IBM System Storage N series Clustered Data ONTAP

Understand Clustered Data ONTAP benefits for dynamic storage solution

Learn about Clustered Data ONTAP features and functions

Design scaleable NAS solutions using N series

Roland Tretau
Christian Fey
Michal Klimes
Steven Pemberton
Tom Provost
Youn-Ho Yang

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM System Storage N series Clustered Data ONTAP

June 2014

SG24-8200-00

Note: Before using this information and the product it supports, read the information in "Notices" on page xix.

First Edition (June 2014)

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

Contents

Figures

Tables

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®
Global Technology Services®
GPFS™
IBM®
Redbooks®
Redpapers™
Redbooks (logo) ®
System Storage®
System x®
Tivoli®

The following terms are trademarks of other companies:

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

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 with Clustered Data ONTAP Version 8.2, which is the first version available on the IBM System Storage N series and, as of October 2013, is also the most current version available. Clustered Data ONTAP is different from previous ONTAP versions by the fact that it offers a storage solution that operates as a cluster with flexible scaling capabilities. Clustered Data ONTAP configurations allow clients to build a scale-out architecture, protecting their investment and allowing horizontal scaling of their environment.

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.

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.

Michal 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 five previous IBM Redbooks publications.

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 IBM Redbooks publications. and IBM Redpapers™ publications. He joined IBM in 2010.

Youn-Ho 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 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
International Technical Support Organization, San Jose Center

Uwe Heinrich Mueller, Uwe Schweikhard
IBM Germany

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`

Part 1

Architectural overview

Data ONTAP 8 for IBM System Storage N series offers two modes of operation: Clustered Data ONTAP and Data ONTAP operating in 7-Mode.

This publication is an overview of IBM System Storage N series Clustered Data ONTAP, including its architecture and core capabilities, positioning it firmly in today's agile data centers.

The amount of data collected by today's systems seems to be growing relentlessly, regardless of company size. This is because on today's smarter planet, instrumented, interconnected and intelligent businesses collect, process, use, and store more information than ever before.

With versatile N series systems, you can combine high-performance solid-state drive (SSD) flash storage or Serial Attached SCSI (SAS) hard drives and large-capacity nearline-SAS or Serial Advanced Technology Attachment (SATA) disk drives in storage tiers to optimize performance and cost. In addition, you can seamlessly consolidate block and file storage on the same system. N series makes this possible by providing native support of the Network File System (NFS) and Common Internet File System (CIFS), Fibre Channel over Ethernet (FCoE), Fibre Channel Protocol (FCP), and Internet Small Computer System Interface (iSCSI) storage protocols, through both Fibre Channel and Ethernet interfaces.

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

- Clustered Data ONTAP: What it is
- Clustered Data ONTAP 8.2 architecture
- Terminology in Clustered Data ONTAP 8.2
- Clustered Data ONTAP compared to 7-Mode
- High availability (HA) pairs and failover behavior
- Physical cluster types and scaling

1

Clustered Data ONTAP: What it is

In this part of the book, we introduce the IBM System Storage N series cluster hardware and software, which provide a range of reliable, scalable storage solutions for a variety of storage requirements.

The following topics are covered:

- Clustered Data ONTAP software
- Clustered Data ONTAP hardware
- Multi-tenancy and cluster components

1.1 Clustered Data ONTAP software

Clustered Data ONTAP is a highly scalable storage operating system that allows multiple IBM System Storage N series controllers to be combined into a single integrated system for horizontal scalability and non-disruptive operations across the hardware lifecycle. It also delivers improved performance, management simplicity, and reliability.

These are some of the key features of Clustered Data ONTAP:

- Support for up to 24 nodes in a single system
- Flexible cluster configurations that combine heterogeneous controllers
- Transparent data movement between nodes (DataMotion for Volumes)
- Single namespace or multiple namespaces
- Non-disruptive operations for storage maintenance, hardware lifecycle operations, and software upgrades, including full technical refresh
- Native multi-tenancy with quality of service (QoS) workload management
- Support for SAN and NAS protocols (FC, FCoE, iSCSI, NFS, CIFS/SMB)
- Full support for Virtual Storage Tier, including Flash Cache, Flash Pool, and Flash Accel software

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), as well as storage area network technologies such as iSCSI, Fibre Channel over Ethernet (FCoE), and Fibre Channel Protocol (FCP).

The N series storage solution supports file and block protocols as shown in Figure 1-1.

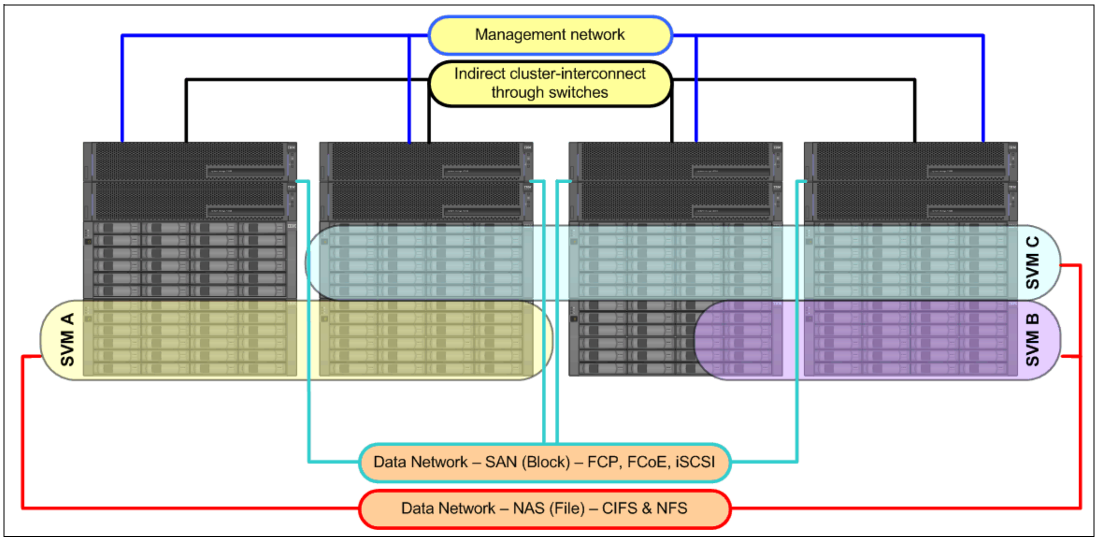

Management network
Indirect cluster-interconnect through switches
SVM A
SVM B
SVM C
Data Network – SAN (Block) – FCP, FCoE, iSCSI
Data Network – NAS (File) – CIFS & NFS

Figure 1-1 N series storage solution - protocols

1.1.1 Multi-protocol unified architecture

Multi-protocol unified architecture is the ability to support multiple data access protocols concurrently in the same storage system, over a whole range of different controller and disk storage types. Data ONTAP 7G and 7-Mode have long been capable of this, and now Clustered Data ONTAP supports an even wider range of data access protocols.

The following protocols are supported in Clustered Data ONTAP 8.2:

- NFS v3, v4 and v4.1, including pNFS:
  - pNFS is a feature of NFSv4.1.
  - pNFS separates a file system's metadata from its physical location in a network.
  - pNFS can be used to deliver increased performance and scalability.
  - Clients need to be running a version of Linux that supports pNFS.
  - pNFS clients are directed to the physical node that hosts a specific file. Data access over pNFS is therefore always direct. If a volume is moved from one physical node to another, the pNFS clients will send I/O requests to the new location.
- SMB 1, 2, 2.1, and 3, including support for non-disruptive failover in Microsoft Hyper-V environments with SMB 3
- iSCSI
- Fibre Channel
- FCoE
- Infinite Volumes can be accessed using NFSv3, NFSv4.1, pNFS, and SMB 1.0.

1.1.2 Clustered Data ONTAP 8.2

Clustered Data ONTAP 8.2 increases the scalability, protocol support, and data protection capabilities of previous releases. It also allows the co-existence of multiple storage virtual machines (SVMs) targeted for various use cases, including large NAS content repositories, general purpose file services, and enterprise applications.

This ability to scale also comes with data mobility features that allow workloads to be balanced across node and disk resources. As a result, updating and upgrading hardware can be done non-disruptively, allowing customers to retain data access even during updates and refreshes.

A cluster can contain up to 24 nodes (unless the iSCSI or FC protocols are enabled, in which case the cluster can contain up to eight nodes). Each node in the cluster can view and manage the same volumes as any other node in the cluster. The total file system namespace, which comprises all of the volumes and their resultant paths, spans through the cluster.

When new nodes are added to a cluster, there is no need to update clients to point to the new nodes. The existence of the new nodes is transparent to the clients.

1.1.3 N series OnCommand

OnCommand is the storage management software for IBM System Storage N series:

- Enables IT storage teams to deliver efficiency savings by integrating operations, provisioning, and protection of their physical and virtual storage resources.
- Enables storage as a service automation through the service catalog, as well as rich third-party virtualization and console integration.
- Delivers storage efficiency, makes the act of storage administration more efficient.
- For Virtual Infrastructure administrators, the use of Virtual Storage Console is advised, since it provides integrated, comprehensive storage management for VMware infrastructure, enabling VMware administrators to manage IBM N series storage trough their vSphere client.

Key points and features are described in Table 1-1.

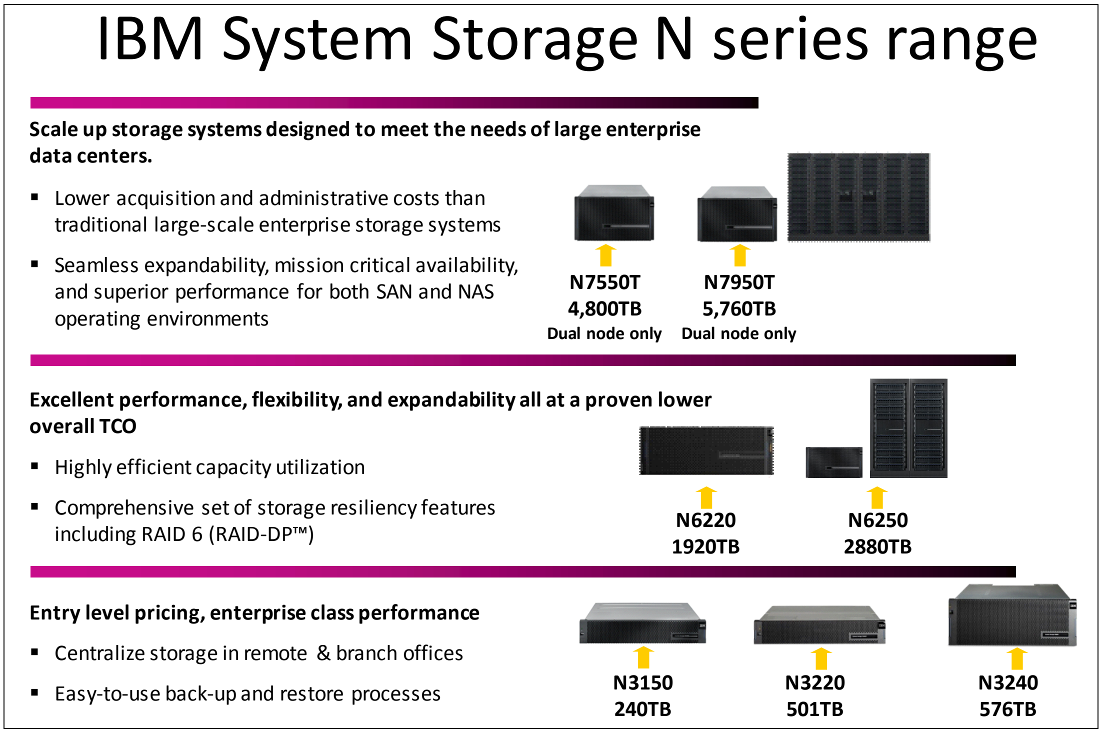
Table 1-1 Key points and features of N series OnCommand

Features	Benefits
Integrated dashboard	Single interface for monitoring and management
UI uses N series Web Framework	Modern look and feel, consistent with other N series interfaces
Integrated policy infrastructure	Provisioning, protection, RBAC other key capabilities are applied uniformly across enterprise-scale environment
Name-able volumes, snapshots	Naming flexibility to align with customer IT naming conventions
Data motion enhancements	Flexibility to migrate data across drive types, and support for flexclones
Dynamic reporting	Customizable, flexible report generation
Open API and free SDK	Extensible infrastructure allows integration with homegrown or 3rd-party management/cloud solutions
64 bit support	Improved system performance
Free licensing	Reduced total cost of data management

1.2 Clustered Data ONTAP hardware

Each system consists of one or more fabric-attached N series storage building blocks, and each is a high-availability pair of controllers (storage nodes). Multiple controller pairs form a single, integrated cluster. Clustered Data ONTAP uses Gigabit (Gb) and 10 Gb Ethernet technology for server connections and for interconnecting N series controllers.

Figure 1-2 shows N series portfolio comparison.



IBM System Storage N series range

Scale up 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 expandability, mission critical availability, and superior performance for both SAN and NAS operating environments

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

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

Entry level pricing, enterprise class performance

- Centralize storage in remote & branch offices
- Easy-to-use back-up and restore processes

Figure 1-2 N series portfolio comparison

Using 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 IBM System Storage N series is designed from the ground up as a stand-alone storage system.

The following expansion units are supported:

- EXN1000 contains up to 14 SATA drives:
  - 1 TB, 2 TB (7,200 rpm)
- EXN3000 contains up to 24 SAS or SATA drives:
  - NL SAS: 4 TB (7,200 rpm)
  - SAS: 300 GB, 450 GB, 600 GB, 600 GB SED (15,000 rpm), 200 GB SSD
  - SATA: 500 GB, 1 TB, 2 TB, 3 TB, 3 TB SED (7,200 rpm), 100GB SSD
- EXN3200 contains up to 48 SATA drives:
  - SATA: 3 TB, 4 TB (7,200 rpm)
- EXN3500 contains up to 24 small form factor drives:
  - SAS (SFF): 450 GB, 600 GB, 600 GB SED, 900 GB, 900 GB SED, 1.2 TB 2.5 inch (10K rpm), 200 GB SSD, 800 GB SSD
- EXN4000 contains up to 14 Fibre Channel drives:
  - 300 GB, 450 GB, 600 GB, 15,000 rpm

Mixed shelf support is provided for the Flash Pool Caching feature:

- 900 GB SAS and 200 GB SSD:
  - EXN3500
  - N3150
  - N3220
- 2 TB SATA and 200 GB SSD:
- EXN3000 w/IOM6:
  - N3240
  - N3150
- 4 TB NL-SAS HDD and 200GB SSD:
  - N3150
  - N3240

1.3 Multi-tenancy and cluster components

A cluster is composed of physical hardware, including storage controllers with attached disk shelves, NICs, and Flash Cache cards, which are optional. Together, these components create a physical resource pool that is virtualized as logical cluster resources to provide data access. Abstracting and virtualizing physical assets into logical resources provide flexibility and potential multi-tenancy in Data ONTAP, as well as the data motion ability that is at the heart of non-disruptive operations.

Physical cluster components

Although there can be different types of storage controllers, they are, by default, all considered equivalent in the cluster configuration; they are all presented and managed as cluster nodes. Individual disks are managed by defining them into aggregates, which are groups of disks of a particular type that are protected using RAID-DP, the same way as it is in Data ONTAP operating in 7-Mode.

NICs and host bus adapters (HBAs) provide physical ports, such as Ethernet and FC, for connection to management and data networks. The physical components are visible only to cluster administrators and not directly to the applications and hosts that use the cluster. The physical components constitute a pool of resources from which are constructed the logical cluster resources. Applications and hosts access data only through virtual servers that contain volumes and logical interfaces.

Logical cluster components

The primary logical cluster component is the storage virtual machine (SVM). Data ONTAP supports from one to hundreds of SVMs in a single cluster. Each SVM enables one or more storage area network (SAN) and network-attached storage (NAS) access protocols and contains at least one volume and at least one logical interface (LIF). Administration of each SVM can also be delegated, if desired, so that separate administrators can be responsible for provisioning volumes and other SVM operations. This is particularly appropriate for multi-tenanted environments or environments where workload separation is desired.

Data ONTAP operating in Cluster-Mode facilitates multi-tenancy at the storage layer by segregating storage entities such as aggregates, LIFs, LUNs, and volumes and containing them in an SVM. Because each SVM operates in its own namespace, each unit/customer mapped to an SVM is completely isolated. Each SVM supports role-based access control (RBAC). Specific protocols such as NFS, CIFS, iSCSI, FC, and FCoE can be assigned to it.

2

Clustered Data ONTAP 8.2 architecture

This chapter describes the architecture of Clustered Data ONTAP, with an emphasis on the separation of physical resources and virtualized containers. Virtualization of storage and network physical resources is the basis for scale-out and non-disruptive operations.

The following topics are covered:

- Hardware support and basic system overview
- Clustered Data ONTAP hardware architecture
- Storage virtual machine
- Understanding cluster quorum, epsilon, and data availability
- Multi-tenancy
- Multi-tenancy
- Clustered Data ONTAP 8.2 licensing

2.1 Hardware support and basic system overview

As seen in Figure 2-1, a clustered ONTAP system consists of IBM N series storage controllers with attached disks. The basic building block is the high availability (HA) pair, a concept familiar from Data ONTAP 7G and 7-Mode environments. An HA pair consists of two identical nodes, or instances of clustered ONTAP. Each node actively provides data services and has redundant cabled paths to the other node's disk storage. If either node is down for any reason, planned or unplanned, its HA partner can take over its storage and maintain access to the data. When the downed system rejoins the cluster, the partner node gives back the storage resources.

Note: A "node" refers to a single instance of Clustered Data ONTAP. Each HA pair contains two nodes.

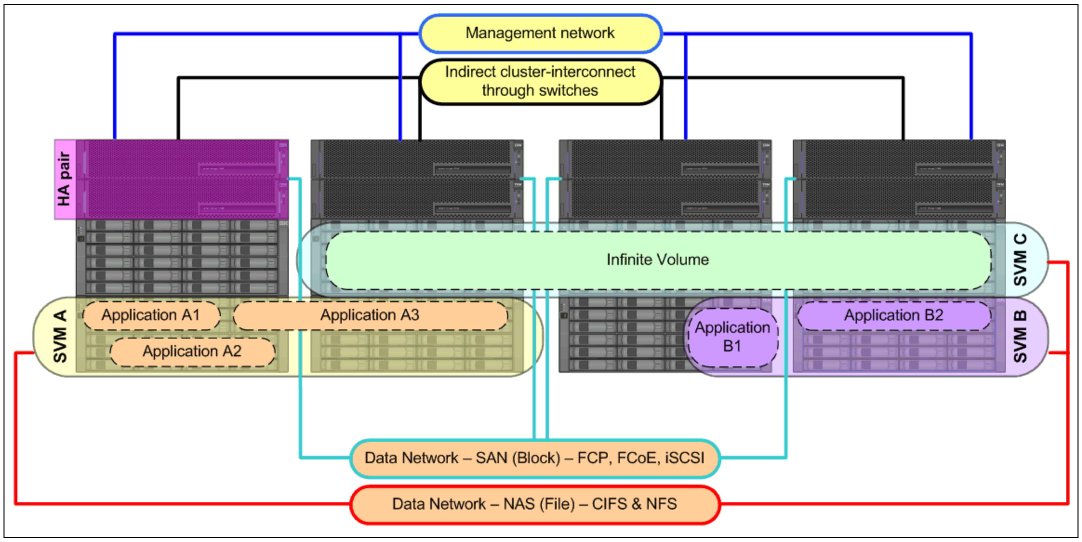

Management network
Indirect cluster-interconnect through switches
HA pair
Infinite Volume
SVM C
SVM A
Application A1
Application A3
Application A2
Application B1
Application B2
SVM B
Data Network – SAN (Block) – FCP, FCoE, iSCSI
Data Network – NAS (File) – CIFS & NFS

Figure 2-1 Four HA partners forming the cluster and running several SVMs

The minimum cluster size starts with 2 matching nodes in an HA pair. Using non-disruptive technology refresh, a two node, entry-level cluster can evolve to the largest cluster size and most powerful hardware. Clusters with SAN protocols are supported up to 8 nodes with mid and high-end controllers. NAS-only clusters of high-end controllers scale up to 24 nodes and over 69 PB of data storage.

Notes:

- Clustered Data ONTAP 8.2 offers the additional option of a single node cluster configuration. This is intended for smaller locations that replicate to a larger data center and enables you to use the cluster ports to serve data traffic.
- The term *cluster* has been used historically to refer to an HA pair running Data ONTAP 7G or 7-Mode. This usage has been discontinued, and HA pair is the only correct term for this configuration. The term *cluster* now refers only to a configuration of one or more HA pairs running Clustered Data ONTAP.

The nodes in a cluster communicate over a dedicated, physically isolated, and secure Ethernet network. The LIFs on each node in the cluster must be on the same subnet.

2.2 Clustered Data ONTAP hardware architecture

Every node in the cluster consists of four major software components called modules. These are accessed only by well defined APIs. Figure 2-2 shows the architecture of Clustered Data ONTAP.

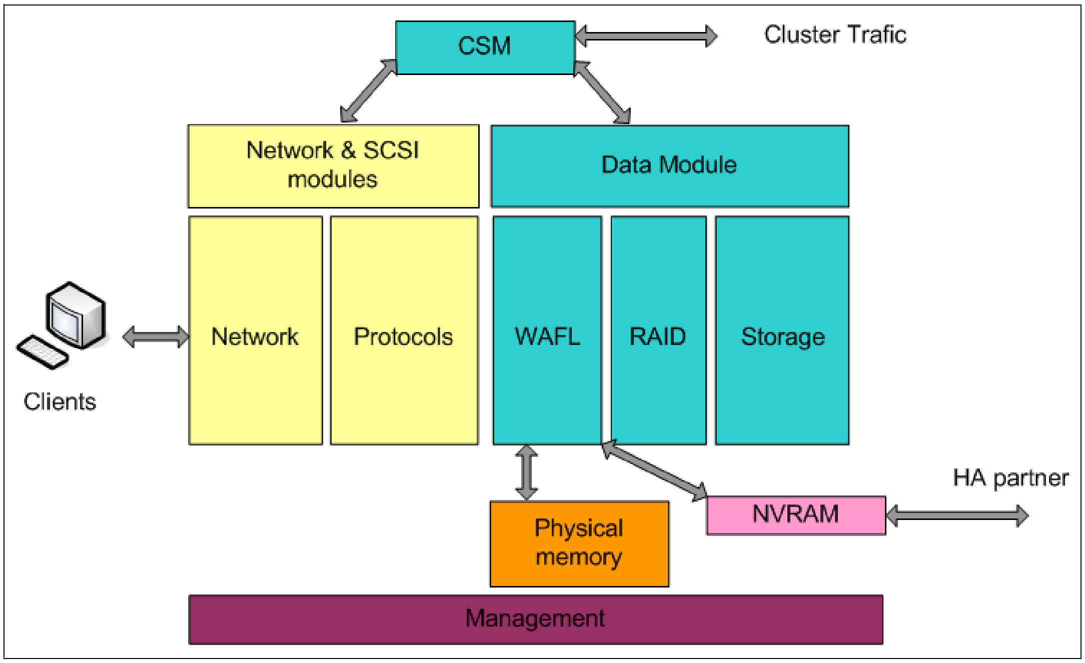


Figure 2-2 Clustered Data ONTAP architecture

Figure 2-3 shows the hardware architecture of the node and dataflow within the modules:

- Network module:
  - Processes the client's data communication through NFS and CIFS and translates it to Spin Network Protocol (SpinNP). Sends it to CSM.
  - Processes the data communication from CSM and translates it from SpinNP to desired protocol.
  - Utilizes TCP/IP and UDP/IP, SpinNP protocols.
- SCSI module:
  - Processes the client's data communication through FC, Fibre Channel over Ethernet (FCoE), and iSCSI and translates it to Spin Network Protocol (SpinNP). Sends it to CSM.
  - Processes the data communication from CSM and translates it from SpinNP to the desired protocol.
  - Utilizes FC, SCSI, SpinNP, and TCP/IP protocols.
- Cluster session manager (CSM):
  - Provides SpinNP communication layer between the network, SCSI, and data modules.
  - Handles all traffic within the node or between the nodes.
  - Utilizes SpinNP protocol.

- Data module:
  - Takes the data from CSM and sends it to disks or tapes through WAFL file system (Write Anywhere File Layout).
  - Reads data from disk and translates it to SpinNP protocol.
  - Manages the WAFL, RAID, and storage.
  - Utilizes SpinNP protocol.
  - Communicates with FC and SAS disks and tape devices.

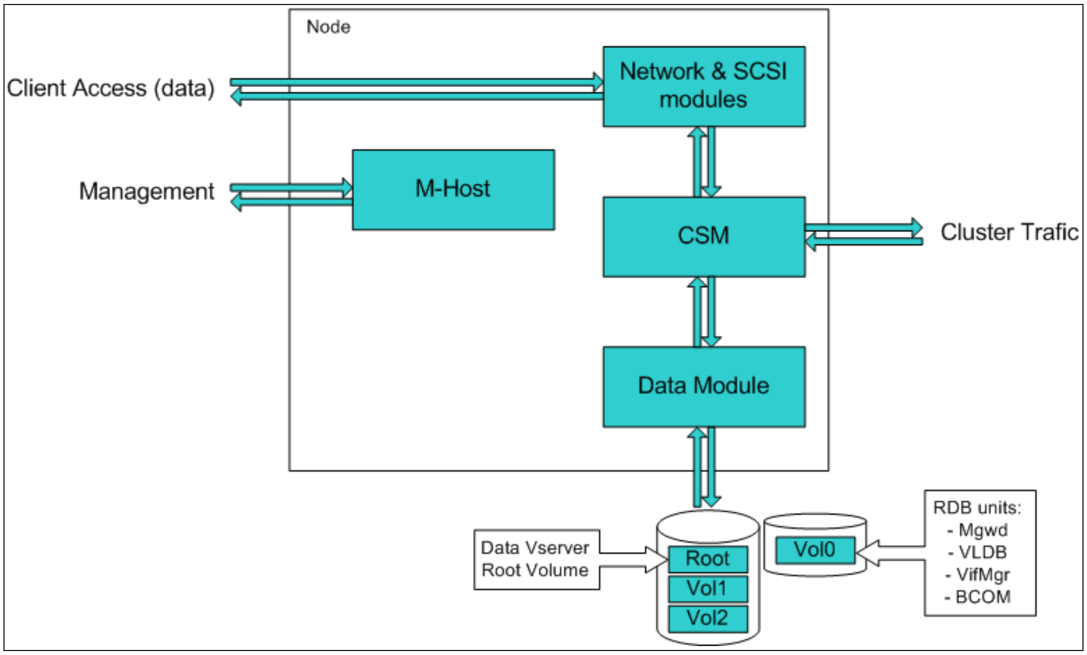


Figure 2-3 HW architecture of Cluster Data ONTAP node and dataflow within

2.2.1 Node Vol0 volume

A node's root volume (also referred to as *node Vol0*) contains special directories and configuration files for that node. The root aggregate contains the root volume.

A node's root volume is a FlexVol volume that is installed at the factory and reserved for system files, log files, and RDB databases. It is not accessible by NAS or SAN clients, nor it is part of the namespace of a storage virtual machine (SVM).

The vol0 volume does not need to be protected by mirroring or backups. In case of the disaster, the cluster automatically rebuilds the vol0.

The following rules govern the node's root volume:

- Do not change the preconfigured size for the root volume or modify the content of the root directory, unless technical support instructs you to do so.

  Editing configuration files directly in the root directory might result in an adverse impact on the health of the node and possibly the cluster. If you need to modify the system configurations, you use Data ONTAP commands to do so.

- Do not store user data in the root volume.

  Storing user data in the root volume increases the storage giveback time between nodes in an HA pair.

- Do not set the root volume's fractional reserve to any value other than 100%.

Note: The root aggregate must be dedicated to the root volume only. You must not include or create datavolumes in the root aggregate.

The cluster can contain multiple volumes with same name, as the volume is bound to the SVM.

The cluster cannot contain multiple aggregates with same name, as the aggregate is bound to the cluster.

When a new node joins the cluster, aggregates belonging to that node are auto renamed.

2.2.2 Local request path

When the data LIF receives the request from the client, it forwards the request to the network module (NAS) or SCSI module (SAN). The associated module then translates the request to SpinNP protocol and sends it through the CSM to data module. The data module translates the SpinNP and forwards the request to nonvolatile RAM (NVRAM), WAFL, and disks. In case of a write request, the acknowledgment is sent to the client. The whole process is shown in Figure 2-4.

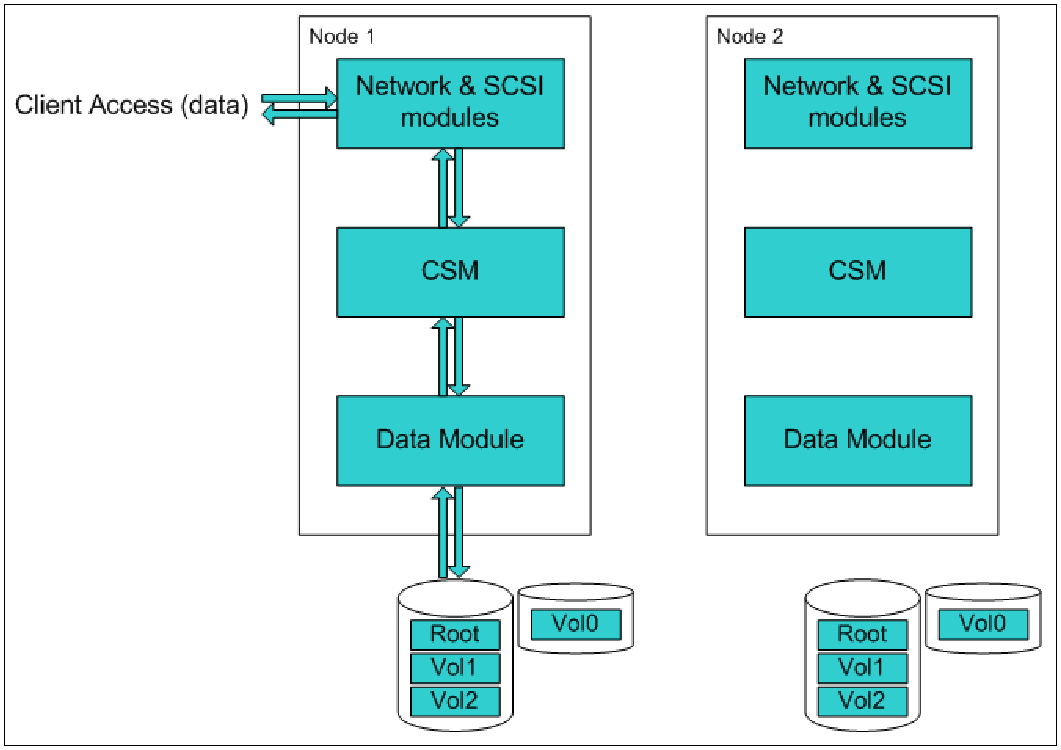


Figure 2-4 Local request path

2.2.3 Remote request path

When the LIF receives the request from the client, it forwards the request to the network module (NAS) or SCSI module (SAN). The associated module then translates the request to SpinNP protocol and sends it through the local CSM module to the remote CSM module and finally to the remote data module. The remote data module translates the SpinNP and forwards the request to nonvolatile RAM (NVRAM), WAFL, and disks. In case of a write request, the acknowledgment is sent to the client.

Communication between two CSM modules is served by cluster interconnect 10 Gbps dedicated switches or cluster interconnect hardwire in the case of a 2-node cluster.

The whole process is shown in Figure 2-5.

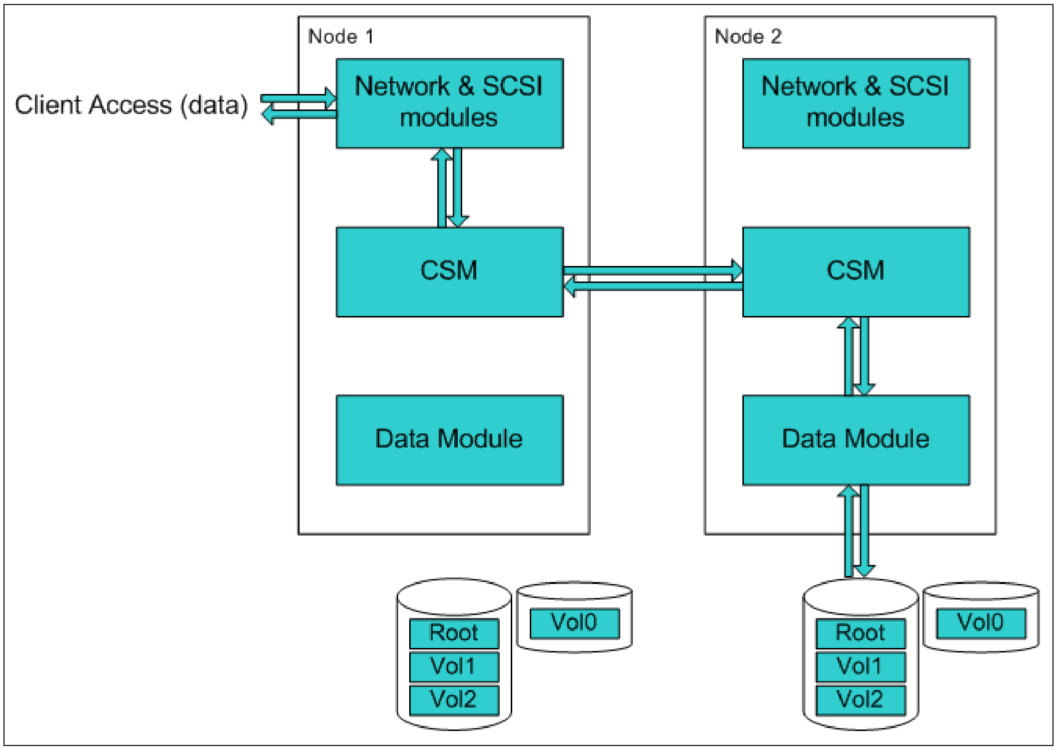


Figure 2-5 Remote request path

2.3 Storage virtual machine

A cluster consists of three types of storage virtual machines (SVMs), which help in managing the cluster and its resources and the data access to the clients and applications.

A cluster contains the following types of SVMs:

- Admin storage virtual machine
- Node storage virtual machine
- Data storage virtual machine

The cluster setup process automatically creates the admin SVM for the cluster. A node SVM is created when the node joins the cluster. The admin SVM represents the cluster, and node SVM represents the individual nodes of the cluster.

> **Note:** Multiple SVMs can coexist in a single cluster without being bound to any node in a cluster. However, they are bound to the physical cluster on which they exist.

2.3.1 Cluster management server

A cluster management server, also called an admin SVM, is a specialized SVM implementation that presents the cluster as a single manageable entity. In addition to serving as the highest-level administrative domain, the cluster management server owns resources that do not logically belong with a data SVM.

The cluster management server is always available on the cluster. You can access the cluster management server through the console, remote LAN manager, or cluster management LIF.

Upon failure of its home network port, the cluster management LIF automatically fails over to another node in the cluster. Depending on the connectivity characteristics of the management protocol you are using, you might or might not notice the failover. If you are using a connectionless protocol (for example, SNMP) or have a limited connection (for example, HTTP), you are not likely to notice the failover. However, if you are using a long-term connection (for example, SSH), then you will have to reconnect to the cluster management server after the failover.

Unlike a data SVM or node SVM, a cluster management server does not have a root volume or host user volumes (though it can host system volumes). Furthermore, a cluster management server can only have LIFs of the cluster management type.

Note: If you run the `vserver show` command, the cluster management server appears in the output listing for that command as shown in Example 2-1.

Example 2-1 Types of SVMs

```
cluster1::>vserver show
Admin Root Name Name
Vserver          Type       State      Volume     Aggregate  Service Mapping
-----------      -------    ---------  ---------- ---------- ------- -------
vs1.example.com  data       running    root_vol1  aggr1      file  file
cluster1         admin      -          -          -          -     -
cluster1-01      node       -          -          -          -     -
cluster1-02      node       -          -          -          -     -
vs2.example.com  data       running    root_vol2  aggr2      file  file
5 entries were displayed.
```

2.3.2 Data storage virtual machine

The data storage virtual machine (data SVM) represents the data serving SVMs. After the cluster setup, a cluster administrator must create data SVMs and add volumes to these SVMs to facilitate data access from the cluster. A cluster must have at least one data SVM to serve data to its clients.

The data storage virtual machine contains the following components:

- LIFs through which it serves data to the clients.
- One or more FlexVol volumes, or a single Infinite Volume.

An SVM securely isolates the shared virtualized data storage and network, and appears as a single dedicated server to its clients. Each SVM has a separate administrator authentication domain and can be managed independently by an SVM administrator.

In a cluster, an SVM facilitates data access. SVMs use the storage and network resources of the cluster. However, the volumes and LIFs are exclusive to the SVM. Multiple SVMs can coexist in a single cluster without being bound to any node in a cluster. However, they are bound to the physical cluster on which they exist.

Storage virtual machine with FlexVol volumes

An SVM with FlexVol volumes (Figure 2-6) in a NAS environment presents a single directory hierarchical view and has a unique namespace. The namespace enables the NAS clients to access data without specifying the physical location of the data. The namespace also enables the cluster and SVM administrators to manage distributed data storage as a single directory with multiple levels of hierarchy.

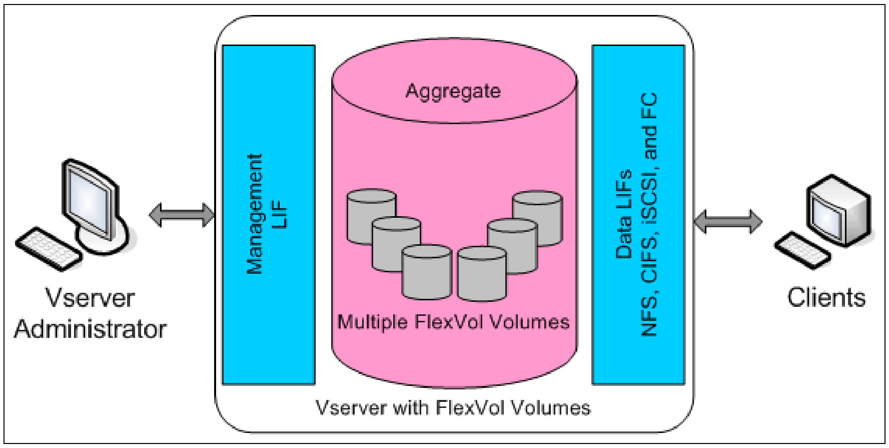


Figure 2-6 Storage virtual machine with FlexVol volumes

The volumes within each NAS SVM are related to each other through junctions and are mounted on junction paths. These junctions present the file system in each volume. The root volume of an SVM is a FlexVol volume that resides at the top level of the namespace hierarchy; additional volumes are mounted to the SVM's root volume to extend the namespace. As volumes are created for the SVM, the root volume of an SVM contains junction paths.

An SVM with FlexVol volumes can contain files and LUNs. It provides file-level data access by using NFS and CIFS protocols for the NAS clients, and block-level data access by using iSCSI, and Fibre Channel (FC) protocol (FCoE included) for SAN hosts.

Storage virtual machine with Infinite Volume

An SVM with Infinite Volume (Figure 2-7) can contain only one Infinite Volume to serve data. An SVM with Infinite Volume includes only one junction path, which has a default value of `/NS`. The junction provides a single mount point for the large namespace provided by the SVM with Infinite Volume.

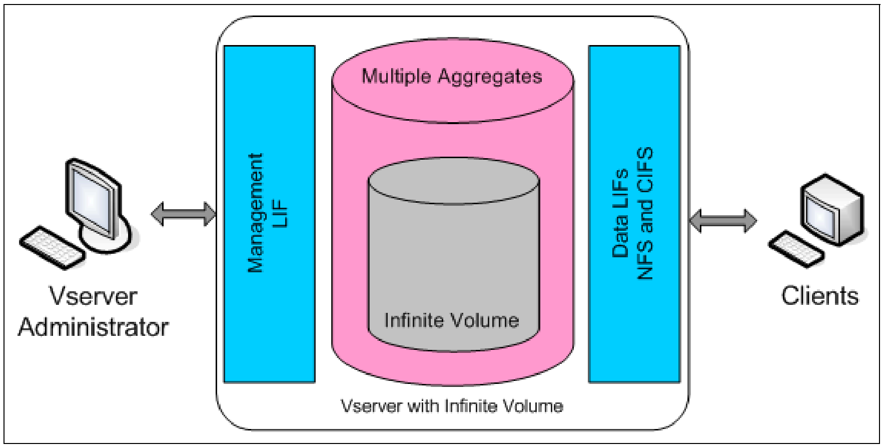


Figure 2-7 Storage virtual machine with Infinite Volume

You cannot add more junctions to an SVM with Infinite Volume. However, you can increase the size of the Infinite Volume.

An SVM with Infinite Volume can contain only files. It provides file-level data access by using NFS and CIFS (SMB 1.0) protocols. An SVM with Infinite Volume cannot contain LUNs and does not provide block-level data access.

2.3.3 The immortal storage virtual machine

Storage virtual machines (SVMs) provide data access to clients without regard to physical storage or controller, similar to any storage system. When you use SVMs, they provide benefits such as non-disruptive operation, scalability, security, and support unified storage.

An SVM has the following benefits:

- Non-disruptive operation:

  SVMs can operate continuously and non-disruptively for as long as they are needed. SVMs help clusters to operate continuously during software and hardware upgrades, addition and removal of nodes, and all administrative operations.

- Scalability:

  SVMs meet on-demand data throughput and the other storage requirements.

- Security:

  An SVM appears as a single independent server, which enables multiple SVMs to coexist while ensuring that no data flows among them.

- Unified storage:

  SVMs can serve data concurrently through multiple data access protocols. An SVM provides file-level data access by using NAS protocols, such as CIFS and NFS, and block-level data access by using SAN protocols, such as iSCSI and FC (FCoE included). An SVM can serve data to SAN and NAS clients independently at the same time.

> **Note:** An SVM with Infinite Volume can serve data only through NFS and CIFS (SMB 1.0) protocols.

- Delegation of management:

  An SVM can have its own user and administration authentication. SVM administrators can manage the SVMs that they are authorized to access. However, SVM administrators have privileges assigned by the cluster administrators.

- Easy Management of large datasets:

  With an SVM with Infinite Volume, management of large and unstructured data is easier, as the SVM administrator has to manage one data container instead of many.

2.3.4 SVM resiliency

The SVM root is the gateway for accessing volumes in NAS environments. If the SVM root is located on the storage of an HA pair where both nodes become unavailable, access to the root of the SVM will be disrupted, which might disrupt access to other volumes in the SVM. Having load-sharing mirrors of the SVM root adds a layer of resiliency for this scenario by allowing the promotion of the load-sharing mirror and resuming NAS access to the SVM. Although load-sharing mirrors can be set up on each HA pair to create the maximum level of resiliency for the SVM root, it is a preferred practice to have at least one load-sharing mirror on another HA pair in the cluster completely separate from the HA pair on which the SVM root resides.

2.3.5 Node (controller) resiliency

In the case of a planned or an unplanned event that takes a node down, HA pair controllers are configured to maintain access to the data. A node within an HA pair does a failover of its storage to the partner node to allow client requests to be processed during the time the node is down.

For each node in the cluster, there is a root aggregate. The root aggregate of each node should be dedicated as a root aggregate without having any user data stored on it. The reason for this is that, during giveback, the root aggregate will be returned to its home node first, assimilating the node back into the cluster and syncing necessary cluster information.

The process during which the root aggregate completes giveback can take time; therefore, any data residing in the root aggregate will not be available during this giveback period. Alternatively, data aggregates are not returned to their home node until the root aggregate is online and the node is back in quorum. Each data aggregate continues to serve data by the partner node while the root aggregate is brought online. When the data aggregates are to be returned to their home node, each data aggregate is sent home serially, minimizing the window during which each data aggregate will be unavailable during the giveback transition. Generally, a data aggregate takes up to 30 seconds to complete a giveback and resume servicing data requests.

Also, LIF failover groups should be defined for network connections to continue processing I/O requests non-disruptively in case of a node failure by migrating to an alternate node within the cluster. More is described in 2.5.4, “LIF resiliency” on page 24.

2.3.6 Storage virtual machine root volume

Every SVM has a root volume (shown in Figure 2-3 on page 12) that serves as the entry point to the namespace provided by that SVM. The root volume of any SVM is a FlexVol volume that resides at the top level of the namespace hierarchy and contains directories that are used as mount points, the paths where data volumes are junctioned into the namespace. These directories do not often change.

In the unlikely event that the root volume of the SVM is unavailable, NAS clients cannot access the namespace hierarchy and therefore cannot access data in the namespace. For this reason, it is considered a preferred practice to create a load-sharing mirror for the root volume on each node of the cluster so that the namespace directory information remains available in the event of a node outage or failover.

> **Note:** It is best not to store user data in the root volume of an SVM. Root volume of an SVM should be used for junction paths, and user data should be stored in non-root volumes of an SVM.

2.3.7 Cluster replication ring

A replication ring is a set of identical processes running on all nodes in the cluster.

The basis of clustering is the replicated database (RDB). An instance of the RDB is maintained on each node in a cluster. There are a number of processes that use the RDB to ensure consistent data across the cluster. These processes include the management application (mgmt), volume location database (vldb), virtual-interface manager (vifmgr), and SAN management daemon (bcomd).

For example, the vldb replication ring for a given cluster consists of all instances of vldb running in the cluster.

RDB replication requires healthy cluster links among all nodes in the cluster. If the cluster network fails in whole or in part, file services can become unavailable. The **`cluster ring show`** displays the status of replication rings and can assist with troubleshooting efforts.

2.3.8 Replicated database

In order for a cluster to remain synchronized, members of each replicated database (RDB) unit on every node are in constant communication. RDB resides on vol0 of each node and does not contain any user data. It contains only data that supports cluster operations.

The four RDB units on each node are as follows:

- Blocks configuration and Operations Manager (BCOM):
  - Hosts the SAN ring that contains information for block and LUN data access.
  - Stores initiator groups (igroups).
- Volume location database (VLDB):
  - Hosts index of nodes, aggregates and volumes.
- Virtual Interface Manager (VifMgr):
  - Responsible for creation and monitoring NFS, CIFS, and iSCSI LIFs.
  - Handles NAS LIF failover and migration of NAS LIFs to other network ports and nodes.
- Management:
  - Enables management of the cluster from any node.
  - Provides the CLI.

Each of these independent units elects one node from the cluster and hosts a “master” copy of its database on this node. The master copy is than replicated throughout the cluster transactionally. This ensures that data is either committed or rolled back. Reads are performed locally on each node. In case of discrepancies, the master copy is used.

Figure 2-8 shows four RDB units and their relation to the one of the nodes in a four-node cluster. Note the following characteristics:

- Node 1 holds Epsilon (tie-breaking ability).
- Each node has one local database for each RDB unit.
- Master databases are shown in blue color. They are independent to other RDB units.

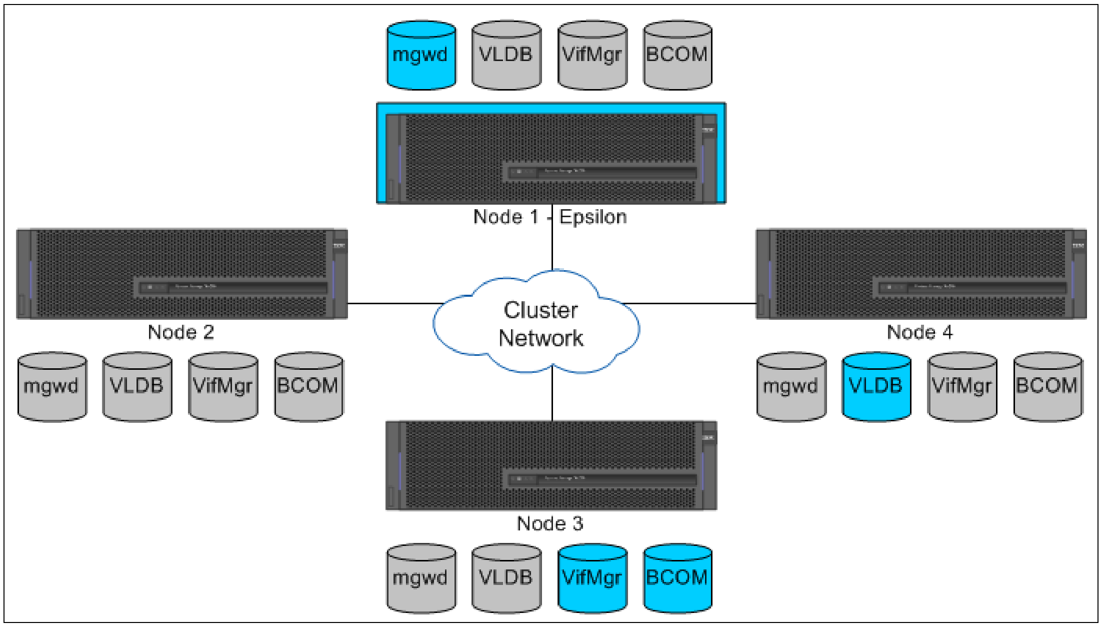


Figure 2-8 Cluster replication ring and RDB databases

2.4 Understanding cluster quorum, epsilon, and data availability

Quorum and epsilon are important measures of cluster health and function that together indicate how clusters address potential communications and connectivity challenges.

2.4.1 Cluster quorum

Quorum is a precondition for a fully-functioning cluster. When a cluster is in quorum, a simple majority of nodes are healthy and can communicate with each other. When quorum is lost, the cluster loses the ability to accomplish normal cluster operations. Only one collection of nodes can have quorum at any one time because all of the nodes collectively share a single view of the data. Therefore, if two non-communicating nodes are permitted to modify the data in divergent ways, it is no longer possible to reconcile the data into a single data view.

Each node in the cluster participates in a voting protocol that elects one node master; each remaining node is a secondary. The master node is responsible for synchronizing information across the cluster. When quorum is formed, it is maintained by continual voting; if the master node goes offline, a new master is elected by the nodes that remain online.

2.4.2 Epsilon

Because there is the possibility of a tie in a cluster that has an even number of nodes, one node has an extra fractional voting weight called epsilon (as shown in Figure 2-8). When the connectivity between two equal portions of a large cluster fails, the group of nodes containing epsilon maintains quorum, assuming that all of the nodes are healthy.

For example, if a single link is established between 12 nodes in one room and 12 nodes in another room to compose a 24-node cluster and the link fails, then the group of nodes that holds epsilon would maintain quorum and continue to serve data while the other 12 nodes would stop serving data. However, if the node holding epsilon was unhealthy or offline, then quorum would not be formed, and all of the nodes would stop serving data.

Epsilon is automatically assigned to the first node when the cluster is created. If the node that holds epsilon becomes unhealthy or is taken over by its high availability partner, epsilon does not move to another node but is rather no longer a factor in determining quorum. Example 2-2 shows how to migrate epsilon between nodes.

Example 2-2 Enabling / Disabling epsilon on given node

```
cluster1::> set -privilege advanced

Warning: These advanced commands are potentially dangerous; use them only
         when directed to do so by IBM personnel.
Do you want to continue? {y|n}: y

cluster1::*> cluster modify -node node0 -epsilon false
cluster1::*> cluster modify -node node1 -epsilon true
```

In general, assuming reliable connectivity among the nodes of the cluster, a larger cluster is more stable than a smaller cluster. The quorum requirement of a simple majority of half the nodes plus epsilon is easier to maintain in a cluster of 24 nodes than in a cluster of two nodes.

- When the cluster is in quorum, the following is true:
  - Configuration is not locked—the surviving nodes may make configuration changes.
  - Epsilon can be reassigned to another node in the cluster.
  - Data services can continue on the nodes that remain running.
  - If storage can be taken over and brought online, it will be.
  - Network interfaces that are configured to fail over to the surviving nodes will be brought online on those nodes.
  - Parts of exported file systems will not be available if storage is not taken over and no LSM exists.
- When the cluster is out of quorum, the following may be true for the cluster:
  - Configuration is locked—data service can continue on the nodes that remain running.
  - If storage can be taken over and brought online, it will be.
  - Network interfaces will not fail over because no quorum exists.
  - Parts of exported file systems will not be available if storage is not taken over and no LSM exists.

Note: A two-node cluster presents some unique challenges for maintaining quorum. In a two-node cluster, neither node holds epsilon; instead, both nodes are continuously polled to ensure that if one node fails, the other has full read-write access to data, as well as access to logical interfaces and management functions.

2.5 Clustered Data ONTAP networking

Nodes in a Clustered Data ONTAP system communicate through a dedicated 10 Gigabit Ethernet (10GbE) interconnect that is supplied with the system.

Starting with Clustered Data ONTAP 8.2, single node clusters without an Ethernet interconnect are supported. Clustered Data ONTAP 8.2 also supports two node switchless clusters.

Figure 2-9 shows the underlying network architecture of Clustered Data ONTAP. Three networks are shown.

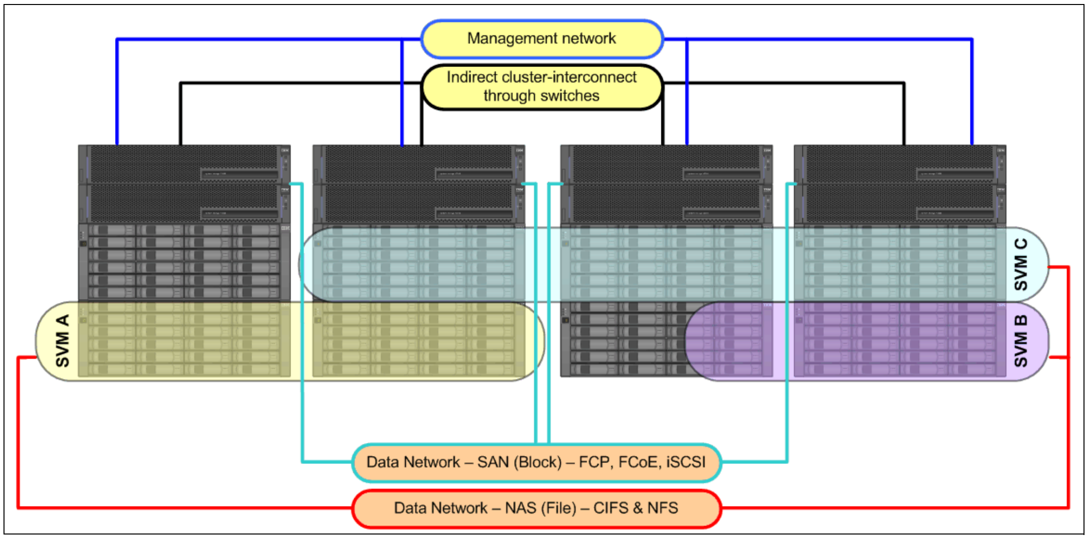

Management network
Indirect cluster-interconnect through switches
SVM C
SVM A
SVM B
Data Network – SAN (Block) – FCP, FCoE, iSCSI
Data Network – NAS (File) – CIFS & NFS

Figure 2-9 Clustered Data ONTAP networking overview.

2.5.1 Cluster-interconnect

A private, dedicated, redundant 10GbE network is used for communication between the cluster nodes and for DataMotion data migration within the cluster. The cluster-interconnect infrastructure is provided with every clustered ONTAP configuration to support this network of four or more nodes. Two-node clusters can be optionally configured without switches, utilizing point-to-point connections used instead of cluster-interconnect. This configuration is known as a switchless cluster. This entry-level configuration gives all the benefits of clustered ONTAP with a simpler infrastructure. Switchless clusters can be non-disruptively upgraded to include a switched cluster-interconnect when the cluster grows beyond two nodes.

The cluster interconnect is used for the following purposes:

- Indirect access to data on any node from any other node in the cluster
- Synchronizing configurations among nodes in the cluster
- Intracluster mirroring
- Volume movement

Only the approved switches that have been qualified for use with Clustered Data ONTAP can be used. The cluster interconnect infrastructure must not be shared for any other purpose.

2.5.2 Management network

All management traffic passes over this network. Management network switches can be included as part of a clustered ONTAP configuration, or customer-provided switches can be used:

- It is used for managing the cluster.
- It is separated from the data network for security.

IBM OnCommand System Manager, OnCommand Unified Manager, and other IBM applications are available for management, configuration, and monitoring of clustered ONTAP systems. System Manager provides GUI management, including a number of easy-to-use wizards for common tasks. Unified Manager provides monitoring and alerts.

2.5.3 Data networks

Provide data access services over Ethernet to NAS clients or Fibre Channel to SAN hosts. These networks are provided by the customer according to requirements and could also include connections to other clusters acting as volume replication targets for data protection.

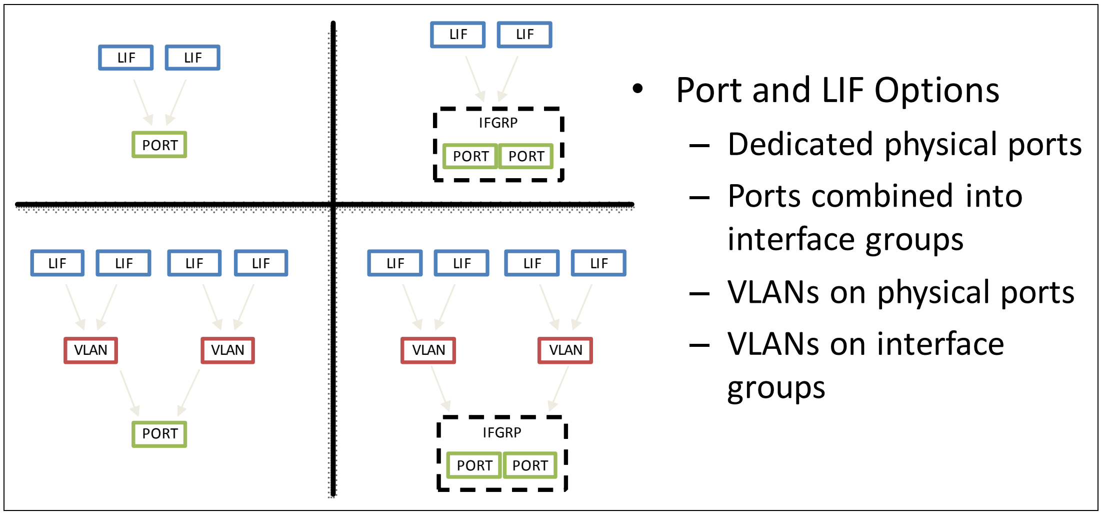
- They are client-facing networks that may consist of 1 GbE, 10 GbE, or Fibre Channel ports.
- The ports on the client-facing network house the LIFs. An LIF belongs to an SVM and has an address.
- A NAS LIF may move to different network ports on different physical node, therefore it is shown as connected to the SVM.
- A SAN LIF is bound to the node ports, therefore it is shown as connected to the node itself. SAN data LIFs (including iSCSI) do not migrate but will instead use ALUA and MPIO processes on the initiators to handle path failures.
- The data network may be segmented into different VLANs.
- Ports in the data network may be divided or joined together in order to use VLAN tagging or link aggregation

Note: Make certain that failover groups and the LIFs residing in them are configured correctly, meaning that you should configure the failover groups to use ports in the same subnet and verify that LIFs are assigned to the correct failover groups. If ports from different subnets are used in the same failover group or if LIFs are not assigned to the correct failover groups and a failover occurs, it will result in loss of network connectivity that will result in the loss of data availability.

2.5.4 LIF resiliency

The LIFs that are created on each node should have failover groups defined for each LIF. In the case of a node failure or a port failure, the LIF can fail over to another network port hosted on a different controller elsewhere in the cluster and continue to serve data. Having failover groups defined appropriately for an LIF that contain the same layer 2 network as the home port of the LIF is critical for achieving transparency for the workload during and after the LIF migration or failover.

LIFs can exist on top of physical ports, interface groups, or VLAN ports as shown in Figure 2-10.



- Port and LIF Options
  - Dedicated physical ports
  - Ports combined into interface groups
  - VLANs on physical ports
  - VLANs on interface groups

Figure 2-10 Port and LIF options

2.6 Multi-tenancy

Customers use multi-tenancy to get the security and management benefits of isolation, with the cost benefits of buying hardware that can be leveraged by multiple groups. It is faster to create a logical partition on an existing physical storage system than it is to procure and deploy a separate physical storage system. Separate physical storage systems for each tenant can be wasteful, as each system must be oversized to accommodate fluctuations in usage.

Multi-tenancy is fundamental to the design of Clustered Data ONTAP. Multiple clients or groups of applications are hosted on a cluster.

With each client’s data in an SVM, the following characteristics apply:

- ► SVM allows for access control separation.
- ► Their data is isolated from one another.
- ► Network resources can be partitioned.
- ► Architecture supports multiple NAS and SAN protocols.

2.7 Clustered Data ONTAP 8.2 licensing

A license is a record of one or more software entitlements. Installing license keys, also known as license codes, enables you to use certain features or services on your cluster.

When you set up a cluster, the setup wizard prompts you to enter the cluster base license key. To use additional features or functionality that require a license, in addition to installing the cluster base license, you must also install a license for the package that offers the functionality.

For the cluster to use licensed functionality, at least one node must be licensed for the functionality. It might be out of compliance to use licensed functionality on a node that does not have an entitlement for the functionality.

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 node entitles the node to use all SnapManager products in the package.

Starting with Data ONTAP 8.2, all license keys are 28 characters in length. Licenses installed prior to Data ONTAP 8.2 continue to work in Data ONTAP 8.2 and later releases. However, if you need to reinstall a license (for example, you deleted a previously installed license and want to reinstall it in Data ONTAP 8.2 or later, or you perform a controller replacement procedure for a node in a cluster running Data ONTAP 8.2 or later), Data ONTAP requires that you enter the license key in the 28-character format.

Notes:

- Starting with Data ONTAP 8.2, the high-availability (HA) functionality no longer requires a license.
- You still must set the **–mode** parameter of the **storage failover modify** command to **ha** to enable the HA functionality.

2.7.1 License management

License management offers the following features

- Add one or more license keys:

  ```
  system license add
  ```

- Display information about installed licenses:

  ```
  system license show
  ```

- Display the packages that require licenses and their current license status on the cluster:

  ```
  system license status show
  ```

- Delete a license from the cluster or a node whose serial number you specify:

  ```
  system license delete
  ```

  The cluster base license is required for the cluster to operate. Data ONTAP does not enable you to delete it.

- Display or remove expired or unused licenses:

  ```
  system license clean-up
  ```

2.7.2 License types

Understanding license types and the licensed method helps you manage the licenses in a cluster.

A package can have one or more of the following types of license installed in the cluster. The **`system license show`** command displays the installed license type or types for a package.

Standard license (license)

A standard license is a node-locked license. It is issued for a node with a specific system serial number (also known as a controller serial number). A standard license is valid only for the node that has the matching serial number.

> **Note:** The **`sysconfig`** command in the nodeshell displays the system serial number of a node. Installing a standard, node-locked license entitles a node to the licensed functionality. For the cluster to use licensed functionality, at least one node must be licensed for the functionality. It might be out of compliance to use licensed functionality on a node that does not have an entitlement for the functionality.

Data ONTAP 8.2 and later releases treat a license installed prior to Data ONTAP 8.2 as a standard license. Therefore, in Data ONTAP 8.2 and later releases, all nodes in the cluster automatically have the standard license for the package that the previously licensed functionality is part of. The **`system license show`** command with the **`-legacy yes`** parameter indicates such licenses.

Site license (site)

A site license is not tied to a specific system serial number. When you install a site license, all nodes in the cluster are entitled to the licensed functionality. The **`system license show`** command displays site licenses under the cluster serial number.

If your cluster has a site license and you remove a node from the cluster, the node does not carry the site license with it, and it is no longer entitled to the licensed functionality. If you add a node to a cluster that has a site license, the node is automatically entitled to the functionality granted by the site license.

Evaluation license (demo)

An evaluation license is a temporary license that expires after a certain period of time (indicated by the **`system license show`** command). It enables you to try certain software functionality without purchasing an entitlement. It is a cluster-wide license, and it is not tied to a specific serial number of a node.

If your cluster has an evaluation license for a package and you remove a node from the cluster, the node does not carry the evaluation license with it.

3

Terminology in Clustered Data ONTAP 8.2

This chapter describes changes in the new terminology introduced with Data ONTAP 8.2.

The following topics are covered:

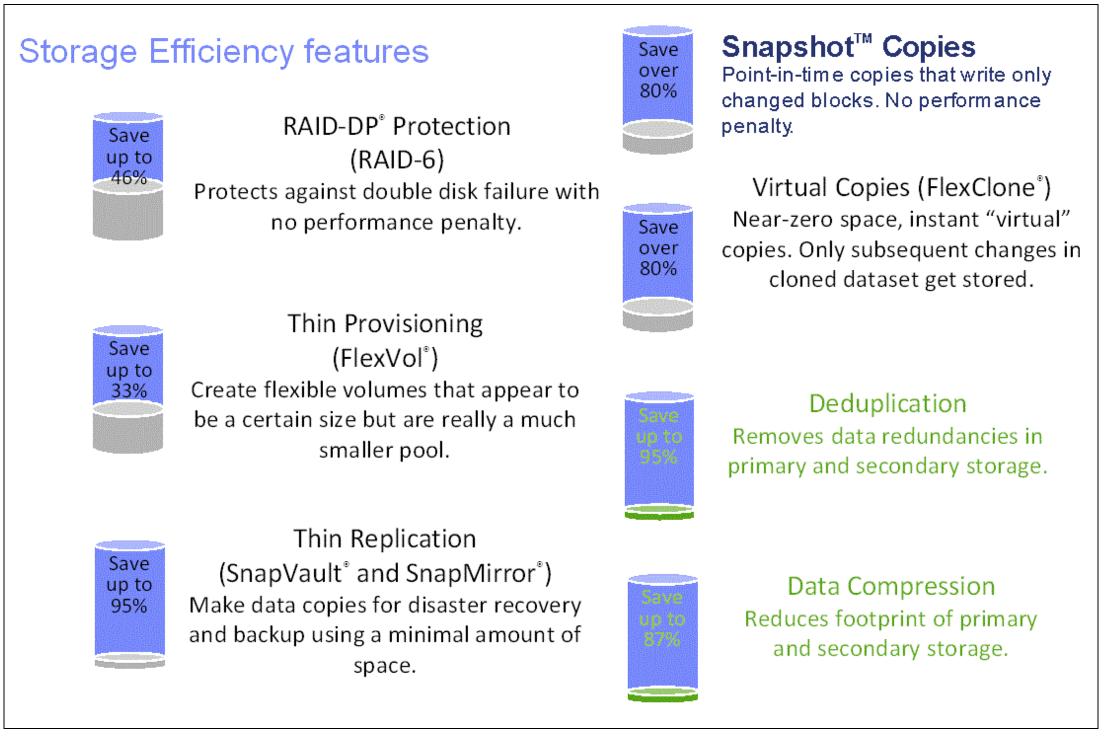
- Storage efficiency features
- Virtual Storage Tier Portfolio
- Cluster and high-availability terms
- New features included in cluster ONTAP 8.2

3.1 Storage efficiency features

In this section, we describe the storage efficiency features offered by this product:

- Deduplication:
  - Transparently eliminates duplicate blocks of data in each flexible volume, while preserving the files and LUNs that use those blocks.
  - Only unique data blocks are stored.
  - Duplicate blocks may be shared by many files or LUNs.
- Compression:
  - Compresses data stored on disk.
- FlexClone:
  - Near-instantaneous flexible volume cloning.
  - The cloned flexible volume will share common blocks with it’s source volume, until those blocks are changed.
- LUN cloning:
  - Near-instantaneous LUN cloning.
  - The cloned LUN will share common blocks with it’s source LUN, until those blocks are changed.
- Thin Provisioning:
  - Allows flexible volumes and LUNs to consume space as it is needed, rather than consuming it all when the volume or LUN is created.
- Virtual Storage Tiering:
  - Allows “hot” data that is frequently accessed to be transparently stored on faster media, like flash.
  - Three varieties are available:
    - Flash Accel: A host-based read cache that maintains Data coherency with the clustered ONTAP system.
    - Flash Cache: A PCI-e based read cache inside nodes that make up the cluster.
    - Flash Pool: A storage (aggregate-level) cache used to improve performance of both reads and writes.
- Snapshot:
  - Makes incremental, data-in-place, point-in-time copies of a LUN or volume with minimal performance impact.
  - Enables frequent, non-disruptive, space-efficient and quickly restorable backups.
- 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.
- RAID disk scrubbing:
  - The process in which a system reads each disk in the RAID group and tries to fix media errors by rewriting the data to another disk area.

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



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 3-1 Storage efficiency features

3.2 Virtual Storage Tier Portfolio

The following features are available:

- Flash Cache:
  - Random reads are cached on PCIe flash cards installed on a cluster node and served with faster response than from a hard disk drives (HDD).
  - Benefits works that have a high mix of repeat random reads, that is, a “hot” dataset that will fit in Flash Cache.
  - All data is stored/written to a single tier of HDD storage (either performance or capacity drives). Data, from any volume behind a controller, that is read randomly by a host or client is automatically cached.
  - Cache capacity is used efficiently because Flash Cache is dedupe and FlexClone aware.
  - Preferred practice for HA pairs is that both nodes have Flash Cache installed.
  - Performance improvement on an existing N Series system that does not already have Flash Cache installed can be modeled using Predictive Cache Statistics (PCS).
- Flash Pool:
  - Similar to Flash Cache, all data from volumes that are provisioned on Flash Pool are cached on aggregate-level cache resource; RAID-protected SSD-based cache.
  - Data that is read randomly by a host, and a recent update (overwrite) of previously written data, is cached. Caching overwrite data offloads IO from HDDs for data that is short-lived before another update. Non-overwrite writes are written to HDD.

  - Data cached in Flash Pool is available and accessible during planned and unplanned controller takeovers. It is also valid and available after a system or controller reboot.
- Flash Accel:
  - Flash Accel SW provides coherence between data stored and cache on Data ONTAP systems and data cached on a VMware vSphere host.
  - Flash Accel is a “write-through” cache, which means all data is stored on the Data ONTAP system, and Data ONTAP data protection and storage efficiency features can be used.
  - Flash Accel is advised as a supplement to storage cache (Flash Cache or Flash Pool), for workloads that benefit from reducing network latency between host and storage. Flash Cache and/or Flash Pool should be installed first.
  - Flash Accel is software only; it works with server cache cards and SSDs from other vendors.

3.3 Cluster and high-availability terms

The following terms are used:

- CFO: The term is now used for controller failover rather than cluster failover.
- SFO - storage failover: In Clustered Data ONTAP, the method of ensuring data availability by transferring the data service of a failed node to another node in an HA pair. Transfer of data service is often transparent to users and applications.
- Takeover: The emulation of the failed node identity by the takeover node in a high availability configuration; the opposite of giveback.
- Takeover mode: The method you use to interact with a node (storage system) when it has taken over its partner. The console prompt indicates when the node is in takeover mode.
- Giveback: The technology that enables two storage systems to return control of each other's data after the issues that caused a controller failover are resolved.
- Heartbeat: A repeating signal transmitted from one storage system to the other that indicates that the storage system is in operation. Heartbeat information is also stored on disk.
- Hot swap: The process of adding, removing, or replacing a disk while the storage system is running.
- Panic: A serious error condition causing the storage system or system to halt. Similar to a software crash in the Windows system environment.
- Partner node: From the point of view of the local node (storage system), the other node in a high-availability configuration.

The following cluster and high-availability terms have changed:

- Cluster:
  - In the Data ONTAP 7.1 release family and earlier releases, this 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, this 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, this 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.
- Beginning with the Data ONTAP 8.2 release, capabilities of the Data ONTAP operating system that were previously designated by the term Cluster-Mode are now referred to as Clustered Data ONTAP.

  For example, you might be upgrading from Data ONTAP 8.1.x running in Cluster-Mode to Clustered Data ONTAP 8.2. The underlying cluster technology is the same.

3.4 New features included in cluster ONTAP 8.2

The following Data ONTAP 8.2 features were introduced:

- Infinite Volume:
  - An Infinite Volume is a single volume that can scale up to 20PB of storage capacity non-disruptively. It provides a single mount point and integrates with features such as deduplication, compression, and data protection features.
- Load-sharing mirror:
  - A load-sharing mirror reduces the network traffic to a FlexVol volume by providing additional read-only access to clients. You can create and manage load-sharing mirrors to distribute read-only traffic away from a FlexVol volume. Load-sharing mirrors do not support Infinite Volumes.
- QoS:
  - QoS stands for “quality of service.” QoS allows an administrator to limit the number of I/O operations per second or raw throughput (MB/s) directed to a policy group that could consist of a single storage virtual machine (SVM), formerly referred to as a Vserver, or a group of LUNs, flexible volumes, or files within an SVM.
- Storage-Efficient SnapVault:
  - Volume-level logical replication for online backup and recovery that preserves compression and deduplication savings.
- Clustered Data ONTAP system and non-disruptive operations:
  - Non-disruptive operation (NDO) is the capability of systems operating in Clustered Data ONTAP to continuously serve data without disruption during maintenance and lifecycle operations as well as unplanned failure events. Software updates and configuration changes occur throughout any system’s lifecycle.

    Above and beyond this, in almost all environments, the hardware infrastructure must be added to and replaced, potentially many times. Many years after the system is originally commissioned, the data has outlived the hardware, so that little or none of the original hardware might remain. Through the NDO capabilities, all of these changes can be achieved without outage or impact on the applications or attached clients and hosts; the cluster entity has persisted intact.

    This concept is also being referred to as *“The Immortal Cluster.”*

- Namespace:
  - In Clustered Data ONTAP, FlexVol volumes containing NAS data are junctioned into the owning SVM in a hierarchy. This hierarchy presents NAS clients with a unified view of the storage, regardless of the physical location of FlexVol volumes inside the cluster. A volume must be junctioned into a namespace in order to be reachable by NAS clients.
  - Junctions allow each FlexVol volume to be browsable like a directory or folder. NFS clients can access multiple FlexVol volumes using a single mount point. CIFS clients can access multiple FlexVol volumes using a single CIFS share. The NAS namespace consists of the hierarchy of FlexVol volumes within a single SVM as presented to the NAS clients.
  - The key benefits are as follows:
    - Clients to not need to remount when a volume's physical location changes.
    - The physical storage layout can be managed independently of the logical storage layout.
    - Datasets can be distributed to increase performance.
- Logical interface (LIF):
  - LIFs are logical resources tied to an SVM. A NAS LIF has an IP address. A Fibre Channel LIF has a WWPN. An LIF is hosted by a physical network port on a node.
  - This separation of the physical port and the LIF allows a NAS LIF to migrate non-disruptively to another physical port on any node within the cluster.
  - Each SVM has its own World Wide Node Name (WWNN).
  - The SAN switch in use must support N-port virtualization (NPIV). NPIV allows multiple World Wide Port Names (WWPNs) on a single, physical Fibre Channel port. The WWPNs on the Fibre Channel port can belong to different SVMs. To the fabric, each "virtualized" WWPN appears as its own port.
- DataMotion / Volume move:
  - Vol move, formally referred to as DataMotion for volumes, is a feature in Clustered Data ONTAP. Vol move allows a volume to be moved non-disruptively from its current aggregate to a different aggregate within the cluster. The source and destination aggregates can be on any node, of any disk type, and of any aggregate type (32-bit or 64-bit). Sufficient space in the destination aggregate to create a volume of the same size as the source must exist.
  - Volumes cannot be moved between SVMs. SnapMirror can be used to copy a volume from one SVM to another.
  - Vol move is supported for all SAN and NAS protocols.
- Storage virtual machine (SVM) / Virtual storage server (Vserver):
  - An SVM contains data volumes and one or more LIFs through which it serves data to the clients. An SVM can either contain one or more FlexVol volumes, or a single Infinite Volume.
  - In a cluster, an SVM facilitates data access. A cluster must have at least one SVM to serve data.
  - A cluster can have one or more SVMs with FlexVol volumes and SVMs with Infinite Volumes.

4

Clustered Data ONTAP compared to 7-Mode

Clustered Data ONTAP offers significant innovations and enhancements over 7-Mode. The latest version of Clustered Data ONTAP offers innovative features such as QoS and in-place controller upgrades, and it also adds new CIFS features that are supported with 7-Mode.

SnapVault in Clustered Data ONTAP is storage efficient. A very small feature set supported in 7-Mode is not supported in Clustered Data ONTAP, but most of these features will be supported in future releases.

The following topics are covered:

- Storage virtual machine versus vFiler unit
- Failover and giveback comparison
- Data protection and load sharing
- Cluster management

4.1 Storage virtual machine versus vFiler unit

A storage virtual machine (SVM) is also referred to as a *Vserver*. It can be thought of as a secure virtual storage system that manages resources, including volumes and logical interfaces (LIFs). Separation of software from the hardware grants an SVM independent mobility of an SVM's LIFs and flexible volumes. An SVM is virtualized because its storage and network connections are not permanently bound to any node or group of nodes:

- The physical resources of the cluster are not bound to any particular SVM. An SVM's volumes can be moved to different physical aggregates without disrupting client access. Similarly, an SVM's LIFs can be moved to different physical network ports without disrupting client access.
- SVMs can be isolated in their own separate VLANs using separate physical ports or using VLAN tagging. Users of one SVM might or might not be granted access to another SVM.
- An SVM can serve both SAN and NAS concurrently.
- An SVM manages volumes. However, a cluster administrator can limit which aggregates an SVM can use for volume creation.
- SVMs can be provisioned on-the-fly for individual departments, companies, or applications. The same physical hardware can be used by many tenants. SVMs provide a secure logical boundary between tenants.

Several similarities between vFiler units in 7-Mode and Clustered Data ONTAP SVMs exist. Both allow a layer of storage virtualization, and both allow administrative control of just that virtual storage unit. However, some differences exist between the two as compared within Table 4-1.

Table 4-1 SVM and vFiler unit differences.

Clustered Data ONTAP SVM	7-Mode MultiStore vFiler Unit
Required. Clustered Data ONTAP needs at least one SVM defined in order to serve data.	Optional. A controller running Data ONTAP in 7-Mode can serve data without a vFiler unit defined.
Serves a single namespace. Multiple flexible volumes can be accessed using a single CIFS share or NFS export.	Volumes must be exported individually.
Uses resources from one or many nodes within the cluster.	Bound to the resources of a single controller.
Supports any Clustered Data ONTAP data protocol, including FCP and FCoE.	Limited to NFS, CIFS, and iSCSI.
Provides fully delegated role-based access control (RBAC).	Provides limited administration delegation.

4.2 Failover and giveback comparison

The HA pair controller architecture is very similar in both 7-Mode and Clustered Data ONTAP, however, there are a few differences.

For customers running within optimal limits, failover is expected to be from 15 through 45 seconds. Most customers can expect unplanned failover times of 30 seconds or less. Approximately 90% of our customer environments are within optimal limits. However, it is still a preferred practice to configure client-side timeouts to withstand a 120-second failover.

For customers that push the system limits by running with the maximum number of spindles, utilizing the maximum system capacity, using a large number of FlexVol volumes on a single node, or consistently running with CPU utilization greater than 50%, unplanned failover times might be longer than 45 seconds. 120 seconds is the preferred client-side timeout setting. This advice is consistent with preferred practices for 7-Mode installations.

Giveback is handled differently in Clustered Data ONTAP than in 7-Mode. During giveback in Clustered Data ONTAP, the root aggregate is sent home first so the node can be assimilated back into the cluster. During this time, all data aggregates continue to serve data using the partner node.

After the home node is ready to serve data, each data aggregate will be given back to the home node serially. Each data aggregate might take up to 30 seconds to complete giveback.

The total giveback time is the time it takes for the root aggregate to do a giveback, the time it takes to assimilate the node back into the cluster, and the time it takes to give back each aggregate. The entire giveback operation might be lengthy, but any single aggregate will only be unavailable for up to 30 seconds.

Planned takeover in Clustered Data ONTAP 8.2 is similar to giveback in previous versions of Clustered Data ONTAP. In these planned takeovers, aggregates are taken over one by one, reducing the total amount of time each aggregate is unavailable. Unplanned takeovers in Clustered Data ONTAP 8.2 behave the same as in previous versions of Clustered Data ONTAP.

Note: Recent versions of Data ONTAP (7-Mode and Clustered Data ONTAP) are configured by default to perform a giveback after a panic-induced takeover.

4.3 Data protection and load sharing

Data protection means backing up data and being able to recover it. You protect the data by making copies of it so that it is available for restoration even if the original is no longer available.

Businesses need data backup and protection for the following reasons:

- To protect data from accidentally deletions, application crashes, data corruption, and so on
- To archive data for future use
- To recover from a disaster

4.3.1 SnapMirror

Only asynchronous SnapMirror mirroring is supported. This can be set both within the cluster (intracluster) as well as between clusters (intercluster). The replication is at the volume level of granularity and is also known as a data protection (DP) mirror. Qtree SnapMirror is not available for Clustered Data ONTAP.

SnapMirror relationships can be throttled to a specific transfer rate using the **`snapmirror modify –throttle`** command.

4.3.2 SnapVault

SnapVault in Clustered Data ONTAP 8.2 delivers much of the same functionality you all may be familiar with from 7-Mode, the ability to store Snapshot copies on a secondary system for a long period of time, without taking up space on your primary system.

However, SnapVault in Clustered Data ONTAP is based on a new engine that uses volume-based logical replication, as opposed to SnapVault in 7-Mode, which used qtree-based replication. Since deduplication and compression operate at the flexible volume level, that represents a big advantage over 7-Mode. Storage efficiency is maintained while data is transferred to the backup system and is also maintained on the backup system. That translates to reduced backup times, and increased storage efficiency in the backup copy.

SnapVault is available in Clustered Data ONTAP 8.2 and above. Intercluster SnapVault is supported. SnapVault relationships between Clustered Data ONTAP and 7-Mode Data ONTAP are not supported.

4.3.3 NDMP

For FlexVol volumes, Data ONTAP supports tape backup and restore through the Network Data Management Protocol (NDMP). For Infinite Volumes, Data ONTAP supports tape backup and restore through a mounted volume. Infinite Volumes do not support NDMP. The type of volume determines what method to use for backup and recovery.

NDMP allows you to back up storage systems directly to tape, resulting in efficient use of network bandwidth. Clustered Data ONTAP supports dump engine for tape backup. Dump is a Snapshot copy-based backup to tape, in which your file system data is backed up to tape. The Data ONTAP dump engine backs up files, directories, and the applicable access control list (ACL) information to tape. You can back up an entire volume, an entire qtree, or a subtree that is neither an entire volume nor an entire qtree. Dump supports level-0, differential, and incremental backups. You can perform a dump backup or restore by using NDMP-compliant backup applications. Starting with Data ONTAP 8.2, only NDMP Version 4 is supported.

4.3.4 Data protection mirror

This feature provides asynchronous disaster recovery. Data protection mirror relationships enable you to periodically create Snapshot copies of data on one volume; copy those Snapshot copies to a partner volume (the destination volume), usually on another cluster; and retain those Snapshot copies. The mirror copy on the destination volume ensures quick availability and restoration of data from the time of the latest Snapshot copy, if the data on the source volume is corrupted or lost.

If you conduct tape backup and archival operations, you can perform them on the data that is already backed up on the destination volume.

4.3.5 Load-sharing mirror

A load-sharing mirror of a source flexible volume is a full, read-only copy of that flexible volume. Load-sharing mirrors are used to transparently off-load client read requests. Client write requests will fail unless directed to a specific writable path.

Load-sharing mirrors can be used to enable the availability of the data in the source flexible volume. Load-sharing mirrors will provide read-only access to the contents of the source flexible volume even if the source becomes unavailable. A load-sharing mirror can also be transparently promoted to become the read-write volume.

A cluster might have many load-sharing mirrors of a single source flexible volume. When load-sharing mirrors are used, every node in the cluster should have a load-sharing mirror of the source flexible volume. The node that currently hosts the source flexible volume should also have a load-sharing mirror. Identical load-sharing mirrors on the same node will yield no performance benefit.

Load-sharing mirrors are updated on demand or on a schedule that is defined by the cluster administrator. Writes made to the mirrored flexible volume will not be visible to readers of that flexible volume until the load-sharing mirrors are updated. Similarly, junctions added in the source flexible volume will not be visible to readers until the load-sharing mirrors are updated. Therefore, it is advised to use load-sharing mirrors for flexible volumes that are frequently read but infrequently written to.

SVM root volumes are typically small, contain only junctions to other volumes, do not contain user data, are frequently read, and are infrequently updated. SVM root volumes must be available for clients to traverse other volumes in the namespace. This makes SVM root volumes good candidates for mirroring across different nodes in the cluster.

In versions of Clustered Data ONTAP prior to 8.2, load-sharing mirrors were used to distribute access to read-only datasets. Clustered Data ONTAP 8.2 introduces FlexCache technology, which can also be used to distribute read access but provides write access and is space efficient.

Load-sharing mirrors are capable of supporting NAS only (CIFS/NFSv3). They do not support NFSv4 clients or SAN client protocol connections (FC, FCoE, or iSCSI).

4.4 Cluster management

One of the major benefits of Clustered Data ONTAP is the ability to manage the cluster as a single entity through the cluster management interface. After a node is joined to the cluster, its components can be managed within the cluster context.

Clustered Data ONTAP systems can be managed in a variety of ways:

- CLI: SSH to cluster, node, or SVMs
- GUI: OnCommand System Manager
- GUI: OnCommand Unified Manager
- GUI: OnCommand Insight
- GUI: OnCommand Balance

In Clustered Data ONTAP 8.2, latency metrics are available for SVMs, protocols, volumes, and nodes, among others. To find these metrics, browse the **`statistics show-periodic`** command directory in the CLI.

5

High availability (HA) pairs and failover behavior

High availability (HA) pairs provide hardware redundancy that is required for non-disruptive operations and fault tolerance. They give each node in the pair the software functionality to take over its partner's storage and subsequently give back the storage.

Takeover and giveback are the operations that let you take advantage of the HA configuration to perform non-disruptive operations and avoid service interruptions. Takeover is the process in which a node takes over the storage of its partner. Giveback is the process in which the storage is returned to the partner. You can initiate the processes in different ways.

The following topics are covered:

- What an HA pair is
- Preferred practices for HA pairs
- Cluster consisting of a single HA pair
- Non-disruptive operations and HA pairs support
- HA pairs within the cluster
- When takeovers occur
- How aggregate relocation works

5.1 What an HA pair is

An HA pair consists of two storage systems (nodes) whose controllers are connected to each other directly. In this configuration, one node can take over its partner's storage to provide continued data service if the partner goes down.

You can configure the HA pair so that each node in the pair shares access to a common set of storage, subnets, and tape drives, or each node can own its own distinct set of storage.

The controllers are connected to each other through an HA interconnect. This allows one node to serve data that resides on the disks of its failed partner node. Each node continually monitors its partner, mirroring the data for each other's nonvolatile memory (NVRAM or NVMEM). The interconnect is internal and requires no external cabling if both controllers are in the same chassis.

Figure 5-1 shows an HA pair in the cluster.

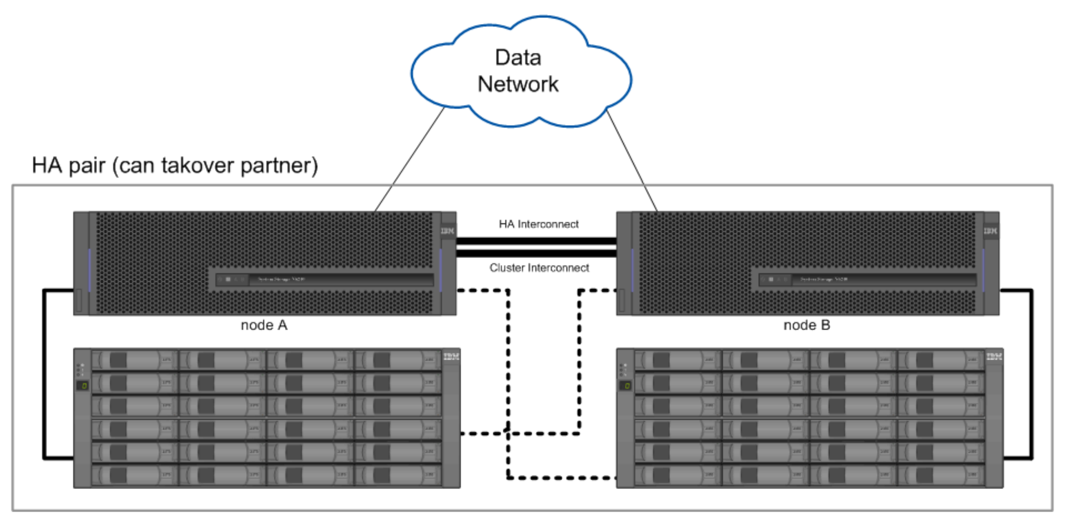


Figure 5-1 HA pair with dedicated HA interconnect and cluster interconnect

5.2 Preferred practices for HA pairs

Here we list some preferred practices for working with HA pairs:

- To ensure that your HA pair is robust and operational, you need to be familiar with configuration preferred practices.
- Do not use the root aggregate for storing data.
- Do not create new volumes on a node when takeover, giveback, or aggregate relocation operations are in progress or pending.
- Make sure that each power supply unit in the storage system is on a different power grid so that a single power outage does not affect all power supply units.
- Use the logical interface (LIF) with defined failover policies to provide redundancy and improve availability of network communication.
- Maintain consistent configuration between the two nodes.

  An inconsistent configuration is often the cause of failover problems.

- Test the failover capability routinely (for example, during planned maintenance) to ensure proper configuration.
- Make sure that each node has sufficient resources to adequately support the workload of both nodes during takeover mode.
- Use the Config Advisor tool to help ensure that failovers are successful.
- If your system supports remote management (through an RLM or Service Processor), make sure that you configure it properly.
- Follow advised limits for FlexVol volumes, dense volumes, Snapshot copies, and LUNs to reduce the takeover or giveback time.

  When adding traditional or FlexVol volumes to an HA pair, consider testing the takeover and giveback times to ensure that they fall within your requirements.
- Multipath HA is required on all HA pairs.
- Avoid using the **`-only-cfo-aggregates`** parameter with the **`storage failover giveback`** command.

5.3 Cluster consisting of a single HA pair

Cluster high availability (HA) is activated automatically when you enable storage failover on clusters that consist of two nodes, and you should be aware that automatic giveback is enabled by default. On clusters that consist of more than two nodes, automatic giveback is disabled by default, and cluster HA is disabled automatically.

A cluster with only two nodes presents unique challenges in maintaining a quorum, the state in which a majority of nodes in the cluster have good connectivity. In a two-node cluster, neither node holds epsilon, the value that designates one of the nodes as the master. Epsilon is required in clusters with more than two nodes. Instead, both nodes are polled continuously to ensure that if takeover occurs, the node that is still up and running has full read-write access to data as well as access to logical interfaces and management functions. This continuous polling function is referred to as cluster high availability (HA).

Cluster HA is different than and separate from the high availability provided by HA pairs and the **`storage failover`** commands. While crucial to full functional operation of the cluster after a failover, cluster HA does not provide the failover capability of the storage failover functionality.

Note: If you have a two-node switchless configuration that uses direct-cable connections between the nodes instead of a cluster interconnect switch, you must ensure that the `switchless-cluster-network` option is enabled. This ensures proper cluster communication between the nodes.

Steps to enable cluster HA and switchless-cluster in a two-node cluster:

1. Enter the following command to enable cluster HA:

   ```
   cluster ha modify -configured true
   ```

   If storage failover is not already enabled, you will be prompted to confirm enabling of both storage failover and auto-giveback.

2. If you have a two-node switchless cluster, enter the following commands to verify that the switchless-cluster option is set:

   a. Enter the following command to change to the advanced-privilege level:

      ```
      set -privilege advanced
      ```

      Confirm when prompted to continue into advanced mode. The advanced mode prompt appears (*>).

   b. Enter the following command:

      ```
      network options switchless-cluster show
      ```

      If the output shows that the value is false, you must issue the following command:

      ```
      network options switchless-cluster modify true
      ```

   c. Enter the following command to return to the admin privilege level:

      ```
      set -privilege admin
      ```

5.4 Non-disruptive operations and HA pairs support

HA pairs provide fault tolerance and let you perform non-disruptive operations, including hardware and software upgrades, relocation of aggregate ownership, and hardware maintenance:

- Fault tolerance:

  When one node fails or becomes impaired and a takeover occurs, the partner node continues to serve the failed node’s data.

- Non-disruptive software upgrades or hardware maintenance:

  During hardware maintenance or upgrades, when you halt one node and a takeover occurs (automatically, unless you specify otherwise), the partner node continues to serve data for the halted node while you upgrade or perform maintenance on the node you halted. During non-disruptive upgrades of Data ONTAP, the user manually enters the **`storage failover takeover`** command to take over the partner node to allow the software upgrade to occur. The takeover node continues to serve data for both nodes during this operation.

  Non-disruptive aggregate ownership relocation can be performed without a takeover and giveback.

The HA pair supplies non-disruptive operation and fault tolerance due to the following aspects of its configuration:

- The controllers in the HA pair are connected to each other either through an HA interconnect consisting of adapters and cables, or, in systems with two controllers in the same chassis, through an internal interconnect.
- The nodes use the interconnect to perform the following tasks:
  - Each node continually checks whether the other node is functioning.
  - They mirror log data for each other’s NVRAM or NVMEM.
  - They use two or more disk shelf loops, or storage arrays, in which the following conditions apply:
    - Each node manages its own disks or array LUNs.
    - In case of takeover, the surviving node provides read/write access to the partner's disks or array LUNs until the failed node becomes available again.

- They own their spare disks, spare array LUNs, or both, and do not share them with the other node.
- They each have mailbox disks or array LUNs on the root volume that perform the following tasks:
  - Maintain consistency between the pair
  - Continually check whether the other node is running or whether it has performed a takeover
  - Store configuration information

5.5 HA pairs within the cluster

HA pairs are components of the cluster, and both nodes in the HA pair are connected to other nodes in the cluster through the data and cluster networks. But only the nodes in the HA pair can takeover each other's storage.

