.1 Remote LAN Module (RLM)

The RLM command line interface (CLI) commands enable you to remotely access and administer the storage system and diagnose error conditions. Also, the RLM extends AutoSupport capabilities by sending alerts and notifications through an AutoSupport message.

The RLM its a management card intended for remote Management of midrange and high-end N series systems.

Tip: Place the interface on a management VLAN or separate network from the user data access path.

The RLM stays operational regardless of the operating state of the storage system. It is powered by a standby voltage, which is available as long as the storage system has input power to at least one of the storage system's power supplies. Therefore you can logon to the RLM card even if a system is unavailable.

The RLM has a single temperature sensor to detect ambient temperature around the RLM board. Data generated by this sensor is not used for any system or RLM environmental policies. It is only used as a reference point that might help you troubleshoot storage system issues. For example, it might help a remote system administrator determine if a system was shut down due to an extreme temperature change in the system.

The following diagram illustrates how you can access the storage system and the RLM (Figure 36-1).

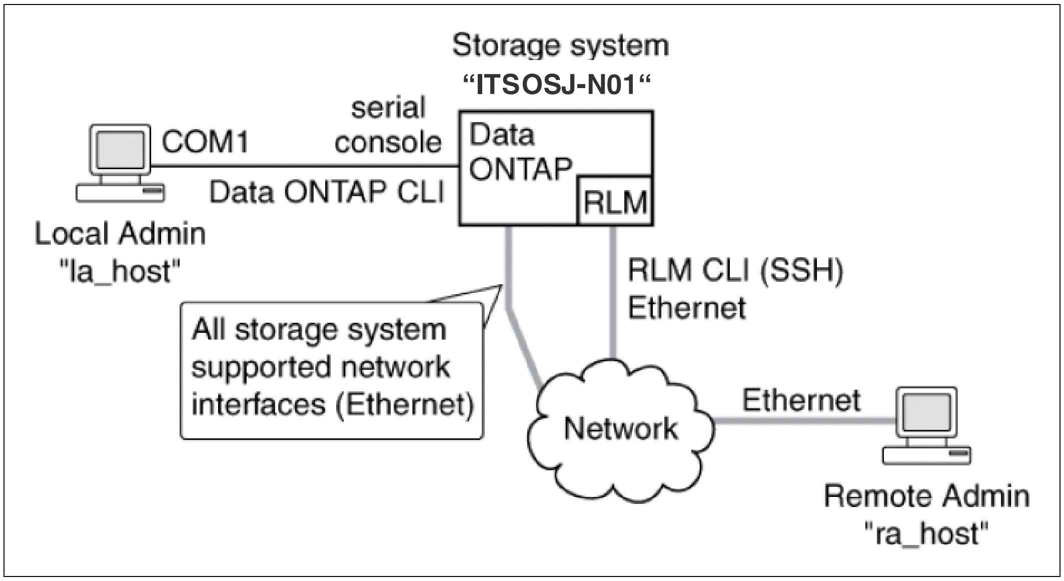


Figure 36-1 RLM diagram

Without the RLM, you can locally access the storage system through the serial console or from an Ethernet connection using any supported network interface. You use the Data ONTAP CLI to administer the storage system.

With the RLM, you can remotely access the storage system through the serial console. The RLM is directly connected to the storage system through the serial console. You use the Data ONTAP CLI to administer the storage system and the RLM.

With the RLM, you can also access the storage system through an Ethernet connection using a secure shell client application. You use the RLM CLI to monitor and troubleshoot the storage system.

If you have a data center configuration where management traffic and data traffic are on separate networks, you can configure the RLM on the management network.

The commands in the RLM CLI enable you to remotely access and administer the storage system and diagnose error conditions. Also, the RLM extends AutoSupport capabilities by sending alerts and notifications through an AutoSupport message.

Using the RLM CLI commands, you can perform the following tasks:

- Remotely administer the storage system by using the Data ONTAP CLI through the RLM's system console redirection feature
- Remotely access the storage system and diagnose error conditions, even if the storage system has failed, by performing the following tasks:
  - View the storage system console messages, captured in the RLM's console log
  - View storage system events, captured in the RLM's system event log
  - Initiate a storage system core dump
  - Power-cycle the storage system (or turn it on or off)
  - Reset the storage system
  - Reboot the storage system

The RLM extends AutoSupport capabilities by sending alerts and "**`down system`**" or "down filer" notifications through an AutoSupport message when the storage system goes down, regardless of whether the storage system can send AutoSupport messages. Other than generating these messages on behalf of a system that is down, and attaching additional diagnostic information to AutoSupport messages, the RLM has no effect on the storage system's AutoSupport functionality. The AutoSupport configuration settings and message content behavior of the RLM are inherited from Data ONTAP.

Tip: The RLM does not rely on the autosupport.support.transport option to send notifications. The RLM uses the Simple Mail Transport Protocol (SMTP).

In addition to AutoSupport messages, the RLM generates SNMP traps to configured trap hosts for all "down system" or "down filer" events, if SNMP is enabled for the RLM.

The RLM has a nonvolatile memory buffer that stores up to 4,000 system events in a system event log (SEL) to help you diagnose system issues. The event list from the SEL is automatically sent by the RLM to specified recipients in an AutoSupport message. The records contain the following data:

- Hardware events detected by the RLM, for example, system sensor status about power supplies, voltage, or other components
- Errors (generated by the storage system or the RLM) detected by the RLM, for example, a communication error, a fan failure, a memory or CPU error, or a "boot image not found" message
- Critical software events sent to the RLM by the storage system, for example, a system panic, a communication failure, an unexpected boot environment prompt, a boot failure, or a user triggered "down system" as a result of issuing the system reset or system power cycle command.

The RLM monitors the storage system console regardless of whether administrators are logged in or connected to the console. When storage system messages are sent to the console, the RLM stores them in the console log. The console log persists as long as the RLM has power from either of the storage system's power supplies. Because the RLM operates with standby power, it remains available even when the storage system is power-cycled or turned off.

Hardware-assisted takeover is available on systems that support the RLM and have the RLM modules set up. For more information about hardware-assisted takeover, see the Data ONTAP 7- Mode High-Availability Configuration Guide.

The RLM supports the SSH protocol for CLI access from UNIX clients and PuTTY for CLI access from PC clients. Telnet and RSH are not supported by the RLM, and system options to enable or disable them have no effect on the RLM.

36.1.1 Ways to configure the RLM

Before using the RLM, you must configure it for your storage system and network. You can configure the RLM when setting up a new storage system with RLM already installed, after setting up a new storage system with RLM already installed, or when adding an RLM to an existing storage system.

You can configure the RLM by using one of the following methods:

- Initializing a storage system that has the RLM pre-installed:

  When the storage system setup process is complete, the **`rlm setup`** command runs automatically. For more information about the entire setup process, see the *Data ONTAP 7-Mode Software Setup Guide*.

- Running the Data ONTAP setup script:

  The setup script ends by initiating the **`rlm setup`** command.

- Running the Data ONTAP **`rlm setup`** command:

  When the **`rlm setup`** script is initiated, you are prompted to enter network and mail host information.

In order to access the storage system through the RLM interface, an account must have **`login-sp`** capability. The storage system **`Administrators`** group has **`login-sp`** capability by default. If the **`root`** local account is disabled, then the **`naroot`** account is disabled and a local user with **`login-sp`** capability can log in to the RLM. It is available on the N62x0 and N7950Tplatforms.

> **Tip:** Determine that the RLM firmware is version 4 or above.

In version 4 firmware, only ssh2 is enabled. The ssh protocol on the RLM is part of the RLM's kernel operating system and therefore segmented for the implementation of ssh by the Data ONTAP operating system.

> **Action:** Disable the **`root`** account and utilize accounts that are members of the storage systems **`Administrators`** group to manage the storage system through the RLM.

> **Tip:** The RLM ignores the **`ssh.idle.timeout`** option and the **`console.timeout`** option. The settings for these options do not have any effect on the RLM.

Attention: RLM firmware 4.0 will track failed SSH login attempts from an IP address. If more than 5 repeated login failures are detected from an IP address in any 10-minute period, the RLM will stop all communication with that IP address for the next 15 minutes. Normal communication will resume after 15 minutes, but repeated login failures are detected again, communication will again be suspended for the next 15 minutes.

For detailed information about the RLM and its capabilities, see the “The Remote LAN Module” section of the *Data ONTAP 8.0 7-Mode System Administration Guide*.

36.1.2 Prerequisites for configuring the RLM

Before you configure the RLM, you must gather information about your network and your AutoSupport settings.

Here is the information you need to gather:

- Network information:

  You can configure the RLM using DHCP or static addressing. If you are using an IPv4 address for the RLM, you need the following information:

  - An available static IP address
  - The netmask of your network
  - The gateway of your network:
    If you are using IPv6 for RLM static addressing, you need the following information:
  - The IPv6 global address
  - The subnet prefix for the RLM
  - The IPv6 gateway for the RLM

- AutoSupport information:

  The RLM sends event notifications based on the following AutoSupport settings:

  - autosupport.to
  - autosupport.mailhost

  It is best that you configure at least the autosupport.to option before configuring the RLM. Data ONTAP automatically sends AutoSupport configuration to the RLM, allowing the RLM to send alerts and notifications through an AutoSupport message to the system administrative recipients specified in the **`autosupport.to`** option. You are prompted to enter the name or the IP address of the AutoSupport mail host when you configure the RLM.x

36.1.3 Setting up the RLM

If you are running RLM firmware version 4.0 or later, and you have enabled IPv6 for Data ONTAP, you have the option to configure the RLM for only IPv4, for only IPv6, or for both IPv4 and IPv6. Disabling IPv6 on Data ONTAP also disables IPv6 on the RLM.

Attention: If you disable both IPv4 and IPv6, and if DHCP is also not configured, the RLM has no network connectivity.

Steps for setting up RLM

Follow these steps:

1. At the storage system prompt, enter *one* of the following commands:

   ```
   setup
   rlm setup
   ```

   If you enter **setup**, the rlm setup script starts automatically after the setup command runs.

2. When the RLM setup asks you whether to configure the RLM, enter **y**.
3. Enter one of the following choices when the RLM setup asks you whether to enable DHCP on the RLM.
   - To use DHCP addressing, enter **y**.
   - To use static addressing, enter **n**.

   > **Tip:** DHCPv6 servers are not currently supported.

4. If you do not enable DHCP for the RLM, the RLM setup prompts you for static IP information.

   Provide the following information when prompted:

   - The IP address for the RLM:

     > **Tip:** Entering 0.0.0.0 for the static IP address disables IPv4 for the RLM.

   - The netmask for the RLM
   - The IP address for the RLM gateway
   - The name or IP address of the mail host to use for AutoSupport

5. If you enabled IPv6 for Data ONTAP, and your RLM firmware version is 4.0 or later, the RLM supports IPv6. In this case, the RLM setup asks you whether to configure IPv6 connections for the RLM. Enter one of the following choices:
   - To configure IPv6 connections for the RLM, enter **y**.

     > **Tip:** You can use the `rlm status` command to find the RLM version information.

   - The subnet prefix for the RLM
   - The IPv6 gateway for the RLM

     > **Tip:** You cannot use the RLM setup to enable or disable the IPv6 router-advertised address for the RLM. However, when you use the ip.v6.ra_enable option to enable or disable the IPv6 router-advertised address for Data ONTAP, the same configuration applies to the RLM.

     For information about enabling IPv6 for Data ONTAP or information about global, link-local, and router-advertised addresses, see the *Data ONTAP 7-Mode Network Management Guide*.

6. At the storage system prompt, enter the following command to verify that the RLM network configuration is correct:

   ```
   rlm status
   ```

7. At the storage system prompt, enter the following command to verify that the RLM AutoSupport function is working properly:

   ```
   rlm test autosupport
   ```

   > **Tip:** The RLM uses the same mail host information that Data ONTAP uses for AutoSupport.

The following message is a sample of the output that Data ONTAP displays:

```
Sending email messages via SMTP server at mailhost@companyname.com. If
autosupport.enable is on, then each email address in autosupport.to should
receive the test message shortly.
```

Connecting to the storage system console from the RLM

The RLM's system console command enables you to log in to the storage system from the RLM.

Follow these steps:

1. Enter the following command at the RLM prompt:

   ```
   system console
   ```

   The message **“`Type Ctrl-D to exit`”** appears.

2. Press Enter to see the storage system prompt.

   You use **`Ctrl-D`** to exit from the storage system console and return to the RLM CLI.

   The storage system prompt appears, and you can enter Data ONTAP commands.

Using online help at the RLM CLI

The RLM online help displays all RLM commands and options when you enter the question mark (?) or help at the RLM prompt.

Follow these steps:

1. To display help information for RLM commands, enter one of the following choices at the RLM prompt:

   ```
   help
   ?
   ```

Example 36-1 shows the RLM CLI online help.

Example 36-1 RLM - Help

```
RLM-itsosj-n01> help
date - print date and time
exit - exit from the RLM command line interface
events - print system events and event information
help - print command help
priv - show and set user mode
rlm - commands to control the RLM
rsa - commands for Remote Support Agent
system - commands to control the system
version - print RLM version
```

Power cycle the N series through RLM

Turn the storage system on or off, or perform a power cycle (which turns system power off and then back on)

```
system power {on | off | cycle}
```

> **Tip:** Standby power stays on, even when the storage system is off. During power-cycling, a brief pause occurs before power is turned back on.

> **Attention:** Using the `system power` command to turn off or power-cycle the storage system might cause an improper shutdown of the system (also called a `dirty` shutdown) and is not a substitute for a graceful shutdown using the Data ONTAP `halt` command.

Display status for each power supply, such as presence, input power, and output power:

```
system power status
```

36.2 Baseboard Management Controller (BMC)

The Baseboard Management Controller (BMC) is a remote management device that is built into the motherboard of N3x00 storage systems. It provides remote platform management capabilities, including remote access, monitoring, troubleshooting, logging, and alerting features.

The Baseboard Management Controller (BMC) is a remote management device that is built into the motherboard of the N3x00 systems. It provides remote platform management capabilities, including remote access, monitoring, troubleshooting, logging, and alerting features.

The BMC stays operational regardless of the operating state of the system. Both the BMC and its dedicated Ethernet NIC use a standby voltage for high availability. The BMC is available as long as the system has input power to at least one of the system's power supplies.

The BMC monitors environmental sensors, including sensors for the temperature of the system's nonvolatile memory (NVMEM) battery, motherboard, and CPU, and for the system's voltage level. When an environmental sensor has reached a critically low or critically high state, the BMC generates AutoSupport messages and shuts down the storage system. The data generated by the sensors can be used as a reference point to help you troubleshoot storage system issues. For example, it can help a remote system administrator determine if a system was shut down due to an extreme temperature change in the system.

The BMC also monitors non-environmental sensors for the status of the BIOS, power, CPU, and serial-attached SCSI (SAS) disks. These sensors are recorded by the BMC to assist support personnel.

You use the BMC sensors show command to display the ID and the current state of the sensors monitored by the BMC, and you use the BMC sensors search command to display information of a sensor by its ID.

Figure 36-2 illustrates how you can access the storage system and the BMC.

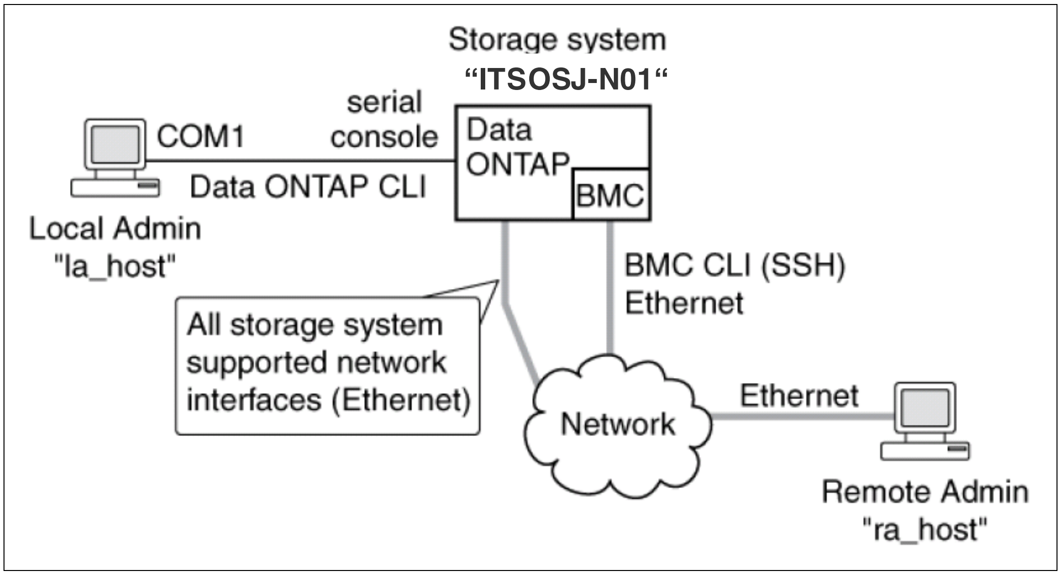


Figure 36-2 BMC diagram

With the BMC, you can access the storage system in these ways:

- Through an Ethernet connection using a secure shell client application:
  You use the BMC CLI to monitor and troubleshoot the storage system.
- Through the serial console:
  You use the Data ONTAP CLI to administer the storage system and the BMC.

If you have a data center configuration where management traffic and data traffic are on separate networks, you can configure the BMC on the management network.

The commands in the BMC CLI enable you to remotely access and administer the storage system and diagnose error conditions. Also, the BMC extends AutoSupport capabilities by sending alerts and notifications through an AutoSupport message.

The BMC provides the following remote management capabilities for the storage system. You use the BMC CLI commands to perform the following tasks:

- Administer the storage system using the Data ONTAP CLI by using the BMC’s system console redirection feature
- Access the storage system and diagnose error conditions, even if the storage system has failed, by performing the following tasks:
  - View the storage system console messages, captured in the BMC's system console log
  - View storage system events, captured in the BMC's system event log
  - Initiate a storage system core dump
  - Power-cycle the storage system (or turn it on or off)
    For instance, when a temperature sensor becomes critically high or low, Data ONTAP triggers the BMC to shut down the motherboard gracefully. The system console becomes unresponsive, but you can still press **`Ctrl-G`** on the console to access the BMC CLI. You can then use the **`system power on`** or **`system power cycle`** command from the BMC to power on or power cycle the system.

- Monitor environmental and non-environmental sensors for the controller module and the NVMEM battery.

You can switch between the primary and the backup firmware hubs to assist in bootup and recovery from a corrupted image in the storage system's primary firmware hub.

The BMC extends AutoSupport capabilities by sending alerts and "**down system**" or "**down filer**" notifications through an AutoSupport message when the storage system goes down, regardless of whether the storage system can send AutoSupport messages. Other than generating these messages on behalf of a system that is down, and attaching additional diagnostic information to AutoSupport messages, the BMC has no effect on the storage system's AutoSupport functionality. The system's AutoSupport behavior is the same as it would be without BMC installed. The AutoSupport configuration settings and message content behavior of the BMC are inherited from Data ONTAP.

Tip: The BMC does not rely on the **autosupport.support.transport** option to send notifications. The BMC uses the Simple Mail Transport Protocol (SMTP).

The BMC has a nonvolatile memory buffer that stores up to 512 system events in a system event log (SEL) to help you diagnose system issues. The records contain the following data:

- Hardware events detected by the BMC, for example, system sensor status about power supplies, voltage, or other components
- Errors (generated by the storage system or the BMC) detected by the BMC, for example, a communication error, a fan failure, a memory or CPU error, or a "**boot image not found**" message
- Critical software events sent to the BMC by the storage system, for example, a system panic, a communication failure, an unexpected boot environment prompt, a boot failure, or a user triggered "down system" as a result of issuing the **system reset** or **system power cycle** command.

The BMC monitors the storage system console regardless of whether administrators are logged in or connected to the console. When storage system messages are sent to the console, the BMC stores them in the system console log. The system console log persists as long as the BMC has power from either of the storage system's power supplies. Because the BMC operates with standby power, it remains available even when the storage system is power-cycled or turned off.

36.2.1 Ways to configure the BMC

Before using the BMC, you must configure it for your storage system and network. You can configure the BMC when setting up a new storage system with BMC already installed or after setting up a new storage system with BMC already installed.

You can configure the BMC by using one of the following methods:

- Initializing a storage system that has the BMC:

  When the storage system setup process is complete, the **bmc setup** command runs automatically. For more information about the entire **setup** process, see the Data ONTAP 7-Mode Software Setup Guide.
- Running the Data ONTAP setup script:

  The setup script ends by initiating the **bmc setup** command.
- Running the Data ONTAP **bmc setup** command:

  When the **bmc setup** script is initiated, you are prompted to enter network and mail host information.

The BMC supports the SSH protocol for CLI access from UNIX clients and PuTTY for CLI access from PC clients. Telnet and RSH are not supported on the BMC, and system options to enable or disable them have no effect on the BMC.

> **Tip:** The BMC ignores the ssh.idle.timeout option and the console.timeout option. The settings for these options do not have any effect on the BMC.

You can use **`root`**, **`naroot`** or **`Administrator`** to log into the BMC. These users have access to all commands available on the BMC. The password for all three account names is the same as the Data ONTAP root password. You cannot add additional users to the BMC.

> **Tip:** The BMC uses the Data ONTAP root password (even if the root account is disabled) to allow access over the LAN with SSH. To access the BMC via SSH, you must configure the Data ONTAP root password. BMC accepts passwords that are no more than 16 characters.

> **Action:** Take great care when using the BMC Management Port on the storage system. Set a strong password on the root account, disable the root account, and reset the root password on a regular basis.

For detailed information about the BMC and its capabilities, see the “The Baseboard Management Controller” section of the *Data ONTAP 8.0 7-Mode System Administration Guide*.

36.2.2 Prerequisites for configuring the BMC

Before you configure the BMC, you need to gather information about your network and your AutoSupport settings.

You need to gather the following information:

- Network information:

  You can configure the BMC using DHCP or static addressing.

- If you are using DHCP addressing, you need the BMC’s MAC address.

  > **Tip:** If you do not provide a valid BMC MAC address, an EMS message shows up to remind you during system bootup or when you use the **`bmc status`** or the **`setup`** command.

- If you are using a static IP address, you need the following information:
  - An available static IP address
  - The netmask of your network
  - The gateway of your network
- AutoSupport settings

  The BMC sends event notifications based on the following Data ONTAP AutoSupport settings:

  - **`autosupport.to`**
  - **`autosupport.mailhost`**

It is best to configure at least the **autosupport.to** option before configuring the BMC. Data ONTAP automatically sends AutoSupport configuration to the BMC, allowing the BMC to send alerts and notifications through an AutoSupport message to the system administrative recipients specified in the **autosupport.to** option. You are prompted to enter the name or the IP address of the AutoSupport mail host when you configure the BMC.

36.2.3 Setting up the BMC

You can use the **setup** command or the bmc **setup** command to configure the BMC.

Before you begin, it is best to configure AutoSupport before configuring the BMC. Data ONTAP automatically sends AutoSupport configuration to the BMC, enabling the BMC to send alerts and notifications through an AutoSupport message.

Steps for setting up the BMC

Follow these steps:

1. At the storage system prompt, enter one of the following commands:

```
setup
bmc setup
```

If you enter **setup**, the **bmc setup** script starts automatically after the **setup** command runs.

2. When the BMC setup asks you whether to configure the BMC, enter **y**.
3. Enter one of the following choices when the BMC setup asks you whether to enable DHCP on the BMC:
   - To use DHCP addressing, enter **y**.
   - To use static addressing, enter **n**.

   Tip: DHCPv6 servers are not currently supported.

4. If you do not enable DHCP for the BMC, the BMC setup prompts you for static IP information. Provide the following information when prompted:
   - The IP address for the BMC
   - The netmask for the BMC
   - The IP address for the BMC gateway
   - The name or IP address of the mail host to use for AutoSupport

   Tip: Currently, you can use only IPv4 addresses to connect to the BMC.

5. Enter the Address Resolution Protocol (ARP) interval for the BMC when you are prompted.
6. If the BMC setup prompts you to reboot the system, enter the following command at the storage system prompt:

   ```
   reboot
   ```

7. At the storage system prompt, enter the following command to verify that the BMC's network configuration is correct:

   ```
   bmc status
   ```

8. At the storage system prompt, enter the following command to verify that the BMC AutoSupport function is working properly:

   `bmc test autosupport`

   > **Tip:** The BMC uses the same mail host information that Data ONTAP uses for AutoSupport. The **`bmc test autosupport`** command requires that you set up the **`autosupport.to`** option properly.

   You have successfully set up the BMC AutoSupport function when the system displays the following output:

   `Please check ASUP message on your recipient mailbox.`

Connecting to the storage system console from the BMC

You can access the BMC CLI from a console session by pressing **`Ctrl-G`**.

Press **`Ctrl-G`** at the storage system prompt to access the BMC CLI.

> **Tip:** Entering **`system console`** at the BMC prompt returns you to the console session.

Only one administrator can log in to an active BMC CLI session at a time. However, the BMC allows you to open both a BMC CLI session and a separate, BMC-redirected system console session simultaneously.

When you use the BMC CLI to start a system console session, the BMC CLI is suspended, and the system console session is started. When you exit the system console session, the BMC CLI session resumes.

The BMC prompt is displayed as **`bmc shell ->.`**

Using online help at the BMC CLI

The BMC help displays all the available BMC commands when you enter the question mark (**`?`**) or **`help`** at the BMC prompt.

Example 36-2 shows the BMC CLI help.

Example 36-2 BMC online help

```
bmc shell -> ?
exit
bmc config
bmc config autoneg [enabled|disabled]
bmc config dhcp [on|off]
bmc config duplex [full|half]
bmc config gateway [gateway]
...
```

If a command has sub-commands, you can see them by entering the command name after the **`help`** command, as shown in Example 36-3).

Example 36-3   BMC help events

```
bmc shell -> help events
events all Print all system events
events info Print SEL(system event log)
information
events latest [N] Print N latest system events
events oldest [N] Print N oldest system events
events search [attr=N] Search for events by
attribute/value pair
events show [N] Print event N
```

Power cycle the N series through BMC

Turn the storage system on or off, perform a power cycle (which turns system power off and then back on), or display the power status:

```
system power {on | off | cycle | status}
```

Tip: Standby power stays on, even when the storage system is off. During power-cycling, there is a brief pause before power is turned back on.

Attention: Using the **`system power`** command to turn off or power-cycle the storage system might cause an improper shutdown of the system (also called a dirty shutdown) and is not a substitute for a graceful shutdown using the Data ONTAP **`halt`** command.

36.3 Service Processor (SP)

The Service Processor (SP) command line interface (CLI) commands enable you to remotely access and administer the storage system and diagnose error conditions. Also, the SP extends AutoSupport capabilities by sending alerts and notifications through an AutoSupport message.

The SP provides the following capabilities:

- The SP enables you to access the storage system remotely to diagnose, shut down, power-cycle, or reboot the system, regardless of the state of the storage controller.
  The SP is powered by a standby voltage, which is available as long as the system has input power to at least one of the system's power supplies.

  The SP is connected to the system through the serial console. You can log in to the SP by using a Secure Shell client application from an administration host. You can then use the SP CLI to monitor and troubleshoot the system remotely. In addition, you can use the SP to access the system console and run Data ONTAP commands remotely.

  You can access the SP from the system console or access the system console from the SP. The SP allows you to open both an SP CLI session and a separate system console session simultaneously.

  For instance, when a temperature sensor becomes critically high or low, Data ONTAP triggers the SP to shut down the motherboard gracefully. The system console becomes unresponsive, but you can still press **`Ctrl-G`** on the console to access the SP CLI. You can then use the **`system power on`** or **`system power cycle`** command from the SP to power on or power cycle the system.

- The SP monitors environmental sensors and logs system events to help you take timely and effective service actions in the event that a system problem occurs.

  The SP monitors the system temperatures, voltages, currents, and fan speeds. When an environmental sensor has reached an abnormal condition, the SP logs the abnormal readings, notifies Data ONTAP of the issue, and sends alerts and “down system” notifications as necessary through an AutoSupport message, regardless of whether the storage system can send AutoSupport messages.

  Other than generating these messages on behalf of a system that is down and attaching additional diagnostic information to AutoSupport messages, the SP has no effect on the storage system’s AutoSupport functionality. The AutoSupport configuration settings and message content behavior are inherited from Data ONTAP.

  > **Tip:** The SP does not rely on the autosupport.support.transport option to send notifications. The SP uses the Simple Mail Transport Protocol (SMTP).

  If SNMP is enabled for the SP, the SP generates SNMP traps to configured trap hosts for all “down system” events.

  The SP also logs system events such as boot progress, Field Replaceable Unit (FRU) changes, Data ONTAP-generated events, and SP command history.

- The SP has a nonvolatile memory buffer that stores up to 4,000 system events in a system event log (SEL) to help you diagnose system issues.

  The SEL stores each audit log entry as an audit event. It is stored in onboard flash memory on the SP. The event list from the SEL is automatically sent by the SP to specified recipients through an AutoSupport message.

  The SEL contains the following data:

  - Hardware events detected by the SP, for example, system sensor status about power supplies, voltage, or other components
  - Errors detected by the SP, for example, a communication error, a fan failure, or a memory or CPU error
  - Critical software events sent to the SP by the storage system, for example, a system panic, a communication failure, a boot failure, or a user-triggered “down system” as a result of issuing the SP system reset or system power cycle command

- The SP monitors the system console regardless of whether administrators are logged in or connected to the console.

  When system messages are sent to the console, the SP stores them in the console log. The console log persists as long as the SP has power from either of the storage system’s power supplies. Because the SP operates with standby power, it remains available even when the storage system is power cycled or turned off.

- Hardware-assisted takeover is available on systems that support the SP and have the SP configured.