Although the controllers in an HA pair are connected to other controllers in the cluster through the cluster network, the HA interconnect and disk-shelf connections are found only between the node and its partner and their disk shelves or array LUNs.

The HA interconnect and each node's connections to the partner's storage provide physical support for high-availability functionality. The high-availability storage failover capability does not extend to other nodes in the cluster.

Note: Network failover does not rely on the HA interconnect and allows data network interfaces to failover to different nodes in the cluster outside the HA pair. Network failover is different than storage failover since it enables network resiliency across all nodes in the cluster.

Non-HA (or stand-alone) nodes are not supported in a cluster containing two or more nodes. Although single node clusters are supported, joining two separate single node clusters to create one cluster is not supported, unless you wipe clean one of the single node clusters and join it to the other to create a two-node cluster that consists of an HA pair.

The diagram in Figure 5-2 shows two HA pairs in the cluster.

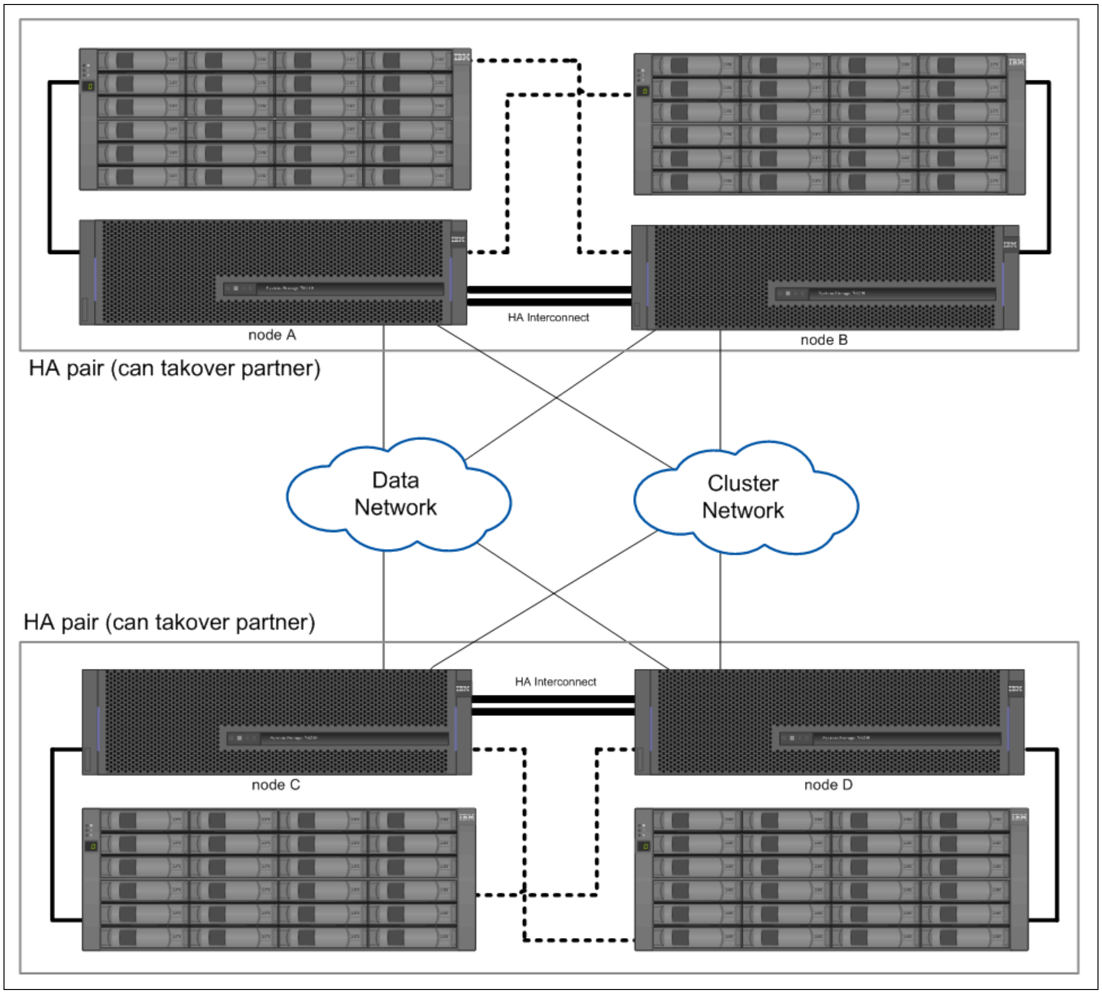


Figure 5-2 Two HA pairs in cluster

5.6 When takeovers occur

Takeovers can be initiated manually or occur automatically when a failover event happens, depending on how you configure the HA pair. In some cases, takeovers occur automatically regardless of configuration.

Takeovers can occur under the following conditions:

- A takeover is manually initiated with the **`storage failover takeover`** command.
- A node is in an HA pair with the default configuration for immediate takeover on panic, and that node undergoes a software or system failure that leads to a panic.

  By default, the node automatically performs a giveback to return the partner to normal operation after the partner has recovered from the panic and booted up.
- A node that is in an HA pair undergoes a system failure (for example, a loss of power) and cannot reboot.

  If the storage for a node also loses power at the same time, a standard takeover is not possible.

- A node does not receive heartbeat messages from its partner.

  This could happen if the partner experienced a hardware or software failure that did not result in a panic but still prevented it from functioning correctly.

- You halt one of the nodes without using the **`-f`** or **`-inhibit-takeover true`** parameter.
- You reboot one of the nodes without using the **`-inhibit-takeover true`** parameter.

  The **`-onreboot`** parameter of the **`storage failover`** command is enabled by default.

- Hardware-assisted takeover is enabled and triggers a takeover when the remote management device (RLM or Service Processor) detects failure of the partner node.

5.6.1 What happens during takeover

When a node takes over its partner, it continues to serve and update data in the partner's aggregates and volumes. To do this, it takes ownership of the partner's aggregates, and the partner's LIFs migrate according to network interface failover rules. Except for specific SMB 3.0 connections, existing SMB (CIFS) sessions are disconnected when the takeover occurs.

The following steps occur when a node takes over its partner:

1. If the negotiated takeover is user-initiated, aggregate relocation is performed to move data aggregates one at a time from the partner node to the node that is doing the takeover.

   The current owner of each aggregate (except for the root aggregate) is changed from the target node to the node that is doing the takeover. There is a brief outage for each aggregate as ownership is changed. This outage is less than that accrued during a takeover that does not use aggregate relocation.

   You can monitor the progress using the **`storage failover show-takeover`** command.

   The aggregate relocation can be avoided during this takeover instance by using the **`-bypass-optimization`** parameter with the **`storage failover takeover`** command.

   To bypass aggregate relocation during all future planned takeovers, set the **`-bypass-takeover-optimization`** parameter of the **`storage failover`** command to **`true`**.

   > **Note:** Aggregates are relocated serially during planned takeover operations to reduce client outage. If aggregate relocation is bypassed, it will result in longer client outage during planned takeover events.

2. If the takeover is user-initiated, the target node gracefully shuts down, followed by takeover of the target node's root aggr and any aggrs that were not relocated in step 1.
3. Data LIFs migrate from the target node to the node doing the takeover, or any other node in the cluster based on LIF failover rules, before the storage takeover begins.

   The LIF migration can be avoided by using the **`-skip-lif-migration`** parameter with the **`storage failover takeover`** command.

4. Existing SMB (CIFS) sessions are disconnected when takeover occurs.

   > **Note:** Due to the nature of the SMB protocol, all SMB sessions except for SMB 3.0 sessions connected to shares with the Continuous Availability property set will be disruptive. SMB 1.0 and SMB 2.x sessions cannot reconnect after a takeover event. Therefore, takeover is disruptive and some data loss could occur.

5. SMB 3.0 sessions established to shares with the *`Continuous Availability`* property set can reconnect to the disconnected shares after a takeover event. If your site uses SMB 3.0

connections to Microsoft Hyper-V and the *Continuous* `Availability` property is set on the associated shares, takeover will be non-disruptive for those sessions.

5.6.2 What happens during giveback

The local node returns ownership of the aggregates and volumes to the partner node after any issues on the partner node are resolved or maintenance is complete. In addition, the local node returns ownership when the partner node has booted up and giveback is initiated either manually or automatically.

The following process takes place in a normal giveback. In this discussion, node A has taken over node B. Any issues on Node B have been resolved and it is ready to resume operations.

1. Any issues on node B have been resolved and it is displaying the following message: `Waiting for giveback`
2. The giveback is initiated by the **`storage failover giveback`** command or by automatic giveback if the system is configured for it. This initiates the process of returning ownership of the node B's aggregates and volumes from node A back to node B.
3. Node A returns control of the root aggregate first.
4. Node B proceeds to complete the process of booting up to its normal operating state.
5. As soon as Node B is at the point in the boot process where it can accept the non-root aggregates, Node A returns ownership of the other aggregates one at a time until giveback is complete.

   You can monitor the progress of the giveback with the **`storage failover show-giveback`** command.

I/O resumes for each aggregate when giveback is complete for that aggregate, therefore reducing the overall outage window of each aggregate.

5.6.3 Displaying the nodes in a cluster

You can display information about the nodes in a cluster and their state. Example 5-1 displays information about all nodes in a four-node cluster.

Example 5-1 Displaying the nodes in a cluster

```
cluster1::> cluster show
Node     Health   Eligibility
--------------------- ------- ------------
node0    true     true
node1    true     true
node2    true     true
node3    true     true
```

The command displays the following information:

- Node name
- Whether the node is healthy
- Whether the node is eligible to participate in the cluster
- Whether the node holds epsilon (advanced privilege level or higher only)

5.6.4 HA policy and giveback of the root aggregate and volume

Aggregates are automatically assigned an HA policy of controller failover (CFO) or storage failover (SFO) that determines how the aggregate and its volumes are given back.

Aggregates created on Clustered Data ONTAP systems (except for the root aggregate containing the root volume) have an HA policy of SFO. During the giveback process, they are given back one at a time after the taken-over system boots.

The root aggregate always has an HA policy of CFO and is given back at the start of the giveback operation. This is necessary to allow the taken-over system to boot. The other aggregates are given back one at a time after the taken-over node completes the boot process.

The HA policy of an aggregate cannot be changed from SFO to CFO in normal operation.

5.7 How aggregate relocation works

Aggregate relocation operations take advantage of the HA configuration to move the ownership of storage aggregates within the HA pair. Aggregate relocation occurs automatically during manually initiated takeover to reduce downtime during planned failover events such as non-disruptive software upgrade, and can be initiated manually for load balancing, maintenance, and non-disruptive controller upgrade. Aggregate relocation cannot move ownership of the root aggregate.

Figure 5-3 shows the relocation of the ownership of aggregate aggr_1 from node1 to node2 in the HA pair.

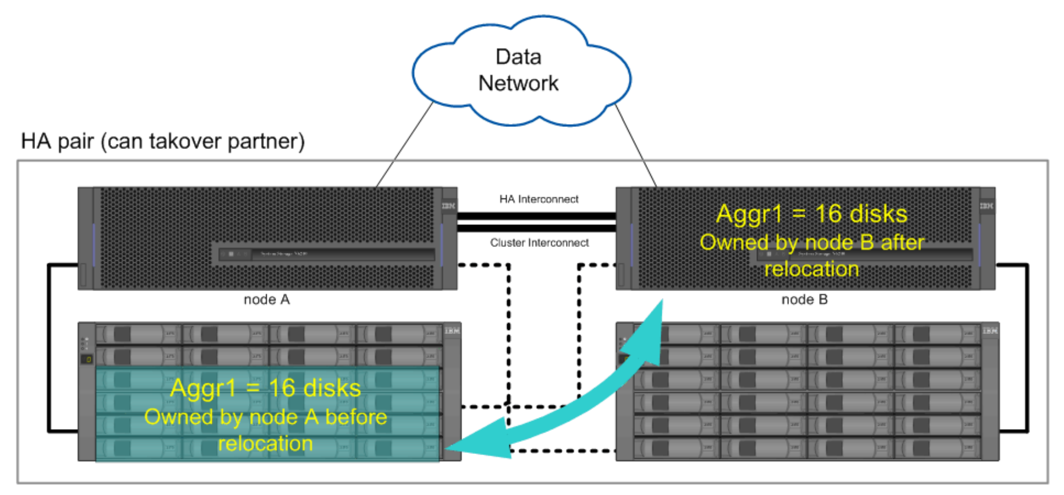

Data Network
HA pair (can takover partner)
HA Interconnect
Cluster Interconnect
Aggr1 = 16 disks
Owned by node B after relocation
node A
node B
Aggr1 = 16 disks
Owned by node A before relocation

Figure 5-3 Aggr1 - relocation of the ownership

The aggregate relocation operation can relocate the ownership of one or more storage failover (SFO) aggregates if the destination node can support the number of volumes in the aggregates. There is only a short interruption of access to each aggregate. Ownership information is changed one by one for the aggregates.

During takeover, aggregate relocation happens automatically when the takeover is initiated manually. Before the target controller is taken over, ownership of the aggregates belonging to that controller are moved one at a time to the partner controller. When giveback is initiated, the ownership is automatically moved back to the original node. The **`-bypass-optimization`** parameter can be used with the **`storage failover takeover`** command to suppress aggregate relocation during the takeover.

The aggregate relocation requires additional steps if the aggregate is currently used by an Infinite Volume with SnapDiff enabled.

Aggregate relocation and Infinite Volumes with SnapDiff enabled:

The aggregate relocation requires additional steps if the aggregate is currently used by an Infinite Volume with SnapDiff enabled. You must ensure that the destination node has a namespace mirror constituent and make decisions about relocating aggregates that include namespace constituents.

6

Physical cluster types and scaling

Clustered Data ONTAP allows the inclusion of different controller types in the same cluster, protecting the initial hardware investment and giving the flexibility to adapt resources to meet the business demands of the workloads. Similarly, support for different disk types, including serial attached SCSI (SAS), serial advanced technology attachment (SATA), and solid state disk (SSD), makes it possible to deploy integrated storage tiering for different data types, together with the transparent DataMotion capabilities of clustered ONTAP.

Flash Cache cards can also be used to provide accelerated read performance for frequently accessed data. Flash Pool intelligent caching is supported, which combines SSD with traditional hard drives for optimal performance and efficiency using virtual storage tiering. The highly adaptable clustered ONTAP architecture is key to delivering maximum, on-demand flexibility for the shared IT infrastructure, offering flexible options to address needs for performance, price, and capacity.

The following topics are covered:

- Physical cluster types
- Supported systems and cluster configurations for Data ONTAP 8.2
- Intelligent scale-out storage versus scale-up
- Non-disruptive operations

6.1 Physical cluster types

Here we outline the implementation of Clustered Data ONTAP network configurations. We provide common Clustered Data ONTAP network deployment scenarios and networking preferred practices as they pertain to a Clustered Data ONTAP environment. A thorough understanding of the networking components of a Clustered Data ONTAP environment is vital to successful implementations.

6.1.1 Single node cluster

A single node cluster as shown in Figure 6-1 is a special implementation of a cluster running on a standalone node. You can deploy a single node cluster if your workload only requires a single node, but does not need non-disruptive operations.

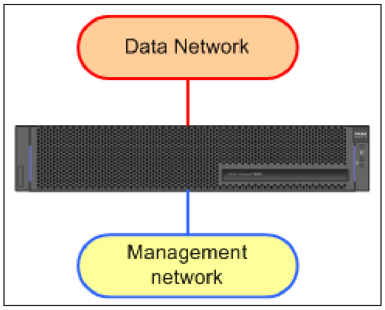


Figure 6-1 Single node cluster

For example, you could deploy a single node cluster to provide data protection for a remote office. In this scenario, the single node cluster would use SnapMirror and SnapVault to replicate the site's data to the primary data center.

In a single node cluster, the high availability (HA) mode is set to standalone, which enables the node to use all of the nonvolatile memory (NVRAM) on the NVRAM card. In addition, single node clusters do not use a cluster network, and you can use the cluster ports as data ports that can host data logical interfaces (LIFs).

Single node clusters are typically configured when the cluster is set up, by using the Cluster Setup wizard. However, you can remove nodes from an existing cluster to create a single node cluster.

The following features and operations are not supported for single node clusters:

- Storage failover and cluster HA:

  Single node clusters operate in a standalone HA mode. If the node goes offline, clients will not be able to access data stored in the cluster.

- Any operation that requires more than one node:

  This includes volume move or performing most copy operations.

- Infinite Volumes:

  Infinite Volumes must contain aggregates from at least two nodes.

- Storing cluster configuration backups in the cluster:

  By default, the configuration backup schedule creates backups of the cluster configuration and stores them on different nodes throughout the cluster. However, if the cluster consists of a single node and you experience a disaster in which the node becomes inaccessible, you will not be able to recover the cluster unless the cluster configuration backup file is stored at a remote URL.

> **Note:** To expand a single node cluster, an identical node must be added to form an HA pair. This is a disruptive operation.

6.1.2 Two-node cluster

This type of cluster takes all benefits of Clustered Data ONTAP operating system, including non-disruptive operations with volume move, LIF migrate, multipath ALUA/MPIO for SAN, online controller hardware upgrade, and so on. A two-node cluster can be seen in Figure 6-2.

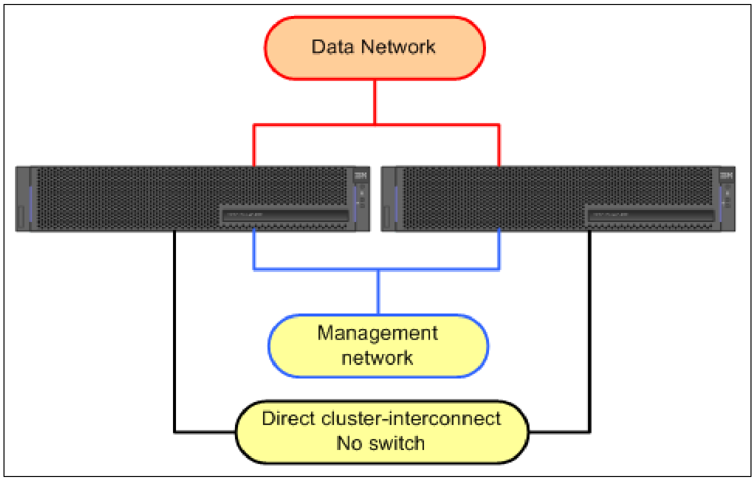


Figure 6-2 Two-node cluster

Non-HA (stand-alone) nodes are not supported in a cluster containing two or more nodes. Although single node clusters are supported, joining two separate single node clusters to create one cluster is not supported, unless you wipe clean one of the single node clusters and join it to the other to create a two-node cluster that consists of an HA pair.

> **Notes:**
>
> - A cluster consisting of only two nodes requires special configuration settings. Cluster high availability (HA) must be configured in a cluster if it contains only two nodes. Cluster HA ensures that the failure of one node does not disable the cluster.
> - Also, if you have a switchless configuration (supported in Clustered Data ONTAP 8.2), in which there is no cluster interconnect switch, you must ensure that the `switchless-cluster-network` option is enabled. This ensures proper cluster communication between the nodes.

6.1.3 Multinode cluster

A multinode cluster can contain up to 24 nodes running a NAS configuration, while running iSCSI or FC protocols within the cluster (Data ONTAP 8.2) will limit this to only eight nodes.

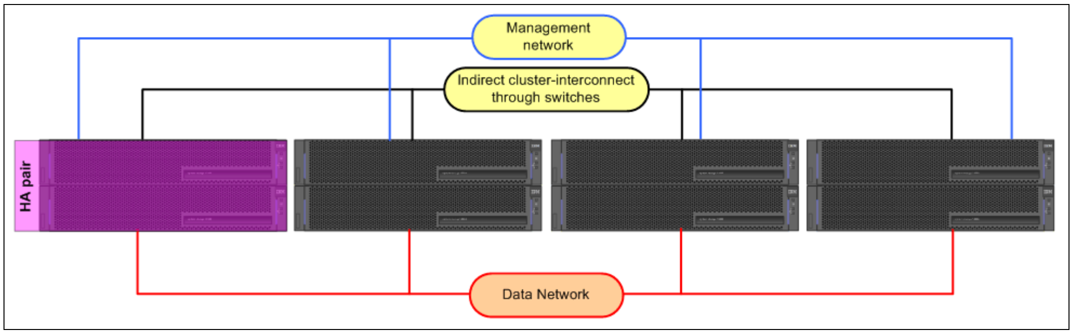


Figure 6-3 Multinode cluster

Note: A non-disruptive procedure describing how to add an interconnect switch is available for two-node clusters. This allows you to seamlessly scale-out the cluster.

6.2 Supported systems and cluster configurations for Data ONTAP 8.2

You need to ensure that you have the required storage systems and firmware to run Clustered Data ONTAP software for the Data ONTAP 8.2 release family.

A cluster can be homogenous (all nodes are the same platform model) or mixed (it has nodes with different platform models); mixed clusters are supported with some restrictions. The maximum number of nodes within a cluster is determined by the platform that supports the fewest number of nodes. The rules on node limits are shown in Table 6-1.

Table 6-1 Rules on node limits

Storage system type	Maximum node count
N3000 series	4
N6000 series	8
N7000 series with no iSCSI or FCP licenses	24
N7000 series with iSCSI or FCP licenses	8
Mixed N6000 series and N7000 series:	8
Mixed N3000 series and N6000 series	4
Mixing N3000 series and N7000 series	Not allowed

Other requirements and points for consideration

Here we list some other requirements for your consideration:

- Single node and two node switchless cluster configurations are designed to allow every customer to run Clustered Data ONTAP from the outset. These configurations also support branch and remote office configurations in which the remote storage is mirrored or vaulted to an IBM storage cluster in a larger data center.
- To expand a single node cluster, an identical node is added to form an HA pair.
- After the first HA pair is created, additional nodes are added in increments of two-node HA pairs.
- HA pair nodes must be homogeneous. Although most of our customers configure clusters using the same midrange or high-end N series platform, heterogeneous platforms in one cluster are supported.
- Software enablement must be homogeneous across the cluster except when upgrading versions.
- Approved cluster infrastructure must be deployed for all new installations. This includes two approved cluster interconnect switches (except for single node and two node switchless configurations) running the firmware release and the Reference Configuration File (RCF) that are prescribed for the version of Clustered Data ONTAP being used.

6.3 Intelligent scale-out storage versus scale-up

Clustered ONTAP can scale both vertically and horizontally via the addition of nodes and storage to the cluster. This scalability, combined with protocol-neutral, proven storage efficiency, can meet the needs of the most demanding workloads.

Scale-out storage used in the Clustered Data ONTAP is the most powerful and flexible way to respond to the inevitable data growth and data management challenges in today's environments. Consider that all storage controllers have physical limits to their expandability (for example, number of CPUs, memory slots and space for disk shelves) that dictate the maximum capacity and performance of which the controller is capable. If more storage or performance capacity is needed, you might be able to upgrade or add CPUs and memory or install additional disk shelves, but ultimately the controller will be completely populated, with no further expansion possible. At this stage, the only option is to acquire one or more additional controllers. Historically this has been achieved by simple "scale-up," with two options, either replace the old controller with a complete technology refresh, or run the new controller side by side with the original. Both of these options have significant shortcomings and disadvantages.

In a technology refresh, data migration is necessary to copy the data from the old to the new controller and reconfigure the environment there. This is time-consuming, planning-intensive, often disruptive, and typically requires configuration changes on all of the attached host systems in order to access the new storage resource. Data migrations have a substantial impact on storage administration costs and administrative complexity. If the newer controller will co-exist with the original controller, there are now two storage controllers to be individually managed, and there are no built-in tools to balance or reassign workloads across them. Data migrations will also be required. The situation becomes worse as the number of controllers increases.

Using scale-up increases the operational burden as the environment grows, and the result is an unbalanced and difficult-to-manage environment. Technology refresh cycles require substantial planning in advance, costly outages, and configuration changes, all of which introduce risk into the system.

With scale-out, as the storage environment grows, additional controllers are added seamlessly to the resource pool residing on a shared storage infrastructure. Scale-out, together with built-in storage virtualization, provides non-disruptive movement of host and client connections, as well as the datastores themselves, anywhere in the resource pool. With these capabilities, new workloads can be easily deployed and existing workloads can be easily and non-disruptively balanced over the available resources. Technology refreshes (replacing disk shelves, adding or completely replacing storage controllers) are accomplished while the environment remains online and serving data.

Figure 6-4 provides a graphical explanation.

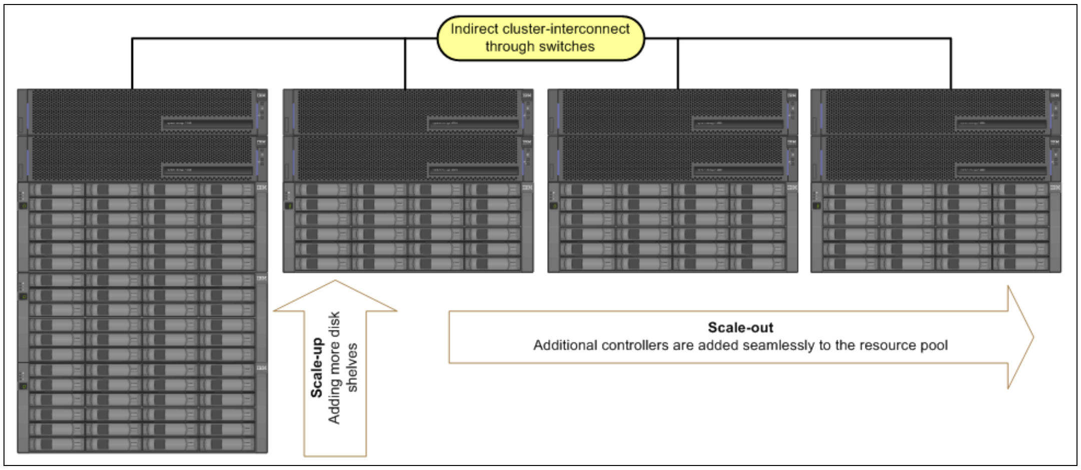

Indirect cluster-interconnect through switches
Scale-up
Adding more disk shelves
Scale-out
Additional controllers are added seamlessly to the resource pool

Figure 6-4 Scale-out storage versus scale-up storage

Although scale-out architecture has been available for some time, it is not in itself an automatic panacea for all of an enterprise’s storage requirements. Many existing scale-out products are characterized by one or more of the following shortcomings:

- Limited protocol support; for example, NAS only
- Limited hardware support: support for only a particular type or a very limited set of storage controllers
- Upgrades dictate scaling in all dimensions based on the available controller configurations, so that capacity, compute power, and I/O all need to be increased, even if only a subset of these is required
- Little or no storage efficiency, such as thin provisioning, deduplication, or compression
- Limited or no data replication capabilities
- Limited flash support

Therefore, although these products may be well positioned for certain specialized workloads, they are not flexible, capable, or robust enough for broad deployment throughout the enterprise.

6.4 Non-disruptive operations

Shared storage infrastructures in today’s 24/7 environments provide services to thousands of individual clients or hosts and support many diverse applications and workloads across multiple business units or tenants. In such environments, downtime is not an option; storage infrastructures must be always on.

Non-disruptive operations (NDO) in clustered ONTAP are intrinsic to its innovative scale-out architecture. NDO is the ability of the storage infrastructure to remain up and serving data through the execution of hardware and software maintenance operations, as well as during other IT lifecycle operations. The goal of NDO is to eliminate downtime, whether preventable, planned or unplanned, and to allow changes to the system to occur at any time.

Clustered ONTAP is highly available by design and can transparently migrate data as well as logical client connections throughout the storage cluster. DataMotion for Volumes is standard and built into clustered ONTAP. It is the ability to non-disruptively move individual data volumes, allowing data to be redistributed across a cluster at any time and for any reason. DataMotion for Volumes is transparent and non-disruptive to NAS and SAN hosts so that the storage infrastructure continues to serve data throughout these changes. Data migration might be performed to rebalance capacity usage, to optimize for changing performance requirements, or to isolate one or more controllers or storage components to execute maintenance or lifecycle operations.

Table 6-2 describes hardware and software maintenance operations that can be performed non-disruptively in a Clustered Data ONTAP environment.

Table 6-2 Non-disruptive hardware and software maintenance operations

Operation	Details
Upgrade software	Upgrade from one version of Data ONTAP to another
Upgrade firmware	System, disk, switch firmware upgrade
Replace a failed controller or a component within a controller	For example, NICs, HBAs, power supplies, and so on
Replace failed storage components	For example, cables, drives, I/O modules, and so on

Table 6-3 describes lifecycle operations that can be performed non-disruptively in a Clustered Data ONTAP environment.

Table 6-3 Non-disruptive lifecycle operations

Operation	Details
Scale storage	Add storage (shelves or controllers) to a cluster and redistribute volumes for future growth
Scale hardware	Add hardware to controllers to increase scalability, performance, or capability (HBAs, NICs, Flash Cache, Flash Pool)
Refresh technology	Upgrade storage shelves, storage controllers, cluster-interconnect switch
Rebalance controller performance and storage utilization	Redistribute data across controllers to improve performance

Part 2

Features

This part of the book discusses the basic features of Clustered Data ONTAP. The features are explained in theory and some examples are provided.

The following topics are covered:

- Physical storage
- Logical storage
- Networking
- NAS protocols
- SAN protocols
- Ancillary protocols
- Storage efficiency
- Data protection
- Disaster recovery
- Performance considerations

7

Physical storage

In this chapter, we talk about the physical storage on which the IBM N series is based. The physical layer is shared by the storage virtual machines (SVMs) to provide the storage to the clients.

The following topics are covered:

- Managing disks
- RAID protection
- Aggregates
- Storage limits

7.1 Managing disks

Disks provide the basic unit of storage for storage systems running Data ONTAP that use native disk shelves. Understanding how Data ONTAP uses and classifies disks will help you manage your storage more effectively.

7.1.1 How Data ONTAP reports drive types

Data ONTAP associates a type with every drive. Data ONTAP reports some drive types differently than the industry standards; you should understand how Data ONTAP drive types map to industry standards to avoid confusion.

When Data ONTAP documentation refers to a drive type, it is the type used by Data ONTAP unless otherwise specified. RAID drive types denote the role a specific drive plays for RAID. RAID drive types are not related to Data ONTAP drive types.

For a specific configuration, the drive types supported depend on the storage system model, the shelf type, and the I/O modules installed in the system.

Table 7-1 shows how Data ONTAP drive types map to industry standard drive types for the SAS and FC-AL storage connection architectures and for storage arrays.

Table 7-1 SAS storage connection architecture

Data ONTAP drive type	Primary drive characteristic	Industry standard drive type	Description
BSAS	Capacity	SATA	Bridged SAS-SATA disks with added hardware to enable them to be plugged into a SAS shelf.
FSAS	Capacity	NL-SAS	Near Line SAS
MSATA	Capacity	SATA	SATA disk in multi-disk carrier disk shelf
SAS	Performance	SAS	
SATA	Capacity	SATA	Available only as internal disks for N3xxx systems.
SSD	High-performance	SSD	Solid-state drives
ATA	Capacity	SATA	
FCAL	Performance	FC	

7.1.2 Storage connection architectures and topologies

Data ONTAP supports two storage connection architectures, serial-attached SCSI (SAS) and Fibre Channel (FC). The FC connection architecture supports three topologies, arbitrated loop, switched, and point-to-point:

- SAS, SATA, BSAS, FSAS, SSD, and MSATA disks use the SAS connection architecture.

- FC and ATA disks use the FC connection architecture with an arbitrated-loop topology (FC-AL).
- Array LUNs use the FC connection architecture, with either the point-to-point or switched topology.

SAS-connected disk shelves are connected to the controller on a daisy chain called a stack. FC- connected disk shelves are connected to the controller on a loop. You cannot combine different connection architectures in the same loop or stack.

Combining disks for the SAS disk connection type

You can combine SAS disk shelves and SATA disk shelves within the same stack, although this configuration is not advised.

Each SAS-connected disk shelf can contain only one type of disk (SAS or SATA). The only exception to this rule is if the shelf is being used for a Flash Pool. In that case, for some SSD sizes and shelf models, you can combine SSDs and HDDs in the same shelf.

Combining disks for the FC-AL disk connection type

You cannot combine disk shelves containing FC disks and disk shelves containing ATA disks in the same loop.

7.1.3 Usable and physical disk capacity by disk size

You cannot use the nominal size of a disk in your aggregate and storage system capacity calculations. You must use either the usable capacity or the physical capacity as calculated by Data ONTAP.

Table 7-2 lists the approximate physical and usable capacities for the disk sizes currently supported by Data ONTAP. The numbers shown are in Mebibytes (MiBs). This unit of measure is equivalent to 2 to the 20th power bytes. (MBs, in contrast, are 10 to the sixth power bytes.)

The physical capacities listed in Table 7-2 are approximations; actual physical disk capacities vary by disk manufacturer. The technical documentation for your disks contains the exact physical disk capacities.

Table 7-2 Usable capacity by disk type

Disk size as described by manufacturer	Physical capacity (MiBs, approximate)	Usable capacity (MiBs)
100 GB SSD (X441A-R5	95,396	95,146
100 GB SSD (X442A-R5)	84,796	84,574
200 GB SSD	190,782	190,532
800 GB SSD	763,097	762,847
300GB SAS/FC	280,104	272,00
450 GB SAS/FC	420,156	418,000
500 GB SATA	423,946	423,111
600 GB SAS/FC	560,208	560,000
900 GB SAS	858,483	857,000

Disk size as described by manufacturer	Physical capacity (MiBs, approximate)	Usable capacity (MiBs)
1.2 TB SAS	1,144,641	1,142,352
1 TB SATA	847,884	847,555
2 TB SATA	1,695,702	1,695,466
3 TB SATA	2,543,634	2,538,546
4 TB NL-SAS	3,815,447	3,807,816

7.1.4 Methods of calculating aggregate and system capacity

You use the physical and usable capacity of the disks you employ in your storage systems to ensure that your storage architecture conforms to the overall system capacity limits and the size limits of your aggregates.

To maintain compatibility across different brands of disks, Data ONTAP rounds down (right-sizes) the amount of space available for user data. In addition, the numerical base used to calculate capacity (base 2 or base 10) also impacts sizing information. For these reasons, it is important to use the correct size measurement, depending on the task you want to accomplish:

- For calculating overall system capacity, you use the physical capacity of the disk, and count every disk that is owned by the storage system.
- For calculating how many disks you can put into an aggregate before you exceed its maximum size, you use the right-sized, or usable capacity of all data disks in that aggregate.

Parity and double parity disks are not counted against the maximum aggregate size.

7.1.5 Disk speeds supported by Data ONTAP

For hard disk drives, which use rotating media, speed is measured in revolutions per minute (rpm). Faster disks provide more disk input/output operations per second (IOPS) and faster response time.

It is best to use disks of the same speed in an aggregate.

Data ONTAP supports the following rotational speeds for disks:

- SAS disks (SAS-connected)
  - 10,000 rpm
  - 15,000 rpm
- SATA, BSAS, FSAS, and MSATA disks (SAS-connected)
  - 7,200 rpm
- FCAL disks (FC-AL connected)
  - 10,000 rpm
  - 15,000 rpm
- ATA disks (FC-AL connected)
  - 5,400 rpm
  - 7,200 rpm

Solid-state disks, or SSDs, are flash memory-based devices and therefore the concept of rotational speed does not apply to them. SSDs provide more IOPS and faster response times than rotating media.

7.1.6 Checksum types and how they affect aggregate and spare management

There are two checksum types available for disks used by Data ONTAP, BCS (block) and AZCS (zoned). Understanding how the checksum types differ and how they impact storage management enables you to manage your storage more effectively.

Both checksum types provide the same resiliency capabilities; BCS optimizes data access speed and capacity for disks that use 520 byte sectors. AZCS provides enhanced storage utilization and capacity for disks that use 512 byte sectors (usually SATA disks, which emphasize capacity).

Aggregates have a checksum type, which is determined by the checksum type of the disks that compose the aggregate. The following configuration rules apply to aggregates, disks, and checksums:

- Checksum types cannot be combined within RAID groups. This means that you must consider checksum type when you provide hot spare disks.
- When you add storage to an aggregate, if it has a different checksum type than the storage in the RAID group to which it would normally be added, Data ONTAP creates a new RAID group.
- An aggregate can have RAID groups of both checksum types. These aggregates have a checksum type of mixed.
- Disks of a different checksum type cannot be used to replace a failed disk.
- You cannot change the checksum type of a disk.

You should know the Data ONTAP disk type and checksum type of all of the disks you manage, because these disk characteristics impact where and when the disks can be used.

Table 7-3 shows the checksum type by Data ONTAP disk type.

Table 7-3 Checksums by disk type

Data ONTAP disk type	Checksum type
SAS or FC-AL	BCS
SATA/BSAS/FSAS/ATA	BCS
SSD	BCS
MSATA	AZCS

7.1.7 Drive name formats

Each drive has a name that differentiates it from all other drives. Drive names have different formats, depending on the connection type (FC-AL or SAS) and how the drive is attached.

Each drive has a universal unique identifier (UUID) that differentiates it from all other drives in the cluster.

The names of unowned drives (broken or unassigned drives) display the alphabetically lowest node name in the cluster that can see that drive.

Table 7-4 shows the various formats for drive names, depending on how they are connected to the storage system.

Note: For internal drives, the slot number is zero, and the internal port number depends on the system model.

Table 7-4 Drive name formats

Drive connection	Drive name	Example
SAS, direct-attached	<node>:<slot><port>.<shelfID>.<bay>	The drive in shelf 2, bay 11, connected to onboard port 0a and owned by node1 is named node1:0a.2.11. The drive in shelf 6, bay 3 connected to an HBA in slot 1, port c, and owned by node 1 is named node1:1c.6.3.
SAS, direct-attached in multi-disk carrier disk shelf	<node>:<slot><port>.<shelfID>.<bay>L<carrierPosition>	Carrier position is 1 or 2.
FC-AL, direct-attached	<node>:<slot><port>.<loopID>	The drive with loop ID 19 (bay 3 of shelf 1) connected to onboard port 0a and owned by node1 is named node1:0a.19. The drive with loop ID 34 connected to an HBA in slot 8, port c and owned by node1 is named node1:8c.34.
FC-AL, switch-attached	<node>:<switch_name>.<switch_port>.<loopID>	The drive with loop ID 51 connected to port 3 of switch SW7 owned by node1 is named node1:SW7.3.51.

7.1.8 Loop IDs for FC-AL connected disks

For disks connected using Fibre Channel-Arbitrated Loop (FC-AL or FC), the loop ID is an integer between 16 and 126. The loop ID identifies the disk within its loop, and is included in the disk name, which identifies the disk uniquely for the entire system.

The loop ID corresponds to the disk shelf number and the bay in which the disk is installed. The lowest loop ID is always in the far right bay of the first disk shelf. The next higher loop ID is in the next bay to the left, and so on. You can view the device map for your disk shelves with the **`fc admin device_map`** command.

7.2 RAID protection

Understanding how RAID protects your data and data availability can help you administer your storage systems more effectively.

For native storage, Data ONTAP uses RAID-DP (double-parity) or RAID Level 4 (RAID4) protection to ensure data integrity within a RAID group even if one or two of those drives fail. Parity drives provide redundancy for the data stored in the data drives. If a drive fails (or, for RAID-DP, up to two drives), the RAID subsystem can use the parity drives to reconstruct the data in the drive that failed.

7.2.1 RAID protection levels for disks

Data ONTAP supports two levels of RAID protection for aggregates composed of disks in native disk shelves: RAID-DP and RAID4. RAID-DP is the default RAID level for new aggregates.

Understanding RAID disk types

Data ONTAP classifies disks as one of four types for RAID: data, hot spare, parity, or double parity. The RAID disk type is determined by how RAID is using a disk; it is different from the Data ONTAP disk type.

Data disk

Holds data stored on behalf of clients within RAID groups (and any data generated about the state of the storage system as a result of a malfunction).

Spare disk

Does not hold usable data, but is available to be added to a RAID group in an aggregate. Any functioning disk that is not assigned to an aggregate but is assigned to a system functions as a hot spare disk.

Parity disk

Stores row parity information that is used for data reconstruction when a single disk drive fails within the RAID group.

dParity disk

Stores diagonal parity information that is used for data reconstruction when two disk drives fail within the RAID group, if RAID-DP is enabled.

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 failure 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 (dParity) disk. However, for non-root aggregates with only one RAID group, you must have at least five disks (three data disks and two parity disks).

If there is a data-disk failure 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.

What RAID4 protection is

RAID 4 provides single-parity disk protection against single-disk failure within a RAID group. If an aggregate is configured for RAID4 protection, Data ONTAP reconstructs the data from a single failed disk within a RAID group and transfers that reconstructed data to a spare disk.

The minimum number of disks in a RAID 4 group is two: at least one data disk and one parity disk. However, for non-root aggregates with only one RAID group, you must have at least three disks (two data disks and one parity disk).

If there is a single data or parity disk failure in a RAID 4 group, Data ONTAP replaces the failed disk in the RAID group with a spare disk and uses the parity data to reconstruct the failed disk's data on the replacement disk. If no spare disks are available, Data ONTAP goes into degraded mode and alerts you of this condition.

Attention: With RAID4, if there is a second disk failure before data can be reconstructed from the data on the first failed disk, there will be data loss. To avoid data loss when two disks fail, you can select RAID-DP. This provides two parity disks to protect you from data loss when two disk failures occur in the same RAID group before the first failed disk can be reconstructed.

7.2.2 Data ONTAP RAID groups

A RAID group consists of one or more data disks or array LUNs, across which client data is striped and stored, and up to two parity disks, depending on the RAID level of the aggregate that contains the RAID group.

RAID-DP uses two parity disks to ensure data recoverability even if two disks within the RAID group fail.

RAID 4 uses one parity disk to ensure data recoverability if one disk within the RAID group fails.

RAID group names

Within each aggregate, RAID groups are named rg0, rg1, rg2, and so on in order of their creation. You cannot specify the names of RAID groups.

RAID group sizes

A RAID group has a maximum number of disks or array LUNs that it can contain. This is called its maximum size, or its size. A RAID group can be left partially full, with fewer than its maximum number of disks or array LUNs, but storage system performance is optimized when all RAID groups are full.

Configuring an optimum RAID group size for an aggregate made up of drives requires a trade-off of factors. You must decide which factor—speed of recovery, assurance against data loss, or maximizing data storage space—is most important for the aggregate that you are configuring.

You change the size of RAID groups on a per-aggregate basis. You cannot change the size of an individual RAID group.

HDD RAID groups

Follow these guidelines when sizing your RAID groups composed of HDDs:

- All RAID groups in an aggregate should have the same number of disks. If this is impossible, any RAID group with fewer disks should have only one less disk than the largest RAID group.
- The advised range of RAID group size is between 12 and 20. The reliability of SAS and FC disks can support a RAID group size of up to 28, if needed.
- If you can satisfy the first two guidelines with multiple RAID group sizes, you should choose the larger size.

SSD RAID groups in Flash Pools

The SSD RAID group size can be different from the RAID group size for the HDD RAID groups in a Flash Pool. Usually, you should ensure that you have only one SSD RAID group for a Flash Pool, to minimize the number of SSDs required for parity.

SSD RAID groups in SSD-only aggregates

Follow these guidelines when sizing your RAID groups composed of SSDs:

- All RAID groups in an aggregate should have the same number of drives. If this is impossible, any RAID group with fewer drives should have only one less drive than the largest RAID group.
- The preferred range of RAID group size is between 20 and 28.

7.2.3 Data ONTAP hot spare disks

A hot spare disk is a disk that is assigned to a storage system but is not in use by a RAID group. It does not yet hold data but is ready for use. If a disk failure occurs within a RAID group, Data ONTAP automatically assigns hot spare disks to RAID groups to replace the failed disks.

How many hot spares you should have

Having insufficient spares increases the risk of a disk failure with no available spare, resulting in a degraded RAID group. The number of hot spares you should have depends on the Data ONTAP disk type.

MSATA disks, or disks in a multi-disk carrier, should have four hot spares during steady state operation, and you should never allow the number of MSATA hot spares to dip below two.

For RAID groups composed of SSDs, you should have at least one spare disk.

For all other Data ONTAP disk types, you should have at least one matching or appropriate hot spare available for each kind of disk installed in your storage system. However, having two available hot spares for all disks provides the best protection against disk failure. Having at least two available hot spares provides the following benefits:

- When you have two or more hot spares for a data disk, Data ONTAP can put that disk into the maintenance center if needed. Data ONTAP uses the maintenance center to test suspect disks and take offline any disk that shows problems.
- Having two hot spares means that when a disk fails, you still have a spare available if another disk fails before you replace the first failed disk.

A single spare disk can serve as a hot spare for multiple RAID groups.

What disks can be used as hot spares

A disk must conform to certain criteria to be used as a hot spare for a particular data disk.

For a disk to be used as a hot spare for another disk, it must conform to the following criteria:

- It must be either an exact match for the disk it is replacing or an appropriate alternative.
- The spare must be owned by the same system as the disk it is replacing.

What a matching spare is

A matching hot spare exactly matches several characteristics of a designated data disk. Understanding what a matching spare is, and how Data ONTAP selects spares, enables you to optimize your spares allocation for your environment.

A matching spare is a disk that exactly matches a data disk for all of the following criteria:

- Effective Data ONTAP disk type:

  The effective disk type can be affected by the value of the raid.mix.hdd.performance and raid.mix.hdd.capacity options, which determine the disk types that are considered to be equivalent.

- Size
- Speed (rpm)
- Checksum type (BCS or AZCS)

What an appropriate hot spare is

If a disk fails and no hot spare disk that exactly matches the failed disk is available, Data ONTAP uses the best available spare. Understanding how Data ONTAP chooses an appropriate spare when there is no matching spare enables you to optimize your spare allocation for your environment.

Data ONTAP picks a non-matching hot spare based on the following criteria:

- If the available hot spares are not the correct size, Data ONTAP uses one that is the next size up, if there is one. The replacement disk is downsized to match the size of the disk it is replacing; the extra capacity is not available.
- If the available hot spares are not the correct speed, Data ONTAP uses one that is a different speed. Using drives with different speeds within the same aggregate is not optimal. Replacing a disk with a slower disk can cause performance degradation, and replacing a disk with a faster disk is not cost-effective.

If no spare exists with an equivalent disk type or checksum type, the RAID group that contains the failed disk goes into degraded mode; Data ONTAP does not combine effective disk types or checksum types within a RAID group.

7.3 Aggregates

To support the differing security, backup, performance, and data sharing needs of your users, you can group the physical data storage resources on your storage system into one or more aggregates. You can then design and configure these aggregates to provide the appropriate level of performance and redundancy.

Each aggregate has its own RAID configuration, plex structure, and set of assigned disks or array LUNs. The aggregate provides storage, based on its configuration, to its associated FlexVol volumes or Infinite Volume.

Aggregates have the following characteristics:

- They are composed of disks.
- They can be in 64-bit or 32-bit format.
- They can be single-tier (composed of only HDDs or only SSDs) or they can be Flash Pools, which include both of those storage types in two separate tiers.

The cluster administrator can assign one or more aggregates to an SVM, in which case you can use only those aggregates to contain volumes for that SVM.

7.3.1 FlexVol and SVM associations

FlexVol volumes are always associated with one SVM, and one aggregate that supplies its storage. The SVM can limit which aggregates can be associated with that volume, depending on how the SVM is configured.

When you create a FlexVol volume, you specify which SVM the volume will be created on, and which aggregate that volume will get its storage from. All of the storage for the newly created FlexVol volume comes from that associated aggregate.

If the SVM for that volume has aggregates assigned to it, then you can use only one of those assigned aggregates to provide storage to volumes on that SVM. This can help you ensure that your SVMs are not sharing physical storage resources inappropriately. This segregation can be important in a multi-tenancy environment, because for some space management configurations, volumes that share the same aggregate can affect each other's access to free space when space is constrained for the aggregate. Aggregate assignment requirements apply to both cluster administrators and SVM administrators.

Volume move and volume copy operations are not constrained by the SVM aggregate assignments, so if you are trying to keep your SVMs on separate aggregates, you must ensure that you do not violate your SVM aggregate assignments when you perform those operations.

If the SVM for that volume has no aggregates assigned to it, then a cluster administrator can use any aggregate in the cluster to provide storage to the new volume. However, an SVM administrator cannot create volumes for an SVM with no assigned aggregates. For this reason, if you want an SVM administrator to be able to create volumes for a specific SVM, then you must assign aggregates to that SVM **(`vserver modify -aggr-list`)**.

Changing the aggregates assigned to an SVM does not affect any existing volumes. For this reason, the list of aggregates assigned to an SVM cannot be used to determine the aggregates associated with volumes for that SVM.

7.3.2 How aggregates work

Aggregates have a single copy of their data, or plex, which contains all of the RAID groups belonging to that aggregate. Mirrored aggregates are not currently supported in Data ONTAP Cluster-Mode.

The diagram in Figure 7-1 shows a non-mirrored aggregate with disks, with its one plex.

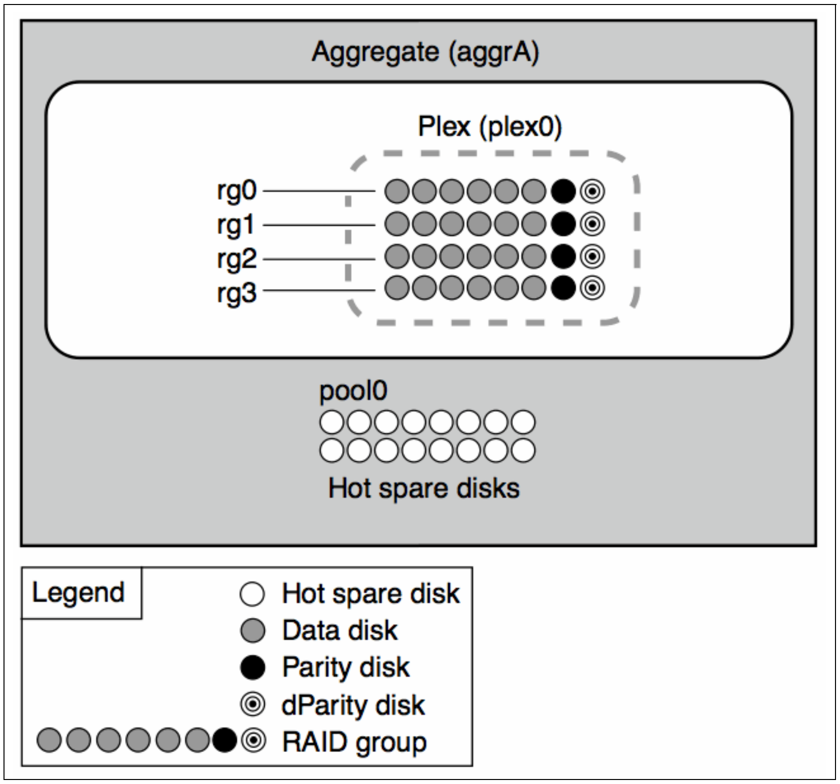


Figure 7-1 None mirrored disk aggregate with one plex

Aggregates are either 64-bit or 32-bit format. 64-bit aggregates have much larger size limits than 32- bit aggregates. 64-bit and 32-bit aggregates can coexist on the same storage system or cluster.

32-bit aggregates have a maximum size of 16 TB; 64-bit aggregates' maximum size depends on the storage system model.

When you create a new aggregate, it is a 64-bit format aggregate.

You can expand 32-bit aggregates to 64-bit aggregates by increasing their size beyond 16 TB. 64-bit aggregates, including aggregates that were previously expanded, cannot be converted to 32-bit aggregates.

You can see whether an aggregate is a 32-bit aggregate or a 64-bit aggregate by using the **`storage aggregate show -fields block-type`** command.

7.3.3 How Flash Pools work

The Flash Pool technology enables you to add one or more RAID groups composed of SSDs to an aggregate that consists of RAID groups of HDDs.

The SSDs provide a high-performance cache for the active data set of the data volumes provisioned on the Flash Pool, which off loads I/O operations from the HDDs to the SSDs. For random workloads, this can increase the performance of the volumes associated with the aggregate by improving the response time and overall throughput for I/O-bound data access operations. (The performance increase is not seen for predominantly sequential workloads.)

The SSD cache does not contribute to the size of the aggregate as calculated against the maximum aggregate size. For example, even if an aggregate is at the maximum aggregate size, you can add an SSD RAID group to it. The SSDs do count toward the overall spindle limit.

The HDD RAID groups in a Flash Pool behave the same as HDD RAID groups in a standard aggregate, following the same rules for mixing disk types, sizes, speeds, and checksums.

The checksum type, RAID type, and RAID group size values can be configured for the SSD cache RAID groups and HDD RAID groups independently.

There is a platform-dependent maximum size for the SSD cache.

Requirements for using Flash Pools

The Flash Pool technology has some configuration requirements that you should be aware of before planning to use it in your storage architecture.

Flash Pools cannot be used in the following configurations:

- 32-bit aggregates
- Aggregates composed of array LUNs
- Aggregates that use the ZCS checksum type

Read-only volumes, such as SnapMirror destinations, are not cached in the Flash Pool cache.

If you create a Flash Pool using an aggregate that was created using Data ONTAP 7.1 or earlier, the volumes associated with that Flash Pool will not support write caching.

How Flash Pool and Flash Cache compare

Both the Flash Pool technology and the family of Flash Cache modules (Flash Cache and Flash Cache 2) provide a high-performance cache to increase storage performance. However, there are differences between the two technologies that you should understand before choosing between them, as shown in Table 7-5.

You can employ both technologies on the same system. However, data stored in volumes associated with a Flash Pool (or an SSD aggregate) is not cached by Flash Cache.

Table 7-5 Differences between Flash Pool and Flash Cache

Criteria	Flash Pool	Flash Cache
Scope	A specific aggregate	All aggregates assigned to a node
Caching types supported	Read and write	Read
Cached data availability during and after takeover events	Cached data is available and unaffected by either planned or unplanned takeover events	Cached data is not available during takeover events. After giveback for a planned takeover, previously cached data that is still valid is recached automatically.
PCIe slot on storage controller required?	No	Yes

About read and write caching for Flash Pools

The Flash Pool technology provides both read caching and write caching for random I/O workloads. You can configure Flash Pool caching on the volume, but for most workloads, the default caching policies result in optimal performance.

Some volumes cannot be enabled for write caching. When you attempt to use an aggregate associated with one or more of these volumes as a Flash Pool, you must force the operation. In this case, writes to that volume would not be cached in the SSD cache, but otherwise the Flash Pool would function normally. You can get more information about why a volume cannot be enabled for write caching by using the **`volume show -instance`** command.

How Flash Pool cache capacity is calculated

Flash Pool cache capacity cannot exceed a platform-dependent limit for the system. Knowing the available cache capacity enables you to determine how many data SSDs you can add before reaching the limit. See Example 7-1 and Example 7-2.

The current cache capacity is the sum of the “used size” capacity of all of the data SSDs used in Flash Pools on the system. Parity SSDs do not count toward the limit.

For systems in an HA configuration, the cache size limits apply to the HA configuration as a whole, and can be split arbitrarily between the two nodes, provided that the total limit for the HA configuration is not exceeded.

If Flash Cache modules are installed in a system, the available cache capacity for Flash Pool use is the Flash Pool cache capacity limit minus the sum of the Flash Cache module cache installed on the node. (In the unusual case where the size of the Flash Cache modules is not symmetrical between the two nodes in an HA configuration, the Flash Pool available cache capacity is decreased by the size of the larger Flash Cache module.)

Example 7-1 Cache size calculation with Flash Cache modules

```
For an HA configuration composed of two storage controllers with a maximum cache
capacity of 12 TB and 2 TB of Flash Cache installed on each node, the maximum
Flash Pool cache capacity for the HA pair would be 12 TB minus 2 TB, or 10 TB.
```

Example 7-2 Cache size calculation with asymmetrically sized Flash Cache modules

```
For an HA configuration composed of two storage controllers with a maximum cache
capacity of 12 TB and 2 TB of Flash Cache installed on one node and 3 TB of Flash
Cache installed on the other node, the maximum Flash Pool cache capacity for the
HA pair would be 12 TB minus 3 TB, or 9 TB.
```

Rules for mixing HDD types in aggregates

You can mix disks from different loops or stacks within the same aggregate. Depending on the value of the **`raid.mix.hdd.disktype`** RAID options, you can mix certain types of HDDs within the same aggregate, but some disk type combinations are more desirable than others.

When the appropriate **`raid.mix.hdd.disktype`** option is set to **`off`**, single-tier aggregates and the HDD tier of Flash Pools can be composed of only one Data ONTAP disk type. This setting ensures that your aggregates are homogeneous, and requires that you provide sufficient spare disks for every disk type in use in your system.

The default value for the **`raid.mix.hdd.disktype.performance`** option is **`off`**, to prevent mixing SAS and FC-AL disks.

The default value for the **raid.mix.hdd.disktype.capacity** option is **on**. For this setting, the SATA, BSAS, FSAS, and ATA disk types are considered to be equivalent for the purposes of creating and adding to aggregates, and spare management.

To maximize aggregate performance and for easier storage administration, you should avoid mixing FC-AL-connected and SAS-connected disks in the same aggregate. This is because of the performance mismatch between FC-AL-connected disk shelves and SAS-connected disk shelves.

When you mix these connection architectures in the same aggregate, the performance of the aggregate is limited by the presence of the FC-AL-connected disk shelves, even though some of the data is being served from the higher-performing SAS-connected disk shelves.

MSATA disks cannot be mixed with any other disk type in the same aggregate.

> **Note:** If you set a **raid.mix.hdd.disktype** option to **off** for a system that already contains aggregates with more than one type of HDD, those aggregates continue to function normally and accept both types of HDDs. However, no other aggregates composed of the specified disk type will accept mixed HDD types asS long as that option is set to **off**.

Rules for mixing drive types in Flash Pools

By definition, Flash Pools contain more than one drive type. However, the HDD tier follows the same disk-type mixing rules as single-tier aggregates. For example, you cannot mix SAS and SATA disks in the same Flash Pool. The SSD cache can contain only SSDs.

7.3.4 Determining space usage in an aggregate

You can view space usage by all volumes in one or more aggregates with the **storage aggregate show-space** command. This helps you see which volumes are consuming the most space in their containing aggregates so that you can take actions to free more space.

The used space in an aggregate is directly affected by the space used in the FlexVol volumes and Infinite Volume constituents it contains. Measures that you take to increase space in a volume also affect space in the aggregate.

When the aggregate is offline no values are displayed. Only non-zero values are displayed in the command output. However, you can use the **-instance** parameter to display all possible feature rows regardless of whether they are enabled and using any space. A value of - indicates that there is no data available to display.

The following rows are included in the **aggregate show-space command** output:

- Volume Footprints:

  The total of all volume footprints within the aggregate. It includes all of the space that is used or reserved by all data and metadata of all volumes in the containing aggregate. It is also the amount of space that is freed if all volumes in the containing aggregate are destroyed. Infinite Volume constituents appear in the output of space usage commands as if the constituents were FlexVol volumes.

- Aggregate Metadata:

  The total file system metadata required by the aggregate, such as allocation bitmaps and inode files.

- Snapshot Reserve:

  The amount of space reserved for aggregate Snapshot copies, based on volume size. It is considered used space and is not available to volume or aggregate data or metadata. The aggregate's Snapshot reserve is set to 0% by default.
- Total Used:

  The sum of all space used or reserved in the aggregate by volumes, metadata, or Snapshot copies.

There is never a row for Snapshot spill.

Example 7-3 shows the **`aggregate show-space command`** output for an aggregate whose Snapshot reserve was increased to 5%. If the Snapshot Reserve was 0, the row would not be displayed, as you can see in this example.

Example 7-3 Storage aggregate show-space command for aggregates with and without snapshot reserve.

```
cdot-cluster01::> storage aggregate show-space

      Aggregate : cdot_cluster01_01_sas450_01

      Feature                                Used      Used%
      --------------------------------   ----------   ------
      Volume Footprints                   509.0GB        7%
      Aggregate Metadata                   4.24MB        0%
      Snapshot Reserve                    404.1GB        5%

      Total Used                          913.1GB       12%


      Aggregate : cdot_cluster01_02_sas450_01

      Feature                                Used      Used%
      --------------------------------   ----------   ------
      Volume Footprints                    2.02GB        0%
      Aggregate Metadata                   2.03MB        0%

      Total Used                           2.02GB        0%


2 entries were displayed.
```

7.3.5 Determining and controlling a volume's space usage in the aggregate

You can determine which FlexVol volumes and Infinite Volume constituents are using the most space in the aggregate and specifically which features within the volume. The **`volume show-footprint`** command provides information about a volume's footprint, or its space usage within the containing aggregate.

The **`volume show-footprint`** command shows details about the space usage of each volume in an aggregate, including offline volumes. This command does not directly correspond to the output of the **`df`** command, but instead bridges the gap between the output of **`volume show-space`** and **`aggregate show-space`** commands. All percentages are calculated as a percent of aggregate size.

Only non-zero values are displayed in the command output. However, you can use the **-instance** parameter to display all possible feature rows regardless of whether they are enabled and using any space. A value of **-** indicates that there is no data available to display.

Infinite Volume constituents appear in the output of space usage commands as if the constituents were FlexVol volumes.

Example 7-4 shows the **volume show-footprint** command output for a volume.

Example 7-4 The show-footprint command example

```
cdot-cluster01::> volume show-footprint vs_01_cifsvol_01

      Vserver : vs_cifs_01
      Volume  : vs_01_cifsvol_01

      Feature                                   Used    Used%
      --------------------------------    ----------    -----
      Volume Data Footprint                   7.52MB       0%
      Volume Guarantee                            0B       0%
      Flexible Volume Metadata               100.4MB       0%
      Delayed Frees                          442.6MB       0%

      Total Footprint                        550.6MB       0%
```

Table 7-6 explains some of the key rows of the output of the **volume show-footprint** command and what you can do to try to decrease space usage by that feature.

Table 7-6 Key rows of output of the volume show-footprint command

Row/feature name	Description/contents of row	Some ways to decrease
Volume Data Footprint	The total amount of space used in the containing aggregate by a volume's data in the active file system and the space used by the volume's Snapshot copies. This row does not include reserved space, so if volumes have reserved files, the volume's total used space in the volume show-space command output can exceed the value in this row.	► Deleting data from the volume. ► Deleting Snapshot copies from the volume.
Volume Guarantee	The amount of space reserved by the volume in the aggregate for future writes. The amount of space reserved depends on the guarantee type of the volume.	Changing the type of guarantee for the volume to none. This row will go to 0. If you configure your volumes with a volume guarantee of none, you should refer to Technical Report 3965 or 3483 for information about how doing so can affect storage availability.
Flexible Volume Metadata	The total amount of space used in the aggregate by the volume's metadata files.	No direct method to control.

Row/feature name	Description/contents of row	Some ways to decrease
Delayed Frees	Blocks that Data ONTAP used for performance and cannot be immediately freed. When Data ONTAP frees blocks in a FlexVol volume, this space is not always immediately shown as free in the aggregate because operations to free the space in the aggregate run in a batch for increased performance. Blocks that are declared free in the FlexVol volume but that are not yet free in the aggregate are called "delayed free blocks" until the associated delayed free blocks are processed. For SnapMirror destinations, this row has a value of 0 and is not displayed.	No direct method to control
Total Footprint	The total amount of space that the volume uses in the aggregate. It is the sum of all of the rows.	Any of the methods used to decrease space used by a volume

7.4 Storage limits

There are limits for storage objects that you should consider when planning and managing your storage architecture.

The limits are listed in the following tables:

- Aggregate limits in Table 7-7
- RAID group limits in Table 7-8
- RAID group sizes in Table 7-9

Table 7-7 Aggregate limits

Limit	Native storage	Notes
Aggregates Maximum per node	100	In an HA configuration, this limit applies to each node individually, so the overall limit for the pair is doubled
Aggregates (32-bit) Maximum size	16TB	
Aggregates (64-bit)	Model-dependent	See the N Series Hardware guide

Limit	Native storage	Notes
Aggregates Minimum size	RAID-DP: 5 disks RAID 4: 3 disks	For root aggregates, the minimum size is 3 disks for RAID-DP and 2 disks for RAID4. If you need a smaller non- root aggregate, you can use the **`force-small-aggregate`** option.
RAID groups Maximum per aggregate	150	

Table 7-8 RAID group limits

Limit	Native storage	Notes
Maximum per system	400	
Maximum per aggregate	150	

Table 7-9 RAID group sizes

RAID type	Default size	Maximum size	Minimum size
RAID-DP	SATA/BSAS/FSAS/ MSATA/ATA: 14 FC/SAS: 16 SSD: 23	SATA/BSAS/FSAS/ MSATA/ATA: 20 FC/SAS: 28 SSD: 28	3
RAID 4	SATA/BSAS/FSAS/ MSATA/ATA: 7 FC/SAS/SSD: 8	SATA/BSAS/FSAS/ MSATA/ATA: 7 FC/SAS/SSD: 14	2

8

Logical storage

Logical storage refers to the storage resources provided by Data ONTAP that are not tied to a physical resource.

Logical storage resources are associated with a storage virtual machine (SVM), and they exist independently of any specific physical storage resource such as a disk, array LUN, or aggregate. Logical storage resources include volumes of all types and qtrees, as well as the capabilities and configurations you can use with these resources, such as Snapshot copies, deduplication, compression, and quotas.

The following topics are covered:

- How volumes work
- FlexVol volumes
- Infinite Volumes
- Storage limits

8.1 How volumes work

Volumes are data containers that enable you to partition and manage your data. Understanding the types of volumes and their associated capabilities enables you to design your storage architecture for maximum storage efficiency and ease of administration.

Volumes are the highest-level logical storage object. Unlike aggregates, which are composed of physical storage resources, volumes are completely logical objects.

Data ONTAP provides two types of volumes, FlexVol volumes and Infinite Volumes. There are also volume variations, such as FlexClone volumes, FlexCache volumes, data protection mirrors, and load-sharing mirrors. Not all volume variations are supported for both types of volumes. Data ONTAP efficiency capabilities, compression and deduplication, are supported for both types of volumes.

Volumes contain file systems in a NAS environment, and LUNs in a SAN environment.

Volumes are associated with one SVM. The SVM is a virtual management entity, or server, that consolidates various cluster resources into a single manageable unit. When you create a volume, you specify the SVM it is associated with. The type of the volume (FlexVol volume or Infinite Volume) is determined by an immutable SVM attribute.

Volumes have a language. The language of the volume determines the character set Data ONTAP uses to display file names and data for that volume. The default value for the language of the volume is the language of the SVM.

Volumes depend on their associated aggregates for their physical storage; they are not directly associated with any concrete storage objects, such as disks or RAID groups. If the cluster administrator has assigned specific aggregates to an SVM, then only those aggregates can be used to provide storage to the volumes associated with that SVM. This impacts volume creation, and also copying and moving FlexVol volumes between aggregates.

8.1.1 What a FlexVol volume is

A FlexVol volume is a data container associated with an SVM with FlexVol volumes. It gets its storage from a single associated aggregate, which it might share with other FlexVol volumes or Infinite Volumes. It can be used to contain files in a NAS environment, or LUNs in a SAN environment.

FlexVol volumes enable you to partition your data into individual manageable objects that can be configured to suit the needs of the users of that data.

A FlexVol volume enables you to take the following actions:

- Create a clone of the volume quickly and without having to duplicate the entire volume by using FlexClone technology.
- Reduce the space requirements of the volume by using deduplication and compression technologies.
- Create a sparse copy of the volume to balance loads or reduce network latency by using FlexCache technology.
- Create a Snapshot copy of the volume for data protection purposes.
- Limit the amount of space a user, group, or qtree can use in the volume by using quotas.
- Partition the volume by using qtrees.

- Create load-sharing mirrors to balance loads between nodes.
- Move the volume between aggregates and between storage systems.
- Make the volume available to client access using any file access protocol supported by Data ONTAP.
- Set up a volume to make more storage available when it becomes full.
- Create a volume that is bigger than the physical storage currently available to it by using thin provisioning.

8.1.2 What an Infinite Volume is

An Infinite Volume is a single, scalable volume that can store up to 2 billion files and tens of petabytes of data.

With an Infinite Volume, you can manage multiple petabytes of data in one large logical entity and clients can retrieve multiple petabytes of data from a single junction path for the entire volume.

An Infinite Volume uses storage from multiple aggregates on multiple nodes. You can start with a small Infinite Volume and expand it non-disruptively by adding more disks to its aggregates or by providing it with more aggregates to use.

Infinite Volumes enable you to store multiple petabytes of data in a single volume that supports multi-protocol access, storage efficiency technologies, and data protection capabilities.

With Infinite Volumes, you can perform the following tasks:

- Manage multiple petabytes of data in a single logical entity with a single junction path and a single namespace.
- Provide multi-protocol access to that data using NFSv3, NFSv4.1, pNFS, and CIFS (SMB 1.0).
- Offer secure multi-tenancy by creating multiple SVMs with FlexVol volumes and multiple SVMs with Infinite Volume in a single cluster.
- Assign more storage to users than is physically available by using thin provisioning.
- Maximize storage efficiency by using deduplication and compression technologies.
- Optimize storage by grouping it into storage classes that correspond to specific goals, such as maximizing performance or maximizing capacity.
- Automatically place incoming files into the appropriate storage class according to rules based on file name, file path, or file owner.
- Protect data by creating Snapshot copies of the volume.
- Create a data protection mirror relationship between two volumes on different clusters, and restore data when necessary.
- Back up data with CIFS or NFS over a mounted volume to tape, and restore data when necessary.
- Increase the physical size of the Infinite Volume by adding more disks to the aggregates used by the Infinite Volume or by assigning more aggregates to the SVM containing the Infinite Volume and then resizing the Infinite Volume.

8.1.3 Comparison of FlexVol volumes and Infinite Volumes

Both FlexVol volumes and Infinite Volumes are data containers. However, they have significant differences that you should consider before deciding which type of volume to include in your storage architecture.

Table 8-1 summarizes the differences and similarities between FlexVol volumes and Infinite Volumes.

Table 8-1 Comparison between FlexVol volumes and Infinite Volumes

Volume capability or feature	FlexVol Volumes	Infinite Volumes	Notes
Maximum number per node	Model-dependent	Model-dependent	
Types of aggregates supported	64-bit or 32-bit	64-bit	
SAN protocols supported	Yes	No	
File access protocols supported	NFS, CIFS	NFS, CIFS	
Deduplication	Yes	Yes	
Compression	Yes	Yes	
FlexClone volumes	Yes	No	
FlexCache volumes	Yes	No	
Quotas	Yes	No	
Qtrees	Yes	No	
Thin provisioning	Yes	Yes	
Snapshot copies	Yes	Yes	
Data protection mirrors	Yes	Yes	For Infinite Volumes, only mirrors between clusters are supported
Load-sharing mirrors	Yes	No	
Tape backup	Yes	Yes	For Infinite Volumes, you must use NFS or CIFS rather than NDMP.
Volume security styles	UNIX, NTFS, mixed	Unified	

8.1.4 How FlexVol volumes and Infinite Volumes share aggregates

Aggregates can be shared among the volumes in a cluster. Each aggregate can contain multiple FlexVol volumes along side multiple constituents of Infinite Volumes.

When you create an Infinite Volume, constituents of the Infinite Volume are placed on aggregates that are assigned to its containing SVM. If the SVM with Infinite Volume includes aggregates that contain FlexVol volumes, one or more of the Infinite Volume's constituents might be placed on aggregates that already include FlexVol volumes, if those aggregates meet the requirements for hosting Infinite Volumes.

Similarly, when you create a FlexVol volume, you can associate that FlexVol volume with an aggregate that is already being used by an Infinite Volume.

Figure 8-1 illustrates aggregate sharing in a four-node cluster that includes both FlexVol volumes and an Infinite Volume. The Infinite Volume uses the aggregates aggrA, aggrB, aggrC, aggrD, aggrE, and aggrG even though the aggregates aggrB, aggrC, and aggrG already provide storage to FlexVol volumes. (For clarity, the individual constituents that make up the Infinite Volume are not shown.)

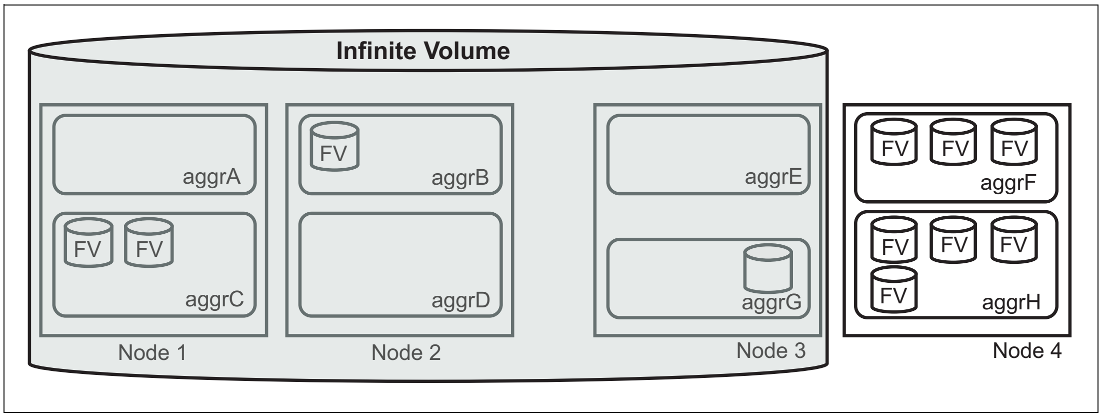


Figure 8-1 Aggregate sharing in a four-node cluster

8.1.5 System volumes

System volumes are FlexVol volumes that contain special metadata, such as metadata for file services audit logs. These volumes are visible in the cluster so that you can fully account for the storage use in your cluster.

System volumes are owned by the cluster management server (also called the admin SVM), and they are created automatically when file services auditing is enabled.

You can view system volumes by using the **`volume show`** command, but most other volume operations are not permitted. For example, you cannot modify a system volume by using the volume modify command.

8.2 FlexVol volumes

Most management tasks for FlexVol volumes are available to the SVM administrator. A few, such as promoting a volume to be the root volume of an SVM and moving or copying volumes, are available only to cluster administrators.

8.2.1 Difference between 64-bit and 32-bit FlexVol volumes

FlexVol volumes are one of two formats: 64-bit or 32-bit. A 64-bit volume has a larger maximum size than a 32-bit volume.

A newly created FlexVol volume is the same format as its associated aggregate. However, a volume can be a different format from its associated aggregate in certain cases.

The maximum size of a 64-bit volume is determined by the size of its associated aggregate, which depends on the storage system model.

A 32-bit volume has a maximum size of 16 TB.

Note: For both volume formats, the maximum size for each LUN or file is 16 TB.

Some Data ONTAP features use two FlexVol volumes; those volumes can be different formats. These features interoperate between the two volume formats.

Table 8-2 Interoperability between 64-bit and 32-bit FlexVol volumes

Data ONTAP feature	Interoperates between 64-bit and 32-bit format?
FlexCache	Y
ndmpcopy	Y
volume copy	Y
Volume SnapMirror	Y
volume move (DataMotion for Volumes)	Y

8.2.2 FlexVol volumes and SVMs

Understanding how FlexVol volumes work with SVMs enables you to plan your storage architecture.

Aggregates association with a FlexVol volume

FlexVol volumes are always associated with one SVM, and one aggregate that supplies its storage. The SVM can limit which aggregates can be associated with that volume, depending on how the SVM is configured.

When you create a FlexVol volume, you specify which SVM the volume will be created on, and which aggregate that volume will get its storage from. All of the storage for the newly created FlexVol volume comes from that associated aggregate.

If the SVM for that volume has aggregates assigned to it, then you can use only one of those assigned aggregates to provide storage to volumes on that SVM. This can help you ensure that your SVMs are not sharing physical storage resources inappropriately. This segregation can be important in a multi-tenancy environment, because for some space management configurations, volumes that share the same aggregate can affect each other's access to free space when space is constrained for the aggregate. Aggregate assignment requirements apply to both cluster administrators and SVM administrators.

Volume move and volume copy operations are not constrained by the SVM aggregate assignments, so if you are trying to keep your SVMs on separate aggregates, you must ensure that you do not violate your SVM aggregate assignments when you perform those operations.

If the SVM for that volume has no aggregates assigned to it, then a cluster administrator can use any aggregate in the cluster to provide storage to the new volume. However, an SVM administrator cannot create volumes for an SVM with no assigned aggregates. For this reason, if you want an SVM administrator to be able to create volumes for a specific SVM, then you must assign aggregates to that SVM (**vserver modify -aggr-list**).

Changing the aggregates assigned to an SVM does not affect any existing volumes. For this reason, the list of aggregates assigned to an SVM cannot be used to determine the aggregates associated with volumes for that SVM.

How to limit the number of FlexVol volumes in an SVM

You can limit the volumes for an SVM with FlexVol volumes to control resource usage or ensure that configuration-specific limits on the number of volumes per SVM are not exceeded.

The maximum volume limit for an SVM is controlled with the `-max-volumes` parameter for the SVM. By default, there is no limit imposed on the number of volumes the SVM can have.

The maximum volume limit for an SVM is applied only if the SVM also has an aggregate list. It is applied for both SVM administrators and cluster administrators.

How the SVM affects the language of the FlexVol volume

The language of the SVM determines the default value for the language of a FlexVol volume, although you can override that value at volume creation time. If you change the language of the SVM, it does not affect its existing FlexVol volumes. You cannot change the language of a FlexVol volume.

For FlexCache volumes and FlexClone volumes, the default language is the language of the parent volume.

8.2.3 Volume junctions

Volume junctions are a way to join individual volumes together into a single, logical namespace. Volume junctions are transparent to CIFS and NFS clients. When NAS clients access data by traversing a junction, the junction appears to be an ordinary directory.

A junction is formed when a volume is mounted to a mount point below the root and is used to create a file-system tree. The top of a file-system tree is always the root volume, which is represented by a slash (/). A junction points from a directory in one volume to the root directory of another volume.

A volume must be mounted at a junction point in the namespace to allow NAS client access to contained data:

- Although specifying a junction point is optional when a volume is created, data in the volume cannot be exported and a share cannot be created until the volume is mounted to a junction point in the namespace.
- A volume that was not mounted during volume creation can be mounted post-creation.
- New volumes can be added to the namespace at any time by mounting them to a junction point.

- Mounted volumes can be unmounted; however, unmounting a volume disrupts NAS client access to all data in the volume and to all volumes mounted at child junction points beneath the unmounted volume.
- Junction points can be created directly below a parent volume junction, or they can be created on a directory within a volume. For example, a path to a volume junction for a volume named “vol3” might be /vol1/vol2/ vol3, or it might be /vol1/dir2/vol3, or even /dir1/dir2/vol3.

8.2.4 Space management

To use the storage provided by FlexVol volumes as effectively as possible, you need to understand the space management capabilities that help you balance overall available storage against required user and application storage needs.

Data ONTAP enables space management using the following capabilities:

- Volume (space) guarantee:

  The volume guarantee, also called space guarantee or just guarantee, determines how much space for the volume is preallocated from the volume's associated aggregate when the volume is created.
- Reservations:

  Reservations, also called space reservations, file reservations, or LUN reservations, determine whether space for a particular file or LUN is preallocated from the volume.
- Fractional reserve:

  Fractional reserve, also called fractional overwrite reserve or LUN overwrite reserve, enables you to control the size of the overwrite reserve for a FlexVol volume.
- Automatic free space preservation:

  Automatic free space preservation can either increase the size of a volume or delete Snapshot copies to prevent a volume from running out of space, all without operator intervention.

These capabilities are used together to enable you to determine, on a volume-by-volume basis, whether to emphasize storage utilization, ease of management, or something in between.

Volume guarantees on FlexVol volumes

Volume guarantees (sometimes called space guarantees) determine how space for a volume is allocated from its containing aggregate whether the space is preallocated for the entire volume or for only the reserved files or LUNs in the volume, or whether space for user data is not preallocated.

The guarantee is an attribute of the volume. You set the guarantee when you create a new volume; you can also change the guarantee for an existing volume by using the **`volume modify`** command with the **`-space-guarantee`** option. You can view the guarantee type and status by using the **`volume show`** command.

Volume guarantee types can be **`volume`** (the default type), **`file`**, or **`none`**.

- A guarantee type of **`volume`** allocates space in the aggregate for the volume when you create the volume, regardless of whether that space is used for data yet. This approach to space management is called thick provisioning. The allocated space cannot be provided to or allocated for any other volume in that aggregate. When you use thick provisioning, all of the space specified for the volume is allocated from the aggregate at volume creation time. The volume cannot run out of space before the amount of data it contains (including Snapshot copies) reaches the size of the volume. However, if your volumes are not very full, this comes at the cost of reduced storage utilization.

- A guarantee type of **`file`** allocates space for the volume in its containing aggregate so that any reserved LUN or file in the volume can be completely rewritten, even if its blocks are being retained on disk by a Snapshot copy. However, writes to any file in the volume that is not reserved could run out of space.
- A guarantee of **`none`** allocates space from the aggregate only as it is needed by the volume. This approach to space management is called thin provisioning. The amount of space consumed by volumes with this guarantee type grows as data is added instead of being determined by the initial volume size, which might leave space unused if the volume data does not grow to that size. The maximum size of a volume with a guarantee of none is not limited by the amount of free space in its aggregate.

Writes to LUNs or files (including space-reserved files) contained by that volume could fail if the containing aggregate does not have enough available space to accommodate the write.

When space in the aggregate is allocated for the guarantee for an existing volume, that space is no longer considered free in the aggregate. Operations that consume free space in the aggregate, such as creation of aggregate Snapshot copies or creation of new volumes in the containing aggregate, can occur only if there is enough available free space in that aggregate; these operations are prevented from using space already allocated to another volume.

When the free space in an aggregate is exhausted, only writes to volumes or files in that aggregate with preallocated space are guaranteed to succeed.

Guarantees are honored only for online volumes. If you take a volume offline, any allocated but unused space for that volume becomes available for other volumes in that aggregate. When you try to bring that volume back online, if there is insufficient available space in the aggregate to fulfill its guarantee, it will remain offline. You must force the volume online, at which point the volume's guarantee will be disabled.

Volume guarantees and space requirements

The amount of space that a FlexVol volume requires from its aggregate varies depending on the volume's guarantee type. Understanding a volume's space requirement helps you predict how much space becomes available or is required when you change its guarantee or delete the volume.

A volume with a guarantee type of **`none`** requires space in the aggregate only for data that is already written to it.

A volume with a guarantee type of **`volume`** requires an amount of space in the aggregate equivalent to the volume's size, regardless of how much data (if any) is actually in the volume.

A volume with a guarantee type of **`file`** requires enough space in the aggregate to enable writes and overwrites to reserved files or LUNs, even if a block being overwritten is locked by a Snapshot copy or other block-sharing technology.

Considerations for using thin provisioning with FlexVol volumes

Using thin provisioning, you can configure your volumes so that they appear to provide more storage than they actually have available, provided that the storage that is actually being used does not exceed the available storage.

To use thin provisioning with FlexVol volumes, you create the volume with a guarantee of none. With a guarantee of none, the volume size is not limited by the aggregate size. In fact, each volume could, if required, be larger than the containing aggregate. The storage provided by the aggregate is used up only as data is written to the LUN or file.

If the volumes associated with an aggregate show more storage as available than the physical resources available to that aggregate, the aggregate is overcommitted. When an aggregate is overcommitted, it is possible for writes to LUNs or files in volumes contained by that aggregate to fail if there is not sufficient free space available to accommodate the write.

If you have overcommitted your aggregate, you must monitor your available space and add storage to the aggregate as needed to avoid write errors due to insufficient space.

Aggregates can provide storage to FlexVol volumes associated with more than one SVM. When sharing aggregates for thin-provisioned volumes in a multi-tenancy environment, be aware that one tenant's aggregate space availability can be adversely affected by the growth of another tenant's volumes.

File and LUN reservations for FlexVol volumes

When reservations are enabled for one or more files or LUNs, Data ONTAP reserves enough space in the volume so that writes to those files or LUNs do not fail because of a lack of disk space.

Reservations are an attribute of the file or LUN; they are persistent across storage system reboots, takeovers, and givebacks. Reservations are enabled for new LUNs by default, but you can create a file or LUN with reservations disabled or enabled. After you create a LUN, you change the reservation attribute by using the `lun modify` command. You change the reservation attribute for files by using the file reservation command.

When a volume contains one or more files or LUNs with reservations 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 reservations enabled continue to succeed.

You can enable reservations for files and LUNs contained by volumes with volume guarantees of any value. However, if the volume has a guarantee of none, reservations do not provide protection against out-of-space errors.

Considerations when using fractional reserve for FlexVol volumes

Fractional reserve, also called LUN overwrite reserve, enables you to control the size of the overwrite reserve for reserved LUNs and files in a FlexVol volume. By using this volume attribute correctly you can maximize your storage utilization, but you should understand how it interacts with other technologies.

The fractional reserve setting is expressed as a percentage; the only valid values are 0% and 100%. You use the `volume modify` command to set fractional reserve.

Setting fractional reserve to 0 increases your storage utilization. However, an application accessing data residing in the volume could experience a data outage if the volume is out of free space, even with the volume guarantee set to `volume`, when any of the following technologies and Data ONTAP features are in use:

- Deduplication
- Compression
- FlexClone files
- FlexClone LUNs
- Virtual environments

If you are using one or more of these technologies with no fractional reserve, and you need to prevent errors due to running out of space, you must use all of the following configuration settings for the volume:

- Volume guarantee of **`volume`**
- File or LUN reservations enabled
- Volume Snapshot copy automatic deletion enabled with a commitment level of **`destroy`** and a destroy list of **`lun_clone`**, **`vol_clone`**, **`cifs_share`**, **`file_clone`**, **`sfsr`**
- Auto grow feature enabled

In addition, you must monitor the free space in the associated aggregate. If the aggregate becomes full enough that the volume is prevented from growing, then data modification operations could fail even with all of the other configuration settings in place.

If you do not want to monitor aggregate free space, you can set the volume's fractional reserve setting to 100. This requires more free space up front, but guarantees that data modification operations will succeed even when the technologies listed above are in use.

The default value and allowed values for the fractional reserve setting depend on the guarantee of the volume.

Table 8-3 Default and allowed fractional reserve values

Volume guarantee	Default fractional reserve	Allowed values
Volume	100	0, 100
None	0	0, 100
File	100	100

Automatically providing more space for full FlexVol volumes

Data ONTAP uses two methods for automatically providing more space for a FlexVol volume when that volume is nearly full, allowing the volume size to increase, and deleting Snapshot copies (with any associated storage object). If you enable both of these methods, you can specify which method Data ONTAP should try first.

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 (known as the autogrow feature).

  This method is useful if the volume's containing aggregate has enough space to support a larger volume. You can configure Data ONTAP to increase the size in increments and set a maximum size for the volume. The increase is automatically triggered based on the amount of data being written to the volume in relation to the current amount of used space and any thresholds set.
- Delete Snapshot copies when the volume is nearly full.

  For example, you can configure Data ONTAP to 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 Data ONTAP to delete first, your oldest or newest Snapshot copies. You can also determine when Data ONTAP should begin deleting Snapshot copies, for example, when the volume is nearly full or when the volume's Snapshot reserve is nearly full.

If you enable both of these methods, you can specify which method Data ONTAP tries first when a volume is nearly full. If the first method does not provide sufficient additional space to the volume, Data ONTAP tries the other method next. By default, Data ONTAP tries to increase the size of the volume first.

How a FlexVol volume can automatically change its size

A volume can be configured to grow and shrink automatically in response to space usage requirements. Automatic growing occurs when used space exceeds an autogrow threshold. Automatic shrinking occurs when used space drops below an autoshrink threshold.

The autosizing feature consists of two possible functionalities:

- The **`autogrow`** functionality grows a volume's size automatically (**`grow`** option). Automatic growing can provide additional space to a volume when it is about to run out of space, as long as there is space available in the associated aggregate for the volume to grow. When the volume's free space percentage is below the specified threshold, it will continue to grow by the specified increment until either the free space percentage arrives at the threshold or the associated aggregate runs out of space.
- The **`autoshrink`** functionality shrinks a volume's size automatically (**`grow_shrink`** option). The autoshrink functionality is only used in combination with autogrow to meet changing space demands and is not available alone. Automatic shrinking helps to more accurately size a volume and prevents a volume from being larger than it needs to be at any given point. The volume shrinks and returns space to the aggregate if the guarantee type is volume.

Because the size of the Snapshot reserve is a percentage of the size of the volume, Snapshot spill can start to occur or increase as a result of a volume shrinking.

A node's root volume does not support the **`grow_shrink autosize`** mode, but you can configure an SVM’s root volume for automatic sizing.

Free space reclamation from FlexClone LUNs

Starting with Data ONTAP 8.2, you can configure the autodelete settings of a FlexVol volume to automatically delete FlexClone LUNs when the free space in a volume decreases below a particular threshold value.

A FlexVol volume that has the autodelete capability enabled resorts to automatic deletion of FlexClone LUNs only in the following situations:

- If the volume does not have Snapshot copies for automatic deletion.
- If the volume has Snapshot copies but the automatic deletion of Snapshot copies does not create sufficient free space in the volume.

Because the autodelete settings enable a FlexVol volume to directly delete FlexClone LUNs when the volume requires free space, you can preserve certain FlexClone LUNs by preventing them from getting automatically deleted.

Note: If the FlexVol volume contains FlexClone LUNs created using Data ONTAP versions earlier than 8.2 and if you want to delete them to increase the amount of free space in the volume, you can specify those FlexClone LUNs for automatic deletion.

Determining space usage in a volume or aggregate

Enabling a feature in Data ONTAP might consume space that you are not aware of or more space than you expected. Data ONTAP helps you determine how space is being consumed by providing three perspectives from which to view space, the volume, a volume's footprint within the aggregate, and the aggregate.

A volume can run out of space due to space consumption or insufficient space within the volume, aggregate, or a combination of both. By seeing a feature-oriented breakdown of space usage from different perspectives, you can assess which features you might want to adjust or turn off, or take other action (such as increase the size of the aggregate or volume).

You can view space usage details from any of these perspectives:

- The volume's space usage:

  This perspective provides details about space usage within the volume, including usage by Snapshot copies. The volume's active file system consists of user data, file system metadata, and inodes. Data ONTAP features that you enable might increase the amount of metadata, and in the case of Snapshot copies, can sometimes spill into the user data portion of the active file system. You can see a volume's space usage by using the **`volume show-space`** command.

- The volume's footprint within the aggregate:

  This perspective provides details about the amount of space each volume is using in the containing aggregate, including the volume's metadata. You can see a volume's footprint with the aggregate by using the **`volume show-footprint`** command.

- The aggregate's space usage:

  This perspective includes totals of the volume footprints of all of the volumes contained in the aggregate, space reserved for aggregate Snapshot copies, and other aggregate metadata. You can see the aggregate's space usage by using the **`storage aggregate show-space`** command.

Certain features, such as tape backup and deduplication, use space for metadata both from the volume and directly from the aggregate. These features show different space usage between the volume and volume footprint perspectives.

Methods to create space in a FlexVol volume

There are multiple ways to create space in a FlexVol volume. Understanding what these methods are and their respective benefits and drawbacks helps you decide which method is best for your requirements.

Some common ways to create space in a volume are as follows:

- Increase the size of the volume. You can do this manually, or automatically by enabling the autogrow functionality.
- Reduce the size of the Snapshot reserve if the **`df`** command shows that the Snapshot reserve is not 100% full. This makes space available to the active file system.
- Make more space in the aggregate. This results directly or indirectly in more space being made for the volume. For example, more space in the aggregate can allow a volume to increase in size automatically with the autogrow capability.
- Enable storage efficiency technologies, such as deduplication and compression.
- Delete volume Snapshot copies if the Snapshot reserve is 100% full and Snapshot copies are spilling into the active file system. You can delete Snapshot copies manually, or automatically by enabling the **`Snapshot autodelete`** capability for the volume.
- Delete FlexClone LUNs manually or enable automatic deletion of FlexClone LUNs.
- (Temporarily) change the fractional reserve to 0% if your volume contains reserved files or LUNs and the fractional reserve is 100%. You should only use this as a temporary measure to create space. When the fractional reserve is set to 0%, overwrites might fail, and in certain deployments write failures might not be acceptable.
- Delete files. If the volume is 100% full, it might not be possible to delete a file if it participates in any block sharing, such as volume Snapshot copies or deduplication, and you cannot recover the space. In addition, modifying a directory to delete a file might require additional space, so deleting the file can actually consume space.

Under these conditions, you can do one of the following actions:

- You can use the **rm** command, available at the advanced privilege level, to delete files even in full volumes with Snapshot copies.
- You can use any of the other methods listed to create more space in the volume and aggregate so that there is enough space available for file deletions.

Methods to create space in an aggregate

If an aggregate runs out of free space, various problems can result that range from loss of data to disabling a volume's guarantee. There are multiple ways to make more space in an aggregate.

All of the methods have various consequences. Prior to taking any action, you should read the relevant section in the documentation.

Here, some common ways to make space in an aggregate are listed in order of least to most consequences:

- Add disks to the aggregate.
- Move some volumes to another aggregate with available space.
- Shrink the size of volumes whose guarantee type is volume in the aggregate.
  You can do this manually or with the autoshrink option of the autosize capability.
- Change volume guarantee types to none on volumes that are using large amounts of space (large volume-guaranteed volumes or file-guaranteed volumes with large reserved files) so that the volumes take up less space in the aggregate.

  A volume with a guarantee **type** of **none** has a smaller footprint in the aggregate than volumes with other **guarantee** types. The Volume Guarantee row of the **volume show-footprint** command output shows whether a volume is reserving a large amount of space in the aggregate due to its guarantee.
- Delete unneeded volume Snapshot copies if the volume's guarantee type is none.
- Delete unneeded volumes.
- Enable space-saving features, such as deduplication or compression.
- (Temporarily) disable features that are using a large amount of metadata (visible with the **volume show-footprint** command).

8.2.5 Rules governing node root volumes and root aggregates

A node's root volume contains special directories and configuration files for that node. The root aggregate contains the root volume. A few rules govern a node's root volume and root aggregate.

A node's root volume is a FlexVol volume that is installed at the factory and reserved for system files, log files, and core files. The directory name is /mroot, which is accessible only through the system shell and with guidance from technical support.

The following rules govern the node's root volume:

- Do not change the preconfigured size for the root volume or modify the content of the root directory, unless technical support instructs you to do so.

  The minimum size for a node's root volume depends on the platform model.

Editing configuration files directly in the root directory might result in an adverse impact on the health of the node and possibly the cluster. If you need to modify system configurations, use Data ONTAP commands to do so.

- Do not store user data in the root volume.

  Storing user data in the root volume increases the storage giveback time between nodes in an HA pair.
- Do not set the root volume's fractional reserve to any value other than 100%.
- Contact technical support if you need to designate a different volume to be the new root volume or move the root volume to another aggregate.

The node's root aggregate contains the node's root volume. Starting with Data ONTAP 8.1, new systems are shipped with the root volume in a dedicated, 64-bit root aggregate that contains three disks. By default, a node is set up to use a hard disk drive (HDD) aggregate for the root aggregate. When no HDDs are available, the node is set up to use a solid-state drive (SSD) aggregate for the root aggregate.

The root aggregate must be dedicated to the root volume only. You must not include or create data volumes in the root aggregate.

8.2.6 Moving and copying volumes (cluster administrators only)

You can move or copy volumes for capacity utilization, improved performance, and to satisfy service-level agreements.

Moving a FlexVol volume

FlexVol volumes are moved from one aggregate or node to another within the same SVM for capacity utilization and improved performance, and to satisfy service-level agreements.

A volume move does not disrupt client access during the move.

> **Note:** You cannot move 64-bit volumes to 32-bit aggregates.

Moving a volume occurs in multiple phases:

- A new volume is made on the destination aggregate.
- The data from the original volume is copied to the new volume.
  During this time, the original volume is intact and available for clients to access.
- At the end of the move process, client access is temporarily blocked.
  During this time, the system performs a final replication from the source volume to the destination volume, swaps the identities of the source and destination volumes, and changes the destination volume to the source volume.
- After completing the move, the system routes client traffic to the new source volume and resumes client access.

The move is not disruptive to client access because the time in which client access is blocked ends before clients notice a disruption and time out. Client access is blocked for 45 seconds by default. If the volume move operation cannot finish in the time that access is denied, the system aborts this final phase of the volume move operation and allows client access. The system runs the final phase of the volume move operation until the volume move is complete or until the default maximum of three attempts is reached. If the volume move operation fails after the third attempt, the process goes into a cut over deferred state and waits for you to initiate the final phase.

You can change the amount of time client access is blocked or the number of times (cut over attempts) the final phase of the volume move operation is run if the defaults are not adequate. You can also determine what the system does if the volume move operation cannot be completed during the time client access is blocked. The volume move start man page contains details about moving a volume without disrupting client access.

Commands for moving volumes

There are specific Data ONTAP commands for managing volume movement.

Table 8-4 Moving volume commands

If you want to...	Use this command
Abort an active volume move operation.	`volume move abort`
Show status of a volume moving from one aggregate to another aggregate.	`volume move show`
Start moving a volume from one aggregate to another aggregate.	`volume move start`
Manage target aggregates for volume move.	`volume move target-aggr`
Trigger cut over of a move job.	`volume move trigger-cutover`

See the man page for each command for more information.

SnapMirror transfers can affect volume move operations

If SnapMirror transfers are running at the same time as volume move operations, the volume move operations are prevented from entering the cut over phase.

The cut over phase cannot occur for the following cases:

- If there are checkpoints on the volume.

  A checkpoint can exist for one of the following reasons:

  - A relationship exists and there is no active transfer, but there is a checkpoint.
  - A relationship has been deleted, but a checkpoint was left.
- If the volume is the destination of an active SnapMirror transfer.

Characteristics of how Data ONTAP copies FlexVol volumes

A copy of a FlexVol volume is a full copy of the original FlexVol volume with the same access (read-only or read-write) as the original volume. Knowing the characteristics of the volume copy can help set your expectations about the volume copy result.

A volume copy has the following characteristics:

- A copy of a volume does not share blocks with the original volume.
  A copy of a 2-GB volume uses 2 GB of disk space.
- After the copy is made, no operations made on the copy or on the original affect the other. For example, if you write data to the original volume, the data is not written to the copy.
- A volume copy is not automatically mounted when it is created.
- A volume copy must occur within the context of the same SVM.
- A volume copy does not copy a volume's SnapMirror labels.
- A 64-bit volume can only be copied to 64-bit aggregates.
  It cannot be copied to a 32-bit aggregate.
- An offline volume cannot be copied.

Copying a FlexVol volume

Copying a volume creates a stand-alone copy of a volume that you can use for testing and other purposes.

About this task

Copying a volume has the following limitations:

- You can copy a volume within an SVM only.
- You can copy a FlexVol volume to a FlexVol volume only.
- If you assigned one or more aggregates to the associated SVM, the destination aggregate must be one of the assigned aggregates.

Steps

To create a stand-alone copy of a volume, follow these steps:

1. Use the **`volume copy start`** command.

   You can make the copy on the same aggregate as the original or on a different aggregate. When the copy is complete, it has no relation to its source volume; changes made to one volume are not propagated to the other.

   Example 8-1 creates a copy of a volume named src_builds on an SVM named vs0. The copy is named builds and is located on an aggregate named aggr4. The copy operation runs as a background process.

Example 8-1 Creating a volume copy

```
cdot-cluster1::> volume copy start -vserver vs0 -volume src_builds
-destination-volume builds -destination-aggregate aggr4 -foreground
false
```

2. Use the **`job show`** command to determine if the volume copy operation is complete.
3. The copy is not automatically mounted; mount it using the **`volume mount`** command.

8.2.7 FlexClone volumes

FlexClone volumes are writable, point-in-time copies of a parent FlexVol volume. FlexClone volumes are space-efficient because they share the same data blocks with their parent FlexVol volumes for common data. The Snapshot copy used to create a FlexClone volume is also shared with the parent volume.

You can clone an existing FlexClone volume to create another FlexClone volume. You can also create a clone of a FlexVol volume containing LUNs and LUN clones.

Starting with Data ONTAP 8.2, you can create two types of FlexClone volumes, read-write FlexClone volumes and data protection FlexClone volumes. While you can create a read-write FlexClone volume of a regular FlexVol volume, you must use only a SnapVault secondary volume to create a data protection FlexClone volume.

Understanding FlexClone volumes

FlexClone volumes can be managed similarly to regular FlexVol volumes, with a few important differences. For example, the changes made to the parent FlexVol volume after the FlexClone volume is created are not reflected in the FlexClone volume.

The following list outlines important facts about FlexClone volumes:

> **Note:** The following statements are applicable to both read-write and data protection FlexClone volumes unless specified otherwise.

- A FlexClone volume is a point-in-time, writable copy of the parent FlexVol volume.
- A FlexClone volume is a fully functional FlexVol volume similar to its parent.
- A FlexClone volume is always created in the same aggregate as its parent.
- A FlexClone volume is always created in the same virtual storage server (SVM) as its parent.
- An Infinite Volume cannot be used as the parent of a FlexClone volume.
- Because a FlexClone volume and its parent share the same disk space for common data, creating a FlexClone volume is instantaneous and requires no additional disk space (until changes are made to the FlexClone volume or its parent).
- A FlexClone volume is created with the same volume guarantee as its parent.
  The volume guarantee setting is enforced for the new FlexClone volume only if there is enough space in the containing aggregate.
- A FlexClone volume is created with the same space reservation and fractional reserve settings as its parent.
- A FlexClone volume is created with the same Snapshot schedule as its parent.
- A FlexClone volume is created with the same language setting as its parent.
- The common Snapshot copy shared between a FlexClone volume and its parent volume cannot be deleted while the FlexClone volume exists.
- While a FlexClone volume exists, some operations on its parent are not allowed, such as deleting the parent volume.
- You cannot create clones of volumes in a storage system that is in a partial giveback state.
- You can break the connection between the parent volume and a read-write FlexClone volume.

  This is called splitting the FlexClone volume. Splitting removes all restrictions on the parent volume and causes the FlexClone volume to use its own additional disk space rather than sharing space with its parent.

  > **Note:** You cannot split a data protection FlexClone volume from its parent volume.

  > **Attention:** Splitting a FlexClone volume from its parent volume deletes all existing Snapshot copies of the FlexClone volume, and disables the creation of new Snapshot copies while the splitting operation is in progress.

  If you want to retain the Snapshot copies of the FlexClone volume, you can move the FlexClone volume to a different aggregate by using the volume move command. During the volume move operation, you can also create new Snapshot copies, if required.
- Quotas applied to the parent volume are not automatically applied to the FlexClone volume.
- When a FlexClone volume is created, any LUNs present in the parent volume are present in the FlexClone volume but are unmapped and offline.

FlexClone volumes and shared Snapshot copies

When volume guarantees are in effect, a new FlexClone volume uses the Snapshot copy it shares with its parent to minimize its space requirements. If you delete the shared Snapshot copy, you might increase the space requirements of the FlexClone volume.

For example, suppose that you have a 100-MB FlexVol volume that has a volume guarantee of volume, with 70 MB used and 30 MB free, and you use that FlexVol volume as a parent volume for a new FlexClone volume. The new FlexClone volume has an initial volume guarantee of volume, but it does not require a full 100 MB of space from the aggregate, as it would if you had copied the volume. Instead, the aggregate needs to allocate only 30 MB (100 MB – 70 MB) of free space to the clone.

Now, suppose that you delete the shared Snapshot copy from the FlexClone volume. The FlexClone volume can no longer optimize its space requirements, and the full 100 MB is required from the containing aggregate.

> **Note:** If you are prevented from deleting a Snapshot copy from a FlexClone volume due to "insufficient space in the aggregate," it is because deleting that Snapshot copy requires the allocation of more space than the aggregate currently has available. You can either increase the size of the aggregate or change the volume guarantee of the FlexClone volume.

You can identify a shared Snapshot copy by using the **`volume snapshot show`** command with the **`-instance`** parameter to list the Snapshot copies in the parent volume. Any Snapshot copy that is marked as busy in the parent volume and is also present in the FlexClone volume is a shared Snapshot copy.

Using volume SnapMirror replication and FlexClone volumes

Because both volume SnapMirror replication and FlexClone volumes rely on Snapshot copies, there are some restrictions on how the two features can be used together. For example, you can create a volume SnapMirror relationship using a FlexClone volume or its parent as the source volume.

However, you cannot create a new volume SnapMirror relationship using either a FlexClone volume or its parent as the destination volume.

Considerations when creating a FlexClone volume

You can create a FlexClone volume from the source or destination volume in an existing volume SnapMirror relationship. However, doing so could prevent future SnapMirror replication operations from completing successfully.

Replication might not work because when you create the FlexClone volume, you might lock a Snapshot copy that is used by SnapMirror. If this happens, SnapMirror stops replicating to the destination volume until the FlexClone volume is destroyed or is split from its parent. You have two options for addressing this issue:

- If you require the FlexClone volume on a temporary basis and can accommodate a temporary stoppage of the SnapMirror replication, you can create the FlexClone volume and either delete it or split it from its parent when possible.

  The SnapMirror replication continues normally when the FlexClone volume is deleted or is split from its parent.

- If a temporary stoppage of the SnapMirror replication is not acceptable, you can create a Snapshot copy in the SnapMirror source volume, and then use that Snapshot copy to create the FlexClone volume. (If you are creating the FlexClone volume from the destination volume, you must wait until that Snapshot copy replicates to the SnapMirror destination volume.)

  This method of creating a Snapshot copy in the SnapMirror source volume allows you to create the clone without locking a Snapshot copy that is in use by SnapMirror.

Splitting a FlexClone volume from its parent

Splitting a read-write FlexClone volume from its parent removes any space optimizations that are currently used by the FlexClone volume. After the split, both the FlexClone volume and the parent volume require the full space allocation determined by their volume guarantees. The FlexClone volume becomes a normal FlexVol volume.

You must be aware of the following considerations related to clone-splitting operations:

- You can split only read-write FlexClone volumes. Data protection FlexClone volumes cannot be split from their parent volumes.
- When you split a FlexClone volume from its parent, all existing Snapshot copies of the FlexClone volume are deleted. If you want to retain the Snapshot copies of the FlexClone volume, you can move the FlexClone volume to a different aggregate by using the **`volume move`** command. During the volume move operation, you can also create new Snapshot copies, if required.
- No new Snapshot copies can be created of the FlexClone volume during the split operation.
- Because the clone-splitting operation is a copy operation that might take considerable time to complete, Data ONTAP provides the **`volume clone split stop`** and **`volume clone split status`** commands to stop or check the status of a clone-splitting operation.
- The clone-splitting operation proceeds in the background and does not interfere with data access to either the parent or the clone volume.
- The FlexClone volume must be online when you start the split operation.
- The parent volume must be online for the split operation to succeed.
- If the FlexClone volume has a data protection or load-sharing mirror, it cannot be split from its parent volume.
- If you split a FlexClone volume from a FlexVol volume that has deduplication and compression enabled, the split volume does not have deduplication and compression enabled.
- After a FlexClone volume and its parent volume have been split, they cannot be rejoined.

FlexClone volumes and LUNs

You can clone FlexVol volumes that contain LUNs and FlexClone LUNs.

> **Note:** 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.

When you create a FlexClone volume, LUNs in the parent volume are present in the FlexClone volume but they are not mapped and they are offline. To bring the LUNs in the FlexClone volume online, you need to map them to initiator groups.

If the parent volume contains FlexClone LUNs, the FlexClone volume also contains FlexClone LUNs, which share storage with the FlexClone LUNs in the parent volume.

Understanding data protection FlexClone volumes

You can use the FlexClone technology to create a space-efficient copy of a data protection volume that is used as a SnapVault secondary volume. The Snapshot copy that establishes the SnapVault relationship between the primary and secondary volumes is the backing Snapshot copy for creating the data protection FlexClone volume.

Data protection FlexClone volumes are similar to read-write FlexClone volumes because they share common blocks with their parent FlexVol volumes. However, you can create a data protection FlexClone volume only from a parent FlexVol volume that is also a secondary SnapVault volume. In addition, you cannot split a data protection FlexClone volume from its parent volume.

How a FlexVol volume can reclaim free space from FlexClone LUNs

Starting with Data ONTAP 8.2, you can configure the auto delete settings of a FlexVol volume to automatically delete FlexClone LUNs when the free space in a volume decreases below a particular threshold value.

A FlexVol volume that has the auto delete capability enabled resorts to automatic deletion of FlexClone LUNs only in the following situations:

- If the volume does not have Snapshot copies for automatic deletion.
- If the volume has Snapshot copies but the automatic deletion of Snapshot copies does not create sufficient free space in the volume.

Because the auto delete settings enable a FlexVol volume to directly delete FlexClone LUNs when the volume requires free space, you can preserve certain FlexClone LUNs by preventing them from getting automatically deleted.

Note: If the FlexVol volume contains FlexClone LUNs created using Data ONTAP versions earlier than 8.2 and if you want to delete them to increase the amount of free space in the volume, you can specify those FlexClone LUNs for automatic deletion.

Features supported with FlexClone files and FlexClone LUNs

FlexClone files and FlexClone LUNs work with different Data ONTAP features such as deduplication, Snapshot copies, quotas, and volume SnapMirror.

The following features are supported with FlexClone files and FlexClone LUNs:

- Deduplication
- Snapshot copies
- Access control lists
- Quotas
- FlexClone volumes
- NDMP
- Volume SnapMirror
- The volume move command
- The volume copy command
- Space reservation
- HA configuration

8.2.8 Qtrees

Qtrees enable you to partition your FlexVol volumes into smaller segments that you can manage individually. You can use qtrees to manage quotas, security style, and CIFS oplocks.

Data ONTAP creates a default qtree, called qtree0, for each volume. If you do not put data into a qtree, it resides in qtree0.

When to use qtrees

Qtrees enable you to partition your data without incurring the overhead associated with a FlexVol volume. You might create qtrees to organize your data, or to manage one or more of the following factors: quotas, security style, and CIFS oplocks setting.

The following list describes examples of qtree usage strategies:

- 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.

- Security style:

  If you have a project that needs to use NTFS-style security, because the members of the project use Windows files and applications, you can group the data for that project in a qtree and set its security style to NTFS, without requiring that other projects also use the same security style.

- CIFS oplocks settings:

  If you have a project using a database that requires CIFS oplocks to be off, you can set CIFS oplocks to `off` for that project's qtree, while allowing other projects to retain CIFS oplocks.

Qtrees compared to FlexVol volumes

In general, qtrees are similar to FlexVol volumes. However, the two technologies have some key differences. Understanding these differences helps you choose between them when you design your storage architecture.

Table 8-5 compares qtrees and FlexVol volumes.

Table 8-5 Comparison between qtrees and FlexVol volumes

Functionality	Qtree	FlexVol volume
Enables organizing user data	Yes	Yes
Enables grouping users with similar needs	Yes	Yes
Accepts a security style	Yes	Yes
Accepts oplocks configuration	Yes	Yes
Can be resized	Yes (using quota limits)	Yes
Supports Snapshot copies	No (qtree data can be extracted from volume Snapshot copies)	Yes
Supports quotas	Yes	Yes
Can be cloned	No (except as part of a FlexVol volume)	Yes

Functionality	Qtree	FlexVol volume
Can serve as the root of an SVM	No	Yes
Can serve as a junction	No	Yes
Can be exported using NFS	No	Yes

Qtree name restrictions

Qtree names can be no more than 64 characters in length. In addition, using some special characters in qtree names, such as commas and spaces, can cause problems with other Data ONTAP capabilities, and should be avoided.

Qtrees and mirrors

You can see but not modify qtrees that exist within a mirror.

For example, you can use the **`volume qtree statistics`** command on the mirror. Note that information displayed about the qtrees (including name, security style, oplock mode, and other attributes) may not be synchronized between the read-write volume and the mirror, depending on the mirror's replication schedule. But after the read-write volume is replicated to the mirror, qtree information is synchronized.

However, you cannot create, modify, or delete the qtrees on the mirror.

Commands for managing qtrees

There are specific Data ONTAP commands for managing and configuring qtrees.

Many qtree commands cannot be performed while a volume move operation is in progress. If you are prevented from completing a qtree command for this reason, wait until the volume move is complete and then retry the command. See Table 8-6.

Table 8-6 Commands for managing qtrees

If you want to...	Use this command
Create a qtree	**`volume qtree create`**
Display a filtered list of qtrees	**`volume qtree show`**
Delete a qtree	**`volume qtree delete`**
Modify a qtree’s UNIX permissions	**`volume qtree modify -unix- permissions`**
Modify a qtree’s CIFS oplocks settings	**`volume qtree oplocks`**
Modify a qtree’s security setting	**`volume qtree security`**
Rename a qtree	**`volume qtree rename`**
Display a qtree’s statistics	**`volume qtree statistics`**
Reset a qtree’s statistics	**`volume qtree statistics -reset`**

8.2.9 Quotas

Quotas provide a way to restrict or track the disk space and number of files used by a user, group, or qtree. Quotas are applied to a specific FlexVol volume or qtree.

Why you use quotas

You can use quotas to limit resource usage in FlexVol volumes, to provide notification when resource usage reaches specific levels, or to track resource usage.

You specify a quota for the following reasons:

- To limit the amount of disk space or the number of files that can be used by a user or group, or that can be contained by a qtree
- To track the amount of disk space or the number of files used by a user, group, or qtree, without imposing a limit
- To warn users when their disk usage or file usage is high

Overview of the quota process

Quotas can be soft or hard. Soft quotas cause Data ONTAP to send a notification when specified thresholds are exceeded, and hard quotas prevent a write operation from succeeding when specified thresholds are exceeded.

When Data ONTAP receives a request to write to a FlexVol volume, it checks to see whether quotas are activated for that volume. If so, Data ONTAP determines whether any quota for that volume (and, if the write is to a qtree, for that qtree) would be exceeded by performing the write operation. If any hard quota is exceeded, the write operation fails, and a quota notification is sent. If any soft quota is exceeded, the write operation succeeds, and a quota notification is sent.

Differences among hard, soft, and threshold quotas

Hard quotas prevent operations while soft quotas trigger notifications.

Hard quotas impose a hard limit on system resources; any operation that would result in exceeding the limit fails. The following settings create hard quotas:

- Disk Limit parameter
- Files Limit parameter

Soft quotas send a warning message when resource usage reaches a certain level, but do not affect data access operations, so you can take appropriate action before the quota is exceeded. The following settings create soft quotas:

- Threshold for Disk Limit parameter
- Soft Disk Limit parameter
- Soft Files Limit parameter

Threshold and Soft Disk quotas enable administrators to receive more than one notification about a quota. Typically, administrators set the Threshold for Disk Limit to a value that is only slightly smaller than the Disk Limit, so that the threshold provides a “final warning” before writes start to fail.

Understanding quota notifications

Quota notifications are messages that are sent to the event management system (EMS) and also configured as SNMP traps.

Notifications are sent in response to the following events:

- A hard quota is reached; in other words, an attempt is made to exceed it
- A soft quota is exceeded
- A soft quota is no longer exceeded

Thresholds are slightly different from other soft quotas. Thresholds trigger notifications only when they are exceeded, not when they are no longer exceeded.

Hard-quota notifications are configurable using the `volume quota modify` command. You can turn them off completely, and you can change their frequency, for example, to prevent sending of redundant messages.

Soft-quota notifications are not configurable because they are unlikely to generate redundant messages and their sole purpose is notification.

Table 8-7 lists the events that quotas send to the EMS system.

Table 8-7 List of events that the system can send to the EMS system

When this occurs...	This event is sent to the EMS...
A hard limit is reached in a tree quota	wafl.quota.qtree.exceeded
A hard limit is reached in a user quota on the volume	wafl.quota.user.exceeded (for a UNIX user) wafl.quota.user.exceeded.win (for a Windows user)
A hard limit is reached in a user quota on a qtree	wafl.quota.userQtree.exceeded (for a UNIX user) wafl.quota.userQtree.exceeded.win (for a Windows user)
A hard limit is reached in a group quota on the volume	wafl.quota.group.exceeded
A hard limit is reached in a group quota on a qtree	wafl.quota.groupQtree.exceeded
A soft limit, including a threshold, is exceeded	quota.softlimit.exceeded
A soft limit is no longer exceeded	quota.softlimit.normal

Table 8-8 lists the SNMP traps that quotas generate.

Table 8-8 List of SNMP traps the system can generate regarding quotas

When this occurs...	This SNMP trap is sent...
A hard limit is reached.	quotaExceeded
A soft limit, including a threshold, is exceeded.	quotaExceeded and softQuotaExceeded
A soft limit is no longer exceeded.	quotaNormal and softQuotaNormal

Note: Notifications contain qtree ID numbers rather than qtree names. You can correlate qtree names to ID numbers by using the `volume qtree show -id` command.

Quota rules, quota policies, and quotas

Quotas are defined in quota rules specific to FlexVol volumes. These quota rules are collected together in a quota policy of a virtual storage server (SVM), and then activated on each volume on the SVM.

A quota rule is always specific to a volume. Quota rules have no effect until quotas are activated on the volume defined in the quota rule.

A quota policy is a collection of quota rules for all the volumes of an SVM. Quota policies are not shared among SVMs. An SVM can have up to five quota policies, which enable you to have backup copies of quota policies. One quota policy is assigned to an SVM at any time.

A quota is the actual restriction that Data ONTAP enforces or the actual tracking that Data ONTAP performs. A quota rule always results in at least one quota, and might result in many additional derived quotas. The complete list of enforced quotas is visible only in quota reports.

Activation is the process of triggering Data ONTAP to create enforced quotas from the current set of quota rules in the assigned quota policy. Activation occurs on a volume-by-volume basis. The first activation of quotas on a volume is called initialization. Subsequent activations are called either reinitialization or resizing, depending on the scope of the changes.

> **Note:** When you initialize or resize quotas on a volume, you are activating the quota rules in the quota policy that is currently assigned to the SVM.

Quota targets and types

Quotas have a type: they can be either user, group, or tree. Quota targets specify the user, group, or qtree for which the quota limits are applied.

Table 8-9 lists the kinds of quota targets, what types of quotas each quota target is associated with, and how each quota target is represented.

Table 8-9 Quota targets and types

Quota target	Quota type	How target is represented	Notes
user	User quota	UNIX user name UNIX UID A file or directory whose UID matches the user Windows user name in pre-Windows 2000 format Windows SID A file or directory with an ACL owned by the user's SID	User quotas can be applied for a specific volume or qtree
group	Group quota	UNIX group name UNIX GID A file or directory whose GID matches the group	Group quotas can be applied for a specific volume or qtree.
qtree	Tree quota	The qtree name	Tree quotas are applied to a particular volume and do not affect qtrees in other volumes.
*	User quota group quota tree quota	The asterisk character (*)	A quota target of * denotes a default quota. For default quotas, the quota type is determined by the value of the type field.

How quotas are applied

Understanding how quotas are applied enables you to configure quotas and set the expected limits.

Whenever an attempt is made to create a file or write data to a file in a FlexVol volume that has quotas enabled, the quota limits are checked before the operation proceeds. If the operation exceeds either the disk limit or the files limit, the operation is prevented.

Quota limits are checked in the following order:

1. The tree quota for that qtree.
   (This check is not relevant if the file is being created or written to qtree0.)
2. The user quota for the user that owns the file on the volume
3. The group quota for the group that owns the file on the volume
4. The user quota for the user that owns the file on the qtree.
   (This check is not relevant if the file is being created or written to qtree0.)
5. The group quota for the group that owns the file on the qtree.
   (This check is not relevant if the file is being created or written to qtree0.)

The quota with the smallest limit might not be the one that is exceeded first. For example, if a user quota for volume vol1 is 100 GB, and the user quota for qtree q2 contained in volume vol1 is 20 GB, the volume limit could be reached first if that user has already written more than 80 GB of data in volume vol1 (but outside of qtree q2).

Considerations for assigning quota policies

A quota policy is a grouping of the quota rules for all the FlexVol volumes of an SVM. You must be aware of certain considerations when assigning the quota policies:

- An SVM has one assigned quota policy at any given time. When an SVM is created, a blank quota policy is created and assigned to the SVM. This default quota policy has the name “default” unless a different name is specified when the SVM is created.
- An SVM can have up to five quota policies. If an SVM has five quota policies, you cannot create a new quota policy for the SVM until you delete an existing quota policy.
- When you need to create a quota rule or change quota rules for a quota policy, you can choose either of the following approaches:
  - If you are working in a quota policy that is assigned to an SVM, then you need not assign the quota policy to the SVM.
  - If you are working in an unassigned quota policy and then assigning the quota policy to the SVM, then you must have a backup of the quota policy that you can revert to if required. For example, you can make a copy of the assigned quota policy, change the copy, assign the copy to the SVM, and rename the original quota policy.
- You can rename a quota policy even when it is assigned to the SVM.

Commands to manage quota rules and quota policies

You can use the **`volume quota policy`** rule commands to configure quota rules, and use the **`volume quota policy`** commands and some SVM commands to configure quota policies.

Note: You can run the commands for managing quota rules only on FlexVol volumes.

See Table 8-10 for commands to manage quota rules.

Table 8-10 Commands to manage quota rules

If you want to...	**Use this command...**
Create a new quota rule	**volume quota policy rule create**
Delete an existing quota rule	**volume quota policy rule delete**
Modify an existing quota rule	**volume quota policy rule modify**
Display information about configured quota rules	**volume quota policy rule show**

See Table 8-11 for commands to manage quota policies.

Table 8-11 Commands to manage quota policies

If you want to...	**Use this command...**
Duplicate a quota policy and the quota rules it contains	volume quota policy copy
Create a new, blank quota policy	volume quota policy create
Delete an existing quota policy that is not currently assigned to the SVM	volume quota policy delete
Rename a quota policy	volume quota policy rename
Display information about quota policies	volume quota policy show
Assign a quota policy to an SVM	vserver modify
Display the name of the quota policy assigned to an SVM	vserver show

See the man page for each command for more information.

Commands to activate and modify quotas

You can use the `volume quota` commands to change the state of quotas and configure message logging of quotas. See Table 8-12.

Table 8-12 Commands to activate and modify quotas

If you want to...	**Use this command...**
Turn quotas on (also called initializing them)	volume quota on
Resize existing quotas	volume quota resize
Turn quotas off	volume quota off
Change the message logging of quotas, turn quotas on, turn quotas off, or resize existing quotas	volume quota modify

See the man page for each command for more information.

8.3 Infinite Volumes

You can use an Infinite Volume to create a large, scalable data container with a single namespace and a single mount point.

8.3.1 Infinite Volume components

An Infinite Volume is made of a group of constituents stitched together into a single volume. When an Infinite Volume is created, it automatically creates the following constituents distributed across nodes:

- Namespace constituent
- One or more namespace mirror constituents
- Data constituents

The namespace constituent contains directory and file names and pointer references to the physical location of the file in the Infinite Volume. It is also the junction path, which is the client-accessible namespace for the entire Infinite Volume. There is one namespace constituent per Infinite Volume, and by default, it is a maximum of 10TB.

The namespace mirror constituent contains an asynchronous volume SnapMirror copy of the namespace constituent. It serves two main purposes, backup for the namespace constituent and enabling support for differential tape backup by using SnapDiff. There is one namespace mirror constituent for backup of the namespace constituent. It is replicated every 5 minutes, and is equal in size to the namespace constituent.

SnapDiff requires that each node contain either a namespace constituent or a namespace mirror constituent. Because one namespace constituent and one namespace mirror constituent are already created by default, adding SnapDiff involves creating additional namespace mirror constituents on each node that contains an Infinite Volume data constituent, but that does not contain a namespace constituent or a namespace mirror constituent. By default, the namespace constituent mirrors created for SnapDiff are replicated once a day, but this can be modified to a value larger than 1 hour. The namespace mirror constituents are equal in size to the namespace constituent.

The data constituents contain the data from files stored in the Infinite Volume. An entire file exists within a single data constituent. Data constituents are created on each node that has at least one aggregate assigned to the Infinite Volume. Upon Infinite Volume creation, equal amounts of usable data constituent space are created on each node that contains an aggregate assigned to the SVM for Infinite Volume. Data constituents can grow up to the maximum supported size for the model of system that contains it.

8.3.2 Requirements for Infinite Volumes

Infinite Volumes require specific platforms, nodes, SVM and aggregate configurations, and junction names. You should review these requirements before you create or expand an Infinite Volume.

Platforms supported by Infinite Volumes

Infinite Volumes are supported only on certain platforms. If you try to create or expand an Infinite Volume that uses aggregates from an unsupported platform, an error message appears and you cannot proceed.

Infinite Volumes support a heterogeneous configuration where you mix different platforms in the same cluster. The nodes that an Infinite Volume spans do not all have to be the same platform.

Node requirements for Infinite Volumes

Unlike FlexVol volumes, Infinite Volumes use multiple nodes in a cluster. Each Infinite Volume can use from 2 through 10 nodes. You should be aware of node requirements and how to meet them before you create or expand an Infinite Volume.

SVM requirements for Infinite Volumes

An Infinite Volume requires a specifically configured SVM that does not contain any other volumes, and the SVM should have a defined list of associated aggregates. You should be aware of these SVM requirements before you create an Infinite Volume.

An Infinite Volume requires an SVM with the following configuration:

- An SVM that has been configured specifically to make it an SVM with Infinite Volume. An Infinite Volume cannot be created inside an SVM with the default configuration, which is called an SVM with FlexVol volumes.

  The **`-is-repository`** parameter is what differentiates an SVM with Infinite Volume from an SVM with FlexVol volumes. The parameter **`istrue`** for SVMs with Infinite Volume and false for SVMs with FlexVol volumes.

- An SVM that contains no volumes.

  An SVM with Infinite Volume cannot contain more than one Infinite Volume, and it cannot contain any FlexVol volumes. The SVM root volume is an exception; every SVM has a root volume for internal management.

- An SVM with a defined list of aggregates (advised).

  After creating an SVM with Infinite Volume, you should define its associated aggregates. Without additional configuration, every SVM is initially associated with all the aggregates in the cluster. A cluster administrator can associate aggregates with an SVM by using the **`-aggr-list`** parameter of the **`vserver modify`** command.

Unlike in earlier versions of Data ONTAP, you are not required to dedicate an entire cluster to the SVM with Infinite Volume. Multiple SVMs with Infinite Volume and multiple SVMs with FlexVol volumes can coexist in the same cluster.

Aggregate requirements for Infinite Volumes

The aggregates that are used by an Infinite Volume must be 64-bit aggregates and must have more than 1.1 TB of available space. If the Infinite Volume uses storage classes, the aggregates must also meet the requirements of the storage class.

The cluster that an Infinite Volume is in can contain 32-bit aggregates, but the aggregates that are associated with any Infinite Volume must all be 64-bit aggregates.

If an aggregate has less than 1.1 TB of available space, it is not used by the SVM with Infinite Volume. The aggregate must have enough space to store a 1 TB data constituent.

If the Infinite Volume uses storage classes, aggregates must meet the requirements of the storage class to be used. For example, if the storage class uses only SAS disks, aggregates created for that storage class must consist entirely of SAS disks.

8.3.3 Before you create an Infinite Volume

Before you create an Infinite Volume, you might want to consider using storage classes, thin provisioning, incremental tape backup, or a data protection mirror relationship with an Infinite Volume that uses a platform with a smaller maximum data constituent size.

Each of these decisions affects the creation of an Infinite Volume in the following ways:

- Storage classes affect the type of aggregates you can use and what tools you can use to create and manage the Infinite Volume.
- With thin provisioning, the size of the Infinite Volume is not tied directly to the available physical space.
- With incremental tape backup, some space in the Infinite Volume is dedicated to additional namespace mirror constituents.
- If the Infinite Volume will be in a data protection mirror relationship with an Infinite Volume that uses a smaller maximum data constituent size, you must restrict the size of the data constituents in the source Infinite Volume when you create the source Infinite Volume.

Considerations when using storage classes

If you want to use storage classes with an Infinite Volume, you should be aware that storage classes can affect the type of aggregates you can use and what tools you can use to create and manage the Infinite Volume. Some command-line interface commands are disabled for Infinite Volumes with storage classes.

Storage class definitions specify what type of aggregate to use. When you create aggregates for an Infinite Volume with storage classes, you must be aware of the definitions of the storage classes that you want to use so that you can create the appropriate type of aggregates. OnCommand Workflow Automation validates that aggregates comply with the restrictions of the storage class before allowing the storage class to use the aggregates. You must use OnCommand Workflow Automation to define workflows for your storage class needs and to assign storage classes to Infinite Volumes.

Storage class definitions can also include some volume settings, such as thin provisioning, compression, and deduplication. When you create an Infinite Volume with storage classes, some settings no longer apply to the entire Infinite Volume. Instead some settings apply to individual storage classes in the Infinite Volume, which allows you to use one or more different storage classes with an Infinite Volume.

When you create an Infinite Volume with two or more storage classes, you should use a data policy to automatically filter incoming data into different storage classes. Otherwise all data written to the Infinite Volume is stored in the first storage class created. You should use OnCommand Unified Manager to modify and manage the data policy for an Infinite Volume.

Considerations when using thin provisioning with Infinite Volumes

You can use thin provisioning with an Infinite Volume, enabling you to allocate more storage to users than is physically available. Before using thin provisioning, you should understand what it is, where and when it is configured, and what its advantages and disadvantages are.

What thin provisioning is

With thin provisioning, the size of an Infinite Volume is not limited by the size of its associated aggregates. You can create a large volume on a small amount of storage, adding disks only as they are required. For example, you can create a 10 TB volume using aggregates that only have 5 TB of available space. The storage provided by the aggregates is used only as data is written. Thin provisioning is also called aggregate overcommitment.

The alternative of thin provisioning is thick provisioning, which allocates physical space immediately, regardless of whether that space is used for data yet. The allocated space cannot be used by any other volumes. When you use thick provisioning, all of the space required for the volume is allocated from the aggregate at the time of creating the volume.

Thin provisioning affects only the data constituents of an Infinite Volume. The namespace constituent and namespace mirror constituents of an Infinite Volume always use thick provisioning. For example, if you create a new Infinite Volume with a 2 TB namespace constituent and use thin provisioning, the namespace constituent will consume 2 TB of space even if the Infinite Volume does not contain any data.

When and where thin provisioning is configured

The way that you configure thin provisioning on an Infinite Volume depends on whether the Infinite Volume uses storage classes.

For an Infinite Volume without storage classes, thick and thin provisioning are configured at the volume level in the following way:

- If you want to use thin provisioning, you typically specify it when you first create the Infinite Volume. By default, an Infinite Volume created through the command line uses thick provisioning.
- You can switch between thick and thin provisioning after the Infinite Volume is created. If you change the setting later, you cannot change the Infinite Volume's size at the same time; you must change the size and guarantee of an Infinite Volume in separate operations. Before changing a volume from thin provisioning to thick provisioning, you must ensure that the physical storage can support the provisioned size.
- You can configure thick and thin provisioning in the command line by using the **`-space-guarantee`** parameter of the **`volume create`** or the **`volume modify`** command. The value none represents thin provisioning and the value volume represents thick provisioning.

For an Infinite Volume with storage classes, thick and thin provisioning are configured at the storage-class level in the following way:

- You can choose to use thick or thin provisioning for each storage class independent of other storage classes.

  For example, one storage class of an Infinite Volume can use thin provisioning while another storage class of the same Infinite Volume uses thick provisioning. When the guarantee differs across storage classes, the guarantee for the entire Infinite Volume is displayed as a dash (—).
- All configuration of thick and thin provisioning is performed by using OnCommand Workflow Automation.
- If an Infinite Volume uses storage classes, it is not possible to configure thick or thin provisioning at the Infinite Volume level.

The **`-space-guarantee`** parameter of the **`volume create`** and the **`volume modify`** commands is disabled for an Infinite Volume with storage classes.

Advantages of thin provisioning

Using thin provisioning with Infinite Volumes provides the following advantages:

- It defers physical storage costs until the storage is actually required.

  Users receive the space allocation that they expect, and valuable resources do not remain unused.
- It facilitates monitoring of aggregate usage.

When you use thin provisioning, information about aggregate usage (for example, the Used Size, Used Percentage, and Available Size) reflects the actual space used to store data. When you use thick provisioning, aggregate usage information reflects the allocated space, which typically differs from the space that is actually used to store data.

- In some cases, it eliminates the need to change the volume size after adding disks.
  If you add more disks to existing aggregates, you do not have to resize the Infinite Volume to make use of the added capacity as long as the total size of the Infinite Volume's associated aggregates is less than the Infinite Volume's size.

Disadvantages of thin provisioning

Thin provisioning includes the following disadvantages:

- If you have overcommitted your aggregate, you must monitor your available space and add storage to the aggregate as needed to avoid write errors due to insufficient space.
- In a multi-tenancy environment, if you share aggregates among volumes that use thin provisioning, be aware that one tenant's aggregate space availability can be adversely affected by the growth of another tenant's volumes.
- The process of balancing incoming files across data constituents is less effective when an Infinite Volume uses thin provisioning because the reported percentage of used space does not always represent the physical used space.

Space considerations for using incremental tape backup

If you want to use incremental tape backup with an Infinite Volume, you should be aware that up to 10 TB of space per node is required starting with the third node that an Infinite Volume uses.

Incremental tape backup requires you to enable SnapDiff, which automatically creates a namespace mirror constituent on each node that the Infinite Volume uses that does not already have either a namespace constituent or a namespace mirror constituent. Because every Infinite Volume already has one namespace constituent and one namespace mirror constituent, SnapDiff requires new namespace mirror constituents on every node beyond two.

Each namespace mirror constituent is the same size as the namespace constituent, which can require up to 10 TB of space.

If you plan to enable SnapDiff eventually, the preferred practice is to enable it when you first create the Infinite Volume.

You can also enable SnapDiff after the Infinite Volume is created. If you enable SnapDiff on an existing Infinite Volume, you must do so in an operation that is separate from any resize operation.

Size requirements for data protection mirror relationships

When the source and destination Infinite Volumes are on platforms that support different maximum data constituent sizes, you must know the maximum data constituent sizes for the different platforms to successfully set up a data protection mirror relationship for Infinite Volumes.

The maximum data constituent size for an Infinite Volume corresponds to the maximum FlexVol volume size for the platform. You must correctly set the maximum data constituent size when you create the source and the destination Infinite Volumes. For example, if you want to create a data protection mirror relationship between a source Infinite Volume on a platform with a large data constituent size and a destination Infinite Volume on a platform with a small data constituent size, the size of the constituents in the source and destination Infinite Volumes are restricted to the smaller maximum data constituent size. You must set the maximum data constituent size for the source and destination Infinite Volumes to be the size of the smaller maximum data constituent size for the two platforms.

You should set the maximum data constituent size when you create the source and destination Infinite Volumes. You can modify the maximum data constituent size for an Infinite Volume. However, you cannot use the setting to shrink the size of existing data constituents. If the current data constituent size is larger than the size that you want to specify, the setting cannot shrink the existing data constituents to the smaller size.

8.3.4 Storage classes and data policies

You can create an Infinite Volume with one or more storage classes and use a data policy with rules to automatically filter incoming data into different storage classes. You should understand what storage classes and data polices are and when they are useful for Infinite Volumes.

What a storage class is

A storage class is a definition of aggregate characteristics and volume settings. You can define different storage classes and associate one or more storage classes with an Infinite Volume. You must use OnCommand Workflow Automation to define workflows for your storage class needs and to assign storage classes to Infinite Volumes.

You can define the following characteristics for a storage class:

- Aggregate characteristics, such as the type of disks to use
- Volume settings, such as compression, deduplication, and volume guarantee

For example, you can define a performance storage class that uses only aggregates with SAS disks and the following volume settings: thin provisioning with compression and deduplication enabled.

How storage classes affect which aggregates for Infinite Volumes

Each storage class definition specifies an aggregate type. When you create an Infinite Volume with a storage class, only the type of aggregate specified for the storage class can supply storage for the volume. You must understand storage class definitions to create aggregates that are appropriate for the storage class.

Storage class definitions are available only in OnCommand Workflow Automation. After you understand the aggregate requirements for each storage class, you can use the command-line interface or OnCommand Workflow Automation to create aggregates for storage classes. However, you must use OnCommand Workflow Automation, not the command-line interface, to create an Infinite Volume with one or more storage classes.

When you use OnCommand Workflow Automation to create an Infinite Volume with a storage class, OnCommand Workflow Automation automatically filters the aggregates available in the cluster based on the storage class that you want to use. If no aggregates meet the requirements of the storage class, you cannot create an Infinite Volume with that storage class.

How storage classes relate to storage services

A storage service is equivalent to a storage class. Data ONTAP uses the term *storage service*, and OnCommand Workflow Automation uses the term *storage class*. If an Infinite Volume uses a storage class, it is considered managed by storage services, and you cannot use some commands for the Infinite Volume.

When you create an Infinite Volume without storage classes, the storage-service label in Data ONTAP is blank. An Infinite Volume without storage classes is considered not managed by storage services, and you can use Data ONTAP commands or OnCommand Unified Manager to monitor and manage the capacity of the Infinite Volume.

Note: You must have diagnostic privilege to view the storage-service label in Data ONTAP.

However, when you create an Infinite Volume with one or more storage classes, the storage-service label in Data ONTAP contains the names of the storage classes, and the Infinite Volume is considered managed by storage services. When an Infinite Volume is managed by storage services, some commands are disabled. Instead of using the commands, you must use OnCommand Workflow Automation to manage the storage classes, and you must use OnCommand Unified Manager to monitor and manage the capacity of the storage classes in the Infinite Volume. When you try to use a disabled command for an Infinite Volume with storage classes, an error message is displayed.

Note: When you upgrade an Infinite Volume from Data ONTAP 8.1.1 and later to Data ONTAP 8.2, the Infinite Volume automatically contains an unnamed storage class. The unnamed storage class is not managed by storage services.

Rules and data policies

A rule determines the placement of files (data) in an SVM with Infinite Volume. A collection of such rules is known as a data policy.

Rule

Rules mainly consist of a set of predefined conditions and information that determine where to place files in the Infinite Volume. When a file is placed in the Infinite Volume, the attributes of that file are matched with the list of rules. If attributes match the rules, then that rule's placement information determines the storage class where the file is placed. A default rule in the data policy is used to determine the placement of files if the attributes do not match any of the rules in the rule list.

For example, if you have a rule, “Place all files of type.mp3 in the bronze storage class.”, all .mp3 files that are written to the Infinite Volume would be placed in the bronze storage class.

Data policy

A data policy is a list of rules. Each SVM with Infinite Volume has its own data policy. Each file that is added to the Infinite Volume is compared to its data policy's rules to determine where to place that file. The data policy enables you to filter incoming files based on the file attributes and place these files in the appropriate storage classes.

Default data policy

A data policy with a default rule is automatically created for an SVM with Infinite Volume when you create the Infinite Volume. The data policy is active and contains a default rule that automatically filters data written to the Infinite Volume. Data policies are useful when you assign two or more storage classes to an Infinite Volume.

The default rule places incoming data as follows for Infinite Volumes with and without storage classes. See Table 8-13.

Table 8-13 Default data policy settings

For an Infinite Volume...	The default rule does this...
Without storage classes	Places all incoming date in the Infinite Volume.
With one storage class	Places all incoming data into the storage class.
With one or more storage classes	Places all incoming data into the first storage class created.

Important: When using more than one storage class, you should modify the data policy as soon as possible to create rules that filter different types of data into the different storage classes. You should modify the data policy by using OnCommand Unified Manager.

You can modify the data policy to create additional rules, but you cannot delete the data policy or its default rule. Changes made to a data policy affect incoming data, not existing data. After data is stored in a storage class in an Infinite Volume, you cannot move the data to another storage class.

Interoperability storage classes and Data ONTAP features

When you use Infinite Volumes with storage classes, you should be aware of how storage classes affect Data ONTAP features and what tools you can use to create and manage Infinite Volumes with storage classes.

Storage classes and client access to Infinite Volumes

Client access is configured in the same way for an Infinite Volume regardless of whether it uses storage classes. However, if an Infinite Volume uses storage classes, you must also create any users and directories that you plan to use in the Infinite Volume's data policy.

An Infinite Volume with storage classes supports the same client access and requires the same client access configuration as an Infinite Volume that does not use storage classes.

You can configure client access to an Infinite Volume with storage classes by using either the command-line interface or OnCommand Workflow Automation. Even if you use OnCommand Workflow Automation to create an Infinite Volume and configure client access, you can further configure client access by using the command-line interface.

When you configure client access to an Infinite Volume that uses storage classes, you must configure the following functionality before you import or activate the Infinite Volume's data policy:

- Any users that you plan to use in rules that filter data based on file owner
- Any directories that you plan to use in rules that filter data based on directory path

For more information about data policies, see the OnCommand Unified Manager Online Help.

Storage classes and Snapshot copies of Infinite Volumes

Snapshot copy creation and management is the same for an Infinite Volume with or without storage classes. You must create Snapshot copies for the entire Infinite Volume. You cannot create Snapshot copies for individual storage classes in Infinite Volumes.

Storage classes and data protection mirror relationships for Infinite Volumes

Setup of data protection mirror relationships differs between Infinite Volumes with and without storage classes. For Infinite Volumes without storage classes, you can use the command-line interface, but for Infinite Volumes with storage classes, you cannot use the command-line interface; you must use OnCommand Workflow Automation instead.

Storage classes affect data protection mirror relationships as follows for Infinite Volumes with and without storage classes. See Table 8-14.

Table 8-14 Infinite Volumes and data protection mirror relations

For an Infinite Volume...	Operations...	Tools supported
Without storage classes	Create and initialize manage data protection relationships. Increase the size of Infinite Volumes in data protection mirror relationships. Recover from disaster.	Command-line interface
With storage classes	Create and initialize data protection mirror relationships. Increase the size of Infinite Volumes in data protection mirror relationships. Recover from disaster.	onCommand Workflow Automation

How storage classes affect tape backup of Infinite Volumes

Setup and management of incremental tape backup is the same for Infinite Volumes with or without storage classes. You must configure tape backup for the entire Infinite Volume. You cannot configure tape backup for individual storage classes in Infinite Volumes.

8.3.5 Managing data policies for an SVM with Infinite Volume

You should use OnCommand Unified Manager to create and modify rules for a data policy associated with an SVM with Infinite Volume. Commands are supported, but not preferred.

Commands to manage a data policy for an SVM with Infinite Volume

You can use the `vserver data-policy` commands to import, export, and validate a data policy for an SVM with Infinite Volume with storage classes. See Table 8-15.

> **Note:** The `vserver data-policy` commands are not supported for SVMs with FlexVol volumes.

Table 8-15 The vserver data-policy commands

To perform this action...	Use this command...
Export a data policy from an SVM with Infinite Volume in JSON format.	**vserver data-policy export**
Ensure that the JSON format in a data policy is valid before importing the data policy to an SVM with Infinite Volume.	**vserver data-policy validate**
Import a data policy in JSON format to an SVM with Infinite Volume.	**vserver data-policy import**

For detailed information about these commands, see the appropriate man page.

Editing rules in a data policy for an SVM with Infinite Volume

You should use OnCommand Unified Manager to edit rules in a data policy for an SVM with Infinite Volume. Changes made to rules in a data policy affect only incoming data and not data that is already stored in the Infinite Volume.

About this task

You can also use the **`vserver data-policy`** command to edit rules in a data policy, but you must work with the data policy and its rules in Java Script Object Notation (JSON) format when you use the **`vserver data-policy`** command.

Steps

Follow these steps:

1. Edit the rules in a data policy and activate the changes by using OnCommand Unified Manager. In OnCommand Unified Manager, you can choose when to activate the changes made to the data policy.
2. If the Infinite Volume is in a data protection mirror relationship, export the updated data policy from the source SVM with Infinite Volume, and import the data policy to the destination SVM with Infinite Volume by using OnCommand Unified Manager. Data policies on the source and destination Infinite Volumes are identical.

Importing a data policy to an SVM with Infinite Volume

You can manually import a data policy in Java Script Object Notation (JSON) format into an SVM with Infinite Volume to ensure that the SVM with Infinite Volume has the latest data policy. The imported data policy immediately starts filtering data written to the Infinite Volume.

Before you begin

Observe these considerations:

- The Infinite Volume contained by the SVM must be mounted.
- The data policy must be in JSON format, and the JSON must be valid as defined for data policies.
- The data policy must contain a rule named default, and the default rule must be the last rule in the data policy.
- If the data policy contains rules that filter data based on directory location, the referenced directories must exist in the Infinite Volume's namespace before you import the data policy.
- If the data policy contains rules that filter data based on users, the user names must exist in the SVM with Infinite Volume or the name service before you import the data policy. For example, if Data ONTAP uses Windows Active Directory authentication to resolve user names, the user name must exist in Windows Active Directory before you import the data policy.

About this task

The entire existing data policy on an SVM with Infinite Volume is replaced when you import a data policy. If the Infinite Volume is in a data protection mirror relationship, you must import the changed data policy into the source and the destination SVMs with Infinite Volume to ensure that both SVMs reference the same data policy.

Steps

Follow these steps:

1. Import the data policy into the SVM with Infinite Volume by using the **`vserver data-policy import`** command.

   The data policy imports and is active for any new data written to the Infinite Volume. The updated data policy does not affect existing data in the Infinite Volume.

2. If the Infinite Volume is in a data protection mirror relationship, import the data policy to the destination SVM with Infinite Volume by using the **`vserver data-policy import`** command.

   The data policies on the source and destination Infinite Volumes are identical.

Considerations for valid JSON formatting in data policies

Data policies are in Java Script Object Notation (JSON) format, and the JSON format must be valid as defined for data policies. You should be aware of what makes JSON formatting valid for data policies.

> **Note:** You should use OnCommand Unified Manager to edit data policies. OnCommand Unified Manager uses valid JSON formatting for data polices.

Consider the following criteria when working with data policies in JSON format:

- A data policy supports a maximum of 100 rules.
- A rule in a data policy supports a maximum of 50 conditions.
- The **`storageservice`** key identifies the name of the storage class.
  You must use OnCommand Workflow Automation to create storage classes that data policies reference.

  > **Note:** The **`storageservice`** key contains a dash (-) when you create an Infinite Volume without storage classes, or when you upgrade an Infinite Volume created in Data ONTAP 8.1.1 and later to Data ONTAP 8.2.x.

- The **`parentURI`** key must end in a forward slash /.
- The **`parentURI`** key supports scope matching with ==, !=, starts, and !starts.

  > **Note:** The **`parentURI`** key does not support ends /! ends.

- The **`metadata/cdmi_owner`** key identifies the owner of a file.
  For example, the CIFS protocol uses a user name to identify the owner of a file.
- The **`metadata/cdmi_owner`** key supports scope matching with == and !=.
- The **`metadata/cdmi_owner`** key must contain a user name that is equal to or less than 192 characters.

Example of JSON data policy: Filtering based on directory

A data policy in Java Script Object Notation (JSON) format can filter data written to the Infinite Volume into different storage classes based on the directory location of the file.

The data policy filters data as follows:

1. If the file is created in or beneath /NS/users/alice, store the data in the gold storage class.
2. If the file is created in and not beneath /NS/users/bob/important, store the data in the gold storage class.
3. If the file is created in or beneath /NS/users/bob, store the data in the silver storage class.
4. Otherwise, store the data in the bronze storage class.

Example 8-2 shows the JSON for the data policy for directory based filtering.

Example 8-2 JSON data policy example, directory based filtering

```
{
   "ruleset_format_version" : "1.0",
   "rules" : [
      {
         "rule_label" : "Alice's stuff",
         "rule_id" : "63cd19f1-74d4-4754-8aca-e0223f6c3923",
         "rule_scope" : [
            { "parentURI" : "starts /NS/users/alice/" }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "gold"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "Bob's important stuff",
         "rule_id" : "792caf8d-b0a1-4881-8529-13daa6b2f49c",
         "rule_scope" : [
            { "parentURI" : "== /NS/users/bob/important/" }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "gold"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "Bob's stuff",
         "rule_id" : "7c007dea-60ee-41d2-9f64-88e18637941c",
         "rule_scope" : [
            { "parentURI" : "starts /NS/users/bob/" }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "silver"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "default",
         "rule_id" : "00fc0534-10dd-7812-ec8c-a9356fa8cd00",
         "rule_scope" : [ ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "bronze"
                  }
               }
            }
         }
      }
   ]
}
```

Example of JSON data policy: File name based filtering

A data policy in JSON format can filter data written to the Infinite Volume into different storage classes based on file name.

The data policy filters data as follows:

1. If the file name ends in .doc or .xls, store its data in the gold storage class.
2. If the file name ends in .mp3, .wav, or .ogg, store its data in the silver storage class.
3. Otherwise, store its data in the bronze storage class.

Example 8-3 shows the JSON for the data policy for file name based filtering.

Example 8-3 JSON data policy example, file name based filtering

```
{
   "ruleset_format_version" : "1.0",
   "rules" : [
      {
         "rule_label" : "office stuff",
         "rule_id" : "63cd19f1-74d4-4754-8aca-e0223f6c3923",
         "rule_scope" : [
            { "objectName" : "ends .doc" },
            { "objectName" : "ends .xls" }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "gold"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "media stuff",
         "rule_id" : "7c007dea-60ee-41d2-9f64-88e18637941c",
         "rule_scope" : [
            { "objectName" : "ends .mp3" },
            { "objectName" : "ends .ogg" },
            { "objectName" : "ends .wav" }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "silver"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "default",
         "rule_id" : "00fc0534-10dd-7812-ec8c-a9356fa8cd00",
         "rule_scope" : [ ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "bronze"
                  }
               }
            }
         }
      }
   ]
}
```

Example of JSON data policy: File owner based filtering

A data policy in Java Script Object Notation (JSON) format can filter data written to the Infinite Volume into different storage classes based on the owner of the file.

The data policy filters data as follows:

1. If Alice owns the file, store the data in the gold storage class.
2. If Bob owns the file, store the data in the silver storage class.
3. If users other than Alice and Bob own the file, store the data in the bronze storage class.
4. Otherwise, store the data in the bronze storage class.

Example 8-4 shows the JSON for the data policy for file owner based filtering.

Example 8-4 JSON data policy example, file owner based filtering

```
{
   "ruleset_format_version" : "1.0",
   "rules" : [
      {
         "rule_label" : "Alice's stuff",
         "rule_id" : "63cd19f1-74d4-4754-8aca-e0223f6c3923",
         "rule_scope" : [
            { "metadata" : { "cdmi_owner" : "== alice" } }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "gold"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "Bob's stuff",
         "rule_id" : "7c007dea-60ee-41d2-9f64-88e18637941c",
         "rule_scope" : [
            { "metadata" : { "cdmi_owner" : "== bob" } }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "silver"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "If not user1, user2, or user3",
         "rule_id" : "963fcb66-1d1c-464e-b08a-6b8b473fac8a",
         "rule_scope" : [
            { "metadata" : { "cdmi_owner" : "!= user1" } },
            { "metadata" : { "cdmi_owner" : "!= user2" } },
            { "metadata" : { "cdmi_owner" : "!= user3" } }
         ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "bronze"
                  }
               }
            }
         }
      },
      {
         "rule_label" : "default",
         "rule_id" : "00fc0534-10dd-7812-ec8c-a9356fa8cd00",
         "rule_scope" : [ ],
         "rule_epoch" : { "epoch_reference" : "cdmi_ctime" },
         "rule_epochs" : {
            "0" : {
               "local" : {
                  "metadata" : {
                     "storageservice" : "bronze"
                  }
               }
            }
         }
      }
   ]
}
```

8.3.6 Infinite Volume constituents

Each Infinite Volume consists of several separate components called constituents. Understanding how these constituents are created and the roles they play can help you plan aggregates for an Infinite Volume, understand its capacity, and interpret the results of operations performed on the Infinite Volume.

Constituents are internal to the Infinite Volume and are not visible to clients as they access the volume. You can display the constituents from the CLI if you substitute the **`-is-constituent true`** parameter for the **`-volume`** parameter in certain show commands. When you use the **`-is- constituent true`** parameter, the output displays the same information for constituents that is displayed for volumes.

Roles of constituents in an Infinite Volume

Constituents play one of the following roles:

- Data constituents, which store data
- Namespace constituent, which tracks file names and directories and the file's physical data location
- Namespace mirror constituents, which are data protection mirror copies of the namespace constituent

Figure 8-2 illustrates the roles of the namespace and data constituents.

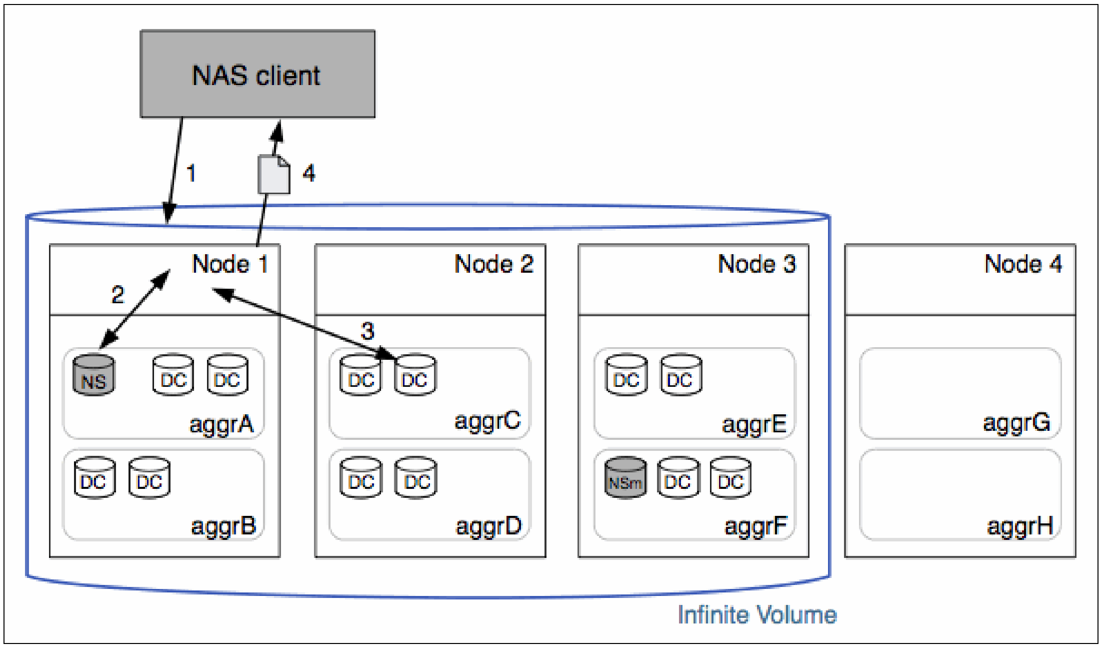


Figure 8-2 Constituents in an Infinite Volume

When a client first retrieves a file from an Infinite Volume, the node goes to the namespace constituent (NS) to look up which data constituent (DC) contains the file's data. The node then retrieves the file's data from the data constituent, sends the file to the client, and caches the file's location to enable the node to more quickly satisfy subsequent requests.

Namespace constituent

Each Infinite Volume has a single namespace constituent that maps directory information and file names to the file's physical data location within the Infinite Volume.

The namespace constituent is essential to the operation of the Infinite Volume because nodes use the information in the namespace constituent to locate file data requested by clients.

Clients are not aware of the namespace constituent and do not interact directly with it. The namespace constituent is an internal component of the Infinite Volume.

Data constituents

In an Infinite Volume, data is stored in multiple separate data constituents. Data constituents store only the data from a file, not the file's name.

Clients are not aware of data constituents. When a client requests a file from an Infinite Volume, the node retrieves the file's data from a data constituent and returns the file to the client.

Each Infinite Volume typically has dozens of data constituents. For example, a 6 PB Infinite Volume that contains 1 billion files might have 60 data constituents located on aggregates from 6 nodes.

Namespace mirror constituent

A namespace mirror constituent is an intracluster data protection mirror copy of the namespace constituent in an Infinite Volume. The namespace mirror constituent performs two roles. It provides data protection of the namespace constituent, and it supports SnapDiff for incremental tape backup of Infinite Volumes.

When namespace mirror constituents are created

When you create an Infinite Volume, one namespace mirror constituent is automatically created to provide data protection for the namespace constituent. When you enable SnapDiff for an Infinite Volume, additional namespace mirror constituents are automatically created for SnapDiff for incremental tape backup of Infinite Volumes.

When you create a read/write Infinite Volume that spans two or more nodes in a cluster, one namespace mirror constituent is automatically created, and a data protection mirror relationship is automatically created between the namespace constituent and the namespace mirror constituent. The data protection mirror relationship is updated every five minutes. The data protection mirror relationship is an automatic process for an Infinite Volume. You cannot use SnapMirror commands to modify or manage the data protection mirror relationship between the namespace constituent and the namespace mirror constituent.

When you enable SnapDiff on an Infinite Volume that spans three or more nodes, additional namespace mirror constituents are automatically created for SnapDiff to use for incremental tape backup of Infinite Volumes. A namespace mirror constituent is created on each node with a data constituent, except the node with the namespace constituent and the node with the namespace mirror constituent that was created to provide data protection for the namespace constituent. Namespace mirror constituents created to support SnapDiff are updated daily or as configured for SnapDiff. A SnapMirror license is not required to enable SnapDiff.

When you create a destination Infinite Volume for a data protection mirror relationship, a namespace mirror constituent is not created on the destination Infinite Volume. However, if you enable SnapDiff on a destination Infinite Volume, namespace mirror constituents are automatically created for use by SnapDiff. You must initialize the data protection mirror relationship between the source and destination Infinite Volumes before you can enable SnapDiff.

Figure 8-3 shows an Infinite Volume in a data protection mirror relationship. SnapDiff is disabled on the source and the destination Infinite Volumes. With SnapDiff disabled, a namespace mirror constituent is created on the source Infinite Volume to provide data protection for the namespace constituent, and no other namespace mirror constituents are created.

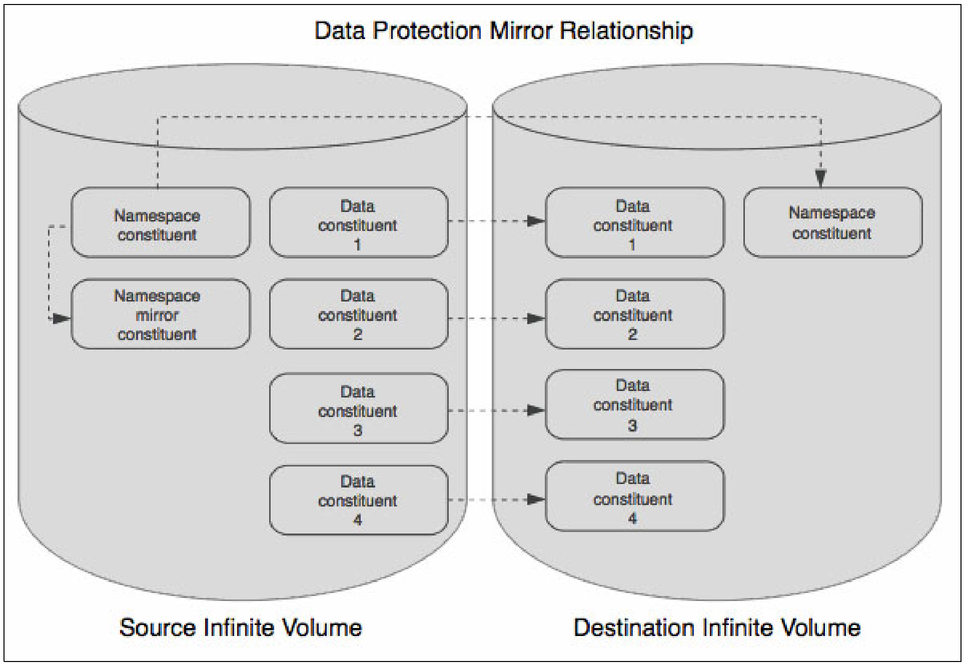


Figure 8-3 Infinite Volume Data Protection Mirror Relationships

8.3.7 Planning aggregates for an Infinite Volume

By learning how an Infinite Volume uses aggregates, you can better prepare aggregates and select them specifically for use by an Infinite Volume.

When you create an Infinite Volume, Data ONTAP automatically selects the best aggregates for each Infinite Volume constituent. While you can control many aspects of how an Infinite Volume uses aggregates, you can also allow Data ONTAP to determine aggregate use.

How aggregates and nodes are associated with Infinite Volumes

The aggregate list of the containing SVM with Infinite Volume determines which aggregates the Infinite Volume uses, as well as who can create an Infinite Volume and which nodes the Infinite Volume uses.

That aggregate list can be specified or unspecified, which is represented as a dash (“-”). By default, when a cluster administrator creates any SVM, its aggregate list is unspecified. After an SVM is created, the cluster administrator can specify the aggregate list by using the **`vserver modify`** command with the **`-aggr-list`** parameter.

Considerations when to specify the aggregate list or leave it unspecified

If you are dedicating an entire cluster to the SVM with Infinite Volume, you can leave the aggregate list of an SVM with Infinite Volume unspecified. In most other situations, you should specify the aggregate list of an SVM with Infinite Volume.

Leaving the aggregate list of an SVM with Infinite Volume unspecified has the following outcomes:

- Only a cluster administrator can create the Infinite Volume, not an SVM administrator.
- When the Infinite Volume is created, it uses all nodes in the cluster.
- When the Infinite Volume is created, it can potentially use all of the aggregates in the cluster.

How the aggregate list contains candidate aggregates

The aggregate list of an SVM with Infinite Volume acts only as a candidate aggregate list for an Infinite Volume. An Infinite Volume uses aggregates according to various factors, including the following requirements:

- When an Infinite Volume is created, at least one data constituent is created on at least one aggregate from each node in the aggregate list.
- An Infinite Volume uses only the aggregates that it requires to meet the capacity requirements for its specified size.

  If the assigned aggregates have far greater capacity than the Infinite Volume requires when it is first created, some aggregates in the aggregate list might not contain any Infinite Volume constituents.

How the aggregate list determines the nodes

An Infinite Volume uses every node that has an aggregate in the aggregate list of an SVM with Infinite Volume.

When changes to the aggregate list take effect

Changes to the aggregate list do not have any immediate effect. The aggregate list is used only when the size of an Infinite Volume changes. For example, if you add an aggregate to the aggregate list of an SVM with Infinite Volume, that aggregate is not used until you modify the size of the Infinite Volume.

If you add aggregates from a new node to the aggregate list and then resize the Infinite Volume, whether the Infinite Volume uses the aggregates from the new node depends on several variables, including the size of existing constituents and how much the Infinite Volume was increased in size.

How the aggregate list can be filtered

The SVMs aggregate list can be filtered by using advanced parameters that control which aggregates are used for each type of constituent, such as data constituents. Unlike the SVMs aggregate list, these aggregate-selection parameters apply only to a single operation. For example, if you use the parameter for data constituent aggregates when you create the Infinite Volume and then resize the Infinite Volume without using the parameter, the Infinite Volume uses the SVM aggregate list.

How much space namespace-related constituents require

The namespace constituent and namespace mirror constituents can each consume up to 10 TB of an Infinite Volume's capacity on two or more nodes, depending on the Infinite Volume's size and configuration. The requirements of these namespace-related constituents affect how space is allocated when you create or expand an Infinite Volume.

In most cases, the namespace constituent is 10 TB and never larger than 10 TB.

However, the namespace constituent is smaller in Infinite Volumes that are relatively small. In small Infinite Volumes, the namespace constituent size is configured so that the combined size of the namespace-related constituents (which includes the namespace constituent and one or more namespace mirror constituents) is equal to 25% of the Infinite Volume's size. For example, if an Infinite Volume is 60 TB and has one namespace mirror constituent, the combined size of the namespace constituent and namespace mirror constituent must be 25% of 60 TB, or 15 TB. That makes the namespace constituent 7.5 TB.

If 25% of the Infinite Volume's size would result in namespace-related constituents being larger than 10 TB, the maximum size takes precedence, and each namespace-related constituent is 10 TB.

How data constituents use aggregate space

Each data constituent is created as large as possible within the following constraints: a hardware- related maximum size, the space available in the containing aggregate, and the requirement to balance capacity across nodes.

Data ONTAP attempts to make every data constituent as big as possible within the following constraints:

- A hardware-related maximum size that is identified in the Hardware Universe.

  If the Infinite Volume uses nodes from multiple platforms, the smallest value is used for all data constituents on all nodes.
- The available space on its containing aggregate.

  The available space on an aggregate is determined by the aggregate's size and the space that is used by other constituents and volumes that share the aggregate. If an aggregate is shared with other SVMs, it can already contain FlexVol volumes, other constituents for the same Infinite Volume, and constituents of other Infinite Volumes.
- The requirement to balance the Infinite Volume's capacity across nodes.

For example, if the maximum data constituent size on a given platform is 100 TB, Data ONTAP attempts to make each data constituent 100 TB.

If data constituents cannot be created at the maximum size, Data ONTAP creates smaller data constituents, to a minimum possible size of 1 TB.

How node balancing affects an Infinite Volume's size and aggregate use

New data constituents of an Infinite Volume are balanced across nodes. This means that the node with the smallest available space determines how much space is used on each node and limits the size of the Infinite Volume that you can create or expand.

What node balancing is

When an Infinite Volume is created or expanded, Data ONTAP ensures that the total size of the volume's data constituents is the same on every node that an Infinite Volume uses. For example, in a 6 PB, six-node Infinite Volume, Data ONTAP attempts to distribute the data constituents so that each node holds 1 PB of data constituents.

The node-balancing requirement means that the size of data constituents on each node is limited by the node with the smallest available space. If one node has only 0.5 PB of available space, every other node used by the Infinite Volume can hold only 0.5 PB of data constituents.

How node balancing affects an Infinite Volume's size

Because the node-balancing requirement controls the amount of data constituents on a node, it also significantly restricts the overall size of the Infinite Volume, both when you first create the volume and when you try to expand it.

For example, if you try to create a 6 PB, six-node Infinite Volume but one of the nodes used by the Infinite Volume has only 0.5 PB of available space, each node can hold only 0.5 PB of data constituents, limiting the total size of the Infinite Volume to approximately 3 PB.

You can roughly determine an Infinite Volume's largest possible size by determining which node has the least amount of available space and multiplying that amount by the number of nodes. You can also determine the largest possible size by running the volume create or volume modify commands; if the requested size cannot be created, the resulting error message indicates the largest possible size of the Infinite Volume given the existing resources.

How node balancing affects the expansion of an Infinite Volume

The node-balancing requirement persists when you expand an Infinite Volume. The source of the new capacity does not necessarily determine where data capacity is placed. Data ONTAP adds data capacity to nodes in a way that equalizes the total size of all data constituents on each node.

In most cases, new data constituents are created. In some circumstances, such as an Infinite Volume that is upgraded from Data ONTAP 8.1, existing data constituents might be expanded evenly until one of them reaches the maximum data constituent size.

8.4 Storage limits

There are limits for storage objects that you should consider when planning and managing your storage architecture.

Limits are listed in the following sections:

- Volume limits are provided here in Table 8-16.
- FlexClone file and FlexClone LUN limits are provided in Table 8-17 on page 130.

Table 8-16 Volume limits

Limit	Native Storage	Notes
Files Maximum size	16TB	
Files Maximum per volume	Volume size dependent, up to 2 billion	2 billion = 2 x 10 to the 9th power. Directories are also counted against this limit
FlexCache volumes Maximum per node	100	for more information see 16.1, “FlexCache” on page 274
FlexClone volumes Hierarchical clone depth	499	The maximum depth of a cested hierarchy of FlexClone volumes that can be created from a single FlexVol volume.
FlexVol volumes Maximum per node	Model-dependent	
FlexVol volumes Maximum per node per SVM	N3150: 200 All other platforms: 500	This limit applies only in SAN environments
FlexVol volumes Minimum size	20 MB	
FlexVol volumes (32-Bit) Maximum size	16TB	

Limit	Native Storage	Notes
FlexVol volumes (64-bit) Maximum size	Model-dependent	
FlexVol node root volumes Minimum size	Model-dependent	
Infinite Volumes Minimum size	1.33 TB * number of nodes used	This size is approximate due to rounding
Infinite Volumes Maximum size	20 PB	
Infinite Volumes Maximum number of nodes	10	This limit is the maximum number of nodes upon which the aggregates associated with an Infinite Volume can be located. The maximum number of nodes in the cluster is unaffected by the presence of Infinite Volumes.
LUNs Maximum per node	► N3150: 200 ► N3220: 2,048 ► N3240: 2,048 ► N6040: 2,048 ► N6060: 2,048 ► N6210: 2,048 ► N6220: 2,048 ► N6240: 2,048 ► All other models: 8,192	
LUNs Maximum per cluster	► N3150: 400 ► N3220: 8,192 ► N3240: 8,192 ► N6040: 8,192 ► N6060: 8,192 ► N6210: 8,192 ► N6220: 8,192 ► N6240: 8,192 ► All other models: 49,152	
LUNs Maximum per volume	► N3150: 200 ► All other models: 512	
LUNs Maximum size	16 TB	
Qtrees Maximum per FlexVol volume	4,995	
Snapshot copies Maximum per volume	255	The use of certain Data ONTAP capabilities could reduce this limit.
Aggregates Maximum per node	100	In an HA configuration, this limit applies to each node individually, so the overall limit for the pair is doubled.

Limit	Native Storage	Notes
Volumes Maximum per cluster for NAS	12,000	Infinite Volumes do not count against this limit, but their constituent volumes do.
Volumes Maximum per cluster with SAN protocols configured	► N3150: 400 ► All other models: 600	Infinite Volumes do not count against this limit, but their constituent volumes do.

Table 8-17 FlexClone file and FlexClone LUN limits

Limit	Native storage	Notes
Maximum per file or LUN	32,767	If you try to create more than 32,767 clones, Data ONTAP automatically creates a new physical copy of the parent file or LUN. This limit might be lower for FlexVol volumes that use deduplication.
Maximum total shared data per FlexVol volume	640 TB	

9

Networking

This chapter describes the networking components of a cluster. You need to understand how to configure networking components of the cluster during and after setting up the cluster.

The following topics are covered:

- Networking components
- Network ports
- Interface groups
- VLANs
- Failover groups
- Logical interfaces
- Routing groups

9.1 Networking components

In Clustered Data ONTAP, the physical networking components of a cluster are abstracted and virtualized into logical components to provide the flexibility and multi-tenancy in a storage virtual machine (SVM).

These are the various networking components in a cluster:

- Ports:
  - Physical ports: Network interface cards (NICs) and HBAs provide physical (Ethernet and Fibre Channel) connections to the physical networks (management and data networks).
  - Virtual ports: VLANs and interface groups (ifgrps) constitute the virtual ports. While interface groups treat several physical ports as a single port, VLANs subdivide a physical port into multiple separate ports.
- Logical interfaces (LIFs):
  - An LIF is an IP address and is associated with attributes such as failover rule lists, firewall rules. An LIF communicates over the network through the port (physical or virtual) it is currently bound to.

    Different types of LIFs in a cluster are data LIFs, cluster management LIFs, node management LIFs, intercluster LIFs, and cluster LIFs. The ownership of the LIFs depends on the SVM where the LIF resides. Data LIFs are owned by data SVM, node management, and cluster LIFs are owned by node SVM, and cluster management LIFs are owned by admin SVM.
- Routing groups:
  - A routing group is a routing table. Each LIF is associated with a routing group and uses only the routes of that group. Multiple LIFs can share a routing group. Each routing group needs a minimum of one route to access clients outside the defined subnet.
- DNS zones:
  - DNS zone can be specified during the LIF creation, providing a name for the LIF to be exported through the cluster's DNS server. Multiple LIFs can share the same name, allowing the DNS load balancing feature to distribute IP addresses for the name according to load. An SVM can have multiple DNS zones.

Figure 9-1 shows the interaction of ports, VLANs, interface groups, and LIFs regarding the whole cluster.

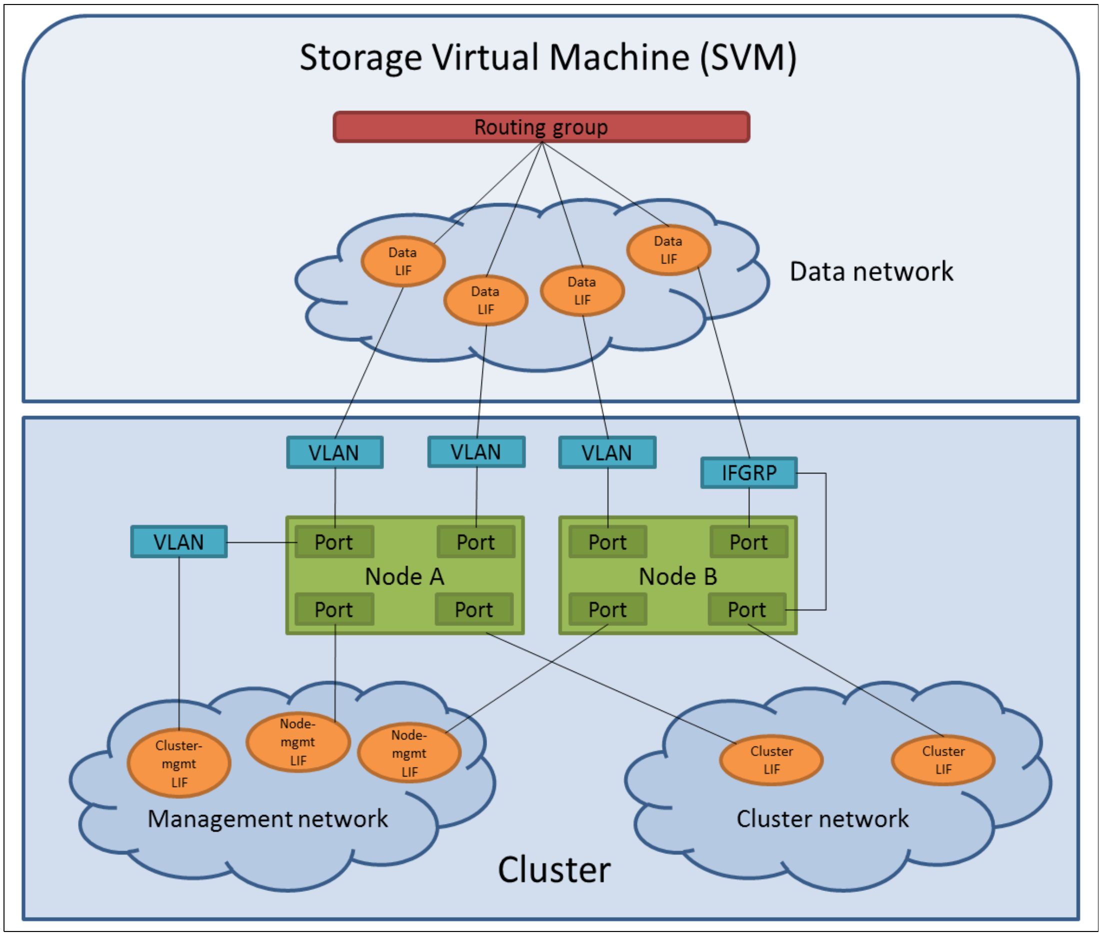


Figure 9-1 Interaction of network objects

9.2 Network ports

Ports are either physical ports (NICs) or virtualized ports, such as interface groups (see 9.3, “Interface groups” on page 135) or VLANs (see 9.4, “VLANs” on page 136). An LIF communicates over the network through the port to which it is currently bound.

Network ports can have roles that define their purpose and their default behavior. Port roles limit the types of LIFs that can be bound to a port. Network ports can have four roles: *node management*, *cluster*, *data*, and *intercluster*. Each network port has a default role. You can modify the roles for obtaining the best configuration. See Table 9-1, which describes the port roles more precisely.

Table 9-1 Network port roles

Role	Description
Node management ports	The ports used by administrators to connect to and manage a node. These ports can be VLAN-tagged virtual ports where the underlying physical port is used for other traffic. The default port for node management differs depending on hardware platform. Some platforms have a dedicated management port (e0M). The role of such a port cannot be changed, and these ports cannot be used for data traffic.
Cluster ports	The ports used for intracluster traffic only. By default, each node has two cluster ports on 10-GbE ports enabled for jumbo frames. You cannot create VLANs or interface groups on cluster ports.
Data ports	The ports used for data traffic. These ports are accessed by NFS, CIFS, FC, and iSCSI clients for data requests. Each node has a minimum of one data port. You can create VLANs and interface groups on data ports. VLANs and interface groups have the data role by default, and the port role cannot be modified.
Intercluster ports	The ports used for cross-cluster communication. An intercluster port should be routable to another intercluster port or the data port of another cluster. Intercluster ports can be on physical ports or virtual ports.

Modifying network port attributes

You can modify the MTU, autonegotiation, duplex, flow control, and speed settings of a physical network or interface group. You can modify only the MTU settings and not other port settings of a VLAN.

You should not modify the following characteristics of a network port:

- The administrative settings of either the 10-GbE or the 1-GbE network interfaces.

  The values that you can set for duplex mode and port speed are referred to as administrative settings. Depending on network limitations, the administrative settings can differ from the operational settings (that is, the duplex mode and speed that the port actually uses).
- The administrative settings of the underlying physical ports in an interface group.

  > **Note:** Use the **`-up-admin`** parameter (available at advanced privilege level) to modify the administrative settings of the port.

- The MTU size of the management port, e0M.
- The MTU size of a VLAN cannot exceed the value of the MTU size of its base port.

See Example 9-1 about how to use the **`network port modify`** command to modify the attributes of a network port. The example shows how to disable the flow control on port e0b by setting it to **`none`**.

> **Note:** Set the flow control of all ports to none. By default, the flow control is set to full.

Example 9-1 Network port modify

```
cdot-cluster01::> network port modify -node nodeA -port e0b -flowcontrol-admin
none
```

9.3 Interface groups

An interface group is a feature in Clustered Data ONTAP that implements link aggregation on your storage system. Interface groups provide a mechanism to group together multiple network ports (links) into one logical interface (aggregate).

After an interface group is created, it is indistinguishable from a physical network interface. Clustered 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 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'

Note: Interface groups cannot be created from other interface groups or VLANs.

9.3.1 Creating interface groups

Use the **`network port ifgrp create`** command to create an interface group *(ifgrp)*.

Interface groups must be named using the syntax `a<number><letter>`. For example, `a0a`, `a0b`, `a1c`, and `a2a` are valid interface group names. For more information about this command, see the man page.

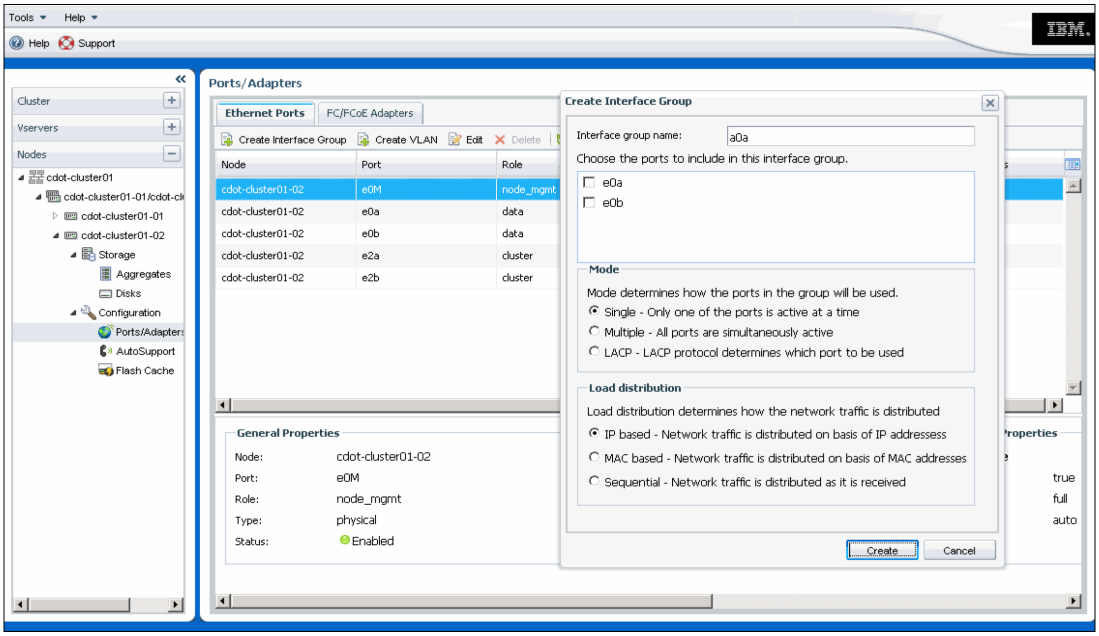
Example 9-2 shows how to create and add ports to an interface group named `a0a` with a distribution function of `ip` and a mode of `multimode`.

Example 9-2 Create interface group

```
cdot-cluster01::> net port show
cdot-cluster01::> network port ifgrp create -node nodeA -ifgrp a0a -distr-func ip
-mode multimode
cdot-cluster01::> network port ifgrp add-port -node nodeA -ifgrp a0a -port e4a
cdot-cluster01::> network port ifgrp add-port -node nodeA -ifgrp a0a -port e4b
```

You can also use the IBM N series OnCommand System Manager to create and edit interface groups. See Figure 9-2 for an example. Follow these steps:

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Nodes** → **Choose the desired node** → **Configuration** → **Ports/Adapters**.
3. Click the **Create Interface Group** button to create the interface group and add ports to it.



Node	Port	Role
cdot-cluster01-02	e0M	node_mgmt
cdot-cluster01-02	e0a	data
cdot-cluster01-02	e0b	data
cdot-cluster01-02	e2a	cluster
cdot-cluster01-02	e2b	cluster

General Properties

Node: cdot-cluster01-02
Port: e0M
Role: node_mgmt
Type: physical
Status: Enabled

true
full
auto

Create Interface Group

Interface group name: a0a

Choose the ports to include in this interface group.

□ e0a
□ e0b

Mode

Mode determines how the ports in the group will be used.

● Single - Only one of the ports is active at a time
○ Multiple - All ports are simultaneously active
○ LACP - LACP protocol determines which port to be used

Load distribution

Load distribution determines how the network traffic is distributed

● IP based - Network traffic is distributed on basis of IP addresses
○ MAC based - Network traffic is distributed on basis of MAC addresses
○ Sequential - Network traffic is distributed as it is received

Figure 9-2 Configure interface groups in system manager

9.3.2 Deleting interface groups

You can delete interface groups if you want to configure LIFs directly on the underlying physical ports or if you decide to change the interface group mode or distribution function. You cannot delete an interface group that has an LIF bound to it.

Use the **`network port ifgrp delete`** command to delete an interface group as shown in Example 9-3.

Example 9-3 Delete interface group

```
cdot-cluster01::> network port ifgrp delete -node cdot-cluster01-02 -ifgrp a0a
```

9.4 VLANs

A VLAN, also as in 7-Mode, is a virtual port that receives and sends VLAN-tagged (IEEE 802.1Q standard) traffic. VLAN port characteristics include the VLAN ID for the port. The underlying physical port or interface group ports are considered to be VLAN trunk ports, and the connected switch ports must be configured to trunk the VLAN IDs.

You can configure an IP address for an interface with VLANs. Any untagged traffic goes to the base interface and the tagged traffic goes to the respective VLAN.

You can configure an IP address for the base interface (physical port) of the VLAN. Any tagged frame is received by the matching VLAN interface. Untagged traffic is received by the native VLAN on the base interface.

Note: You should not create a VLAN on a network interface with the same identifier as the native VLAN of the switch. For example, if the network interface e0b is on native VLAN 10, you should not create a VLAN `e0b-10` on that interface.

You cannot bring down the base interface that is configured to receive tagged and untagged traffic. You must bring down all VLANs on the base interface before you bring down the interface. However, you can delete the IP address of the base interface.

9.4.1 Creating a VLAN

You can create a VLAN for maintaining separate broadcast domains within the same network domain by using the **network port vlan create** command. Before you begin, make sure that the following requirements are met:

- The switches deployed in the network either comply with IEEE 802.1Q standards or have vendor-specific implementation of VLANs.
- For supporting multiple VLANs, an end-station is statically configured to belong to one or more VLANs.

Note: You cannot create a VLAN on cluster management and node management ports. If you want to use a cluster management LIF on a VLAN port, that should have the data role.

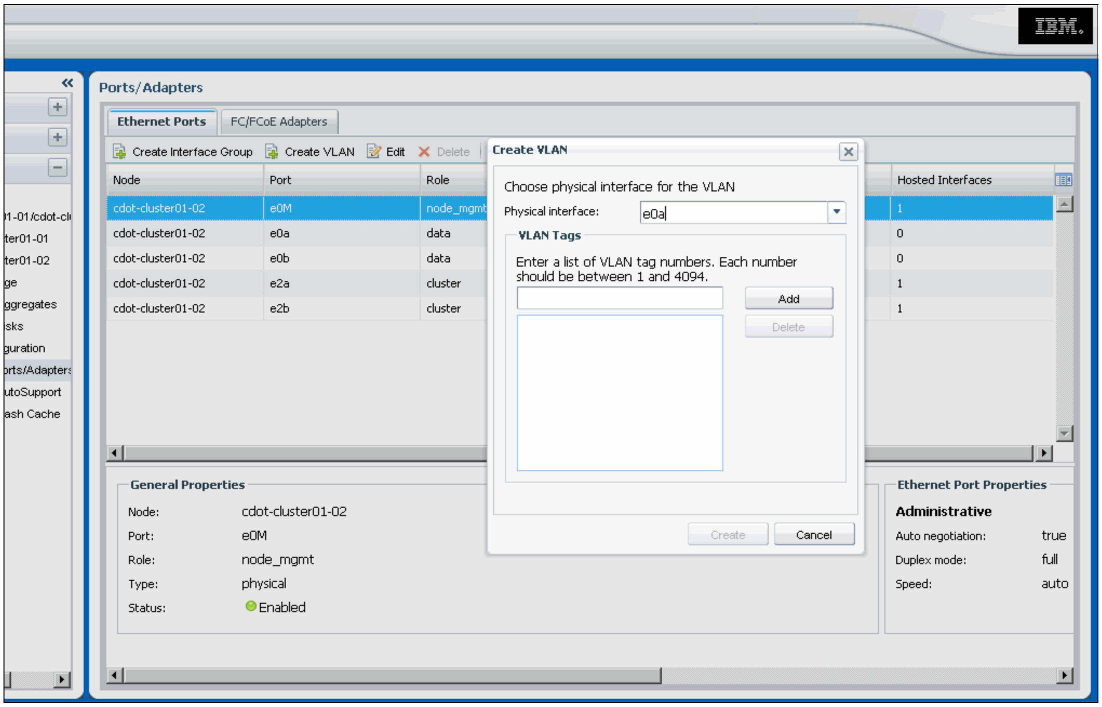
Example 9-4 shows how to create a VLAN a0a-1051 attached to network port a0a on the node nodeA.

Example 9-4 Create VLAN

```
cluster1::> network port vlan create -node nodeA -port a0a -vlan-id 1051
cluster1::> net port vlan show -node nodeA -port a0a
  (network port vlan show)
                 Network Network
Node   VLAN Name Port    VLAN ID  MAC Address
------ --------- ------- -------- -----------------
nodeA
       a0a-1051
                 a0a     1051     02:a0:98:1a:2a:48
```

You can also use the IBM N series OnCommand System Manager to create VLANs. See Figure 9-3 for an example. Follow these steps:

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Nodes** → **Choose the desired node** → **Configuration** → **Ports/Adapters**.
3. Click the **Create VLAN** button to create the VLAN interface.



Node	Port	Role	Hosted Interfaces
cdot-cluster01-02	e0M	node_mgmt	1
cdot-cluster01-02	e0a	data	0
cdot-cluster01-02	e0b	data	0
cdot-cluster01-02	e2a	cluster	1
cdot-cluster01-02	e2b	cluster	1

Choose physical interface for the VLAN

Physical interface: e0a

Enter a list of VLAN tag numbers. Each number should be between 1 and 4094.