  For more information about hardware-assisted takeover, see the Data ONTAP 7-Mode High-Availability Configuration Guide.

Figure 36-3 illustrates access to the storage system and the SP.

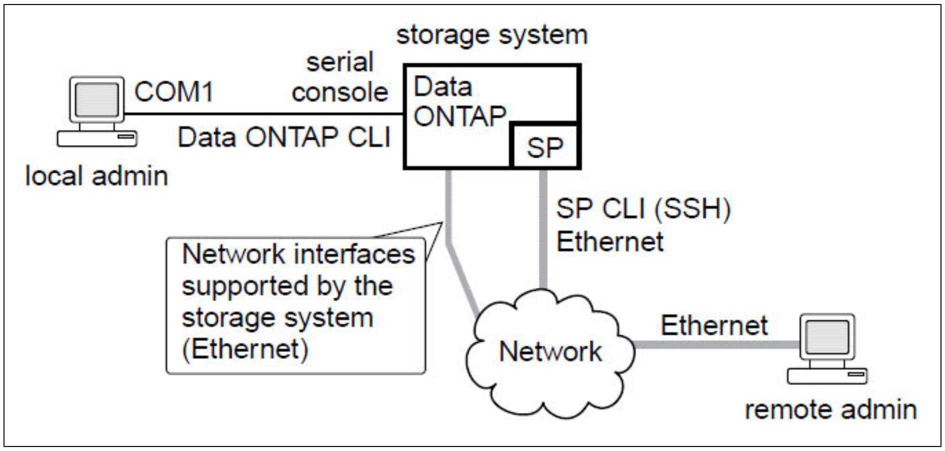


Figure 36-3 Service Processor diagram

36.3.1 Ways to configure the SP

Configuring the SP for your storage system and network enables you to log in to the SP over the network. It also enables the SP to send an AutoSupport message in the event of a problem. You can configure the SP when you set up a new storage system. You can also configure the SP by running the **`setup`** or the **`sp setup`** command.

On a storage system that comes with the SP, you can configure the SP by using one of the following methods:

- Initializing a new storage system:

  When you power on a storage system for the first time, the setup command begins to run automatically. When the storage system setup process is complete, the sp setup command runs automatically and prompts you for SP configuration information. For more information about the system setup process, see the Data ONTAP 7-Mode Software Setup Guide.

- Running the Data ONTAP setup command:

  If you want to change both system setup and SP configuration, you use the **`setup`** command. The **`system setup`** process ends by initiating the **`sp setup`** command.

- Running the Data ONTAP sp setup command directly:

  If the storage system has been set up and you want to reconfigure only the SP, you can use the **`sp setup`** command, which omits system setup and prompts you directly for SP configuration information.

In order to access the storage system through the SP interface an account must have **`login-sp`** capability. The storage system Administrators group has **`login-sp`** capability by default. If the **`root`** local account is disabled, then the **`naroot`** account is disabled and a local user with login-sp capability can log in to the SP.

SP firmware 1.2 and later will track failed SSH login attempts from an IP address. If more than 5 repeated login failures are detected from an IP address in any 10-minute period, the RLM will stop all communication with that IP address for the next 15 minutes. Normal communication will resume after 15 minutes, but, if repeated login failures are detected again, communication will again be suspended for the next 15 minutes.

For detailed information about the SP and its capabilities, see the “Using the service processor for remote system management” section of the *Data ONTAP 8.1 7-Mode System Administration Guide*.

36.3.2 Prerequisites for configuring the SP

You need the following information about your network and AutoSupport settings when you configure the SP:

- Network information:

  If you are using an IPv4 address for the SP, you need the following information:

  - An available static IP address for the SP
  - The netmask of your network
  - The gateway IP of your network

  If you are using IPv6 for SP static addressing, you need the following information:

  - The IPv6 global address
  - The subnet prefix for the SP
  - The IPv6 gateway IP for the SP

  For information about network interfaces and management, see the *Data ONTAP 7-Mode Network Management Guide*.

- AutoSupport information:

  The SP sends event notifications based on the settings of the following AutoSupport options:

  - **autosupport.to**
  - **autosupport.mailhost**

  At the minimum, consider configuring the **autosupport.to** option before configuring the SP. Data ONTAP automatically sends AutoSupport configuration to the SP, allowing the SP to send alerts and notifications through an AutoSupport message to the system administrative recipients specified in the **autosupport.to** option. You are prompted to enter the name or the IP address of the AutoSupport mail host when you configure the SP.

36.3.3 Setting up the SP

You can use the **setup** command or the **sp setup** command to configure the SP, depending on whether you want to change the system setup besides configuring the SP. You can configure the SP to use either a static or a DHCP address.

If you have enabled IPv6 for Data ONTAP, you have the option to configure the SP for only IPv4, for only IPv6, or for both IPv4 and IPv6. Disabling IPv6 on Data ONTAP also disables IPv6 on the SP. If you disable both IPv4 and IPv6, and if DHCP is also not configured, the SP will not have network connectivity.

The firewall for IPv6 is configured to accept a maximum of 10 Internet Control Message Protocol (ICMP) packets in a one-second interval. If your system has management software that frequently performs diagnostic checks, this limit can cause false positive errors to be generated. Consider increasing the software's ping interval or tuning the software's report to expect the false positive errors caused by the ICMP limit.

Steps for setting up the SP

Follow these steps:

1. At the storage system prompt, enter one of the following commands:

   – **`setup`**

     If you want to change both system setup and SP configuration, you use the **`setup`** command. When the storage system setup process is complete, the **`sp setup`** command runs automatically and prompts you for SP configuration information.

     For information about system setup, see the *Data ONTAP 7-Mode Software Setup Guide*.

   – **`sp setup`**

     If the storage system has been set up and you want to configure only the SP, you use the **`sp setup`** command, which omits system setup and prompts you directly for SP configuration information.

2. When the SP setup asks you whether to configure the SP, enter **`y`**.
3. Enter one of the following choices when the SP setup asks you whether to enable DHCP on the SP:

   – To use DHCP addressing, enter **`y`**.

     Tip: The SP supports DHCPv4 but not DHCPv6.

   – To use static addressing, enter **`n.`**

4. If you do not enable DHCP for the SP, provide the following static IP information when the SP setup prompts you to enter it:

   – The IP address for the SP

     Tip: Entering **`0.0.0.0`** for the static IP address disables IPv4 for the SP. If you enter **`0.0.0.0`** for the static IP address, you must enter **`0.0.0.0`** also for the netmask and the IP address for the SP gateway.

   – The netmask for the SP

   – The IP address for the SP gateway

   – The name or IP address of the mail host to use for AutoSupport

5. If you have enabled IPv6 for Data ONTAP, the SP supports IPv6. In this case, the SP setup asks you whether to configure IPv6 connections for the SP. Enter one of the following choices:

   – To configure IPv6 connections for the SP, enter **`y`**.

   – To disable IPv6 connections for the SP, enter **`n`**.

6. If you choose to configure IPv6 for the SP, provide the following IPv6 information when the SP setup prompts you to enter it:

   – The IPv6 global address:

     Even if no IPv6 global address is assigned for the SP, the link-local address is present on the SP. The IPv6 router-advertised address is also present if the **`ip.v6.ra_enable`** option is set to **`on`**.

   – The subnet prefix for the SP

   – The IPv6 gateway for the SP

> **Tip:** You cannot use the SP setup to enable or disable the IPv6 router-advertised address for the SP. However, when you use the **`ip.v6.ra_enable`** option to enable or disable the IPv6 router-advertised address for Data ONTAP, the same configuration applies to the SP

For information about enabling IPv6 for Data ONTAP or information about global, link-local, and router-advertised addresses, see the *Data ONTAP 7-Mode Network Management Guide*.

7. At the storage system prompt, enter the following command to verify that the SP network configuration is correct:

   ```
   sp status
   ```

8. At the storage system prompt, enter the following command to verify that the SP AutoSupport function is working properly:

   ```
   sp test autosupport
   ```

   > **Tip:** The SP uses the same mail host information that Data ONTAP uses for AutoSupport.

The following message is a sample of the output Data ONTAP displays:

```
Sending email messages via SMTP server at mailhost@companyname.com. If
autosupport.enable is on, then each email address in autosupport.to should
receive the test message shortly.
```

Accessing the SP from the system console

You can access the SP from the system console to perform monitoring or troubleshooting tasks.

To access the SP CLI from the system console, press **`Ctrl-G`** at the storage system prompt. The SP prompt appears, indicating that you have access to the SP CLI.

> **Tip:** You can press **`Ctrl-D`** and then press Enter to return to the system console.

Only one administrator can log in to an active SP CLI session at a time. However, the SP allows you to open both an SP CLI session and a separate system console session simultaneously.

The SP prompt appears with SP in front of the hostname of the storage system. For example, if your storage system is named itsosj-n01, the storage system prompt is itsosj-n01> and the prompt for the SP session is **`SP itsosj-n01>.`**

If an SP CLI session is currently open, you or another administrator with privileges to log in to the SP can close the SP CLI session and open a new one. This feature is convenient if you logged in to the SP from one computer and forgot to close the session before moving to another computer, or if another administrator takes over the administration tasks from a different computer.

You can use the SP's **`system console`** command to connect to the storage system console from the SP. You can then start a separate SSH session for the SP CLI, leaving the system console session active. When you press **`Ctrl-D`** to exit from the storage system console, you automatically return to the SP CLI session. If an SP CLI session already exists, the following message appears:

```
User username has an active console session.
Would you like to disconnect that session, and start yours [y/n]?
```

If you enter **`y`**, the session owned by **`username`** is disconnected and your session is initiated. This action is recorded in the SP's system event log.

Using online help at the SP CLI

The SP online help displays the SP CLI commands and options when you enter the question mark (**`?`**) or **`help`** at the SP prompt.

1. To display help information for the SP commands, enter one of the following at the SP prompt:

   `help`

   `?`

Example 36-4 shows the SP CLI online help.

Example 36-4   SP help

```
SP itsosj-n01> help
date - print date and time
exit - exit from the SP command line interface
events - print system events and event information
help - print command help
priv - show and set user mode
sp - commands to control the SP
rsa - commands for Remote Support Agent
system - commands to control the system
version - print SP version
```

2. To display help information for the option of an SP command, enter the following command at the SP prompt:

   `help SP_command`

Example 36-5 shows the SP CLI online help for the SP events command.

Example 36-5   SP help events

```
SP itsosj-n01> help events
events all - print all system events
events info - print system event log information
events newest - print newest system events
events oldest - print oldest system events
events search - search for and print system events
```

Power cycle the N series through SP

Turn the storage system on or off, or perform a power cycle (turning system power off and then back on):

```
system power{on|off|cycle}
```

Tip: The standby power stays on to keep the SP running without interruption. During the power cycle, a brief pause occurs before power is turned back on.

Attention: Using the `system power` command to turn off or power-cycle the storage system might cause an improper shutdown of the system (also called a dirty shutdown) and is not a substitute for a graceful shutdown using the Data ONTAP `halt` command

36.4 CLI administration

In this section, we introduce various ways to administer N series systems through CLI (command line interface). We cover the following network protocols:

- telnet
- SSH
- RSH
- Audit Logging

On storage systems shipped with Data ONTAP 8.0 or later, secure protocols are enabled and non-secure protocols are disabled by default. SecureAdmin is set up automatically on storage systems shipped with Data ONTAP 8.0 or later. These systems have the following default security settings:

- Secure protocols (including SSH, SSL, and HTTPS) are enabled by default.
- Non-secure protocols (including RSH, Telnet, FTP, and HTTP) are disabled by default.

We advise that you configure and enable SecureAdmin. immediately after initially setting up Data ONTAP. This preferred practice enables SSH and SSL encryption for secure administration of the N series storage system. Also, use only the SSH version 2 protocol and using SSH public key authentication. For more information about SecureAdmin, see the Data ONTAP System Administration Guide.

Although SSH version 1 is supported in Data ONTAP, it has known exploitable vulnerabilities that can be prevented only by using SSH version 2 exclusively. SSH public keys provide a stronger and more granular method of SSH access to N series storage systems.

In Data ONTAP version 7.3.4 the option to disable sslv2 (`options ssl.v2.enable off)` was added.

Audit logging

An audit log is a record of commands executed at the console through a telnet shell or an SSH shell or by using the **rsh** command. All the commands executed in a source file script are also recorded in the audit log. Administrative HTTP operations, such as those resulting from the use of System Manager or another SDK ONTAPIR application, are logged. All login attempts to access the storage system, with success or failure, are also audit logged.

In addition, changes made to configuration and registry files are audited. Read-only APIs by default are not audited but you can enable auditing with the **auditlog.readonly_api.enable** option. By default, Data ONTAP is configured to save an audit log. The audit log data is stored in the /**etc/log** directory in a file called auditlog. For configuration changes, the audit log shows the following information:

- Which configuration files were accessed
- When the configuration files were accessed
- What was changed in the configuration files

For commands executed through the console, a telnet shell, or an SSH shell or by using the **rsh** command, the audit log shows the following information:

- Which commands were executed
- Who executed the commands
- When the commands were executed

You can access the audit log files using your NFS or CIFS client, or HTTP(s).

For detailed information about audit logging and its capabilities, see the “Audit logging” section of the *Data ONTAP 8.x 7-Mode System Administration Guide*.

> **Tip:** There is no option to extend the maximum audit log entry character limit. The limit is 511 characters.

Preferred practice

Audit logging must always be enabled. This logs administrative access from the console and from remote shell sessions. Log file size depends on corporate security policy, but it must be large enough to record several days' worth of administrative usage at a minimum. A preferred practice is to set log file size to a large value (several megabytes, at least) and then adjust the size after monitoring growth of the log file.

Some corporate security policies might dictate central log collection and analysis. Data ONTAP does support the sending of Data ONTAP audit logs to an external syslog host. Although we do not advise using an external syslog as a preferred practice, consider this option as a way to collect historical data; see **syslog.conf** for details.

37

Command line administration

This chapter introduces various ways to administer N series systems through the command line interface (CLI).

The following network protocols are covered:

- Introduction to CLI administration
- Telnet
- SSH
- RSH

37.1 Introduction to CLI administration

On storage systems shipped with Data ONTAP 8.0 or later, secure protocols are enabled and non-secure protocols are disabled by default. SecureAdmin is set up automatically on storage systems shipped with Data ONTAP 8.0 or later. These systems have the following default security settings:

- Secure protocols (including SSH, SSL, and HTTPS) are enabled by default.
- Non-secure protocols (including RSH, Telnet, FTP, and HTTP) are disabled by default.

We advise that you configure and enable SecureAdmin. immediately after initially setting up Data ONTAP. This preferred practice enables SSH and SSL encryption for secure administration of the N series storage system. Also, use only the SSH version 2 protocol and using SSH public key authentication. For more information about SecureAdmin, see the *Data ONTAP System Administration Guide*.

Although SSH version 1 is supported in Data ONTAP, it has known exploitable vulnerabilities that can be prevented only by using SSH version 2 exclusively. SSH public keys provide a stronger and more granular method of SSH access to N series storage systems.