Node: cdot-cluster01-02
Port: e0M
Role: node_mgmt
Type: physical
Status: Enabled

Auto negotiation: true
Duplex mode: full
Speed: auto

Figure 9-3 Create VLANs in system manager

9.4.2 Deleting a VLAN

You can delete a VLAN with the **`network port vlan delete`** command. When you delete a VLAN, it is automatically removed from all failover rules and groups that use it. You cannot delete a VLAN that has an LIF bound to it.

Example 9-5 deletes the VLAN `e0a-1065` from network port `e0a` on node `cdot-cluster01-02.`

Example 9-5 Delete VLAN

```
cdot-cluster01::> network port vlan delete -node cdot-cluster01-02 -vlan-name
e0a-1065
```

9.5 Failover groups

LIF failover in Clustered Data ONTAP refers to the automatic migration of an LIF in response to a link failure on the LIF's current network port. When such a port failure is detected, the LIF is migrated to a working port.

A failover group contains a set of network ports (physical, VLANs, and interface groups) on one or more nodes. An LIF can subscribe to a failover group. The network ports that are present in the failover group define the failover targets for the LIF.

You can manage failover groups by adding ports to them, removing ports from them, renaming them, and displaying information about them.

Failover groups for LIFs can be **`system-defined`** or **`user-defined`**. Additionally, a failover group called **`clusterwide`** exists and is maintained automatically. See Table 9-2, which explains the different types.

Table 9-2 Failover group types

Failover group type	Notes
`system-defined`	Failover groups that automatically manage LIF failover targets on a per-LIF basis. This is the default failover group for data LIFs in the cluster. For example, when the value of the failover-group option is `system-defined`, the system will automatically manage the LIF failover targets for that LIF, based on the home node or port of the LIF.
`user-defined`	Failover groups that automatically manage LIF failover targets on a per-LIF basis. This is the default failover group for data LIFs in the cluster. For example, when the value of the failover-group option is system-defined, the system will automatically manage the LIF failover targets for that LIF, based on the home node or port of the LIF.
`clusterwide`	Failover group that consists of all the data ports in the cluster. This is the default failover group for the cluster management LIFs only. For example, when the value of the failover-group option is cluster-wide, every data port in the cluster will be defined as the failover targets for that LIF.

Note: It is preferred practice to create user-defined failover groups for every network type that is existing in the cluster. Create a failover-group for all ports, VLANs, or interface groups that are connected for example to the accounting network and create another failover group that contains all ports, VLANs, or interface groups that are connected to the storage network. Make sure that in a case of failover, each interface that is part of the same failover group can access the same network segment in order to serve data to the clients. Failover groups do not apply in a SAN iSCSI or FC environment.

9.5.1 Creating failover groups

Use the **`network interface failover-groups create`** command to create a failover group or add a port to an existing failover group like shown in Example 9-6.

Example 9-6 Create failover groups

```
cluster1::> network interface failover-groups create -failover-group storage-lan
-node nodeA -port e1b
cluster1::> network interface failover-groups create -failover-group storage-lan
-node nodeA -port e1b
cluster1::> network interface failover-groups show
```

Tip: You might have to check whether the failover rules of an LIF are configured correctly. In order to prevent mis-configuration of the failover rules, you can view the failover target for an LIF or all LIFs.

Use the **failover** option of the **network interface show** command to view the failover targets of an SVMs LIFs as shown in Example 9-7.

Example 9-7 Display information about failover targets

```
cdot-cluster01::> net int show -failover -vserver vs_cifs_01
  (network interface show)
        Logical        Home                 Failover       Failover
Vserver  Interface      Node:Port            Policy         Group
-------- -------------- -------------------- -------------- --------------
vs_cifs_01
         vs_cifs_01_cifs_lif1 cdot-cluster01-02:e0a
                                             nextavail      storage-lan
                        Failover Targets: cdot-cluster01-02:e0a,
                                          cdot-cluster01-02:e0b,
                                          cdot-cluster01-01:e0a,
                                          cdot-cluster01-01:e0b
```

9.5.2 Deleting failover groups

To remove a port from a failover group or to delete an entire failover group, you use the **network interface failover-groups delete** command.

Example 9-8 shows how to delete port e0b of node cdot-cluster01-02 from a failover group named storage-lan.

Example 9-8 Delete port from failover group

```
cdot-cluster01::> net int failover-groups delete -failover-group storage-lan -node
cdot-cluster01-02 -port e0b
```

You can also delete the whole failover group by using wildcards as shown in Example 9-9.

Example 9-9 Delete whole failover group

```
cdot-cluster01::> net int failover-groups delete -failover-group storage-lan
[-node | -port] *
```

9.6 Logical interfaces

A logical interface (LIF) is an IP address with associated characteristics, such as a role, a home port, a home node, a routing group, a list of ports to fail over to and a firewall policy. You can configure multiple LIFs on ports over which the cluster sends and receives communications over the network.

If there is any component failure, an LIF can fail over to or be migrated to a different physical port, thereby continuing to communicate with the cluster.

LIFs can be hosted on the following ports:

- Physical ports that are not part of interface groups
- Interface groups
- VLANs
- Physical ports or interface groups that host VLANs

While configuring SAN protocols such as FC on an LIF, it will be associated with a WWPN.

9.6.1 LIF types

An LIF role determines the kind of traffic that is supported over the LIF, along with the failover rules that apply and the firewall restrictions that are in place. An LIF can have any one of the five roles: node management, cluster management, cluster, intercluster, and data.

- Node management LIF:

  The LIF that provides a dedicated IP address for managing a particular node and gets created at the time of creating or joining the cluster. These LIFs are used for system maintenance, for example, when a node becomes inaccessible from the cluster. Node management LIFs can be configured on either node management or data ports.

  The node management LIF can fail over to other data or node management ports on the same node.

  Sessions established to SNMP and NTP servers use the node management LIF. AutoSupport requests are sent from the node management LIF.

- Cluster management LIF:

  The LIF that provides a single management interface for the entire cluster. Cluster management LIFs can be configured on node management or data ports.

  The LIF can fail over to any node management or data port in the cluster. It cannot fail over to cluster or intercluster ports.

- Cluster LIF:

  The LIF that is used for intracluster traffic. Cluster LIFs can be configured only on cluster ports.

  Cluster LIFs must always be created on 10-GbE network ports and can fail over between cluster ports on the same node, but they cannot be migrated or failed over to a remote node. When a new node joins a cluster, IP addresses are generated automatically.

  Note: If you want to assign IP addresses manually to the cluster LIFs, you must ensure that the new IP addresses are in the same subnet range as the existing cluster LIFs.

- Data LIF:

  The LIF that is associated with an SVM and is used for communicating with clients. Data LIFs can be configured only on data ports.

  You can have multiple data LIFs on a port. These interfaces can migrate or fail over throughout the cluster.

  Sessions established to NIS, LDAP, Active Directory, WINS, and DNS servers use data LIFs.

- Intercluster LIF:

  The LIF that is used for cross-cluster communication, backup, and replication.

  Intercluster LIFs can be configured on data ports or intercluster ports. You must create an intercluster LIF on each node in the cluster before a cluster peering relationship can be established.

9.6.2 LIF limitations

There are limits (see Table 9-3) on each type of LIF that you should consider when planning your network. Also be aware of the effect of the number of LIFs in your cluster environment.

The maximum number of LIFs that are supported on a node is 262. You can create additional cluster, cluster management, and intercluster LIFs, but creating these LIFs requires a reduction in the number of data LIFs.

Table 9-3 LIF limits

LIF type	Min	Max
Data LIFs	1 per SVM	128 per node with failover enabled 256 per node with failover disabled
Cluster LIFs	2 per Node	-
Node management LIFs	1 per Node	1 per port and per subnet
Cluster management LIFs	1 per cluster	-
Intercluster LIFs	0 without cluster peering 1 per node with cluster peering	-

9.6.3 Creating an LIF

There are certain guidelines that you should consider before creating an LIF. Consider the following points while creating an LIF:

- In data LIFs used for file services, the default data protocol options are NFS and CIFS.
- FC LIFs can be configured only on FC ports. iSCSI LIFs cannot coexist with any other protocols.
- NAS and SAN protocols cannot coexist on the same LIF.
- You can create both IPv4 and IPv6 LIFs on the same network port.

To create an LIF use the **`network interface create`** command as shown in Example 9-10. See **`man network interface create`** for further information.

Example 9-10 Create data LIF

```
cdot-cluster01::> network interface create -vserver vs_cifs_01 -lif
vs_cifs_01_lif1 -role data -home-node cdot-cluster01-01 -home-port e0b -address
9.155.66.26 -netmask 255.255.255.0
```

You can display information about all LIFs in the cluster with the **`network interface show`** command (see Example 9-11).

Example 9-11 Network interface show

```
cdot-cluster01::> net int show
(network interface show)
            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
cdot-cluster01
            cluster_mgmt up/up    9.155.66.34/24     cdot-cluster01-02
                                                                   e0a     false
cdot-cluster01-01
```

```
            cdot-cluster01-01_intercluster_lif1
                         up/up    9.155.66.65/24     cdot-cluster01-01
                                                                   e0a     true
            clus1        up/up    169.254.216.146/16 cdot-cluster01-01
                                                                   e2a     true
            clus2        up/up    169.254.67.99/16   cdot-cluster01-01
                                                                   e2b     true
            mgmt1        up/up    9.155.90.168/24    cdot-cluster01-01
                                                                   e0M     true
cdot-cluster01-02
            cdot-cluster01-02_intercluster_lif2
                         up/up    9.155.66.66/24     cdot-cluster01-02
                                                                   e0a     true
            clus1        up/up    169.254.45.229/16  cdot-cluster01-02
                                                                   e2a     true
            clus2        up/up    169.254.250.11/16  cdot-cluster01-02
                                                                   e2b     true
            mgmt1        up/up    9.155.90.169/24    cdot-cluster01-02
                                                                   e0M     true
vs_cifs_01
            vs_cifs_01_cifs_lif1
                         up/up    9.155.66.31/24     cdot-cluster01-02
                                                                   e0a     true
10 entries were displayed.
```

Note: You can use the **`ping`** command from a client or **`network ping`** command from a node in the cluster to verify if the configured addresses are reachable. Every LIF should be pingable.

9.6.4 Modifying an LIF

You can modify an LIF by changing the attributes such as the home node or the current node, administrative status, IP address, netmask, failover policy, or the firewall policy. You can also modify the address family of an LIF from IPv4 to IPv6. However, you cannot modify the data protocol that is associated with an LIF when the LIF was created.

Use the **`network interface modify`** command to modify an LIF's attributes. See man page for further information.

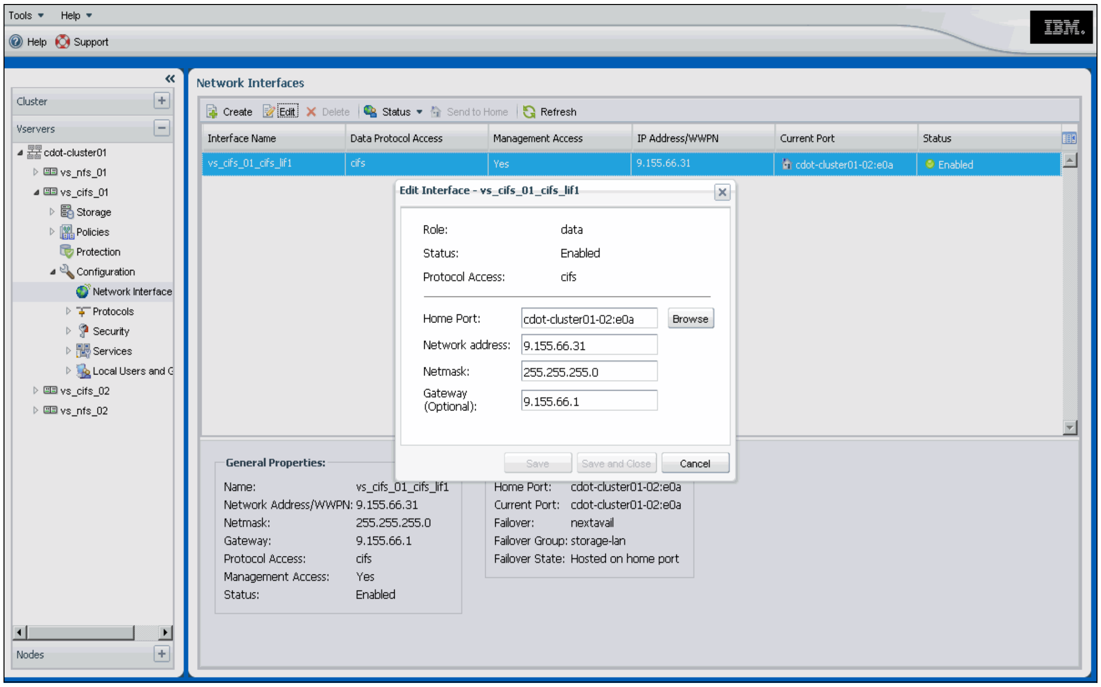
Example 9-12 shows how to modify the LIF `vs_cifs_01_cifs_lif1` that is owned by SVM `vs_cifs_01` in order to change it‘s IP address.

Example 9-12 Modify network interface

```
cdot-cluster01::> net int modify -vserver vs_cifs_01 -lif vs_cifs_01_cifs_lif1
-address 9.155.66.41 -netmask 255.255.255.0
```

You can also use the IBM N series OnCommand System Manager to modify the LIF. See Figure 9-4 for an example. Follow these steps:

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Vservers** → **Choose the desired SVM** → **Configuration** → **Network Interfaces**.
3. Select the LIF you want to modify and click **Edit** to edit the LIF.



Interface Name	Data Protocol Access	Management Access	IP Address/WWPN	Current Port	Status
vs_cifs_01_cifs_lif1	cifs	Yes	9.155.66.31	cdot-cluster01-02:e0a	Enabled

Edit Interface - vs_cifs_01_cifs_lif1

Role: data
Status: Enabled
Protocol Access: cifs

Home Port: cdot-cluster01-02:e0a
Network address: 9.155.66.31
Netmask: 255.255.255.0
Gateway (Optional): 9.155.66.1

General Properties:

Name: vs_cifs_01_cifs_lif1
Network Address/WWPN: 9.155.66.31
Netmask: 255.255.255.0
Gateway: 9.155.66.1
Protocol Access: cifs
Management Access: Yes
Status: Enabled

Home Port: cdot-cluster01-02:e0a
Current Port: cdot-cluster01-02:e0a
Failover: nextavail
Failover Group: storage-lan
Failover State: Hosted on home port

Figure 9-4 Modify LIF in system manager

9.6.5 Migrating an LIF

You might have to migrate an LIF to a different port on the same node or a different node within the cluster, if the port is either faulty or requires maintenance.

Before you migrate the LIF, you should ensure that the destination node and ports are operational and able to access the same network as the source port. Also failover groups must have been set up for the LIFs. See 9.5, “Failover groups” on page 138 for further information about failover groups.

> **Note:** You cannot migrate iSCSI LIFs from one node to another node. VMware VAAI copy offload operations fail when you migrate the source or the destination LIF.

You can either migrate a specific LIF or migrate all data and cluster management LIFs away from one node.

Example 9-13 shows how to migrate an LIF named `vs_cifs_01_cifs_lif1` on the SVM `vs_cifs_01` to the port `e0a` on node `cdot-cluster01-01`.

Example 9-13 Migrate specific LIF

```
cdot-cluster01::> net interface migrate -vserver vs_cifs_01 -lif
vs_cifs_01_cifs_lif1 -dest-node cdot-cluster01-01 -dest-port e0a
```

Example 9-14 shows how to migrate all LIFs away from node `cdot-cluster01-01.`

Example 9-14 Migrate all LIFs away from specific node

```
cdot-cluster01::> network interface migrate-all -node cdot-cluster01-01
```

9.6.6 Reverting an LIF

You can revert an LIF to its home port after it fails over or is migrated to a different port either manually or automatically. If the home port of a particular LIF is unavailable, the LIF remains at its current port and is not reverted.

You can use the **network interface show** command to find LIFs that are not on their home-port and have a Is Home state of false. These LIFs need to be reverted (see Example 9-15).

Example 9-15 Display LIFs that are not on their home-port

```
cdot-cluster01::> net int show -is-home false
  (network interface show)
            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
vs_cifs_01
            vs_cifs_01_cifs_lif1
                         up/up    9.155.66.31/24     cdot-cluster01-01
                                                                   e0a     false
```

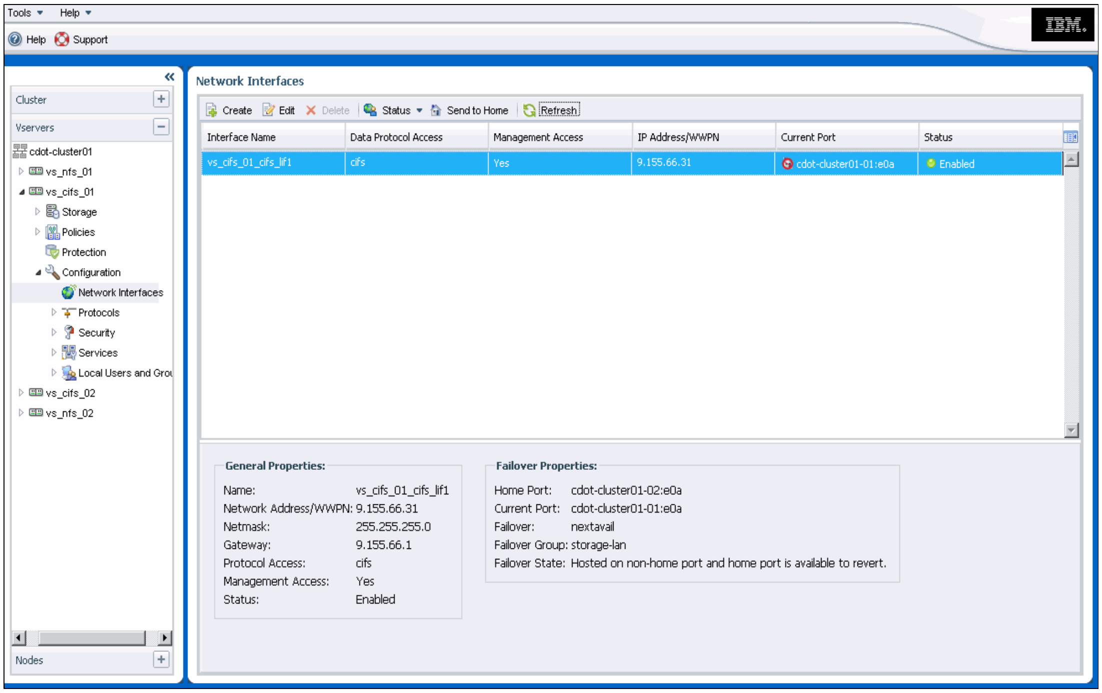
Revert the interface with the **network interface revert** command as shown in Example 9-16.

Example 9-16 Revert LIFs

```
cdot-cluster01::> net interface revert -vserver vs_cifs_01 -lif
vs_cifs_01_cifs_lif1
```

You can also use the IBM N series OnCommand System Manager to revert the LIF. See Figure 9-5 for an example. Follow these steps:

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Vservers** → **Choose the desired SVM** → **Configuration** → **Network Interfaces**.
3. Select the LIF you want to revert and click **Send to Home** to revert the LIF.



Network Interfaces

Interface Name	Data Protocol Access	Management Access	IP Address/WWPN	Current Port	Status
vs_cifs_01_cifs_lif1	cifs	Yes	9.155.66.31	cdot-cluster01-01:e0a	Enabled

General Properties:

Name:	vs_cifs_01_cifs_lif1
Network Address/WWPN:	9.155.66.31
Netmask:	255.255.255.0
Gateway:	9.155.66.1
Protocol Access:	cifs
Management Access:	Yes
Status:	Enabled

Failover Properties:

Home Port:	cdot-cluster01-02:e0a
Current Port:	cdot-cluster01-01:e0a
Failover:	nextavail
Failover Group:	storage-lan
Failover State:	Hosted on non-home port and home port is available to revert.

Figure 9-5 Revert LIF in system manager

9.6.7 Deleting an LIF

You can delete an LIF that is not required.

Use the **`network interface delete`** command to delete the LIF. You can specify * as a wildcard in order to delete all LIFs.

Example 9-17 shows how to delete an LIF named `vs_cifs_01_cifs_lif1` on the SVM `vs_cifs_01`.

Example 9-17 Delete LIF

```
cdot-cluster01::> network interface delete -vserver vs_cifs_01 -lif
vs_cifs_01_cifs_lif1
```

9.7 Routing groups

You can control how LIFs in an SVM use your network for outbound traffic by configuring routing groups and static routes. A set of common routes are grouped in a routing group that makes the administration of routes easier.

When an LIF is created, an associated routing group is automatically created. You cannot modify or rename an existing routing group; therefore, you might have to create a routing group.

9.7.1 Creating routing groups

When you create a new routing group, observe the following rules:

- The routing group and the associated LIFs should be in the same subnet.
- All LIFs sharing a routing group must be on the same IP subnet.
- All next-hop gateways must be on that same IP subnet.
- An SVM can have multiple routing groups, but a routing group belongs to only one SVM.
- The routing group name must be unique in the cluster and should not contain more than 64 characters.
- You can create a maximum of 256 routing groups per node.

To create a routing group, use the **`network routing-groups create`** command. Example 9-18 shows how to create a routing group named `d9.155.66.0` for the SVM `vs_cifs_01` and the subnet `9.155.66.0/24`. See **`man network routing-groups create`** for further information.

Example 9-18 Create routing group

```
cdot-cluster01::> network routing-groups create -vserver vs_cifs_01 -routing-group
d9.155.66.0 -subnet 9.155.66.0/24 -role data
```

9.7.2 Deleting routing groups

When you delete a routing group, observe the following rules:

- There must be no LIF that is using the routing group.
- The routing group must not contain any routes.

To delete an existing routing group, use the **`network routing-groups delete`** command as shown in Example 9-19.

Example 9-19 Delete routing group

```
cdot-cluster01::> network routing-groups delete -vserver vs_cifs_01 -routing-group
d9.155.66.0
```

9.7.3 Creating a static route

Use the **`network routing-groups route create`** command to create a route within the routing group.

Example 9-20 shows how to create a default gateway route for the routing group named `d9.155.66.0` for the SVM `vs_cifs_01` and the subnet `9.155.66.0/24`.

Example 9-20 Create route within routing group

```
cdot-cluster01::> network routing-groups route create -vserver vs_cifs_01
-routing-group d9.155.66.0 -destination 0.0.0.0/0 -gateway 9.155.66.1
```

9.7.4 Deleting a static route

You can delete an unneeded static route from a routing group.

Example 9-21 shows how to delete a static route for the routing group named `d9.155.66.0` for the SVM `vs_cifs_01` and the subnet `9.155.66.0/24`.

Example 9-21 Delete static route

```
cdot-cluster01::> network routing-groups route delete -vserver vs_cifs_01
-routing-group d9.155.66.0 -destination 0.0.0.0/0
```

10

NAS protocols

This chapter describes NAS network protocols for the IBM System Storage N series Clustered Data ONTAP storage system and characteristics for each network protocol.

Read this chapter to determine what is the correct protocol for your environment and which protocol should be used for each case.

N series Clustered Data ONTAP now supports DNS NAS load balancing, automatic logical interface (LIF) rebalancing, and flex cache volumes for NAS protocols.

The following topics are covered:

- Supported network protocols and characteristics
- CIFS
- NFS
- pNFS
- Further information

10.1 Supported network protocols and characteristics

You can use Table 10-1 to check supported protocols for each of the Clustered Data ONTAP features shown.

Table 10-1 Supported network protocols for each feature

DNS NAS load balancing	Automatic LIF rebalancing	Flex cache volumes	Infinite Volumes
NFS v3, v4, and v4.1	NFS v3 only	NFS v3, v4	NFS v3, v4, v4.1, and pNFS
SMB v1.0 and v2.0	N/A	SMB v1.0, v2.x, and v3.0	SMB v1.0 only

Before describing each NAS protocol, we introduce some characteristics for NAS protocols.

10.1.1 Connectionless and connection-oriented

If you want to know the detailed information for TCP, UDP, three-way hand-shake, and so on, refer to the OSI seven-layer specification. The International Standards Organization (ISO) developed the seven-layer Open Systems Interconnect (OSI) Reference Model to describe communication between network devices. The model is explained in virtually every networking reference text, and so it is not covered in this chapter.

There are two protocols in the transport layer. One is connectionless and the other is connection-oriented, as shown in Table 10-2.

Table 10-2 Connection-oriented and connectionless protocols

	Connection-oriented	Connectionless
Transport layer protocols	TCP	UDP
Application layer protocols	All of CIFS, NFS v4, NFS v3 over TCP	NFS v3 over UDP (default)

As you can see, NFS v3 over UDP is the only connectionless protocol and most users are using this default NFS v3 for their UNIX or Linux hosts.

Connectionless protocols are very popular protocols. Packets are sent over the network without regarding whether they actually arrive at their destinations. There are no acknowledgments, but a datagram can be sent to many different destinations at the same time. Connectionless protocols are faster than other connection-oriented protocols because no time is used in establishing and tearing down connections and so on.

Connectionless protocols are also referred to as best-effort protocols. But typically they are only used for a Local Area Network (LAN) configuration. Normally if the network congestion occurs or the traffic passes through a Wide Area Network (WAN), lots of UDP packets would be lost because the operating system (OS) does not know which packets should be retransmitted or not. The following note expands on these concepts.

Note: Use the following guidelines to determine when you should use the UDP transport or the TCP transport to improve filer performance:

- Use the TCP transport over a WAN network.
- Use the UDP transport over a LAN network.
- Use the TCP transport if you are using the UDP transport and you experience packet loss, especially during periods of heavy write traffic.

You can specify the transport using the options **`nfs.tcp.enable`** command for 7-Mode or below. You can apply this feature in Clustered Data ONTAP as follows:

```
vserver nfs modify -vserver [vserver name] -tcp enabled
```

10.1.2 Physical ports configuration for performance or redundancy

Normally you can assign an IP address for one physical port. And the IP address can be worked through the partner node even if the node having the configured physical port is down or the specific network port is down. But for the best performance or redundancy, configure interface groups *(ifgrps)* as you did with previous Data ONTAP versions.

You can refer to the network management guide for each version. For Clustered Data ONTAP, documents for the Clustered Data ONTAP v8.2 or above are suitable.

For example, here is the link to all publications for v8.1.3 for 7 Mode. You can refer to documents for the cluster Data ONTAP as well through the similar link when Clustered Data ONTAP is announced officially:

https://www.ibm.com/support/entdocview.wss?uid=ssg1S7004497

Note: The *ifgrps* provide three modes, single, multi, and LACP.

Keep in mind that there are several types of *ifgrps* and you can select one of them for the best performance and redundancy.

iSCSI also uses Ethernet like other NAS protocols, but it is only for block I/O. It will be reviewed in Chapter 11, “SAN protocols” on page 169.

10.2 CIFS

Server Message Block (SMB) is a remote file-sharing protocol used by Microsoft Windows clients and servers starting in the mid-1980s. The original SMB 1.0 (also known as Common Internet File System or CIFS) was designed and implemented to support file-serving solutions based on the assumptions existing at that time.

From Clustered Data ONTAP point of view, CIFS architecture in Clustered Data ONTAP has been drastically changed from the traditional architecture used in Data ONTAP 7 and in Data ONTAP 8.1 operating in 7-Mode. The entire CIFS infrastructure has been reconstructed for scalability and robustness in the cluster organization.

10.2.1 Change history

For the past decade or so, some minor changes and tweaks have been made to the protocol to support some new functionality such as network resiliency, scalability, and so on. SMB 2.0, introduced with Windows Vista, was the first major redesign that considered the needs of the next generation of file servers and clients. These needs included a redesign for modern networking environments such as wide area networks (WANs), possible high-loss networks, time-outs, high latency, and so on. SMB 2.1 was a new revision built on top of SMB 2.0, with additional features. SMB 2.1 is inclusive of SMB 2.0; when a CIFS server or client is said to support SMB 2.1, this server/client also supports SMB 2.0.

Microsoft's ongoing efforts to evolve SMB 2 have positioned the protocol as the next generation of the previous CIFS (SMB 1.0) protocol. It was first introduced with SMB 2.0 in Windows Vista in 2007 and updated with the release of Windows Server 2008 and Windows Vista SP1 in 2008. SMB 2.1 was then introduced in Windows 7 and Windows Server 2008 R2. Microsoft plans to support SMB 2 as the file system protocol of choice on all future releases of Microsoft operating systems.

Microsoft announced SMB v3.0 (also formerly known as v2.2) and introduced with Windows server 2012.

Note: For Data ONTAP 8.2, SMB 3.0 features are supported only for Hyper-V clients.

Clustered Data ONTAP v8.2.x also supports all CIFS protocol versions, but it depends on which Clustered Data ONTAP features are required or which Windows servers are attached.

Data ONTAP supports several versions of the Server Message Block (SMB) protocol on your VSM's CIFS server. Data ONTAP support for SMB for Vservers with FlexVol volumes and VSMs with Infinite Volumes differ. You need to be aware of which versions are supported for each type of VSM.

10.2.2 SMB 2.0 and 3.0 enhancements

SMB 2.0 is a major revision of the SMB 1.0 protocol and includes a complete reworking of the packet format. SMB 2.0 introduces several performance improvements over earlier versions:

- More efficient network use
- Compounding of requests
- Larger reads and writes
- File-property and directory-property caching
- Durable file handles
- Hash-based message authentication code (HMAC) SHA-256 signing

SMB 2.1 provides these additional performance enhancements in addition to SMB 2.0:

- The client opportunistic lock (oplock) leasing model
- Support for large maximum transmission unit (MTU)
- Support for earlier versions of SMB

To enable SMB 2.0 and 2.1, use the following command as shown in Example 10-1.

```
cluster::> vserver cifs options modify -vserver <vserver name> -smb2-enabled true
```

※ privilege: **advanced**

Example 10-1 Enable SMB 2.0 and 2.1

```
cdot-cluster01::> set -privilege advanced

Warning: These advanced commands are potentially dangerous; use them only when
         directed to do so by IBM personnel.
Do you want to continue? {y|n}: y

cdot-cluster01::*> vserver cifs options modify -vserver vs_cifs_02 -smb2-enabled
true

cdot-cluster01::*> set -privilege admin
```

The SMB 2.1 protocol provides several minor enhancements to the SMB 2.0 specification. Clustered Data ONTAP supports mots of the SMB 2.1 features. Support for SMB 2.1 is automatically enabled when you enable the SMB 2.0 protocol on a storage virtual machine (SVM). You can use the command shown in Example 10-1 to enable SMB 2.x for an SVM.

One of the most important features in the SMB 2.1 protocol is the opportunistic lock (oplock) leasing model. Leasing enables a client to hold oplocks over a wider range of scenarios. The feature offers enhanced file caching and metadata caching opportunities for the SMB client and provides major performance benefits by limiting the amount of data that must be transferred between the client computer and the server. This enhancement particularly benefits works with high latency. Additionally, because the number of operations that must be directed toward an SMB file server is reduced, the SMB file server scalability is increased.

The new leasing model in SMB 2.1 enables greater file-caching and handle-caching opportunities for an SMB 2.1 client computer while preserving data integrity and requiring no current application changes to take advantage of this capability.

Clustered Data ONTAP 8.2 introduces support for SMB 3.0 enhancements that provide BranchCache Version 2, witness protocol, remote VSS for SMB shares, persistent file handles, ODX copy offload, and continuously available shares. But SMB 3.0 is supported only for Hyper-V clients.

10.2.3 Supported SMB versions for each Data ONTAP feature

Data ONTAP supports the following SMB versions for SVMs with FlexVol volumes and SVMs with Infinite Volumes, as listed in Table 10-3.

Table 10-3 Supported SMB versions for the SVM on your Clustered Data ONTAP

SMB version	Supported on SVM with FlexVol volumes	Supported on SVM with Infinite Volumes
SMB 1.0	Yes	Yes
SMB 2.0	Yes	No
SMB 2.1	Yes	No
SMB 3.0	Yes	No

10.2.4 Takeover in a CIFS environment

When a node takes over its partner, it continues to server and update data in partner's aggregates and volumes. To do this, it takes ownership of the partner's data aggregates, and the partner's LIFs migrate according to network interface failover rule. Except for specific SMB 3.0 connections, exiting SMB(CIFS) sessions are disconnected when the takeover occurs.

> **Note:** Due to the nature of the SMB protocol, all SMB sessions, except for SMB 3.0 sessions connected to shares with the `Continuous Availability` property set, will be disruptive.

10.2.5 CIFS configuration in Clustered Data ONTAP

We show you a detailed configuration for CIFS shares in Chapter 23, "CIFS storage" on page 383.

That chapter explains what functions were added for Clustered Data ONTAP and which ones are configured for CIFS shares.

Creating an SVM

You already knew that the Clustered Data ONTAP needs an SVM for each data service. Before creating any CIFS shares, you should define at least one Storage Virtual Server for NAS services.

Configuring an SVM with name service switches

There are three types of name service switches (file, nis, and ldap). If you use Lightweight Directory Access Protocol (LDAP), an LDAP domain must be associated with this SVM.

- LDAP domains must already exist outside of Data ONTAP architecture.
- Multiple LDAP domain configurations can exist for an SVM, but only one can be active at a time.
- Use the **`vserver services ldap`** command to view and modify LDAP settings in an SVM.

> **Note:** Multiple configurations can be created within an SVM and for multiple SVMs. Any or all of those configurations can use the same LDAP domain or different ones. Only one LDAP domain configuration can be active for an SVM at one time.

Configuring a SVM with a Domain controller

There are two types of CIFS domains.

- NT LAN Manager (NTLM), prior to Windows 2000
- Active Directory for Windows 2000 and later

Active Directory uses Kerberos authentication. NT LAN Manager (NTLM) is provided for backward compatibility with Windows clients earlier than Windows 2000 Server. When configuring CIFS, the domain controller information is automatically discovered, and the account on the domain controller is created for you. Domain configuration grants clients single-sign-on access to the file server.

To configure the CIFS services with a Domain controller, you need this information:

- The domain name
- The NetBIOS name for the Data ONTAP CIFS "server"

- An administrator domain controller password so that a “computer” entity can be automatically created

Example 10-1 shows a way to configure a CIFS server in Clustered Data ONTAP.

Example 10-2 CIFS configuration example

```
cdot-cluster01::> vserver cifs create -vserver vs_cifs_02 -domain test.ibm.local
-cifs-server MYCIFS

cdot-cluster01::> vserver cifs share create -vserver vs_cifs_02 -share-name -path
/ -share-properties browsable

cdot-cluster01::> vserver cifs share create -vserver vs_cifs_02 -share-name ~%w
-path /user/%w -share-properties browsable,homedirectory
```

Kerberos Authentication

Here are the characteristics for Kerberos Authentication. This authentication method is for both NFS and CIFS. And Microsoft Active Directory also uses Kerberos. Note the following characteristics:

- Kerberos can be used with NFS and CIFS.
- Microsoft Active Directory Kerberos and MIT Kerberos are supported.

Kerberos realms have the following characteristics:

- A Kerberos realm must already exist outside of Data ONTAP architecture.
- Kerberos realms can be used by any NFS configurations.
- CIFS does not require a separate Kerberos realm definition.
- Use the **vserver services kerberos-realm** command to view and modify the Kerberos settings in an SVM.

When the storage server’s clock is not in sync within 5 minutes of the domain controllers, the authentication service through Kerberos stops working. The security service logs the time skew error messages.

Adjust the cluster timer by using the following command:

cluster::> **date** *[[[[[cc]yy]mm]dd]hhmm[.ss]]*

Or you can set up the Network Time Protocol (NTP) server to synchronize the cluster clock with the target server.

> **Note:** Kerberos has strict time requirements, which means that the clocks of the involved hosts must be synchronized within configured limits. The cluster timer must be in sync with the Windows domain controller to avoid time-related problems. To set up the NTP server, use this command:
>
> ```
> cluster::> system services ntp server
> ```

Creating an export policy for an SVM

Each volume has an export policy that is associated with it. Each policy can have rules that govern access to the volume based on criteria such as a client’s IP address or network, the protocol that is used (CIFS or any), and more. A “default” export policy exists, which contains no rules.

Each export policy is associated with one data SVM. An export policy need to be unique only within an SVM. When an SVM is created, the default export policy is created for it.

Changing the export rules within an export policy changes the access for every volume that uses that export policy.

Here are some guidelines for export policies and rules:

- Export policies and rules enable the administrator to restrict access to volumes based on client IP address and authentication types.
- Each volume has an export policy. (The default is **`default`**.)
- Export policies only apply at the volume (not the qtree) level.

Use these commands to manage policies and rules:

- Use the **`vserver export-policy`** command to view and modify export policy settings in an SVM.
- Use the **`vserver export-policy rule`** command to view and modify export policy rule settings in an SVM.

Follow these guidelines for export rule configuration:

- Set everything wide open to verify that it is set up properly.
- Tighten down security.
- Add only one export policy per volume.

Note: Rules can be added to control specific access.

Creating name mapping rules for CIFS

The Data ONTAP storage system is designed to support either CIFS credentials (NTFS), UNIX credentials (UNIX), or both (MIXED), depending on the volume security type. Both of the supported network file systems (CIFS and NFS) can be used to access the files on any of the volume types. In a multiprotocol environment, the name-mapping mechanism plays a key role in controlling the behavior that allows a user with credentials from one network file system type to be mapped to a user with credentials on another network file system type.

The name mapping procedure is accomplished in three sequential steps, as described in the following sections.

1. Explicit name mapping

Explicit name mapping can be defined either locally through the clusterwide name-mapping table or through LDAP. The name-mapping service switch can be defined by using the following command:

```
cluster::> vserver modify -vserver <vserver> -fields nm-switch {file|ldap}
```

2. Implicit name mapping

When explicit name mapping does not find a matching entry, the system tries implicit name mapping.

Suppose that the Windows domain for the CIFS server is called DOMAIN.COM. Then:

- Windows user DOMAIN.COM\user1 is mapped to the UNIX user called user1.
- UNIX user2 is mapped back to Windows user DOMAIN.COM\user2.

Preferred practice:

- Define an entry that maps Windows user *\root to a UNIX user nobody or pcuser to prevent the NT root user from being mapped to UNIX UID 0 (root user, superuser).
- Define an entry that maps the UNIX user administrator to a Windows Guest user to prevent spoofing the NT administrator account from UNIX.

3. Default name mapping

When both explicit and implicit name mapping fail, the default name mapping is used, if defined, as the last resort.

Preferred practice:

To prevent unexpected mapping, set up a default name mapping rule through the **`vserver cifs options`** command and only give the Guest account to the default mapping account.

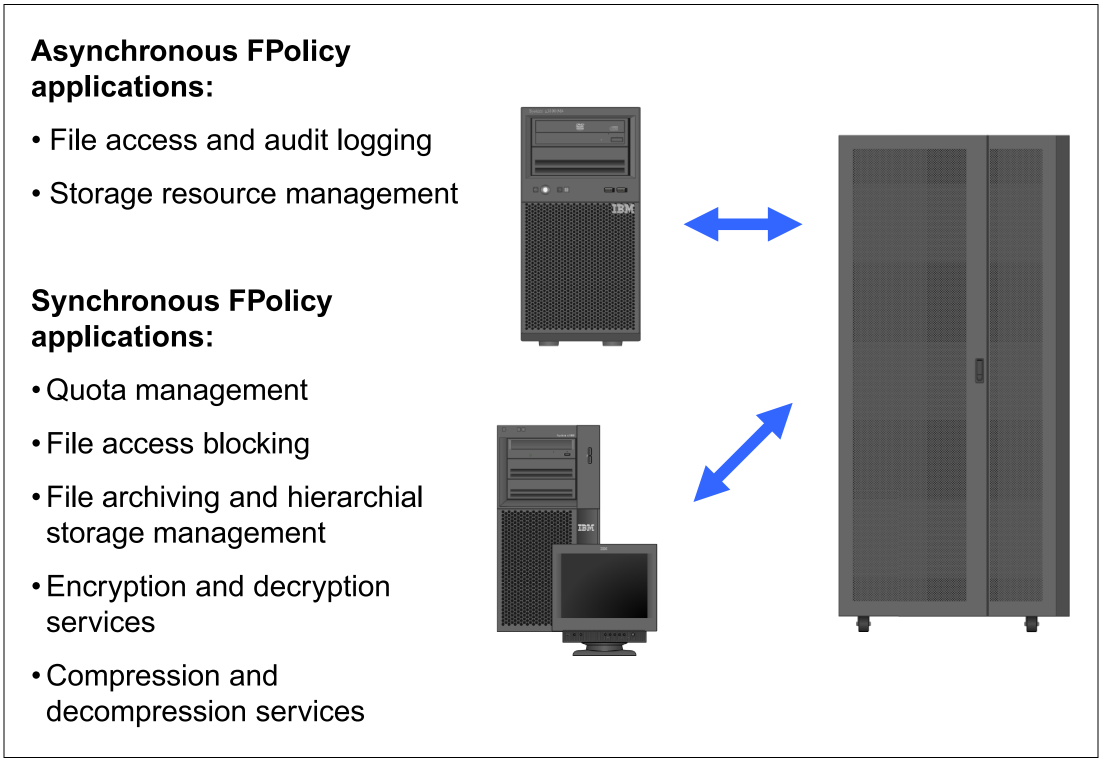
FPolicy framework

Data ONTAP 8.2 and later releases provide support for the FPolicy feature on SVMs with FlexVol volumes. FPolicy is a file access notification framework that uses policies to monitor and manage NFS and SMB file access events.The FPolicy feature supports event notification for files and directories in FlexVol volumes that are accessed using NFSv3, NFSv4, or CIFS. Currently, the FPolicy feature does not work with Infinite Volumes.

The Data ONTAP framework creates and maintains the FPolicy configuration, monitors file events, sends notifications to external FPolicy servers, and manages connections between Vservers and the external FPolicy servers. External FPolicy servers are applications servers that can provide file monitoring and management services such as monitoring and recording file access events, providing quota services, performing file blocking based on defined criteria, and providing data migration services using hierarchical storage management applications.

If you do not want to use external FPolicy servers, FPolicy also provides native file screening, which you can use to configure simple file blocking based on file extensions.

See Figure 10-1 to understand how FPolicy Framework works.



Asynchronous FPolicy applications:

- File access and audit logging
- Storage resource management

Synchronous FPolicy applications:

- Quota management
- File access blocking
- File archiving and hierarchial storage management
- Encryption and decryption services
- Compression and decompression services

Figure 10-1 FPolicy Framework

10.3 NFS

Actually, the first remote file system shipped with System V was RFS. Although it had excellent UNIX semantics, its performance was poor, so it met with little use. The most commercially successful and widely available remote-file system protocol is the network file system (NFS) designed and implemented by Sun Microsystems 1985. There are two important components to the success of NFS. First, SUN placed the protocol specification for NFS in the public domain. Second, Sun sells that implementation to all people who want it, for less than the cost of implementing it themselves. Thus, most vendors chose to buy the Sun implementation.

10.3.1 Change history

The NFS implementation that appears in 4.4BSD was written by Rick Macklem at the University of Guelph using the specifications of the Version 2 protocol published by Sun Microsystems, 1989.

Many of the extensions were incorporated into the revised NFS Version 3 specification by Sun Microsystems, 1993. The main problem with using TCP transport with Version 2 of NFS is that it is supported between only BSD and a few other vendors clients and servers. However, the clear superiority demonstrated by the Version 2 BSD TCP implementation of NFS convinced the group at Sun Microsystems implementing NFS Version 3 to make TCP the default transport. Thus, a Version 3 Sun client will first try to connect using TCP; only if the server refuses will it fall back to using UDP.

In the summer of 1998, Sun Microsystems ceded change control of NFS to the Internet Engineering Task Force [RFC2339]. IETF assumed the responsibility to create a new version of NFS for use on the Internet.

Prior to the formation of the IETF NFS Version 4 working group, Sun Microsystems deployed portions of the technology leading up to NFS Version 4, notably WebNFS [RFC2054, RFC2055] and strong authentication with Kerberos [MIT] within a GSS-API framework [RFC2203].

Following discussions in the working group, and contributions by many members, prototype implementations of the protocol began to prove out the concepts. Initial implementation testing of prototypes (including a Java prototype) based on the working drafts occurred in October 1999 to verify the design. The specification was submitted to the Internet Engineering Steering Group for consideration as a Proposed Standard in February 2000. Further implementation work and interoperability testing occurred early March 2000.

Note: With faster NICs and switches, you are advised to support NFSv2 or NFS v3 protocol over TCP rather than over UDP. NFS v4 is supported over TCP only.

10.3.2 NFS v3, v4, and v4.1 protocols: Enhancements

Now we introduce the NFS protocols, including NFS v3, v4, and v4.1.

NFS v3

NFS Version 3 was designed to be easy to implement, given an NFS Version 2 implementation. As we were told in the previous 10.3.1, “Change history”, NFS v3 has supported TCP transport that it is supported for all UNIX-based clients and servers.

The default transport protocol is UDP and users can select TCP transport protocol as well. Version 3 incorporated many performance improvements over Version 2. But it 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.

Here are the significant changes for Version 3 from Version 2:

- Large block transfers
- Safe asynchronous writes
- Improved attribute returns
- The `readdirplus` operation

You can find detailed information regarding the enhancements for Version 3 in the Redbooks publication, *IBM System Storage N series Software Guide,* SG24-7129.

NFS v4

NFSv4 introduced the following new features in addition to the previous Version 3:

- 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
- Referrals

Also, as this protocol is working on the TCP transport only, then the NFS Version 4 protocol is stateful. NFS is a distributed file system designed to be operating system independent. It achieves this by being relatively simple in design and not relying too heavily on any particular file system model.

The first major structural change to NFS compared to prior versions is the elimination of ancillary protocols. In NFS Versions 2 and 3, the Mount protocol was used to obtain the initial filehandle, while file locking was supported via the Network Lock Manager protocol. NFS Version 4 is a single protocol that uses a well-defined port, which, coupled to the use of TCP, allows NFS to easily transit firewalls to enable support for the Internet.

Another structural difference between NFS Version 4 and its predecessors is the introduction of a COMPOUND RPC procedure that allows the client to group traditional file operations into a single request to send to the server. In NFS Versions 2 and 3, all actions were RPC procedures. NFS Version 4 is no longer a “simple” RPC-based distributed application. In NFS Version 4, work is accomplished via operations.

See the additional explanation regarding the enhancements for Version 4 provided in the Redbooks publication, *IBM System Storage N series Software Guide*, SG24-7129-06.

NFS v4.1

NFS Version 4.1 is a minor revision and extension, not a modification, of NFS Version 4.0. Also, NFS v4.1 is fully compliant with the current NFS v4 protocol specification. NFS v4.1 extends delegation beyond files to directories and symbolic links (symlinks), introduces NFS sessions for enhanced efficiency and reliability, and provides parallel NFS (pNFS). NFS v4.1 also fixes some problems with NFS v4 as well.

Briefly, here we list what was changed.

- A minor revision of NFS v4
- An extension, not a modification, of NFS v4
- Delegations on directories and symbolic links
- A new NFS session model
- pNFS
- Fixes to NFS v4

In 10.4, “pNFS” on page 163, we look into parallel NFS, which is one of features in v4.1, in more detail.

10.3.3 Supported NFS versions for each Data ONTAP feature and referrals

FlexCache volumes support client accessing using the protocols of NFS v3 and NFS v4. And Infinite Volumes are supported by the protocols of NFS v3 and NFS v4.1.

With DNS load balancing enabled, supported NFS protocols include NFS v3, NFS v4, and NFS v4.1.

In automatic LIF rebalancing, LIFs are automatically migrated to a less-utilized port, based on the configured failover rule. Automatic LIF rebalancing allows even distribution of the current load. NFS v3 is the only supported protocol.

Clustered Data ONTAP 8.2 introduced support for the NFS v4 protocol specification and for elements of NFS v4.1. Clustered Data ONTAP continues to support NFS v3 fully. NFS v4 support brings the Data ONTAP 8.2 operating system in parity with the Data ONTAP 7.3 operating system.

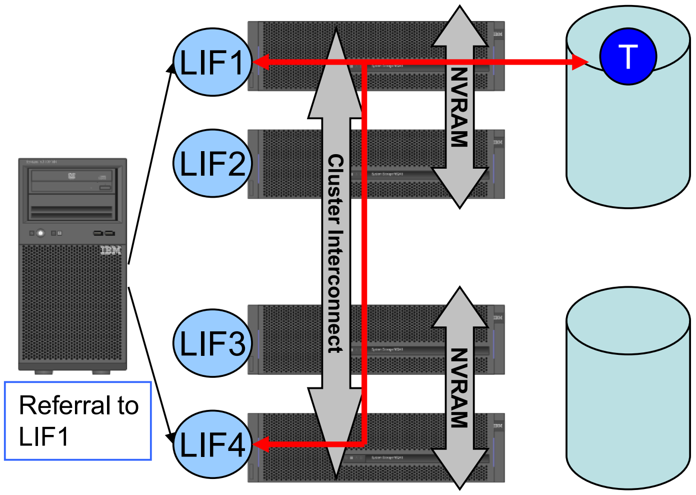
One of the key features for NFS v4 is the concept of *referrals*.

Clustered Data ONTAP 8.2 introduced NFS v4 referrals. When referrals are enabled in an SVM, Clustered Data ONTAP provides referrals within that SVM to NFS v4 clients. An intra-SMV referral occurs when a cluster node that is receiving the NFS v4 request refers the NFS v4 client to another LIF in the SVM. The NFS v4 client uses this referral to direct its access over the referred path at the target LIF from that point onward. The original cluster node issues a referral when it determines that an LIF exists in the SVM and is a resident on the cluster node on which the data volume resides. In other worlds, if a cluster node receives an NFS v4 request for a nonlocal volume, it can refer the client to the local path for that volume through the LIF.

This therefore allows clients faster access to the data and avoids extra traffic on the cluster interconnect.

By default, NFS v4 referrals are enabled on Linux clients like Red Hat Enterprise Linux 5.4 and later releases.

See Figure 10-2 to understand the concept of referrals.



- The initial request for T is through LIF4.
- Access to T is over the cluster interconnect.
- LIF4 issues a referral back to the client to use LIF1.
- The client begins to use LIF1 to access T.
- This is supported in clustered Data ONTAP 8.2 and later.
- This is supported in Red Hat 5.4 and later.

Figure 10-2 NFS v4 referrals

10.3.4 NFS configuration in Clustered Data ONTAP

We show you a detailed configuration for NFS shares in Chapter 24, “NFS storage” on page 411.

That chapter explains what functions were added for Clustered Data ONTAP and which ones are configured for NFS shares.

Creating an SVM

You already know that the Clustered Data ONTAP needs an SVM for each data service. Before creating any NFS shares, you should define at least one Storage Virtual Server for NAS services.

Configuring an SVM with name service switches

There are three types of name service switches (file, nis, and ldap). If you use a NIS domain already, a NIS domain must be associated with this SVM.

NIS domains must already exist outside of Data ONTAP architecture. The NIS domain cannot be created within a Data ONTAP cluster. If the customer has a NIS domain, then a configuration can be created to associate the domain with data SVMs within Clustered Data ONTAP.

Multiple NIS domain configurations can exist for an SVM, but only one can be active at a time.

To view and modify NIS settings in an SVM, use the following command:

```
cluster::> vserver services nis-domain
```

> **Note:** Export policies and rules, Kerberos, and Kerberos Realm for NFS are exactly same as explained for the previous CIFS configuration. See the previous discussion in 10.2.5, “CIFS configuration in Clustered Data ONTAP” on page 154.

Configuring an SVM for NFS

An NFS configuration is limited in scope to a data SVM. An SVM does not have an NFS configuration. As such, an SVM must exist before an NFS configuration can be created.

If Kerberos is to be used for an NFS configuration, Kerberos must be configured at the LIF level within the SVM. Therefore, if two data LIFs are associated with an SVM that is running NFS, Kerberos can be configured for one of the LIFs but not for the other one. Each LIF can use a different Kerberos configuration and therefore a different Kerberos realm.

Here are the configuration options for NFS that can be enabled or disabled for each NFS configuration:

- NFS v3
- NFS v4
- NFS v4.1
- NFS v4.1 parallel NFS (pNFS)
- TCP
- UDP

Example 10-3 shows how the NFS configuration for each SVM can be configured with the options that are listed here. Each SVM can use different NFS options.

Example 10-3 NFS configuration options

```
cluster1::> Vserver nfs modify -
-Vserver               -access               -v3
-v4.0                  -udp                  -tcp
-spinauth              -default-win-user     -v4.0-acl
−4.0-read-delegatoin   -v4.0-writedelegation -v4-id-domain
-v4.1                  -rquota               -v4.1-pnfs
-v4.1-acl              -vstorage
```

10.4 pNFS

This section discusses the purpose of pNFS.

10.4.1 pNFS: What it is, and why it was created

Parallel NFS (pNFS) is a new standard documented in RFC 5661 by the Internet Engineering Task Force (IETF) in conjunction with NFS Version 4.1 (NFS v4.1). pNFS offers an industry-standard framework for shared, high-performance parallel I/O suitable for use with data-intensive scientific, engineering, and other applications running on large compute clusters. pNFS overcomes the single-file server design of standard NFS by utilizing a metadata server in combination with multiple data servers. pNFS client systems are able to I/O in parallel to file data striped across multiple data servers. File-based, block-based, and object-based data servers are currently supported by the protocol.

> **Note:** Currently pNFS works with only Red Hat Enterprise Linux v6.2, Fedora 14, or any kernel Version 2.6.39 and higher. Refer to the interoperability site before applying pNFS at your site.

Compute clusters running computer-aided engineering (CAE), seismic data processing, bioinformatics, and other science and engineering applications often rely on NFS to access shared data. However, because all files in a file system must be accessed through a single file server, NFS can result in significant bottlenecks for applications such as these.

One common way to scale performance is to “scale up” the performance of that single file server. Another way to scale performance is to “scale out” storage; clients connect to a single file server, but data is distributed across multiple servers using a clustered file system. For some applications, this can increase performance, but it might not accelerate access to single, large files or eliminate all bottlenecks. These limitations have created a strong need for an industry-standard method to “scale out” shared storage, analogous to the way servers are scaled out in a compute cluster.

See Figure 10-3 to discover what is different between pNFS and the previous NFS v4.

With standard NFS, all data must be accessed through a single server, which might become a bottleneck for certain application. In pNFS, data is striped across multiple data servers. Clients access data from data server directly based on information received from a metadata server.

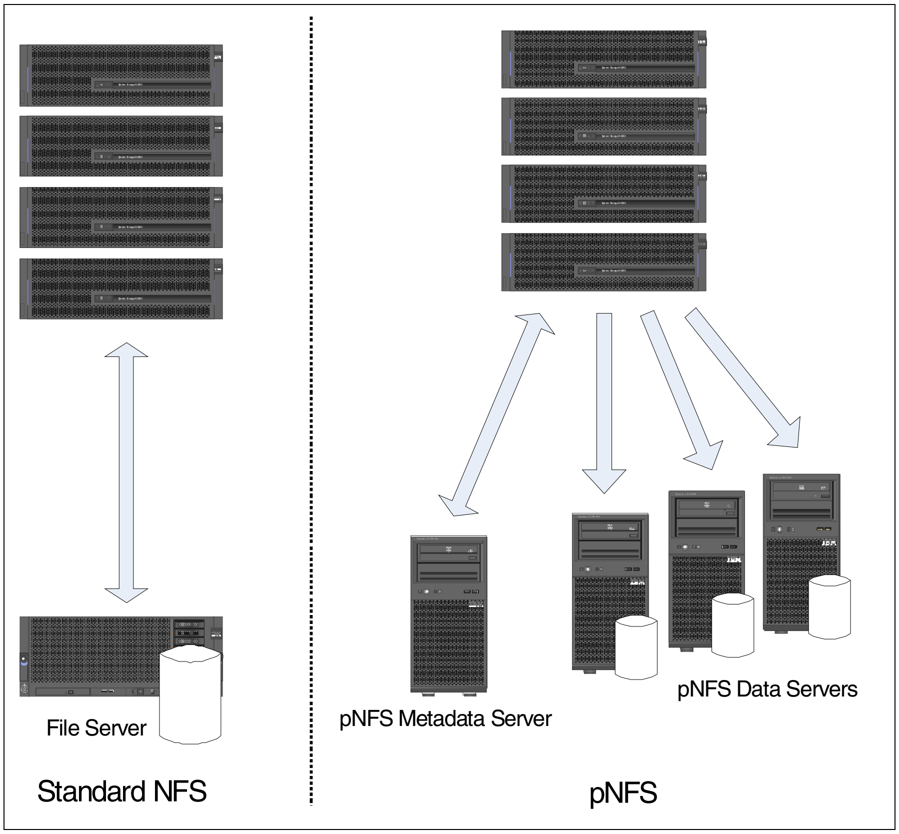


Figure 10-3 Standard NFS versus pNFS

pNFS will offer a number of advantages over existing shared file system options. In addition to parallel I/O and support for a broad range of hardware, pNFS provides these capabilities:

- Application transparency:

  Applications will be able to access pNFS data sources without any code changes.

- Integration:

  Because this feature is included in OS, it will not require any installation, recompiling, debugging, and so on.

- No client-side software:

  There is a significant consideration for large compute clusters. It can take hours to install software and drivers on large numbers of clients, not to mention ongoing maintenance, patches, and so on.

10.4.2 pNFS: How it works

In this section we discuss various considerations regarding pNFS.

pNFS: Part 1 (the NFS v3 process)

First of all, take a look at Figure 10-4 to check the normal path on the NFS Version 3. When a client tries to access a volume through a data LIF, the request traverses the cluster network. The result is returned to the client along the same path. In normal cases, it should be OK, but if there is a lot of traffic on the cluster network, it might impact the performance for the whole system.

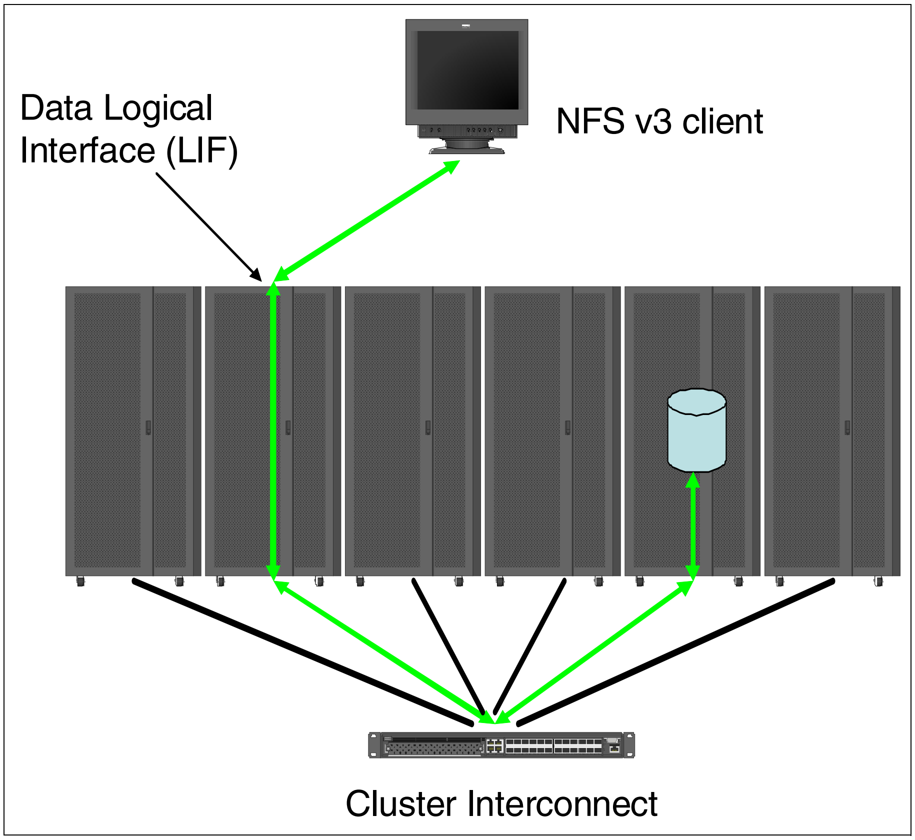


Figure 10-4 The NFS v3 process

pNFS: Part 2 (pNFS original data path)

With pNFS, when a file is opened by an NFS v4.1 pNFS client, the mounted data LIF on the second node serves as a metadata path because this path is used to carry out discovery of the target volume's location. If the data is hosted by the second node, the operation is managed locally. In this case, the local node discovers that the data is on the fifth node. Based on the pNFS protocol, the client is redirected to an LIF that is hosted on the fifth node. This request and subsequent requests to the volume are serviced locally and bypass the cluster network. See Figure 10-5.

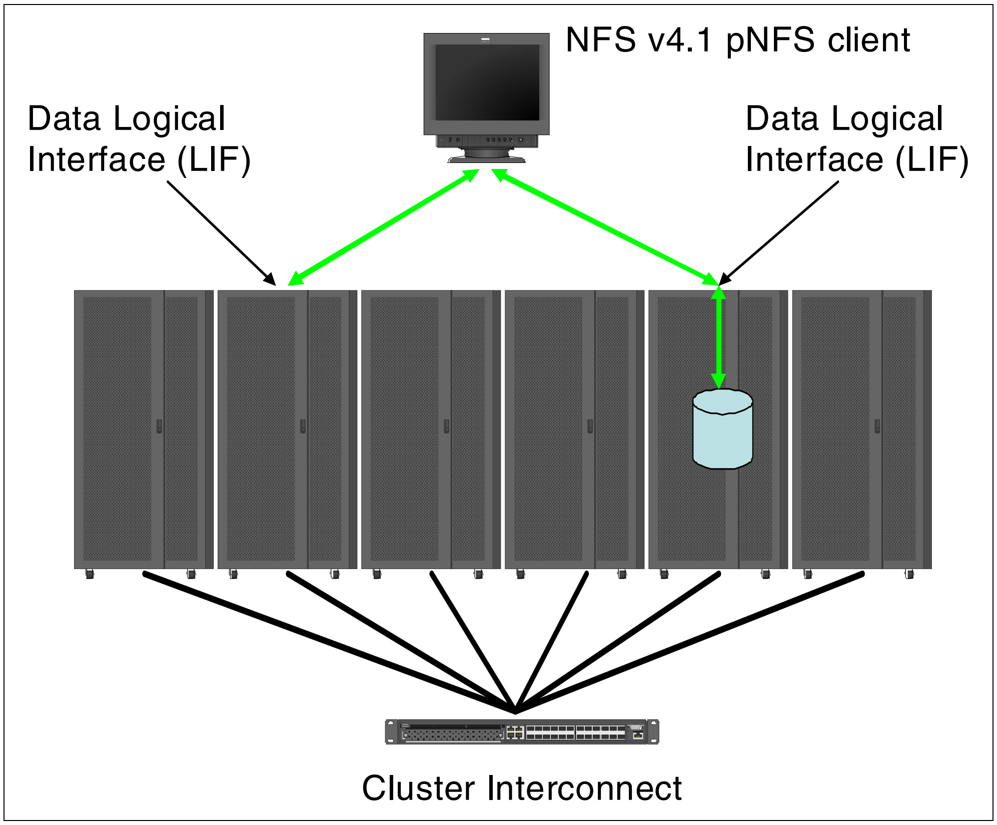


Figure 10-5 pNFS original data path

pNFS: Part 3 (pNFS new data path)

When a volume is moved to an aggregate on a different node, the pNFS client data path is redirected to a data LIF hosted on the destination node. In this example, because the volume is moved to the third node, the client is redirected to an LIF that is hosted on the third node automatically. All of the jobs were done by pNFS protocol. As we know, the client software does not require any code changes and so on. And other traffic will not waste the bandwidth for the network or cluster interconnect. See Figure 10-6.

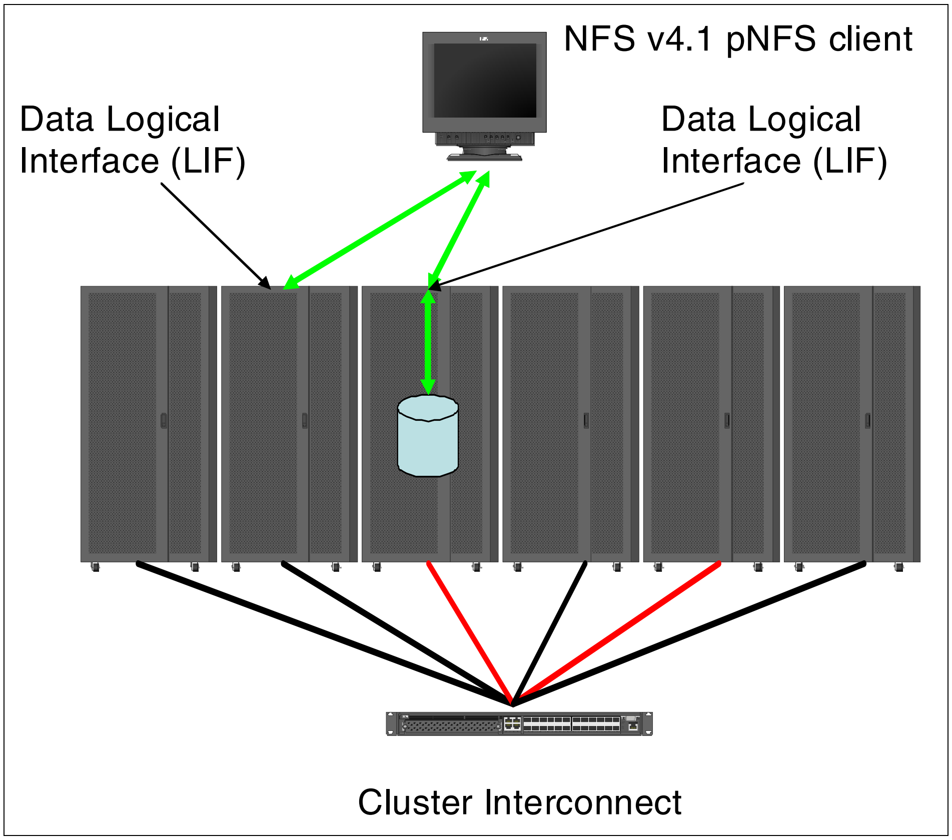


Figure 10-6 pNFS new data path

10.4.3 pNFS configuration in Clustered Data ONTAP

With emergent availability of an end-to-end pNFS solution, you might want to consider including pNFS in your storage plans for scientific, engineering, business, and enterprise workloads. You can begin to prepare for the transition as follows:

- Review how your file data is stored and served now and how it will need to be structured in the future.
- Talk to your operating system and other related vendors to find out their plans for pNFS. Your application vendors might also provide guidance about storage needs going forward.
- If you will be implementing an N series cluster Data ONTAP solution, talk to your representative to understand how you can smoothly transition to scale-out storage.
- Understand NFSv4 and NFSv4.1. If you begin the transition now, not only will you gain the advantages of NFSv4, but also it will simplify the process of transitioning client systems to pNFS.

By taking a few appropriate steps now, you will be able to make a smooth transition to pNFS, with less disruption and better results.

To enable pNFS feature on the NAS side, do the command shown in Example 10-4.

Example 10-4 Enable a pNFS feature

```
cdot-cluster01::> vserver nfs modify -v4.1-pnfs enabled
```

Also, pNFS cannot coexist with referrals, meaning that the client does not need a remount and so on.

pNFS is supported by Red Hat Enterprise Linux 6.2, Fedora 14, or any kernel Version 2.6.39 and higher. But at the time of writing this chapter, Red Hat Enterprise Linux 6.2 or above is certified only with Clustered Data ONTAP. Check for the latest interoperability before your implementation of pNFS.

10.5 Further information

More details on SAN and the Fibre Channel protocol can be found in the following Redbooks publications:

- *Clustered Data ONTAP 8.2 File Access and Protocols Management Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

- *Data ONTAP 8.2 Release notes*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

- *OnCommand System Manager 3.0 Help for Use with Clustered Data ONTAP*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

11

SAN protocols

This chapter describes SAN network protocols for the IBM System Storage N series Clustered Data ONTAP storage system and characteristics for each SAN protocol.

We provide information to help you determine what is the correct protocol for your environment and which protocol should be used for each case.

A SAN is a block-based storage system that uses Fibre Channel (FC), Fibre Channel over Ethernet (FCoE), and iSCSI. In the Data ONTAP 8.2 operating system, SAN is currently supported in clusters of up to eight nodes.

The following topics are covered:

- Clustered Data ONTAP with SAN protocols
- Fiber Channel (FC)
- iSCSI
- FCoE
- Further information

11.1 Clustered Data ONTAP with SAN protocols

In Clustered Data ONTAP 8.2, scalable SAN support has been expanded to up to eight nodes, increasing capacity for storage, CPU cycles, and network bandwidth for clustered SAN solutions, with no need to increase management and administrative resources.

Clustered Data ONTAP 8.2 continues to support Windows, Red Hat Linux, VMWare ESX, and Solaris hosts, and also added support for IBM AIX® as a SAN host. To function with scalable SAN, all SAN client stacks must support Asymmetric Logical Unit Access (ALUA).

Clustered Data ONTAP 8.2 and Data ONTAP operating in 7-Mode show some differences due to the various architectures. See Table 11-1.

Table 11-1 Configuration differences between Clustered Data ONTAP and 7-Mode

Configuration detail	Data ONTAP 7-Mode	Clustered Data ONTAP
iSCSI secondary paths	Failover to partner interface	ALUA based
Fibre Channel indirect paths	Over NVRAM interconnect	Over cluster network
Fibre Channel ports	Physical ports	Virtual ports (NPIVs)
Portsets	Optical	Advised
interface with SanpDrive	Any active Ethernet port	SVM management LIF
FC and iSCSI service scope	Per-node/per-HA pair	Per SVM

Preferred practice:

When creating iSCSI or Fibre Channel logical interfaces (LIFs) for the first time for an existing storage virtual machine (SVM), make sure that the Fibre Channel and/or iSCSI service for the SVM has been created and is turned on by using the `fcp show` or `iscsi show` command, or by navigating to the **Cluster** → **Vserver** → **Configuration** → **Protocols** pane in OnCommand System Manager.

Note: This step is not necessary if the SVM was originally set up to serve these protocols by using either the `vserver setup` command or System Manager Vserver Setup Wizard.

11.1.1 Volume configuration

When provisioning volumes in a cluster or in Data ONTAP operating in 7-Mode, many considerations regarding deduplication, space reservations, and storage efficiency are the same. One major difference is that volumes in Clustered Data ONTAP are oriented to SVM containers instead of to individual nodes, and a side effect is that they can be mapped into an SVM-wide global namespace for the purpose of exporting file systems by using NFS or CIFS protocols. However, the presence or absence of a given volume in the global namespace has no effect on data that is served by using Fibre Channel or iSCSI.

Note: Volumes that contain LUNs do not need to be junctioned to the global namespace to serve data by using block protocols; they only require an igroup-to-LUN mapping.

11.1.2 Host connectivity

Hosts that access data served by Clustered Data ONTAP using a block protocol are expected to make use of the ALUA extension to the SCSI protocol to determine which paths are direct and which are indirect. The ALUA standard refers to direct paths as active/optimized and to indirect paths as active/nonoptimized. All ALUA information is requested and delivered in-band, using the same iSCSI or Fibre Channel connection that used for data.

The status of a given path is discoverable by a host that sends a path status inquiry down each of the paths it has discovered. This path status inquiry can be triggered when the storage system sends extra data along with the result of a SCSI request to inform a host that path statuses have been updated and that their priorities should be rediscovered.

ALUA is a well-known and widely deployed standard and is a requirement for access to data served by Clustered Data ONTAP. Any operating systems tested and qualified to work with Clustered Data ONTAP block access protocols will support ALUA.

See Figure 11-1 to see how ALUA works.

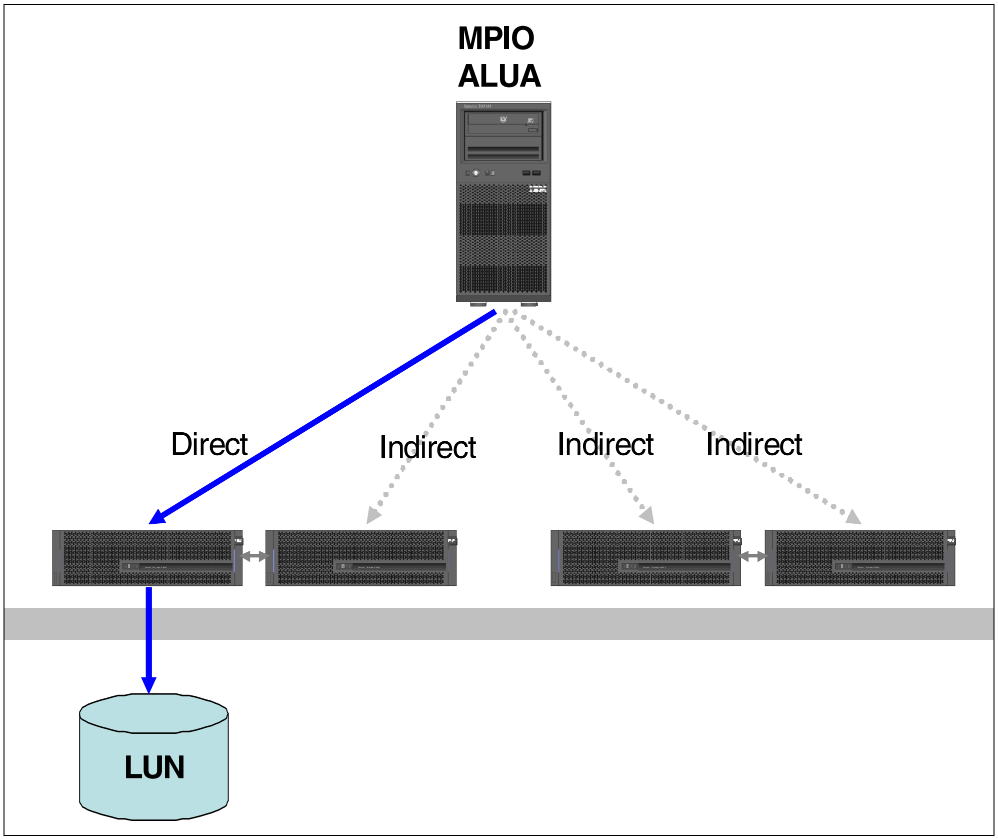


Figure 11-1 ALUA and MPIO with direct and indirect paths

11.1.3 Path selection

You must configure SAN clients to use the following features:

- Multipath I/O (MPIO) to access LUNs
- ALUA to determine the state of a given data path to LUNs

Even though every LIF owned by an SVM accepts writes and read requests for its LUNs, only one of the cluster nodes actually owns the disks backing that LUN at any given moment. This effectively divides available paths to a LUN into two types, direct paths and indirect paths.

The “active-optimized” path to a LUN means the path for which the LIF and LUN are hosted by the same node. The “active-nonoptimized path” represents the path for which the LIF and LUN are hosted on separate nodes.

As you already saw the previous Figure 11-1 on page 171, a *direct path* for a LUN is a path where an SVM’s LIFs and the LUN being accessed reside on the same node. To go from a physical target port to disk, it is not necessary to traverse the cluster network.

Indirect paths are data paths where an SVM’s LIFs and the LUN being accessed reside on different nodes. Data must traverse the cluster network in order to go from a physical target port to disk. Because the cluster network is fast and highly available, this does not add a great deal of time to the round trip, but it is not the maximally efficient data path.

Unlike NAS LIFs, SAN LIFs do not migrate between interfaces or nodes. Therefore, the client uses ALUA to determine the most efficient path (or paths) to communicate to the LUN. The active-optimized paths become the primary paths for data transfer between the host and the LUN.

Here are possible path priority selection states through ALUA:

- Active-optimized (direct)
- Active-nonoptimized (indirect)
- Standby (not implemented in the Data ONTAP operating system)
- Unavailable

11.1.4 Path selection changes

There are three events that could change the path selected by a host to access data on a cluster.

HA failover

In an HA failover event, LIFs on the down node are seen as offline, and LIFs on the HA partner that has taken over for the down node are now direct paths. This state changes automatically due to ALUA path inquiry, and no administrative changes are necessary. See Figure 11-2.

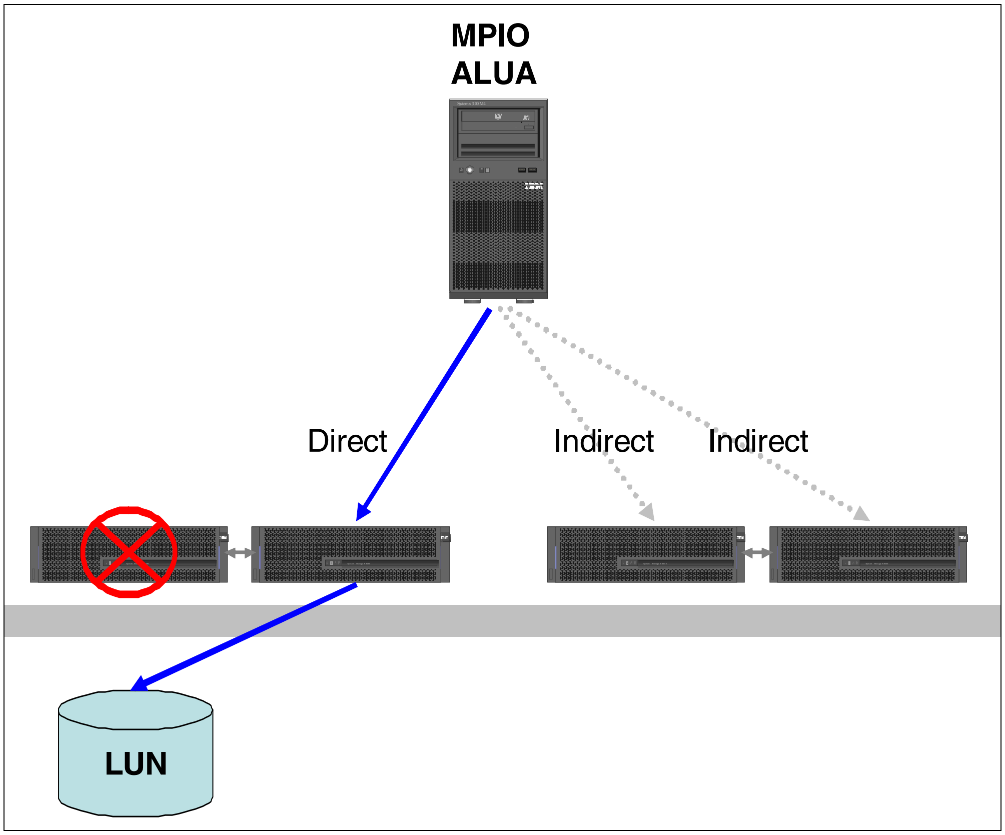


Figure 11-2 MPIO and ALUA path changes during an HA failover

Port or switch failover

In a port or switch failure, no more direct paths are available. Path priority remains the same, and MPIO software running on the host selects alternate indirect paths until a direct path becomes available again.

The ALUA path states do not change.

See Figure 11-3.

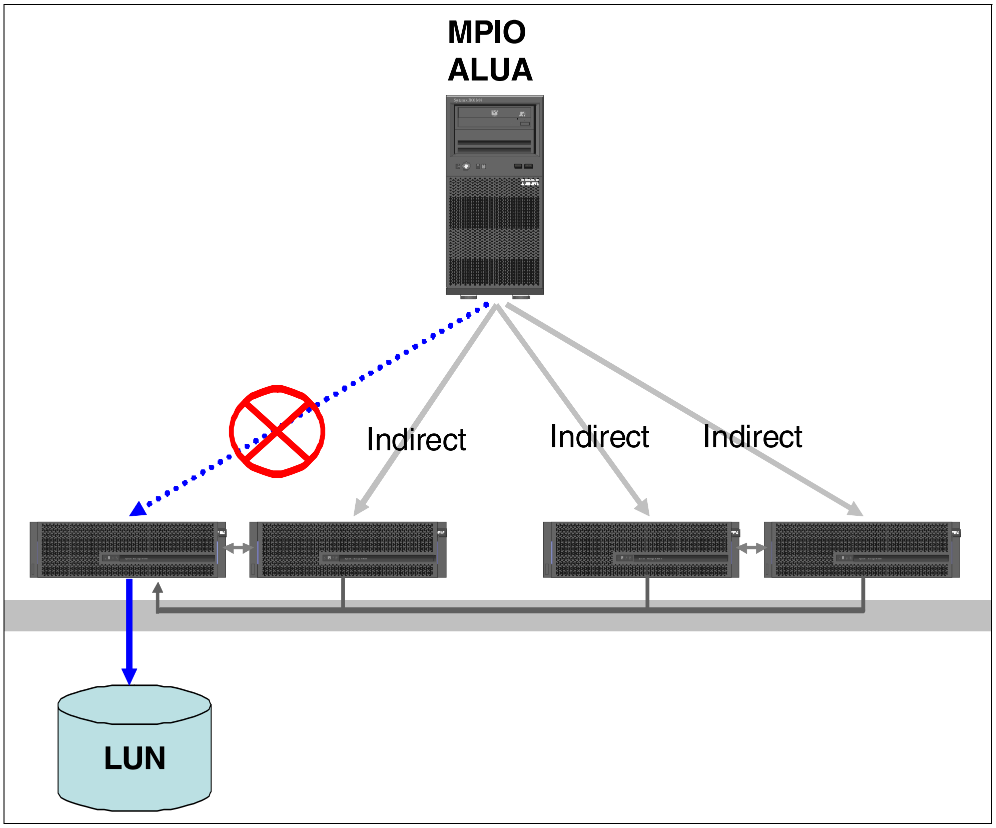


Figure 11-3 Port or switch failure

Volume movement

A volume is moved transparently between nodes by using **`volume move`** functionality. When the cutover occurs and the volume's new node begins to handle read and write requests, the path status is updated so that the new node has direct paths and the old node has indirect paths. All paths remain available at all times.

Figure 11-4 illustrates the original "direct path" before the volume movement.

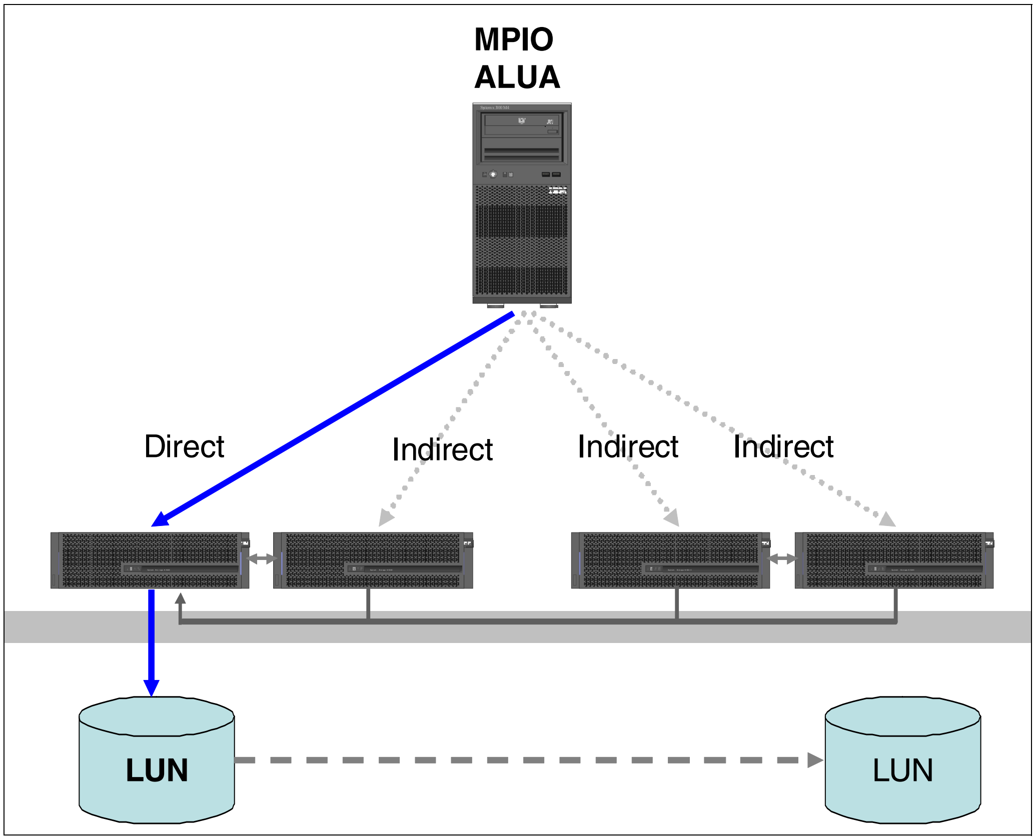


Figure 11-4 MPIO and ALUA path changes during a volume movement

After completing the volume movement, the direct path will be changed to the node which has a physical path where the LUN moved. See Figure 11-5.

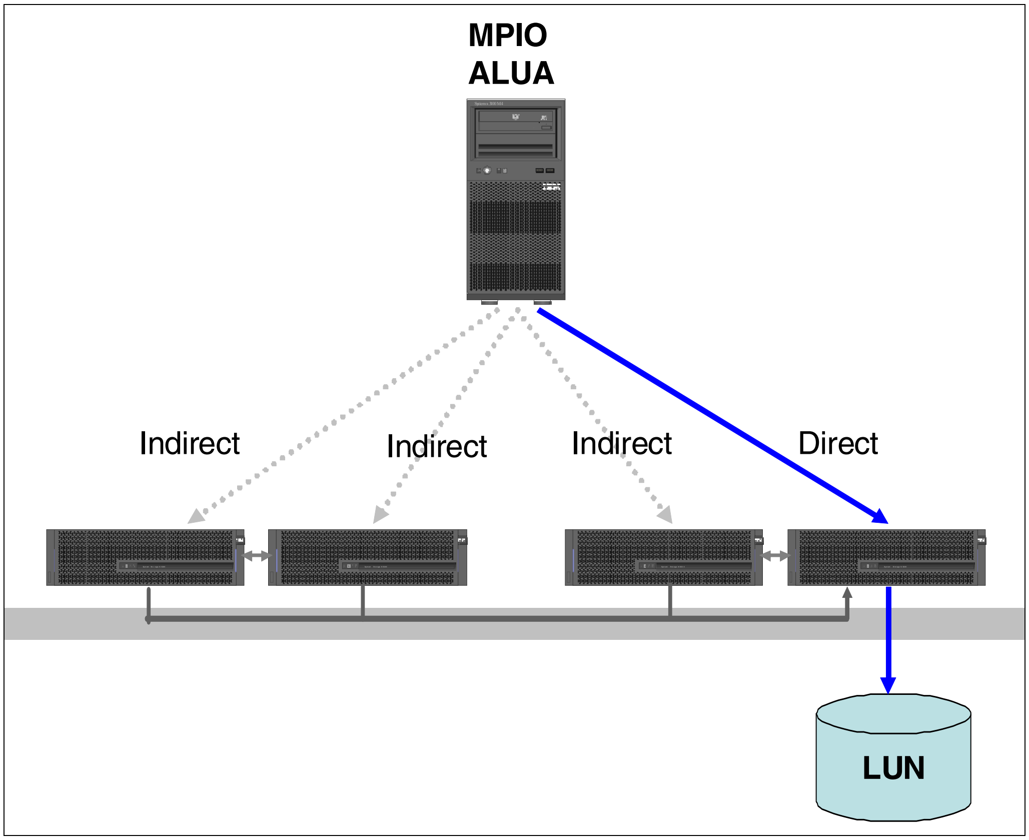


Figure 11-5 MPIO and ALUA path changes after the volume movement

iSCSI and the partner interface

Data ONTAP operating in 7-Mode uses a partner interface to provide redundancy for iSCSI connections. During an HA failover event, the HA partner of the affected storage controller has a predesignated partner interface that is brought online with the IP address of the taken-over HA partner’s iSCSI target interface, and I/O resumes.

In Clustered Data ONTAP, iSCSI partner interfaces are not assigned when configuring iSCSI. Instead, iSCSI connections are made to the node where the data resides, to its HA partner for redundancy, and optionally to additional nodes in the same cluster. Instead of the path reappearing after an HA takeover on the HA partner, MPIO on the host determines that the primary path is no longer available and resumes I/O over the already-established connection to the taken over node’s HA partner. This means that, unlike in 7-Mode, iSCSI connections to a cluster have both direct and indirect paths to data managed by host MPIO software, operating in the same way as Fibre Channel paths.

Fibre Channel and NPIV

Clustered Data ONTAP uses N_Port ID Virtualization (NPIV) to permit every logical interface to log into an FC fabric with its own World Wide Port Name (WWPN), rather than using a single WWNN and associated WWPNs based on the address of the HA pair's physical FC target adapters, as when operating in 7-Mode. This permits a host connected to the same FC fabric to communicate with the same SCSI target regardless of which physical node owns which LIF. The virtual port presents the SCSI target service and sends and receives data.

Note: NPIV is required for Fibre Channel LIFs to operate correctly. Before creating FC LIFs, make sure that any fabrics attached to a Clustered Data ONTAP system have NPIV enabled.

In the case of Cisco NX-OS switches, you can check NPIV enabled as shown in Example 11-1.

Example 11-1 Check NPIV enabled in Cisco NX-OS switches

```
NX-OS# show npiv status
NPIV is enabled
```

If the SAN switches are using Brocade Fabric OS, use the command **portcfgshow** to check NPIV capability and status.

From the storage administration console, it is not possible to inquire about NPIV status on an attached switch directly, but examining the local FC topology can show whether fabric switch ports have NPIV enabled.

In Example 11-2, NPIV must be enabled, because port 2/1 has multiple attached WWPNs, the home of which are virtual ports.

Example 11-2 Check NPIV enabled through the cluster shell

```
cdot-cluster01::> node run -node node01 fcp topology show
Switch Name: N5K-A
Switch Vendor: Cisco Systems, Inc.
Switch Release: 5.0(2)N2(1a)
Switch Domain: 200
Switch WWN: 20:66:00:0d:ec:b4:94:01
Port Port WWPN              State  Type   Attached WWPN           Port ID
--------------------------------------------------------------------------------
2/1 20:01:00:0d:ec:b4:94:3f Online F-Port 50:0a:09:83:8d:4d:bf:f1 0xc80033
                                          20:1c:00:a0:98:16:54:8c 0xc80052*
                                          20:0e:00:a0:98:16:54:8c 0xc80034*
                                          20:10:00:a0:98:16:54:8c 0xc8003f
2/3 20:03:00:0d:ec:b4:94:3f Online F-Port 50:0a:09:83:8d:3d:c0:1c 0xc8002c
```

Note: Physical WWPNs (beginning with 50:0a:09:8x) do not represent a SCSI target service and should not be included in any zone configuration on the FC fabric. Use only virtual WWPNs (visible in the output of the **network interface show** command and in the **System Manager Cluster** → **Vserver** → **Configuration** → **Network Interface** pane.

Portsets

Clusters with more than two nodes are likely to have more paths than has commonly been the case in the past. Clusters attached to more than one fabric, or with nodes attached more than once per fabric, can quickly multiply the number of potential paths available.

Portsets permit administrators to mask an interface group (igroup) so that the LUNs that are mapped to it are available on a subset of the total number of available target ports. This functionality is available in both Clustered Data ONTAP and 7-Mode. Portsets are much more likely to be useful in Clustered Data ONTAP, because higher potential path counts are supported.

Note: An LIF that is currently a member of a portset cannot be modified until it is removed from the portset. It can be added to the portset after modification, but care should be taken to leave enough LIFs in the portset to satisfy host requirements for a path to data.

Management interfaces

Because LIFs belonging to SVMs that serve data by using block protocols cannot also be used for administration purposes, and because the logical unit of management in Clustered Data ONTAP is the SVM, every SVM must have a management interface in addition to interfaces that are serving data using block protocols.

This interface should have the following properties:

- An LIF type of data
- No data protocols assigned (-data-protocols none)
- A firewall policy that permits management access (-firewall-policy mgmt)

Note: The management LIF should be assigned to a failover group that makes it accessible to hosts that might need to contact it for data management purposes, such as creating or managing Snapshot copies by using SnapDrive. For more information about failover groups, see *Configuring Failover Groups for LIFs* in *the Clustered Data ONTAP Network Management Guide*.

With the foregoing changes and characteristics in Clustered Data ONTAP, you will configure SAN protocols. Refer to each chapter for the actual exercises.

11.2 Fiber Channel (FC)

When an SVM is first created and a block protocol (FC or iSCSI) is enabled, the SVM gets either a Fibre Channel worldwide name (WWN) or an iSCSI qualified name (IQN), respectively. This identifier is used irrespective of which physical node is being addressed by a host, with Data ONTAP making sure that SCSI target ports on all of the cluster nodes work together to present a virtual, distributed SCSI target to hosts that are accessing block storage.

In this section, you will check three SAN protocols and any special enhancements in Clustered Data ONTAP v8.2.

Note: IBM AIX using Fibre Channel to access data on a cluster is supported beginning with Clustered Data ONTAP 8.2. For the supported AIX technology levels and service packs, refer to the Interoperability Matrix. See the following link:

http://www.ibm.com/systems/storage/network/interophome.html

11.2.1 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.

11.2.2 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.

11.2.3 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.

11.2.4 How FC nodes are identified

Each FC node is identified by a worldwide node name (WWNN) and a worldwide port name (WWPN).

11.2.5 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.

11.3 iSCSI

In this section, we introduce high-level iSCSI concepts.

Note: The iSCSI protocol is currently not supported with AIX and Clustered Data ONTAP.

11.3.1 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 N3xxx, N5xxx, and N6xxx 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).

11.3.2 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.

11.3.3 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 11-3.

Example 11-3 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
```

11.3.4 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
```

11.3.5 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

11.3.6 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.

11.3.7 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.

11.3.8 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.

11.3.9 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.

11.3.10 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.

11.3.11 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.

11.4 FCoE

This section provides a detailed discussion of Fibre Channel over Ethernet (FCoE).

11.4.1 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).

11.4.2 Fibre Channel over Ethernet

Fibre Channel over Ethernet 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 these 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 100 Gbps 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 11-6 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.