In Data ONTAP version 7.3.4, the option to disable sslv2 (**`options ssl.v2.enable off)`** was added.

37.1.1 Audit logging

An audit log is a record of commands executed at the console through a telnet shell or an SSH shell or by using the **`rsh`** command. All the commands executed in a source file script are also recorded in the audit log. Administrative HTTP operations, such as those resulting from the use of System Manager or another SDK ONTAPIR application, are logged. All login attempts to access the storage system, with success or failure, are also audit logged.

In addition, changes made to configuration and registry files are audited. Read-only APIs by default are not audited but you can enable auditing with the **`auditlog.readonly_api.enable`** option. By default, Data ONTAP is configured to save an audit log. The audit log data is stored in the **`/etc/log`** directory in a file called auditlog. For configuration changes, the audit log shows the following information:

- Which configuration files were accessed
- When the configuration files were accessed
- What was changed in the configuration files

For commands executed through the console, a telnet shell, or an SSH shell or by using the **`rsh`** command, the audit log shows the following information:

- Which commands were executed
- Who executed the commands
- When the commands were executed

You can access the audit log files using your NFS or CIFS client, or HTTP(s).

For detailed information about audit logging and its capabilities, see the “Audit logging” section of the *Data ONTAP 8.x 7-Mode System Administration Guide*.

> **Tip:** There is no option to extend the maximum audit log entry character limit. The limit is 511 characters.

37.1.2 Preferred practice

Audit logging must always be enabled. It logs administrative access from the console and from remote shell sessions. Log file size depends on corporate security policy, but it must be large enough to record several days' worth of administrative usage at a minimum. A preferred practice is to set log file size to a large value (several megabytes, at least) and then adjust the size after monitoring growth of the log file.

Some corporate security policies might dictate central log collection and analysis. Data ONTAP does support the sending of Data ONTAP audit logs to an external syslog host. Although we do not advise using an external syslog as a preferred practice, consider this option as a way to collect historical data; see `syslog.conf` for details.

37.2 Telnet

You can access a storage system from a client through a Telnet session if you enabled Telnet.

37.2.1 Telnet session options

A Telnet session must be reestablished before any of the following options command values take effect:

- autologout.console.enable
- autologout.console.timeout
- autologout.telnet.enableautologout.telnet.timeout
- telnet.distinct.enable

For more information about these options, see the na_options(1) man page.

> **Tip:** Telnet and RSH are not supported on the BMC, and system options to enable or disable them have no effect on the BMC.

Clear text passwords are passed between the client and the storage system.

The `telnet.distinct.enable` option enables making the Telnet and console separate user environments. If it is off, then Telnet and console share a session. The two sessions view each other's inputs/outputs and both acquire the privileges of the last user to log in. If this option is toggled during a Telnet session, then it goes into effect on the next Telnet login. Valid values for this option are `On` or `Off`. This option is set to On if a user belonging to Compliance Administrators is configured and cannot be set to Off until the user is deleted. The default setting is `On`.

You configure a banner message to appear at the beginning of a Telnet session to a storage system by creating a file called `/etc/issue.` The message only appears at the beginning of the session. It is not repeated if there are multiple failures when attempting to log in.

> **Tip:** The `/etc/issue` file can be created from the storage system CLI using the `wrfile` command. For more information about how it is accomplished, see the "Writing a WAFL file" section of the *Data ONTAP 8.x 7-Mode System Administration Guide.*

There are two option settings that control the auto logout of the Telnet session. They are **`autologout.telnet.enable`** and **`autologout.telnet.timeout`**. Auto logout for the Telnet session is enabled by default with a timeout setting of 60 minutes.

> **Suggestions:** If Telnet is used, set the session timeout to a value of 5 minutes and take precautions to ensure that the accounts and passwords are not compromised in transit from the client to the storage controller. Set a banner message through the creation of the **`/etc/issue`** file.

37.2.2 Starting a Telnet session

You need to start a Telnet session to connect to the storage system.

The following requirements must be met before you can connect to a storage system using a Telnet session:

- The **`telnet.enable`** option must be set to **`on`**. You can verify that the option is on by entering the options **`telnet`** command. You set the option to on by entering the options **`telnet.enable on`** command. For more information, see the na_options(1) man page.
- The **`telnet.access`** option must be set so that the protocol access control defined for the storage system allows Telnet access. For more information, see the na_options(1) and na_protocolaccess(8) man pages.

About this task

Only one Telnet session can be active at a time. You can, however, open a console session at the same time a Telnet session is open.

Steps

Follow these steps:

1. Open a Telnet session on a client.
2. Connect to the storage system using its name.
3. If the storage system displays the login prompt, do one of the following actions:
   - To access the storage system with the system account, enter this account name:

     `root`
   - To access the storage system with an alternative administrative user account, enter the appropriate name:

     `username`

     Where **`username`** is the administrative user account.

     The storage system responds with the password prompt.
4. Enter the password for the root or administrative user account.

> **Tip:** If no password is defined for the account, press Enter.

5. When you see the storage system prompt followed by a system message, press Return to get to the storage system prompt.

   Example 37-1 shows the output after initiating a telnet session to an N series system.

Example 37-1 Telnet

```
itsosj-n01> Thu Jun 14 3:26:55 PST [itsosj-n01: telnet_0:info]: root logged in
from host: itsosj_unix01.xxx.yyy.com
```

   Press Enter.

   itsosj-n01>

   Tip: You can abort commands entered through a Telnet session by pressing **Ctrl-C**.

For detailed information about Telnet and its capabilities, see the “Telnet sessions and storage system access” section of the *Data ONTAP 8.x 7-Mode System Administration Guide*.

37.3 SSH

The **secureadmin setup ssh** command configures the SSH server. The administrator specifies the key strength for the RSA host and server keys. The keys can range in strength from 384 to 2,048 bits.

37.3.1 SSH options

If your storage system does not have SSH enabled, you can set up SecureAdmin to enable secure sessions using SSH. A few options enable you to control password-based authentication and public key authentication, control access to a storage system, and assign the port number to a storage system.

SecureAdmin is set up automatically on storage systems shipped with Data ONTAP 8.0 or later.

A post-log-in banner is available for the sshv2 protocol. The banner that is used is read from the **/etc/motd** file. To activate this banner set the option **ssh2.banner.enable** to **On**. This option does not exist until it is created.

Tip: The **/etc/motd** file can be created from the storage system CLI using the **wrfile** command. For more information about how it is accomplished, see the “Writing a WAFL file” section of the “Data ONTAP 8.x 7-Mode System Administration Guide.”

Action: Ensure that ssh1 is disabled; only ssh2 is enabled by default. Then check the status of ssh and ssl using the **secureadmin status** command at the storage system CLI.

The ssh session timeout is defaulted to 600 seconds (10 minutes).

Action: Set the options **ssh.idle.timeout** to a value of **300** (5 minutes).

The **telnet.distinct.enable** option enables making the ssh and console separate user environments.

Action: Set the **telnet.distinct.enable** option to **On**.

For detailed information about SSH and its capabilities, see the “SSH protocol” section of the *Data ONTAP 8.x 7-Mode System Administration Guide*.

For detailed information about the **secureadmin** command, see the “secureadmin” section of the *Data ONTAP 8.0 7-Mode Commands: Manual Page Reference, Volume 1.*

Note that ssh is enabled by invoking the **secureadmin setup ssh** command at the CLI or through the System Manager under <storage controller name> => Configuration => Security => SSH/SSL. Example 37-2 shows how ssh can be set up.

Example 37-2 Secureadmin setup

```
itsosj-n01> secureadmin setup ssh
SSH Setup
---------
Determining if SSH Setup has already been done before...no

SSH server supports both ssh1.x and ssh2.0 protocols.

SSH server needs two RSA keys to support ssh1.x protocol. The host key is
generated and saved to file /etc/sshd/ssh_host_key during setup. The server key is
re-generated every hour when SSH server is running.

SSH server needs a RSA host key and a DSA host key to support ssh2.0 protocol.
The host keys are generated and saved to /etc/sshd/ssh_host_rsa_key and
/etc/sshd/ssh_host_dsa_key files respectively during setup.

SSH Setup will now ask you for the sizes of the host and server keys.
   For ssh1.0 protocol, key sizes must be between 384 and 2048 bits.
   For ssh2.0 protocol, key sizes must be between 768 and 2048 bits.
   The size of the host and server keys must differ by at least 128 bits.

Please enter the size of host key for ssh1.x protocol [768] :
Please enter the size of server key for ssh1.x protocol [512] :
Please enter the size of host keys for ssh2.0 protocol [768] :

You have specified these parameters:
        host key size = 768 bits
        server key size = 512 bits
        host key size for ssh2.0 protocol = 768 bits
Is this correct? [yes]

Setup will now generate the host keys. It will take a minute.
After Setup is finished the SSH server will start automatically.

itsosj-n01> Tue Jun 14 10:00:41 PST [secureadmin.ssh.setup.success:info]: SSH
setup is done and ssh2 should be enabled. Host keys are stored in
/etc/sshd/ssh_host_key, /etc/sshd/ssh_host_rsa_key, and
/etc/sshd/ssh_host_dsa_key.
```

Table 37-1 lists the SSH related options.

Table 37-1 SSH related options

Option	Default	Preferred	Setting / CLI command
ssh.access	*	Hosts or IP range	**options ssh.access host=<hostname>** **options ssh.access host=aa.bb.cc.dd/mm** Refer to the Manual Page Reference, Volume 2 - na_protocolaccess(8), for valid values.
ssh.enable	On	On	**options ssh.enable on**
ssh.passwd_auth.enable	On	On	**options ssh.passwd_auth.enable on**
ssh.idle.timeout	0	60	Controls orphaned connection—disconnect value in seconds. **options ssh.idle.timeout 60**
ssh.port	22	22	**options ssh.port 22**
ssh.pubkey_auth.enable	On	On	**options ssh.pubkey_auth.enable on**
ssh1.enable	Off	Off	**options ssh1.enable off** **s**
sh2.enable	On	On	**options ssh2.enable on**
telnet.distinct.enable	Off	On	Enables making the ssh and the console separate user environments; if set to OFF, ssh and the console will share the session. **options telnet.distinct.enable on**
autologout.telnet.enable	On	On	Enables the automatic disconnect of inactive SSH interactive sessions. **options autologout.telnet.enable on**
autologout.telnet.timeout	60	5	Timeout time in minutes. **options autologout.telnet.timeout 5**

37.3.2 Interactive SSH support for vFiler units

Starting with Data ONTAP 8.1, you can establish interactive SSH sessions with vFiler units. A vFiler unit can have only one active interactive SSH session established at a time. You can also use IPv6 addresses to establish interactive SSH sessions.

Depending on the number of vFiler units allowed on that storage system, there are limits on the number of concurrent interactive SSH sessions that you can run on a storage system.

For more information about limits on the number of interactive SSH sessions, see the *Data ONTAP 8.x 7-Mode MultiStore Management Guide*.

37.4 RSH

Note: SSH will be used to connect via RSH. The previous section discusses SSH in detail.

SSH improves security by providing a means for a storage system to authenticate the client and by generating a session key that encrypts data sent between the client and storage system. SSH performs public-key encryption using a host key and a server key.

Data ONTAP supports password authentication and public-key-based authentication. Data ONTAP does not support the use of a `.rhosts` file or the use of a `.rhosts` file with RSA host authentication.

Data ONTAP supports the following encryption algorithms:

- RSA/DSA 1024 bit
- 3DES in CBC mode
- HMAC-SHA1
- HMAC-MD5

Data ONTAP supports the SSH 1.x protocol and the SSH 2.0 protocol.

Data ONTAP supports the following SSH clients:

- OpenSSH client version 4.4p1 on UNIX platforms
- SSH Communications Security client (SSH Tectia client) version 6.0.0 on Windows platforms
- Vandyke SecureCRT version 6.0.1 on Windows platforms
- PuTTY version 0.6.0 on Windows platforms
- F-Secure SSH client version 7.0.0 on UNIX platforms

SSH uses three keys to improve security:

- Host key:

  SSH uses the host key to encrypt and decrypt the session key. You determine the size of the host key, and Data ONTAP generates the host key when you configure SecureAdmin.

  Tip: SecureAdmin is set up automatically on storage systems shipped with Data ONTAP 8.0 or later.

- Server key:

  SSH uses the server key to encrypt and decrypt the session key. You determine the size of the server key when you configure SecureAdmin. If SSH is enabled, Data ONTAP generates the server key when any of the following events occur:

  - When you start SecureAdmin
  - When an hour elapses
  - When the storage system reboots

- Session key:

  SSH uses the session key to encrypt data sent between the client and storage system. The session key is created by the client. To use the session key, the client encrypts the session key using the host and server keys and sends the encrypted session key to the storage system, where it is decrypted using the host and server keys. After the session key is decrypted, the client and storage system can exchange encrypted data.

You can use an RSH connection to access a storage system from a UNIX client to perform administrative tasks. Before you begin, the **`rsh.enable`** option must be set to **`on`**.

If you access the storage system by using its IPv6 address, the **`ip.v6.enable`** option must be set to **`on`** for the system and the UNIX client you use must support IPv6.

The maximum for concurrent RSH sessions is 24 per system / 4 per vFiler unit.

To access, do one of these tasks:

- If the UNIX host name or the user name you use is not specified in the /etc/hosts.equiv file on the root volume of the storage system, enter the **`rsh`** command in the following format:

  ```
  rsh hostname_or_ip -l username:password command
  ```

- If the UNIX host name and the user name you use are specified in the **`/etc/hosts.equiv`** file on the root volume of the storage system, enter the **`rsh`** command in the following format:

  ```
  rsh hostname_or_ip [-l username] command
  ```

  hostname_or_ip is the host name, IPv4 address, or IPv6 address of the storage system.

  Tip: You can also specify the IP address by using the rsh.access option.

  Here, **`command`** is the Data ONTAP command you want to run over the RSH connection.

Clear text passwords are passed between the client and the storage system.

Tip: Take care when using this protocol to maintain the storage and take precautions so that your passwords and user IDs are not compromised in transit from the client to the storage system.

To disable RSH, enter the following command: **`options rsh.enable off`**

For detailed information about RSH and its capabilities, see the “How to access a storage system using a Remote Shell connection” section of the *Data ONTAP 8.x 7-Mode System Administration Guide*.

Tip: Telnet and RSH are not supported on the BMC, and system options to enable or disable them have no effect on the BMC.

RSH (Remote shell) is disabled by default. If these services are not required in your infrastructure, we advise that they be disabled.

Table 37-2 contains the **`rsh`** services that are on by default and the preferred settings.

Table 37-2 Non-secure rsh settings

Option	Default	Preferred	Setting / CLI Command
rsh.access	legacy	Host or None	**`options rsh.access host=–`** Refer to the Manual Page Reference, Volume 2 - na_protocolaccess(8), for valid values.
rsh.enable	off	off	**`options rsh.enable off`**

Clear text passwords are passed between the client and the storage system.

Action: Take care when using this protocol to maintain the storage and take precautions to ensure that your passwords and user IDs are not compromised in transit from the client to the storage system.

For detailed information about RSH and its capabilities, see the “How to access a storage system using a Remote Shell connection” section of the *Data ONTAP 8.x 7-Mode System Administration Guide*.

N series System Manager

This chapter describes the IBM N series System Manager (NSM) software.

System Manager is a Web-based graphical management interface that enables you to perform many common tasks:

- Configure and manage storage objects, such as disks, aggregates, volumes, qtrees, and quotas.
- Configure protocols, such as CIFS and NFS and provision file sharing.
- Configure protocols, such as FC and iSCSI for block access.
- Create and manage vFiler units.
- Set up and manage SnapMirror relationships.
- Manage HA configurations and perform takeover and giveback operations.
- Perform cluster management, storage node management, and vServer management operations in a cluster environment.

System Manager replaces FilerView as the tool to manage storage systems running Data ONTAP 8.1.

The following topics are covered:

- Introduction to N series System Manager (NSM)
- Installing the N series System Manager
- Getting started with NSM

38.1 Introduction to N series System Manager (NSM)

System Manager enables you to manage storage systems and storage objects, such as disks, volumes, and aggregates. System Manager is a Web-based graphical management interface to manage common functions related to storage systems from a Web browser.

You can download and install System Manager on a desktop or laptop that is running a Windows or a Linux operating system.

The N series system manager is supported on the following platforms:

- Microsoft Windows:
  - Windows XP
  - Windows Vista
  - Windows 7
  - Windows Server 2003
  - Windows Server 2008
- Linux:
  - Red Hat Enterprise Linux 5
  - SUSE Linux Enterprise Server 11

Because NSM is a Java application with a web-browser GUI, it might be possible, though it is not officially supported, to run it on other platforms, such as Ubuntu Linux or Mac OS X.

You can use System Manager to manage storage systems and HA configurations running the following versions of Data ONTAP:

- Data ONTAP 7.x (starting from 7.2.3)
- Data ONTAP 8.x 7-Mode

You can also manage N series gateway systems.

> **Tip:** System Manager replaces FilerView as the tool to manage storage systems running Data ONTAP 8.1.

38.2 Installing the N series System Manager

Before you install System Manager, you must download the software from the IBM N series Support Site:

http://www.ibm.com/systems/support/storage/nas/

The software is available to all registered clients as a complimentary download.

> **Important:** IBM clients must register their N series system with the IBM support website to be granted access for complimentary downloads and software updates.

38.2.1 Installing NSM on Windows

You can install System Manager on your Windows system by using the wizard-based installer:

1. Run the System Manager setup (.exe) file from the directory where you downloaded and saved the software.
2. Follow the on-screen prompts to complete your installation.

You can now launch System Manager and start managing your storage systems and objects.

38.2.2 Installing System Manager on Linux

You can install System Manager on your Linux system by using Red Hat Package Manager (RPM):

1. Install System Manager by performing the appropriate action:
   - From the Linux desktop:

     Double-click the RPM package file.
   - From the command line interface:

     Enter the following command:

     ```
     rpm -i downloaded_rpm_file_name
     ```
2. Optional: Check the progress of the installation by using the following command:

   ```
   rpm -ivv downloaded_rpm_file_name
   ```

You can launch System Manager and start managing your storage systems and objects.

38.3 Getting started with NSM

The System Manager user interface enables you to configure your storage systems and manage storage objects such as disks, aggregates, volumes, quotas, qtrees, and LUNs; protocols such as CIFS, NFS, iSCSI, and FCP; vFiler units; vServers; HA configurations; V-Series systems; and SnapMirror relationships.

For more information about how to configure and manage your storage systems from System Manager, see the *System Manager Help*. You can access the Help in PDF format from the IBM Support Site or from the Help provided with the System Manager software.

http://www-01.ibm.com/support/docview.wss?uid=ssg1S7003448

Before you can start managing a storage system from System Manager, you have to add it to System Manager.

38.3.1 Starting NSM

Although NSM is a Java application with a web browser GUI, it is started just like a native application. It will automatically start the web browser when the NSM java daemon initializes.

The NSM application can be started from the Windows desktop menu:

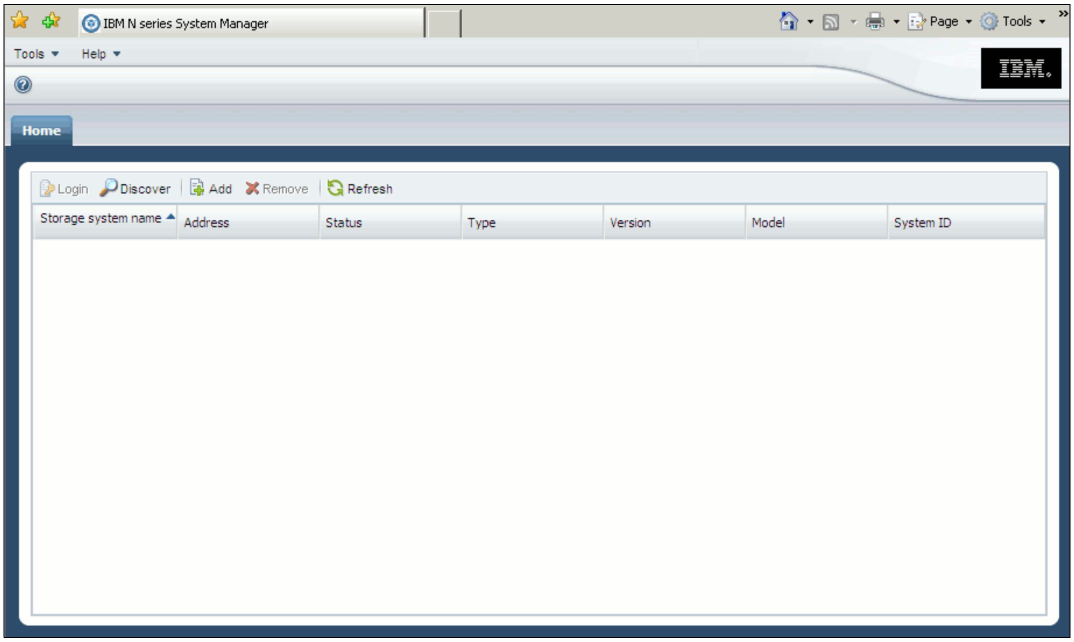
Start Menu → Programs → IBM → N series OnCommand System Manager → IBM N series OnCommand System Manager 3.0

Similar to the Windows example, you can start the NSM application from the Linux desktop menu.

Alternatively, you can also start NSM from the command line:

```
cd /opt/IBM/oncommand_system_manager/3.0
java -jar SystemManager.jar
```

In either Windows or Linux, it will then spawn a Web browser running the NSM interface (see Figure 38-1).



Storage system name ▲	Address	Status	Type	Version	Model	System ID

Figure 38-1 Initial NSM interface

Next, you need to add a storage system to the NSM interface.

38.3.2 Adding a storage system

Before you can use System Manager to manage your storage systems and objects, you have to add them to System Manager. You can also add storage systems that are in a high-availability (HA) configuration.

Check the following items before you begin:

- Your storage systems must be running a supported version of Data ONTAP.
- If your storage system is running a Data ONTAP release in the Data ONTAP 8.0 release family, SSL must be enabled on the storage system.

If you are adding one of the storage systems from an HA pair, the partner node is automatically added to the list of managed systems. If a high-availability partner node is down, you can add the working storage node.

Perform the following steps to add a storage system:

1. From the **Home** tab, click **Add**.
2. Type the fully qualified DNS host name, or the IPv4 address of the storage system.

   You can specify the IPv6 address of the storage system, if you are adding a system that is running a supported version of Data ONTAP 7-Mode.
3. Click the **More** arrow.
4. Select the method for discovering and adding the storage system or cluster:
   - SNMP:

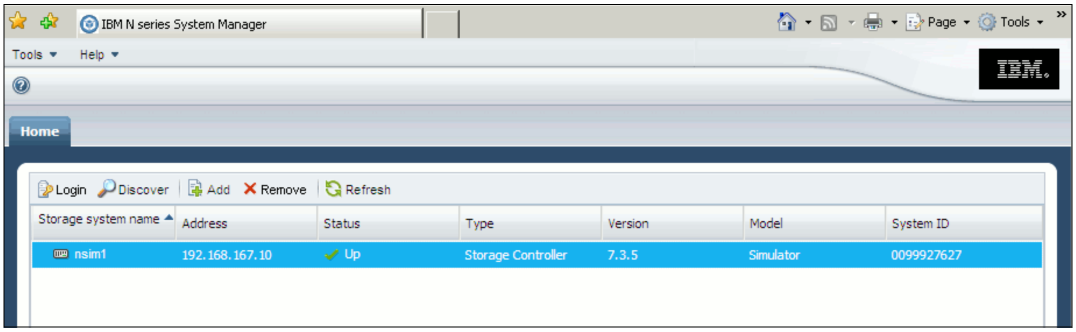
     Specify the SNMP community and SNMP version.

     If the storage system is running a Data ONTAP release in the Data ONTAP 7.2 release family, use SNMP version 1.
   - Credentials:

     Specify the user name and password.
5. 5. Click **Add**.

Alternatively, you can use the **Discover Storage Systems** dialog box to automatically discover storage systems or high-availability (HA) pair of storage systems on a network subnet and add them to the list of managed systems.

You will then see the N series controller in the NSM interface (Figure 38-2).



Storage system name	Address	Status	Type	Version	Model	System ID
nsim1	192.168.167.10	✓ Up	Storage Controller	7.3.5	Simulator	0099927627

Figure 38-2 Adding a controller to NSM

The first time that you double-click the new storage controller, you need to provide the correct username and password credentials.

38.3.3 Configuring a storage system

You can use the Storage Configuration wizard to configure your storage system or a high-availability configuration. You must separately configure each storage system when you configure an HA configuration.

Your storage systems must be running one of the following versions of Data ONTAP:

- Data ONTAP 7.x (starting from 7.2.3)
- Data ONTAP 8.x 7-Mode

Perform the following steps to configure a storage system:

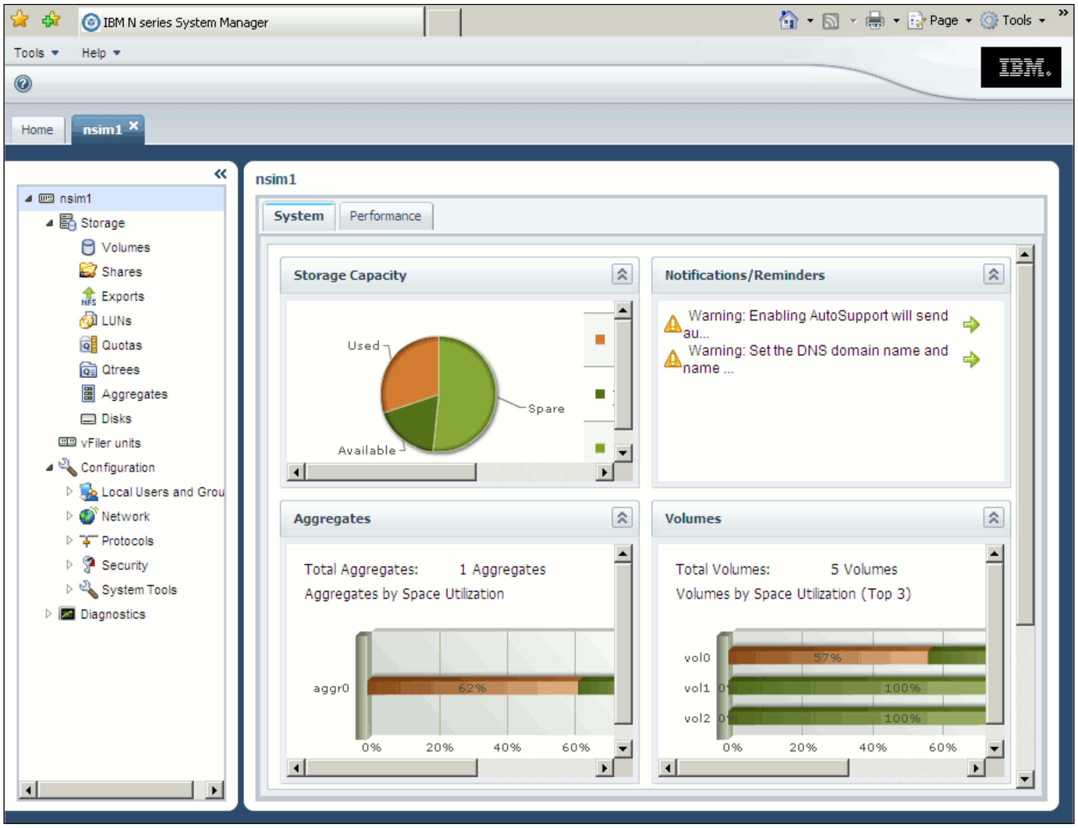
1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Storage**.
3. Click the desired **Frequent Tasks** wizard.

   These choices are available:

   – Create Aggregate
   – Create Volume
   – Create LUN
   – Create Qtree
   – Create Share
   – Create Export
   – Provision storage for VMware

4. Type or select information as requested by the wizard.
5. Confirm the details and click **Finish** to complete the wizard.

Figure 38-3 shows the main dashboard window, with the navigation panes on the left side.



Warning: Enabling AutoSupport will send au...

Warning: Set the DNS domain name and name ...

Total Aggregates: 1 Aggregates

Aggregates by Space Utilization

Total Volumes: 5 Volumes

Volumes by Space Utilization (Top 3)

Figure 38-3 Dashboard view in NSM

39

AutoSupport

This chapter describes the AutoSupport feature, and provides several examples of its configuration and usage.

AutoSupport is an integrated and efficient monitoring and reporting feature of the Data ONTAP operating system that runs on the IBM System Storage N series storage systems. It continuously monitors the health of your N series storage system, and it will automatically notify IBM Service and Support when certain problems are detected.

The following topics are covered:

- Overview of AutoSupport
- What is new in 8.2
- How AutoSupport works
- High level perspective
- Detailed perspective

39.1 Overview of AutoSupport

AutoSupport is one of the most important troubleshooting tools for N series customers. AutoSupport allows the system to send messages directly to IBM Service and Support. It provides these key features:

- Enables sophisticated monitoring for faster incident management
- Provides automated “call home” about critical events, even opening a support case automatically, such as hardware replacement requests
- Delivers non-intrusive alerting to notify you of a problem and provide information for IBM to take corrective action
- Enables AutoSupport analysis tools to monitor messages for known configuration issues
- Performs on-going health check analysis of 600 system parameters
- Sends system alerts to IBM Support and specified customer contacts

39.2 What is new in 8.2

Data ONTAP 8.2 supports several enhancements to the delivery of AutoSupport messages.

- AutoSupport On Demand is a new feature:

  This a new AutoSupport feature that periodically sends HTTPS requests to technical support to obtain delivery instructions. The delivery instructions can include requests to generate new AutoSupport messages, retransmit previously generated messages, and disable delivery of messages for specific trigger events.

  AutoSupport On Demand is enabled by default.