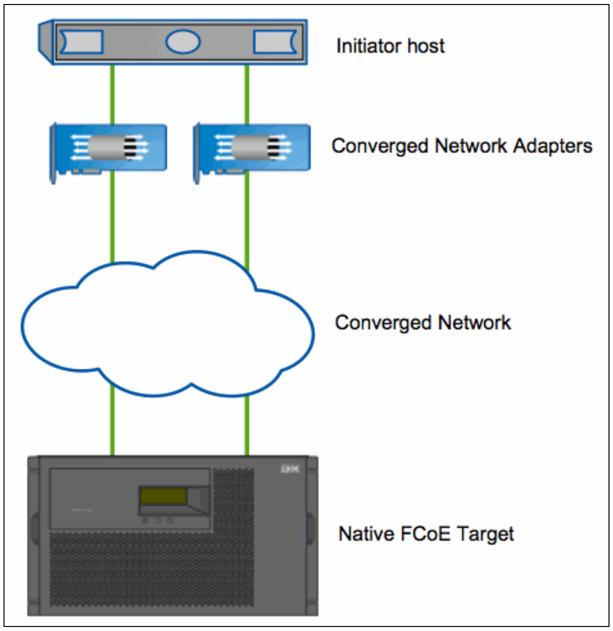


Figure 11-6 Implemented converged network

11.4.3 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.

Because the FC frames are transported with the FC header all encapsulated in the Ethernet frame (see Figure 11-7), 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.

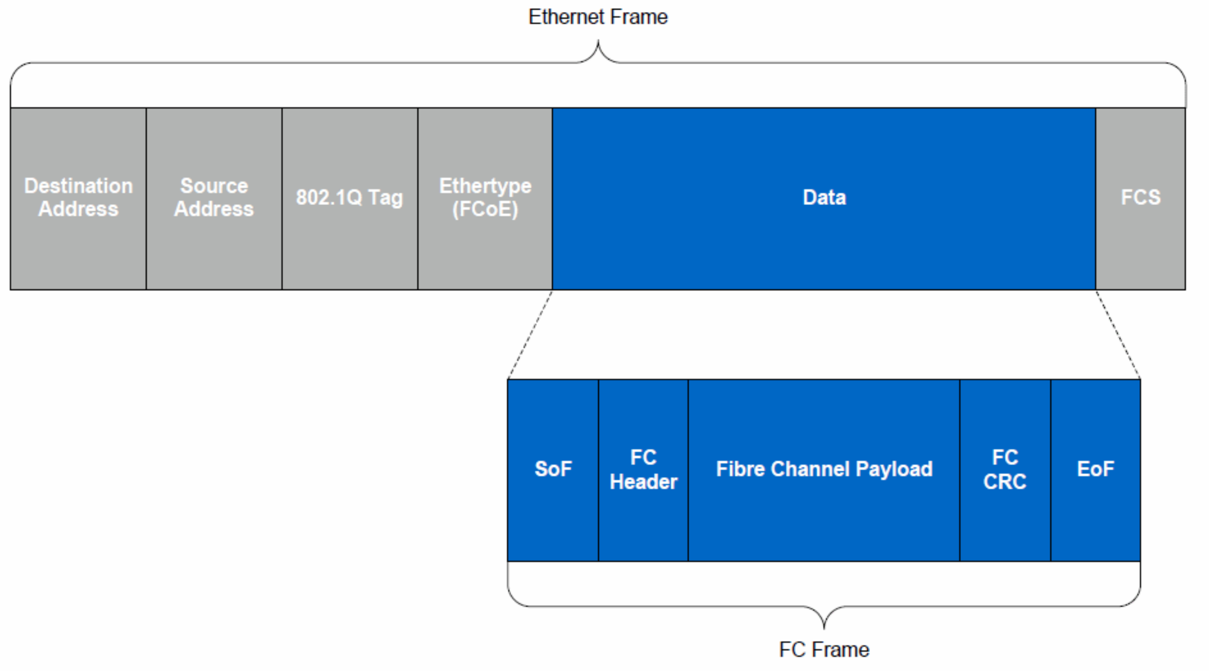


Figure 11-7 FCoE sample frame

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 some 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 an 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).

11.5 Further information

More details on SAN and the Fibre Channel protocol can be found here:

- *Designing an IBM Storage Area Network*, SG24-5758, which is located at this website:

  `http://www.redbooks.ibm.com/abstracts/sg245758.html?Open`

- *Introduction to Storage Area Networks and System Networking*, SG24-5470, which is located at this website:

  `http://www.redbooks.ibm.com/abstracts/sg245470.html?Open`

More details on the iSCSI protocol can be found here:

- *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240, which is located at this website:

  `http://www.redbooks.ibm.com/abstracts/sg246240.html?Open`

More details on converged networking and the FCoE protocol can be found here:

- *Storage and Network Convergence Using FCoE and iSCSI,* SG24-7986, which is located at this website:

  `http://www.redbooks.ibm.com/abstracts/sg247986.html?Open`

12

Ancillary protocols

This chapter describes other protocols that can be used with N series systems. Being a Unified Storage solution, the N series provides more by far than CIFS, NFS, FCP, and iSCSI access. Clustered Data ONTAP solutions can scale from 1 to 24 nodes, and from the customer point of view, data backup is also crucial. Network Data Management Protocol (NDMP) is one possible backup solution.

Clustered Data ONTAP 8.2 does not currently support data access over FTP, SFTP, or HTTP.

In this chapter, we describe only an NDMP backup solution for Clustered Data ONTAP.

The following topics are covered:

- Network Data Management Protocol (NDMP)
- Further information about SAN and Fibre Channel

12.1 Network Data Management Protocol (NDMP)

Clustered Data ONTAP provides a robust feature set, including data protection features such as Snapshot copies, intracluster asynchronous mirroring, SnapVault backups, and NDMP backups. SnapMirror is used for disaster recovery and maintains only one read-only replica of the source volume. SnapVault is used for creating read-only archival copies of your source volume. And Data ONTAP uses NDMP with third-party software for disaster recovery.

> **Note:** NDMP Version 3 is not supported beginning with Data ONTAP 8.2; only NDMP Version 4 is supported.

12.1.1 About NDMP modes of operation

Starting with Data ONTAP 8.2, you can choose to perform tape backup and restore operations either at a node level as you have been doing until now or at a storage virtual machine (SVM) level. To perform these operations successfully at the SVM level, NDMP service must be enabled on the SVM.

> **Note:** You cannot use System Manager to enable, disable, or stop the NDMP service on storage systems running Clustered Data ONTAP 8.2. However, for storage systems running Clustered Data ONTAP 8.1 or a later version in the 8.1 release family, NDMP management is supported.
>
> For storage systems running Data ONTAP 8.2, you should use the *command-line interface* to manage the NDMP service.

In a mixed cluster where nodes are running different versions of Data ONTAP 8.2 and earlier versions in the Data ONTAP 8.x release family, NDMP follows the node-scoped behavior (at the node level). This NDMP behavior continues even after upgrading to Data ONTAP 8.2. You must explicitly disable the node-scoped NDMP mode to perform tape backup and restore operations in the SVM aware mode.

In a newly installed cluster where all nodes are running Data ONTAP 8.2, NDMP is in the SVM aware mode (at the SVM level) by default. To perform tape backup and restore operations in the node-scoped NDMP mode, you must explicitly enable the node-scoped NDMP mode.

> **Note: The NDMP default password must be changed.**
>
> - In the node-scoped NDMP mode, you must use NDMP specific credentials to access a storage system in order to perform tape backup and restore operations.
> - The default user ID is “root”. Before using NDMP on a node, you must change the default NDMP password associated with the NDMP user. You can also change the default NDMP user ID.

12.1.2 Tape backup of FlexVol volumes with NDMP

Clustered Data ONTAP uses NDMP with third-party software for disaster recovery. No native tape backup or restore commands are currently available in Clustered Data ONTAP. All tape backups and restores are performed through third-party NDMP applications.

> **Note:** For Infinite Volumes, Data ONTAP supports tape backup and restore through a mounted volume. Infinite Volumes do not support NDMP.

Tape backup of FlexVol volumes using NDMP

NDMP allows you to back up storage systems directly to tape, resulting in efficient use of network bandwidth. Clustered Data ONTAP supports dump engine for tape backup. Dump is a Snapshot copy-based backup to tape, in which your file system data is backed up to tape. The Data ONTAP dump engine backs up files, directories, and the applicable access control list (ACL) information to tape. You can back up an entire volume, an entire qtree, or a subtree that is neither an entire volume nor an entire qtree. Dump supports level-0, differential, and incremental backups.

You can perform a dump backup or restore by using NDMP-compliant backup applications. Starting with Data ONTAP 8.2, only NDMP Version 4 is supported.

Tape backup of Infinite Volumes using a mounted volume

You can back up and restore Infinite Volumes using any data management application that can back up files over a volume mounted with the NFS or CIFS protocols and that supports SnapDiff.

However, you cannot back up or restore Infinite Volumes with NDMP.

Tape backup and restore workflow for FlexVol volumes

The following high-level tasks are required to perform a tape backup and restore operation:

1. Set up a tape library configuration choosing an NDMP-supported tape topology.
2. Enable NDMP services on your storage system.

   You can enable the NDMP services either at a node level or at an SVM level. This depends upon the NDMP mode in which you choose to perform a tape backup and restore operation.
3. Use NDMP options to manage NDMP on your storage system.

   You can use NDMP options either at a node level or at an SVM level. This depends upon the NDMP mode in which you choose to perform a tape backup and restore operation.
4. Perform a tape backup or restore operation by using NDMP-enabled backup application.

   Clustered Data ONTAP supports dump engine for tape backup and restore. For more information about using the backup application (also called Data Management Applications or DMAs) to perform backup or restore operations, see your backup application documentation.

A Clustered Data ONTAP system can be an **ndmpcopy** source or destination. The NDMP destination path is of the format */svm_name/volume_name*.

Local, three-way, and remote NDMP backups

NDMP includes the following types of backups:

- Local backup: This is the simplest configuration. A backup application backs up data from an N series system to a locally attached (or SAN attached) tape device.
- Three-way backup: This configuration allows storage system data to be backed up from the storage system to another storage system that has a locally attached (or SAN attached) tape device.
- Remote NDMP backup: Like three-way backup, this configuration needs TCP/IP network and bandwidth. The data is backed up from a storage system to backup application host that has a tape divce(tape drive or tape library).

Figure 12-1 illustrates these types of backups.

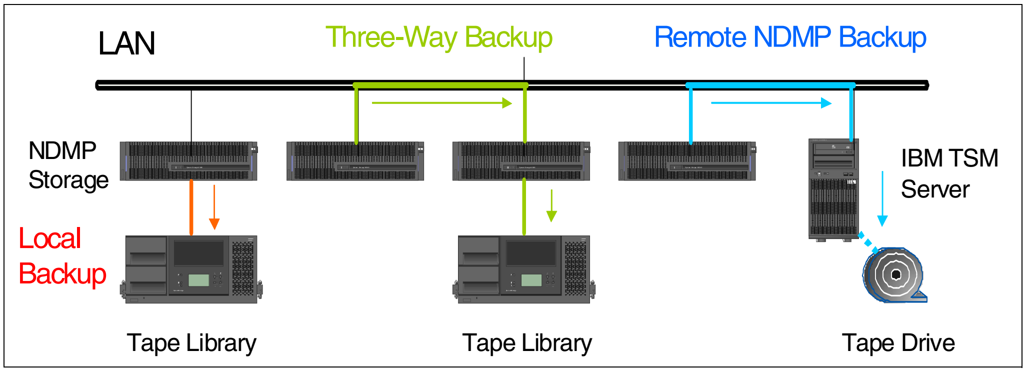


Figure 12-1 Local, remote, and three-way NDMP backup

Direct Access Recovery (DAR)

Data ONTAP 8.0 or later supports enhanced DAR.

DAR is the ability of a data management application to restore a selected file of selected files without the need to sequentially read the entire tape or tapes are involved in a backup.

Enabling enhanced DAR functionality might impact the backup performance because an offset map has to be created and written onto tape. You can enable or disable enhanced DAR in both the node-scoped and SVM aware NDMP modes.

12.1.3 SVM-aware NDMP

Before introducing SVM-aware NDMP, we need to know what Cluster Aware Backup extension does.

What Cluster Aware Backup is

Cluster Aware Backup (CAB) is an NDMP Version 4 protocol extension. This extension enables the NDMP server to establish a data connection on a node that owns a volume. This also enables the backup application to determine if volumes and tape devices are located on the same node in a cluster.

To enable the NDMP server to identify the node that owns a volume and to establish a data connection on such a node, the backup application must support the CAB extension. CAB extension requires the backup application to inform the NDMP server about the volume to be backed up or restored prior to establishing the data connection. This allows the NDMP server to determine the node that hosts the volume and appropriately establish the data connection.

With the CAB extension supported by the backup application, the NDMP server provides affinity information about volumes and tape devices. Using this affinity information, the backup application can perform a local backup instead of a three-way backup if a volume and tape device are located on the same node in a cluster.

SVM-aware NDMP

Starting with Data ONTAP 8.2, you can perform tape backup and restore operations at an SVM level successfully if the NDMP service is enabled on the SVM. You can back up and restore all volumes hosted across different nodes in a cluster of an SVM if the backup application supports the CAB extension.

You can add NDMP in the allowed or disallowed protocols list by using the **`vserver modify`** command. By default, NDMP is in the allowed protocols list. If NDMP is added to the disallowed protocols list, NDMP sessions cannot be established.

An NDMP control connection can be established on different logical interface (LIF) types. In an SVM aware NDMP mode, these LIFs belong to either the data SVM or admin SVM. Data LIFs belong to the data SVM and the intercluster LIF. Node management LIFs and cluster management LIFs belong to the admin SVM.

The NDMP control connection can be established on an LIF only if the NDMP service is enabled on the SVM that owns this LIF. In an SVM context, the availability of volumes and tape devices for backup and restore operations depends upon the LIF type on which the NDMP control connection is established and the status of the CAB extension. If your backup application supports the CAB extension and a volume and tape device share the same affinity, then the backup application can perform a local backup or restore operation instead of a three-way backup or restore operation.

You can also manage NDMP on a per SVM basis by using the NDMP options and commands. In the SVM aware NDMP mode, user authentication is integrated with the role-based access control mechanism. To perform tape backup and restore operations in the node-scoped NDMP mode, you must explicitly enable the node-scoped NDMP mode.

Note: Backups do not traverse junctions; you must list every volume to be backup up.

12.1.4 Configuring for NDMP

Enable and configure NDMP on the node or nodes:

```
cdot-cluster01::> system services ndmp modify
```

Identify tape and library attachments:

```
cdot-cluster01::> system node hardware tape drive show
```

```
cdot-cluster01::> system node hardware tape library show
```

Configure the data management application (such as IBM Tivoli® Storage Manager) for NDMP.

12.1.5 Clustered Data ONTAP and NDMP

Clustered Data ONTAP supports the IBM Tivoli Storage Manager and Symantec NetBackup data management applications, and more are being added.

Clustered Data ONTAP supports local NDMP, remote NDMP, and three-way NDMP backup.

A data management application with DAR can restore selected files without sequentially reading entire tapes.

12.1.6 Preferred practices for disaster recovery with NDMP

Here are some ideas to consider to plan for disaster recovery with NDMP:

- Enable Snapshot copies and data-protection mirror copies for critical volumes:
  - Consider putting data-protection mirror copies on SATA disks.
  - Use data-protection mirror copies on SATA disks as a disk-based backup solution.
  - Use intercluster data-protection mirror copies for off-site backups.
- Plan disaster-recovery implementations carefully by considering taking quorum and majority rules. (You can recover an out-of-quorum site, but doing so is not customer-friendly.)
- Use NDMP to back up important volumes to tape.
- Have a policy for rotating backups off-site for disaster recovery.

Tape backup policy during volume move, SFO, and ARL

You can continue performing incremental tape backup and restore operations after volume move, storage failover (SFO), and aggregate relocation (ARL) operations in the SVM aware NDMP mode only if your backup application supports the Cluster Aware Backup (CAB) extension.

If your backup application does not support the CAB extension or if you are using the node-scoped NDMP mode, then you can continue performing an incremental backup operation only if you migrate the LIF that is configured in the backup policy to the node that hosts the destination aggregate. Else, after volume migration, you must perform a baseline backup prior to performing the incremental backup operation.

> **Note:** For SFO operations, the LIF configured in the backup policy must migrate to the partner node.

For more information about this behavior, see the Clustered *Data ONTAP Data Protection Tape Backup and Recovery Guide.*

Considerations when using NDMP

Keep in mind the following considerations for NDMP:

- NDMP services can generate file history data at the request of NDMP backup applications.

  File history is used by backup applications to enable optimized recovery of selected subsets of data from a backup image. File history generation and processing might be time-consuming and CPU-intensive for both the storage system and the backup application.

If your data protection needs are limited to disaster recovery, where the entire backup image will be recovered, you can disable file history generation to reduce backup time. See your backup application documentation to determine if it is possible to disable NDMP file history generation.

- Firewall policy for NDMP is enabled by default on all LIF types. For information about managing firewall service and policies, see the *Clustered Data ONTAP System Administration Guide for Cluster Administrators.*
- In the node-scoped NDMP mode, to back up a FlexVol volume you must use the backup application to initiate a backup on a node that owns the volume. However, you cannot back up a node root volume.
- You can perform NDMP backup from any LIF as permitted by the firewall policies. If you use a data LIF, you must select one that is not configured for failover. If a data LIF fails over during an NDMP operation, the NDMP operation fails and must be re executed.
- In the node-scoped NDMP mode, NDMP data connection uses the same LIF as the NDMP control connection.
- NDMP backup path is of the format /vserver_name/volume_name/path_name, where path_name is the path of the directory, file, or Snapshot copy.
- When using **`ndmpcopy`** command for transferring data between a storage system running Data ONTAP operating in 7-Mode and a storage system running Clustered Data ONTAP:
- The ndmpcopy command must be initiated from a storage system running Data ONTAP operating in 7-Mode
- In the node-scoped mode, the destination IP address is the address of an LIF on the node on which the target volume is located
- Destination path is of the format /vserver_name/volume_name.

> **Note:** You should not use the ndmpcopy command for restoring a LUN between a storage system running Data ONTAP operating in 7-Mode and a storage system running Clustered Data ONTAP because the LUN is restored as a file on the destination storage system.

For the syntax and examples of the ndmpcopy command, see the *Data ONTAP Data Protection Tape Backup and Recovery Guide for 7-Mode.*

- When a SnapMirror destination is backed up to tape, only the data on the volume is backed up. The SnapMirror relationships and the associated metadata are not backed up to tape. Therefore, during restore, only the data on that volume is restored and the associated SnapMirror relationships are not restored.

12.2 Further information about SAN and Fibre Channel

More details on SAN and the Fibre Channel protocol can be found in the following IBM support documents:

- *Data ONTAP Data Protection Tape Backup and Recovery Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

- *Data ONTAP 8.2 Release notes*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

13

Storage efficiency

In this chapter, we provide more information about the storage efficiency features that you can find in Clustered Data ONTAP. We explain how they work with FlexVol volumes and Infinite Volumes.

The following topics are covered:

- Thin provisioning
- Deduplication
- Compression
- Storage efficiency on Infinite Volumes with storage classes

13.1 Thin provisioning

Although thin provisioning is a commonly used storage term, implementations will vary depending on the specific vendor, product, and even application. This section provides an overview of the IBM N series thin implementation and its benefits. It includes key points, such as the ease of enabling thin provisioning, and how IBM N series thin provisioning architecture allows performance levels to be maintained. Thin provisioning is the most efficient way to provision storage, because the storage is not all preallocated up front. In other words, when a volume or LUN is created using thin provisioning, no space on the storage system is used. The space remains unused until data is written to the LUN or the volume, at which time only enough space to store the data will be used. To take full advantage of these and other benefits that thin provisioning has to offer, a mature storage management capability is necessary.

The alternative to thin provisioning is thick provisioning, which is the traditional approach in which storage is allocated at the time of volume or LUN creation. In this case, the space is dedicated to a specific LUN or volume, and it cannot be shared with other LUNs or volumes, even if the space is sitting empty and unused.

13.1.1 Thin provisioning defined

IBM N series thin provisioning has two main abilities:

- The ability to provision more logical storage space to the hosts connecting to the storage system than is actually available on the storage system
- The ability to allocate storage on demand as the data comes in, instead of preallocating it

Consider an example in which there are 15 hosts, and each host estimates that they will need 500 GB of storage space for the duration of the project over 3 years, totaling 7500 GB. The storage administrator configures a storage system using thin provisioning in which the storage system only has 5000 GB of usable storage. The resulting effect is that each host sees the capacity they requested and the storage system has not used any space for the initial creation of that storage.

In contrast to thin provisioning, a thick-provisioned storage system would attempt to preallocate the storage required for each host and not be capable of servicing the requests for all projects. Thick provisioning would require the full 7500 GB of storage to be available up front. Instead, with thin provisioning, 5000 GB of storage remain available to all projects until a project writes data to the storage, at which time that specific storage is allocated to that project and no longer available to other projects. To avoid running out of space, the administrator monitors the storage system and adds storage as needed.

13.1.2 Thin provisioning architecture

IBM N series FlexVol technology provides the ability to create efficient flexible data containers that separate the content from the constraints of the physical storage. The FlexVol technology decouples the physical barrier between data containers and their associated physical disks. The result is a significant increase in flexibility and storage utilization. With the Data ONTAP FlexVol technology, you can shrink or grow data containers based on immediate needs. Adding disks can be done on the fly, without disrupting the system or the associated applications.

Thin provisioning on IBM N series systems consist of some key components:

- FlexVol volumes: FlexVol volumes are a standard free component included with IBM N series storage systems. They can be configured to automatically allocate space as data is written to them, and can be configured to work with file-level protocols (NAS).
- LUN: LUNs are a standard free component included with IBM N series storage systems. They provide support for block-level protocols (SAN). Within Data ONTAP, LUNs are simply objects within a volume, providing additional levels of flexibility and efficiency.
- Aggregate: 32-bit aggregates and 64-bit aggregates are free components on IBM N series storage systems, and are used as the storage pool for volumes and LUNs using thin provisioning. Multiple aggregates can be used simultaneously on a single system, and each can be expanded with additional storage if more space is needed in the storage pool.

Figure 13-1 illustrates how these components can be combined to create a solution based on a single aggregate. In practice, it is common for a single system to contain multiple aggregates and various combinations of volumes and LUNs within the aggregates.

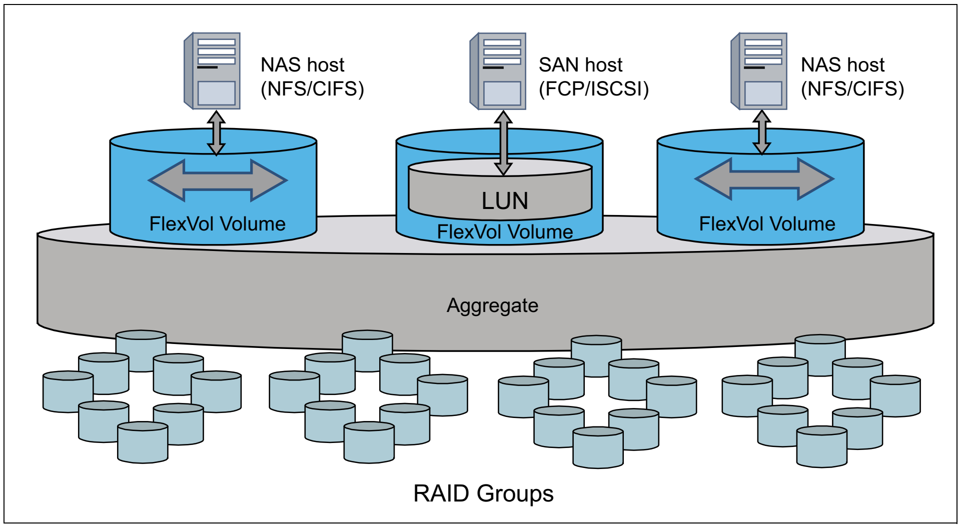


Figure 13-1 Thin provisioning

The Data ONTAP architecture uses aggregates to virtualize the physical storage into pools for logical allocation. The volumes and LUNs see the logical space, and the aggregate controls the physical space. This architecture provides the flexibility to create multiple volumes and LUNs that can exceed the physical space available in the aggregate. All volumes and LUNs in the aggregate will use the available storage within the aggregate as a shared storage pool. This will allow them to efficiently allocate the space available in the aggregate as data is written to it, rather than preallocating (reserving) the space.

Preallocating space will often result in space that sits unused for long periods of time, resulting in lower storage utilization rates, also known as wasted storage. In the case that some volumes or LUNs do indeed have a valid reason for preallocating space, you have the flexibility to allow all other volumes and LUNs to carve out preallocated space in the aggregate. This can be achieved while simultaneously supporting thin provisioning within the same aggregate. More clearly stated, the aggregate is the shared storage pool for volumes and LUNs using thin provisioning, and can also include the volumes and LUNs using preallocated storage, simultaneously. The aggregate can be expanded by simply adding more disks to it. Monitoring of the storage utilization takes place at the aggregate level to predict when the aggregate should be expanded.

13.2 Deduplication

Deduplication is a Data ONTAP feature that reduces the amount of physical storage space required by eliminating duplicate data blocks within a FlexVol volume or an Infinite Volume. You should not enable deduplication on the root volume.

You can decide to deduplicate only the new data that is written to the volume after enabling deduplication, or both the new data and the data existing in the volume prior to enabling deduplication.

13.2.1 How deduplication works

Deduplication operates at the block level within the entire FlexVol volume or an Infinite Volume, eliminating duplicate data blocks, and storing only unique data blocks.

Each block of data has a digital signature that is compared with all other signatures in a data volume. If an exact block signature match exists, a byte-by-byte comparison is done for all the bytes in the block, and the duplicate block is discarded and its disk space is reclaimed.

Deduplication removes data redundancies, as shown in Figure 13-2.

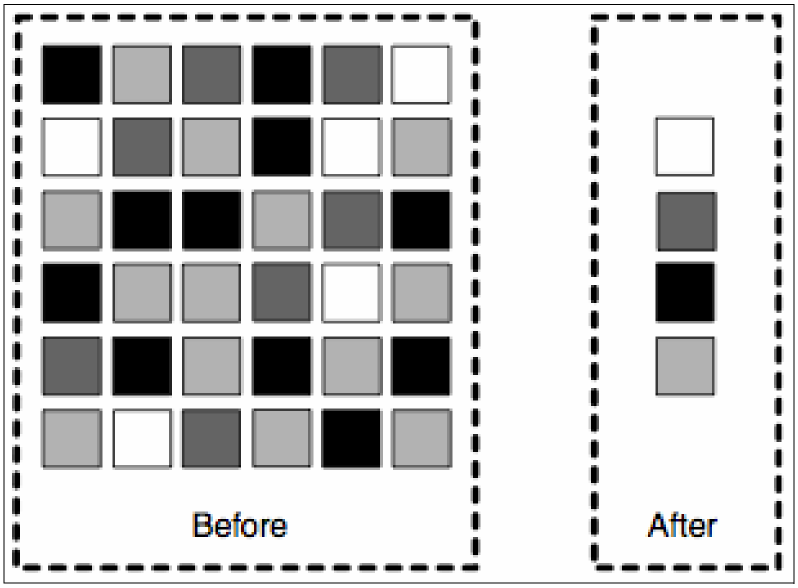


Figure 13-2 Visible deduplication example

Data ONTAP writes all data to a storage system in 4-KB blocks. When deduplication runs for the first time on a volume with existing data, it scans all the blocks in the volume and creates a digital fingerprint for each of the blocks. Each of the fingerprints is compared to all the other fingerprints within the volume. If two fingerprints are found to be identical, a byte-by-byte comparison is done for all data within the block. If the byte-by-byte comparison detects identical data, the pointer to the data block is updated, and the duplicate block is removed.

Note: When deduplication is run on a volume with existing data, it is best to configure deduplication to scan all the blocks in the volume for better space savings.

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.

13.2.2 Deduplication and Infinite Volumes

Deduplication of an Infinite Volume is configured at the level of the Infinite Volume, but the actual deduplication process occurs within each data constituent.

The scope of deduplication

For an Infinite Volume, deduplication occurs within each data constituent, not across the Infinite Volume. For example, if two files on the same data constituent contain the same block, deduplication discards the duplicate block. If two files in separate data constituents contain the same block, deduplication does not discard the duplicate block.

The namespace constituent and namespace mirror constituents of an Infinite Volume are not deduplicated.

How deduplication is configured

Deduplication is configured at the Infinite Volume level. When you enable or disable deduplication on an Infinite Volume, deduplication is enabled or disabled on the data constituents of the Infinite Volume.

You can see whether deduplication is enabled by viewing the State field in the output of the **`volume efficiency show`** command. (The terms “efficiency state” and “deduplication state” are sometimes used interchangeably.) By default, the field shows the deduplication state of the Infinite Volume as a whole. If the State field contains a dash (“-”), one or more data constituents are offline or have a different deduplication state than the other data constituents. If you add the **`-is-constituent true`** parameter to the **`volume efficiency show`** command, the output displays the deduplication state of each individual data constituent.

How deduplication operations are run

When a deduplication operation is run on an Infinite Volume, separate deduplication operations run on each data constituent in the Infinite Volume. For example, if an Infinite Volume has 100 data constituents, a deduplication operation on the Infinite Volume triggers a deduplication operation on each of the 100 data constituents.

Deduplication operations are combined with post process compression into a queue of efficiency operations. A maximum of eight efficiency operations per node occur at any one time. If more than eight efficiency operations per node are scheduled to run at any one time, they are queued and run as each operation finishes.

If an operation succeeds overall but fails on one or more constituents, the names of the failed constituents are reported to the event management system, along with the reason for failure.

Information about deduplication operations, such as the status and progress, is not available for an Infinite Volume as a whole. You can see the status and progress of post process compression operations on individual data constituents by using the volume efficiency show command with the **`-is-constituent true`** parameter.

How space savings are reported

The space savings gained by deduplication on an Infinite Volume reflect the total space savings of all of the volume's data constituents. You can see the space savings by using the **`volume show`** command with either the -instance or -fields parameter.

Space savings information is available only when all data constituents are online.

How volume state affects efficiency

You can run efficiency operations and view efficiency information only when an Infinite Volume is online, which means that every constituent in the Infinite Volume must be online.

Free space is required for deduplication.

Deduplication has the following free space requirements:

- Each aggregate that contains deduplication-enabled data constituents or deduplication-enabled FlexVol volumes must have free space that is equivalent to 3% of the total logical data contained within all of the deduplicated data constituents and deduplicated FlexVol volumes on the aggregate.
- Each data constituent in the deduplicated Infinite Volume must have free space that is equivalent to 4% of the data in the data constituent.

If the data constituents or their containing aggregates lack adequate free space, the affected data constituents are skipped while deduplication continues to run on the other data constituents.

13.2.3 Deduplication metadata

The deduplication metadata includes the fingerprint file and change logs. Fingerprints are the digital signatures for every 4-KB data block in a FlexVol volume or an Infinite Volume.

The deduplication metadata contains two change log files. When deduplication is running, the fingerprints of the new data blocks from one change log file are merged into the fingerprint file, and the second change log file stores the fingerprints of the new data that is written to the volume during the deduplication operation. The roles of the change log files are reversed when the next deduplication operation is run.

In Data ONTAP 8.0.1, the deduplication metadata is located within the aggregate. Starting with Data ONTAP 8.1, two copies of deduplication metadata are maintained per volume. A copy of the deduplication metadata resides in the volume and another copy is in the aggregate. The deduplication metadata in the aggregate is used as the working copy for all the deduplication operations. An additional copy of the deduplication metadata resides in the volume.

When a volume is moved, the deduplication metadata is also transferred with the volume. If the volume ownership changes, the next time deduplication is run, then the deduplication metadata which resides in the aggregate is created automatically by using the copy of deduplication metadata in the volume. This method is a faster operation than creating a new fingerprint file.

Starting with Data ONTAP 8.2, the fingerprints are stored for each physical block, this reduces the amount of space required to store the deduplication metadata.

Deduplication metadata can occupy up to 7% of the total physical data contained within the volume, as follows:

- In a volume, deduplication metadata can occupy up to 4% of the total amount of data contained within the volume. For an Infinite Volume, the deduplication metadata within an individual data constituent can occupy up to 4% of the total amount of data contained within the data constituent.
- In an aggregate, deduplication metadata can occupy up to 3% of the total physical data contained within the volume.

You can use the **`storage aggregate show`** command to check the available space in an aggregate and the volume show command to check the available space in a volume. See Example 13-1. For more information about these commands, see the man pages.

Example 13-1 Metadata calculation

```
A 2 TB aggregate has four volumes, each 400 GB in size, in the aggregate. You need
three volumes to be deduplicated with varying savings percentage on each volume.
The space required in the different volumes for deduplication metadata is as
follows:
   2 GB [4% × (50% of 100 GB)] for a 100 GB of logical data with 50 percent
   savings
   6 GB [4% × (75% of 200 GB)] for a 200 GB of logical data with 25 percent
   savings
   3 GB [4% × (25% of 300 GB)] for a 300 GB of logical data with 75 percent
   savings
The aggregate needs a total of 8.25 GB [(3% × (50% of 100 GB)) + (3% × (75% of 200
GB)) + (3% × (25% of 300 GB)) = 1.5+4.5+2.25= 8.25 GB] of space available in the
aggregate for deduplication metadata.
```

13.2.4 Guidelines for using deduplication

Deduplication runs as a system operation and consumes system resources when the deduplication operation is running on FlexVol volumes or Infinite Volumes.

If the data does not change often in a volume, it is advised to run deduplication less frequently. If you run multiple concurrent deduplication operations on a storage system, these operations lead to a higher consumption of system resources. It is advised to begin with fewer concurrent deduplication operations. Increasing the number of concurrent deduplication operations gradually enables you to better understand the impact on the system.

Note: It is advised not to have multiple volumes with the volume size nearing the logical data limit of a volume with deduplication enabled.

Various factors affect the performance of deduplication. You should 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 can affect the performance of deduplication:

- 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
- The nature of data layout in the volume
- The amount of changed data between deduplication operations

- The number of concurrent deduplication operations
- Hardware platform (system memory and CPU module)
- Load on the system
- Disk types (for example, ATA/FC, and rpm of the disk)

13.3 Compression

Data compression enables you to store more data in less space. Further, you can use data compression to reduce the time and bandwidth required to replicate data during volume SnapMirror transfers. Data compression can save space on regular files or LUNs.

However, storage system internal files, Windows NT streams, and volume metadata are not compressed.

Data compression works by compressing a small group of consecutive blocks known as a compression group. Data compression can be done in the following ways:

- Inline compression:

  If inline compression is enabled on a volume, during subsequent data writes the compressible data is compressed and written to the volume. However, data which cannot be compressed or data bypassed by inline compression is written in the uncompressed format to the volume.

- Post process compression:

  If post process compression is enabled on a volume, the new data writes to the volume which were not compressed initially (if inline compression is enabled), are rewritten as compressed data to the volume when post process compression is run. The post process compression operation runs as a low-priority background process.

If both inline and post process compression are enabled, then post process compression compresses only the blocks on which inline compression was not run. This includes blocks that were bypassed by inline compression such as small, partial compression group overwrites.

Note: You cannot enable data compression on the storage system root volumes or on the volumes that are contained within 32-bit aggregates.

13.3.1 How compression works on Infinite Volumes

Data compression of an Infinite Volume is configured at the level of the Infinite Volume, but the actual compression processes occur within each data constituent.

The scope of compression

Compression runs on the data files within the Infinite Volume, not on the information contained within the namespace constituent.

Configuring compression and inline compression

Data compression of an Infinite Volume is configured at the Infinite Volume level. When you enable or disable compression and inline compression on an Infinite Volume, compression and inline compression are enabled or disabled on the data constituents of the Infinite Volume.

You can see whether post process compression and inline compression are enabled by viewing the Compression and Inline Compression fields that are displayed when you use the **`-instance`** parameter with the **`volume efficiency show`** command. By default, the fields show the compression state and inline compression state of the Infinite Volume as a whole. If either field contains a dash ("-"), one or more data constituents are offline or have a different state than the other data constituents. If you add the **`-is-constituent true`** parameter to the **`volume efficiency show`** command, the output displays the post process compression and inline compression state of each individual data constituent.

Running post process compression operations

When a post process compression operation runs on an Infinite Volume, separate compression operations run on each data constituent in the Infinite Volume. For example, if an Infinite Volume has 100 data constituents, a post process compression operation on the Infinite Volume triggers 100 post process compression operations on the data constituents.

If an operation succeeds overall but fails on one or more constituents, the names of the failed constituents are reported to the event management system, along with the reason for failure.

Information about post process compression operations, such as the status and progress, is not available for an Infinite Volume as a whole. You can see the status and progress of post-process compression operations on individual data constituents by using the **`volume efficiency show`** command with the **`-is-constituent true`** parameter.

How space savings are reported

The space savings gained by compression on an Infinite Volume reflect the total space savings of all of the volume's data constituents. You can see the space savings by using the **`volume show`** command. If you add the **`-is-constituent true`** parameter to the **`volume show`** command, the output displays the space savings of each individual data constituent.

Space savings information is available only when all the constituents are online.

How volume state affects efficiency

You can run efficiency operations and view efficiency information only when an Infinite Volume is online, which means that every constituent in the Infinite Volume must be online.

13.3.2 Detecting incompressible data

Incompressible data detection allows you to check if a file is compressible and for large size files, you can check if a compression group within a file is compressible. Allowing incompressible data to be detected saves the system resources used by inline compression trying to compress incompressible files or compression groups.

For files with size less than 500 MB, inline compression checks if a compression group can be compressed. If incompressible data is detected within a compression group, then a flag is set for the file containing the compression group to indicate that the file is incompressible. During subsequent compression attempts, inline compression first checks if the incompressible data flag is set for the file. If the flag is set, then inline compression is not attempted on the file.

For files with size equal to or greater than 500 MB, inline compression performs a quick check on the first 4 KB block of each compression group to determine if it can be compressed. If the 4 KB block cannot be compressed, the compression group is left uncompressed. However, if compression of the 4 KB block is successful, then compression is attempted on the whole compression group.

Post process compression runs on all files irrespective of whether the file is compressible or not. If post process compression compresses at least one compression group in an incompressible file, then the incompressible data flag for that file is cleared. During the next compression attempt, inline compression can run on this file to achieve space savings.

For more information about enabling or disabling incompressible data detection and modifying the minimum file size to attempt quick check on a file, see the **`volume efficiency modify`** command man page.

13.4 Storage efficiency on Infinite Volumes with storage classes

When an Infinite Volume uses storage classes, space-saving technologies are configured at the storage-class level to reflect the service-level objectives of each storage class. This affects the way that efficiency is configured and the way that efficiency information is displayed.

13.4.1 Configuring efficiency on an Infinite Volume with storage classes

Instead of configuring deduplication and compression technologies for the entire Infinite Volume, deduplication and compression settings are configured within the storage classes that an Infinite Volume uses.

For example, an Infinite Volume might have two storage classes, one that emphasizes capacity and another that emphasizes performance. While storage classes might have deduplication enabled, compression might be enabled only on the storage class that emphasizes capacity. The storage class that emphasizes capacity might have a schedule that runs efficiency operations daily at off-peak hours, while the storage class that emphasizes performance might have a schedule that runs efficiency operations only on weekends.

To configure storage classes, you must use OnCommand Workflow Automation. You can use the command-line interface to create and modify efficiency policies, but you must use OnCommand Workflow Automation to apply an efficiency policy to a storage class.

13.4.2 Checking efficiency state on an Infinite Volume with storage classes

When an Infinite Volume uses storage classes, efficiency state information is not available at the Infinite Volume level. For an Infinite Volume with storage classes, the state of deduplication, compression, and in-line compression are displayed as a dash (-).

Instead, you can display efficiency state information for each individual data constituent.

Space savings information is displayed for an Infinite Volume with storage classes. When an Infinite Volume uses storage classes, space savings information is available both for the entire Infinite Volume and for individual data constituents.

When you view space savings information at the Infinite Volume level, you should keep in mind that the information reflects a summary of the savings across all data constituents in all storage classes. If efficiency technologies are not enabled on all storage classes, the savings for the entire Infinite Volume might be lower than you expect. For example, if one storage class has 50% space savings and another storage class has 0% space savings (because efficiency technologies are disabled), the savings for the entire Infinite Volume are 25%.

14

Data protection

Data protection means backing up data and being able to recover it. You protect the data by making copies of it so that it is available for restoration even if the original is no longer available.

Businesses need data backup and protection for the following reasons:

- To protect data from accidentally deleted files, application crashes, data corruption, and viruses
- To archive data for future use
- To recover from a disaster

Data ONTAP provides a variety of tools that you can use to build a comprehensive strategy to protect your company's data:

- Snapshot copies: Enable you to manually or automatically create, schedule, and maintain multiple backups
- SnapRestore: Enables you to perform fast, space-efficient, on-request Snapshot recovery from Snapshot copies on an entire volume.
- SnapVault backups: Provide storage-efficient and long-term retention of backups.
- Volume copy: Enables you to perform fast block-copy of data from one volume to another.

The following topics are covered:

- Snapshot
- Snapshot introduction
- Creation of Snapshot copy schedules
- Snapshot for Infinite Volume
- Snapshot process: Basic operation
- Understanding Snapshots in detail
- Snapshot data structures and algorithms
- SnapVault
- SnapVault basics
- 7-Mode versus Clustered Data ONTAP SnapVault
- How a SnapVault backup works
- Supported data protection deployment configurations
- Protecting data on FlexVol volumes by using SnapVault
- Managing backup operations for SnapVault backups
- Managing restore operations for SnapVault backups
- Managing storage efficiency for SnapVault secondaries

14.1 Snapshot

Snapshot copies are the first line of defense for data protection. Data ONTAP maintains a configurable Snapshot schedule that creates and deletes Snapshot copies automatically for each FlexVol volume and Infinite Volume. You can also create and delete Snapshot copies, and manage Snapshot schedules based on your requirements.

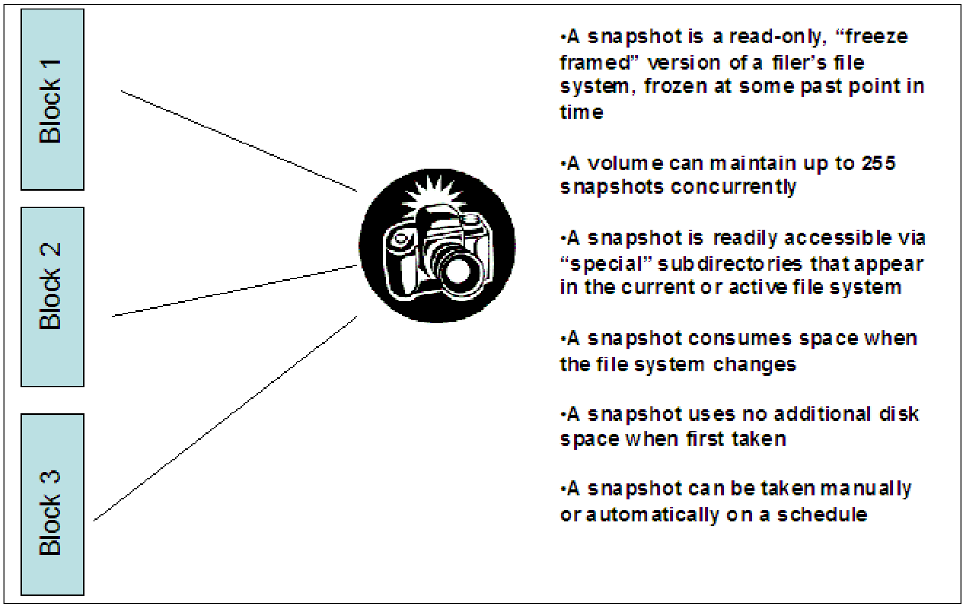
14.2 Snapshot introduction

A Snapshot is a point-in-time copy of a FlexVol volume representing the volume's contents at a particular point in time. You can view the contents of the Snapshot copy and use the Snapshot copy to restore data that you lost recently.

A Snapshot copy of a volume is located on the parent volume but has read-only access. It represents the contents of the original volume at a particular point in time. A parent volume and a Snapshot copy of it share disk space for all blocks that have not been modified between the creation of the volume and the time the Snapshot copy is made, thereby making Snapshot copies lightweight.

Similarly, two Snapshot copies share disk space for those blocks that were not modified between the times that the two Snapshot copies were created. You can create a chain of Snapshot copies to represent the state of a volume at a number of points in time. Users can access Snapshot copies online, enabling users to retrieve their own data from past copies, rather than asking a system administrator to restore data from tape. Administrators can restore the contents of a volume from a Snapshot copy.

Snapshot features are shown in Figure 14-1.



- 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 14-1 Snapshot features

14.2.1 Key features of Snapshot software

Here we provide a brief overview of the Snapshot features:

- Stability:

  A Snapshot copy is a read-only, static, and immutable copy. It enables organizations to perform consistent backups from an N series storage system while applications run.

- Performance:

  Storing a Snapshot copy on an N series system has no performance impact. In addition, creating and deleting a Snapshot copy have little performance impact on a properly configured system. Even if you use a competing primary storage system, N series storage can be used for replication, backup, and archival.

- Scalability:

  N series storage volumes support up to 255 Snapshot copies. The ability to store a large number of low-impact, frequently created Snapshot copies increases the likelihood that the desired version of data can be successfully recovered.

- User visibility and file recoverability:

  Snapshot high performance, scalability, and stability provides an ideal online backup for user-driven recovery. Additional solutions allow you to copy backups to offline disk or to tape and archive them in provably immutable form for compliance or e-discovery.

- Efficient storage utilization:

  Two Snapshot copies taken in sequence differ only by the blocks added or changed in the time interval between the two. This block-incremental behavior limits associated storage capacity consumption. Some alternative implementations can consume storage volumes rivaling that of the active file system, raising storage capacity requirements.

14.2.2 User access to Snapshot copies

Each volume has a *.snapshot* directory that is accessible to NFS users by using the **ls** command and to CIFS users by double-clicking the *~snapshot* folder. This access to Snapshot copies can be turned off.

Snapshot files carry the same read permissions as the original file. A user who has permission to read a file in the volume can read that file in a Snapshot copy. A user without read permission to the volume cannot read that file in a Snapshot copy. Snapshot copies do not have write permissions.

Example 14-1 shows the contents of the *.snapshot* directory within an NFS share. It is a set of subdirectories, labeled by type, date, and time.

Example 14-1 Listing .snapshot folder on NFS share

```
$ ls .snapshot
hourly.2013-05-15_1006/
hourly.2013-05-15_1106/
hourly.2013-05-15_1206/
hourly.2013-05-15_1306/
hourly.2013-05-15_1406/
hourly.2013-05-15_1506/
daily.2013-05-14_0013/
daily.2013-05-15_0012/
weekly.2013-05-13_0019/
```

Each subdirectory of the *.snapshot* directory includes a list of the parent volume's files and directories. If users accidentally delete or overwrite a file, they can locate it in the most recent Snapshot directory and restore it to their main read-write volume simply by copying it back to the main directory. Example 14-2 shows how an NFS user can locate and retrieve a file named my.txt from the *.snapshot* directory.

Example 14-2 Recovering my.txt file from latest snapshot

```
$ ls my.txt
ls: my.txt: No such file or directory
$ ls .snapshot
hourly.2013-05-15_1006/
hourly.2013-05-15_1106/
hourly.2013-05-15_1206/
hourly.2013-05-15_1306/
hourly.2013-05-15_1406/
hourly.2013-05-15_1506/
daily.2013-05-14_0013/
daily.2013-05-15_0012/
weekly.2013-05-13_0019/
$ ls .snapshot/hourly.2013-05-15_1506/my.txt
my.txt
$ cp .snapshot/hourly.2013-05-15_1506/my.txt .
$ ls my.txt
my.txt
```

Note: The *.snapshot* directory is always visible to NFSv2 and NFSv3 clients and available from within the volume, and not visible but still available from any other volume. For NFSv4 clients, the *.snapshot* directory is not visible, but accessible in all paths of a volume.

Some Snapshot copies in the *.snapshot* directory are used only to support internal system processes for the volume, such as data protection of the namespace constituent for an Infinite Volume, and you cannot access these Snapshot copies. You can access any of the Snapshot copies for a volume that are displayed when you use the **`volume snapshot show`** command. The command hides the types of Snapshot copies that you cannot access.

14.2.3 Maximum number of Snapshot copies

You should know what the maximum number of Snapshot copies you can accumulate is to minimize the possibility that you do not have Snapshot copies available when you need them.

- You can accumulate a maximum of 255 Snapshot copies of a FlexVol volume.
- If the FlexVol volume is in a data protection mirror relationship, the maximum number of Snapshot copies is 254 because one Snapshot copy is reserved for use by the relationship during recovery operations.
- If the FlexVol volume is in a disk to disk backup relationship, the maximum number of Snapshot copies is 251.
- If the Infinite Volume is in a data protection mirror relationship, the maximum number of Snapshot copies is reduced by two for each namespace mirror constituent and another 2 if you have a SnapMirror relationship between Infinite Volumes

Over time, automatically generated hourly, weekly, and monthly Snapshot copies accrue. Having a number of Snapshot copies available gives you a greater degree of accuracy if you have to restore a file.

The number of Snapshot copies can approach the maximum if you do not remove older Snapshot copies. You can configure Data ONTAP to automatically delete older Snapshot copies of volumes as the number of Snapshot copies approaches the maximum.

The following data protection mirror copies affect the maximum number of Snapshot copies available to a volume:

- A FlexVol volume in a data protection mirror relationship
- A FlexVol volume with a load-sharing mirror copy
- An Infinite Volume with one or more namespace mirror constituents

Each namespace mirror constituent uses two Snapshot copies. By default, a read/write Infinite Volume contains one namespace mirror constituent. If you enable SnapDiff on an Infinite Volume, each additional namespace mirror uses two Snapshot copies.

An Infinite Volume also uses up to four Snapshot copies when technical support runs some commands that require diagnostic privilege. You must keep the number of Snapshot copies far enough below the limit to ensure that technical support can run commands.

14.3 Creation of Snapshot copy schedules

Data ONTAP provides a default Snapshot copy schedule for each FlexVol volume and Infinite Volume. You can create schedules to fit your needs if the default Snapshot copy schedule is not adequate.

For FlexVol volumes, the default Snapshot copy schedule automatically creates one daily Snapshot copy Monday through Saturday at midnight, an hourly Snapshot copy five minutes past the hour, every hour, and a weekly Snapshot copy. Data ONTAP retains the two most recent nightly Snapshot copies and the six most recent hourly Snapshot copies, and deletes the oldest nightly and hourly Snapshot copies when new Snapshot copies are created.

14.3.1 Types of user-specified Snapshot copy schedules

Data ONTAP contains weekly, daily, and hourly Snapshot copy schedules that you can use to create Snapshot copy policies that retain the number and type of Snapshot copies you want.

Table 14-1 describes the available types of Snapshot copy schedules.

Table 14-1 Snapshot copy schedules

Type	Description
Weekly	Data ONTAP creates these Snapshot copies every Sunday at 15 minutes after midnight. Weekly Snapshot copies are named *weekly.n*, where *n* is the date in year-month-day format followed by an underscore (_) and the time. For example, a weekly Snapshot copy created on 25 November 2012 is named `weekly.2012-11-25_0015`.
Daily	Data ONTAP creates these Snapshot copies every night at 10 minutes after midnight. Daily Snapshot copies are named *daily.n*, where *n* is the date in year-month-day format followed by an underscore (_) and the time. For example, a daily Snapshot copy created on 4 December 2012 is named `daily.2012-12-04_0010`.

Type	Description
Hourly	Data ONTAP creates these Snapshot copies five minutes after the hour. Hourly Snapshot copies are named *hourly.n*, where *n* is the date in year-month-day format followed by an underscore (_) and the time. For example, an hourly Snapshot copy created on 4 December 2012 at 1:00 (1300) is named `hourly.2012-12-04_1305`.

14.3.2 Creating a Snapshot copy schedule

If the default Snapshot copy schedule does not meet your needs, you can create a schedule that does by using the **`job schedule cron create`** command or the **`job schedule interval create`** command.

Depending on how you want to implement the schedule, run one of the following commands:

- The **`job schedule cron create`** command creates a cron schedule. A cron schedule, like a UNIX cron job, runs at a specified time. You can also specify months, days of the month, or days of the week on which the schedule will run.

  If you specify values for both days of the month and days of the week, they are considered independently. For example, a cron schedule with the day specification Friday, 13 runs every Friday and on the 13th day of each month, not just on every Friday the 13th.

  Example 14-3 creates a cron schedule named `weekendcron` that runs on weekend days (Saturday and Sunday) at 3:00 a.m.

  Example 14-3 Cron schedule example

  ```
  cluster1::> job schedule cron create -name weekendcron -dayofweek "Saturday,
  Sunday" -hour 3 -minute 0
  ```

- The **`job schedule interval create`** creates an interval schedule. An interval schedule runs jobs at specified intervals after the previous job finishes.

  For example, if a job uses an interval schedule of 12 hours and takes 30 minutes to complete, the job runs at the following times:

  - Day one at 8:00 a.m. (the job's initial run)
  - Day one at 8:30 p.m.
  - Day two at 9:00 a.m.
  - Day two at 9:30 p.m.

  Example 14-4 creates an interval schedule named `rollingdaily` that runs six hours after the completion of the previous occurrence of the job.

  Example 14-4 Interval schedule example

  ```
  cluster1::> job schedule interval create -name rollingdaily -hours 6
  ```

  See the man page for each command to determine the command that meets your needs.

14.3.3 Deleting Snapshot copies automatically

You can automatically delete Snapshot copies from read-write volumes and FlexClone LUNs from read-write parent volumes. You cannot set up automatic deletion of Snapshot copies from Infinite Volumes or from read-only volumes, for example, SnapMirror destination volumes.

You define and enable a policy for automatically deleting Snapshot copies by using the **volume snapshot autodelete modify** command.

Example 14-5 enables the automatic deletion of Snapshot copies and sets the trigger to snap_reserve for the vol3 volume, which is part of the vs0 storage virtual machine (SVM).

Example 14-5 Enabling automatic deletion of Snapshot copies

```
cluster1::> volume snapshot autodelete modify -vserver vs0 -volume vol3 -enabled
true -trigger snap_reserve
```

Example 14-6 enables the automatic deletion of Snapshot copies and of FlexClone LUNs for the vol3 volume, which is part of the vs0 SVM.

Example 14-6 Enabling the automatic deletion of Snapshot copies and of FlexClone LUNs

```
cluster1::> volume snapshot autodelete modify -vserver vs0 -volume vol3 -enabled
true -trigger volume -commitment try -delete-order oldest_first -destroy-list
lun_clone,file_clone
```

See the **volume snapshot autodelete modify** man page for information about the parameters that you can use with this command to define a policy that meets your needs.

Notes:

- You can view the settings for the automatic deletion of Snapshot copies to help you when you are deciding if the settings are meeting your needs.
- View the settings for the automatic deletion of Snapshot copies by using the **volume snapshot autodelete show** command.

14.4 Snapshot for Infinite Volume

Snapshot copies are managed at the Infinite Volume level, not at the individual data constituent level or the storage class level. Similarly to the way he uses a FlexVol volume, an administrator can create, delete, and restore data files by using Snapshot on the Infinite Volume along with the similar (hourly, nightly, weekly) Snapshot schedule management.

The Snapshot copy of Infinite Volume differs from that of a FlexVol volume in terms of performance, because all the constituents have to be fenced, Snapshot copies created, and unfenced together. The latency that this incurs for file operations increases linearly with the number of data constituents.

Note: Infinite Volume-level Snapshot copies are crash-consistent Snapshot copies. The Snapshot process coordinates with each individual data constituent on each node to start the Snapshot creation process by fencing all I/Os while performing the Snapshot creation.

- A Snapshot copy can incur latency of up to 30 seconds if the number of data constituents is <=50.
- It can incur latency up to 2 minutes if the number of data constituents is 200.

The *.snapshot* directory for a FlexVol volume shows the files in the Snapshot copies only for that specific FlexVol volume, not for the entire namespace that is using numerous junctions in that hierarchy. A Snapshot copy of an Infinite Volume shows the files for the entire namespace at the Infinite Volume level.

14.4.1 Snapshot for FlexVol volume versus Infinite Volume

Table 14-2 compares the Snapshot functionality for the FlexVol volumes with that for the Infinite Volume.

Table 14-2 Snapshot for FlexVol volume versus Infinite Volume

Snapshot functionality	FlexVol volume	Infinite Volume
Create / Delete / Restore	Yes	Yes
Single File SnapRestore	Yes	No
Compute space reclaimable	Yes	No
Rename	Yes	No
Show	Yes	Yes
Latencies less than 2 minutes for commands	Yes	No
Snapshot policy (includes any of the defined policies in volume snapshot policy)	Yes	Yes
Subvolume Snapshot copies (partial behavior)	N/A	No
Auto Snapshot delete	Yes	No
Manage ONTAP (ZAPI) support	Yes	Yes

14.4.2 Snapshot copies for Infinite Volume states

Snapshot copies of an Infinite Volume are restorable and fully accessible to clients only when the Snapshot copies are in a valid state.

The availability of a Snapshot copy of an Infinite Volume is indicated by its state, as explained in Table 14-3.

Table 14-3 Snapshot copies for Infinite Volume states

State	Description	Client access to the Snapshot copy	Impact on restore
Valid	The copy is complete.	Fully accessible to clients	Can be restored
Partial	Data is missing or incomplete.	Partially accessible to clients	Cannot be restored without assistance from technical support
Invalid	Namespace information is missing or incomplete.	Inaccessible to clients	Cannot be restored

The validity of a Snapshot copy is not tied directly to the state of the Infinite Volume. A valid Snapshot copy can exist for an Infinite Volume with an offline state, depending on when the Snapshot copy was created compared to when the Infinite Volume went offline. For example, a valid Snapshot copy exists before a new constituent is created. The new constituent is offline, which puts the Infinite Volume in an offline state. However the Snapshot copy remains valid because it references its needed pre-existing constituents. The Snapshot copy does not reference the new, offline constituent.

To view the state of Snapshot copies, you can use the `volume snapshot show` command.

14.4.3 Guidelines for working with Snapshot copies of Infinite Volumes

You can create, manage, and restore Snapshot copies of Infinite Volumes. However, you should be aware of the factors affecting the Snapshot creation process and the requirements for managing and restoring the copies.

Guidelines for creating Snapshot copies of Infinite Volumes

Observe the following guidelines:

- The volume must be online.

  You cannot create a Snapshot copy of an Infinite Volume if the Infinite Volume is in a Mixed state because a constituent is offline.

- The Snapshot copy schedule should not be less than hourly.

  It takes longer to create a Snapshot copy of an Infinite Volume than of a FlexVol volume. If you schedule Snapshot copies of Infinite Volumes for less than hourly, Data ONTAP tries but might not meet the schedule. Scheduled Snapshot copies are missed when the previous Snapshot copy is still being created.

- Time should be synchronized across all the nodes that the Infinite Volume spans.

  Synchronized time helps schedules for Snapshot copies run smoothly and restoration of Snapshot copies function properly.

- The Snapshot copy creation job can run in the background.

  Creating a Snapshot copy of an Infinite Volume is a cluster-scoped job (unlike the same operation on a FlexVol volume). The operation spans multiple nodes in the cluster. You can force the job to run in the background by setting the **`-foreground`** parameter of the **`volume snapshot create`** command to **`false`**.

- After you create Snapshot copies of an Infinite Volume, you cannot rename the copy or modify the comment or SnapMirror label for the copy.

Guidelines for managing Snapshot copy disk consumption

Observe the following guidelines:

- You cannot calculate the amount of disk space that can be reclaimed if Snapshot copies of an Infinite Volume are deleted.
- The size of a Snapshot copy for an Infinite Volume excludes the size of namespace mirror constituents.
- If you use the **`df`** command to monitor Snapshot copy disk consumption, it displays information about consumption of the individual data constituents in an Infinite Volume, not for the Infinite Volume as a whole.
- To reclaim disk space used by Snapshot copies of Infinite Volumes, you must manually delete the copies.

  You cannot use a Snapshot policy to automatically delete Snapshot copies of Infinite Volumes. However, you can manually delete Snapshot copies of Infinite Volumes, and you can run the delete operation in the background.

Guidelines for restoring Snapshot copies of Infinite Volumes

Observe the following guidelines:

- You must restore the entire Snapshot copy of the Infinite Volume.

  You cannot restore single files or parts of files. You also cannot restore a Snapshot copy of a single constituent.

- The Snapshot copy must be in a valid state.

  You cannot use admin privilege to restore a Snapshot copy of an Infinite Volume if the copy is in a partial or invalid state because the commands require diagnostic privilege. However, you can contact technical support to run the commands for you.

14.5 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 14-2).

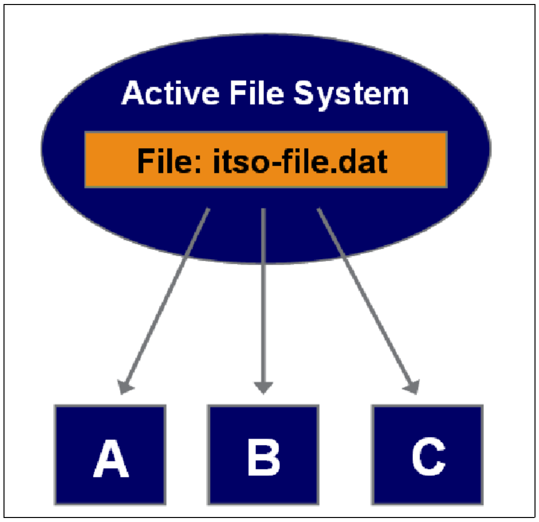


Figure 14-2 Identify active data to be snapped

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 14-3). These pointers are contained within metadata.

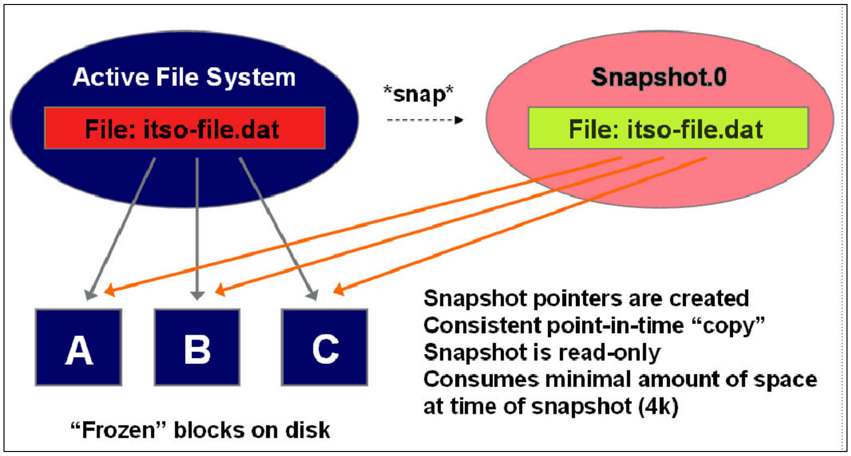


Figure 14-3 Pointers are created

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 14-4).

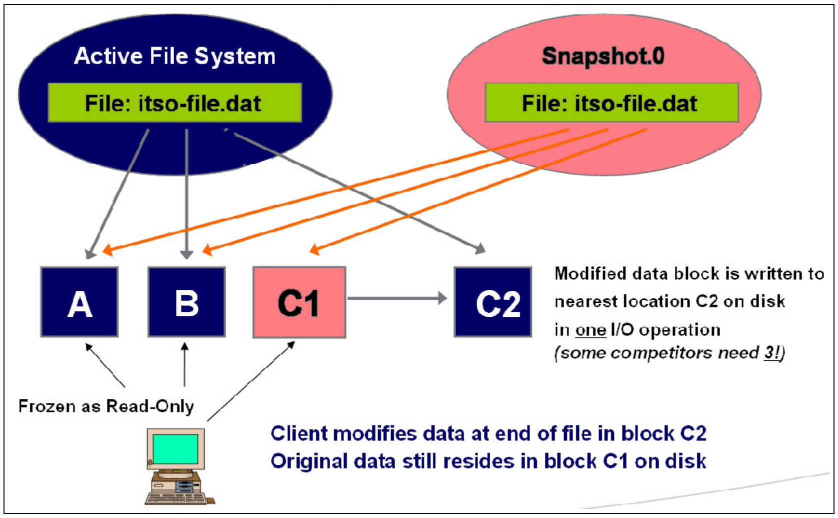

Active File System
File: itso-file.dat
Snapshot.0
File: itso-file.dat
A
B
C1
C2
Modified data block is written to nearest location C2 on disk in one I/O operation
(some competitors need 3!)
Frozen as Read-Only
Client modifies data at end of file in block C2
Original data still resides in block C1 on disk

Figure 14-4 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 14-5).

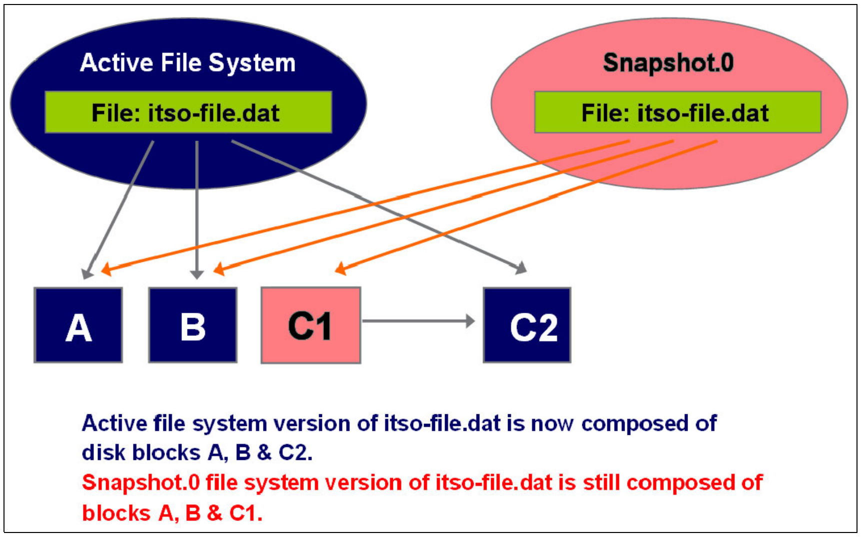

Active File System
File: itso-file.dat
Snapshot.0
File: itso-file.dat
A
B
C1
C2
Active file system version of itso-file.dat is now composed of disk blocks A, B & C2.
Snapshot.0 file system version of itso-file.dat is still composed of blocks A, B & C1.

Figure 14-5 Final result showing active points

14.6 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 14-6 illustrates how this works.

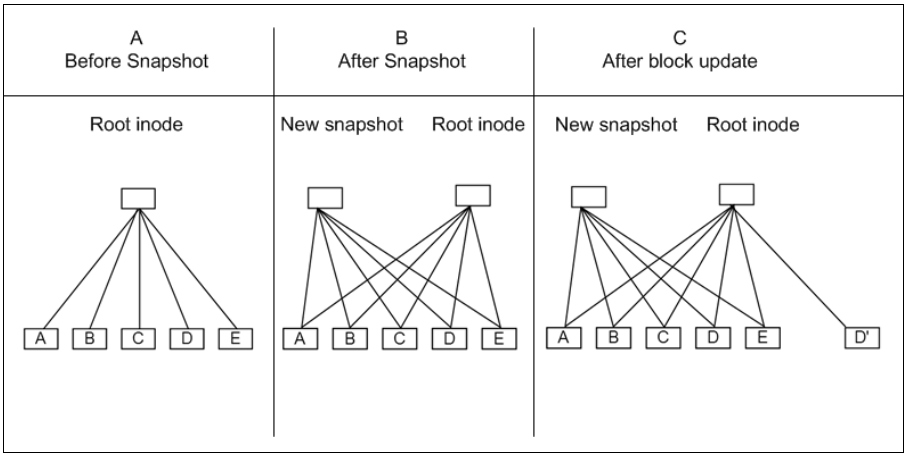


Figure 14-6 WAFL creates a Snapshot by duplicating the root inode

Column A in Figure 14-6 represents the basic situation before the Snapshot.

Column B in Figure 14-6 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 14-6 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 14-6 on page 221 to column C is illustrated in more detail here in Figure 14-7. 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.

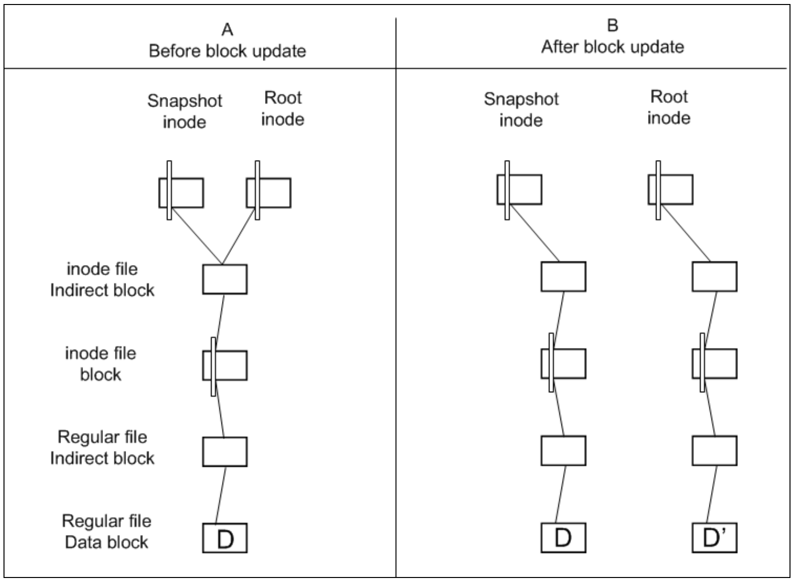


Figure 14-7 Block updates

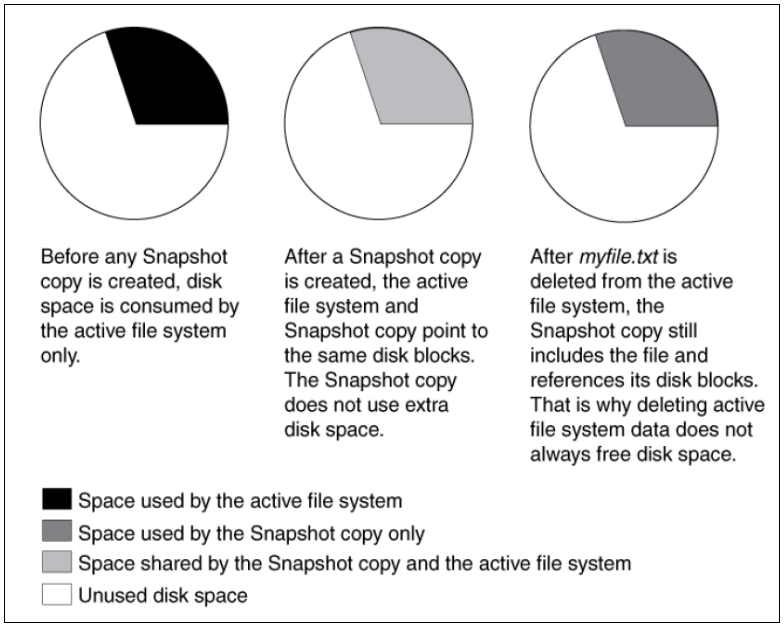
WAFL might be inefficient if it wrote this many blocks for each Network File System (NFS) write request. Instead, WAFL gathers up many hundreds of NFS requests before scheduling a write episode. During a write episode, WAFL allocates disk space for all the unclean data in the cache and schedules the 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.

14.6.1 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 14-8 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 14-8 How Snapshot copies consume disk space

14.6.2 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 14-9 shows how changing file content consumes disk space.

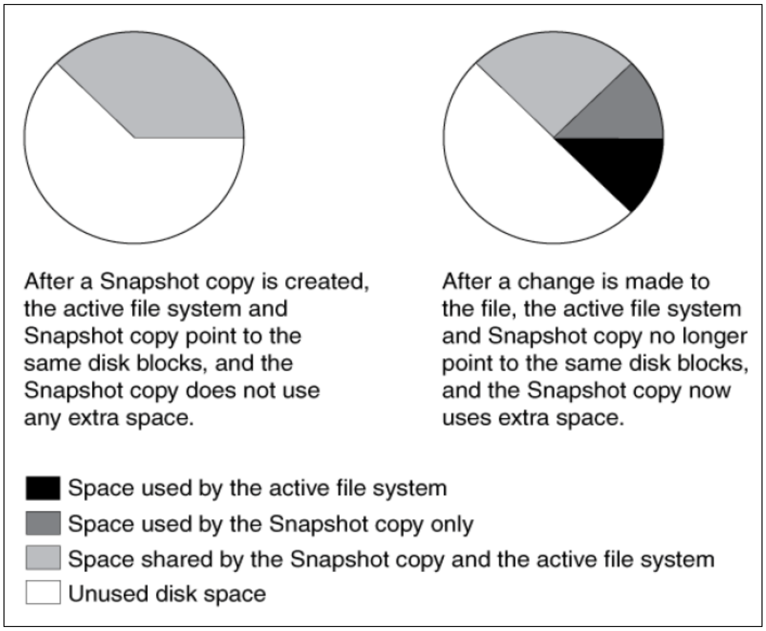


Figure 14-9 How changing file content consumes disk space

14.6.3 What the Snapshot copy reserve is

The Snapshot copy reserve sets a specific percentage of the disk space for Snapshot copies. For FlexVol volumes, the default Snapshot copy reserve is set to 5% of the disk space. By default, the Snapshot copy reserve is 5% of the disk space for a FlexVol volume and 0% for aggregates.

The active file system cannot consume the Snapshot copy reserve space, but the Snapshot copy reserve, if exhausted, can use space in the active file system.

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.

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 14-7) displays statistics about the amount of disk space on a volume.

Example 14-7 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 14-8).

Example 14-8 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 14-9.

Example 14-9 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 14-10) lists this unavailable space as used even though no files are stored in the active file system.

Example 14-10 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 in Example 14-11.

Example 14-11 Command output after 500,000 KB of data is added to the active file system

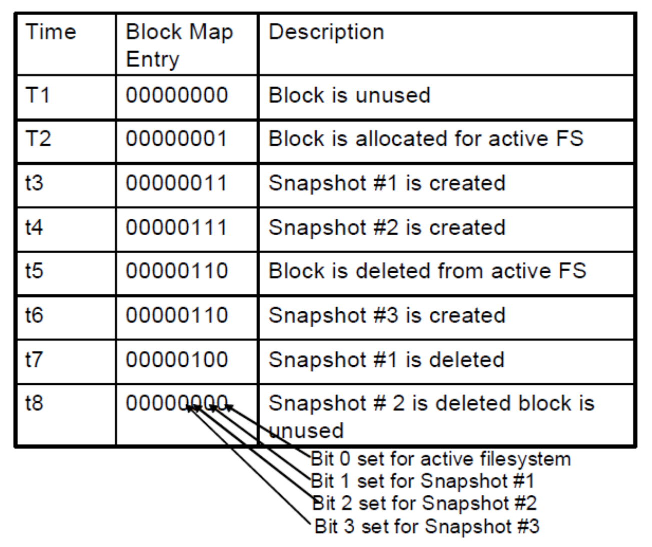
```
itsonas1*>  df /vol/LUN1
Filesystem           kbytes        used      avail   capacity
/vol/LUN1/           3000000       2500000   0       100%
/vol/LUN1/.snapshot  1000000       3500000   0       350%
itsonas1*>
```

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.

14.7 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 14-10 shows the lifecycle 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 14-10 Lifecycle 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 14-10, it occurs at time t8, after both Snapshots that reference the block have been removed.

14.7.1 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.

To create the Snapshot on an SVM `vs1`, volume `vol1`, named `snap1` manually, enter the command shown in Example 14-12.

Example 14-12 Creating the snapshot

```
cluster1::> volume snapshot create -vserver vs1 -volume vol1 -snapshot snap1
```

Notes:

- Due to the fact that you can run deduplication only on the active file system, Snapshot copies created before you run deduplication, locks the data in Snapshot, resulting in reduced space savings.
- To avoid conflicts between deduplication and Snapshot copies, run deduplication before creating new Snapshot copies or remove unnecessary Snapshot copies stored in deduplicated volumes.

14.7.2 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 `snap1` on volume `vol1`, SVM `vs1`, enter the command shown in Example 14-13.

Example 14-13 Manual deletion of snapshot

```
cluster1::> volume snapshot delete -vserver vs1 -volume vol1 -snapshot snap1
```

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.

14.8 SnapVault

SnapVault is a separately licensed feature in Cluster Data ONTAP that provides disk-based space-efficient data protection for storage systems. It performs asynchronous replication using snapshot copies of a primary volume.

A SnapVault backup is a collection of Snapshot copies on a FlexVol volume that you can restore data from if the primary data is not usable. Snapshot copies are created based on a Snapshot policy. The SnapVault backup backs up Snapshot copies based on its schedule and SnapVault policy rules.

You can convert SnapMirror to SnapVault relationship if needed. The major operational difference is that SnapVault allows you to have different retention schedules for Snapshot copies on the primary and secondary volume. You can keep up to 251 Snapshot copies per volume.

14.9 SnapVault basics

SnapVault, as shown in Figure 14-11, protects data on IBM N series storage systems by maintaining a number of read-only versions of that data on a SnapVault secondary system and the SnapVault primary system.

Note: SnapVault relationships are not supported on Infinite Volumes.

SnapVault is a disk-based storage backup feature of Cluster 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.

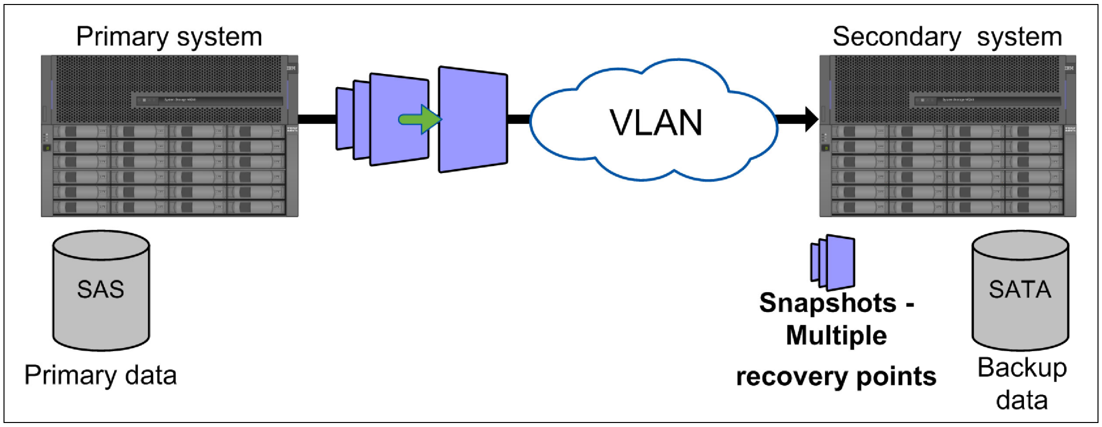


Figure 14-11 SnapVault technology overview

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.

Note: Notice that SnapVault has a read-only destination. This is different from 7-Mode, where the destination could be converted to RW using SnapMirror, this is not possible in Clustered Data ONTAP. You can create a FlexClone copy of a SnapVault backup.

14.9.1 SnapVault terms

The following terms are used to describe the SnapVault feature:

- Primary system:

  A system whose data is to be backed up.

- Primary volume:

  A volume that contains data that is to be backed up. In cascade or fan-out backup deployments, the primary volume is the volume that is backed up to a SnapVault backup, regardless of where in the chain the SnapVault source is. In a cascade chain configuration in which A has a mirror relationship to B and B has a SnapVault relationship to C, B serves as the source for the SnapVault backup even though it is a secondary destination in the chain.

- Secondary system:

  A system to which data is backed up.

- Secondary volume:

  A volume to which data is backed up from a primary volume. Such a volume can be a secondary or tertiary (and onward) destination in a cascade or fan-out backup configuration. The SnapVault secondary system maintains Snapshot copies for long-term storage and possible restore operations.

- SnapMirror label:

  An attribute that identifies Snapshot copies for the purpose of selection and retention in SnapVault backups. Each SnapVault policy configures the rules for selecting Snapshot copies on the primary volume and transferring the Snapshot copies that match a given SnapMirror label.

- SnapVault relationship:

  The backup relationship between a FlexVol on a primary system and its corresponding secondary system FlexVol.

- Snapshot copy:

  The backup images on the source volume that are created manually or automatically as scheduled by an assigned policy. Baseline Snapshot copies contain a copy of the entire source data being protected; subsequent Snapshot copies contain differential copies of the source data. Snapshot copies can be stored on the source volume or on a different destination volume in a different SVM or cluster.

  Snapshot copies capture the state of volume data on each source system. For SnapVault and mirror relationships, this data is transferred to destination volumes.

- SnapVault Snapshot basename:

  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 FlexVol to a corresponding FlexVol 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.

14.9.2 Which data gets backed up and restored from a SnapVault backup

You create SnapVault relationships to back up and restore volumes. You can select the Snapshot copies that the SnapVault relationship uses to backup and restore volumes.

The SnapVault operation backs up a specified volume on the primary system to the associated volume on the SnapVault secondary system. If necessary, data is restored from the SnapVault secondary volume back to the associated primary volume or to a different volume.

The Snapshot policy assigned to the source volume specifies when Snapshot copies are performed. The SnapVault policy assigned to the SnapVault relationship specifies which of the source volume Snapshot copies are replicated to the SnapVault backup.

14.9.3 Which data does not get backed up to a SnapVault backup

If you back up an entire SVM to a SnapVault backup by establishing a SnapVault relationship for each volume in the SVM, namespace and root information is not backed up. To protect namespace and root information for an SVM, you must manually create the namespace and root on the SnapVault secondary volume. When backing up LUNs to a SnapVault secondary volume, not all LUN information is replicated.

In SAN environments, the following LUN attributes are not replicated to the secondary volume:

- Path:

  The LUN in the SnapVault secondary volume can be in a different SVM or volume from the source LUN. Path-related metadata, such as persistent reservations, are not replicated to the SnapVault primary volume.

- Serial number
- Device ID
- UUID
- Mapped status
- Read Only state:

  The Read Only state is always set to true on the destination LUN.

- NVFAIL attribute:

  The NVFAIL attribute is always set to false on the destination LUN.

You can set persistent reservations for LUNs on the SnapVault secondary volume.

14.9.4 Clustered Data ONTAP SnapVault highlights

SnapVault allows you to have asymmetric snapshot retention on your primary and secondary volumes. The typical use case for SnapVault is that you want to keep backups for a short period of time (maybe only a week) on your primary system, but you want to keep backups on your secondary for a long period of time (possibly several years):

- Replication based on Snapshot
- One baseline, forever incremental backups
- Multiple recovery points (each incremental copy = recovery point)
- Supported over any distance
- In case of data loss, recovery is fast and simple
- Storage efficiency preserved over the wire
- Take advantage of high-density storage (SATA)
- Reduce reliance on tape
- End user browse and restore - ability to use NFS and CIFS to allow end users restores
- Simple management

14.10 7-Mode versus Clustered Data ONTAP SnapVault

It is not possible to transition 7-Mode SnapVault relationships to Clustered Data ONTAP without needing a rebaseline. New SnapVault relationships must be created in Clustered Data ONTAP. This requires a baseline transfer.

Table 14-4 compares the 7-Mode and Clustered Data ONTAP SnapVault. Most of the points are the same from 7-Mode SnapVault except for storage efficiency preserved over the wire, which is new for Clustered Data ONTAP.

Table 14-4 7-Mode versus Clustered Data ONTAP SnapVault

Feature	7-Mode	Clustered Data ONTAP
Replication Granularity	qtree	Flex Volume
Baseline and Incremental Backup	Yes	Yes
Baseline and Incremental Restore	Yes	Yes
Single File/LUN Restore (Using NDMP Copy)	Yes	Yes
Data ONTAP Version Interoperability	Yes	Yes
Schedule-Driven Update	Yes	Yes, policy driven
Secondary Snapshot Management	Yes	Yes
Usable Replica (read access)	Yes	Yes
Tape Integration	Yes	Yes, using NDMP or SMTape
Primary/Secondary Deduplication or Compression	Yes, but savings lost over wire	Yes, savings preserved over wire
Auto Grow Secondary	No	Yes
SnapMirror to SnapVault Conversion	No	Yes, but SnapVault to SnapMirror not possible

Notes:

- SnapVault relationships between a 7-Mode and Clustered Data ONTAP system are not possible. Also, OSSV is not supported in Clustered Data ONTAP 8.2.
- Clustered Data ONTAP SnapVault supports 64-bit aggregates only.

Even when creating a SnapVault relationship and not a SnapMirror relationship the commands for both is `snapmirror`. When using the `snapmirror` commands, you will have to specify the type of relationship you are creating (`-type DP` or `-type XDP`). A SnapVault relationship is type `XDP`.

14.11 How a SnapVault backup works

Backing up volumes to a SnapVault backup involves starting the baseline transfers, making scheduled incremental transfers, and restoring data upon request.

14.11.1 Baseline transfers

A baseline transfer occurs when you initialize the SnapVault relationship. When you do this, Data ONTAP creates a new Snapshot copy. Data ONTAP transfers the Snapshot copy from the primary volume to the secondary volume. This Snapshot copy is the baseline of the volume at the time of the transfer and is a complete transfer, not an incremental transfer. As a result, none of the other Snapshot copies on the primary volume are transferred as part of the initial SnapVault transfer, regardless of whether they match rules specified in the SnapVault policy.

14.11.2 Incremental transfers

The source system creates incremental Snapshot copies of the source volume as specified by the Snapshot policy that is assigned to the primary volume. Each Snapshot copy for a specific volume contains a label that is used to identify it.

The SnapVault secondary system selects and retrieves specifically labeled incremental Snapshot copies, according to the rules that are configured for the SnapVault policy that is assigned to the SnapVault relationship. The Snapshot label is retained to identify the backup Snapshot copies.

Snapshot copies are retained in the SnapVault backup for as long as is needed to meet your data protection requirements. The SnapVault relationship does not configure a retention schedule, but the SnapVault policy does specify number of Snapshot copies to retain.

14.11.3 SnapVault backup updates

At the end of each Snapshot copy transfer session, which can include transferring multiple Snapshot copies, the most recent incremental Snapshot copy in the SnapVault backup is used to establish a new common base between the primary and secondary volumes and is exported as the active file system.

14.11.4 Data restore

If data needs to be restored to the primary volume or to a new volume, the SnapVault secondary transfers the specified data from the SnapVault backup.

14.11.5 SnapVault backups with data compression

SnapVault relationships preserve storage efficiency when replicating data from the source to the SnapVault secondary volume except when additional data compression is enabled.

If additional compression is enabled on the SnapVault secondary volume, storage efficiency is affected as follows:

- Storage efficiency is not preserved for data transfers between the primary and secondary volumes.
- You do not have the option of returning to replicating data while preserving storage efficiency.

14.11.6 Data protection of SVM namespace and root information

Backups to secondary volumes in SnapVault relationships between FlexVol volumes replicate only volume data, not the SVM namespace or root information.

SnapVault relationships replicate only volume data. If you want to back up an entire SVM to a SnapVault secondary volume, you must first create SnapVault relationships for every volume in the SVM.

To provide data protection of the SVM namespace information, you must manually create the namespace on the SnapVault secondary immediately after the first data transfer is completed for all of the volumes in the SVM and while the source SVM volumes are still active. When subsequent changes are made to the namespace on the source SVM, you must manually update the namespace on the destination SVM.

You cannot create the namespace for an SVM on a SnapVault secondary volume if only a subset of the SVM volumes are in a SnapVault relationship, or if only a subset of the SVM volumes have completed the first data transfer.

14.12 Supported data protection deployment configurations

A simple data protection deployment consists of a FlexVol volume or Infinite Volume in a single mirror relationship or a FlexVol volume in a SnapVault relationship. More complex deployment configurations that provide additional data protection consist of a cascade chain of relationships between FlexVol volumes or a set of fan-out relationships for a FlexVol volume or Infinite Volume.

Although a single volume-to-volume relationship does provide data protection, your data protection needs might require the additional protection that is provided by more complex cascade and fan-out configurations.

An example of a *cascade chain* is an A to B to C configuration. In this example, A is the source that is replicated to B as a data protection mirror, and B is the primary that is backed up to C as a SnapVault backup. Cascade chains can be more complex than A to B to C, but the more relationships that are involved in the chain, the greater the number of temporary locks on volumes while replication or update operations are in progress.

The three types of cascade chain relationships that you can configure are as follows:

- Mirror-mirror cascade (for FlexVol volumes only)
- Mirror-SnapVault cascade (for FlexVol volumes only)
- SnapVault-mirror cascade (for FlexVol volumes only)

An example of a *fan-out* is an A to B and A to C backup or mirror replication configuration. In this example, A is the primary source that is replicated to both B (either in a mirror or SnapVault relationship) and C.

In a fan-out relationship structure, the source is replicated to multiple destinations, which can be mirror or SnapVault destinations. Only one SnapVault relationship is allowed in a fan-out.

- Mirror-SnapVault fan-out (for FlexVol volumes only)
- Multiple-mirrors fan-out (for FlexVol volumes and Infinite Volumes)

Note: Only one SnapVault relationship is supported in a cascade chain configuration, but many SnapVault relationships are supported in a fan-out configuration; multiple mirror relationships are supported.

14.12.1 Basic data protection configuration

A basic data protection deployment (Figure 14-12) consists of two volumes, either FlexVol volumes or Infinite Volumes, in a one-to-one, source-to-destination relationship. This deployment backs up data to one location, which provides a minimal level of data protection.

In a data protection configuration, source volumes are the data objects that need to be replicated. Typically, users can access and write to source volumes.

Destination volumes are data objects to which the source volumes are replicated. Destination volumes are read-only. Destination FlexVol volumes are usually placed on a different SVM from the source SVM. Destination Infinite Volumes must be placed on a different SVM from the source SVM. Destination volumes can be accessed by users in case the source becomes unavailable. The administrator can use SnapMirror commands to make the replicated data at the destination accessible and writable.

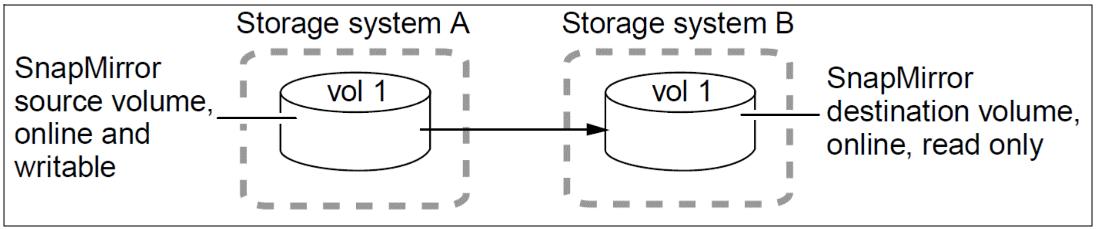

Storage system A
Storage system B
SnapMirror source volume, online and writable
vol 1
vol 1
SnapMirror destination volume, online, read only

Figure 14-12 Basic data protection deployment

14.12.2 Source to destination to tape backup

A common variation of the basic data protection backup deployment adds a tape backup of a destination FlexVol volume as shown in Figure 14-13. By backing up to tape from the destination volume, you do not subject the heavily accessed source volume to the performance degradation and complexity of a direct tape backup.

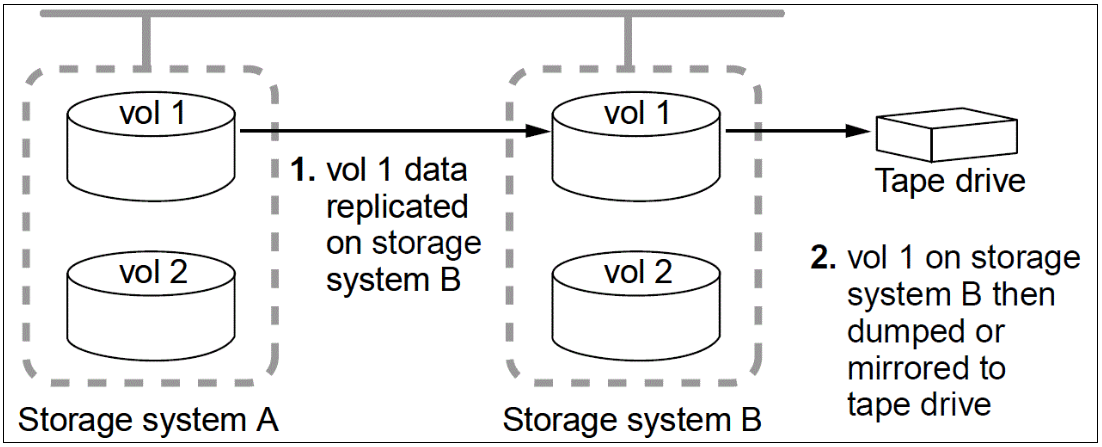

vol 1
vol 2
1. vol 1 data replicated on storage system B
vol 1
vol 2
Tape drive
2. vol 1 on storage system B then dumped or mirrored to tape drive
Storage system A
Storage system B

Figure 14-13 Data protection chain deployment with a tape backup

NDMP is required for this configuration, and Infinite Volumes do not support NDMP.

14.12.3 Mirror to mirror cascade

A mirror-mirror cascade (Figure 14-14) deployment is supported on FlexVol volumes and consists of a chain of mirror relationships in which a volume is replicated to a secondary volume and the secondary is replicated to a tertiary volume. This deployment adds one or more additional backup destinations without degrading performance on the source volume.

By replicating source A (as shown in this illustration) to two different volumes (B and C) in a series of mirror relationships in a cascade chain, you create an additional backup. The base for the B-to-C relationship is always locked on A to ensure that the backup data in B and C always stay synchronized with the source data in A.

If the base Snapshot copy for the B-to-C relationship is deleted from A, the next update operation from A to B fails and an error message is generated that instructs you to force an update from B to C. The forced update establishes a new base Snapshot copy and releases the lock, which enables subsequent updates from A to B to finish successfully.

If the volume on B becomes unavailable, you can synchronize the relationship between C and A to continue protection of A without performing a new baseline transfer. After the resynchronize operation finishes, A is in a direct mirror relationship with C, bypassing B.

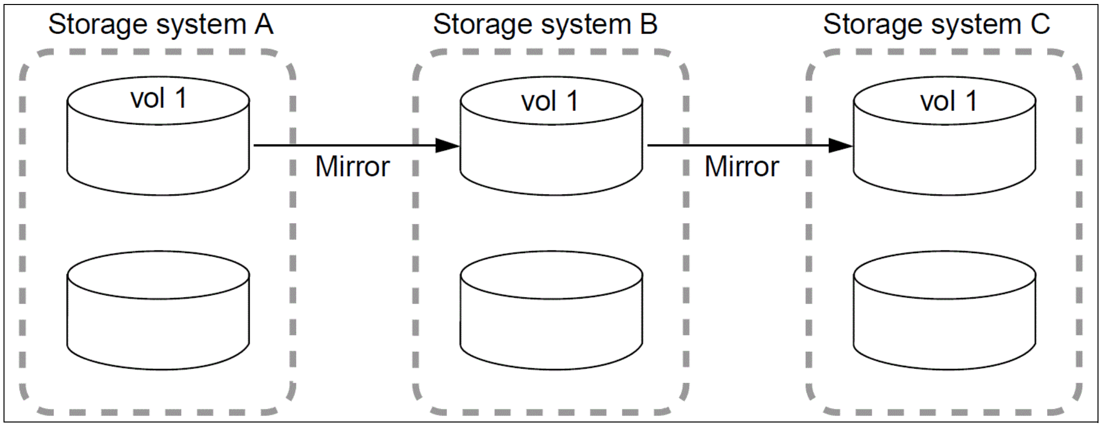


Figure 14-14 Mirror to mirror cascade chain

14.12.4 Mirror to SnapVault cascade

A mirror-SnapVault cascade (Figure 14-15) deployment is supported on FlexVol volumes and consists of a chain of relationships in which a volume is replicated to a destination volume and then the destination volume becomes the primary for a SnapVault backup on a tertiary volume. This deployment adds a SnapVault backup, which fulfills more strict protection requirements.

In a typical mirror-SnapVault cascade, only the exported Snapshot copies from the mirror destination are transferred to the SnapVault secondary when the SnapVault update occurs. These exported Snapshot copies are created by Data ONTAP and have a prefix of snapmirror and a hidden **`snapmirror-label`** called `sm_created`. The SnapVault backup, using a SnapVault policy and applying a rule that identifies Snapshot copies with the `sm_created` **`snapmirror-label`**, backs up the exported Snapshot copies. Only in the case of mirror-SnapVault cascades is the **`snapmirror-label`** `sm_created` used.

Note: A cascade chain can contain multiple mirror relationships but only one SnapVault relationship. The SnapVault relationship can occur anywhere in the chain, depending on your data protection requirements.

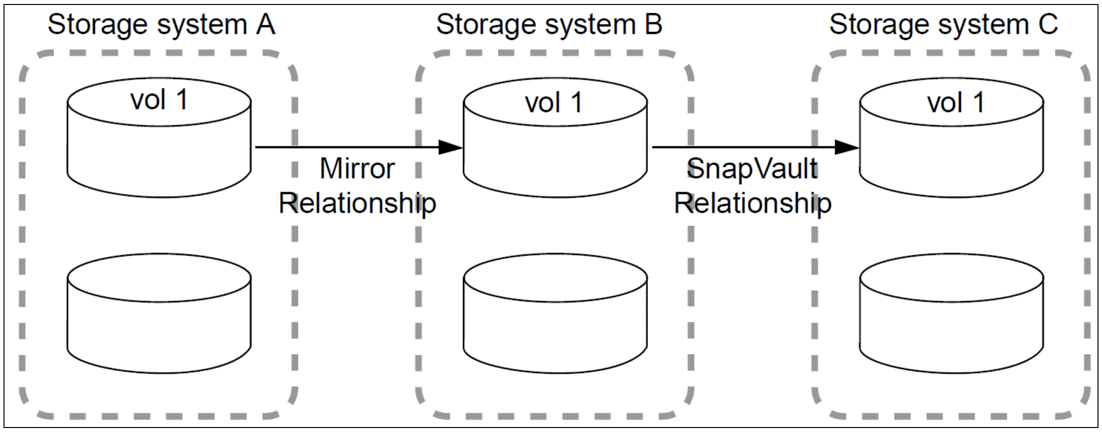


Figure 14-15 Mirror to SnapVault cascade chain

14.12.5 SnapVault to mirror cascade

A SnapVault-mirror cascade (Figure 14-16) consists of a chain of relationships in which a volume has a SnapVault backup on a secondary volume, and then that secondary volume data is replicated to a tertiary volume. In effect, this deployment provides two SnapVault backups.

A SnapVault-mirror cascade deployment is only supported on FlexVol volumes if the first leg in the cascade is a SnapVault backup. In cascade chains that include a SnapVault relationship, updates to the SnapVault backup always include the Snapshot copy base of the SnapVault relationship in addition to the Snapshot copies that are selected in conformance with the SnapVault policy that is assigned to the relationship. This ensures that the common Snapshot copy for B is always available on A (as shown in the following illustration), which enables you to establish a direct relationship for A to C, if necessary. The extra base Snapshot copy is replaced with a newer common Snapshot copy at every subsequent SnapVault update.

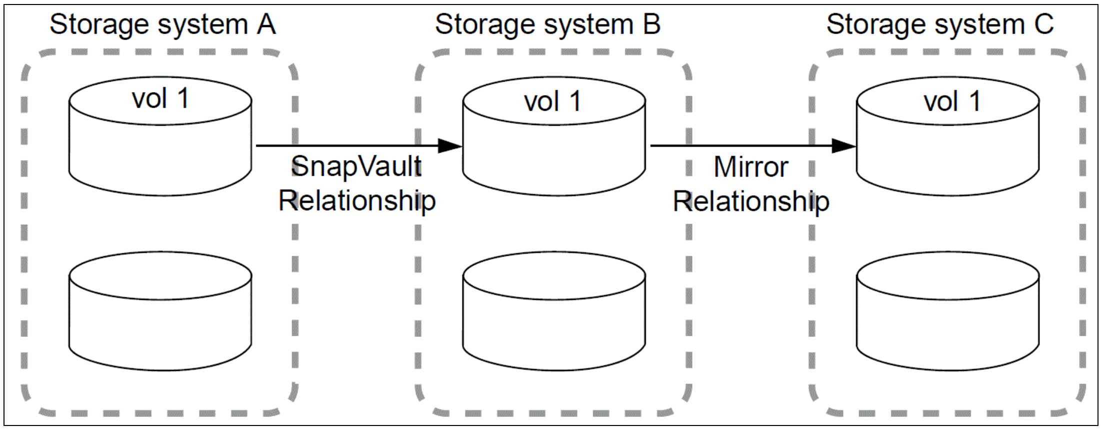


Figure 14-16 SnapVault to mirror cascade chain

14.12.6 Fan-in and fan-out deployments

Since replication is now done at the volume level, you cannot have multiple source volumes backing up to the same destination volume similar to the way you could have multiple source qtrees backing up to one volume with 7-Mode SnapVault. You can have volumes from different SVMs backing up to volumes on the same SVM. Note that in 8.2 the number of cluster peers is limited to 8. This means that volumes from a maximum of 7 different clusters can back up to a single destination cluster.

The fan-out limit of 1:4 applies to the combined number of SnapMirror and SnapVault relationships. One volume can have a maximum of 4 relationships of any combination of SnapMirror and SnapVault.

A mirror-SnapVault fan-out deployment is supported on FlexVol volumes and consists of a source volume that has a direct mirror relationship to a secondary volume and also a direct SnapVault relationship to a different secondary volume.

A multiple-mirrors fan-out deployment is supported on FlexVol volumes and Infinite Volumes, and consists of a source volume that has a direct mirror relationship to multiple secondary volumes.

Note: A fan-out deployment might not provide as much data protection as a cascade chain.

Fan-in and fan-out limitations are shown in Figure 14-17.

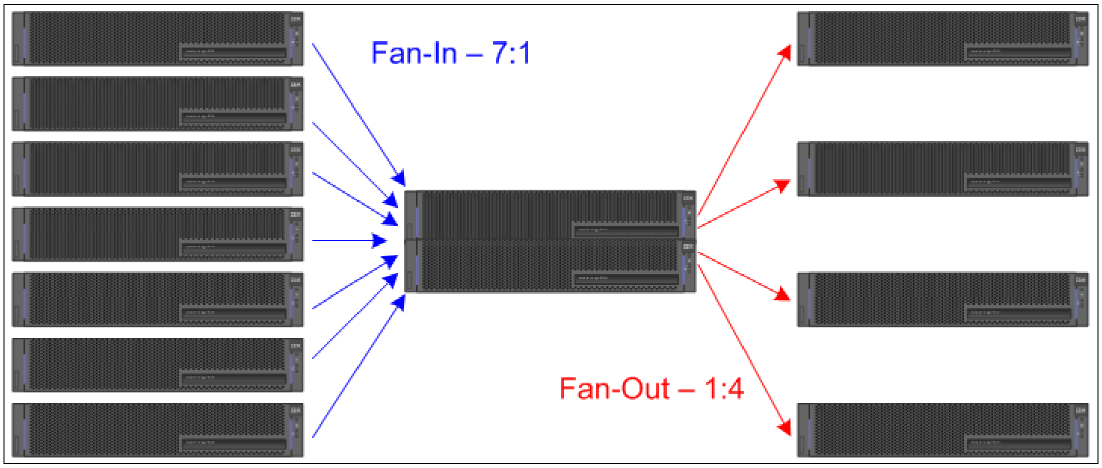


Figure 14-17 Fan-in and fan-out limitations

14.13 Protecting data on FlexVol volumes by using SnapVault

You can create a SnapVault relationship between FlexVol volumes and assign a SnapVault policy to it to create a SnapVault backup. A SnapVault backup contains a set of read-only backup copies, located on a secondary volume.

A SnapVault backup differs from a set of Snapshot copies or a set of mirror copies on a destination volume. In a SnapVault backup, the data in the secondary volume is periodically updated to keep the data in the secondary volume up to date with changes made in the primary data.

14.13.1 Creating SnapVault backups on FlexVol volumes

You configure a SnapVault relationship and assign a SnapVault policy to the relationship to establish a SnapVault backup.

General guidelines for creating a SnapVault relationship

The following guidelines apply to all SnapVault relationships:

- A volume can be in multiple relationships, either as the secondary or the primary.

  A volume can be the primary for multiple secondaries and also the secondary for another primary.

- A volume can be the secondary for only one SnapVault relationship.
- You cannot configure SnapVault relationships from multiple primary volumes to a single SnapVault secondary volume.

  For example, if you want to back up an entire SVM to a SnapVault backup, then you must create a separate secondary volume for each volume in the SVM, and create a separate SnapVault relationship for each primary volume.

- You can configure SnapVault relationships to be used simultaneously with data protection mirror relationships.
- Primary or secondary volumes cannot be 32-bit volumes.
- The primary of a SnapVault backup should not be a FlexClone volume.

  The relationship will work, but the efficiency provided by FlexClone volumes is not preserved.
- A SnapVault secondary volume cannot be the primary volume of FlexCache volumes.
- Primary and secondary volumes must have the same `vol lang` settings.
- After you establish a SnapVault relationship, you cannot change the language assigned to the secondary volume.
- A SnapVault relationship can be only one leg of a cascade chain.
- After you establish a SnapVault relationship, you can rename primary or secondary volumes.

  If you rename a primary volume, it can take a few minutes for the relationship to recover from the name change.

Guidelines for a SnapVault relationship to a prepopulated secondary

Typically, you create a prepopulated secondary volume when you copy a primary volume to a secondary volume using tape. This process is known as tape seeding.

If the SnapVault secondary volume already contains data, you can create a SnapVault relationship by using the **`snapmirror resync`** command with the **`-type XDP`** option.

Before creating a SnapVault relationship to a prepopulated secondary, observe the following guidelines:

- The primary and secondary volumes must have a common Snapshot copy.
- Snapshot copies on the secondary volume that are newer than the common Snapshot copy are deleted.

  When a SnapVault relationship is created, all Snapshot copies on the secondary volume that are more recent than the common Snapshot copy and that are not present on the primary volume are deleted. Newer Snapshot copies on the primary volume that match the configured SnapVault policy are transferred to the secondary volume according to the SnapVault policy.

  You can use the **`-preserve`** option to keep any Snapshot copies that are more recent than the common Snapshot copy on the SnapVault secondary volume and that are not present on the primary volume.

  When you use the **`-preserve`** option, data on the secondary volume is logically made the same as the common Snapshot copy. All newer Snapshot copies on the primary volume that match the SnapVault policy are transferred to the secondary volume.

  This option is useful when the latest common Snapshot copy is deleted from the primary volume but another, older common Snapshot copy between the primary and secondary volumes still exists.

Notes:

- If the aggregate that contains the secondary volume of the SnapVault backup is out of space, SnapVault updates fail, even if the secondary volume has space.
- Ensure that there is free space in the aggregate and the volume for transfers to succeed.

Prepopulated SnapVault secondary scenarios

There are several ways in which a secondary FlexVol volume for a SnapVault relationship might be prepopulated with data.

In the following scenarios, a SnapVault secondary might be populated before a SnapVault relationship is created:

- You used tape backups to provide a baseline transfer to a secondary volume.
- A SnapVault primary volume in a cascade becomes unavailable.

  You have a data protection mirror relationship between a source and a destination volume (a mirror relationship from A to B) and a SnapVault relationship between the secondary destination volume and a tertiary destination volume (a SnapVault relationship from B to C). The backup cascade chain is A mirror to B and B SnapVault backup to C. If the volume on B becomes unavailable, you can configure a SnapVault relationship directly from A to C. The cascade chain is now A SnapVault backup to C, where C was prepopulated with data.
- You created a SnapVault relationship between two flexible clones.

  You create a SnapVault relationship between two flexible clones for which their respective parent volumes are already in a SnapVault relationship.
- You extended the SnapVault backup protection beyond 251 Snapshot copies.

  To extend the SnapVault backup protection beyond the volume limit of 251 Snapshot copies, you can clone the secondary volume. The original SnapVault secondary volume is the parent volume for the new flexible clone.
- You restored data from a SnapVault secondary to a new primary volume.

  You have a SnapVault relationship from A to B. A becomes inaccessible, so the SnapVault secondary volume (B) is used for a baseline restore operation to a new SnapVault secondary volume (C).

  After the restore operation finishes, you establish a new SnapVault relationship from the new secondary volume (C), which now becomes the primary volume, and the original SnapVault secondary volume (in other words, C to B). The disk to disk backup relationship is now C to B, where B was prepopulated with data.
- You deleted the base Snapshot copy from the primary volume.

  You deleted the base Snapshot copy from the primary volume that was used for a SnapVault transfer, but another, older Snapshot copy exists that is common between the primary and secondary volumes.

14.13.2 Creating a SnapVault backup in an empty FlexVol volume

You can protect data that has long-term storage requirements on a FlexVol volume by replicating selected Snapshot copies to a SnapVault backup on another SVM or cluster.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster, and SVM administrator privileges to perform this task for an SVM.
- If the primary and secondary volumes are in different SVMs, the SVMs must be in a peer relationship.

  If the primary and secondary volumes are in different clusters, the clusters must be in a peer relationship.
- A SnapVault policy must exist.

  You must either create one or accept the default SnapVault policy (named *XDPDefault*) that is automatically assigned.

  Only Snapshot copies with the labels configured in the SnapVault policy rules are replicated in SnapVault operations.
- The Snapshot policy assigned to the primary volume must include the **snapmirror-label** attribute.

  You can create a new Snapshot policy by using the **volume snapshot policy add-schedule** command, or you can modify an existing policy by using the **volume snapshot policy modify-schedule** command to set the **snapmirror-label** attribute for the set of Snapshot copies that you want backed up to the SnapVault secondary volume. Other Snapshot copies on the primary volume are ignored by the SnapVault relationship.
- Your work environment must be able to accommodate the time it might take to transfer a baseline Snapshot copy with a large amount of data.

> **Note:** Even when creating a SnapVault relationship and not a SnapMirror relationship, the command for both is **snapmirror**. When using the **snapmirror** commands, you will have to specify the type of relationship you are creating (**-type DP** or **-type XDP**). A SnapVault relationship is type **XDP**.

Steps

Follow these steps:

1. On the destination SVM, create a SnapVault secondary volume with a volume type DP.
2. Create a schedule that Data ONTAP uses to update the SnapVault relationship by using the **job schedule cron create** command.

   Example 14-14 creates a schedule that runs on the weekend at 3 a.m.

   Example 14-14 Creating a schedule

   ```
   vserverB::> job schedule cron create -name weekendcron -dayofweek "Saturday,
   Sunday" -hour 3 -minute 0
   ```
3. On the source SVM, create a Snapshot copy policy that contains the schedule of when Snapshot copies with **snapmirror-label** attributes occur by using the **volume snapshot policy create** command with the **snapmirror-label** parameter, or use the default Snapshot copy policy called **default**.

Example 14-15 creates a Snapshot copy policy called `keep-more-snapshot`.

Example 14-15 Creating snapshot policy

```
vserverB::> snapshot policy create -vserver vs1 -policy keep-more-snapshot
-enabled true -schedule1 weekly -count1 2 -prefix1 weekly -schedule2 daily
-count2 6 -prefix2 daily -schedule3 hourly -count3 8 -prefix3 hourly
```

The name specified in the **`snapmirror-label`** attribute for the new Snapshot policy must match the **`snapmirror-label`** attribute that is specified in the SnapVault policy. This ensures that all subsequent Snapshot copies created on the primary volume have labels that are recognized by the SnapVault policy.

The default Snapshot copy policy has two **`snapmirror-label`** attributes associated with it, daily and weekly.

4. Create a SnapVault policy by using the **`snapmirror policy create`** command, or use the default SnapVault policy called **`-XDPDefault`**.

   Example 14-16 creates a SnapVault policy called `vserverB-vault-policy.`

   Example 14-16 Creating a SnapVault policy

   ```
   vserverB::> snapmirror policy create -vserver vserverB -policy
   vserverB-vault-policy
   ```

5. Add the **`snapmirror-label`** attribute to the SnapVault policy you created by using the **`snapmirror policy add-rule`** command.

   If you used the XDPDefault SnapMirror policy, you do not need to perform this step. The XDPDefault SnapVault policy uses the daily and weekly **`snapmirror-label`** attributes specified by the default Snapshot copy policy.

   Example 14-17 adds a rule to the `vserverB-vault-policy` to transfer Snapshot copies with the `weekly` **`snapmirror-label`** attribute and to keep 40 Snapshot copies.

   Example 14-17 Adding rules to SnapVault policy

   ```
   vserverB::> snapmirror policy add-rule -vserver vserverB -policy
   vserverB-vault-policy -snapmirror-label weekly -keep 40
   ```

6. On the destination SVM, create a SnapVault relationship and assign a SnapVault policy by using the **`snapmirror create`** command with the **`-type XDP`** parameter and the **`-policy`** parameter.

   In the path specification, a single name is interpreted as a volume name in the SVM from which the command is executed. To specify a volume in a different SVM or in a different cluster, you must specify the full path name.

   Example 14-18 creates a SnapVault relationship between the primary volume `srcvolA` on SVM `vserverA` and the empty secondary volume `dstvolB` on SVM `vserverB`. It assigns the SnapVault policy named `vserverB-vault-policy` and uses the `weekendcron` schedule.

   Example 14-18 Creating a SnapVault relationship and assigning a SnapVault policy

   ```
   vserverB::> snapmirror create -source-path vserverA:srcvolA -destination-path
   vserverB:dstvolB -type XDP -policy vserverB-vault-policy -schedule weekendcron
   ```

7. On the destination SVM, initialize the SnapVault relationship by using the **snapmirror initialize** command to start a baseline transfer.

   The command creates a new Snapshot copy that is transferred to the secondary volume and used as a baseline for subsequent incremental Snapshot copies. The command does not use any Snapshot copies that currently exist on the primary volume.

   Example 14-19 begins the relationship initialization by creating and transferring a baseline Snapshot copy to the destination volume dstvolB on SVM vserverB.

   Example 14-19 Beginning the relationship initialization

   ```
   vserverB::> snapmirror initialize -destination-path vserverB:dstvolB
   ```

   Note: Creating a baseline for a large amount of data might take a while.

14.13.3 Creating the SnapVault relationship of a mirror-SnapVault cascade

The SnapVault relationship of a mirror-SnapVault cascade requires a different configuration from a SnapVault relationship that is not a part of a mirror-SnapVault cascade.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster, and SVM administrator privileges to perform this task for an SVM.
- If the primary and secondary volumes are in different SVMs, the SVMs must be in a peer relationship.

  If the primary and secondary volumes are in different clusters, the clusters must be in a peer relationship.

About this task

The Snapshot copies that are exported to the mirror destination are ones that are created by Data ONTAP. These Snapshot copies have a **snapmirror-label** called sm_created associated with them. Only these Snapshot copies are replicated from the mirror to the SnapVault backup. To configure the SnapVault relationship of the mirror-SnapVault cascade, the SnapVault policy associated with the SnapVault relationship must have the sm_created **snapmirror-label** in a rule to restrict the number of Snapshot copies retained on the SnapVault backup.

Steps

Follow these steps:

1. On the destination SVM, create a SnapVault secondary volume with a volume type DP.
2. Create a SnapVault policy by using the **snapmirror policy create** command, or use the default SnapVault policy called XDPDefault.
3. Add the sm_created **snapmirror-label** to the SnapVault policy by using the **snapmirror policy add-rule** command.

   Only the sm_created rule will be observed. Any other rules associated with the SnapVault policy, like the daily or weekly rule, will be disregarded.

   Example 14-20 adds a rule to the XDPDefault policy to transfer Snapshot copies with the sm_created **snapmirror-label** and to keep 40 Snapshot copies.

Example 14-20 Adding rule to policy

```
vserverB::> snapmirror policy add-rule -vserver vserverC -policy XDPDefault
-snapmirror-label sm_created -keep 40
```

4. On the destination SVM, create a SnapVault relationship and assign a SnapVault policy by using the **snapmirror create** command with the **-type XDP** parameter and the **-policy** parameter.

   The Example 14-21 creates a SnapVault relationship between the primary volume `srcvolB` on SVM `vserverB` and the empty secondary volume `dstvolC` on SVM `vserverC`. It assigns the SnapVault policy named `XDPDefault`.

   Example 14-21 Creating a SnapVault relationship

   ```
   vserverC::> snapmirror create -source-path vserverB:srcvolB -destination-path
   vserverC:dstvolC -type XDP -policy XDPDefault
   ```

5. On the destination SVM, initialize the SnapVault relationship by using the **snapmirror initialize** command to start a baseline transfer.

   The Example 14-22 begins the relationship initialization by creating and transferring a baseline Snapshot copy to the secondary volume `dstvolC` on SVM `vserverC`.

   Example 14-22 Initializing the SnapVault relationship

   ```
   vserverC::> snapmirror initialize -destination-path vserverC:dstvolC
   ```

   Note: Creating a baseline for a large amount of data might take a while.

14.13.4 Preserving a Snapshot copy on the primary source volume

In a mirror-SnapVault cascade, you must preserve a Snapshot copy on the primary source volume until it transfers to the secondary volume of the SnapVault backup. For example, you want to ensure that application-consistent Snapshot copies are backed up.

Before you begin

You must have created the mirror-SnapVault cascade.

Steps

Follow these steps:

1. Ensure that the Snapshot copy you want to preserve has a **snapmirror-label** by using the **volume snapshot show** command.
2. If the Snapshot copy does not have a **snapmirror-label** associated with it, add one by using the **volume snapshot modify** command.

   Example 14-23 adds a **snapmirror-label** called `exp1` to the Snapshot copy called `snapappa`.

   Example 14-23 Adding snapmirror-label to Snapshot copy

   ```
   clust1::> volume snapshot modify -volume vol1 -snapshot snapappa
   -snapmirror-label exp1
   ```

3. Preserve the Snapshot copy on the source volume by using the **snapmirror snapshot-owner create** command to add an owner name to the Snapshot copy.

Example 14-24 adds `ApplicationA` as the owner name to the `snapappa` Snapshot copy in the `vol1` volume on the `vs1` SVM.

Example 14-24 Adding owner name to Snapshot copy

```
clust1::> snapmirror snapshot-owner create -vserver vs1 -volume vol1 -snapshot
snapappa -owner ApplicationA
```

4. Update the destination volume of the data protection mirror relationship by using the **`snapmirror update`** command.

   Alternatively, you can wait for the scheduled update of the data protection mirror relationship to occur.

5. Update the secondary volume of the SnapVault relationship to transfer the specific Snapshot copy from the SnapMirror destination volume to the SnapVault secondary volume by using the **`snapmirror update`** command with the **`-source-snapshot`** parameter.

6. Remove the owner name from the primary source volume by using the **`snapmirror snapshot-owner delete`** command.

   Example 14-25 removes `ApplicationA` as the owner name to the `snapappa` Snapshot copy in the `vol1` volume on the `vs1` SVM.

   Example 14-25 Removing owner name

```
clust1::> snapmirror snapshot-owner delete -vserver vs1 -volume vol1 -snapshot
snapappa -owner ApplicationA
```

14.13.5 Creating a SnapVault backup in a prepopulated FlexVol volume

You can protect data that has long-term storage requirements on a FlexVol Volume by replicating selected Snapshot copies to a SnapVault backup on another SVM or cluster. The SnapVault secondary volume might contain data that already exists from a previous data protection mirror or SnapVault relationship or has been loaded from a tape backup.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster, and you must have SVM administrator privileges to perform this task for an SVM.
- If the primary and secondary volumes are in different SVMs, the SVMs must be in a peer relationship.

  If the primary and secondary volumes are in different clusters, the clusters must be in a peer relationship.
- The secondary volume must be prepopulated with data.
- A SnapVault policy must exist.

  You must either create one or accept the default SnapVault policy (named `XDPDefault`) that is automatically assigned.

  The SnapVault policy configuration includes the **`snapmirror-label`** attribute that is used to select Snapshot copies on the primary volume and match Snapshot copies between the primary and secondary volumes. Only Snapshot copies with the labels configured in the SnapVault policy rules are replicated in SnapVault operations.
- The Snapshot policy assigned to the primary volume must include the **`snapmirror-label`** attribute.

The name specified in the **snapmirror-label** attribute for the new Snapshot policy must match the **snapmirror-label** attribute that is specified in the SnapVault policy. This ensures that all subsequent Snapshot copies created on the primary volume have labels that are recognized by the SnapVault policy.

You can create a new Snapshot policy by using the **volume snapshot policy add-schedule**, or you can modify an existing Snapshot policy by using the **volume snapshot policy modify-schedule** command to set the **snapmirror-label** attribute for the set of Snapshot copies that you want replicated to the SnapVault secondary volume. Other Snapshot copies on the primary volume are ignored by the SnapVault relationship.

- Your work environment must be able to accommodate the time it might take to transfer a baseline Snapshot copy with a large amount of data.

Step

On the destination SVM, establish the relationship by using the **snapmirror resync** command and the **-type XDP** parameter.

If the most recent common Snapshot copy between the primary and the secondary is deleted from the primary but there exists another, older common Snapshot copy, you can also use the **-preserve** option. This option performs a logical local rollback to make the data in the primary and the secondary the same, and then it replicates all newer Snapshot copies from the source that match the SnapVault policy.

Example 14-26 creates a SnapVault relationship between the primary volume `srcvolA` on SVM `vserverA` and the prepopulated secondary volume `dstvolB` on SVM `vserverB`.

Example 14-26 Creating a SnapVault relationship between primary volume and prepopulated secondary volume

```
vserverB::> snapmirror resync -source-path vserverA:srcvolA -destination-path
vserverB:dstvolB -type XDP
```

14.13.6 Creating a destination baseline using a tape backup

You can perform a baseline transfer from local tape copies to a SnapVault secondary volume to manage your bandwidth or timing constraints over a network.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster.
- You must have SVM administrator privileges to perform this task for an SVM.
- The destination volume must not contain data.

About this task

This operation physically copies data from tape to one or more secondary volumes. When the operation finishes, the secondary volume contains all the Snapshot copies that existed on the primary volume at the time the tape copy was created.

Steps

Follow these steps:

1. Create a copy of the primary volume on the tape by using the **system smtape backup** command.

2. Restore the data to the empty secondary volume from the tape copy.
3. Initialize the SnapVault relationship by using the **`snapmirror resync`** command with the **`-typeXDP`** parameter on the secondary volume, and enable incremental updates.

14.13.7 Converting a data protection destination to a SnapVault secondary

You convert a data protection destination volume to a SnapVault secondary volume after a tape seeding operation or after you lose a SnapVault secondary volume in a backup to disaster protection mirror cascade.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster.
- You must have SVM administrator privileges to perform this task for an SVM.

About this task

In the case of tape seeding, after you transfer the data from the tape to the volume, the volume is a data protection destination volume.

In the case of a SnapVault secondary volume to disaster protection volume cascade, if the SnapVault secondary volume is lost, you can resume SnapVault protection by creating a direct relationship between the SnapVault primary volume and the disaster protection destination volume. You must make the disaster protection destination volume a SnapVault secondary volume to do this.

Steps

Follow these steps:

1. Break the data protection mirror relationship by using the **`snapmirror break`** command.

   The relationship is broken and the disaster protection volume becomes a read-write volume.

2. Delete the existing data protection mirror relationship, if one exists, by using the **`snapmirror delete`** command.
3. Remove the relationship information from the source SVM by using the **`snapmirror release`** command.

   This also deletes the Data ONTAP created Snapshot copies from the source volume.

4. Create a SnapVault relationship between the primary volume and the read-write volume by using the **`snapmirror create`** command with the **`-type XDP`** parameter.
5. Convert the destination volume from a read-write volume to a SnapVault volume and establish the SnapVault relationship by using the **`snapmirror resync`** command.

14.14 Managing backup operations for SnapVault backups

You configure SnapVault relationships on FlexVol volumes to establish SnapVault backups. You manage SnapVault relationships to optimize the performance of the relationships.

14.14.1 How an out-of-order Snapshot copy transfer works

The transfer of a Snapshot copy that does not conform to the usual sequence scheduled by a SnapVault policy is an out-of-order Snapshot copy transfer.

In SnapVault relationships, Snapshot copies are selected and transferred from the primary volume to the secondary volume, according to the configured SnapVault policy. Only Snapshot copies that are newer than the common Snapshot copy between the primary and secondary volume are transferred.

However, you can use the **`snapmirror update`** command to initiate the transfer of a Snapshot copy that was not originally selected and transferred.

When you initiate an out-of-order transfer, an older Snapshot copy is used to establish the base. To avoid subsequent transfers of Snapshot copies that already exist on the SnapVault secondary volume, the list of Snapshot copies that are selected for transfer in this update cycle are reconciled against the Snapshot copies that are already present on the secondary volume. Snapshot copies that are already present on the secondary volume are discarded from the transfer list.

Establishing new base from an out-of-order Snapshot copy transfer

In this example, the SnapVault policy has a schedule in which only the even-numbered Snapshot copies on the primary volume are transferred to the secondary volume. Before the out-of-order transfer begins, the primary volume contains Snapshot copies 2 through 6; the secondary volume contains only the even-numbered Snapshot copies (noted as “SC” in the figures). Snapshot copy 4 is the common Snapshot copy that is used to establish the base, as shown in Figure 14-18.

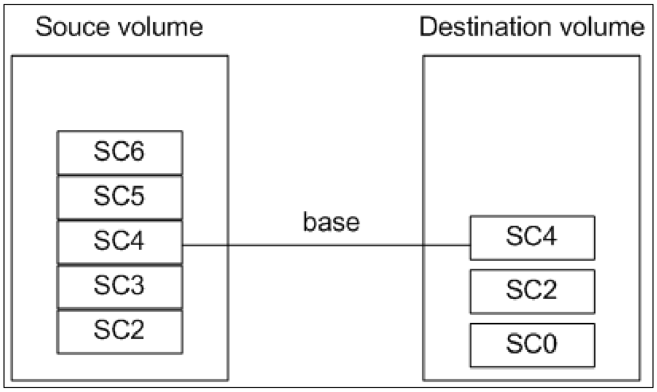


Figure 14-18 Common Snapshot copy used to establish the base

After Snapshot copy 3 is transferred to the secondary volume, out of order, it becomes the new common Snapshot copy that is used to establish the base, as shown in Figure 14-19.

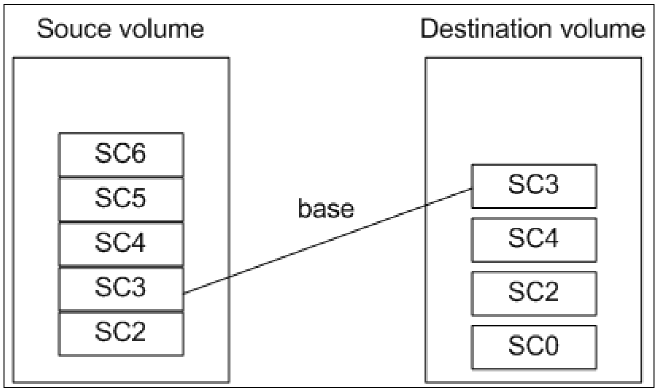


Figure 14-19 New common Snapshot copy after Snapshot copy 3 is transferred

> **Note:** Although Snapshot copy 3 is now the base, the exported Snapshot copy is still Snapshot copy 4.

When Snapshot copies are selected for subsequent updates according to the SnapVault policy, the policy selects Snapshot copy 4 and Snapshot copy 6 for transfer to the secondary volume.

When the transfer list is reconciled, Snapshot copy 4 is removed from the transfer list because it already exists on the secondary volume. Only Snapshot copy 6 is transferred, which becomes the new common Snapshot copy that is used to establish the base, as shown in Figure 14-20.

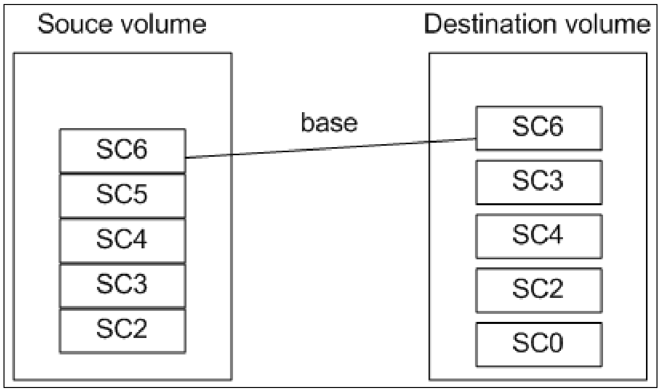


Figure 14-20 Snapshot copy 6 is transferred and becomes the new common Snapshot copy

14.14.2 Backing up from a Snapshot copy that is older than base Snapshot copy

You might want to replicate a special, manually initiated Snapshot copy to the SnapVault backup. The Snapshot copy is one that is not in the sequence scheduled by the SnapVault policy assigned to the SnapVault relationship.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster.
- You must have SVM administrator privileges to perform this task for an SVM.

Step

Begin the backup transfer of the older Snapshot copy by using the **snapmirror update** command.

Example 14-27 starts an out-of-order transfer of Snapshot copy SC3 from the source volume srcvolA on SVM vserverA and the secondary volume dstvolB on SVM vserverB.

Example 14-27 Backup transfer of the older Snapshot copy

```
vserverA::> snapmirror update -source-path vserverA:srcvolA -destination-path
vserverB:dstvolB -snapshot SC3
```

Result

After the backup finishes, the transferred Snapshot copy becomes the base.

14.14.3 Backing up FlexVol volumes with the maximum limit of Snapshot copies

To work around the limit of 251 Snapshot copies per volume, you can create a new destination volume clone, then establish a SnapVault relationship with the new clone.

Before you begin

You must have cluster administrator privileges to perform this task for a cluster. You must have SVM administrator privileges to perform this task for an SVM.

About this task

Creating a new SnapVault relationship to a new volume clone enables you to continue SnapVault protection with minimum disruption on the clone volume and without starting a new baseline transfer. Because the source clone and the volume clone share the latest common Snapshot copy, subsequent updates are performed as usual, according to the policy assigned to the SnapVault relationship.

Steps

Follow these steps:

1. Quiesce the SnapVault relationship between the primary volume and the secondary volume by using the **snapmirror quiesce** command.

   This step prevents updates from starting until after the task is complete.

2. Verify that there are no active transfers on the relationship by using the **snapmirror show** command.

   The Relationships field should be Idle.

3. Create a volume clone based on the most recent common Snapshot copy between the SnapVault primary volume and the SnapVault secondary volume by using the **volume clone create** command with the **-type DP** parameter.

4. Establish the SnapVault relationship between the primary volume and the newly created secondary volume clone by using the **snapmirror resync** command and the **-type XDP** parameter.

5. Delete the SnapVault relationship between the primary volume and the original SnapVault secondary volume by using the **snapmirror delete** command.

14.15 Managing restore operations for SnapVault backups

The restore operation from a SnapVault backup copies a single, specified Snapshot copy from a SnapVault secondary volume to a specified volume. Restoring a volume from a SnapVault secondary volume changes the view of the active file system but preserves all earlier Snapshot copies in the SnapVault backup.

> **Note:** You cannot make the SnapVault destination volume readable and writable. If you need disaster recovery capabilities, use volume SnapMirror. However, it is possible to create a FlexClone volume of the Snapshot copies on the SnapVault destination to gain read and write access to the data.

14.15.1 Guidelines for restoring the active file system

Before restoring a volume, you must shut down any application that accesses data in a volume to which a restore is writing data. Therefore, you must dismount the file system, shut down any database, and deactivate and quiesce the Logical Volume Manager (LVM) if you are using an LVM.

The restore operation is disruptive. When the restore operation finishes, the cluster administrator or SVM administrator must remount the volume and restart all applications that use the volume.

The restore destination volume must not be the destination of another mirror or the secondary of another SnapVault relationship.

You can restore to the following volumes:

- Original source volume

  You can restore from a SnapVault secondary volume back to the original SnapVault primary volume.

- New, empty secondary volume

  You can restore from a SnapVault secondary volume to a new, empty secondary volume. You must first create the volume as a data protection (DP) volume.

- New secondary that already contains data

  You can restore from a SnapVault secondary volume to a volume that is prepopulated with data.

  The volume must have a Snapshot copy in common with the restore primary volume and must not be a DP volume.

14.15.2 Guidelines for restoring LUNs in SAN environments

The restore operation from a SnapVault backup copies a single, specified LUN from a SnapVault secondary volume to a specified volume. Restoring a LUN from a SnapVault secondary volume changes the view of the active system on the volume to which data is being restored, preserving all earlier Snapshot copies.

The following guidelines apply only to SAN environments:

- You can restore a single file or single LUN from a SnapVault secondary volume by using the IBM OnCommand management software online management tools.
- When LUNs are restored to existing LUNs, new access controls do not need to be configured. You must configure new access controls for the restored LUNs only when restoring LUNs as newly created LUNs on the volume.
- If LUNs on the SnapVault secondary volume are online and mapped before the restore operation begins, they remain so for the duration of the restore operation and after the operation finishes.
- The host system can discover the LUNs and issue non-media access commands for the LUNs, such as inquiries or commands to set persistent reservations, while the restore operation is in progress.
- You cannot create new LUNs in a volume during a restore operation with the **`lun create`** command.
- Restore operations from tape and from a SnapVault backup are identical.
- You cannot restore a single LUN from a SnapVault secondary volume that is located on a system that is running in 7-Mode.

14.15.3 How restore operations work from a SnapVault backup

A restore operation from a SnapVault backup consists of a series of actions performed on a temporary restore relationship and on the secondary volume.

During a restore operation, the following actions occur:

1. A new temporary relationship is created from the restore source (which is the original SnapVault relationship secondary volume) to the restore destination.

   The temporary relationship is a restore type (RST). The **`snapmirror show`** command displays the RST type while the restore operation is in progress.

   The restore destination might be the original SnapVault primary or might be a new SnapVault secondary.
2. During the restore process, the restore destination volume is changed to read-only.
3. When the restore operation finishes, the temporary relationship is removed and the restore destination volume is changed to read-write.

14.15.4 Restoring a volume from a SnapVault backup

If the data on a volume becomes unavailable, you can restore the volume to a specific time by copying a Snapshot copy in the SnapVault backup. You can restore data to the same primary volume or to a new location. This is a disruptive operation.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster.
- You must have SVM administrator privileges to perform this task for an SVM.
- CIFS traffic must not be running on the SnapVault primary volume when a restore operation is running.

About this task

Example 14-28 shows how to restore a whole volume from a SnapVault backup. To restore a single file or LUN, you can restore the whole volume to a different, non-primary volume, and then select the file or LUN, or you can use the IBM OnCommand management software online management tools.

Steps

Follow these steps:

1. If the volume to which you are restoring has compression enabled and the secondary volume from which you are restoring does not have compression enabled, disable compression.

   You disable compression to retain storage efficiency during the restore.

2. Restore a volume by using the **snapmirror restore** command as in Example 14-28.

   Example 14-28 Restoring a volume from a SnapVault backup

   ```
   vs1::> snapmirror restore -destination-path vs1:vol1 -source-path
   vs2:vol1_dp_mirror2 -source-snapshot snap3
   Warning: All data newer than Snapshot copy snap6 on volume vs1:vol1 will be
   deleted.
   Do you want to continue? {y|n}: y
   [Job 34] Job is queued: snapmirror restore from source vs2:vol1_dp_mirror2 for
   the snapshot snap3.
   ```

3. Remount the restored volume and restart all applications that use the volume.
4. If you previously disabled compression, reenable compression on the volume.

14.15.5 Managing the SnapVault-mirror cascade when the SnapVault backup is unavailable

You can manipulate relationships in a SnapVault-mirror cascade to maintain data backup relationships if the secondary of the SnapVault relationship becomes unavailable.

Before you begin

You must have a SnapVault-mirror cascade already configured.

About this task

The destination of the SnapVault relationship is the middle of the SnapVault-mirror cascade. If it becomes unavailable, you might have the following issues:

- You cannot update the SnapVault backup.
- You cannot update the mirror copy of the SnapVault secondary.

To manage this issue, you can temporarily remove the SnapVault secondary volume from the cascade and establish a SnapVault relationship to the mirror copy of the SnapVault secondary volume. When the unavailable secondary volume becomes available, you can reestablish the original cascade configuration.

In the following steps, the primary volume of the cascade is called “A”, the secondary volume of the SnapVault relationship is called “B”, and the destination volume of the data protection mirror relationship is called “C”.

Steps

Follow these steps:

1. Identify the current exported Snapshot copy on C by using the **volume snapshot show** command with the **-fields busy** parameter.

   The busy field is set to **true** for the exported Snapshot copy:

   ```
   volume snapshot show C -fields busy
   ```

2. Break the data protection mirror relationship by **snapmirror break** command on C:

   ```
   snapmirror break C
   ```

3. Create a dummy **snapmirror-label** on the exported Snapshot copy you previously identified by using the **volume snapshot modify** command with the **-snapmirror-label** parameter.

   If a **snapmirror-label** already exists for the exported Snapshot copy, you do not need to perform this step:

   ```
   volume snapshot modify -volume C -snapshot name -snapmirror-label exp1
   ```

4. Create a Snapshot owner on the exported Snapshot copy of C by using the **snapmirror snapshot-owner create** command.

   This prevents Clustered Data ONTAP from deleting the Snapshot copy:

   ```
   snapmirror snapshot-owner create -volume C -snapshot exported -owner admin1
   ```

5. Delete the data protection mirror relationship between B and C by using the **snapmirror delete** command:

   ```
   snapmirror delete C
   ```

6. Create the SnapVault relationship between A and C by using the **snapmirror resync** command and the **-type XDP** parameter:

   ```
   snapmirror resync -source-path A -destination-path C -type XDP
   ```

   You can maintain this SnapVault relationship until you recover the original SnapVault secondary volume. At that time, you can reestablish the original cascade relationship by using the steps that follow this step.

7. Delete the data protection mirror relationship between A and B by using the **snapmirror delete** command.

8. Perform a disaster recovery resynchronization from C to B by using the **snapmirror resync** command:

   ```
   snapmirror resync –source-path C –destination-path B
   ```

   This step copies from C to B, all of the Snapshot copies made after B became unavailable.

9. Identify the current exported Snapshot copy on B by using the **volume snapshot show** command with the **-fields busy** parameter:

   ```
   volume snapshot show B -fields busy
   ```

   The **busy** field is set to **true** for the exported Snapshot copy.

10. Break the data protection mirror by using the **snapmirror break** command on B:

    ```
    snapmirror break B
    ```

11. Create a dummy **snapmirror-label** on the exported Snapshot copy you previously identified by using the **volume snapshot modify** command with the **-snapmirror-label** parameter:

    ```
    volume snapshot modify -volume B -snapshot name -snapmirror-label exp2
    ```

If a `snapmirror-label` already exists for the exported Snapshot copy, you do not need to perform this step.

12. Create a Snapshot owner on the exported Snapshot copy of B by using the `snapmirror snapshot-owner create` command.

    This prevents Clustered Data ONTAP from deleting the Snapshot copy:

    ```
    snapmirror snapshot-owner create -volume B -snapshot exported -owner admin1
    ```

13. Delete the data protection mirror relationship between C and B by using the `snapmirror delete` command.
14. Perform a SnapVault resynchronization from A to B by using the `snapmirror resync` command and the `-type XDP` parameter:

    ```
    snapmirror resync –source-path A –destination-path B –type XDP
    ```

    New Snapshot copies that meet the Snapshot policy of the SnapVault relationship are transferred from A to B.

15. Delete the data protection mirror relationship between A and C by using the `snapmirror delete` command.
16. Perform a disaster recovery resynchronization from B to C by using the `snapmirror resync` command.

    This step copies from B to C, all of the Snapshot copies made after reestablishing the A to B relationship without deleting any Snapshot copies on C:

    ```
    snapmirror resync –source-path B –destination-path C
    ```

17. Remove the Snapshot copy owner from volumes B and C by using the `snapmirror snapshot-owner delete` command:

    ```
    snapmirror snapshot-owner delete -volume B -snapshot exported_snap
    ```

18. Remove SnapMirror labels that you created from volumes B and C by using the snapshot modify command:

    ```
    snapshot modify -volume B -snapshot exported_snap -snapmirror-label text
    snapshot modify -volume C -snapshot exported_snap -snapmirror-label text
    ```

14.16 Managing storage efficiency for SnapVault secondaries

SnapVault relationships preserve storage efficiency when backing up data from the primary volume to the secondary volume, with one exception: If post-process and optionally inline compression are enabled on the secondary volume, storage efficiency is not preserved for data transfers between the primary and secondary volumes.

14.16.1 Guidelines for managing storage efficiency for SnapVault backups

If both the primary and secondary volumes in a SnapVault relationship have storage efficiency enabled, then data transfers to the SnapVault secondary volume preserve storage efficiency. If the primary volume does not have storage efficiency enabled, you might want to enable storage efficiency only on the secondary volume.

Because SnapVault secondary volumes typically contain a large amount of data, storage efficiency on SnapVault secondary volumes can be very important.

Notes:

- If the primary volume in a SnapVault relationship is enabled for storage efficiency, all data backup operations preserve the storage efficiency.
- If the primary volume in a SnapVault relationship does not have storage efficiency enabled, you might want to enable storage efficiency for the secondary volume because it is likely to contain a large amount of data over time.

You can use the **`volume efficiency`** command to start a scan on the volume if there is already data on the volume from transfers. If this is a new relationship with no transfers, then there is no need to run the scan manually.

Changes to the volume's efficiency schedule do not take effect for a SnapVault secondary volume. Instead, when storage efficiency is enabled, the SnapVault relationship manages the schedule. When a data transfer begins, the storage efficiency process automatically pauses until the transfer is finished, and then automatically begins again after the data transfer is complete. Because data transfers to a SnapVault secondary volume might include more than one Snapshot copy, the storage efficiency process is paused for the entire duration of the update operation. After the transfer is finished and the post-transfer storage efficiency process is complete, the last Snapshot copy created in the secondary volume is replaced by a new, storage-efficient Snapshot copy.

If the last Snapshot copy that is created in the secondary volume is locked before it can be replaced by a new, storage-efficient Snapshot copy, then a new, storage-efficient Snapshot copy is still created, but the locked Snapshot copy is not deleted. That Snapshot copy is deleted later during the storage-efficient cleanup process after a subsequent update to the SnapVault secondary volume and after the lock is released. A Snapshot copy in a SnapVault secondary volume might be locked because the volume is the source in another relationship, such as a data protection mirror relationship.

Note: If the secondary volume has additional compression enabled, storage efficiency is not preserved.

Storage efficiency on all data transfers in SnapVault relationships is not preserved when the secondary volume has additional compression enabled. Because of the loss of storage efficiency, a warning message is displayed when you enable compression on a SnapVault secondary volume. After you enable compression on the secondary volume, you can never have storage-efficient transfers.

14.16.2 Enabling storage efficiency on a SnapVault secondary volume

If the primary volume does not have storage efficiency enabled, you can enable storage efficiency on a SnapVault secondary volume by enabling storage efficiency on the volume.

Before you begin

Observe these prerequisites:

- You must have cluster administrator privileges to perform this task for a cluster.
- You must have SVM administrator privileges to perform this task for an SVM.

Steps

Follow these steps:

1. Use the **`volume efficiency`** command with the **`-on`** parameter to enable storage efficiency.
2. If the volume already has data which you want to make storage efficient, use the **`volume efficiency`** command with the **`-start`** and **`-scan-old-data`** parameters to start a scan of the volume.

15

Disaster recovery

In this chapter, we discuss disaster recovery, more specifically SnapMirror, and all of its features.

Although the load sharing SnapMirror, being a disaster recovery feature, might not appear to fit in this chapter, we have included it because it uses the same technology as the data protection SnapMirror.

The following topics are covered:

- SnapMirror overview
- SnapMirror Data Protection (SnapMirror DP)
- SnapMirror Load Sharing (SnapMirror LS)

15.1 SnapMirror overview

SnapMirror in Clustered Data ONTAP provides asynchronous volume-level replication based on a configured replication update interval. SnapMirror uses Snapshot technology as part of the replication process.

Clustered Data ONTAP 8.1 onward provides the following replication capabilities:

- Data protection mirrors. Replication to create a backup copy within the same cluster (intracluster) or to create a DR copy in a different cluster (intercluster).
- Load-sharing mirrors. Replication from one volume to multiple volumes in the same cluster to distribute a read-only workload across a cluster.

Snapmirror replication basically consists of the folowing steps, triggered either by the scheduler or on demand by the user:

1. A new Snapshot copy is created on the source volume.
2. The block-level difference between the new Snapshot copy and the last replication Snapshot copy is determined and then transferred to the destination volume. This transfer includes other Snapshot copies that were created between the last replication Snapshot copy and the new one.
3. When the transfer is complete, the new Snapshot copy exists on the destination volume.

A SnapMirror destination volume is available for read-only access if it is shared using Common Internet File System (CIFS) protocol, exported using Network File System (NFS) protocol. A logical unit number (LUN) in the replicated volume can be made available to a client that supports connection to read-only LUNs.

Replication occurs at the volume level. Qtrees can be created in Clustered Data ONTAP and replicated along with the replicated volume; however, individual qtrees cannot be separately replicated.

DP relationships can be resynchronized in either direction after a failover without recopying the entire volume. If a relationship is resynchronized in the reverse direction, only new data written since the last successful synchronization snapshot will be sent back to the destination.

SnapMirror relationships in Clustered Data ONTAP 8.1 must be managed by a cluster administrator; administration cannot be delegated to a storage virtual machine administrator. Starting with Clustered Data ONTAP 8.2, a cluster administrator can delegate the management of SnapMirror relationships to a storage virtual machine administrator.

15.2 SnapMirror Data Protection (SnapMirror DP)

Data protection mirrors can be performed as intercluster or intracluster:

- Intercluster DP mirrors: Replication between volumes in two different storage virtual machines in different clusters operating in Clustered Data ONTAP. They are primarily used for providing DR to another site or location.
- Intracluster DP mirrors: Replication between two volumes in different storage virtual machines in the same cluster, or between two volumes in the same storage virtual machine. They are primarily used for maintaining a local backup copy.

DP mirror relationships have the same characteristics regardless of whether intracluster or intercluster is being replicated. These characteristics are included:

- DP mirror relationships are created and managed on the destination cluster.
- DP mirror relationship transfers are triggered by the scheduler in the destination cluster.
- Each DP mirror destination volume is a separate SnapMirror relationship that is performed independently of other DP mirror volumes; however, the same Clustered Data ONTAP schedule entry can be used for different DP mirror relationships.
- Destination volumes for both DP- and LS-type mirrors must be created with a volume type (-type option) of DP. The storage administrator cannot change the volume -type property after the volume has been created.
- DP mirror destination volumes are read-only until failover.
- DP mirror destination volumes can be failed over using the SnapMirror break operation, making the destination volume writable. The SnapMirror break must be performed separately for each volume.
- DP mirror destination volumes can be mounted into a storage virtual machine namespace while still read-only, but only after the initial transfer is complete.
- An intercluster DP mirror destination volume cannot be mounted in the same namespace as the source volume, because intercluster DP mirror relationships are to a different cluster and therefore to a different storage virtual machine, which is a different namespace.
- An intracluster DP mirror destination volume can be mounted in the same namespace as the source volume if both the source and destination volumes exist in the same storage virtual machine; however, they cannot be mounted to the same mount point.
- LUNs contained in DP mirror destination volumes can be mapped to igroups and connected to clients; however, the client must be able to support connection to a read-only LUN.
- DP mirror relationships can be managed using the Clustered Data ONTAP command line interface (CLI), IBM N series OnCommand System Manager 3.0, and IBM N series OnCommand Unified Manager 6.0.
- If an in-progress transfer is interrupted by a network outage or aborted by an administrator, a subsequent restart of that transfer can automatically continue from a saved restart checkpoint.

Clustered Data ONTAP 8.2 onward provides an additional SnapMirror relationship, XDP vault. For more information, see the SnapVault section in the data protection section, 14.8, “SnapVault” on page 229.

Intercluster and intracluster DP SnapMirror relationships are different based on the network that is used for sending data. Intercluster DP SnapMirror relationships use the intercluster network defined by intercluster LIFs. Figure 15-1 illustrates an intercluster network for SnapMirror.

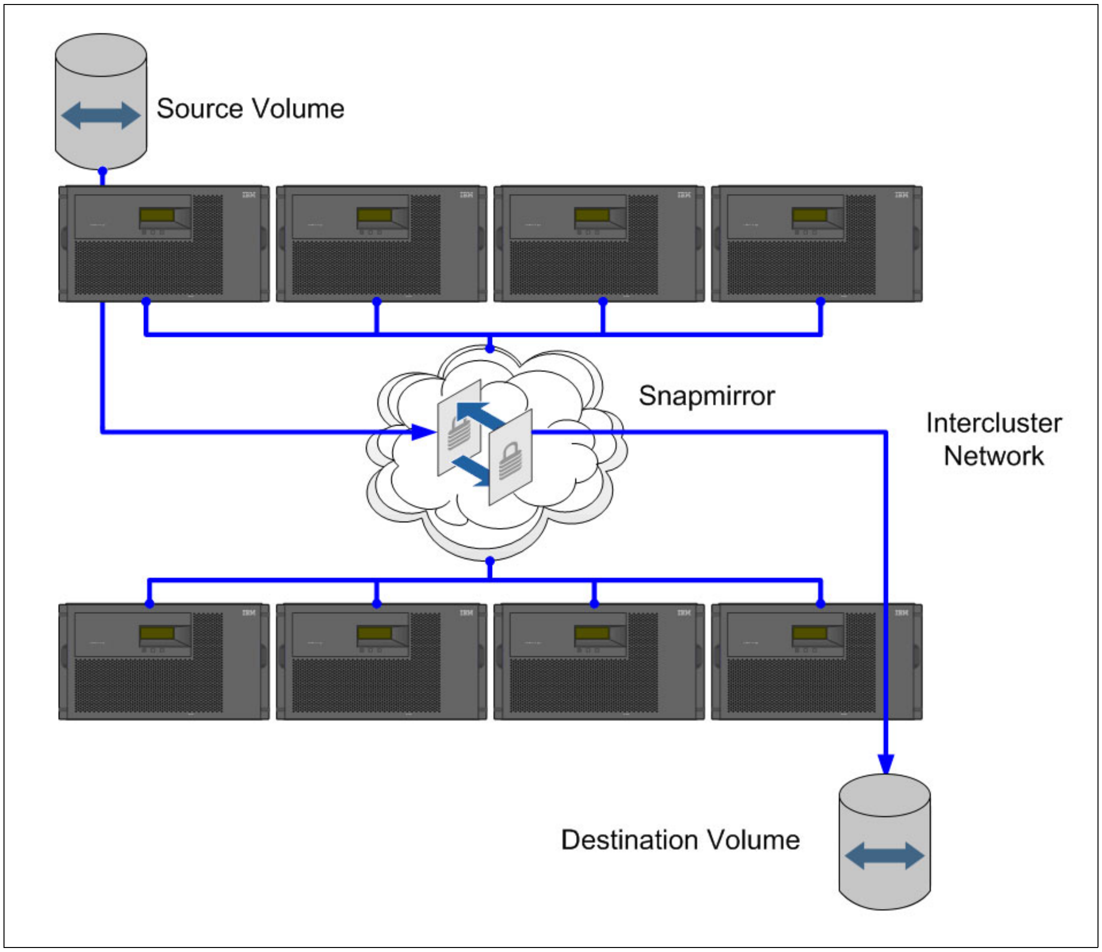


Figure 15-1 Intercluster SnapMirror

Intracluster DP mirror relationships use the cluster interconnect, which is the private connection used for communication between nodes in the same cluster. Figure 15-2 illustrates a cluster interconnect for intercluster SnapMirror.

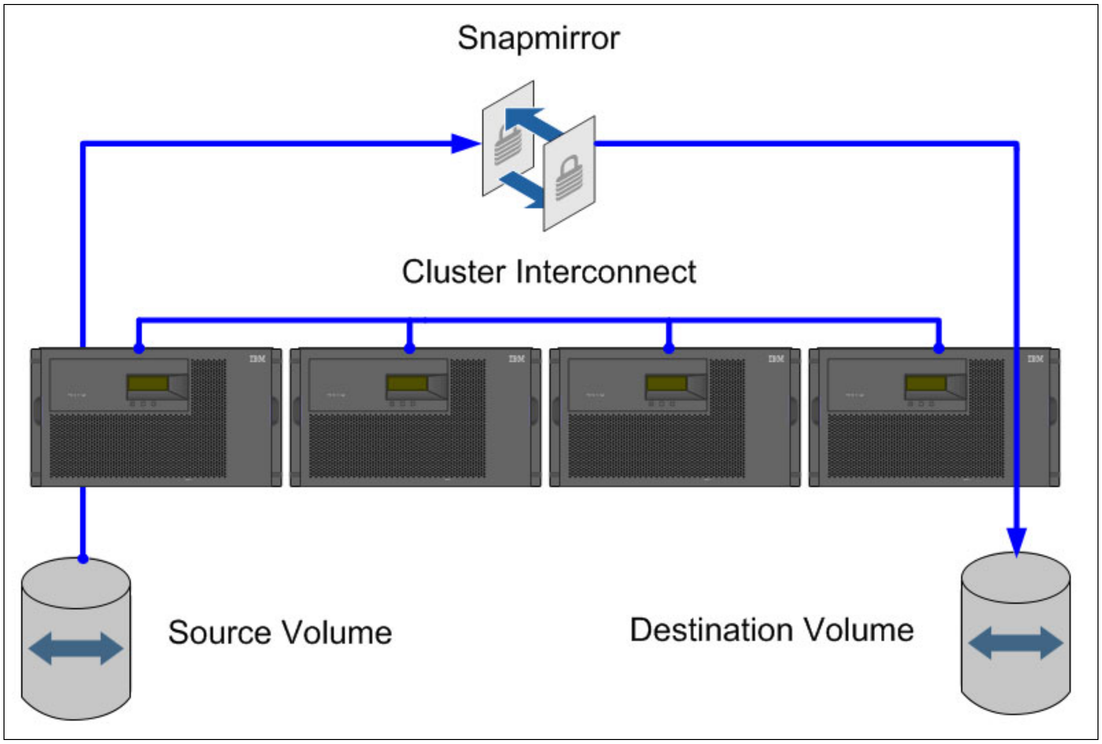


Figure 15-2 Intracluster SnapMirror

15.2.1 SnapMirror data protection relationships

After the cluster peer relationship and storage virtual machine peer relationship have been successfully created between the two clusters, create the intercluster SnapMirror relationships. A peer relationship is not required to mirror data between two storage virtual machines in the same cluster or between two volumes in the same storage virtual machine.

Both the source and destination storage virtual machines must have the same language type setting to be able to replicate between them. A storage virtual machine language type cannot be changed after it has been created.

Intercluster SnapMirror relationships are primarily used to provide DR capability in another site or location. If all necessary volumes have been replicated to a DR site with SnapMirror, then a recovery can be performed so that operations can be restored from the DR site.

The creation of SnapMirror relationships in Clustered Data ONTAP does not depend on storage virtual machine hostname to IP address resolution. While the cluster names are resolved through the peer relationship, the storage virtual machine names are internally resolved through the clusters. The host names of the source and destination storage virtual machine and cluster are used to create SnapMirror relationships in Clustered Data ONTAP; it is not necessary to use the IP address of an LIF.

Intercluster SnapMirror requirements

Complete the following requirements before creating an intercluster SnapMirror relationship:

- Configure the source and destination nodes for intercluster networking.
- Configure the source and destination clusters in a peer relationship.
- Create a destination storage virtual machine that has the same language type as the source storage virtual machine; volumes cannot exist in Clustered Data ONTAP without a storage virtual machine.
- Configure the source and destination storage virtual machine in a peer relationship.
- Create a destination volume with a type of DP, with a size equal to or greater than that of the source volume.
- Assign a schedule to the SnapMirror relationship in the destination cluster to perform periodic updates. If any of the existing schedules are not adequate, a new schedule entry must be created.

Storage virtual machine fan-out and fan-in

It is possible to fan-out or fan-in volumes between different storage virtual machines. For example, multiple different volumes from a single storage virtual machine in the source cluster might be replicated with each volume replicating into a different storage virtual machine in the destination cluster, referred to as fan-out. Alternatively, multiple different volumes might also be replicated, each existing in a different storage virtual machine in the source cluster, to a single storage virtual machine in the destination cluster, referred to as fan-in.

Tip: When replicating to provide DR capabilities, mirror all required volumes from a given storage virtual machine in the source cluster to a particular matching storage virtual machine in the destination cluster. Design considerations that determine that a given set of volumes should reside in the same storage virtual machine should also apply to keeping those same volumes in a like storage virtual machine at a DR site. In order for different volumes to be accessible in the same namespace, they must exist in the same storage virtual machine (a storage virtual machine is a namespace).

Volume fan-out and fan-in

For SnapMirror DP relationships, a single FlexVol volume can be replicated to up to four different destination volumes. Each destination volume can exist in a different storage virtual machine or all can exist in the same storage virtual machine; this is referred to as volume fan-out. Volume fan-in, which is replication of multiple different volumes into the same destination volume, is not possible.

Cascade relationships or multihop replication

Starting in Clustered Data ONTAP 8.2, SnapMirror relationships can be cascaded. However, only one of the relationships in the cascade configuration can be a SnapVault relationship.

Cascading is defined as replicating from established replicas. Suppose there are three storage systems, A, B, and C. Replicating from A to B and from B to C is considered a cascade configuration.

An example cascade configuration with two hops is shown in Figure 15-3.

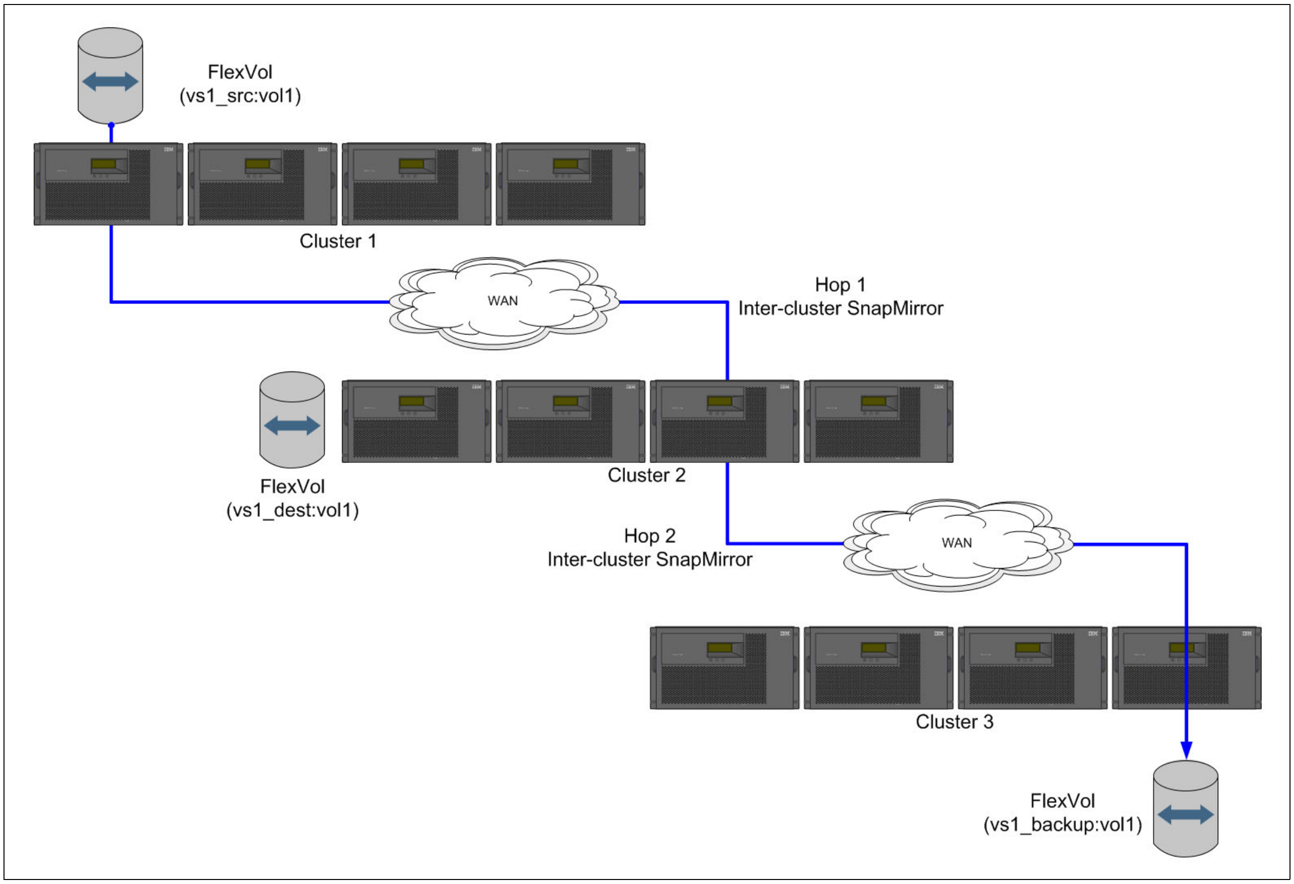


Figure 15-3 Cascaded volume replication using SnapMirror

Dual-hop volume SnapMirror

This configuration involves volume SnapMirror replication among three clusters.

```
vs1_src:vol1 → vs1_dest:vol1 → vs1_backup:vol1
```

> **Note:** In the preceding configuration, vs1_src:vol1 to vs1_dest:vol1 and vs1_dest:vol1 to vs1_backup:vol1 transfers can occur at the same time.

Table 15-1 Snapshot copy propagation for dual-hop volume SnapMirror

Timeline	Snapshot copies on Cluster 1	Snapshot copies on Cluster 2	Snapshot copies on Cluster 3
1. After volume initialization on Cluster 2	hourly.2013-10-28_1050 snapmirror.c72d52b9v1	hourly.2013-10-28_1050 snapmirror.c72d52b9v1	
2. Volume SnapMirror update on Cluster 2	hourly.2013-10-28_1050 snapmirror.c72d52b9v2	hourly.2013-10-28_1050 snapmirror.c72d52b9v1 snapmirror.c72d52b9v2	
3. After volume initialization on Cluster 3	hourly.2013-10-28_1050 snapmirror.c72d52b9v2	hourly.2013-10-28_1050 snapmirror.c72d52b9v1 snapmirror.c72d52b9v2	hourly.2013-10-28_1050 snapmirror.c72d52b9v1 snapmirror.c72d52b9v2

Timeline	Snapshot copies on Cluster 1	Snapshot copies on Cluster 2	Snapshot copies on Cluster 3
4. Volume SnapMirror update on Cluster 2	hourly.2013-10-28_1050 snapmirror.c72d52b9v2 snapmirror.c72d52b9v3	hourly.2013-10-28_1050 snapmirror.c72d52b9v2 snapmirror.c72d52b9v3	hourly.2013-10-28_1050 snapmirror.c72d52b9v1 snapmirror.c72d52b9v2
5. Volume SnapMirror update on Cluster 3	hourly.2013-10-28_1050 snapmirror.c72d52b9v2 snapmirror.c72d52b9v3	hourly.2013-10-28_1050 snapmirror.c72d52b9v2 snapmirror.c72d52b9v3	hourly.2013-10-28_1050 snapmirror.c72d52b9v2 snapmirror.c72d52b9v3

15.2.2 Scheduling SnapMirror updates

Clustered Data ONTAP has a built-in scheduling engine similar to cron. Periodic replication updates in Clustered Data ONTAP can be scheduled by assigning a schedule to a SnapMirror relationship in the destination cluster. Create a schedule through the command line using the **`job schedule cron create`** command. Example 15-1 demonstrates the creation of a schedule called Hourly_SnapMirror that runs at the top of every hour (on the zero minute of every hour).

Example 15-1 SnapMirror schedule creation

```
cdot-cluster01::> job schedule cron create Hourly_SnapMirror -minute 0
cdot-cluster01::> job schedule cron show
Name                Description
----------------    ---------------------------------------------------
5min                @:00,:05,:10,:15,:20,:25,:30,:35,:40,:45,:50,:55
8hour               @2:15,10:15,18:15
Hourly_SnapMirror   @:00
avUpdateSchedule    @2:00
daily               @0:10
hourly              @:05
weekly              Sun@0:15
```

The schedule can then be applied to a SnapMirror relationship at the time of creation using the **`-schedule`** option or to an existing relationship using the **`snapmirror modify`** command and the **`-schedule`** option. In Example 15-2, the **`Hourly_SnapMirror`** schedule is applied to an existing relationship.

Example 15-2 Alter SnapMirror schedule

```
cdot-cluster01::> snapmirror modify -destination-path cluster02://vs1/vol1
–schedule Hourly_SnapMirror
```

Schedules can also be managed and applied to SnapMirror relationships using OnCommand System Manager 3.0.

15.2.3 Converting a SnapMirror relationship to a SnapVault relationship

Here is one scenario where you would want to convert an existing SnapMirror relationship to a SnapVault relationship. An existing customer using SnapMirror in Clustered Data ONTAP 8.1 wants to make use of SnapVault in Clustered Data ONTAP 8.2 for longer retention.

Upgrade your source and destination clusters to Clustered Data ONTAP 8.2. Your existing SnapMirror relationships will continue to remain cluster scope and will behave as they did in Clustered Data ONTAP 8.1. They will not benefit from the scalability improvements unless they are deleted and recreated. However, both Clustered Data ONTAP 8.1 and Clustered Data ONTAP 8.2 use the block-level engine for mirrors, and it is important to note that no new baseline will be required, only resync.

Converting the relationship from SnapMirror to SnapVault would consist of the following steps:

1. Delete the mirror (DR) relationship.
2. Break the mirror destination.
3. Create an XDP (vault) relationship between the same endpoints.
4. Perform resync between the endpoints. This will convert a DR destination to a vault destination without having to do a new baseline.

15.3 SnapMirror Load Sharing (SnapMirror LS)

SnapMirror LS mirrors increase performance and availability for NAS clients by distributing a storage virtual machine namespace root volume to other nodes in the same cluster and distributing data volumes to other nodes in the cluster to improve performance for large read-only workloads.

Note: SnapMirror LS mirrors are capable of supporting NAS only (CIFS/NFSv3). LS mirrors do not support NFSv4 clients or SAN client protocol connections (FC, FCoE, or iSCSI).

15.3.1 Administering load-sharing mirrors

LS mirror relationships can only be managed by the Data ONTAP. Currently, LS mirror relationships cannot be managed using System Manager.

One way in which LS mirror relationships differ from DP relationships is that additional commands are provided to manage the LS mirror's **`snapmirror initialize-ls-set`**, **`update-ls-set`**, and **`promote`** commands. A group of LS mirror destination volumes that replicate from the same source volume is referred to as an LS mirror set.

When an LS mirror set is created, each destination volume must be created in the appropriate aggregate, creating the destination volumes with a type of DP. In this example, two volumes named **`vs1_ls_a`** and **`vs1_ls_b`** are created as LS mirror destination volumes for the storage virtual machine root volume named **`vs1`**. See Example 15-3.

Example 15-3 Creation of the Load-sharing volumes

```
cdot-cluster01::> vol create -vserver vs1 -volume vs1_ls_a -aggregate aggr1 -size
20MB -type DP
cdot-cluster01::> vol create -vserver vs1 -volume vs1_ls_b -aggregate aggr1 -size
20MB -type DP
```

After all LS mirror destination volumes are created, each SnapMirror relationship can be created with a type of LS. In Example 15-4, an LS SnapMirror relationship is created for each of the destination volumes, **`vs1_ls_a`** and **`vs1_ls_b`**, with an hourly update schedule.

Example 15-4 Creation of the snapmirror relationships

```
cdot-cluster01::> snapmirror create -source-path vs1:vs1 -destination-path
vs1:vs1_ls_a -type LS
cdot-cluster01::> snapmirror create -source-path vs1:vs1 -destination-path
vs1:vs1_ls_b -type LS –schedule hourly
```

LS mirror relationships can be updated manually or by setting the desired schedule in the **`-schedule`** option. For LS mirror relationships, this is done by setting the desired schedule on any one of the destinations in the LS mirror set. Data ONTAP automatically applies that schedule to all destinations in that LS mirror set. A later change to the update schedule for any of the destination volumes in the LS mirror set applies the new schedule to all volumes in that LS mirror set. Therefore, in the previous example, the **`-schedule`** option was used only in the creation of the last relationship, which applied the schedule to both relationships.

All destination volumes can then be initialized for a particular LS mirror set in one operation using the **`snapmirror initialize-ls-set`** command, as shown in Example 15-5. When using this command, specify the source path to identify the LS mirror set instead of a destination path, because in an LS mirror set the source path is common to all relationships that are being initialized.

Example 15-5 Load-sharing snapmirror initialization and on demand update

```
cdot-cluster01::> snapmirror initialize-ls-set -source-path cluster01://vs1/vs1
cdot-cluster01::> snapmirror show
                                                                      Progress
Source         Destination   Mirror   Relationship  Total             Last
Path     Type  Path          State    Status        Progress Healthy  Updated
-------  ----  ------------  -------  ------------  -------- -------  --------
cluster01://vs1/vs1
         LS    cluster01://vs1/vs1_ls_a
                             Snapmirrored
                                      Transferring  -        false    -
               cluster01://vs1/vs1_ls_b
                             Snapmirrored
                                      Transferring  -        false    -
cdot-cluster01::> snapmirror update-ls-set -source-path vs1:vs1
```

LS mirror relationships can be updated on demand using the snapmirror update-ls-set command, as shown in the following example. When using this command, specify the source path to identify the LS mirror set instead of a destination path, because in an LS mirror set the source path is common to all relationships that are being updated. Data ONTAP updates all destination volumes for the LS set in one operation.

15.3.2 Accessing load-sharing mirror volumes

By default, all client requests for access to a volume in an LS mirror set are granted read-only access. Read-write access is granted by accessing a special administrative mount point, which is the path that servers requiring read-write access into the LS mirror set must mount. All other clients will have read-only access. After changes are made to the source volume, the changes must be replicated to the rest of the volumes in the LS mirror set using the **`snapmirror update-ls-set`** command, or with a scheduled update.

Volumes can be mounted inside other volumes, also referred to as a nested volume. When a new volume is mounted inside a volume that is configured in an LS mirror set, clients cannot see the new mount point until after the LS mirror set has been updated. This can be performed on demand using the **`snapmirror update-ls-set`** command or when the next scheduled update of the LS mirror is set to occur.

Access to the volume in read-write mode is granted to different CIFS and NFS clients.

To allow a CIFS client to connect to the source volume with read-write access, create a CIFS share for the admin mount point by adding **`/.admin`** to the volume path. In the following example, a CIFS share called **`report_data_rw`** is created that allows read-write access to a volume called **`report_data`**, which is part of an LS mirror set.

Use the path in Example 15-6 to access the read-write admin share of an LS mirror set using CIFS.

Example 15-6 Accessing the read-write admin share of an LS mirror set

```
cdot-cluster01::> vserver cifs share create -vserver vs1 -share-name
report_data_rw -path/.admin/report_data
```

Any CIFS client requiring read-write access must connect to the read-write path.

To connect to the source volume of an LS mirror set with read-write access, from an NFS client, mount the NFS export path and add /.admin to the volume path.

Any process or application running on the **`nfs_client`** system must use the path **`/client_rw_mountpoint`** for read-write access.

15.3.3 Load-sharing mirrors for storage virtual machine namespace root volumes

A namespace root volume is very small, containing only directories that are used as mount points, the paths where data volumes are junctioned (mounted) into the namespace. However, they are extremely important for NAS clients, which are not able to access data if the storage virtual machine root volume is unavailable.

> **Note:** SAN client connections (FC, FCoE, or iSCSI) do not depend on the storage virtual machine root volume.

New volumes can be mounted into a namespace root volume that is configured in an LS mirror set. After mounting the volume, clients cannot see the new mount point until after the LS mirror set has been updated. This can be performed on demand using the **`snapmirror update-ls-set`** command, or when the next scheduled update of the LS mirror is set to occur.

As previously mentioned, LS mirror volumes are read-only. When the LS mirror volume is a namespace root volume, that volume is read-only; however, data volumes mounted in the namespace are read-write or read-only, depending on the individual volume characteristics and permissions set on files and folders within those volumes.

Tip: Create an LS mirror of a NAS storage virtual machine namespace root volume on every node in the cluster so that the root of the namespace is available, regardless of node outages or node failovers.

When a client requests access to a volume configured with an LS mirror set, Data ONTAP directs all client connections only to the LS mirror destination volumes; therefore, a destination volume on the same node where the source volume resides should be created, allowing the namespace to provide a direct data access path to data volumes on that node.

15.3.4 Load-sharing mirrors for read-only workloads

LS mirrors can also be used to distribute data volumes to other nodes in the cluster to improve performance for read-only workloads. LS mirrors of data volumes are used in cases in which one or a few clients have read-write access to the dataset and many clients have read-only access. For example, a few servers that generate a large amount of test output data into the source volume and many servers that have read-only access to the test data can process it to output reports with increased performance because the read-only workload has been distributed across the cluster.

In the configuration shown in Figure 15-4, all volumes appear as one volume and are presented as read-only to all clients. In the same way that LS mirrors of a namespace root volume can distribute connections across a cluster, the read-only LS mirror volumes can be made available on every node, so that all clients can connect to the read-only volume by a direct data access path.

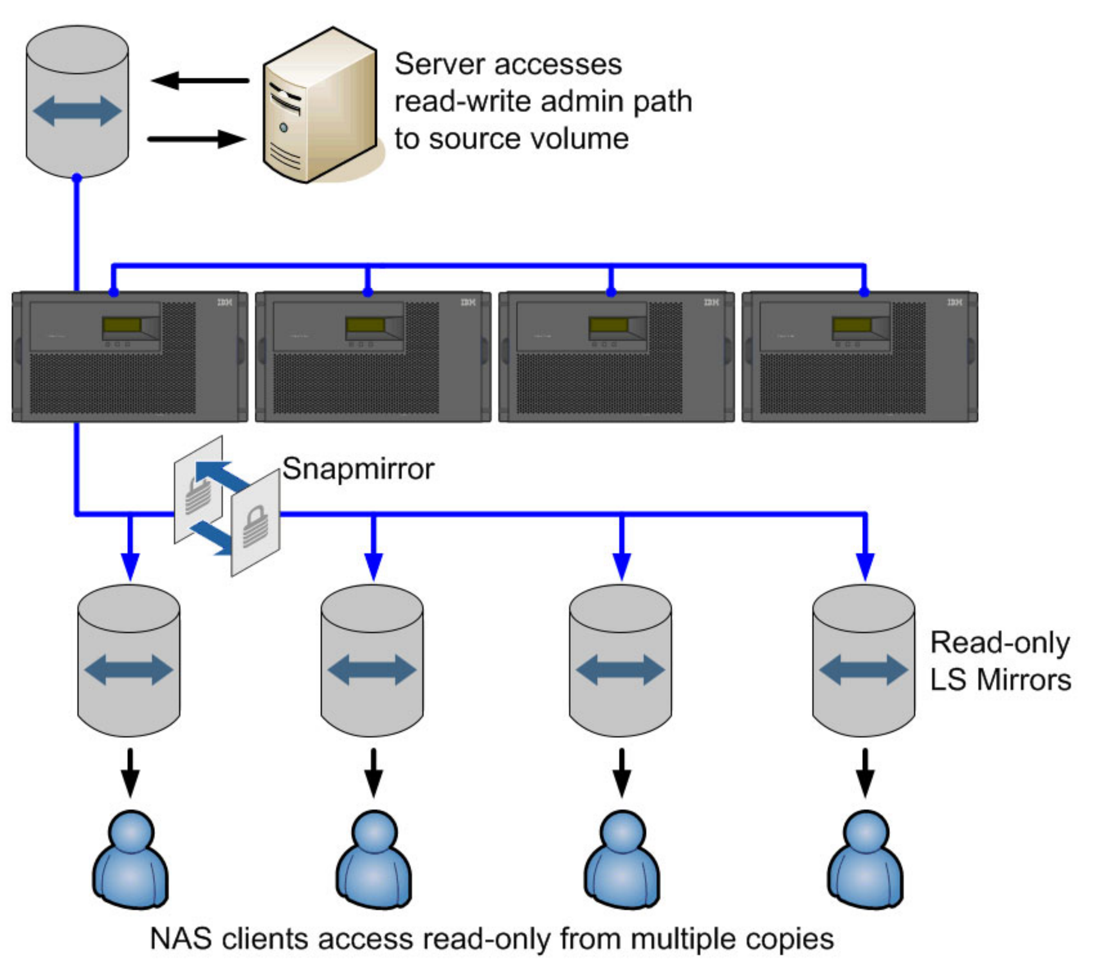


Figure 15-4 LS mirrors for read-only workloads

16

Performance considerations

This chapter focuses on performance, more specifically, the tools that the IBM N series provides to simplify performance increase by configuring certain features.

The following topics are covered:

- FlexCache
- Virtual Storage Tiering
- Storage Quality of Service (QoS)

16.1 FlexCache

A FlexCache volume is a sparsely-populated volume on a cluster node, that is backed by a volume, usually present on a different node within that cluster. A sparsely-populated volume or a sparse volume provides access to data in the backing volume (also called the origin volume) without requiring that all the data be in the sparse volume.

You can use only FlexVol volumes to create FlexCache volumes. However, many of the regular FlexVol volumes features are not supported on FlexCache volumes, such as Snapshot copy creation, deduplication, compression, FlexClone volume creation, volume move, and volume copy.

You can use FlexCache volumes to speed up access to data, or to offload traffic from heavily accessed volumes. FlexCache volumes help improve performance, especially when clients need to access the same data repeatedly, because the data can be served directly without having to access the source. Therefore, you can use FlexCache volumes to handle system workloads that are read- intensive.

Cache consistency techniques help in ensuring that the data served by the FlexCache volumes remains consistent with the data in the origin volumes.

16.1.1 Contents of a cached file

When the client requests a data block of a specific file from a FlexCache volume, then the attributes of that file and the requested data block are cached. The file is then considered to be cached, even if all its data blocks are not present in the FlexCache volume. If the requested data is cached and valid, a read request for that data is fulfilled without access to the origin volume.

16.1.2 Serving read requests

A FlexCache volume directly serves read requests if it contains the data requested by the client. Otherwise, the FlexCache volume requests the data from the origin volume and stores the data before serving the client request. Subsequent read requests for the data are then served directly from the FlexCache volume.

FlexCache volumes serve client read requests as follows:

1. A cluster node, which corresponds to the logical interface (LIF) on which the client sends its read request, accepts the request.
2. The node responds to the read request based on the types of volumes it contains. See Table 16-1 for the FlexCache volume behaviors to certain actions.

Table 16-1 FlexCache volume behaviors

If the node contains...	Then
A FlexCache volume that contains the requested data and the origin volume	The data is served from the origin volume
A FlexCache volume that contains the requested data but not the origin volume	The data is served from the FlexCache volume.
A FlexCache volume that does not contain the requested data	The FlexCache volume retrieves the requested data from a volume that contains the data, stores the data, and serves the client request

If the node contains...	Then
A volume that is the primary source of the requested data but does not contain a FlexCache volume	The data is served directly from the volume containing the requested data.

Note: If the node does not contain either the primary source of the data or a FlexCache volume, the client request is directly passed to a node that contains the primary source of the data.

16.1.3 Why using FlexCache volumes

FlexCache volumes are used to improve performance and balance resources during data read operations.

- Performance scaling:

  A data volume when created is stored on a specific node of the cluster. That volume can move within the cluster, but at any point in time, only one node contains the source data. If there is intensive access to the data on that volume, then that node in the cluster can get overloaded, and develop a performance bottleneck.
  FlexCache volumes scale performance by enabling multiple nodes of a cluster to respond to read requests efficiently without having to overload the node containing the source data and without having to send data over the cluster interconnect (for cache hits).

- Resource balancing:

  Certain nodes of a cluster can encounter spikes of high performance during certain tasks or activities to a specific data set. By caching copies of data throughout the cluster, FlexCache volumes efficiently enable each node in the cluster to handle the workload. This approach spreads the workload across the cluster, smoothing out the performance created by heavy read or metadata access.

16.1.4 Considerations for working with FlexCache volumes

Take into account the following considerations when creating and working with FlexCache volumes:

- You do not need to install any license for creating FlexCache volumes.
- Using Clustered Data ONTAP, you can cache a FlexVol volume within the storage virtual machine (SVM) that contains the origin volume.

  You must use a caching system running Data ONTAP operating in 7-Mode if you want to cache a FlexVol volume outside the cluster.

- You cannot use Infinite Volumes as the caching or origin volume.
  You can use only FlexVol volumes to cache data in other FlexVol volumes.
- To cache a FlexVol volume within a cluster, you must ensure that the FlexCache volumes and the origin volumes are created on storage systems supported by Clustered Data ONTAP 8.2 or later.

  Note: For information about the requirements that the storage system running Data ONTAP operating in 7-Mode must meet for caching a Clustered Data ONTAP FlexVol volume, see the *Data ONTAP Storage Management Guide for 7-Mode*.

- You can create FlexCache volumes on a specific cluster node or on all the cluster nodes spanned by the SVM that contains the origin volume.
- FlexCache volumes are created with a space guarantee type of partial.
  The partial guarantee type is a special guarantee type that cannot be changed. You cannot view the space guarantee type for a FlexCache volume from the command line interface. When you use commands such as volume show to view the volume's space guarantee type, the value for the particular field shows a dash.
- There is no specific limit on the size of the origin volume that a FlexCache volume can cache.
- A FlexCache volume is created with the same language setting as its corresponding origin volume.
- Flash Cache is supported on nodes with FlexCache volumes, and optimizes performance and efficiency accordingly for all volumes on the node, including FlexCache volumes.
- Storage Accelerator (SA) systems do not support Clustered Data ONTAP. SA systems support only Data ONTAP operating in 7G or 7-Mode.
- FlexCache volumes support client access using the following protocols, NFSv3, NFSv4.0, and CIFS (SMB 1.0, 2.x, and 3.0).

  In addition, FlexCache volumes can retrieve from the origin volumes the Access Control Lists (ACLs) and the stream information for the cached data, depending on the protocol.
- For better performance of FlexCache volumes in a cluster, you must ensure that data LIFs are properly configured on the cluster nodes that contain the FlexCache volumes.

16.1.5 Limitations of FlexCache volumes

You can have a maximum of 100 FlexCache volumes on a cluster node. In addition, certain features of Data ONTAP are not available on FlexCache volumes, and other features are not available on origin volumes that are backing the FlexCache volumes.

You cannot use the following Data ONTAP capabilities on FlexCache volumes (these limitations do not apply to origin volumes):

- Compression:

  Compressed origin volumes are supported.
- Snapshot copy creation
- SA systems
- SnapManager
- SnapRestore
- SnapMirror
- FlexClone volume creation
- The ndmp command
- Quotas
- Volume move
- Volume copy
- Cache load balancing
- Qtree creation on cache:

  Qtree management must be done on the origin.

- Deduplication
- Mounting the FlexCache volume as a read-only volume
- I2P:

  The cache volume will not synchronize I2P information from the origin; any requests for this information are always forwarded to the origin.
- Storage QoS policy group:

  An origin volume can be assigned to a policy group, which controls the origin volume and its corresponding FlexCache volumes.

> **Note:** The following N series systems do not support FlexCache volumes in a Clustered Data ONTAP environment:N6040, N6060, N6210, and N6240.

The following limitations apply to origin volumes:

- You must map FlexCache volumes to an origin volume that is inside the same SVM. FlexCache volumes in a Clustered Data ONTAP environment cannot point to an origin volume that is present outside the SVM, such as a 7-Mode origin volume.
- You cannot use a FlexCache volume to cache data from Infinite Volumes. The origin volume must be a FlexVol volume.
- A load-sharing mirror volume or a volume that has load-sharing mirrors attached to it cannot serve as an origin volume.
- Any volume in a SnapVault relationship cannot serve as an origin volume.
- You cannot use an origin volume as the destination of a SnapMirror migrate command.
- A FlexCache volume cannot be used as an origin volume.

16.1.6 Comparison of FlexCache volumes and load-sharing mirrors

Both FlexCache volumes and load-sharing mirror volumes can serve hosts from a local node in the cluster, instead of using the cluster interconnect to access the node storing the primary source of data. However, you need to understand the essential differences between them and use them in your storage system.

Table 16-2 explains the differences between load-sharing mirror volumes and FlexCache volumes.

Table 16-2 Comparison between Load-sharing mirror volumes and FlexCache volumes

Load-sharing mirror volumes	FlexCache volumes
The data that load-sharing mirror volumes use to serve client requests is a complete copy of the source data.	The data that FlexCache volumes use to serve client requests is cached copy of the source data, containing only data blocks that are accessed by clients.
Can be used as a disaster-recovery solution by promoting a load-sharing mirror to a source volume.	Cannot be used for disaster recovery. A FlexCache volume does not contain a complete copy of the source data.
Are read-only volumes, with the exception of admin privileges for write access or bypass of the load-sharing mirror.	Are read and write-through cache volumes.

Load-sharing mirror volumes	FlexCache volumes
A user creates one load-sharing mirror volume at a time.	A user can create one FlexCache volume at a time, or can simultaneously create FlexCache volumes on all the nodes spanned by the SVM that contains the origin volume.

16.2 Virtual Storage Tiering

The Virtual Storage Tier is the IBM N series approach to automated storage tiering. We had several important goals in mind when we set out to design Virtual Storage Tier components:

- Use storage system resources as efficiently as possible, especially by minimizing I/O to disk drives Provide a dynamic, real-time response to changing I/O demands of applications
- Fully integrate storage efficiency capabilities so efficiency is not lost when data is promoted to the Virtual Storage Tier
- Use fine data granularity so that cold data never gets promoted with hot data thus making efficient use of expensive Flash media
- Simplify deployment and management

The Virtual Storage Tier is a self-managing, data-driven service layer for storage infrastructure. It provides real-time assessment of workload priorities and optimizes I/O requests for cost and performance without the need for complex data classification and movement.

The Virtual Storage Tier leverages key storage efficiency technologies, intelligent caching, and simplified management. You simply choose the default media tier you want for a volume or LUN (SATA,FC or SAS). Hot data from the volume or LUN is automatically promoted on demand to flash-based media.

The Virtual Storage Tier promotes hot data without the data movement overhead associated with other approaches to automated storage tiering. Any time a read request is received for a block on a volume or LUN where the Virtual Storage Tier is enabled, that block is automatically subject to promotion. Note that promotion of a data block to the Virtual Storage Tier is not data migration because the block remains on hard disk media when a copy is made to the Virtual Storage Tier.

With the Virtual Storage Tier, data is promoted to Flash media after the first read from hard disk drives. This approach to data promotion means that additional disk I/O operations are not needed to promote hot data. By comparison, other implementations may not promote hot data until it has been read from disk many times, and then additional disk I/O is still required to accomplish the promotion process.

Our algorithms distinguish high-value data from low-value data and then retain that data in the Virtual Storage Tier. Metadata, for example, is always promoted when read for the first time. In contrast, sequential reads are normally not cached in the Virtual Storage Tier unless specifically enabled because they tend to crowd out more valuable data.You can change the behavior of the intelligent cache to meet the requirements of applications with unique data access requirements. For example, you can configure the Virtual Storage Tier to cache incoming random writes as they are committed to disk and to enable the caching of sequential reads.

You can optionally create different classes of service by enabling or disabling the placement of data into the Virtual Storage Tier on a volume-by-volume basis.

16.3 Storage Quality of Service (QoS)

Storage QoS is a new feature in Data ONTAP that provides the ability to group storage objects and set throughput limits on the group. With this ability, a storage administrator can separate workloads by organization, application, business unit, or production versus development environments.