- The fully qualified domain name is now sent to SMTP mail servers:

  When sending connection requests to SMTP mail servers, AutoSupport now specifies the fully qualified domain name of the node, if you configured DNS. In previous releases, AutoSupport specified only the host name.

39.3 How AutoSupport works

AutoSupport is a call home feature in the Data ONTAP operating software for all IBM N series systems, providing an integrated efficient monitoring and reporting capability that continuously checks the health of your system. AutoSupport provides you with detailed knowledge of your N series environment. Information is sent to IBM Support and other designated addresses for quick incident resolution and preventative support.

AutoSupport also sends weekly diagnostic data back to IBM where it is automatically analyzed for any issues that might impact future system stability and performance.

39.4 High level perspective

Figure 39-1 provides a high level perspective of how AutoSupport functions:

1. An AutoSupport message is generated by the client filer.
2. The AutoSupport message is sent to IBM via the Internet (HTTPS and SMTP).
3. An automated solution finder analyzes the AutoSupport message to locate existing solutions.
4. IBM opens a problem management record (PMR).
5. The IBM Support Center will resolve the issue with the customer (by voice and/or email).

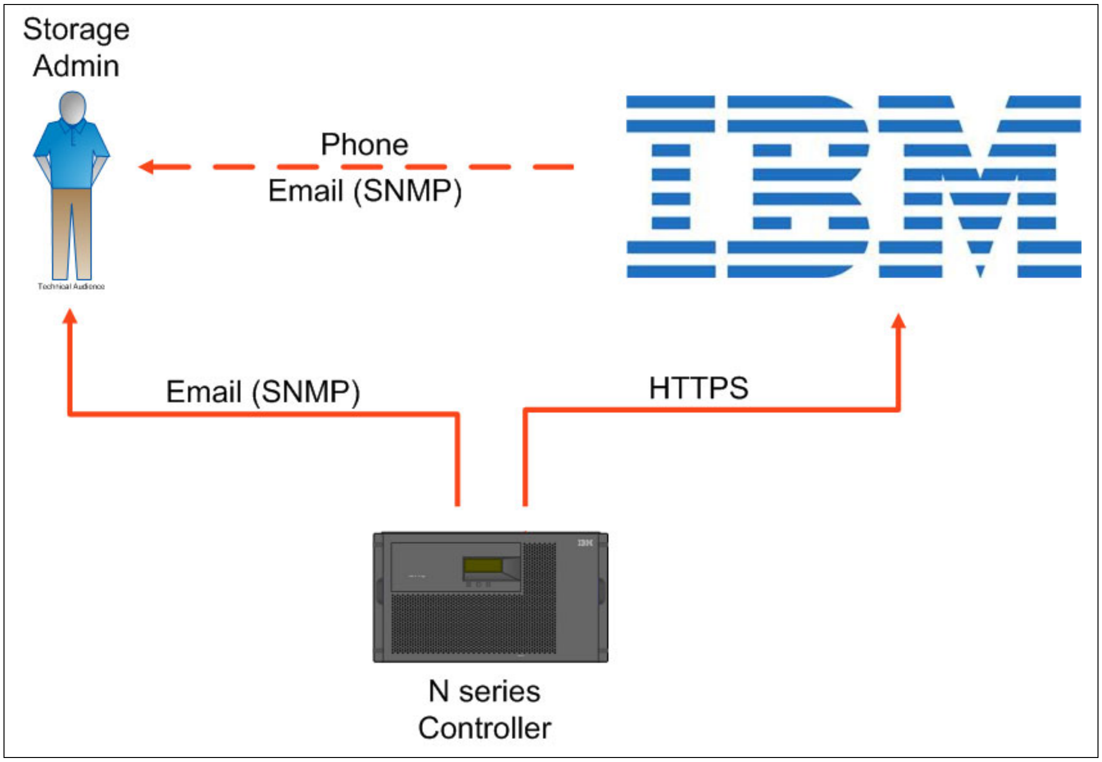


Figure 39-1 High-level perspective of AutoSupport

39.5 Detailed perspective

The AutoSupport daemon is enabled by default on the N series storage systems with Data ONTAP versions 7.1H2 and later. The AutoSupport options control how the N series storage system sends automatic status messages. Autosupport options can be set from FilerView or from the command line using the `Option` commands.

Architecture

AutoSupport has a new architecture in DOT 8 as shown in Figure 39-2. AutoSupport is now an M-Host (user space) process called notified. It collects information from the D-Blade, from Management Gateway (mgwd), from BSD commands, and from files.

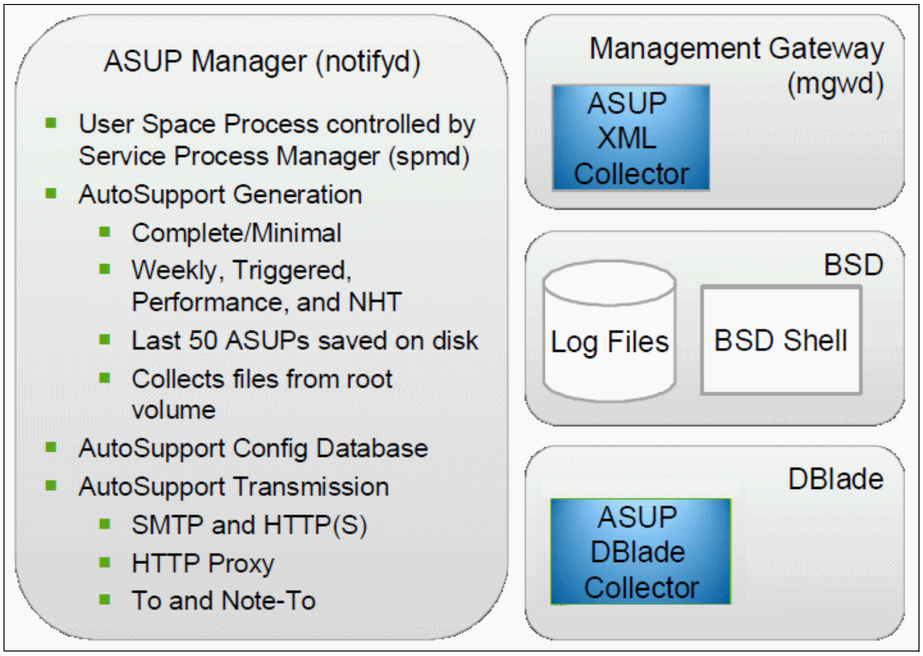

ASUP Manager (notifyd)
User Space Process controlled by Service Process Manager (spmd)
AutoSupport Generation
Complete/Minimal
Weekly, Triggered, Performance, and NHT
Last 50 ASUPs saved on disk
Collects files from root volume
AutoSupport Config Database
AutoSupport Transmission
SMTP and HTTP(S)
HTTP Proxy
To and Note-To
Management Gateway (mgwd)
ASUP XML Collector
BSD
Log Files
BSD Shell
DBlade
ASUP DBlade Collector

Figure 39-2 AutoSupport Architecture

The following sections provide more information about AutoSupport functionality and features.

AutoSupport daemon

N series storage controllers use an AutoSupport daemon to control how messages are sent from the customer’s system to IBM Support. The AutoSupport daemon is enabled by default on N series storage systems.

AutoSupport mechanism

The AutoSupport mechanism functions in the following ways:

- Data ONTAP triggers the AutoSupport mechanism automatically once a week to send information to IBM as well as the email addresses specified in autosupport.to. In addition, the options command can be used to invoke the AutoSupport mechanism to send this information on-demand.
- An AutoSupport message is sent in response to events that require corrective action from the system administrator or IBM Support.
- An AutoSupport message is sent when the system reboots.

Flexible transport options

AutoSupport can be configured with SMTP, HTTP, or HTTPS as the transport option, depending on network and security preferences. The transport option selected applies to the **`autosupport.support.to`** field. All other optional configured AutoSupport destinations are sent through SMTP and depend on a mailhost.

To improve email security, IBM supports Transport Layer Security (TLS) through VeriSign 256-bit digital certificates for encryption and authentication between mail server gateways through the following process:

- Ask your email administrator to enable TLS and install digital certificates on your mail servers. This will provide authentication against “man-in-the-middle” attacks.
- TLS encrypts the AutoSupport email content between your email server and the IBM email servers.

> **Tip:** HTTPS is the default transport protocol for AutoSupport, and it is the only transport protocol that IBM supports for Data ONTAP 8.x 7-Mode. To verify which specific transport protocols are supported on your Data ONTAP release, see the Data ONTAP product documentation.

Interaction with mail hosts

Storage systems do not function as mail hosts, instead they rely on another mailhost at the customer site that listens on the SMTP port 25 to send mail. To receive AutoSupport messages, the storage system requires access to an SMTP server or a mail forwarder, such as the sendmail program or Microsoft Exchange server. The administration host defined during setup is used as the default mail host unless otherwise specified. Customers can specify additional mail hosts, if desired.

AutoSupport messages

AutoSupport email options support up to five email addresses, which can include distribution email aliases for each AutoSupport option. These options are defined during the initial configuration and setup of the system and there are no requirements for AutoSupport email notifications to be customer or partner specific. Private email addresses can be defined. Refer to the Data ONTAP System Administration Guide for the version of Data ONTAP you have installed for detailed information about AutoSupport message capability, configuration and troubleshooting. See the AutoSupport tool section for more information.

Email messages

Up to five email addresses can be set for each AutoSupport mail option:

- The **`autosupport.noteto`** option provides for a “short note” email message containing the reason for the notification in the subject line and the time of failure. These messages are triggered only by specific urgent events and are easily viewed on a cell phone or other text device. This option is useful for system administrators who read email messages on alphanumeric pagers.
- The **`autosupport.partner.to`** option defines the list of email addresses that will receive all AutoSupport email notifications regardless of the severity level. This option is typically used by IBM support partners.
- The **`autosupport.to`** option defines the list of email addresses that will receive only critical AutoSupport email notifications; however, all AutoSupport notifications, regardless of their level of severity, continue to be sent to technical support as displayed by the read-only options: autosupport.support.to or autosupport.support.url.

Subject line of AutoSupport messages

The subject line of messages sent by the AutoSupport mechanism contains a text string that identifies the reason for the notification. The format of the subject line is as follows:

```
System Notification from <System_Name>(message)<Severity>
```

The subject line of messages sent by storage systems configured for high availability (HA) start with “Cluster Notification” (Data ONTAP 7.x) or “HA Group Notification” (Data ONTAP 8.x).

IBM Support response

IBM Support will address cases triggered by AutoSupport. Note that not all AutoSupport messages will result in a case. Some messages are for your information only.

Supported systems

AutoSupport is supported on any IBM N series systems running Data ONTAP 7.1 and later. IBM currently supports the 7.1, 7.2, 7.3, and 8.0, 8.1 (7-Mode) release families.

See the *Data ONTAP System Administration Guide* for the version of Data ONTAP you have installed to get more information about the AutoSupport proactive health check capability. The Data ONTAP publication matrices can be found on the IBM N series support site:

http://www.ibm.com/storage/support/nas

AutoSupport message content

AutoSupport collects configuration, status, and performance information about the storage system for IBM Support without the need for your involvement. Always see the *Data ONTAP System Administration Guide* for the most current AutoSupport information.

Types of information

Each AutoSupport message contains the following types of information. Items in the list marked with an asterisk (*) are suppressed when autosupport.content is set to minimal format. Items marked with two asterisks (**) are partially displayed in the autosupport.content minimal format:

- Date and timestamp of the message
- Data ONTAP software version
- IBM N series machine type - model
- Serial number of the storage system
- Encrypted software licenses*
- Host name of the storage system*
- SNMP contact name and location (if specified)*
- Console encoding type
- Output of commands that provide system information
- Checksum status
- Error-Correcting Code (ECC) memory scrubber statistics
- The following information, if High Availability (HA) configuration is licensed**
  - System ID of the partner in an HA pair
  - Host name of the partner in an HA pair
  - HA node status, including the HA monitor and HA interconnect statistics

- Contents of selected /etc directory files
- Expiry date of all SnapLock volumes on the system*
- Registry information
- Usage information*
- Service statistics
- Boot time statistics*
- NVLOG statistics*
- WAFL check log
- Modified configurations
- X-header information
- Information about the boot device (such as the CompactFlash card)

In addition, the contents of the /etc/messages and /etc/log/ems files are sent with each AutoSupport message as .gz attachments. You can specify the value of the autosupport.content option as complete or minimal to control the detail level of event messages and weekly reports.

Minimal autosupport

The **`autosupport.content`** option can be set to **`minimal`** (default is **`complete`**) to remove sensitive data from the autosupport messages.

The following information is removed when the **`autosupport.content`** option is set to **`minimal`**:

- Encrypted software licenses
- Host name of the storage system
- SNMP contact name and location (if specified)
- The following information, if High Availability (HA) configuration is licensed: (partially displayed):
  - System ID of the partner in an HA pair
  - HA node status, including the HA monitor and HA interconnect statistics
- Expiry date of all SnapLock volumes on the system
- Usage information
- Boot time statistics
- NVLOG statistics

In addition, file attachments that are normally sent are not attached to a minimal AutoSupport message. Minimal AutoSupport messages also omit sections and values that might be considered sensitive information and significantly reduce the amount of information sent.

40

OnCommand

OnCommand provides visibility across your storage environment by continuously monitoring and analyzing its health.

This chapter gives you an overview of what is deployed and how it is being utilized, enabling you to improve your storage capacity utilization and increase the productivity and efficiency of your IT administrators.

The following topics are covered:

- Introduction to OnCommand
- Key functionality

40.1 Introduction to OnCommand

OnCommand Storage Management software groups many products into a single family. It unifies multiple capabilities into a single product, in order to help customers achieve the storage efficiency they require.

Tip: OnCommand is included with Data ONTAP essentials.

OnCommand is a family of products designed to make N series storage the best for physical, virtual and cloud environments (Figure 40-1).

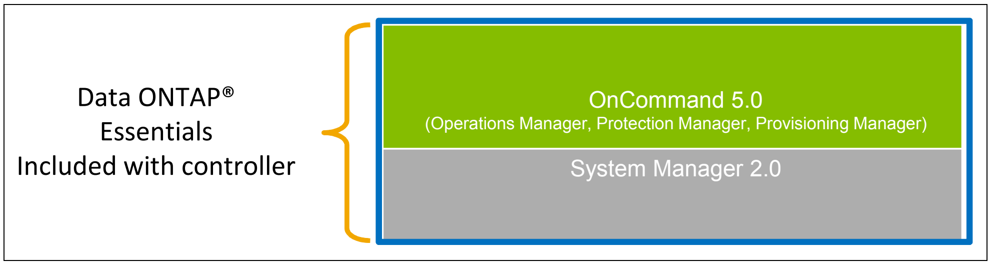


Figure 40-1 OnCommand portfolio

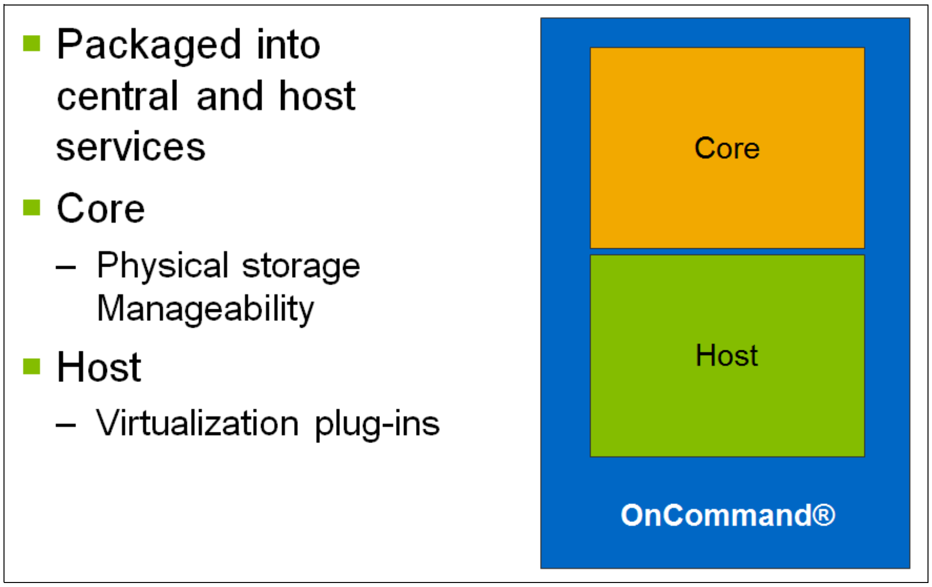
OnCommand provides the following capabilities:

- You can control your N series storage with System Manager and AutoSupport.
- System Manager provides simple, workflow-based wizards to automate common device management tasks. Admins can quickly set up and efficiently manage N series SAN and NAS systems.
- You can automate your N series storage infrastructure with OnCommand unified manager and SnapManager Software.
- OnCommand unified manager integrates the functions of Provisioning Manager, Protection Manager, and Operations Manager into a single user interface. Through a single view, you can monitor your entire shared storage environment, as well as drill down to define storage service levels and policy based workflows. Also included here are SnapManager software that provides the ability to connect to, and manage from, virtualization and other platforms.

Attention: Starting with Data ONTAP 8.1, the FilerView N series system management tool has been discontinued and is longer part of Data ONTAP. OnCommand replaces the FilerView tool.

40.1.1 OnCommand architecture

Figure 40-2 introduces the OnCommand main components, Core and Host.



- Packaged into central and host services
- Core
  - Physical storage Manageability
- Host
  - Virtualization plug-ins

Figure 40-2 OnCommand components

The architecture diagram in Figure 40-3 shows the basic components of the OnCommand Core and Host packages. The color-coding shows the Core (orange) versus Host (green) components.

Solid boxes show front-end GUIs with direct user interaction, versus dashed boxes, which are back-end server or services that are not directly visible to the user.

The OnCommand GUI console provides GUI management for Hyper-V objects and an alternative GUI for VMware objects. It launches the Operations Manager console and N series Management Console to manage the physical environment.

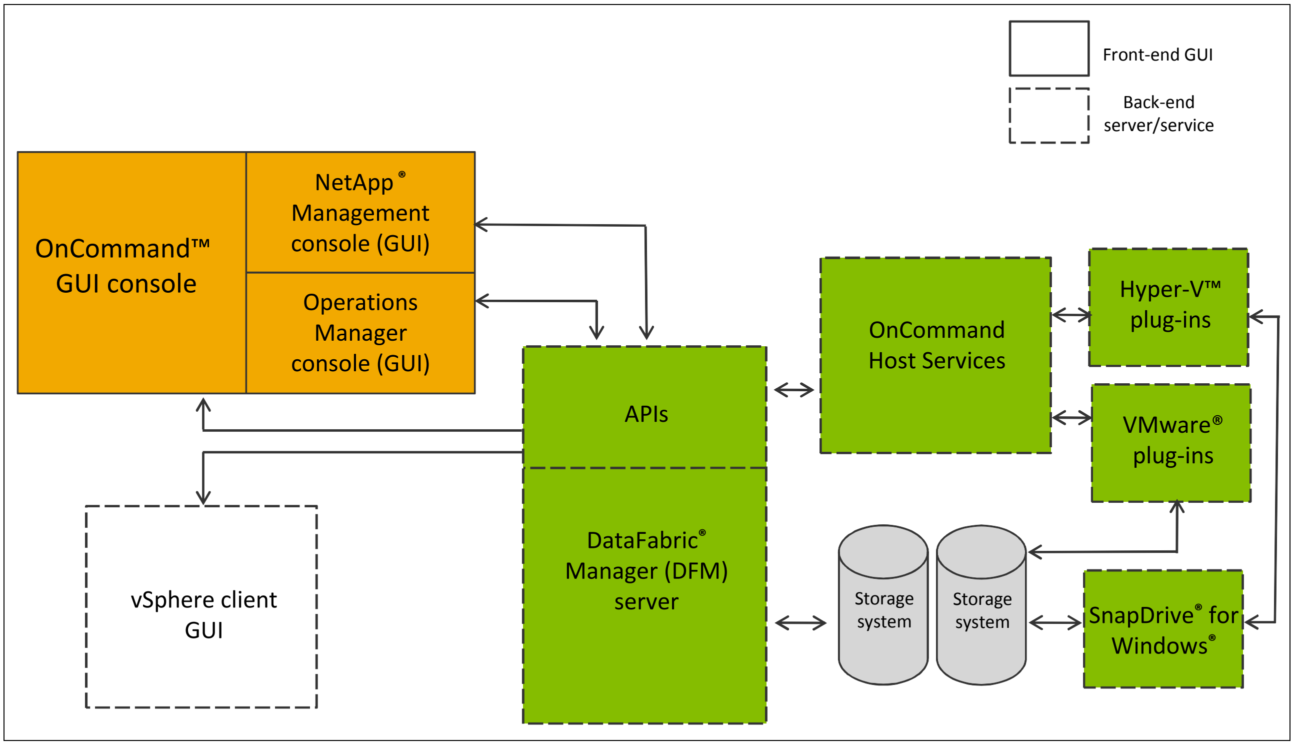


Figure 40-3 OnCommand architecture core and host components

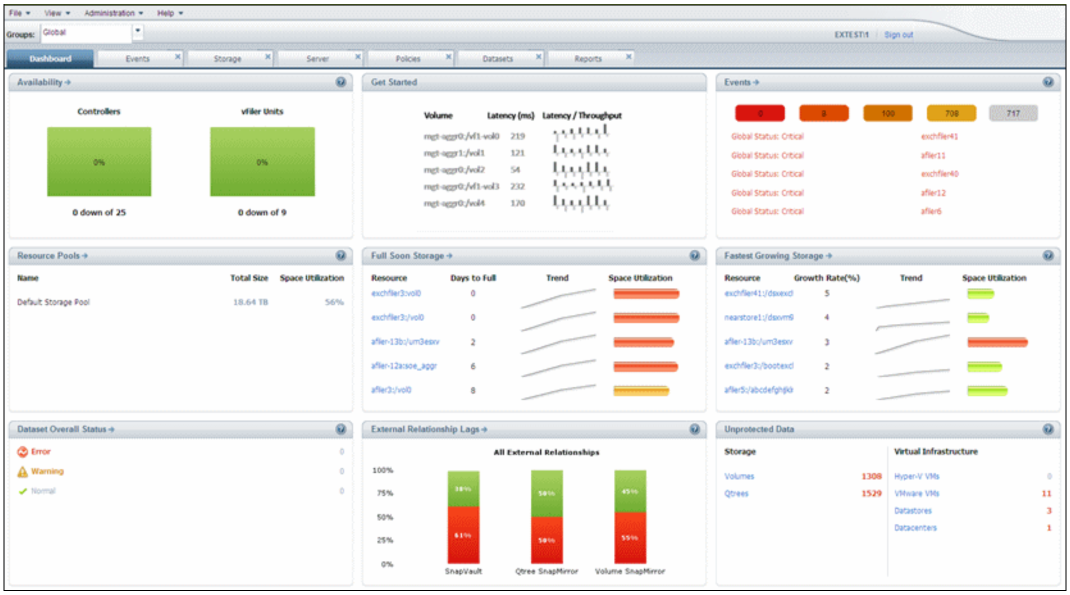
DataFabric Manager (DFM) can be installed either in the Standard or Express editions. OnCommand Host services caches schedules, catalogs, and events for short periods; enables execution without DataFabric Manager Server.

The Plug-ins for Hyper-V and VMware are collections of primitives that enable connection into these environments. SnapDrive for Windows is software used only within the Hyper-V environment for storage discovery and to manage LUNs or Snapshot copies.

The vSphere Client GUI is native VMware software that is used by the Vmware admin for virtual environment administration; it is provided access into the Storage environment through OnCommand.

40.1.2 Dashboard

OnCommand now provides a single, unified dashboard to view all storage resources for at-a-glance status and metrics, and also enables other interface choices (see Figure 40-4).



Volume	Latency (ms)
mgt-aggr0:/vf1-vol0	219
mgt-aggr1:/vol1	121
mgt-aggr0:/vol2	54
mgt-aggr0:/vf1-vol3	232
mgt-aggr0:/vol4	170

Global Status: Critical	exchfiler41
Global Status: Critical	afiler11
Global Status: Critical	exchfiler40
Global Status: Critical	afiler12
Global Status: Critical	afiler6

Name	Total Size	Space Utilization
Default Storage Pool	18.64 TB	56%

Resource	Days to Full
exchfiler3:vol0	0
exchfiler3:/vol0	0
afiler-13b:/um3esxv	2
afiler-12a:soe_aggr	6
afiler3:/vol0	8

Resource	Growth Rate(%)
exchfiler41:/dsxexcl	5
nearstore1:/dsxvm6	4
afiler-13b:/um3esxv	3
exchfiler3:/bootexcl	2
afiler5:/abcdefghjkl	2

Error	0
Warning	0
Normal	0

Storage	
Volumes	1308
Qtrees	1529

Virtual Infrastructure	
Hyper-V VMs	0
VMware VMs	11
Datastores	3
Datacenters	1

Figure 40-4 Dashboard

OnCommand provides visibility across your environment by continuously monitoring and analyzing its health. You get a view of what is deployed and how it is being utilized, enabling you to improve your storage capacity utilization and increase the productivity and efficiency of your IT administrators.

The dashboard contains information panels providing information about the system. N series OnCommand has various dashboard panels to provide cumulative information about various aspects of your environment:

- Availability panel: Information about storage controllers and vFiler units discovered and monitored by OnCommand. You can also view the number of controllers and vFiler units in a down state.
- Events panel: Status of the storage and server objects by listing the top five events based on their severity.
- Full Soon Storage panel: Aggregates and volumes reaching their capacity. Based on the number of days in which this threshold will be breached.
- Fastest Growing Storage panel: Aggregates and volumes for which space usage is increasing rapidly. Also displays growth rate, trend, and for a specific aggregate or volume.
- Dataset Overall Status panel: Overall status of the environment.
- Resource Pools panel: Resource pools facing potential space shortages based on the current usage levels.
- External Relationship Lags panel: Relative percentages of external SnapVault, Qtree SnapMirror, and volume SnapMirror relationships with lag times in error, warning and normal status.
- Unprotected Data panel: Number of unprotected storage and server objects being monitored.

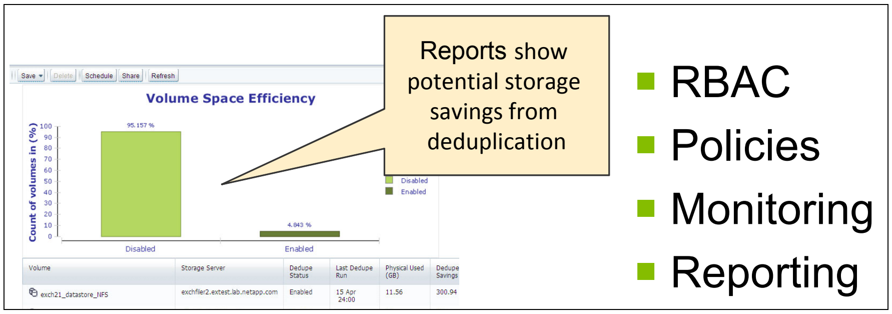
In addition, views are available through virtualization platforms based on SnapManagers, self-service customer portals via the Service Catalog capability, or through integrated partner frameworks.

40.2 Key functionality

This section provides an overview of the key functionalities of OnCommand.

40.2.1 Operations

OnCommand simplifies and standardizes storage operations. Standardized configuration accelerates deployment and mitigates operational risks. OnCommand delivers storage management features that enable business policy compliance. It is achieved by using enterprise-wide configuration management, distributed policy setting, and customized reporting (see Figure 40-5).



Volume	Storage Server	Dedupe Status	Last Dedupe Run	Physical Used (GB)	Dedupe Savings
exch21_datastore_NFS	exchfiler2.extest.lab.netapp.com	Enabled	15 Apr 24:00	11.56	300.94

Reports show potential storage savings from deduplication

- RBAC
- Policies
- Monitoring
- Reporting

Figure 40-5 OnCommand operations

OnCommand is intuitive and helps improve the productivity of storage administrators. The operations capability of the product helps storage administrators resolve problems faster and improve capacity utilization by providing a full picture of N series storage resources. With just a few clicks, administrators can drill down to detailed storage system information. And by replacing repetitive, time-intensive tasks with policy-based automation, they become more productive.

Role-based access control on the centralized console makes it possible for server and database administrators to perform self-service provisioning. Because these tasks are only performed within the limits of policies defined by IT architects and based on company business requirements, the system remains stable, efficiently configured, and under control. Policies that can be ascribed to datasets include capacity, storage reliability, space provisioning requirements, access mechanisms and security settings.

Another valuable dimension of operations management is monitoring and reporting.

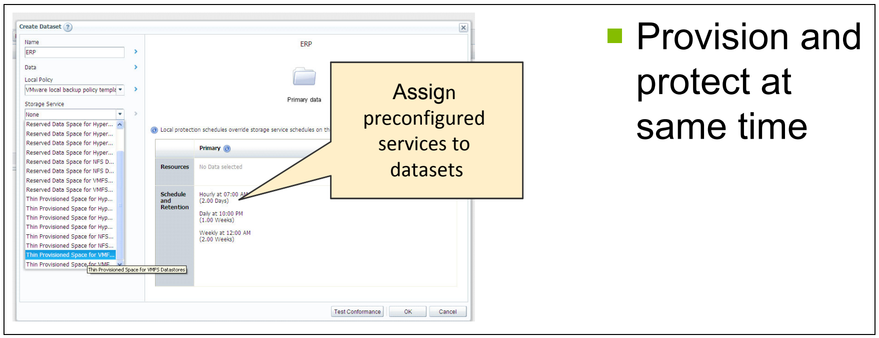
With OnCommand, you can continuously monitor and analyze the health of your storage environment, and can thus maintain visibility of what is being deployed and how it is being utilized. It improves both storage capacity utilization as well as administrator efficiency.