In enterprise environments, storage QoS offers these benefits:

- Helps to prevent user workloads from affecting each other
- Helps to protect critical applications critical applications that have specific response times that must be met

In IT as a service (ITaaS) environments, storage QoS offers these benefits:

- Helps to prevent tenants from affecting each other
- Helps to avoid performance degradation with each new tenant

Storage QoS differs from the FlexShare quality-of-service tool. FlexShare is only in 7-Mode, and storage QoS is only in Clustered Data ONTAP. FlexShare works by setting relative priorities on workloads or resources, and attaining the desired results can be complicated. Storage QoS sets hard limits on collections of one or more objects and replaces complex relative priorities with very specific limits to throughput.

Clustered Data ONTAP 8.2 provides Storage QoS policies on cluster objects. An entire SVM, or a group of volumes or LUNS within an SVM, can be dynamically assigned to a policy group, which specifies a throughput limit, defined in terms of IOPS or MB/sec. This can be used to reactively or proactively throttle rogue workloads and prevent them from affecting the rest of the workloads. QoS policy groups can also be used by service providers to prevent tenants from affecting each other, as well as to avoid performance degradation of the existing tenants when a new tenant is deployed on the shared infrastructure.

Part 3

Cluster setup

This part of the book provides guidance and checklists for planning and implementing the initial hardware installation and cluster setup.

To help perform the initial hardware and cluster setup, we also describe the administrative interfaces and non-disruptive operations.

The following topics are covered:

- Physical installation
- Non-disruptive operations
- Command Line Interface (CLI)
- N series OnCommand System Manager 3.0

17

Physical installation

This chapter describes the planning, prerequisite tasks, and implementation tasks that need to be completed for a successful IBM N series Clustered Data ONTAP implementation.

The following topics are covered:

- Installation prerequisites
- Configuration worksheet
- Initial hardware setup
- Setting up the cluster and joining nodes
- Setting up the cluster base
- Creating an SVM
- Post-Installation and verification checklist

17.1 Installation prerequisites

This section describes, at a high level, some of the planning and prerequisite tasks that need to be completed for a successful N series implementation.

For more information, see the *N series Introduction and Planning Guide* at this website:

http://www-304.ibm.com/support/docview.wss?crawler=1&uid=ssg1S7001913

17.1.1 Pre-installation checklist

Before arriving at the customer site, send the customer the relevant system specifications, and a pre-installation checklist to complete. This list should contain environmental specifications for N series equipment:

- Storage controller weight, dimensions, and rack units.
- Power requirements
- Network connectivity

The customer completes the pre-installation checklist with all the necessary information about their environment, such as host name, IP, DNS, AD, and Network.

Work through this checklist with the customer and inform them about the rack and floor space requirements. This process speeds up the installation time because all information has been collected beforehand.

After this process is complete and equipment is delivered to the customer, you can arrange an installation date.

17.1.2 Before arriving on site

Before arriving at the customer site, ensure that you have the following tools and resources:

- Required software and firmware:
  - Data ONTAP software (take note of the storage platform)
  - Latest firmware files:
    - Expansion shelf firmware
    - Disk firmware
    - SP / RLM firmware
    - System firmware
- Appropriate tools and equipment:
  - Pallet jack, forklift, or hand truck, depending on the hardware that you receive
  - #1 and #2 Phillips head screwdrivers, and a flathead screwdriver for cable adapters
  - A method for connecting to the serial console:
    - A USB-to-Serial adapter
    - Null modem cable (with appropriate connectors)
- Documentation stored locally on your mobile computer such as ONTAP documentation, HW documentation
- Sufficient people to safely install the equipment into a rack:
  - Two or three people are required, depending on the hardware model
  - See the specific hardware installation guide for your equipment

17.2 Configuration worksheet

Before powering on your storage system for the first time, use the configuration worksheet (Table 17-1) to gather information for the software setup process.

Table 17-1 Configuration worksheet

Type of information		**Your values**
Management and Cluster Interconnect switches should be configured to use the same Name Server, NTP and Mail Host details as the cluster nodes. These details are recorded in a later section of this document.		
Switch Information	Management Switch A Name	
	Management Switch A IP Address	
	Management Switch A Netmask	
	Management Switch A Gateway	
	Management Switch B Name	
	Management Switch B IP Address	
	Management Switch B Netmask	
	Management Switch B Gateway	
	Cluster Interconnect Switch A Name	
	Cluster Interconnect Switch A IP Address	
	Cluster Interconnect Switch A Netmask	
	Cluster Interconnect Switch A Gateway	
	Cluster Interconnect Switch B Name	
	Cluster Interconnect Switch B IP Address	
	Cluster Interconnect Switch B Netmask	
	Cluster Interconnect Switch B Gateway	
Cluster	Cluster Name	
	Cluster Base Aggregate	
License	Cluster-Base	

Type of information		Your values
Admin SVM	Cluster administrator password	
	Cluster management LIF IP address	
	Cluster management LIF netmask	
	Cluster management LIF default gateway	
	DNS domain name	
	Name server IP addresses	
Time Synchronization	Time services protocol (NTP)	
	Time Servers (up to 3 IP Addresses)	
	Max time skew (<5 minutes for CIFS)	
	Time Zone (for example Europe/Berlin)	
Node Information	Node 1 Name	
	Node 1 Serial Number	
	Node 2 Name	
	Node 2 Serial Number	
	Node 3 Name	
	Node 3 Serial Number	
	Node 4 Name	
	Node 4 Serial Number	
The following sections contain all items for one object only. If there is more than one object (which likely will be the case), replicate the appropriate section to add additional objects (for example VLANs).		
Interface Groups (IFGRP)	IFGRP Name	
	Node	
	Distribution function	
	Mode	
	Ports	
Virtual LANs (VLANs)	VLAN Name	
	Node	
	Associated Network Port	
	Switch VLAN ID	

Type of information		Your values
Logical Interfaces (LIFs)	LIF name	
	Home node	
	Home port	
	Netmask	
	Routing group	
	Failover group	
Routing Groups	Vserver	
	Routing Group	
	Subnet	
	Destination Network	
	Gateway	
	Metric	
	Role	
Intercluster Network	Node name	
	Port	
	LIF name	
	IP address	
	Netmask	
Service Processor (SP)	IP address	
	Network mask	
	Gateway	

17.3 Initial hardware setup

The following three sections describe how the N series Clustered Data ONTAP nodes, also known as HA pairs, are set up and how they are connected to the cluster network switches.

17.3.1 HA pairs

An N series Clustered Data ONTAP system consists of one or multiple HA pairs, which are all connected to a shared cluster network. Although the controllers in an HA pair are connected to other controllers in the cluster through the cluster network, the HA interconnect and disk-shelf connections are found only between the node and its partner and their disk shelves, hence only the nodes in the HA pair can take over each other's storage.

Figure 17-1 illustrates the functional design of multiple HA pairs in a cluster.

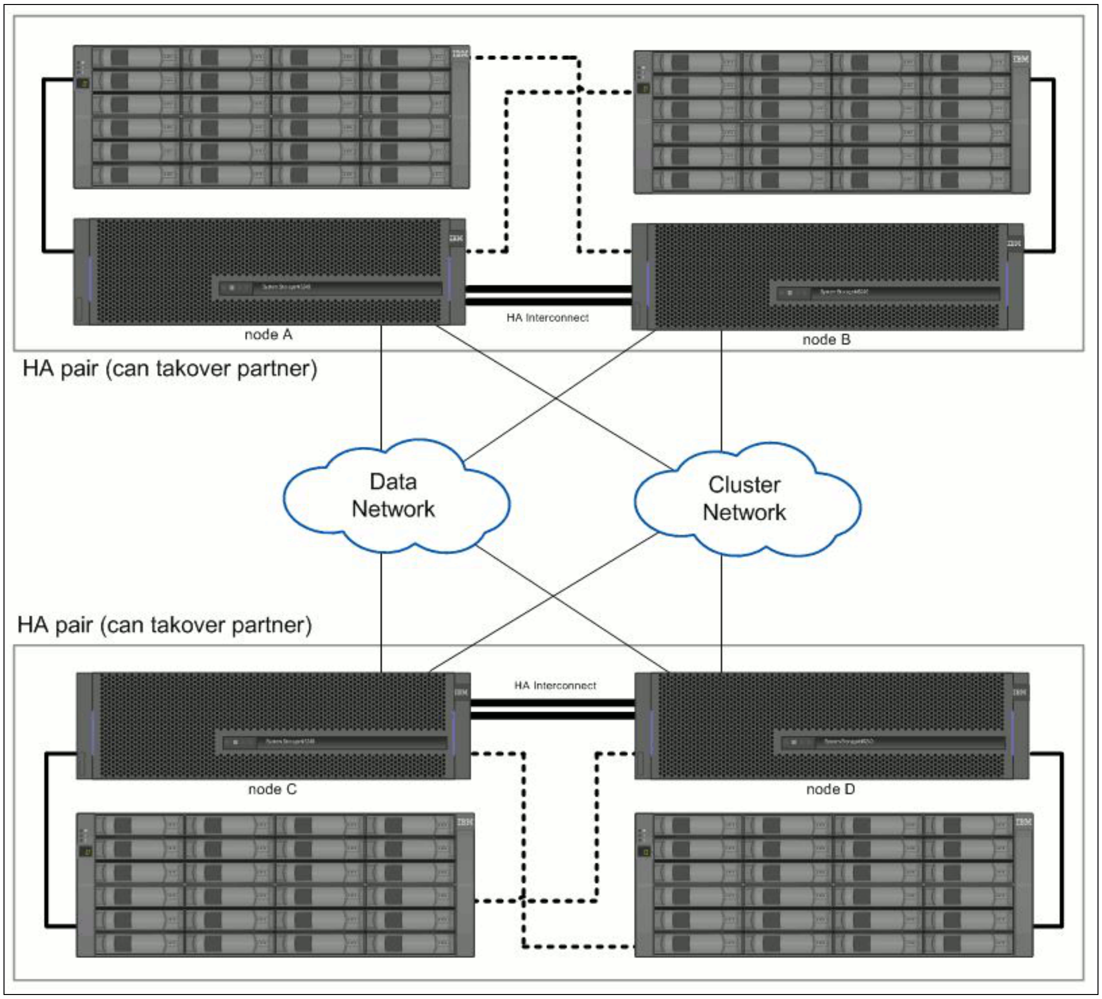


Figure 17-1 Functional design of multiple HA pairs

Regarding only the initial hardware setup of a single HA pair, nothing has changed in comparison to an N series 7-Mode system. Therefore, we refer you to the *IBM System Storage N series Hardware Guide,* SG24-7840, regarding the hardware setup, available at the following website:

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

17.3.2 Cluster network

The cluster network consists of two CN1610 managed Layer 2 switches where each provides 16 10 GE Small Form-Factor Plugable Plus (SFP+) ports and features four ISL ports with an inband/outband management port.

These switches are designed to work in clusters ranging from two to eight nodes as a supported configuration, although there are ports to connect 12 nodes. Four ports are reserved for further use.

Each of the controllers has to be connected to every switch. It is leading practice to use dedicated 10 GE cards to connect to the cluster network if possible (for example e1a, e2a).

See Figure 17-2 for a cabling example.

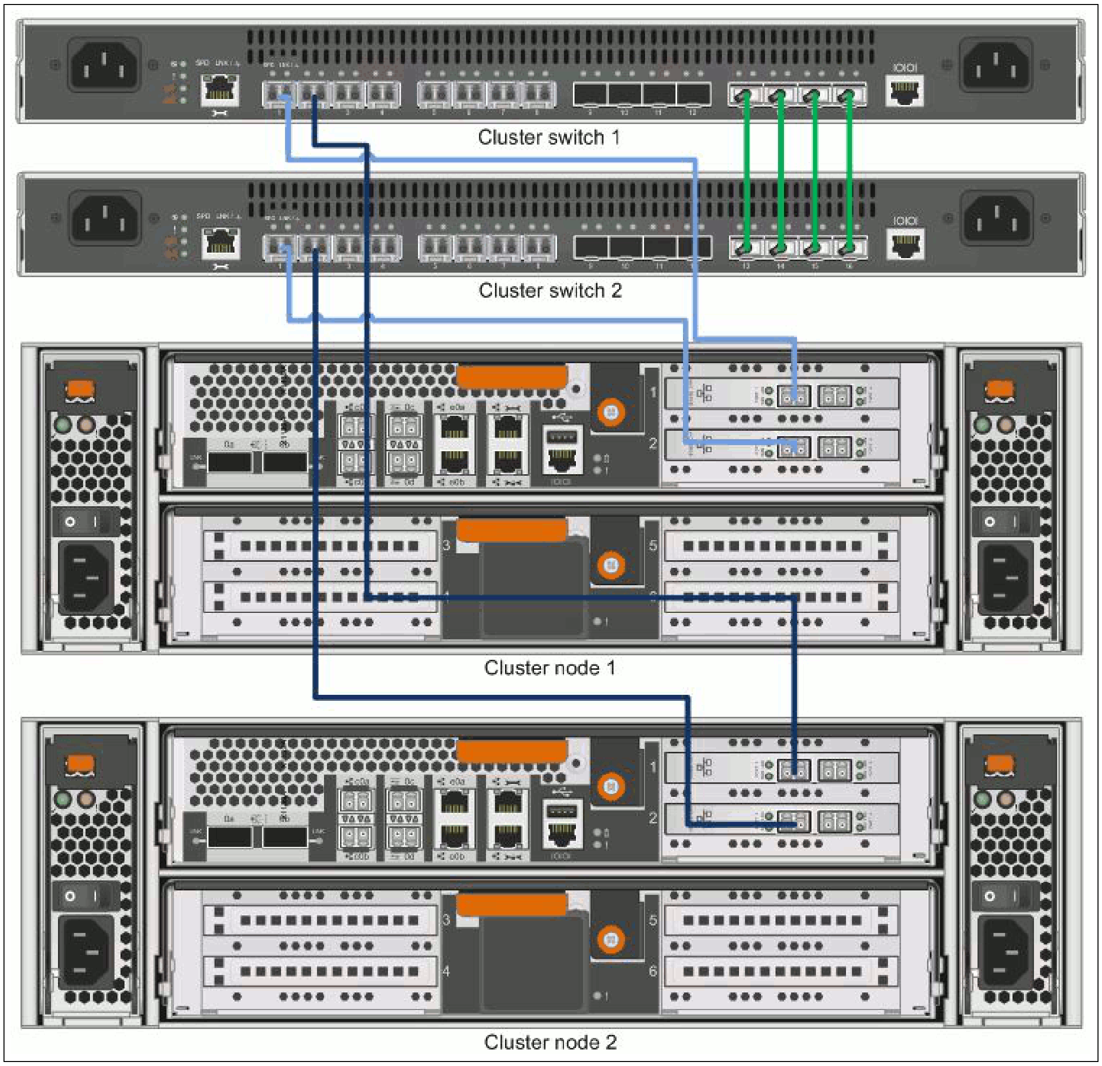


Figure 17-2   Cabling example cluster network

The following port assignment table (Table 17-2) provides the preferred port assignments from the CN1610 switches to the controllers.

Table 17-2   Cluster network port assignment table

CN1610 cluster switch A		**CN1610 cluster switch B**	
Switch port	**Node/port usage**	**Switch port**	**Node/port usage**
1	Node 1 cluster port 1	1	Node 1 cluster port 2
2	Node 2 cluster port 1	2	Node 2 cluster port 2
3	Node 3 cluster port 1	3	Node 3 cluster port 2
4	Node 4 cluster port 1	4	Node 4 cluster port 2
5	Node 5 cluster port 1	5	Node 5 cluster port 2
6	Node 6 cluster port 1	6	Node 6 cluster port 2

CN1610 cluster switch A		CN1610 cluster switch B	
7	Node 7 cluster port 1	7	Node 7 cluster port 2
8	Node 8 cluster port 1	8	Node 8 cluster port 2
9 (reserved)	Node 9 cluster port 1	9 (reserved)	Node 9 cluster port 2
10 (reserved)	Node 10 cluster port 1	10 (reserved)	Node 10 cluster port 2
11 (reserved)	Node 11 cluster port 1	11 (reserved)	Node 11 cluster port 2
12 (reserved)	Node 12 cluster port 1	12 (reserved)	Node 12 cluster port 2
13	ISL to switch B port 13	13	ISL to switch A port 13
14	ISL to switch B port 14	14	ISL to switch A port 14
15	ISL to switch B port 15	15	ISL to switch A port 15
16	ISL to switch B port 16	16	ISL to switch A port 16

CN1610 initial setup

Follow the general summary of this process to customize the switches for your environment's needs. All of your cluster network switches should arrive with the standard factory default configuration installed on them. These switches should also have the current version of the firmware and reference configuration files (RCFs) loaded.

To complete the configuration, follow these steps:

1. Power on the switches.
2. Connect the serial port (the RJ-45 socket on the right side of the switch) to the host or serial port of your choice.
3. Connect the management port (the RJ-45 wrench port on the left side of the switch) to your management network.
4. At the console, set the host side serial settings:
   a. 9600 baud
   b. 8 data bits
   c. 1 stop bit
   d. Parity: none
   e. Flow control: none
5. Log in to the switch. The switch user is *admin* and there is no password by default. At the (CN1610) prompt, enter **enable**. This gives you access to Privileged EXEC mode, which allows you to configure the network interface as shown in Example 17-1.

   Example 17-1 Log in to the switch

   ```
   User:admin
   Password:
   (CN1610) >enable
   Password:
   ```

6. Prepare to connect to the management network. If you are using DHCP, you do not need to do this. The service port is set to use DHCP by default. The network management port will be set to none for the IPv4 and IPv6 protocol settings. If your wrench port is connected to the network that has a DHCP server, that part is done. If you are setting a static IP address, use the **serviceport protocol**, **network protocol**, and **serviceport ip** commands as shown in Example 17-2.

Example 17-2 Configure service port

```
(CN1610) #serviceport protocol none
(CN1610) #network protocol none
(CN1610) #serviceport ip ipaddr netmask gateway
```

7. To verify the results, use the **show serviceport** command as shown in Example 17-3.

Example 17-3 Show serviceport

```
(CN1610) #show serviceport
Interface Status............................... Up
IP Address..................................... 10.x.x.x
Subnet Mask.................................... 255.255.255.0
Default Gateway................................ 10.x.x.x
IPv6 Administrative Mode....................... Enabled
IPv6 Prefix is ................................
fe80::2a0:98ff:fe4b:abfe/64
Configured IPv4 Protocol....................... None
Configured IPv6 Protocol....................... None
IPv6 AutoConfig Mode........................... Disabled
Burned In MAC Address.......................... 00:A0:98:4B:AB:FE
```

8. Set an appropriate host name, to be able to identify the switch with the **hostname** command as shown in Example 17-4.

Example 17-4 Setting hostname

```
(CN1610) #hostame clusterswitch1
```

9. Configure the embedded ntp client with the **sntp** command as shown in Example 17-5.

Example 17-5 Configure sntp

```
(clusterswitch1) #sntp server 10.10.10.230 1
(clusterswitch1) #show sntp server
```

10. Generate crypto-keys and enable the ssh server. You have to switch into the configure mode to generate the keys. After you have done that, save the changes with **write memory**, switch back with **exit** and enable the ssh server with the **ip** command as shown in Example 17-6.

Example 17-6 Configure ssh

```
(clusterswitch1) #configure
(clusterswitch1) config#crypto key generate dsa
(clusterswitch1) config#crypto key generate rsa
(clusterswitch1) config#write memory
(clusterswitch1) #exit
(clusterswitch1) #ip ssh server enable
```

11. Check the settings by using the **`show running-config`** command as shown in Example 17-7.

Example 17-7 Display running-config

```
(clusterswitch1) #show running-config
```

12. If you are satisfied with the configuration, write it to memory with the **`write memory`** command as shown in Example 17-8. Enter **`y`** when prompted to save the configuration.

Example 17-8 Write config to memory

```
(clusterswitch1) #write memory
This operation may take a few minutes.
Management interfaces will not be available during this time.
Are you sure you want to save? (y/n) y
Config file 'startup-config' created successfully.
```

17.3.3 Switchless cluster

You can optionally configure two-node clusters without cluster network switches. Instead, you can apply the networking switchless-cluster option and use direct, back-to-back connections between the nodes.

If you have a two-node switchless configuration in which there is no cluster interconnect switch, you must ensure that the switchless-cluster-network option is enabled (see Example 17-9). This ensures proper cluster communication between the nodes.

Example 17-9 Configuring switchless cluster

```
cluster1::> set -privilege advanced
cluster1::> network options switchless-cluster show
cluster1::> network options switchless-cluster modify true
```

See Figure 17-3 for a cabling example.

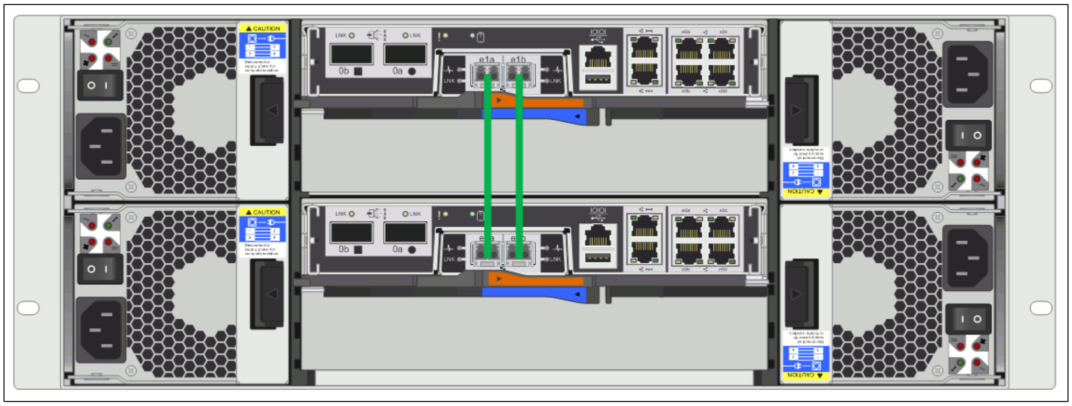


Figure 17-3 Cabling example switchless cluster

Note: In a two-node cluster, if the cluster network between the two nodes fails, one of the nodes will panic and the other node will take over its services and maintain client access to the data.

17.4 Setting up the cluster and joining nodes

Setting up the cluster involves creating the cluster on the first node, joining any remaining nodes to the cluster, and configuring a number of features, such as synchronizing the system time, which enable the cluster to operate non-disruptively.

17.4.1 Creating a cluster on one node

You use the Cluster Setup wizard to create the cluster on the first node. The wizard helps you to configure the cluster network that connects the nodes, create the cluster admin storage virtual machine (SVM), add feature license keys, and create the node management interface for the first node.

You should have completed the configuration worksheet (see 17.2, “Configuration worksheet” on page 285), the storage system hardware should be installed and cabled (see 17.7.1, “Physical installation” on page 309), and the console should be connected to the node on which you intend to create the cluster.

If you do not want to stay near the nodes during the setup steps, you can configure the service processor (or rlm) from the loader prompt in order to connect via remote console later on. To do that, you need to cancel the autoboot process by pressing **`ctrl+c`** and the node will drop to the `loader>` prompt. From there you can set up the service processor (or rlm) with the commands shown in Example 17-10.

Example 17-10 Configuring service-processor

```
LOADER-A> sp setup
Would you like to configure the SP? y
Would you like to enable DHCP on the SP LAN interface? [yN] N
Please enter the IP address for the SP [10.10.101.80]:
Please enter the netmask for the SP [255.255.255.0]:
Please enter the IP address for the SP gateway [10.10.101.254]:
Do you want to enable IPv6 on the SP? [yN]

Service Processor New Network Configuration

Ethernet Link:      up, 100 Mb, full duplex, auto-neg complete
Mgmt MAC Address:   00:A0:98:1A:2A:4E
IPv4 Settings
 Using DHCP:        NO
 IP Address:        10.10.101.80
 Netmask:           255.255.255.0
 Gateway:           10.10.101.254
IPv6:               Disabled
```

Now, as you have established either direct or console access to the node, you can continue with the setup:

1. To start, power on the first node (if you already have powered on, type **`autoboot`** in the `loader>` prompt) and wait until the system has finished booting and the cluster setup wizard appears on the console, as shown in Example 17-11.

Example 17-11 Cluster setup wizard

```
Welcome to the cluster setup wizard.

You can enter the following commands at any time:
  "help" or "?" - if you want to have a question clarified,
  "back" - if you want to change previously answered questions, and
  "exit" or "quit" - if you want to quit the cluster setup wizard.
     Any changes you made before quitting will be saved.

You can return to cluster setup at any time by typing "cluster setup".
To accept a default or omit a question, do not enter a value.

Do you want to create a new cluster or join an existing cluster? {create, join}:
```

2. Type in **`create`** to create a new cluster with the wizard.
3. Follow the prompts to complete the cluster setup. To accept the default value for a prompt, press Enter. Example 17-12 shows a cluster setup procedure on a N6270. If the cluster setup wizards asks you to reboot the node in order to activate HA, you should do that. During the cluster setup wizard you can choose if you want to create a single node or switchless cluster.

Example 17-12 Create cluster

```
Do you want to create a new cluster or join an existing cluster? {create, join}:
create

Do you intend for this node to be used as a single node cluster? {yes, no} [no]:
no

System Defaults:
Private cluster network ports [e2a].
Cluster port MTU values will be set to 9000.
Cluster interface IP addresses will be automatically generated.
The cluster will be connected using network switches.

Do you want to use these defaults? {yes, no} [yes]: no

Step 1 of 5: Create a Cluster
You can type "back", "exit", or "help" at any question.

List the private cluster network ports [e2a]: e2a,e2b
Enter the cluster ports' MTU size [9000]:
Enter the cluster network netmask [255.255.0.0]:

Generating a default IP address. This can take several minutes...
Enter the cluster interface IP address for port e2a [169.254.216.146]:

Generating a default IP address. This can take several minutes...
Enter the cluster interface IP address for port e2b [169.254.67.99]:

Will the cluster network be configured to use network switches? [yes]:
no

Enter the cluster name: cdot-cluster1
Enter the cluster base license key: XXXXXXXXXXXXXXXXXXXXX
```

```
Creating cluster cdot-cluster01

Network set up
Creating cluster
System start up

Cluster cdot-cluster01 has been created.
Name of primary contact: ITSO Team
Phone number of primary contact: +49 1234567
Alternate phone number of primary contact:
Primary contact e-mail address (or IBM WebID): itsoteam@neries.local
Name of secondary contact:
Phone number of secondary contact:
Alternate phone number of secondary contact:
Secondary contact e-mail address (or IBM WebID):
Business name: IBM
Business address: Hechtsheimer Str.2
City where the business resides: Mainz
State where the business resides: RLP
2-character country code: DE
Postal code where the business resides: 55131

Step 2 of 5: Add Feature License Keys
You can type "back", "exit", or "help" at any question.

Enter an additional license key []:

Step 3 of 5: Set Up a Vserver for Cluster Administration
You can type "back", "exit", or "help" at any question.

Enter the cluster administrator's (username "admin") password:

Retype the password:

Enter the cluster management interface port [e0a]:
Enter the cluster management interface IP address: 9.155.66.34
Enter the cluster management interface netmask: 255.255.255.0
Enter the cluster management interface default gateway: 9.155.66.1

A cluster management interface on port e0a with IP address 9.155.66.34 has been
created.  You can use this address to connect to and manage the cluster.

Enter the DNS domain names: nseries.local
Enter the name server IP addresses: 9.155.113.200
DNS lookup for the admin Vserver will use the nseries.local domain.

Step 4 of 5: Configure Storage Failover (SFO)
You can type "back", "exit", or "help" at any question.

SFO will be enabled when the partner joins the cluster.

Step 5 of 5: Set Up the Node
You can type "back", "exit", or "help" at any question.
```

```
Where is the controller located []: Lab Mainz
Enter the node management interface port [e0M]:
Enter the node management interface IP address: 9.155.90.168
Enter the node management interface netmask: 255.255.255.0
Enter the node management interface default gateway: 9.155.90.1
A node management interface on port e0M with IP address 9.155.90.168 has been
created.

This system will send event messages and weekly reports to IBM Technical Support.
To disable this feature, enter "autosupport modify -support disable" within 24
hours.
Enabling AutoSupport can significantly speed problem determination and resolution
should a problem occur on your system.
For further information on AutoSupport, please see:
http://www.ibm.com/systems/storage/network/software/autosupport/
Press enter to continue:

Cluster setup is now complete.
```

17.4.2 Joining a node to the cluster

After creating a new cluster, for each remaining node, you use the cluster setup wizard to join the node to the cluster and create its node management interface. Therefore, power on each node that you want to join, connect to its console, and wait until the cluster setup wizard appears (see Example 17-11).

1. Type in **`join`** to join the node to the cluster.
2. Follow the prompts to complete the cluster setup. To accept the default value for a prompt, press Enter. Example 17-13 shows a cluster join procedure on an N6270.

Example 17-13 Join additional node

```
Do you want to create a new cluster or join an existing cluster? {join}:
join

System Defaults:
Private cluster network ports [e2a].
Cluster port MTU values will be set to 9000.
Cluster interface IP addresses will be automatically generated.

Do you want to use these defaults? {yes, no} [yes]: no

Step 1 of 3: Join an Existing Cluster
You can type "back", "exit", or "help" at any question.

List the private cluster network ports [e2a]: e2a,e2b
Enter the cluster ports' MTU size [9000]:
Enter the cluster network netmask [255.255.0.0]:

Generating a default IP address. This can take several minutes...
Enter the cluster interface IP address for port e2a [169.254.45.229]:

Generating a default IP address. This can take several minutes...
Enter the cluster interface IP address for port e2b [169.254.250.11]:
```

```
Enter the name of the cluster you would like to join [cdot-cluster01]:

Joining cluster cdot-cluster01

Network set up
Node check
Joining cluster
System start up
Starting cluster support services

This node has joined the cluster cdot-cluster01.

Step 2 of 3: Configure Storage Failover (SFO)
You can type "back", "exit", or "help" at any question.

SFO is enabled.

Step 3 of 3: Set Up the Node
You can type "back", "exit", or "help" at any question.

Notice: HA is configured in management.

Enter the node management interface port [e0M]:
Enter the node management interface IP address: 9.155.90.169
Enter the node management interface netmask [255.255.255.0]:
Enter the node management interface default gateway [9.155.90.1]:

A node management interface on port e0M with IP address 9.155.90.169 has been
created.

This system will send event messages and weekly reports to IBM Technical Support.
To disable this feature, enter "autosupport modify -support disable" within 24
hours.
Enabling AutoSupport can significantly speed problem determination and resolution
should a problem occur on your system.
For further information on AutoSupport, please see:
http://www.ibm.com/systems/storage/network/software/autosupport/
Press enter to continue:

Cluster setup is now
complete.
```

17.5 Setting up the cluster base

After the cluster has been created and after all nodes are joined to the cluster, you can go on setting up the cluster base. You should now already be able to log in either via one of the node management IPs or via the cluster management IP.

17.5.1 Storage failover (SFO)

Storage failover takeover and giveback are the operations that let you take advantage of the HA configuration to perform non-disruptive operations and avoid service interruptions. Takeover is the process in which a node takes over the storage of its partner. Giveback is the process in which the storage is returned to the partner.

After finishing the cluster setup, you should check if storage failover is configured with the **storage failover show** command link as shown in Example 17-14.

Example 17-14 Storage failover show

```
cdot-cluster01::> storage failover show
                              Takeover
Node           Partner        Possible State Description
-------------- -------------- -------- -------------------------------------
cdot-cluster01-01
               cdot-          true     Connected to cdot-cluster01-02
               cluster01-02
cdot-cluster01-02
               cdot-          true     Connected to cdot-cluster01-01
               cluster01-01
```

If the storage failover was not enabled during the cluster setup wizard, you can enable it with the **storage failover modify** command on every node as shown in Example 17-15.

Example 17-15 Enable storage failover

```
cdot-cluster01::> storage failover modify -node node1 -enabled true
cdot-cluster01::> storage failover modify -node node2 -enabled true
```

Note: If you have a two-node cluster, you have to make sure that high availability is configured with the **ha modify -configured true** command. If you add more nodes to the cluster, you have to disable this again.

You can configure hardware-assisted takeover to speed up takeover times. Hardware-assisted takeover uses the remote management device to quickly communicate local status changes to the partner node. Hardware-assisted takeover has to be enabled for every node after completing the cluster setup wizard with the **storage failover modify** command. See Example 17-16, which shows how to enable hardware-assisted takeover.

Example 17-16 Enable hwassist

```
cdot-cluster01::> storage failover modify -node nodeA -hwassist true
-hwassist-partner-ip <ip-of-nodeB>
cdot-cluster01::> storage failover hwassist show
```

17.5.2 Synchronizing the system time across the cluster

Synchronizing the time ensures that every node in the cluster has the same time, and prevents CIFS and Kerberos failures.

You can either use the CLI (see Example 17-17) or the IBM N series OnCommand System Manager to configure the NTP client.

Example 17-17 Setting up time services

```
cdot-cluster01::> system services ntp config modify -enabled true
cdot-cluster01::> system services ntp server create -node nodeA -server
ntp.domain.local
cdot-cluster01::> system services ntp server show
Node   Server                        Version
------ ----------------------------- ----------------------------------------
```

```
nodeA
       ntp.domain.local               3
nodeB
       ntp.domain.local               3
2 entries were displayed.
```

17.5.3 Event management system

You can configure EMS to receive event messages, and to set up the event destinations and the event routes for a particular event severity.

1. As a first step, you should configure the sender and the mail server (see Example 17-18).

   Example 17-18 Configure event management system

   ```
   cdot-cluster01::> event config modify -mailserver mailhost.domain.local
   -mailfrom sender@domain.local

   cdot-cluster01::> event config show
   Mail From: sender@domain.local
   Mail Server: mailhost.domain.local
   ```

2. You can now modify an existing destination in order to receive e-mails on defined events (see Example 17-19).

   Example 17-19 Configure event destinations

   ```
   cdot-cluster01::> event destination modify -name criticals -mail
   admin@domain.local
   cdot-cluster01::> event destination show
                                                                            Hide
   Name             Mail Dest.        SNMP Dest.         Syslog Dest.       Params
   ---------------- ----------------- ------------------ ------------------ ------
   allevents        -                 -                  -                  false
   asup             -                 -                  -                  false
   criticals        admin@domain.local
                                      -                  -                  false
   pager            -                 -                  -                  false
   traphost         -                 -                  -
   false
   ```

3. You can also define a new destination that includes custom routes for special events or severities. Example 17-20 shows the creation of a new destination that sends an e-mail for every event that is more significant than `CRITICAL`.

   Example 17-20 Create event destination

   ```
   cdot-cluster01::> event destination create -name alerts -mail
   admin@domain.local
   cdot-cluster01::> event route add-destinations {-severity <=CRITICAL}
   -destinations alerts
   ```

4. If you want to enable the system to send traps to a traphost, you have to configure the following coding as shown in Example 17-21.

Example 17-21 Configure snmp

```
cdot-cluster01::> system snmp contact -contact "Team Storage, 0900/123456"
cdot-cluster01::> system snmp location -location "Datacenter North, Row 5, Rack
1-5"

cdot-cluster01::> system snmp traphost add -peer-address 1.2.3.4
cdot-cluster01::> system snmp init -init 1
```

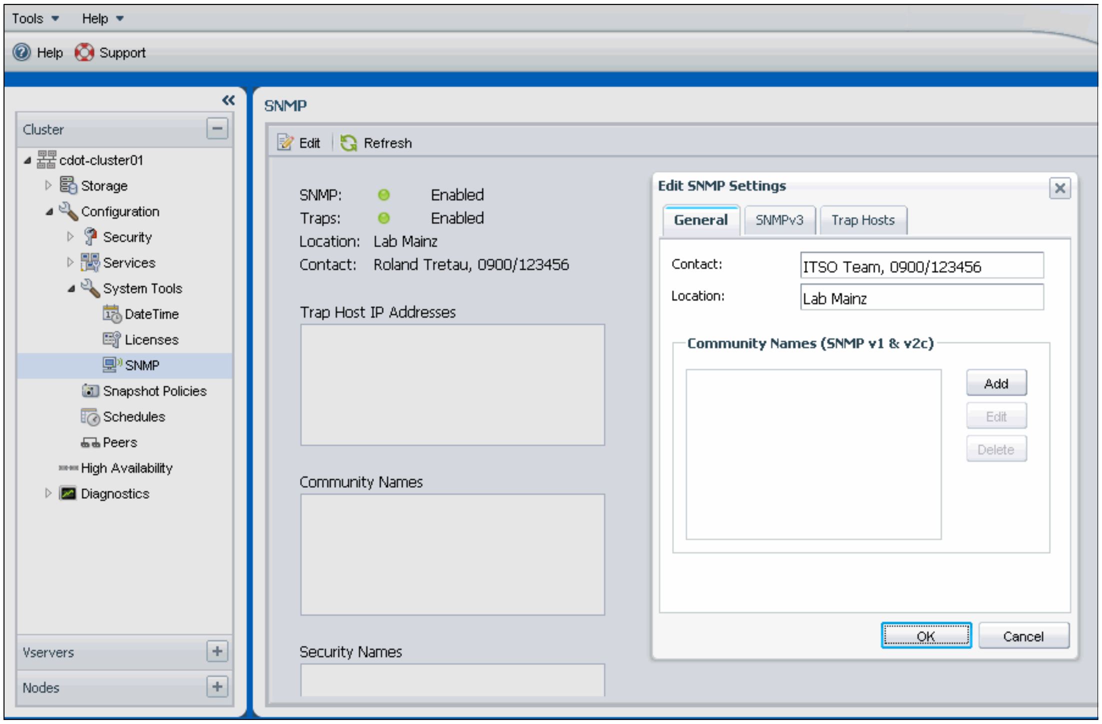
5. There is also the possibility to configure owner and location on per node basis as shown in Example 17-22. This information will be visible in alert emails.

Example 17-22 Configure node information

```
cdot-cluster01::> system node modify -node nodeA -owner "Team Storage,
0900/123456"
cdot-cluster01::> system node modify -node nodeA -location "Datacenter North,
Row 5, Rack 3"
```

You can also use the IBM N series OnCommand System Manager to edit the snmp settings. See Figure 17-4 for an example:

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Cluster** → **Configuration** → **System Tools** → **SNMP**.
3. Click the **Edit** button to configure the snmp settings.



SNMP: Enabled
Traps: Enabled
Location: Lab Mainz
Contact: Roland Tretau, 0900/123456

Contact: ITSO Team, 0900/123456
Location: Lab Mainz

Figure 17-4 Configure snmp in system manager

17.5.4 AutoSupport

To ensure proper callhome and support, it is mandatory to configure autosupport. AutoSupport is a mechanism that proactively monitors the health of your system and automatically sends messages to IBM support, your internal support organization, and a support partner.

Although AutoSupport messages to technical support are enabled by default, you must set the correct options and have a valid mail host to have messages sent to your internal support organization:

1. Ensure AutoSupport is enabled by setting the **-state** parameter of the **system node autosupport modify** command to enable it as shown in Example 17-23.

   Example 17-23 Enabling autosupport

   ```
   cdot-cluster01::> system node autosupport modify -node * -state enable
   ```

2. If you want technical support to receive AutoSupport messages, set the following parameters of the **system node autosupport modify** command as shown in Example 17-24.

   Example 17-24 Enabling autosupport to technical support

   ```
   cdot-cluster01::> system node autosupport modify –node * -support enable
   ```

3. If you want your internal support organization or a support partner to receive AutoSupport messages, you can use the parameters shown in Table 17-3.

Table 17-3 Autosupport recipient types

Set this parameter	To this
-to	Up to five comma-separated individual email addresses or distribution lists in your internal support organization that will receive key AutoSupport messages
-noteto	Up to five comma-separated individual email addresses or distribution lists in your internal support organization that will receive a shortened version of key AutoSupport messages designed for cell phones and other mobile devices
-partnerto	Up to five comma-separated individual email addresses or distribution lists in your support partner organization that will receive all AutoSupport messages

4. Select a transport protocol for messages by setting **-transport** to **smtp**, **http**, or **https** as shown in Example 17-25.

   Example 17-25 Configure transport protocol

   ```
   cdot-cluster01::> system node autosupport modify -node * -transport smtp
   ```

5. If you selected SMTP transport for messages to technical support or you are sending messages to your internal support organization, configure SMTP by setting the parameters in Example 17-26 of the **system node autosupport modify** command.

Example 17-26 Configure smtp

```
cdot-cluster01::> system node autosupport modify -node * -mail-hosts
mailrelay.domain.local
cdot-cluster01::> system node autosupport modify -node * -from
filer@domain.local
```

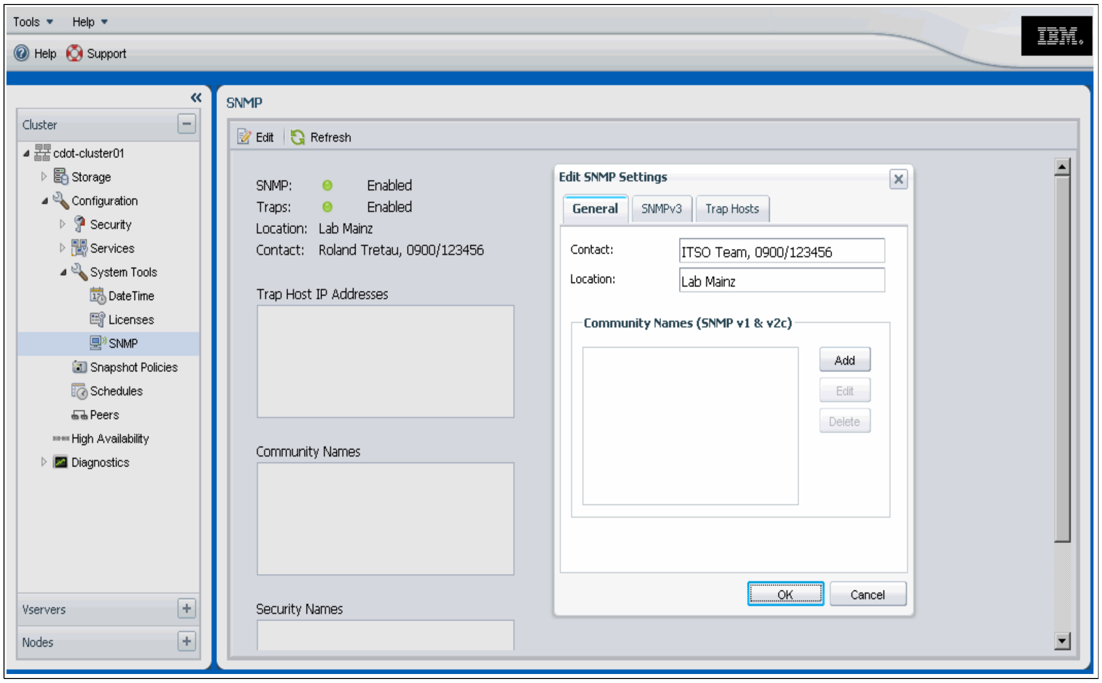
6. See Example 17-27 about how to test that AutoSupport messages are being sent and received.

Example 17-27 Test autosupport

```
cdot-cluster01::> system node autosupport invoke -node * -type all -message
„testing ASUP“
```

You can also use the IBM N series OnCommand System Manager to edit the autosupport settings. See Figure 17-5 for an example.

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Nodes** → **Choose the desired node** → **Configuration** → **AutoSupport.**
3. Click the **Edit** button to configure the autosupport settings.



SNMP: Enabled
Traps: Enabled
Location: Lab Mainz
Contact: Roland Tretau, 0900/123456

Contact: ITSO Team, 0900/123456
Location: Lab Mainz

Figure 17-5 Configure autosupport in system manager

17.5.5 Networking

Before you continue creating your first SVM, you should also configure the following networking components:

- 9.2, “Network ports” on page 133
- 9.3, “Interface groups” on page 135
- 9.4, “VLANs” on page 136
- 9.5, “Failover groups” on page 138

17.5.6 Aggregates

You need to create an aggregate to provide storage to one or more SVM root volumes, FlexVol volumes and Infinite Volumes. Aggregates are a physical storage object; they are associated with a specific node in the cluster.

Drives and array LUNs are owned by a specific node; when you create an aggregate, all drives in that aggregate must be owned by the same node, which becomes the home node for that aggregate.

You can display a list of the available spares by using the **`storage disk show -spare`** command. This command displays the spare drives for the entire cluster. If you are logged in to the cluster on the cluster management interface, you can create an aggregate on any node in the cluster. To ensure that the aggregate is created on a specific node, use the **`-node`** option or specify drives that are owned by that node.

Aggregate names must conform to the following requirements:

- Begin with either a letter or an underscore (_).
- Contain only letters, digits, and underscores.
- Contain no more than 250 characters.

> **Note:** It is preferred practice to use a proper naming scheme for every aggregate in the cluster. Beside the general aggregate naming requirements, we advise to include the name of the home node in the aggregate name, for example, `node1_sas600_01`. You can also rename existing aggregates with the **`storage aggregate rename`** command.

Create the aggregate by using the **`storage aggregate create`** command.

You can specify the following options:

- Aggregate's home node (that is, the node on which the aggregate is located unless the aggregate fails over to the node's storage failover partner)
- List of specific drives or array LUNs that are to be added to the aggregate
- Number of drives to include
- Checksum style to use for the aggregate
- Type of drives to use
- Size of drives to use
- Disk speed to use
- RAID type for RAID groups on the aggregate
- Maximum number of drives or array LUNs that can be included in a RAID group
- Whether drives with different rpm are allowed

For more information about these options, see the **`storage aggregate create`** man page.

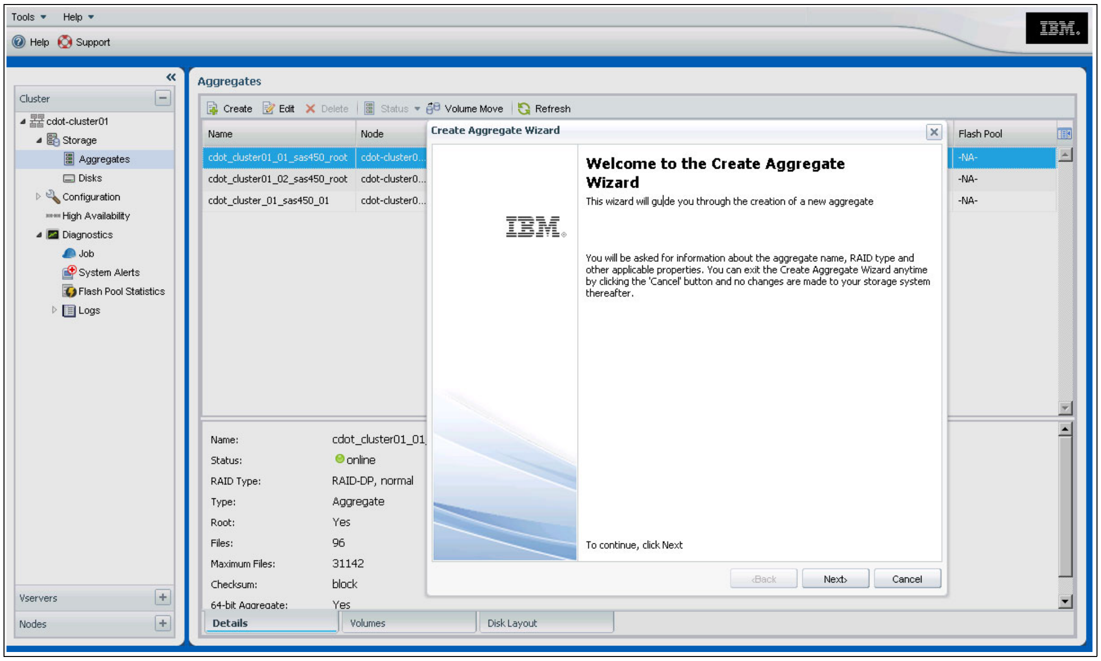
Example 17-28 creates a 64-bit raid-dp aggregate named `cdot_cluster01_01_sas450_01` on node `cdot-cluster01-01` that is composed of 24 SAS disks and has a RAID group size of 24.

Example 17-28 Create aggregate

```
cdot-cluster01::> storage aggregate create -aggregate cdot_cluster_01_sas450_01
-diskcount 24 -disktype SAS -maxraidsize 24 -raidtype raid_dp -nodes
cdot-cluster01-01
```

You can also use the IBM N series OnCommand System Manager to create aggregates. See Figure 17-6 for an example:

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Cluster** → **Storage** → **Aggregates.**
3. Click the **Create** button to start the Create Aggregate Wizard.



Name	Node	Flash Pool
cdot_cluster01_01_sas450_root	cdot-cluster0...	-NA-
cdot_cluster01_02_sas450_root	cdot-cluster0...	-NA-
cdot_cluster_01_sas450_01	cdot-cluster0...	-NA-

Name: cdot_cluster01_01
Status: online
RAID Type: RAID-DP, normal
Type: Aggregate
Root: Yes
Files: 96
Maximum Files: 31142
Checksum: block
64-bit Aggregate: Yes

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 17-6 Create aggregate wizard in system manager

17.6 Creating an SVM

You can create and configure SVMs with FlexVol volumes fully to start serving data immediately or with minimal configuration to delegate administration to the SVM administrator by using the `vserver setup` command.

By using the `vserver setup` command, which launches a CLI wizard, you can perform the following tasks:

- Creating and configuring an SVM fully
- Creating and configuring an SVM with minimal network configuration
- Configuring existing SVM
- Setting up a network interface
- Provisioning storage by creating volumes
- Configuring services
- Configuring protocols

Example 17-29 shows the creation of an SVM that can provide nfs shares, including a root volume of the SVM, a data volume, and a logical interface (LIF) for client access. You can also use the IBM N series OnCommand System Manager to create the SVM.

Example 17-29 SVM setup CLI

```
cdot-cluster01::> vserver setup
Welcome to the Vserver Setup Wizard, which will lead you through
the steps to create a virtual storage server that serves data to clients.

You can enter the following commands at any time:
"help" or "?" if you want to have a question clarified,
"back" if you want to change your answers to previous questions, and
"exit" if you want to quit the Vserver Setup Wizard. Any changes
you made before typing "exit" will be applied.

You can restart the Vserver Setup Wizard by typing "vserver setup". To accept a
default
or omit a question, do not enter a value.

Vserver Setup wizard creates and configures only data Vservers.
If you want to create a Vserver with Infinite Volume use the vserver create
command.

Step 1. Create a Vserver.
You can type "back", "exit", or "help" at any question.

Enter the Vserver name: vs_nfs_01
Choose the Vserver data protocols to be configured {nfs, cifs, fcp, iscsi, ndmp}:
nfs
Choose the Vserver client services to be configured {ldap, nis, dns}: dns
Enter the Vserver's root volume aggregate {cdot_cluster01_02_sas450_01,
cdot_cluster_01_sas450_01} [cdot_cluster_01_sas450_01]:
Enter the Vserver language setting, or "help" to see all languages [C]: de.UTF-8
Enter the Vserver root volume's security style {mixed, ntfs, unix} [unix]:
Vserver creation might take some time to finish....

Vserver vs_nfs_01 with language set to de.UTF-8 created.  The permitted protocols
are nfs.

Step 2: Create a data volume
You can type "back", "exit", or "help" at any question.

Do you want to create a data volume? {yes, no} [yes]: yes
Enter the volume name [vol1]: vs_04_nfsvol_01
Enter the name of the aggregate to contain this volume
{cdot_cluster01_02_sas450_01, cdot_cluster_01_sas450_01}
[cdot_cluster_01_sas450_01]:
Enter the volume size: 500g
Enter the volume junction path [/vol/vs_04_nfsvol_01]:
It can take up to a minute to create a volume...

Volume vs_04_nfsvol_01 of size 500GB created on aggregate
cdot_cluster_01_sas450_01 successfully.
Do you want to create an additional data volume? {yes, no} [no]: no

Step 3: Create a logical interface.
You can type "back", "exit", or "help" at any question.

Do you want to create a logical interface? {yes, no} [yes]:
```

```
Enter the LIF name [lif1]: cdot_cluster_01_01_lif_01
Which protocols can use this interface {nfs, cifs, iscsi}: nfs
Enter the home node {cdot-cluster01-02, cdot-cluster01-01} [cdot-cluster01-01]:
Enter the home port {e0a, e0b} [e0a]:
Enter the IP address: 9.155.66.26
Enter the network mask: 255.255.255.0
Enter the default gateway IP address: 9.155.66.1

LIF cdot_cluster_01_01_lif_01 on node cdot-cluster01-01, on port e0a with IP
address 9.155.66.26 was created.
Do you want to create an additional LIF now? {yes, no} [no]: no

Step 4: Configure DNS (Domain Name Service).
You can type "back", "exit", or "help" at any question.

Do you want to configure DNS? {yes, no} [yes]:
Enter the comma separated DNS domain names: nseries.local
Enter the comma separated DNS server IP addresses: 9.155.113.200

DNS for Vserver vs_nfs_01 is configured.

Step 5: Configure NFS.
You can type "back", "exit", or "help" at any question.

NFS configuration for Vserver vs_nfs_01 created successfully.

Vserver vs_nfs_01, with protocol(s) nfs, and service(s) dns has been configured
successfully.
```

Note: After the SVM is created, we advise to protect the SVM root volume with a load-sharing mirror (17.6.3, “Protecting the SVM root volume” on page 309), assign an appropriate failover group to the LIF (17.6.1, “Assigning a failover group” on page 306), and modify the export policy (17.6.2, “Assigning an export policy” on page 307) in order to restrict client access.

17.6.1 Assigning a failover group

Every newly created LIFs has the `system-defined` failover group assigned, which might not be the correct one according to the network topology. If custom failover groups have been created they have to be assigned to the appropriate LIF.

Example 17-30 shows how to use the **`network interface`** command to assign a failover group to an LIF and how to check if is has been assigned correctly.

Example 17-30 Assign failover group

```
cdot-cluster01::> network interface modify -vserver vs_nfs_01 -lif
cdot_cluster_01_01_lif_01 -failover-group storage-lan

cdot-cluster01::> network interface show -fields lif,failover-group -role data
vserver    lif                 failover-group
---------- ------------------- --------------
vs_cifs_01 vs_cifs_01_cifs_lif1 storage-lan
```

17.6.2 Assigning an export policy

You can use export policies to restrict NFS access to volumes to clients that match specific parameters.

Export policies contain one or more export rules that process each client access request. The result of the process determines whether the client is denied or granted access and what level of access. An export policy with export rules must exist on an SVM for clients to access data.

You associate exactly one export policy with each volume to configure client access to the volume. An SVM can contain multiple export policies. This enables you to do the following tasks for SVMs with multiple volumes:

- Assign different export policies to each volume of an SVM for individual client access control to each volume in the SVM.
- Assign the same export policy to multiple volumes of an SVM for identical client access control without having to create a new export policy for each volume.

If a client makes an access request that is not permitted by the applicable export policy, the request fails with a permission-denied message. If a client does not match any rule in the volume's export policy, then access is denied. If an export policy is empty, then all accesses are implicitly denied.

See the man page for the **`export policy`** command for a brief description about the parameters.

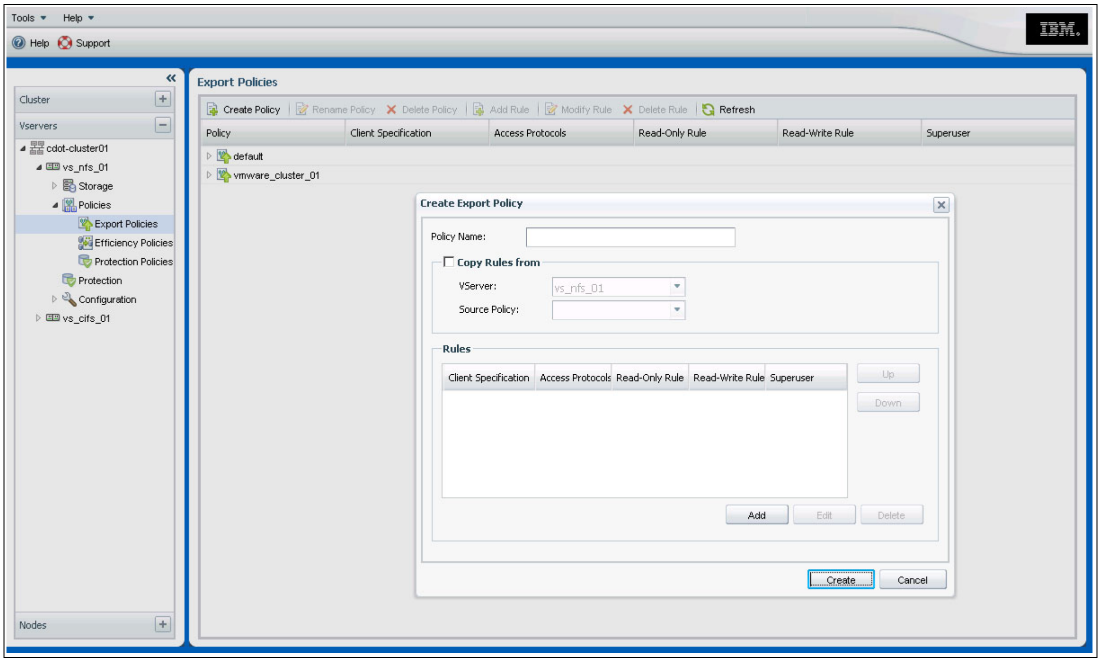
Example 17-31 creates an export policy, adds a rule to it that allows read and write access to a subnet and then assigns it to the according volume.

Example 17-31 Create export policy cli

```
cdot-cluster01::> export-policy create -vserver vs_nfs_01 -policyname
vmware_cluster_01
cdot-cluster01::> export-policy rule create -vserver vs_nfs_01 -policyname
vmware_cluster_01 -clientmatch 9.155.0.0/16 -protocol nfs -rorule any -rwrule any
cdot-cluster01::> vol modify -vserver vs_nfs_01 -volume vs_04_nfsvol_01 -policy
vmware_cluster_01
```

You can also use the IBM N series OnCommand System Manager to create and assign export policies.

1. From the **Home** tab, double-click the appropriate storage system.
2. In the navigation pane, click **Vservers** → **Select the Vserver** → **Policies** → **Export Policies.**
3. Click the **Create Policy** button to create a new policy and add a rule to it. See Figure 17-7 for an example.



Policy	Client Specification	Access Protocols	Read-Only Rule	Read-Write Rule	Superuser
default					
vmware_cluster_01					

Policy Name:

□ Copy Rules from

VServer: vs_nfs_01

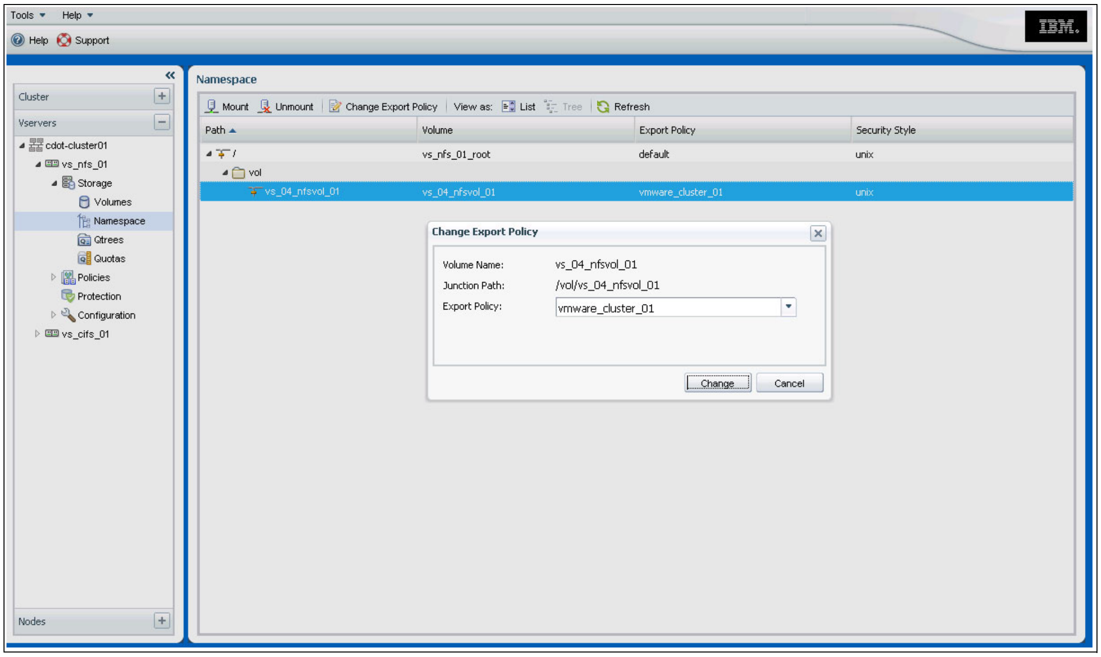
Source Policy:

Client Specification	Access Protocols	Read-Only Rule	Read-Write Rule	Superuser

Figure 17-7 Create export policy in system manager

4. After the policy is created, go back to the navigation pane and click **Vservers** → **Select the Vserver** → **Storage** → **Namespace.**

5. Select the volume you want to change and after that, click **Change Export Policy.**See Figure 17-8 on page 308 for an example.



Path	Volume	Export Policy	Security Style
/	vs_nfs_01_root	default	unix
vol			
vs_04_nfsvol_01	vs_04_nfsvol_01	vmware_cluster_01	unix

Volume Name: vs_04_nfsvol_01

Junction Path: /vol/vs_04_nfsvol_01

Export Policy: vmware_cluster_01

Figure 17-8 Assign export policy in system manager

17.6.3 Protecting the SVM root volume

To protect the SVM namespace root volume, you can create a load-sharing mirror volume on every node in the cluster, including the node in which the root volume is located. Then you create a mirror relationship to each load-sharing mirror volume, and initialize the set of load-sharing mirror volumes.

Example 17-32 shows the following steps:

1. Creation of a FlexVol and designating it as a load-sharing mirror with the **vol create** command.
2. Creation of the load-sharing mirror relationship with the **snapmirror create** command.
3. Initialization of the load-sharing mirror relationship with the **snapmirror initialize** command.

For further information, see the man page for each command.

Example 17-32 Create load-sharing mirrors to protect SVM root volume

```
cdot-cluster01::> vol create -vserver vs_nfs_01 -volume vs_nfs_01_ls01 -aggregate
cdot_cluster01_01_sas450_01 -size 1GB -type DP
cdot-cluster01::> vol create -vserver vs_nfs_01 -volume vs_nfs_01_ls02 -aggregate
cdot_cluster01_02_sas450_01 -size 1GB -type DP
cdot-cluster01::> snapmirror create -source-path vs_nfs_01:rootvol
-destination-path vs_nfs_01:vs_nfs_01_ls01 -type LS -schedule hourly
cdot-cluster01::> snapmirror create -source-path vs_nfs_01:rootvol
-destination-path vs_nfs_01:vs_nfs_01_ls02 -type LS -schedule hourly
cdot-cluster01::> snapmirror initialize-ls-set -source-path vs_nfs_01:rootvol
```

17.7 Post-Installation and verification checklist

After you finished each of the installation steps, go trough the associated checklist and check the status of the points mentioned.

17.7.1 Physical installation

Table 17-4 describes the checks that should be done after the completion of the physical installation.

Table 17-4 Physical installation checklist

Physical Installation	Status
Check and verify that all ordered components were delivered to the customer site.	
Confirm that the N series controllers are properly installed in the cabinets.	
Confirm that there is sufficient airflow and cooling in and around the N series system.	
Confirm that all power connections are secured adequately.	
Confirm that the racks are grounded.	
Confirm that there is sufficient power distribution to N series controllers and expansion units.	

Physical Installation	**Status**
Confirm that power cables are properly arranged in the cabinet.	
Confirm that LEDs and LCDs are displaying the correct information.	
Confirm that cables from N series controllers to expansion units and among expansion units are not crimped or stretched (fiber cable service loops should be bigger than your fist).	
Confirm that fiber cables laid between cabinets are properly connected and are not prone to physical damage.	
Confirm that expansion unit IDs are set correctly.	
If EXN4000: Confirm that Fiber Channel 2Gb/4Gb loop speeds are set correctly on expansion units and proper LC-LC cables are used.	
If SAS Expansion: Confirm that ACP is cabled from e0P.	
Confirm that Ethernet cables are arranged and labeled properly.	
Confirm that at least e0M (SP) and another port (e0a) are cabled.	
Confirm that all fiber cables are arranged and labeled properly.	
Confirm that the Cluster Interconnect Cables are connected (for HA pairs).	
Confirm that there is sufficient space behind the cabinets to perform hardware maintenance.	
Confirm that the Cluster Interconnect switches are properly placed in the cabinet.	
Confirm that the management ports of the Cluster Network switches are cabled and configured.	
Confirm that the latest “Reference Configuration File” for the Cluster Interconnect switches has been installed.	
Confirm that the setup of the cluster network as in 17.3.2, “Cluster network” on page 288 has been completed.	
Confirm that any VLANs or Port-Channels required have been defined to the appropriate switches.	
Power up the disk shelves to ensure that the disks spin up and are initialized properly.	
Connect the console to the serial port cable and establish a console connection using a terminal emulator such as PuTTY.	
Power on the controllers.	

17.7.2 Cluster base

Table 17-5 describes the checks that should be done after the completion of 17.4, “Setting up the cluster and joining nodes” on page 293 and 17.5, “Setting up the cluster base” on page 297.

Table 17-5 Cluster base checklist

Cluster base	Status
Confirm that IBM N series OnCommand System Manager 3.0 is installed on a Windows or Linux system.	
Confirm that you can connect to the service-processor.	
Confirm that all nodes are visible and all licenses are added.	
Confirm that storage-failover and hwassist are configured an working.	
If 2-Node cluster: Confirm that HA is enabled and epsilon is false on all nodes.	
If switchless cluster: Confirm that `network` **`options switchless-cluster show`** is configured.	
Confirm that NTP is configured and working.	
Confirm that DNS is configured and working.	
Confirm that event management is configured.	
Confirm that snmp is configured.	
Confirm that autosupport is configured and a test has been triggered.	
If needed: Confirm that the interface-groups are configured and active.	
If needed: Confirm that VLANs are created.	
Confirm that all used interfaces where put into appropriate failover-groups.	
Confirm that at least one data aggregate has been created.	

17.7.3 SVM checklist

Table 17-6 describes the checks that should be done after the completion of 17.6, “Creating an SVM” on page 304.

Table 17-6 SVM checklist

SVM setup	Status
Confirm that the first SVM has been created.	
Confirm that DNS lookup has been configured.	
Confirm that an LIF has been created and is assigned to appropriate failover-group.	
Confirm that UNIX-users and UNIX-groups have been created.	

18

Non-disruptive operations

Non-Disruptive Upgrade (NDU) began as the process of upgrading Data ONTAP software on the two nodes in an HA pair controller configuration without interrupting I/O to connected client systems.

The overall objective is to enable upgrade and maintenance of the storage system without affecting the system’s ability to respond to foreground I/O requests. This does not mean that there is no interruption to client I/O. Rather, the I/O interruptions are brief enough so that applications continue to operate without the need for downtime, maintenance, or user notification.

The following topics are covered:

- Adding or removing a node
- Software updates

18.1 Adding or removing a node

After a cluster is created, you can add nodes to it or remove nodes from it by using the Cluster Setup Wizard or the CLI. For example, this feature can be used to do hardware refresh or scale out the cluster without the need to bring the services down.

18.1.1 Adding a node

Prior adding nodes to the cluster, the following conditions must be met:

- If you are adding nodes to a multiple-node cluster, more than half of the existing nodes in the cluster must be healthy (indicated by the **`cluster show`** command).
- If you are adding nodes to a two-node cluster, cluster HA must be disabled. See 17.5.1, "Storage failover (SFO)" on page 297 for further information about HA.
- Nodes must be in even numbers so that they can form HA pairs.

See 17.4.2, "Joining a node to the cluster" on page 296 for a brief description about how to add a new node to the existing cluster.

18.1.2 Removing a node

You can remove unwanted nodes from the cluster. You can remove only one node at a time. After you remove a node, you must also remove its failover partner.

The following steps describe the process of removing a node from the cluster:

1. Set the privilege level to **`advanced`** and use the **`cluster unjoin`** command as shown in Example 18-1.

   Example 18-1 Remove node

   ```
   cdot-cluster01::> set advanced
   Warning: These advanced commands are potentially dangerous; use them only when
            directed to do so by IBM personnel.
   Do you want to continue? {y|n}: y

   cdot-cluster01::*> cluster unjoin -node cdot-cluster01-02
   Warning: This command will unjoin node "cdot-cluster01-02" from the cluster.
   You must unjoin the failover partner as well. After the node is successfully
   unjoined, erase its configuration and initialize all disks by using the "Clean
   configuration and initialize all disks (4)" option from the boot menu.
   Do you want to continue? {y|n}:
   ```

   A failure message is generated if you have conditions that you must address before removing the node. For example, the message might indicate that the node has shared resources that you must remove or that the node is in a cluster HA configuration or storage failover configuration that you must disable.

   Note: If a failure message indicates error conditions, address those conditions and rerun the **`cluster unjoin`** command. The node is automatically rebooted after it is successfully removed from the cluster.

2. If the node will rejoin the same cluster or join a new cluster, do the following steps after the node is rebooted:
   a. During the boot process, press Ctrl-C to display the boot menu when prompted to do so.
   b. Select boot menu option **`(4) Clean configuration and initialize all disks`** to erase the node's configuration and initialize all disks.
3. Repeat the preceding steps to remove the failover partner from the cluster.

18.2 Software updates

This section provides an overview of the process for updating Clustered Data ONTAP 8.2 to a newer release, including the preparation steps.

Table 18-1 provides information about what needs to be considered prior to the update.

Table 18-1 Update considerations

Related to	Description
Update	**Volume/aggregate online check** Use the **`vol show -state !online`** command to check if volumes are present that are not online. Volumes that are not online must be onlined before proceeding with the update. If you proceed with offline volumes, that data will be unavailable and the ability to revert the system will be compromised.
Update	**Protocol considerations** In general, services based on stateless protocols - such as NFS, FCP, and iSCSI - are less susceptible to service interruptions during upgrades than session-oriented protocols - such as CIFS and NDMP. Use the **`vserver show -type data  -fields allowed-protocols`** command to check which protocols your SVMs provide.
Update	**Compatibility review** Confirm interoperability of NAS Clients (NFS/SMB), SAN Switch OS/ SAN Clients (iSCSI, FCP, FCOE) Backup/Recovery Applications and Management Software such as SnapManager or OnCommand Products. See the *IBM System Storage N series Interoperability Matrix* for further information.
Update	**Disk firmware update** Disk firmware is bundled with the Data ONTAP system files and updated automatically during Data ONTAP upgrades. By default, disk firmware updates take place automatically in the background, thus ensuring the continuity of storage system services.
Update	**CPU / disk utilization** You must ensure that CPU and disk utilization does not exceed 50% before beginning an update. Use the **`dashboard performance show`** command to obtain information about the CPU usage.
Update	**SP firmware update** SP firmware is bundled with the Data ONTAP system files and updated automatically during Data ONTAP upgrades.
Update	**System health** You must ensure that the nodes in the cluster are healthy before beginning a update. Use the **`system health status show`** and system health alert show to ensure that the system is healthy and there are no alerts.

Related to	Description
Update	**Estimating the duration of the update process** For each HA pair, you should plan for approximately 30 minutes to complete preparatory steps, 60 minutes to perform the upgrade, and 30 minutes to complete post-upgrade steps.

Before starting to update the latest Clustered Data ONTAP release on your cluster, you must perform several cluster health checks, including ensuring that the cluster is running and healthy, verifying that the cluster is in a quorum, and verifying the health of the storage virtual machine (SVM). The steps in Table 18-2 guide you through these checks and through the update process.

Table 18-2 Update plan

Step	Action

1	**Verifying that the cluster is in an RDB quorum**
	Before upgrading, you must ensure that all nodes are participating in a replicated database (RDB) quorum and that all rings are in the quorum. You must also verify that the per-ring quorum master is the same for all nodes:
	1. Log in to the cluster shell, and set the privilege level to advanced:
	`set -privilege advanced`
	2. Enter y to continue.
	3. Display the RDB processes for the management application (mgmt), volume location database (vldb), and virtual-interface manager (vifmgr), and SAN management daemon (bcomd):
	`cluster ring show -unitname vldb`
	`cluster ring show -unitname mgmt`
	`cluster ring show -unitname vifmgr`
	`cluster ring show -unitname bcomd`
	4. For each process, verify the following configuration:
	- The relational database epoch and database epochs match for each node.
	- The per-ring quorum master is the same for all nodes.
	Note that each ring might have a different quorum master.
	5. Return to the admin privilege level:
	`set -privilege admin`

Step	Action
2	**Verifying cluster and SVM health** Before and after you update, you should verify that the nodes are healthy and eligible to participate in the cluster, and the aggregates and volumes are online: 1. Verify that the nodes in the cluster are online and are eligible to participate in the cluster: **`cluster show`** 2. Verify that all aggregates are online and display the state of physical and logical storage, including storage aggregates: **`storage aggregate show -state !online`** This command displays the aggregates that are not online. 3. To verify that all volumes are online, display any volumes not online by entering the following command: **`volume show -state !online`** 4. To verify that all cluster SVMs are up and running, enter the following command: **`vserver show`**
3	**Monitor CPU and disk utilization** Before updating Clustered Data ONTAP, monitor CPU and disk utilization for 30 seconds by entering the following command at the console of each storage controller: **`node run -node <nodename> -command sysstat -c 10 -x 3`** The values in the CPU and Disk Util columns are strongly advised not to exceed 50% for all ten measurements reported. Ensure that no additional load is added to the storage system until the update completes.
4	**Enabling and reverting LIFs to home ports** During a reboot, some logical interfaces (LIFs) might have been migrated to their assigned failover ports. Before and after you upgrade, revert, or downgrade a cluster, you must enable and revert any LIFs that are not on their home ports. The **`net interface revert`** command reverts an LIF that is not currently on its home port back to its home port, provided that the home port is operational: 1. Display all LIFs that are currently not on its home port: **`net interface show -is-home false`** 2. If any LIFs appear with an `Is home` status of false, continue with the next step. 3. Revert LIFs to their home ports: **`net interface revert *`** 4. Verify that all LIFs are in their home ports: **`net interface show -is-home false`** The output of this command should show up no LIFs.

Step	Action
5	**Verifying the LIF failover configuration** Before you perform an upgrade, use the `network interface show` command to verify the LIF failover configuration: 1. Display the failover policy for each data port: `network interface show -role data -failover` 2. For each LIF, verify that the `Failover Targets` field includes data ports from a different node that will remain up while the LIFs home node is being upgraded. Verify also that the `Failover Targets` are ports that are capable to host the LIF that might fail over to them.

6	**Verifying the system time**
	You should verify that NTP is configured, and that the time is synchronized across the cluster:
	1. Verify that each node is associated with an NTP server:
	`system services ntp server show`
	2. Verify that each node has the same date and time:
	`system node date show`

7	**Installing Clustered Data ONTAP 8.2 software images in the cluster** You must copy the software image (for example 82P4_q_image.tgz) from the IBM fix central site to an HTTP server on your network so that nodes can access the images: 1. Download and install the software image in the same operation: `system node image update -node * -package http://9.155.66.130/82P4_q_image.tgz -replace-package true` This command installs the software image on each node one at a time. 2. To install the image on each node simultaneously, set the `-background` parameter to `true`.
8	**How Clustered Data ONTAP software images are stored and alternated in the cluster** Each node in the cluster can hold two Data ONTAP software images, the current image that is running, and an alternate image that you can boot. You can view the software images on each node in the cluster by using the command: `system node image show`

Step	Action

9	**Ensuring that no jobs are running**
	You must verify the status of cluster jobs before updating to a different Clustered Data ONTAP release. If any aggregate, volume, NDMP (dump or restore), SnapMirror copy, or Snapshot jobs are running or queued (such as `create`, `delete`, `move`, `modify`, `replicate`, and `mount` jobs), allow the jobs to finish successfully or stop the queued entries:
	1. Review the list of any running or queued aggregate, volume, NDMP (dump or restore), SnapMirror copy, or Snapshot jobs:
	`job show`
	2. Delete any running or queued aggregate, volume, SnapMirror copy, or Snapshot jobs:
	`job delete *`
	3. If jobs are displayed in the `Queued` or `Dormant` states, you might need to delete them individually by using the Job ID listed in the job show command output:
	`job delete -id job_id`

10	**Verifying that the cluster is ready to be upgraded**
	You must verify that the target Data ONTAP software is installed, storage failover is enabled, and if necessary, cluster HA is enabled:
	1. Verify that the Data ONTAP 8.2 software is installed:
	`system node image show`
	2. Set the new Clustered Data ONTAP 8.2 software image to be the default image:
	`system image modify {-node * -iscurrent false} -isdefault true`
	3. Verify that the Clustered Data ONTAP 8.2 software image is set as the default image:
	`system node image show`
	4. Verify that storage failover is enabled and possible:
	`storage failover show`

Step	Action

11	**Upgrading a Data ONTAP cluster non-disruptively by using the rolling upgrade method**
	Start by upgrading the first node in an HA pair:
	1. Disable automatic giveback on both nodes of the HA pair if it is enabled by entering the following command on each node:
	`storage failover modify -node <nodename> -auto-giveback false`
	2. Verify that automatic giveback is disabled for both nodes:
	`storage failover show -auto-giveback false`
	3. Migrate LIFs away from the node:
	`network interface migrate-all -node <nodenameA>`
	Data LIFs for SAN protocols are not migrated. As long as these LIFs exist on each node in the cluster, data can be served through alternate paths during the upgrade process.
	If you are connected to the cluster through the cluster management LIF, and if this node currently hosts the cluster management LIF, then your SSH session will be temporarily disconnected while the LIF is migrated.
	4. Verify that the LIFs migrated to the proper ports on the node's partner:
	`network interface show -role data -curr-node <nodenameB>`
	5. Trigger an AutoSupport notification:
	`system node autosupport invoke -node <nodenameA> -type all -message "starting_update"`
	This AutoSupport notification includes a record of the system status just prior to upgrade. It saves useful troubleshooting information in case there is a problem with the upgrade process.
	6. Initiate a takeover:
	`storage failover takeover -bynode <nodenameB>`
	Do not specify the parameter `-option immediate`, because a normal takeover is required for the node that is being taken over to boot onto the new software image. The first node boots up to the “Waiting for giveback” state.
	7. Verify that the takeover was successful:
	`storage failover show`
	8. Wait 8 minutes to ensure that client multipathing (if deployed) is stabilized and clients are recovered from the pause in I/O that occurs during takeover.
	The recovery time is client-specific and may take longer than 8 minutes depending on the characteristics of the client applications.
	9. Return the aggregates to the first node:
	`storage failover giveback -fromnode <nodenameB>`

Step | **Action**

Attention: If the giveback is not initiated, an error message is returned, and an event is generated if any conditions such as these are detected:
- Long-running operations (such as ASUP generation)
- Operations that cannot be restarted (such as aggregate creation)
- Error conditions (such as a disk connectivity mismatch between the nodes)

If giveback is not initiated, address the “veto” condition described in the error message, ensuring that any identified operations are terminated gracefully. After that, reenter the giveback command:

```
storage failover giveback -fromnode <nodenameB>
```

Alternatively, you can analyze the messages and events for relevance in your environment.

If you determine that the veto conditions are not significant, you can override the giveback veto by entering the following command:

```
storage failover giveback -fromnode <nodenameB> -override-vetoes true
```

10. Verify that all aggregates have been returned:

    ```
    storage failover show-giveback
    ```

    If the Giveback Status field indicates that the node that was taken over is in partial giveback, then complete the following actions before proceeding:

    Determine which aggregates have been returned:

    ```
    storage aggregate show -node <nodenameA>
    ```

    Check EMS logs for errors and take corrective action:

    ```
    event log show
    ```

    Repeat the `storage aggregate show command` to verify any corrections.

11. Revert the LIFs that were migrated away back to the node:

    ```
    network interface revert *
    ```

12. Verify that data is being served to clients:

    ```
    network interface show
    network port show
    ```

13. Trigger an AutoSupport notification:

    ```
    system node autosupport invoke -node <nodenameA> -type all -message
    "finishing_NDU"
    ```

14. Upgrade the partner node in an HA pair.

    After upgrading the first node in a HA pair, you upgrade its partner by initiating a takeover on it. The first node serves the partner's data while the partner node is upgraded.

15. Migrate LIFs away from the node:

    ```
    network interface migrate-all -node <nodenameB>
    ```

16. Verify that the LIFs migrated to the proper ports on the node's partner:

    ```
    network interface show -data-protocol nfs -role data -curr-node <nodenameA>
    ```

17. Trigger an AutoSupport notification:

    ```
    system node autosupport invoke -node <nodenameB> -type all -message
    "starting_NDU"
    ```

Step	Action

	18. Initiate a takeover:
	`storage failover takeover -bynode <nodenameA>`
	19. Verify that the takeover was successful:
	`storage failover show`
	20. Wait 8 minutes to ensure that client multipathing (if deployed) is stabilized and clients are recovered from the pause in I/O that occurs during takeover.
	21. Return the aggregates to the partner node:
	`storage failover giveback -fromnode <nodenameA>`
	Attention: See step 9 on page 320 and 10 on page 321 about how to check if all aggregates were returned to the partner node and which steps to take regarding corrective actions.
	22. Revert the LIFs that were migrated away back to the node:
	`network interface revert *`
	23. Verify that data is being served to clients:
	`network interface show`
	`network port show`
	24. Trigger an AutoSupport notification:
	`system node autosupport invoke -node <nodenameA> -type all -message "finishing_NDU"`
	25. Confirm that the new Data ONTAP 8.2 software is running on both nodes of the HA pair:
	`system node image show`
	26. Re-enable automatic giveback on both nodes if it was previously disabled:
	`storage failover modify -node <nodename> -auto-giveback true`

Step	Action
12	**Post upgrade cluster health verification** You should verify that the nodes in the cluster are online and healthy, determine the health of any SVMs, identify any storage aggregates that are not online and healthy, and identify any LIFs that are not on their home servers: 1. Set the privilege level to advanced: **`set advanced`** 2. Ensure that upgrade status is complete for each node by running the following command: **`system node upgrade-revert show`** The status for each node should be listed as complete. If the status for any node is not successful, run the following command: **`system node upgrade-revert upgrade`** If this command does not complete the node's upgrade, contact technical support immediately. 3. Return to the admin privilege level: **`set -privilege admin`** 4. Verify that the nodes in the cluster are online and are eligible to participate in the cluster: **`cluster show`** If any nodes are not online or not eligible to participate in the cluster, resolve the issue before continuing. 5. Display information about any storage aggregates that have a state other than online: **`storage aggregate show -state !online`** If any storage aggregates appear with a status other than online, note the issue and continue. 6. Display information about any volumes that have a state other than online: **`volume show -state !online`** If any volumes appear with a status other than online, resolve the issue before continuing. 7. Enabling and reverting LIFs to home ports: **`network interface show`** If any LIFs appear with a `Status Admin` status of `down` or with an `Is home` status of `false`, continue with the next command: **`network interface modify -lif * -status-admin up`** 8. Verify that all LIFs are on their home ports and active: **`network interface show`**

19

Command Line Interface (CLI)

This chapter introduces various ways to administer N series systems through the command line interface (CLI).

The following topics are covered:

- Introduction to CLI administration
- New features in Clustered Data ONTAP CLI
- Audit logging
- Accessing the cluster by using SSH
- Enabling Telnet or RSH access
- 7-Mode to Clustered Data ONTAP

19.1 Introduction to CLI administration

The CLI provides a command-based mechanism that is similar to the UNIX tcsh shell. On storage systems shipped with Clustered Data ONTAP 8.2 and later, secure protocols are enabled and non-secure protocols are disabled by default:

- Secure protocols (SSH v2 only) are enabled by default.
- Non-secure protocols (including RSH, Telnet) are disabled by default.

We advise using only the SSH Version 2 protocol and using SSH public key authentication. SSH public keys provide a stronger and more granular method of SSH access to N series storage systems.

The cluster has three different shells for CLI commands, the *clustershell*, the *nodeshell*, and the *systemshell*. Depending on the task you perform, you might need to use different shells to execute different commands:

- The clustershell is the native shell that is started automatically when you log in to the cluster. It provides all the commands you need to configure and manage the cluster.

  The clustershell CLI help (triggered by **`?`** at the clustershell prompt) displays available clustershell commands. The **`man command_name`** command in the clustershell displays the man page for the specified clustershell command.
- The nodeshell is a special shell containing a subset of commands that are available in 7-Mode Data ONTAP systems that take effect only at the node level. The nodeshell is accessible through the **`system node run`** command.

  The nodeshell CLI help (triggered by **`?`** or **`help`** at the nodeshell prompt) displays available nodeshell commands. The **`man command_name`** command in the nodeshell displays the man page for the specified nodeshell command.
- The systemshell is a low-level shell that is used only for diagnostic and troubleshooting purposes.

  The systemshell is not intended for general administrative purposes. You access the systemshell only with guidance from technical support.

19.2 New features in Clustered Data ONTAP CLI

Starting with Clustered Data ONTAP, a new CLI was introduced. Commands in the CLI are now organized into a hierarchy by command directories. You can run commands in the hierarchy either by entering the full command path or by navigating through the directory structure.

You can use the **`top`** command to go to the top level of the command hierarchy, and the **`up`** command or **`..`** command to go up one level in the command hierarchy.

It is also possible to abbreviate commands by entering only the minimum number of letters in a command that makes the command unique. For example, to abbreviate the command **`network interface show`**, you can enter **`n i show`**.

The CLI is also capable of command-line completion. You can type the first few characters of a command and press the completion key *(Tab key)* to fill out the rest of the command. In case of multiple possible completions, the CLI will list all possible parameters beginning with the specified characters. You can type more characters and press the *Tab key* again to complete the command or display possible parameters, including defaults, until you have built up the whole command.

Press the **?** key at any time to get context-sensitive help.

Command history

Each CLI session keeps a history of all commands issued in it. You can view the command history of the session that you are currently in. You can also reissue commands. To view the command history, you can use the **history** command.

To reissue a command, you can use the **redo** command with one of the following arguments:

- A string that matches part of a previous command. For example, if the only volume command you have run is **volume show**, you can use the **redo volume** command to reexecute the command.
- The ID of a previous command, as listed by the **history** command. For example, you can use the **redo 4** command to reissue the fourth command in the history list.
- A negative offset from the end of the history list. For example, you can use the **redo -2** command to reissue the command that you ran two commands ago.

Query operators

The management interface supports queries, UNIX-style patterns and wildcards to match multiple values in command-parameter arguments.

See Table 19-1, which describes the possible query operators.

Table 19-1 Query operators

Operator	Description
*	Wildcard that matches all entries. For example, the command **volume show -volume *backup*** displays a list of all volumes whose names include the string **backup**.
!	NOT operator. For example, the command **volume show -state !online** displays a list of all volumes which are in a state that does not match the value **online**.
|	OR operator. Separates two values that are to be compared. For example, the command **vserver show -vserver *cifs*|vs_*** displays a list of all SVMs that contain the value **cifs** or begin with **vs_**.
..	Range operator. For example, the command **vol show -size 1GB..10GB** displays all volumes that have a size between and including 1 GB and 10 GB.
<	Less-than operator. For example, the command **vol show -size <10GB** displays all volumes smaller than 10 GB.
>	Greater-than operator. For example, the command **vol show -size >10GB** displays all volumes bigger than 10 GB.
<=	Less-than-or-equal-to operator. For example, the command **vol show -size <=10GB** displays all volumes smaller or equal than 10 GB.
>=	Greater-than-or-equal-to operator. For example, the command **vol show -size >=10GB** displays all volumes bigger or equal than 10 GB.

Tip: You can use multiple query operators in one command line. For example, the command **`volume show -size >10GB -percent-used >80 -vserver vs_cifs*`** displays all volumes that are greater than 10 GB in size, more than 80% utilized, and in a storage virtual machine (SVM) whose name begins with "vs_cifs."

Extended queries

You can use extended queries to match and perform operations on objects that have specified values.

You specify extended queries by enclosing them within curly brackets (**`{}`**). An extended query must be specified as the first argument after the command name, before any other parameters.

Example 19-1 sets offline all volumes whose names include the string **`temp`**.

Example 19-1 Extended query example 1

```
cluster01::> volume modify {-volume *temp*} -state offline
```

Example 19-2 enables compression on all volumes where it is disabled.

Example 19-2 Extended query example 2

```
cluster01::> sis modify {-compression false} -compression true
```

19.3 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. Administrative HTTP operations, such as those resulting from the use of System Manager or ONTAP API application, are logged. All login attempts to access the storage system, with success or failure, are also audit logged.

Clustered Data ONTAP enables you to audit two types of requests, set requests and get requests. A set request typically applies to non-display commands, such as creating, modifying, or deleting an object. A get request occurs when information is retrieved and displayed to a management interface. This is the type of request that is issued when you run a **`show`** command, for example.

You can configure auditing with the **`security audit modify`** command. By default, Clustered Data ONTAP is configured to save an audit log. The audit log data is stored in the /**`etc/log/mlog`** directory of every node in the files shown in Table 19-2.

Table 19-2 Audit log files

file	explanation
mgwd.log	Depending on the **security audit** settings, this file can include the following audit information: ► Set requests for the CLI (audited by default) ► Set requests for the ONTAP API (audited by default) ► Get requests for the CLI ► Get requests for the ONTAP API The file also includes information about login attempts or failures. For more information, see the man pages for the **security audit** commands.
command-history.log	Regardless of the settings for the **security audit** commands, set requests are always recorded in this file. This file is included in the AutoSupport to the specified recipients.

The **command-history.log** and **mgwd.log** files are rotated when they reach 100 MB in size, and their previous 34 copies are preserved (with a maximum total of 35 files, respectively).

You can display the content of the **/etc/log/mlog** directory by using a web browser if your cluster user account and the required web services have been configured for the access.

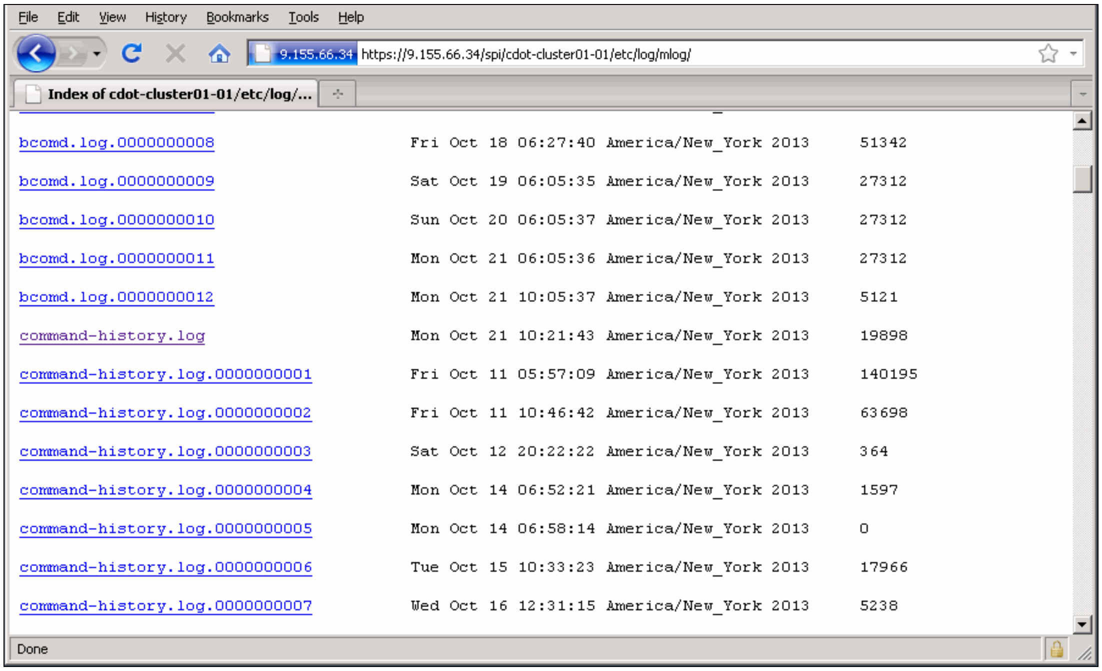
Example 19-3 shows how to create a user with read-only permissions to the web services and enable the web services so that the user is able to browse and download files from the **/etc/log/mlog** directory with a browser.

Example 19-3 Enable spi webservice

```
cdot-cluster01::> security login create -username webuser -application http
-authmethod password -role readonly -vserver cdot-cluster01
Please enter a password for user 'webuser':
Please enter it again:
cdot-cluster01::> vserver services web modify -vserver cdot-cluster01 -name spi
-enabled true
cdot-cluster01::> vserver services web access create -vserver cdot-cluster01 -name
spi -role readonly
```

After the web service has been enabled, it is possible to browse the directory with the http(s) link as shown in Figure 19-1:

```
http(s)://cluster-mgmt-ip/spi/node-name/etc/log/mlog
```


File Edit View History Bookmarks Tools Help
9.155.66.34 https://9.155.66.34/spi/cdot-cluster01-01/etc/log/mlog/
Index of cdot-cluster01-01/etc/log/...
Done

```
bcomd.log.0000000008           Fri Oct 18 06:27:40 America/New_York 2013    51342
bcomd.log.0000000009           Sat Oct 19 06:05:35 America/New_York 2013    27312
bcomd.log.0000000010           Sun Oct 20 06:05:37 America/New_York 2013    27312
bcomd.log.0000000011           Mon Oct 21 06:05:36 America/New_York 2013    27312
bcomd.log.0000000012           Mon Oct 21 10:05:37 America/New_York 2013    5121
command-history.log            Mon Oct 21 10:21:43 America/New_York 2013    19898
command-history.log.0000000001 Fri Oct 11 05:57:09 America/New_York 2013    140195
command-history.log.0000000002 Fri Oct 11 10:46:42 America/New_York 2013    63698
command-history.log.0000000003 Sat Oct 12 20:22:22 America/New_York 2013    364
command-history.log.0000000004 Mon Oct 14 06:52:21 America/New_York 2013    1597
command-history.log.0000000005 Mon Oct 14 06:58:14 America/New_York 2013    0
command-history.log.0000000006 Tue Oct 15 10:33:23 America/New_York 2013    17966
command-history.log.0000000007 Wed Oct 16 12:31:15 America/New_York 2013    5238
```

Figure 19-1 Browse /etc/log/mlog directory

19.4 Accessing the cluster by using SSH

You can issue SSH requests to the cluster to perform administrative tasks. SSH is enabled by default.