You can increase operating efficiency and eliminate hands-on complexity by streamlining provisioning with OnCommand. Complexity of the underlying storage can be removed for easier down-stream administration. OnCommand allows you to provision and protect data at the same time. The moment that you provision storage, you protect it. No additional steps or time are required.

40.2.2 Provisioning

OnCommand increases operating efficiency and eliminate hands-on complexity by streamlining provisioning. It allows the ability to provision and protect data at the same time; no additional steps or time are required.

Provisioning with OnCommand allows the automation of complex provisioning processes. Services can be defined granularly by the storage architect, and then be easily and consistently selected by down-stream administrators (see Figure 40-6).



Name: ERP

Data

Local Policy: VMware local backup policy templa

Storage Service: None

- Reserved Data Space for Hyper...
- Reserved Data Space for Hyper...
- Reserved Data Space for Hyper...
- Reserved Data Space for Hyper...
- Reserved Data Space for NFS D...
- Reserved Data Space for NFS D...
- Reserved Data Space for VMFS...
- Reserved Data Space for VMFS...
- Thin Provisioned Space for Hyp...
- Thin Provisioned Space for Hyp...
- Thin Provisioned Space for Hyp...
- Thin Provisioned Space for Hyp...
- Thin Provisioned Space for NFS...
- Thin Provisioned Space for NFS...
- Thin Provisioned Space for VMF...
- Thin Provisioned Space for VMF...

Thin Provisioned Space for VMFS Datastores

ERP

Local protection schedules override storage service schedules on th

	Primary
Resources	No Data selected
Schedule and Retention	Hourly at 07:00 AM (2.00 Days) Daily at 10:00 PM (1.00 Weeks) Weekly at 12:00 AM (2.00 Weeks)

Assign preconfigured services to datasets

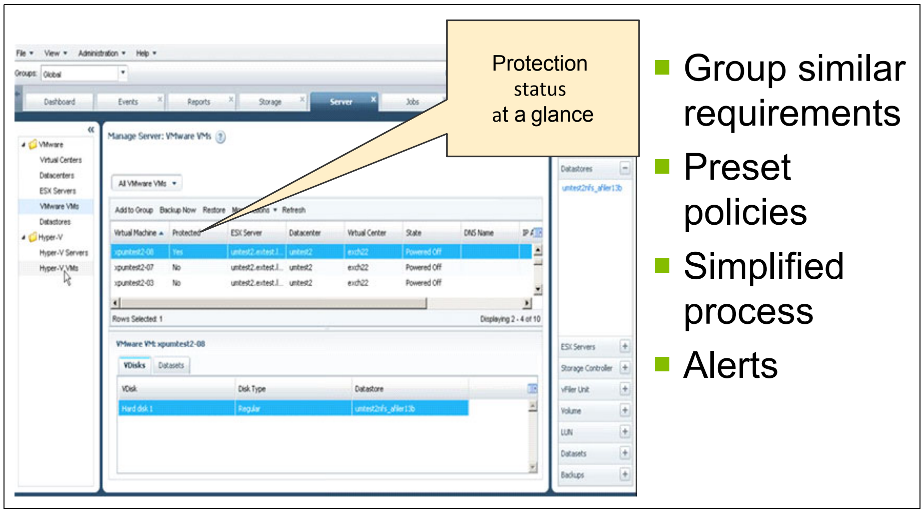
- Provision and protect at same time

Figure 40-6 Provisioning

To maximize use of your resources, OnCommand automates N series storage efficiency features including thin provisioning and primary data deduplication. This eliminates unnecessary and wasteful over-provisioning and provides storage only when needed.

40.2.3 Protection

OnCommand simplifies the process of protecting enterprise data by allowing administrators to group data with similar protection requirements into datasets and then apply preset policies (see Figure 40-7). They no longer need knowledge or expertise or underlying storage infrastructure.



Groups: Global

Manage Server: VMware VMs

Virtual Machine	Protected	ESX Server	Datacenter	Virtual Center	State	DNS Name
xpumtest2-08	Yes	umtest2.extest.l...	umtest2	exch22	Powered Off	
xpumtest2-07	No	umtest2.extest.l...	umtest2	exch22	Powered Off	
xpumtest2-03	No	umtest2.extest.l...	umtest2	exch22	Powered Off	

Rows Selected: 1

Displaying 2 - 4 of 10

VMware VM: xpumtest2-08

VDisk	Disk Type	Datastore
Hard disk 1	Regular	umtest2nfs_afiler13b

Datastores: umtest2nfs_afiler13b

Protection status at a glance

- Group similar requirements
- Preset policies
- Simplified process
- Alerts

Figure 40-7 OnCommand protection

OnCommand protects data by providing administrators with an easy-to-use management console to quickly configure and control all SnapMirror, SnapVault, Open Systems SnapVault (OSSV), and SnapManager family operations. It allows administrators to apply data protection policies consistently, automate complex protection processes, and pool backup and replication resources.

OnCommand simplifies the process of protecting enterprise data by allowing administrators to group data with similar protection requirements into datasets and then apply preset policies. It automatically correlates datasets and underlying physical storage resources, so administrators do not need to think in terms of the storage infrastructure.

A simple dashboard depicts comprehensive data protection information at a glance, including unprotected data. The software allows administrators to apply predefined policies to the data, minimizing the potential for error inherent in manual management. OnCommand also provides e-mail alerting to allow rapid analysis and correction of issues before they have a significant impact on data protection.

40.2.4 Plug-ins

OnCommand plug-ins for VMware and Microsoft provide access to OnCommand control and automation features from those respective management frameworks.

40.2.5 N series Operations Manager

Operations Manager is a Web-based UI of the DataFabric Manager server.

> **Attention:** Operations Manager is part of the Data ONTAP essentials base package. Operations Manager is one part of the back-end for the OnCommand management software.

You can use Operations Manager for the following day-to-day activities on storage systems:

- Discover storage systems.
- Monitor the device or the object health, the capacity utilization, and the performance.
- Determine characteristics of a storage system.
- View or export reports.
- Configure alerts and thresholds for event managements.
- Group devices, vFiler units, host agents, volumes, qtrees, and LUNs.
- Run Data ONTAP CLI commands simultaneously on multiple systems.
- Configure role-based access control (RBAC).
- Manage host users, user groups, domain users, local users, and host roles.

The OnCommand GUI supports the following applications and capabilities:

- Monitoring and reporting: The DataFabric Manager server discovers the storage systems supported on your network. The DataFabric Manager server periodically monitors data that it collects from the discovered storage systems, such as CPU usage, interface statistics, free disk space, qtree usage, and chassis environmental. The DataFabric Manager server generates events when it discovers a storage system, when the status is abnormal, or when a predefined threshold is breached. If configured to do so, the DataFabric Manager server sends a notification to a recipient when an event triggers an alarm.
- Alarm configuration: The DataFabric Manager server uses alarms to tell you when events occur. The DataFabric Manager server can send alarm notification to one or more specified recipients: an e-mail address, a pager number, an SNMP traphost, or a script that you write.

  You can customize which events cause alarms, whether the alarm repeats until it is acknowledged, and how many recipients an alarm has. Not all events are severe enough to require alarms, and not all alarms are important enough to require acknowledgment. Nevertheless, you must configure the DataFabric Manager server to repeat notification until an event is acknowledged, to avoid multiple responses to the same event.

  The DataFabric Manager server does not automatically send alarms for the events. You must configure alarms for the events that you specify.
- Backup Manager: You can manage disk-based backups for your storage systems using Backup Manager. You can access it from the Backup tab in Operations Manager. Backup Manager provides tools for selecting data for backup, scheduling backup jobs, backing up data, and restoring data.

  The DataFabric Manager server uses the SnapVault technology of Data ONTAP to manage the backup and restore operations. This tab is displayed only when the Business Continuance Option license is installed.
- Disaster Recovery Manager: Disaster Recovery Manager is an application within the DataFabric Manager server that enables you to manage and monitor multiple SnapMirror relationships from a single interface.

  Disaster Recovery Manager provides a simple, Web-based method of monitoring and managing SnapMirror relationships between volumes and qtrees on your supported storage systems and vFiler units. This tab is displayed only when the Business Continuance Option license is installed.
- Quota management: Quotas can cause Data ONTAP to send a notification (soft quota) or to prevent a write operation from succeeding (hard quota) when quotas are exceeded. You can use the OnCommand GUI to view user quota summary reports, chargeback reports, user details, quota events, and so on.

For more details about Operations Manager, see the Redbooks publication, *Managing Unified Storage with IBM System Storage N series Operation Manager*, SG24-7734, which is available at this website:

http://www.redbooks.ibm.com/abstracts/sg247734.html?Open

40.2.6 N series Management Console

The ISM N series Management Console (NMC) is a client platform that supports N series Manageability Software capabilities. Together, the NMC and the associated OnCommand console, enable you to implement both policy-based and on-demand protection, provisioning, migration, and restoration of data contained in your physical and virtual storage systems.

The NMC supports the following applications and capabilities:

- Performance Advisor: This application provides a single location from which you can view comprehensive information about storage system and MultiStore vFiler unit performance and perform short-trend analysis. The application also helps you identify the data infrastructure causes and potential causes of reduced performance.
- Protection Manager: The N series Management Console data protection capability provides a policy-based management tool to help you unify and automate backup and mirroring operations. This capability uses a holistic approach to data protection. It provides end-to-end, workflow-based design and seamless integration of SnapVault, SnapMirror, and Open Systems SnapVault to enable you to manage large-scale deployments easily.

  The disaster recovery feature of the N series Management Console data protection capability enhances your data protection services by enabling you to continue to provide data access to your users, even in the event of mishap or disaster that disables or destroys the storage systems in your primary data node. You can quickly enable your secondary storage systems to provide primary data storage access to your users with little or no interruption, until your primary storage systems are renewed or replaced.
- Provisioning Manager: The N series Management Console provisioning capability helps you simplify and automate the tasks of provisioning and managing storage. It provides policy-based provisioning and conformance of storage in datasets. This capability also enables you to manually add volumes or qtrees to a dataset at any time, provides manual controls for space and capacity management of existing storage and newly provisioned storage, and allows you to migrate datasets and vFiler units to a new storage destination.

The deduplication feature of the N series Management Console provisioning capability enhances your data provisioning services by enabling you to eliminate duplicate data blocks to reduce the amount of storage space used to store active data.

40.2.7 Host Agent

The Host Agent is software that resides on a Windows, Linux, or Solaris host. It collects information, such as OS name, version, HBA information, and file system metadata, and then sends that information back to the DataFabric Manager Server. Users can create reports of the collected information by using the Operations Manager UI or the DataFabric Manager Server CLI.

Related publications

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

IBM Redbooks publications

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 System Storage N Series Hardware Guide*, SG24-7840
- *IBM System Storage N series MetroCluster,* REDP-4259
- *IBM System Storage N series Clustered Data ONTAP*, SG24-8200
- *IBM System Storage N series Reference Architecture for Virtualized Environments*, REDP-4865
- *IBM System Storage N series Reference Architecture for Virtualized Environments*, SG24-8155
- *Managing Unified Storage with IBM System Storage N series Operation Manager*, SG24-7734
- *Using the IBM System Storage N series with IBM Tivoli Storage Manager*, SG24-7243
- *IBM System Storage N series and VMware vSphere Storage Best Practices,* SG24-7871
- *IBM System Storage N series with VMware vSphere 5*, SG24-8110
- *Designing an IBM Storage Area Network*, SG24-5758
- *Introduction to Storage Area Networks and System Networking*, SG24-5470
- *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240
- *Storage and Network Convergence Using FCoE and iSCSI,* SG24-7986
- *IBM Data Center Networking: Planning for Virtualization and Cloud Computing*, SG24-7928
- *IBM N Series Storage Systems in a Microsoft Windows Environment*, REDP-4083
- *Using an IBM System Storage N series with VMware to Facilitate Storage and Server Consolidation*, REDP-4211
- *IBM System Storage N series with FlexShare*, REDP-4291
- *IBM System Storage N series A-SIS Deduplication Deployment and Implementation Guide*, REDP-4320
- *IBM N Series Storage Systems in a Microsoft Windows Environment*, REDP-4083
- *IBM System Storage N series with VMware vSphere 4.1*, SG24-7636
- *IBM System Storage N series with VMware vSphere 4.1 using Virtual Storage Console 2*, REDP-4863
- *Introduction to IBM Real-time Compression Appliances,* SG24-7953
- *Designing an IBM Storage Area Network*, SG24-5758

- *Introduction to Storage Area Networks and System Networking*, SG24-5470
- *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240
- *Storage and Network Convergence Using FCoE and iSCSI,* SG24-7986
- *IBM Data Center Networking: Planning for Virtualization and Cloud Computing*, SG24-7928.

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

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- Network-attached storage:

  http://www.ibm.com/systems/storage/network/

- IBM Support: Documentation:

  http://www.ibm.com/support/entry/portal/Documentation

- IBM Storage – Network Attached Storage: Resources:

  http://www.ibm.com/systems/storage/network/resources.html

- IBM System Storage N series Machine Types and Models (MTM) Cross Reference:

  – http://www-304.ibm.com/support/docview.wss?uid=ssg1S7001844

- IBM N Series to NetApp Machine type comparison table:

  – http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105042

- Interoperability matrix:

  – http://www-304.ibm.com/support/docview.wss?uid=ssg1S7003897

Online resources

These websites are also relevant as further information sources:

- IBM NAS support website:

  http://www.ibm.com/storage/support/nas/

- NAS product information:

  http://www.ibm.com/storage/nas/

- IBM Integrated Technology Services:

  http://www.ibm.com/planetwide/

Help from IBM

IBM Support and downloads:

ibm.com/support

IBM Global Services:

ibm.com/services

IBM System Storage N series Software Guide

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

IBM System Storage N series Software Guide

Learn about Data ONTAP 8.2 7-mode features and functions

See storage efficiency features embedded in N series systems

Understand unified N series storage architecture

Corporate workgroups, distributed enterprises, and small to medium-sized companies are increasingly seeking to network and consolidate storage to improve availability, share information, reduce costs, and protect and secure information. These organizations require enterprise-class solutions capable of addressing immediate storage needs cost-effectively, while providing an upgrade path for future requirements. IBM System Storage N series storage systems and their software capabilities are designed to meet these requirements.

IBM System Storage N series storage systems offer an excellent solution for a broad range of deployment scenarios. IBM System Storage N series storage systems function as a multiprotocol storage device that is designed to allow you to simultaneously serve both file and block-level data across a single network. These activities are demanding procedures that, for some solutions, require multiple, separately managed systems. The flexibility of IBM System Storage N series storage systems, however, allows them to address the storage needs of a wide range of organizations, including distributed enterprises and data centers for midrange enterprises. IBM System Storage N series storage systems also support sites with computer and data-intensive enterprise applications, such as database, data warehousing, workgroup collaboration, and messaging.

This IBM Redbooks publication explains the software features of the IBM System Storage N series storage systems. This book also covers topics such as installation, setup, and administration of those software features from the IBM System Storage N series storage systems and clients and provides example scenarios.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7129-07 ISBN 0738439649