Take note of the following information regarding SSH access:

- You must have a user account that is configured to use `ssh` as an access method. See the **`-application`** parameter in **`man security login`** for more information.
- The Clustered Data ONTAP 8.2 release family supports OpenSSH client version 5.4p1 and OpenSSH server version 5.4p1.
- Only the SSH v2 protocol is supported; SSH v1 is not supported.
- Clustered Data ONTAP supports a maximum of 64 concurrent SSH sessions per node.
- If the cluster management logical interface (LIF) resides on the node, it shares this limit with the node management LIF.
- If the rate of in-coming connections is higher than 10 per second, the service is temporarily disabled for 60 seconds.
- Clustered Data ONTAP supports only the AES and 3DES encryption algorithms (also known as ciphers) for SSH.
- If you want to access the Clustered Data ONTAP CLI from a Windows host, you can use a third-party utility such as PuTTY.

From an administration host, enter the **ssh** command in one of the following formats to access the cluster:

ssh username@hostname_or_IP [command]

ssh -l username hostname_or_IP [command]

If you are using an AD domain user account, you must specify username in the format of domainname\\AD_accountname (with double backslashes after the domain name) or "domainname\AD_accountname" (enclosed in double quotation marks and with a single backslash after the domain name). This does not apply for a session from a Windows host with PuTTY. Use the username as it was specified when the user was created.

> **Note:** If you use an Active Directory (AD) domain user account to access the cluster, an authentication tunnel for the cluster must have been set up through a CIFS-enabled SVM, and your AD domain user account must also have been added to the cluster with ssh as an access method and domain as the authentication method. Only one SVM is allowed to be used as a tunnel. See the man page of **security login domain-tunnel** for more information.

Note that hostname_or_IP is the host name or the IP address of the cluster management LIF or a node management LIF. Use of the cluster management LIF is advised. Use of command is not required for SSH-interactive sessions.

Example 19-4 shows how the user account named "admin" can issue an SSH request to access a cluster whose cluster management LIF is 9.155.66.26.

Example 19-4 SSH login to clustershell

```
$ ssh -l admin 9.155.66.26 date
Password:
Node      Date                     Time zone
--------- ------------------------ -------------------------
cdot-01   Tue Oct 22 11:25:03 2013 Europe/Berlin
cdot-02   Tue Oct 22 11:25:03 2013 Europe/Berlin
2 entries were displayed.

$ ssh -l admin 9.155.66.26
Password:
cdot::> cluster show
Node                 Health  Eligibility
-------------------- ------- ------------
cdot-01              true    true
cdot-02              true    true
2 entries were displayed.
```

19.5 Enabling Telnet or RSH access

Telnet and RSH are disabled in the predefined management firewall policy (`mgmt`). To enable the cluster to accept Telnet or RSH requests, you must create a new management firewall policy that has Telnet or RSH enabled and then associate the new policy with the cluster management LIF.

See the following steps for information about creating a new firewall policy and assigning it the cluster management LIF:

1. Use the **`system services firewall policy clone`** command to create a new management firewall policy based on the mgmt management firewall policy as shown in Example 19-5.

   Example 19-5 Clone firewall policy

   ```
   cluster01::> system services firewall policy clone -policy mgmt
   -new-policy-name mgmt_rsh_telnet
   ```

2. Use the **`system services firewall policy create`** command to enable Telnet or RSH in the new management firewall policy as shown in Example 19-6.

   Example 19-6 Enable Telnet and RSH

   ```
   cluster01::> system services firewall policy create -policy mgmt_rsh_telnet
   -service telnet -action allow -ip-list 0.0.0.0/0

   cluster01::> system services firewall policy create -policy mgmt_rsh_telnet
   -service rsh -action allow -ip-list 0.0.0.0/0
   ```

3. Use the **`network interface modify`** command to associate the new policy with the cluster management LIF as shown in Example 19-7.

   Example 19-7 Modify management LIF

   ```
   cluster01::> network interface modify -vserver cluster01
   -lif cluster_mgmt -firewall-policy mgmt_rsh_telnet
   ```

Note: Telnet and RSH are not supported on the service-processor (SP) or remote-lan-module (RLM). Enabling the protocols does not have an effect on SP or RLM.

19.5.1 Accessing the cluster by using Telnet

You can issue Telnet requests to the cluster to perform administrative tasks. Telnet is disabled by default.

Consider the following information regarding SSH access:

- You must have a cluster local user account that is configured to use Telnet as an access method. See the **`-application`** parameter in **`man security login`** for more information.
- Telnet must already be enabled in the management firewall policy that is used by the cluster or node management LIFs. See 19.5, “Enabling Telnet or RSH access” on page 332 for more information.
- Clustered Data ONTAP supports a maximum of 50 concurrent Telnet sessions per node.

- If the cluster management LIF resides on the node, it shares this limit with the node management LIF.
- If the rate of in-coming connections is higher than 10 per second, the service is temporarily disabled for 60 seconds.
- If you want to access the Clustered Data ONTAP CLI from a Windows host, you can use a third-party utility such as PuTTY.

From an administration host, enter the **`telnet`** command in one of the following formats to access the cluster:

```
telnet username@hostname_or_IP
```

Note that `hostname_or_IP` is the host name or the IP address of the cluster management LIF or a node management LIF. Using the cluster management LIF is advised.

> **Note:** Telnet is not a secure protocol. 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.

19.5.2 Accessing the cluster by using RSH

You can issue Telnet requests to the cluster to perform administrative tasks. Telnet is disabled by default.

Consider the following information regarding SSH access:

- You must have a cluster local user account that is configured to use Telnet as an access method. See the **`-application`** parameter in **`man security login`** for more information.
- RSH must already be enabled in the management firewall policy that is used by the cluster or node management LIFs. See 19.5, “Enabling Telnet or RSH access” on page 332 for more information.
- Clustered Data ONTAP supports a maximum of 50 concurrent RSH sessions per node.
- If the cluster management LIF resides on the node, it shares this limit with the node management LIF.
- If the rate of in-coming connections is higher than 10 per second, the service is temporarily disabled for 60 seconds.

From an administration host, enter the **`rsh`** command in one of the following formats to access the cluster:

```
rsh username@hostname_or_IP command
```

Note that `hostname_or_IP` is the host name or the IP address of the cluster management LIF or a node management LIF. Using the cluster management LIF is advised. Rather, **`command`** is the command you want to execute over RSH.

> **Note:** RSH is not a secure protocol. 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.

19.6 7-Mode to Clustered Data ONTAP

If you are moving from Data ONTAP running in 7-Mode to Clustered Data ONTAP, this chapter can provide you information about equivalents in Clustered Data ONTAP regarding some 7-Mode commands and configuration files.

19.6.1 Configuration files

In Data ONTAP operating in 7-Mode, you typically use flat files to configure the storage system. In Clustered Data ONTAP, you use configuration commands. Table 19-3 shows how 7-Mode configuration files map to Clustered Data ONTAP configuration commands.

Table 19-3 Configuration files map to Clustered Data ONTAP commands

7-Mode configuration file	Clustered Data ONTAP configuration command
`/etc/cifs_homedir.cfg`	`vserver cifs home-directory search-path`
`/etc/exports`	`vserver export-policy`
`/etc/hosts`	`vserver services dns hosts`
`/etc/hosts.equiv`	Not applicable. Use **`security login`** commands to create user access profiles.
`/etc/messages`	`event log show`
`/etc/nsswitch.conf`	`vserver modify -ns-switch	-nm-switch`
`/etc/rc`	In Clustered Data ONTAP, the retention of node configuration information processed at boot is transferred to other internal files that retain the configuration information. This contrasts with Data ONTAP operating in 7-Mode, in which features configured in memory are also retained in the **`/etc/rc`** file to be replayed at boot and reconfigured.
`/etc/quotas`	`volume quota`
`/etc/resolv.conf`	vserver services dns modify
`/etc/snapmirror.allow`	Intercluster relationships exist between two clusters. Intracluster relationships exist between two nodes on the same cluster. Authentication of the remote cluster occurs during the creation of the cluster peering relationship. Intracluster **`snapmirror create`** can be performed only by the cluster administrator to enforce per Vserver security.
`/etc/snapmirror.conf`	`snapmirror create`
`/etc/symlink.translations`	`vserver cifs symlink`
`/etc/usermap.cfg`	`vserver name-mapping create`

19.6.2 Commands

You can use Table 19-4 to determine the Clustered Data ONTAP equivalents of some useful 7-Mode commands.

Table 19-4 7-Mode commands map to Clustered Data ONTAP commands

7-Mode command	Clustered Data ONTAP command
aggr status	storage aggregate show
aggr show_space	system node run -node *nodename* aggr show_space
aggr status -f	disk show -broken
cf status	storage failover show
cf takeover	storage failover takeover -ofnode *nodename*
cf giveback	storage failover giveback -ofnode *nodename*
cifs domaininfo	vserver cifs domain discovered-servers show
cifs resetdc	vserver cifs domain discovered-servers reset-servers
disk zero spares	storage disk zerospares
environment status	system node environment sensors show
exportfs	vserver export-policy
halt	system node halt -node *nodename*
ifconfig -a	network interface show or network port show
license	system license show
nfsstat	statistics show -object nfs*
passwd	security login password
ping {*host*}	network ping -node *nodename* -destination {*host*}
sp status	system node service-processor show
sysstat	statistics show-periodic
uptime	node show -node *nodename* -fields uptime
useradmin user list	security login show
version -b	system image show

20

N series OnCommand System Manager 3.0

This chapter describes Version 3.0 of the IBM N series OnCommand System Manager software.

The following topics are covered:

- Introduction to N series OnCommand System Manager
- Installing the N series OnCommand System Manager
- Getting started with OnCommand System Manager

OnCommand 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.
- Set up and manage SnapMirror and SnapVault relationships.
- Verify and configure network configuration settings in the storage systems.
- Perform cluster management, storage node management, and storage virtual machine (SVM) management operations in a cluster environment.
- Create and configure SVMs, manage storage objects associated with an SVM, and manage SVM services.

20.1 Introduction to N series OnCommand System Manager

OnCommand System Manager enables you to manage storage systems and storage objects, such as disks, volumes, and aggregates from a Web browser.

As a cluster administrator, you can use OnCommand System Manager to administer the entire cluster and its resources.

You can download and install OnCommand System Manager on a desktop or notebook that is running a Windows or a Linux operating system.

The N series OnCommand System Manager is supported on the following platforms:

- Microsoft Windows:
  - Windows XP
  - Windows Vista
  - Windows 7
  - Windows 8
  - Windows Server 2003
  - Windows Server 2008
  - Windows Server 2008 R2
  - Windows Server 2012
- Linux:
  - Red Hat Enterprise Linux 5 or 6
  - SUSE Linux Enterprise Server 11

Because OnCommand System Manager 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 OnCommand System Manager to manage storage systems and HA configurations running the following versions of Clustered Data ONTAP:

- Data ONTAP 8.1.2 and 8.1.3
- Data ONTAP 8.2

20.2 Installing the N series OnCommand System Manager

Before you install OnCommand 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.

20.2.1 Installing OnCommand System Manager on Windows

You can install OnCommand System Manager on your Windows system by using the wizard-based installer:

1. Run the OnCommand 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 OnCommand System Manager and start managing your storage systems and objects.

20.2.2 Installing OnCommand System Manager on Linux

You can install OnCommand System Manager on your Linux system by using Red Hat Package Manager (command **`rpm`**):

1. Install OnCommand 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 OnCommand System Manager and start managing your storage systems and objects.

20.3 Getting started with OnCommand System Manager

The OnCommand 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; SVMs; HA configurations; and SnapMirror relationships.

For more information about how to configure and manage your storage systems from OnCommand System Manager, see the *OnCommand System Manager Help*. You can access the Help in PDF format from the IBM Support Site or from the Help provided with the OnCommand System Manager software.

http://www-01.ibm.com/support/docview.wss?uid=ssg1S7003448

Before you can start managing a storage system from OnCommand System Manager, you have to add it to OnCommand System Manager.

20.3.1 Starting OnCommand System Manager

Although OnCommand System Manager 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 OnCommand System Manager Java daemon initializes.

The OnCommand System Manager application can be started from the Windows desktop menu:

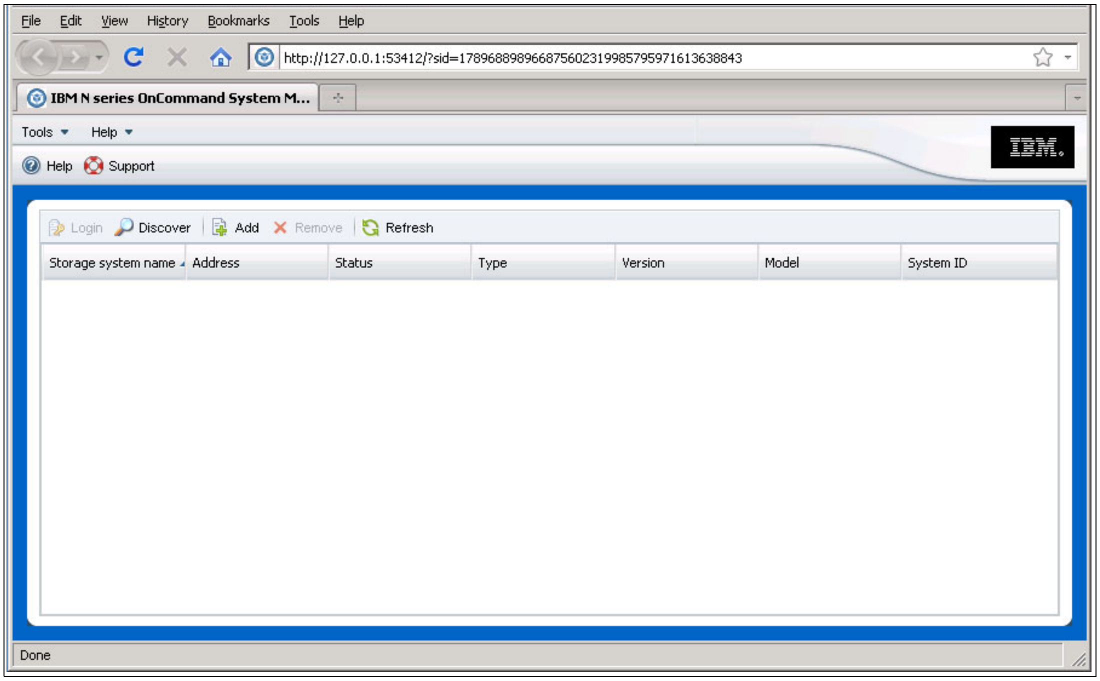
`Start --> Programs --> IBM --> N series OnCommand System Manager --> IBM N series OnCommand System Manager 3.0`

Similar to the Windows example, you can start the OnCommand System Manager application from the Linux desktop menu.

Alternatively, you can also start OnCommand System Manager 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 OnCommand System Manager interface (see Figure 20-1).



http://127.0.0.1:53412/?sid=17896889896687560231998579597161363843

Figure 20-1 Initial OnCommand System Manager Interface

Next, you need to add a storage system to the OnCommand System Manager interface.

20.3.2 Adding a storage system

Before you can use OnCommand System Manager to manage your storage systems and objects, you have to add them to OnCommand System Manager.

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.

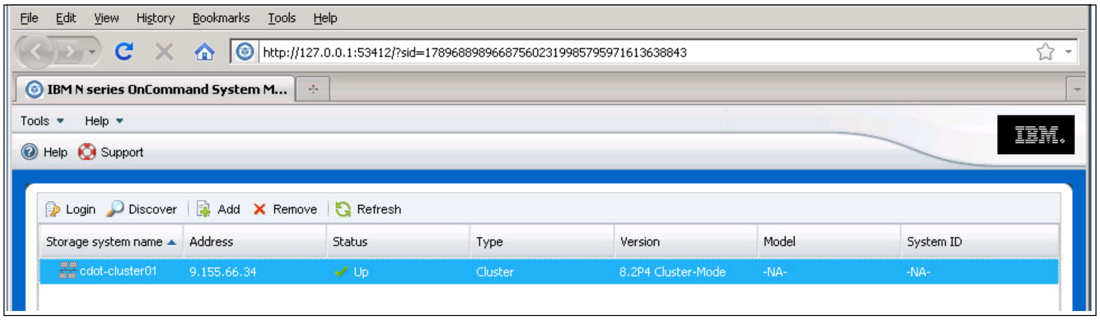
   You can specify the IPv6 address of the storage system, if you are adding a system that is running a supported version of Clustered Data ONTAP.
3. Click the **More** arrow.
4. Select the method for discovering and adding the storage system or cluster:
   - SNMP:

     Specify the SNMP community and SNMP version.
   - Credentials:

     Specify the user name and password.
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 OnCommand System Manager interface (see Figure 20-2).


File
Edit
View
History
Bookmarks
Tools
Help
http://127.0.0.1:53412/?sid=1789688989668756023199857959716l3638843
IBM N series OnCommand System M...
Tools
Help
IBM.
Help
Support
Login
Discover
Add
Remove
Refresh

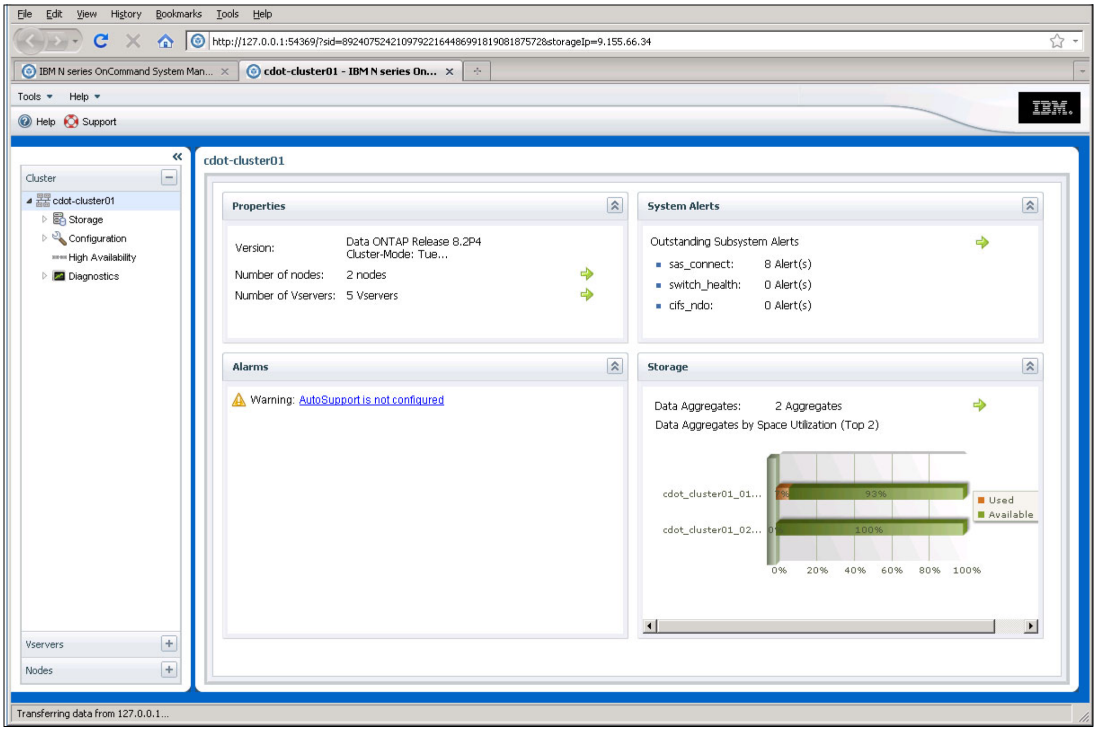
Storage system name ▲	Address	Status	Type	Version	Model	System ID
cdot-cluster01	9.155.66.34	✓ Up	Cluster	8.2P4 Cluster-Mode	-NA-	-NA-

Figure 20-2 Adding a controller to OnCommand System Manager

The first time that you double-click the new storage controller, you need to provide the correct username and password credentials.

20.3.3 Configuring a storage system

Figure 20-3 shows the main dashboard window, with the navigation panes on the left side.


File Edit View History Bookmarks Tools Help
http://127.0.0.1:54369/?sid=892407524210979221644869918190818757 28storageIp=9.155.66.34
IBM N series OnCommand System Man...
cdot-cluster01 - IBM N series On...
Tools
Help
Help
Support
IBM.
Cluster
cdot-cluster01
Storage
Configuration
High Availability
Diagnostics
Vservers
Nodes
cdot-cluster01
Properties
System Alerts
Alarms
Storage
cdot_cluster01_01...
cdot_cluster01_02...
93%
100%
0%
20%
40%
60%
80%
100%
Used
Available
Transferring data from 127.0.0.1...

Version: Data ONTAP Release 8.2P4 Cluster-Mode: Tue...

Number of nodes: 2 nodes

Number of Vservers: 5 Vservers

Outstanding Subsystem Alerts

- sas_connect: 8 Alert(s)
- switch_health: 0 Alert(s)
- cifs_ndo: 0 Alert(s)

Warning: AutoSupport is not configured

Data Aggregates: 2 Aggregates

Data Aggregates by Space Utilization (Top 2)

Figure 20-3 Dashboard view in OnCommand System Manager

For further information about configuration use cases, see Part 4, “Storage virtual machine use cases” on page 343.

Part 4

Storage virtual machine use cases

This part of the book addresses examples that guide you through the implementation of typical use cases for a storage virtual machine (SVM) and therefore serves as a how-to guide. The examples are based on N series OnCommand System Manager and also on commands in the clustershell.

The following topics are covered:

- Data protection
- iSCSI/FC storage
- CIFS storage
- NFS storage

21

Data protection

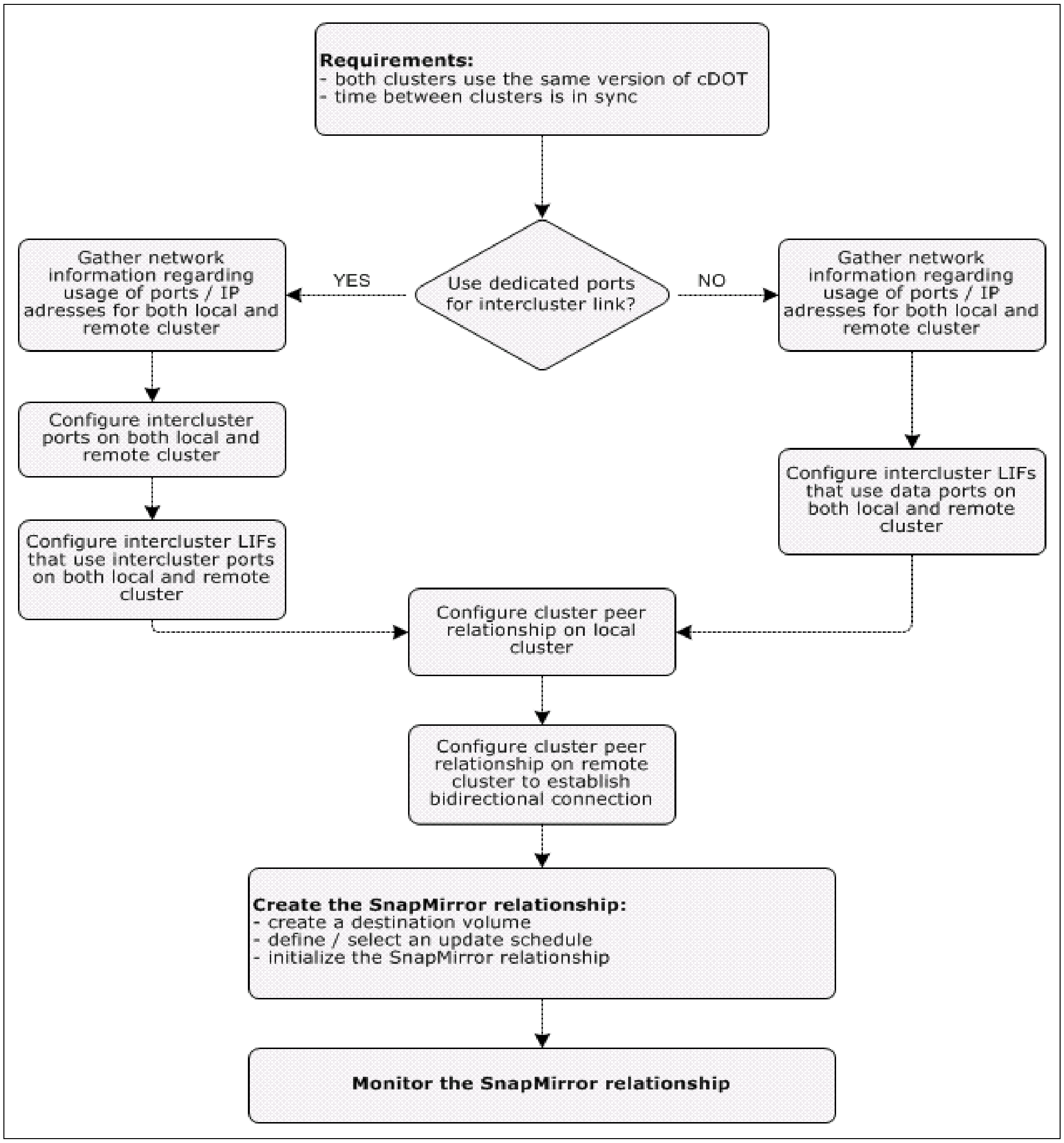
This chapter describes how to configure and monitor SnapMirror relationships between volumes in different Clustered Data ONTAP systems with IBM N series OnCommand System Manager. The example described herein can serve as a guide if you want to configure and monitor SnapMirror relationships for disaster recovery. The conceptual background for the task is not part of this example.

The following topics are covered:

- Configuration workflow
- Protecting a volume with SnapMirror

21.1 Configuration workflow

The workflow diagram in Figure 21-1 describes the process of creating a mirror relationship between volumes hosted by storage virtual machines (SVMs) in different clusters.



Requirements:
- both clusters use the same version of cDOT
- time between clusters is in sync
Use dedicated ports for intercluster link?
YES
NO
Gather network information regarding usage of ports / IP adresses for both local and remote cluster
Configure intercluster ports on both local and remote cluster
Configure intercluster LIFs that use intercluster ports on both local and remote cluster
Gather network information regarding usage of ports / IP adresses for both local and remote cluster
Configure intercluster LIFs that use data ports on both local and remote cluster
Configure cluster peer relationship on local cluster
Configure cluster peer relationship on remote cluster to establish bidirectional connection
Create the SnapMirror relationship:
- create a destination volume
- define / select an update schedule
- initialize the SnapMirror relationship
Monitor the SnapMirror relationship

Figure 21-1 SnapMirror configuration workflow

21.2 Protecting a volume with SnapMirror

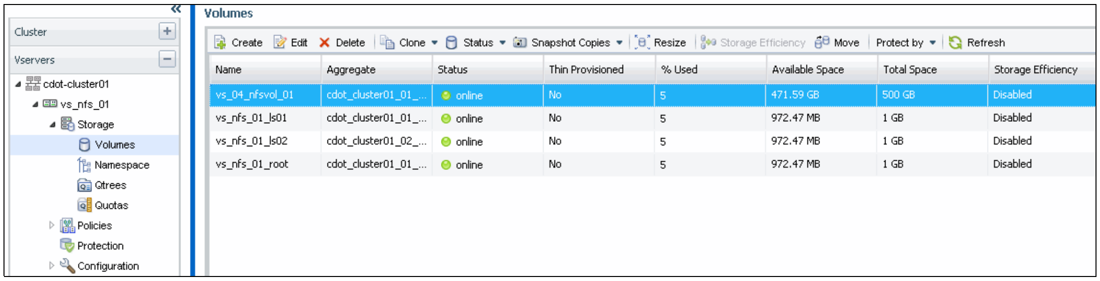
The goal of this example is to create a SnapMirror relationship between volumes on different clusters for disaster recovery. This includes creating a peer relationship between the source and the destination cluster.

You should check if the following requirements are met before continuing with the example:

- You must have the administrator user name and password for the source and destination clusters.
- The source and destination clusters must be added to System Manager.
- The SnapMirror license must be enabled on both the source and the destination clusters.

The following steps will guide you through every task that needs to be completed in order to protect a volume with SnapMirror:

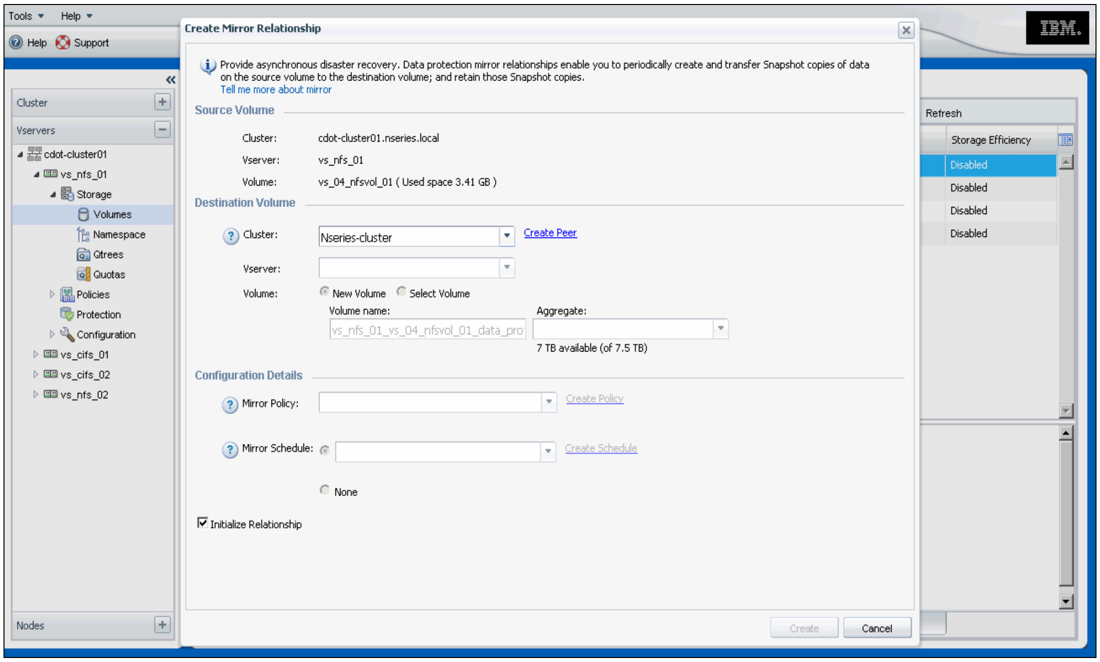
1. From the System Manager home page, double-click the appropriate cluster.
2. Expand the **SVMs** hierarchy in the left navigation pane.
3. Select the source SVM from the left navigation pane that contains the volume you want to protect, and then select **Storage** → **Volumes**.
4. In the Details tab, select the volume you want to protect, and then click **Protect by** → **Mirror** (see Figure 21-2).



Name	Aggregate	Status	Thin Provisioned	% Used	Available Space	Total Space	Storage Efficiency
vs_04_nfsvol_01	cdot_cluster01_01_...	online	No	5	471.59 GB	500 GB	Disabled
vs_nfs_01_ls01	cdot_cluster01_01_...	online	No	5	972.47 MB	1 GB	Disabled
vs_nfs_01_ls02	cdot_cluster01_02_...	online	No	5	972.47 MB	1 GB	Disabled
vs_nfs_01_root	cdot_cluster01_01_...	online	No	5	972.47 MB	1 GB	Disabled

Figure 21-2 Protect volume with mirror

5. The Create Mirror Relationship window is displayed (see Figure 21-3).



Create Mirror Relationship

Provide asynchronous disaster recovery. Data protection mirror relationships enable you to periodically create and transfer Snapshot copies of data on the source volume to the destination volume; and retain those Snapshot copies.
Tell me more about mirror

Source Volume

Cluster: cdot-cluster01.nseries.local
Vserver: vs_nfs_01
Volume: vs_04_nfsvol_01 (Used space 3.41 GB)

Destination Volume

Cluster: Nseries-cluster ▾ Create Peer
Vserver: ▾
Volume: ● New Volume ○ Select Volume
Volume name: vs_nfs_01_vs_04_nfsvol_01_data_pro
Aggregate: ▾
7 TB available (of 7.5 TB)

Configuration Details

Mirror Policy: ▾ Create Policy
Mirror Schedule: ● ▾ Create Schedule
○ None

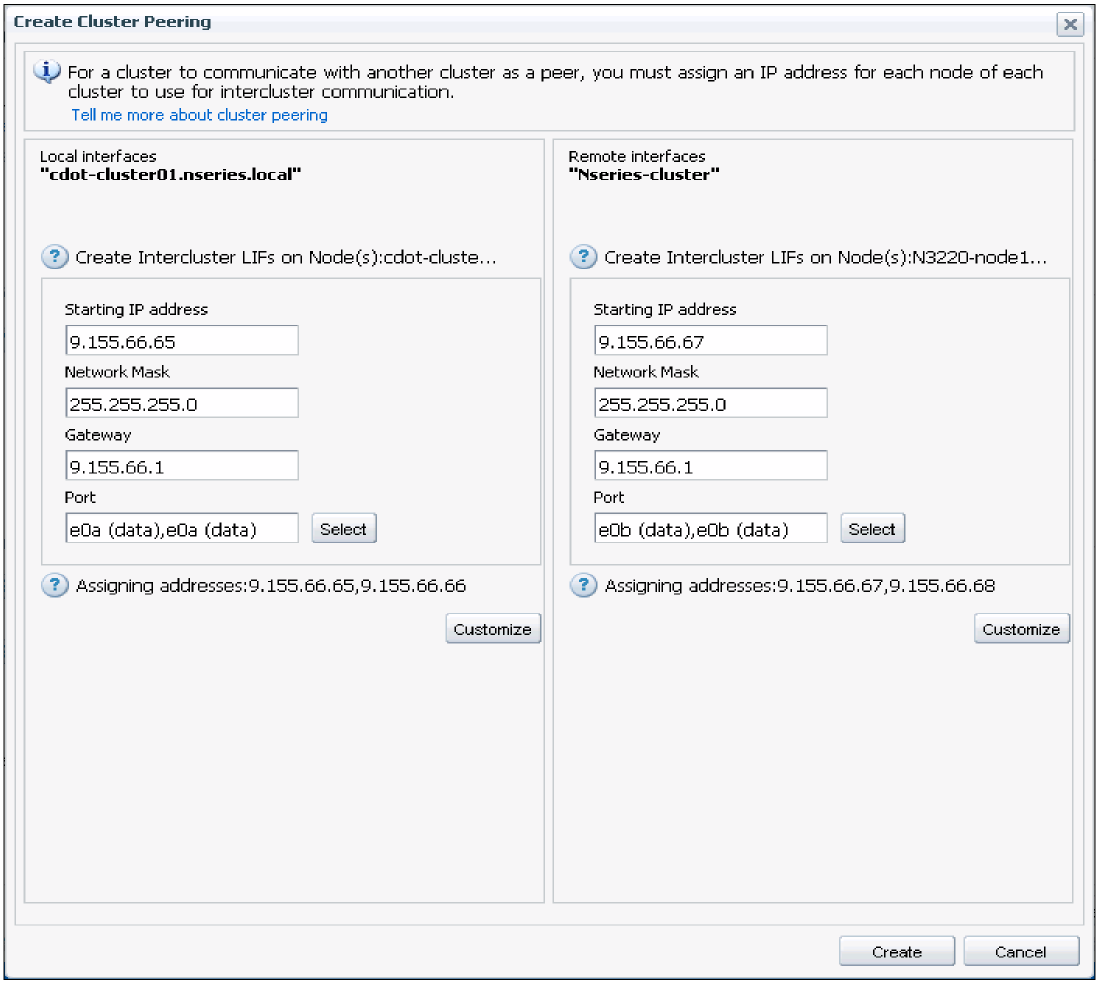
☑ Initialize Relationship

Create Cancel

Figure 21-3 Create Mirror Relationship window

6. In the **Destination Volume** area, select the remote cluster from the **Cluster** list and click **Create Peer** to establish a peer relationship with the remote cluster.

7. The Create Cluster Peering window is displayed (see Figure 21-4).



Create Cluster Peering

For a cluster to communicate with another cluster as a peer, you must assign an IP address for each node of each cluster to use for intercluster communication.
Tell me more about cluster peering

Local interfaces
"cdot-cluster01.nseries.local"

Create Intercluster LIFs on Node(s):cdot-cluste...

Starting IP address: 9.155.66.65
Network Mask: 255.255.255.0
Gateway: 9.155.66.1
Port: e0a (data),e0a (data)

Assigning addresses:9.155.66.65,9.155.66.66

Remote interfaces
"Nseries-cluster"

Create Intercluster LIFs on Node(s):N3220-node1...

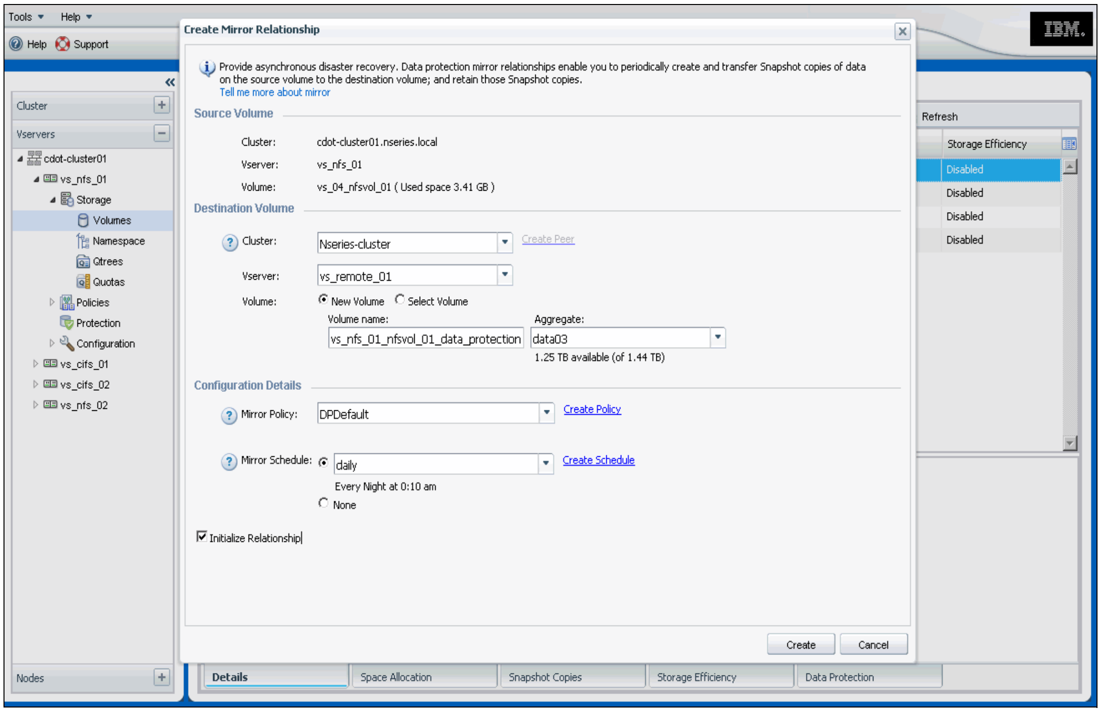
Starting IP address: 9.155.66.67
Network Mask: 255.255.255.0
Gateway: 9.155.66.1
Port: e0b (data),e0b (data)

Assigning addresses:9.155.66.67,9.155.66.68

Figure 21-4 Create Cluster Peering window

8. Fill in the required network information and select the physical ports you want to use for the intercluster logical interfaces (LIFs). Click **Create** to close this window and continue with the configuration of the mirror relationship (see Figure 21-5).

> **Note:** You can edit the automatically assigned addresses to fulfill your needs (for example, if you cannot specify a range of addresses). Click **Customize** after you have filled out every field and choose appropriate addresses.



Create Mirror Relationship

Provide asynchronous disaster recovery. Data protection mirror relationships enable you to periodically create and transfer Snapshot copies of data on the source volume to the destination volume; and retain those Snapshot copies.
Tell me more about mirror

Source Volume

Cluster: cdot-cluster01.nseries.local
Vserver: vs_nfs_01
Volume: vs_04_nfsvol_01 (Used space 3.41 GB)

Destination Volume

Cluster: Nseries-cluster Create Peer
Vserver: vs_remote_01
Volume: ● New Volume ○ Select Volume
Volume name: vs_nfs_01_nfsvol_01_data_protection
Aggregate: data03
1.25 TB available (of 1.44 TB)

Configuration Details

Mirror Policy: DPDefault Create Policy
Mirror Schedule: ● daily Create Schedule
Every Night at 0:10 am
○ None

☑ Initialize Relationship

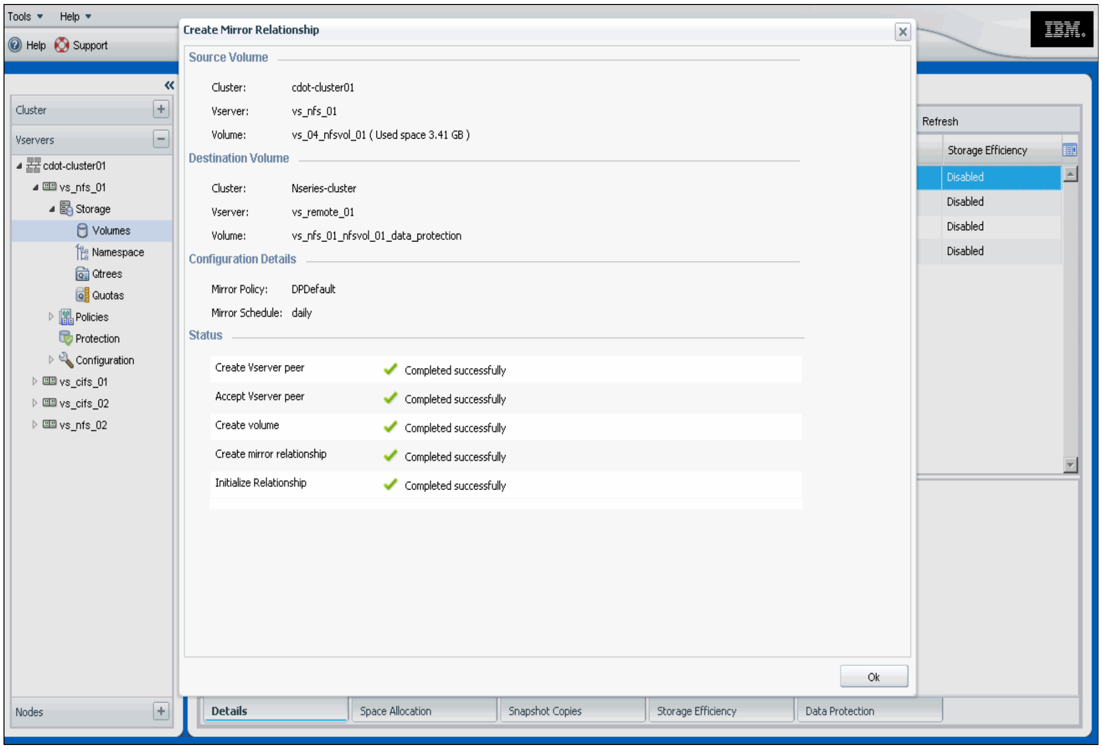
Create Cancel

Figure 21-5 Create Mirror Relationship

9. Select the desired destination SVM in the **Destination Volume** area. You also need to choose if you want to create a new destination volume or use an existing one. In this example, we create a new one in aggregate *data03*. In the **Configuration Details** area, select the **Mirror Policy** and the **Mirror Schedule**. You can use the predefined policies or create new ones. If you want your relationship to be initialized after it is created, select the **Initialize Relationship** check box.

10. Click **Create** in the lower right to finish the wizard (see Figure 21-6).

> **Note:** Consider that the initialization of the mirror relationship might need several hours, depending on the amount of data that is stored in the source volume.



Create Mirror Relationship

Source Volume

Cluster: cdot-cluster01
Vserver: vs_nfs_01
Volume: vs_04_nfsvol_01 (Used space 3.41 GB)

Destination Volume

Cluster: Nseries-cluster
Vserver: vs_remote_01
Volume: vs_nfs_01_nfsvol_01_data_protection

Configuration Details

Mirror Policy: DPDefault
Mirror Schedule: daily

Status

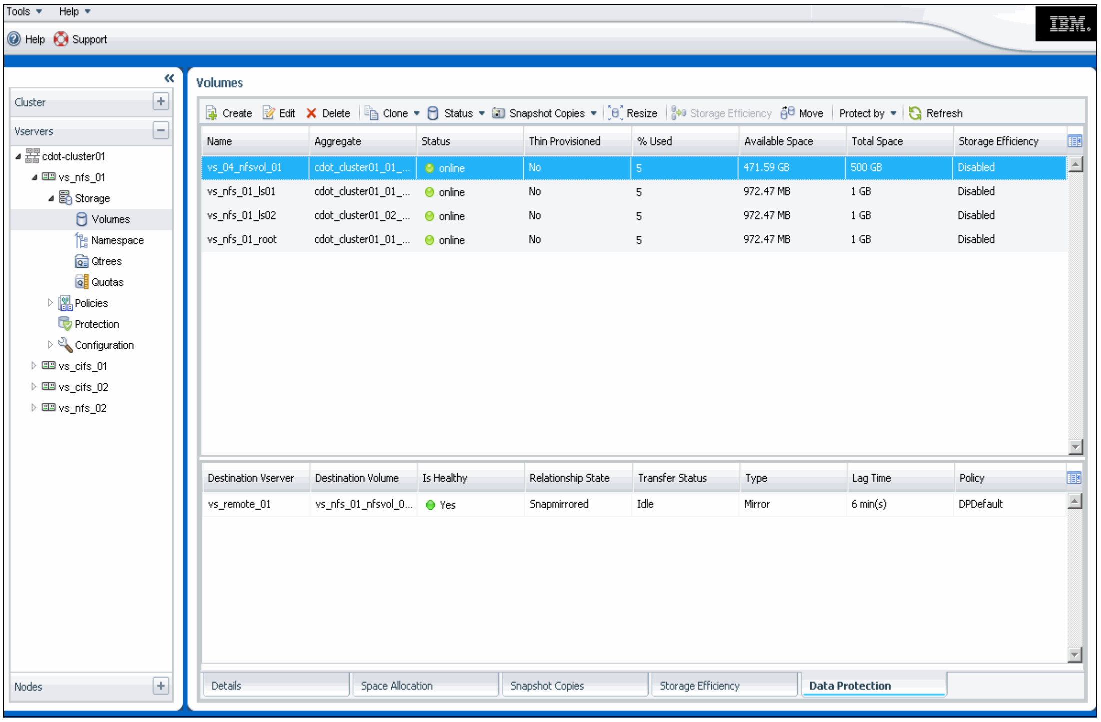
Create Vserver peer	✓ Completed successfully
Accept Vserver peer	✓ Completed successfully
Create volume	✓ Completed successfully
Create mirror relationship	✓ Completed successfully
Initialize Relationship	✓ Completed successfully

Figure 21-6 Create Mirror Relationship

11. After the wizard has completed all tasks, click **Ok** to close the window.

12. Select the volume from the Volumes list and click **Data protection**.

13. In the Data protection tab, verify that the SnapMirror relationship you created is listed and the **Relationship State** is *Snapmirrored* and the **Is Healthy** state is *Yes*. See Figure 21-7.



Name	Aggregate	Status	Thin Provisioned	% Used	Available Space	Total Space	Storage Efficiency
vs_04_nfsvol_01	cdot_cluster01_01_...	● online	No	5	471.59 GB	500 GB	Disabled
vs_nfs_01_ls01	cdot_cluster01_01_...	● online	No	5	972.47 MB	1 GB	Disabled
vs_nfs_01_ls02	cdot_cluster01_02_...	● online	No	5	972.47 MB	1 GB	Disabled
vs_nfs_01_root	cdot_cluster01_01_...	● online	No	5	972.47 MB	1 GB	Disabled

Destination Vserver	Destination Volume	Is Healthy	Relationship State	Transfer Status	Type	Lag Time	Policy
vs_remote_01	vs_nfs_01_nfsvol_0...	● Yes	Snapmirrored	Idle	Mirror	6 min(s)	DPDefault

Figure 21-7 Check SnapMirror state for volume

22

iSCSI/FC storage

This chapter shows a realistic example for block based storage using iSCSI storage that consists of a Clustered Data ONTAP system and a Windows 2008 server having several Ethernet ports. The approach for Fibre Channel (FC) is very similar.

We show you how to configure a storage virtual machine (SVM) for a Windows iSCSI host.

In the Clustered Data ONTAP 8.2 operating system, iSCSI is currently supported in clusters of up to eight nodes.

The following topics are covered:

- Planning and checking the iSCSI environment
- Configuring iSCSI for both Windows 2008 and Clustered Data ONTAP
- Accessing the iSCSI LUN on the Windows 2008 Host
- Further information

22.1 Planning and checking the iSCSI environment

In this chapter, you will create an SVM for a Windows 2008 iSCSI host through an OnCommand System Manager GUI. According to the official procedure, you will check the interoperability and configure both a Windows 2008 host and a Clustered Data ONTAP system.

22.1.1 Checking your environment before configuration of iSCSI

Before configuring iSCSI, you will need to perform several checks as described next.

Checking the interoperability

Here is the link to the N series interoperability matrix:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003897

In this case, we will check the interoperability between Windows 2008 and Clustered Data ONTAP. You can check iSCSI interoperability with the following link, which is included in the prior link:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003657

The support information is organized into the following topics in this interoperability matrix:

- Host Utilities version
- Certified Windows version
- Supported CPU types
- Information and alerts:

  In this topic, you will find important information regarding OS patches, considerations, and so on.

> **Note:** It is crucial to check the interoperability before the implementation and the planning phase. If you cannot find any certified solution, you must submit a SCORE/RPQ (Request for Price Quotations). To submit a SCORE/RPQ, contact your IBM Representative.

22.2 Configuring iSCSI for both Windows 2008 and Clustered Data ONTAP

In a Clustered Data ONTAP environment, you must create an SVM to provide iSCSI services to the host. We assume that you already know how to configure the basic cluster system.

22.2.1 Creating an SVM for iSCSI

To provide iSCSI services to specific hosts, you must create at least one SVM for iSCSI services. In this case, you will create one SVM for one Windows 2008 server.

Creating an SVM for iSCSI

To create an SVM, proceed as follows:

1. Click **Vservers** in the System Manager navigation frame.
2. Click the **Create** button on the System Manager Vservers page as shown in Figure 22-1.

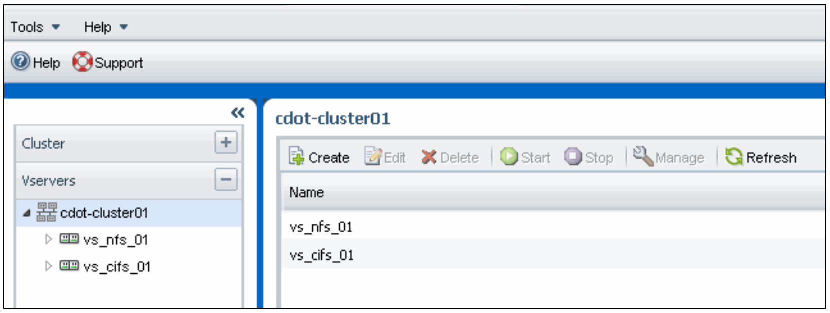


Figure 22-1 Create an SVM for iSCSI

3. Next you see the Vserver Setup wizard as shown in Figure 22-2. You need to enter an SVM name and several options in this window.

   Here we list some values to be entered:

   – Vserver Name: Define a name for the SVM that you want to name.

   – Data Protocols: Each SVM can be a server that provides multiple protocols or at least one protocol. In this chapter, select iSCSI only.

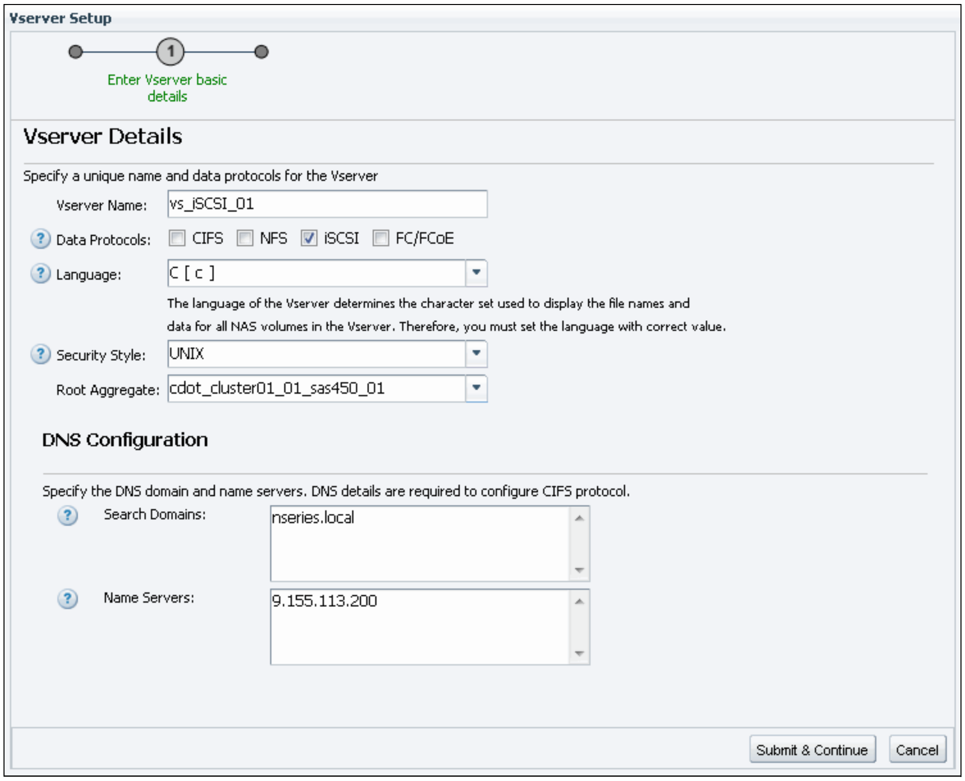
   – Language: This parameter depends on the specific characters that are used by your language environment. The default language is set to C.UTF-8. If your country uses two-byte characters, select a correct one.

   > **Note:** The language of an SVM with FlexVol volumes can be modified after the SVM is created.

   – Security style: There are three security styles. UNIX, NTFS, and Mixed are the options. You can decide what security style to use on a volume, and you should consider two factors. The primary factor is the type of administrator that manages the file system. The secondary factor is the type of user or service that accesses the data on the volume.

   – Root Aggregate: Select a root aggregate from one of the node root aggregates.

4. When you have made your selections, click the **Submit & Continue** button.



Vserver Details

Specify a unique name and data protocols for the Vserver

Vserver Name: vs_iSCSI_01

Data Protocols: □ CIFS □ NFS ☑ iSCSI □ FC/FCoE

Language: C [c]

The language of the Vserver determines the character set used to display the file names and data for all NAS volumes in the Vserver. Therefore, you must set the language with correct value.

Security Style: UNIX

Root Aggregate: cdot_cluster01_01_sas450_01

DNS Configuration

Specify the DNS domain and name servers. DNS details are required to configure CIFS protocol.

Search Domains: nseries.local

Name Servers: 9.155.113.200

Figure 22-2 VServer Setup - Details Setup

After completing the prior steps, you can configure iSCSI logical interfaces (LIFs) as follows (see Figure 22-3):

1. In this panel, you can configure a data interface LIF for iSCSI.

> **Note:** You can configure the iSCSI protocol while creating the SVM or you can do so at a later time.

Here we explain some values that you can choose:

– Target Alias: Specify an alias for the iSCSI target.

> **Note:** The maximum number of characters for an alias name is 128. If you do not specify a target alias, the SVM name is used as an alias.

– LIFs Per Node: Specify the number of iSCSI LIFs that can be assigned to a single node. The minimum number for LIFs per node is two. The maximum number is the minimum of all the ports in the up state across the nodes. If the maximum value is an odd number, the previous even number is considered as the maximum value. You can choose any even number in the minimum and maximum value range.

> **Note:** Do not manually enter the value. You must the select a value from the list.
>
> **Example:** A 4-node cluster has node1, node2, and node3 with 4 ports each in the up state, and node4 with 5 ports in the up state. The effective maximum value for the cluster is 4.

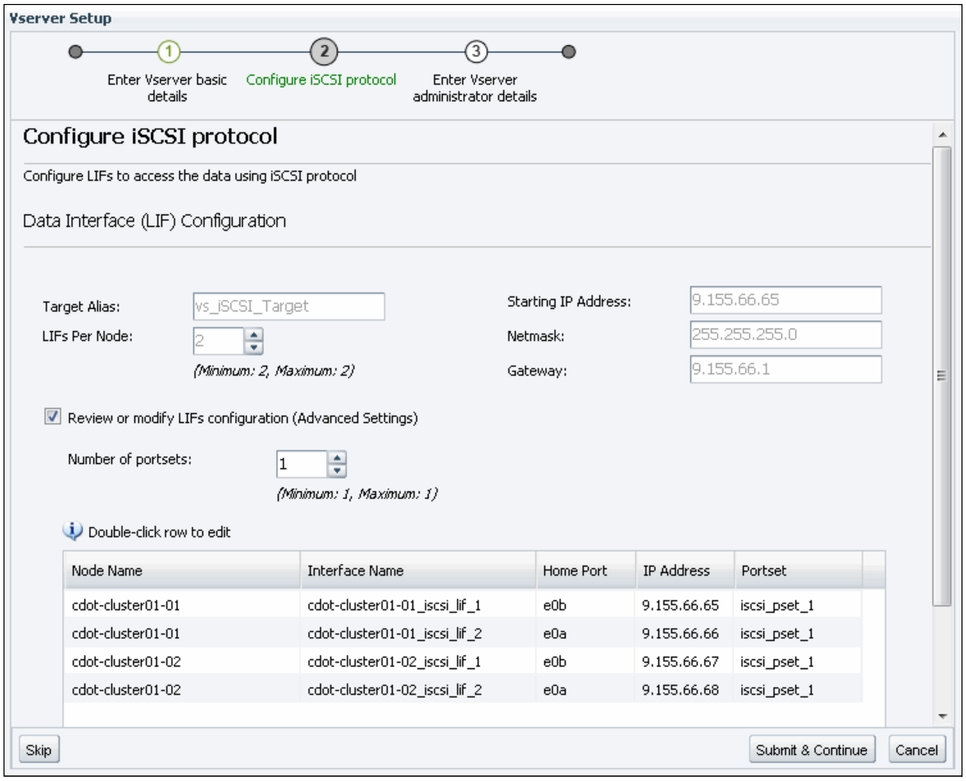
- Starting IP Address: Specify the starting network IP address. The starting IP address specifies the first of the contiguous addresses in the LIF IP address pool.

> **Note:** If you specify the number 4 for LIFs Per Node, the GUI will use a series of IP addresses as the starting IP address specified by you.
>
> **Example**: If you specify the IP address of 9.155.66.65, the rest of the IP addresses should be 9.155.66.66, 67, 68. If you do not want to use a series of IP addresses, make the LIF manually through the CLI or click the check box, **Review or Modify LIFs configuration (Advanced Settings)**.

- Netmask: Specify the subnet mask.
- Gateway:Specify the default gateway address.
- Number of portsets: If you want to verify or modify the automatically generated iSCSI LIFs configuration, select **Review or Modify LIFs configuration (Advanced Settings)**. You can modify only the LIF name, home port, and LIF IP address. By default, the portsets are set to the minimum value.

  You must ensure that you do not specify duplicate entries. If you specify duplicate LIF names, the System Manager appends numeric values to the duplicate LIF name. Based on the selected portset, the LIFs are distributed across the portsets using a round-robin method to ensure redundancy in case of node or port failure.

2. Then click the **Submit & Continue** button.


Vserver Setup
1 Enter Vserver basic details
2 Configure iSCSI protocol
3 Enter Vserver administrator details

Configure iSCSI protocol

Configure LIFs to access the data using iSCSI protocol

Data Interface (LIF) Configuration

Target Alias: vs_iSCSI_Target

LIFs Per Node: 2 (Minimum: 2, Maximum: 2)

Starting IP Address: 9.155.66.65

Netmask: 255.255.255.0

Gateway: 9.155.66.1

☑ Review or modify LIFs configuration (Advanced Settings)

Number of portsets: 1 (Minimum: 1, Maximum: 1)

Double-click row to edit

Node Name	Interface Name	Home Port	IP Address	Portset
cdot-cluster01-01	cdot-cluster01-01_iscsi_lif_1	e0b	9.155.66.65	iscsi_pset_1
cdot-cluster01-01	cdot-cluster01-01_iscsi_lif_2	e0a	9.155.66.66	iscsi_pset_1
cdot-cluster01-02	cdot-cluster01-02_iscsi_lif_1	e0b	9.155.66.67	iscsi_pset_1
cdot-cluster01-02	cdot-cluster01-02_iscsi_lif_2	e0a	9.155.66.68	iscsi_pset_1

Figure 22-3 Configure iSCSI protocol

3. After clicking the **Submit & Continue** button, the iSCSI LIFs and portsets are created with the specified configuration. The LIFs are distributed accordingly among the portsets. The iSCSI service is started if all the LIFs are successfully created. See Example 22-1 to check the result through the CLI.

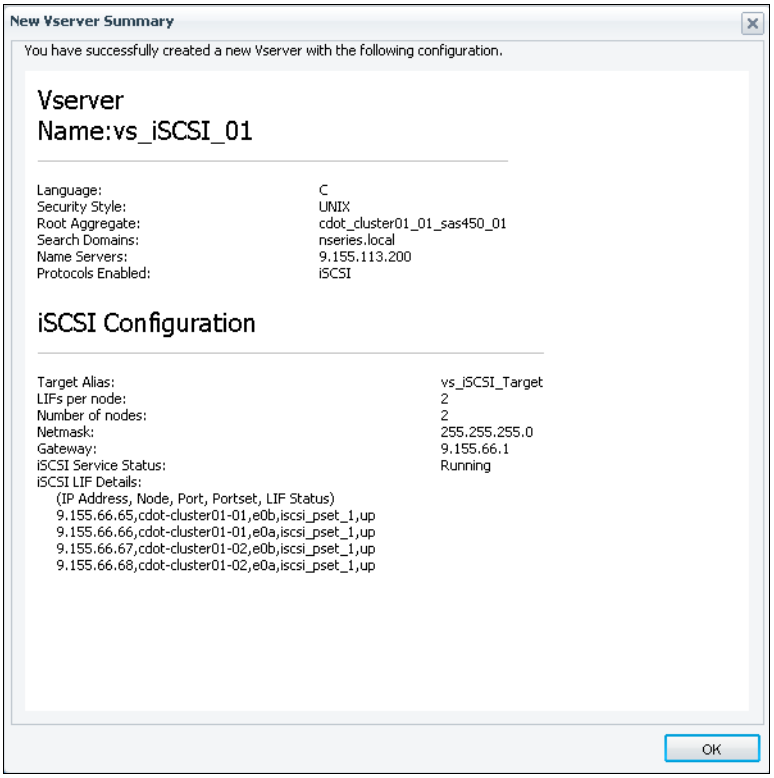
Example 22-1 A command to check iSCSI LIFs

```
cdot-cluster01::> network interface show
            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
cdot-cluster01
            cluster_mgmt up/up    9.155.66.34/24     cdot-cluster01-02
                                                                   e0a     false
cdot-cluster01-01
            clus1        up/up    169.254.216.146/16 cdot-cluster01-01
                                                                   e2a     true
            clus2        up/up    169.254.67.99/16   cdot-cluster01-01
                                                                   e2b     true
            mgmt1        up/up    9.155.90.168/24    cdot-cluster01-01
                                                                   e0M     true
cdot-cluster01-02
            clus1        up/up    169.254.45.229/16  cdot-cluster01-02
                                                                   e2a     true
            clus2        up/up    169.254.250.11/16  cdot-cluster01-02
                                                                   e2b     true
            mgmt1        up/up    9.155.90.169/24    cdot-cluster01-02
                                                                   e0M     true
vs_cifs_01
            vs_cifs_01_cifs_lif1
                         up/up    9.155.66.31/24     cdot-cluster01-02
Press <space> to page down, <return> for next line, or 'q' to quit...

                                                                   e0a     true

            Logical    Status     Network            Current       Current Is
Vserver     Interface  Admin/Oper Address/Mask       Node          Port    Home
----------- ---------- ---------- ------------------ ------------- ------- ----
vs_iSCSI_01
            cdot-cluster01-01_iscsi_lif_1
                         up/up    9.155.66.65/24     cdot-cluster01-01
                                                                   e0b     true
            cdot-cluster01-01_iscsi_lif_2
                         up/up    9.155.66.66/24     cdot-cluster01-01
                                                                   e0a     true
            cdot-cluster01-02_iscsi_lif_1
                         up/up    9.155.66.67/24     cdot-cluster01-02
                                                                   e0b     true
            cdot-cluster01-02_iscsi_lif_2
                         up/up    9.155.66.68/24     cdot-cluster01-02
                                                                   e0a     true
vs_nfs_01
            cdot_cluster_01_01_lif_01
                         up/up    9.155.66.26/24     cdot-cluster01-02
                                                                   e0a     false
13 entries were displayed.
```

4. On the next Vserver administration page, click **Skip**.

   In this panel, you can add an SVM user account and specify a login user method to access the storage system. If needed, you can configure them. Also, you can configure them through the CLI as well.

5. After completing the configuration, you can see the panel shown in Figure 22-4.



New Vserver Summary

You have successfully created a new Vserver with the following configuration.

Vserver
Name:vs_iSCSI_01

Language:	C
Security Style:	UNIX
Root Aggregate:	cdot_cluster01_01_sas450_01
Search Domains:	nseries.local
Name Servers:	9.155.113.200
Protocols Enabled:	iSCSI

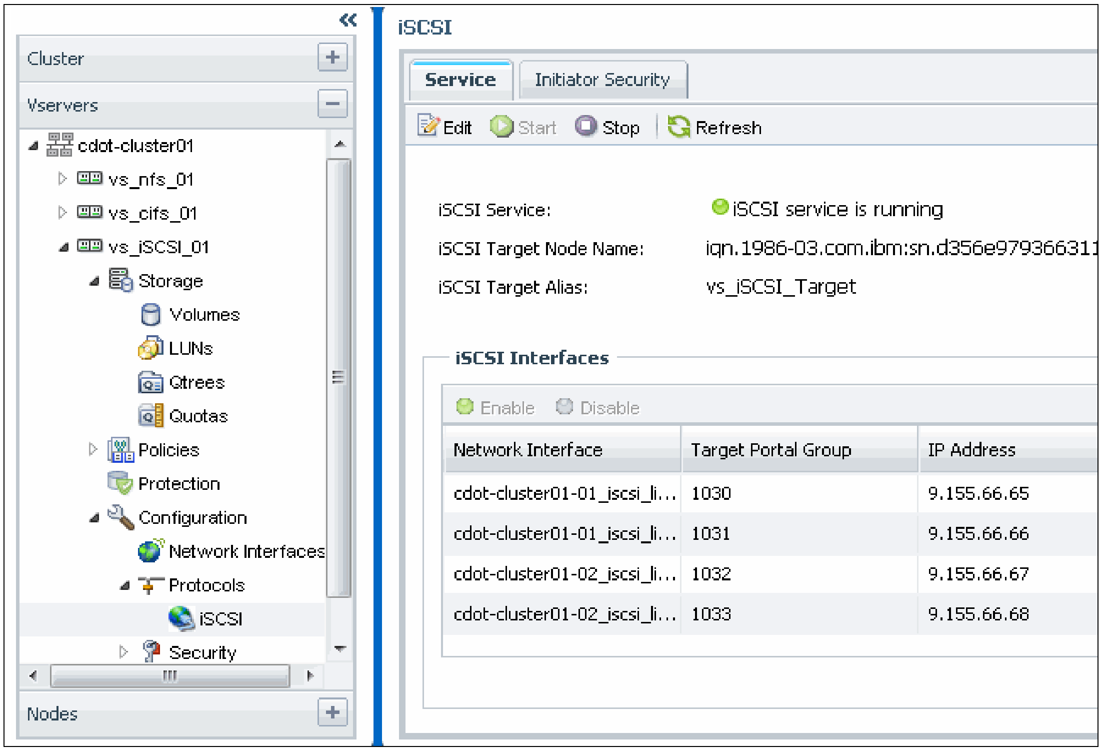
iSCSI Configuration

Target Alias:	vs_iSCSI_Target
LIFs per node:	2
Number of nodes:	2
Netmask:	255.255.255.0
Gateway:	9.155.66.1
iSCSI Service Status:	Running
iSCSI LIF Details:	

```
(IP Address, Node, Port, Portset, LIF Status)
9.155.66.65,cdot-cluster01-01,e0b,iscsi_pset_1,up
9.155.66.66,cdot-cluster01-01,e0a,iscsi_pset_1,up
9.155.66.67,cdot-cluster01-02,e0b,iscsi_pset_1,up
9.155.66.68,cdot-cluster01-02,e0a,iscsi_pset_1,up
```

Figure 22-4 New Vserver Summary

6. Also, you can see the status of iSCSI LIFs and working condition through the OnCommand System Manager as shown in Figure 22-5.



iSCSI Service: iSCSI service is running

iSCSI Target Node Name: iqn.1986-03.com.ibm:sn.d356e979366311

iSCSI Target Alias: vs_iSCSI_Target

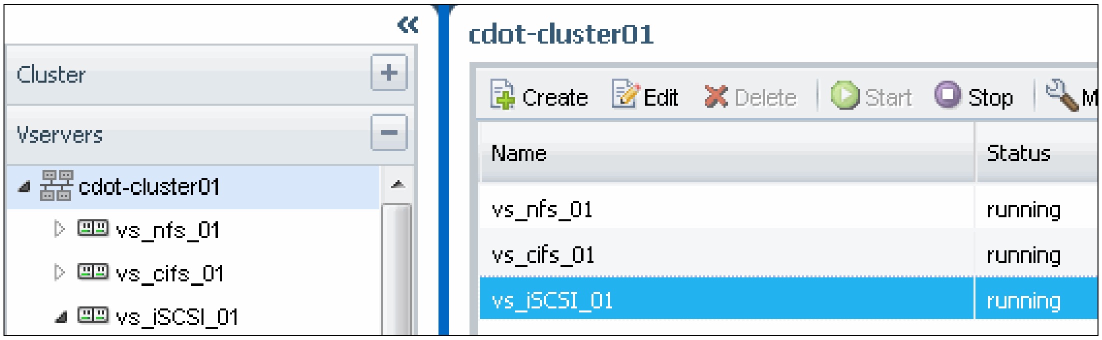
Network Interface	Target Portal Group	IP Address
cdot-cluster01-01_iscsi_li...	1030	9.155.66.65
cdot-cluster01-01_iscsi_li...	1031	9.155.66.66
cdot-cluster01-02_iscsi_li...	1032	9.155.66.67
cdot-cluster01-02_iscsi_li...	1033	9.155.66.68

Figure 22-5 iSCSI Service status on OnCommand System Manager

Assigning aggregates as a resource for an SVM

In these steps, you can learn how to assign aggregates as a resource for a specific SVM to prioritize storage resource allocation to the highest-value workloads. Proceed as follows:

1. Select the name of the iSCSI SVM that you made, and click **Edit**. See Figure 22-6.

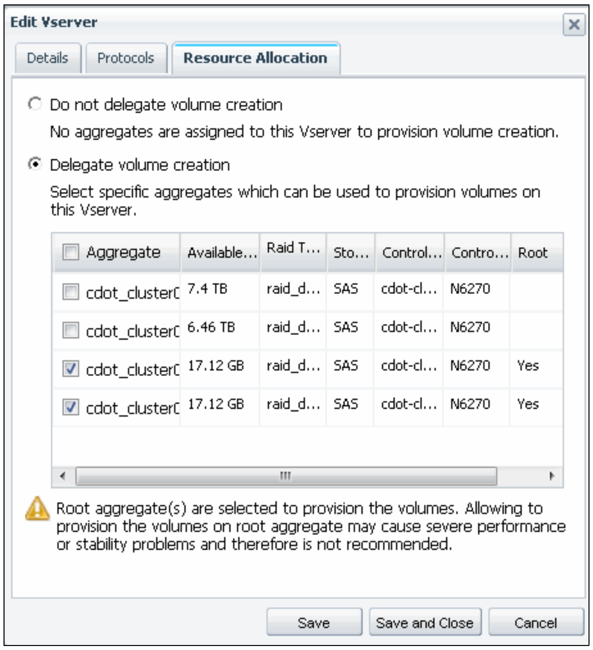


Name	Status
vs_nfs_01	running
vs_cifs_01	running
vs_iSCSI_01	running

Figure 22-6 Editing iSCSI properties

2. Click the **Resource Allocation** tab when the Edit Vserver dialog box appears. You can select any specific aggregates for the current iSCSI SVM.

   See Figure 22-7. This is a *bad* example for a configuration when choosing root aggregates.



○ Do not delegate volume creation

No aggregates are assigned to this Vserver to provision volume creation.

● Delegate volume creation

Select specific aggregates which can be used to provision volumes on this Vserver.

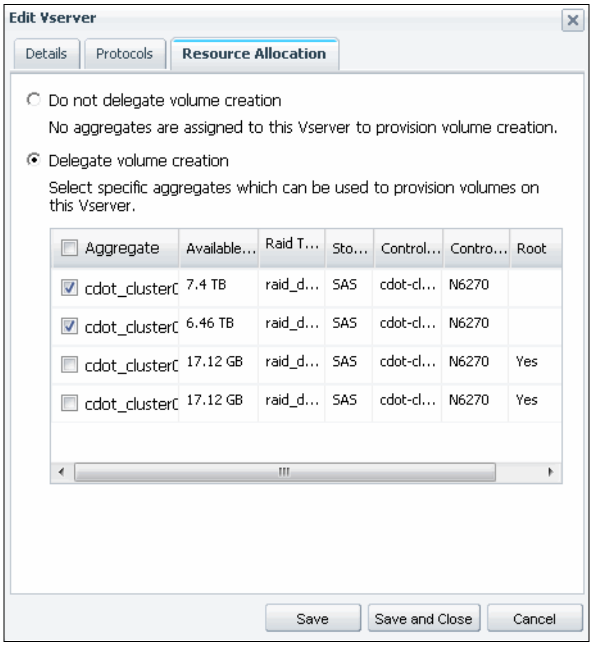
□ Aggregate	Available...	Raid T...	Sto...	Control...	Contro...	Root
□ cdot_cluster(	7.4 TB	raid_d...	SAS	cdot-cl...	N6270	
□ cdot_cluster(	6.46 TB	raid_d...	SAS	cdot-cl...	N6270	
☑ cdot_cluster(	17.12 GB	raid_d...	SAS	cdot-cl...	N6270	Yes
☑ cdot_cluster(	17.12 GB	raid_d...	SAS	cdot-cl...	N6270	Yes

⚠ Root aggregate(s) are selected to provision the volumes. Allowing to provision the volumes on root aggregate may cause severe performance or stability problems and therefore is not recommended.

Figure 22-7 Bad example: Assign a resource on the root aggregates

As you can see here, OnCommand System Manager warns against this kind of configuration. To prevent any future performance issues or reduced recovery time when the system is in HA condition and so on, you need to specify suitable data aggregates for your SVM.

3. If you specify any data aggregate that does not include a root aggregate, you will not see any warnings, as shown in Figure 22-8. This is a *good* example for a configuration.



Edit Vserver

Details | Protocols | Resource Allocation

○ Do not delegate volume creation

No aggregates are assigned to this Vserver to provision volume creation.

● Delegate volume creation

Select specific aggregates which can be used to provision volumes on this Vserver.

□ Aggregate	Available...	Raid T...	Sto...	Control...	Contro...	Root
☑ cdot_clusterC	7.4 TB	raid_d...	SAS	cdot-cl...	N6270	
☑ cdot_clusterC	6.46 TB	raid_d...	SAS	cdot-cl...	N6270	
□ cdot_clusterC	17.12 GB	raid_d...	SAS	cdot-cl...	N6270	Yes
□ cdot_clusterC	17.12 GB	raid_d...	SAS	cdot-cl...	N6270	Yes

Figure 22-8 Good example: Assign a resource on the data aggregates

22.2.2 Configuring a Windows 2008 Host for iSCSI

We have already checked the interoperability and configured iSCSI SVM on Clustered Data ONTAP. In these steps, we are going to configure the Windows 2008 iSCSI host.

Configuring iSCSI on the Windows 2008 host

On the Windows desktop, proceed as follows:

1. Open the Server Manager to activate an iSCSI feature, and select **Add Features** as follows. On the Server Manager, choose **Action** → **Add Features**. See Figure 22-9.

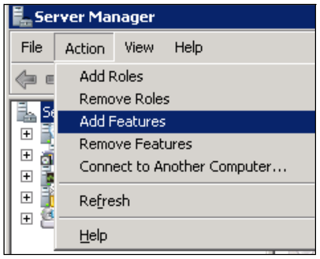
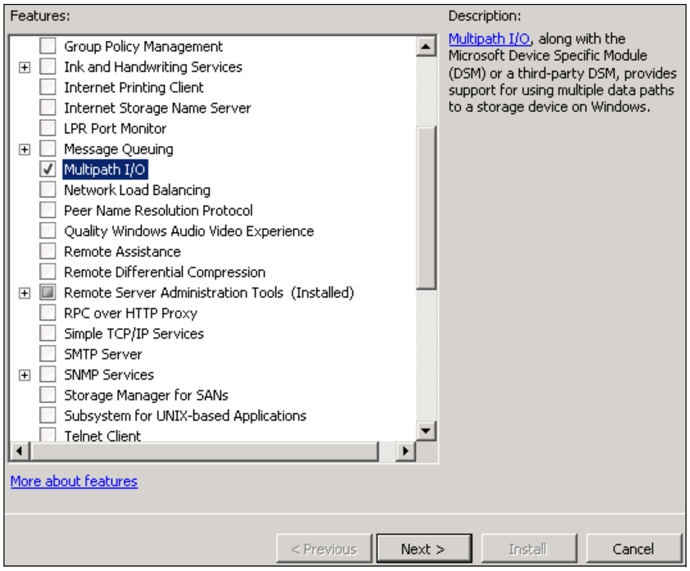


Figure 22-9 Add Features in Server Manager MMI

2. Then select **Multipath I/O** on the Select Features page and click **Next**. See Figure 22-10.



Features:

- ☐ Group Policy Management
- ⊞ ☐ Ink and Handwriting Services
- ☐ Internet Printing Client
- ☐ Internet Storage Name Server
- ☐ LPR Port Monitor
- ⊞ ☐ Message Queuing
- ☑ Multipath I/O
- ☐ Network Load Balancing
- ☐ Peer Name Resolution Protocol
- ☐ Quality Windows Audio Video Experience
- ☐ Remote Assistance
- ☐ Remote Differential Compression
- ⊞ ▣ Remote Server Administration Tools (Installed)
- ☐ RPC over HTTP Proxy
- ☐ Simple TCP/IP Services
- ☐ SMTP Server
- ⊞ ☐ SNMP Services
- ☐ Storage Manager for SANs
- ☐ Subsystem for UNIX-based Applications
- ☐ Telnet Client

Description:

Multipath I/O, along with the Microsoft Device Specific Module (DSM) or a third-party DSM, provides support for using multiple data paths to a storage device on Windows.

More about features

Figure 22-10 Selecting Multipath I/O

3. Then, to activate Multipath I/O, click the **Install** button. After clicking several following buttons, the Multipath I/O configuration will complete.

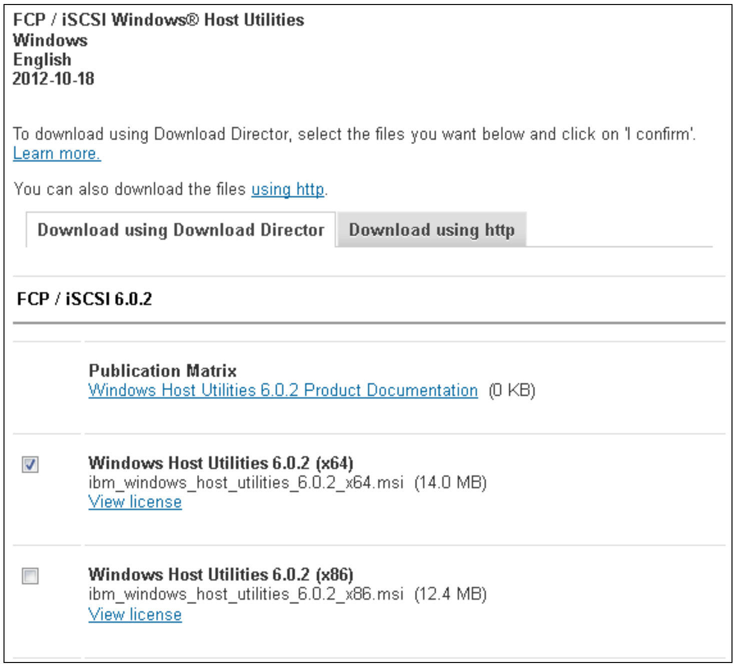
Installing FCP / iSCSI Windows Host Utilities

As we have already checked the suitable version for the host utilities, you must download the certified host utilities files from an IBM official site:

1. Use the following link to get N series software packages:

   http://www.ibm.com/support/docview.wss?uid=ssg1S7003280

2. If you access the site with an authority, you can download the file as in Figure 22-11.



FCP / iSCSI Windows® Host Utilities
Windows
English
2012-10-18

To download using Download Director, select the files you want below and click on 'I confirm'.
Learn more.

You can also download the files using http.

FCP / iSCSI 6.0.2

Publication Matrix
Windows Host Utilities 6.0.2 Product Documentation (0 KB)

☑ **Windows Host Utilities 6.0.2 (x64)**
ibm_windows_host_utilities_6.0.2_x64.msi (14.0 MB)
View license

☐ **Windows Host Utilities 6.0.2 (x86)**
ibm_windows_host_utilities_6.0.2_x86.msi (12.4 MB)
View license

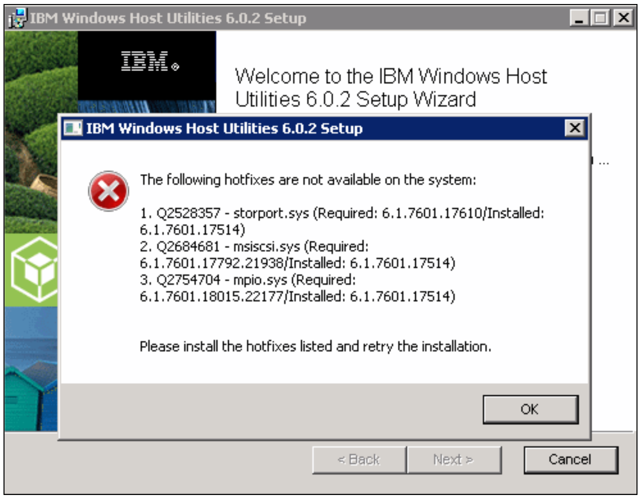
Figure 22-11 Windows FCP / iSCSI Windows Host Utilities download

3. The Windows Host Utilities enable you to connect a Windows host computer to the N series storage systems.

   The Windows Host Utilities include an installation program that sets the required Windows registry and HBA values. Starting with version 5.3, the diagnostic programs for troubleshooting problems with Windows hosts connected to N series storage systems were replaced by the nSANity program. You can obtain the nSANity program from your technical support representative.

4. To install the FCP/iSCSI Windows Host Utilities, double-click the file downloaded from the IBM official site in Figure 22-11.

5. If your system has any issues regarding use of iSCSI with N series storage systems, you can see error messages as shown in Figure 22-12.


IBM Windows Host Utilities 6.0.2 Setup
IBM.
Welcome to the IBM Windows Host Utilities 6.0.2 Setup Wizard
IBM Windows Host Utilities 6.0.2 Setup
OK
< Back
Next >
Cancel

The following hotfixes are not available on the system:

1. Q2528357 - storport.sys (Required: 6.1.7601.17610/Installed: 6.1.7601.17514)
2. Q2684681 - msiscsi.sys (Required: 6.1.7601.17792.21938/Installed: 6.1.7601.17514)
3. Q2754704 - mpio.sys (Required: 6.1.7601.18015.22177/Installed: 6.1.7601.17514)

Please install the hotfixes listed and retry the installation.

Figure 22-12 Host Utilities installation error due to Windows 2008 hotfix issues

   As you can see here, the installation will not be able to complete due to some hotfix issues.

6. According to this description, the customer should install all of the hotfixes as mentioned.

> **Note:** Here is a tip to download a hotfix. Just type the hotfix number at the end of the following link: **`http://support.microsoft.com/kb/`**
>
> **Example**: For Q2528357, just visit the following Microsoft site:
>
> `http://support.microsoft.com/kb/2528357`
>
> Then you can request the hotfix you want.

   a. After installing all of the hotfixes, you can complete the installation of FCP / iSCSI Windows Host Utilities.

Configuring the iSCSI software initiator in Windows 2008

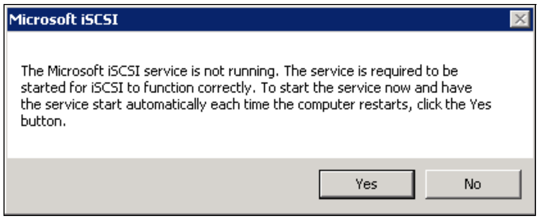
In the following steps, you will configure the MPIO to use and manage iSCSI connections. We will not configure the Multiple Connections per Session (MCS) technique for multipathing. If you need more information about MCS, see the Microsoft MCS articles.

In order to check the current iSCSI configuration, proceed as follows:

1. Open the control panel and double-click **iSCSI Initiator**.

   To make it easier to find the **iSCSI initiator** icon, you can select **View by small icons** for viewing for the control panel.

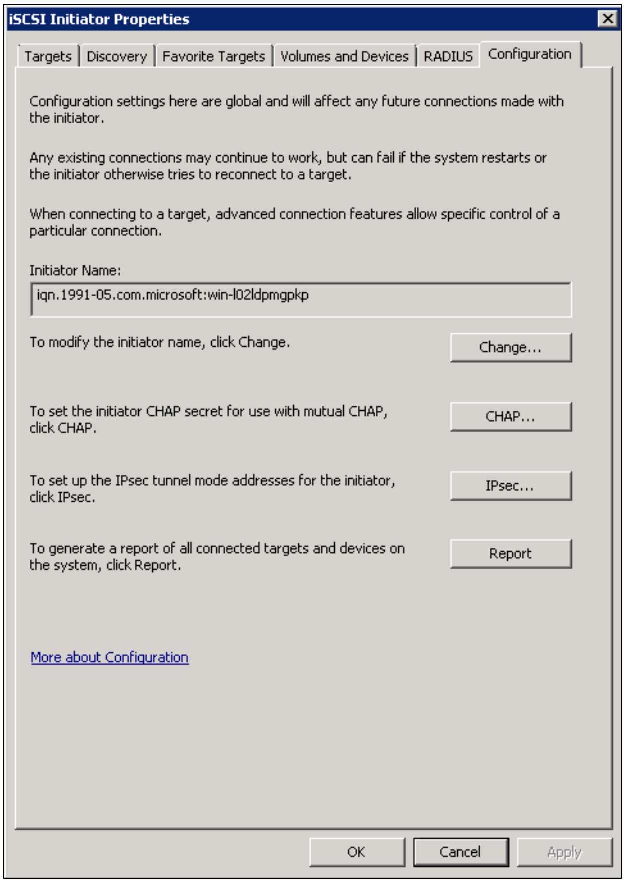
2. If an error message appears to indicate that the Microsoft ISCSI service is not running, click **Yes** to start the service. See Figure 22-13.



The Microsoft iSCSI service is not running. The service is required to be started for iSCSI to function correctly. To start the service now and have the service start automatically each time the computer restarts, click the Yes button.

Figure 22-13 Starting Microsoft iSCSI

3. When the iSCSI Initiator Properties dialog box appears, click the **Configuration** tab as shown in Figure 22-14.



Configuration settings here are global and will affect any future connections made with the initiator.

Any existing connections may continue to work, but can fail if the system restarts or the initiator otherwise tries to reconnect to a target.

When connecting to a target, advanced connection features allow specific control of a particular connection.

Initiator Name:
iqn.1991-05.com.microsoft:win-l02ldpmgpkp

To modify the initiator name, click Change.

To set the initiator CHAP secret for use with mutual CHAP, click CHAP.

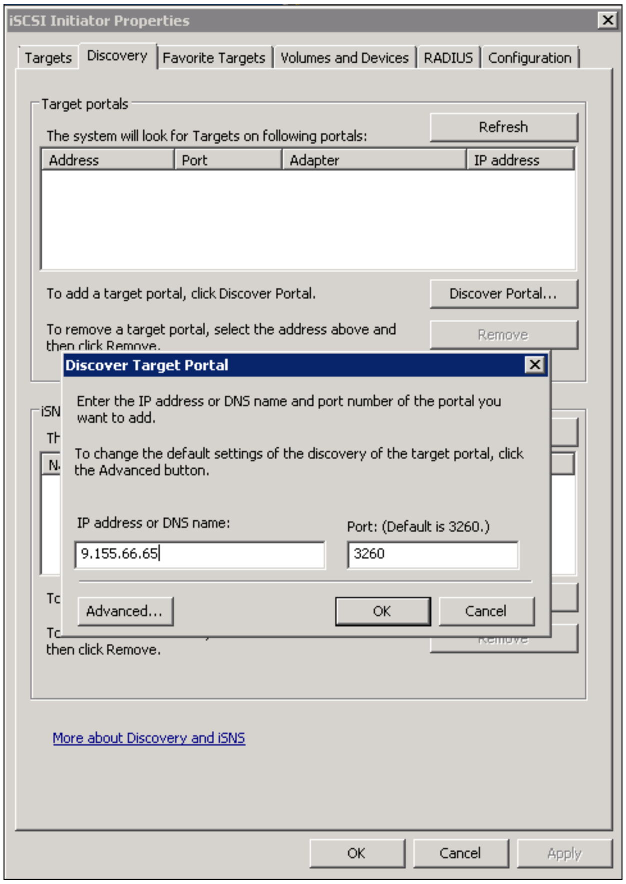
To set up the IPsec tunnel mode addresses for the initiator, click IPsec.

To generate a report of all connected targets and devices on the system, click Report.

Figure 22-14 Configuration tab in iSCSI Initiator Properties

4. Record the name of the Initiator (IQN). In this picture, we can see that the IQN name is **`iqn.1991-05.com.microsoft:win-1021dpmgpkp`**. We need this IQN name when we configure the iSCSI connection on the SVM in Clustered Data ONTAP.

5. To check the iSCSI Ethernet links that you made before in this chapter, click the **Discovery** tab. Click **Discover Portal**, enter the IP address of one of the ports within the iscsi_pset_1 port set, and click **OK**. See Figure 22-15.



The system will look for Targets on following portals:

To add a target portal, click Discover Portal.

To remove a target portal, select the address above and then click Remove.

Enter the IP address or DNS name and port number of the portal you want to add.

To change the default settings of the discovery of the target portal, click the Advanced button.

IP address or DNS name: 9.155.66.65

Port: (Default is 3260.) 3260

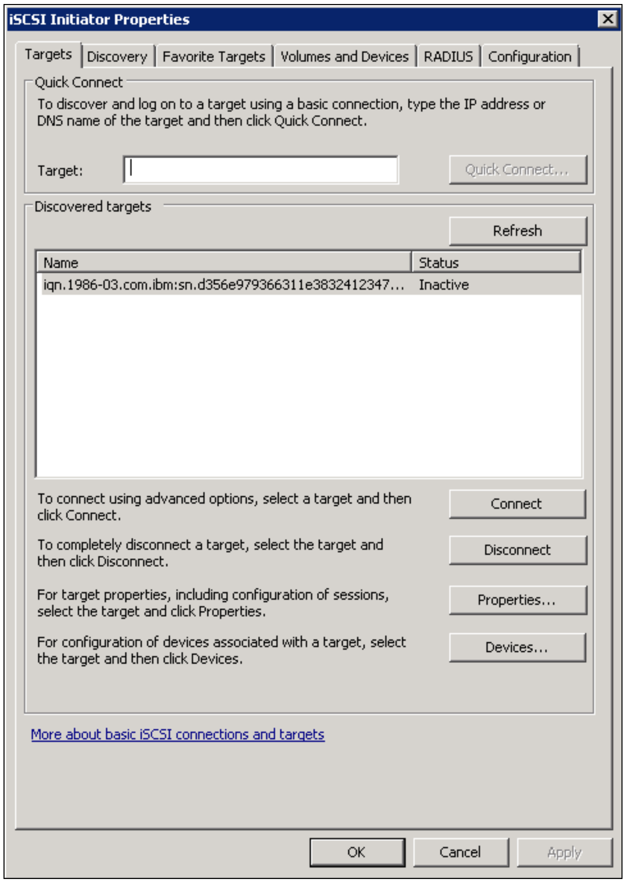
Figure 22-15 Discover Target Portal

6. If you want to check the port set name or other information, you can use the cluster shell as well. See Example 22-2.

Example 22-2 The command of portset show

```
cdot-cluster01::> portset show
Vserver   Portset      Protocol Port Names              Igroups
--------- ------------ -------- ----------------------- ------------
vs_iSCSI_01
          iscsi_pset_1 iscsi    cdot-cluster01-01_iscsi_lif_1,
cdot-cluster01-01_iscsi_lif_2, cdot-cluster01-02_iscsi_lif_1,
cdot-cluster01-02_iscsi_lif_2
                                                        -
```

7. Then click the **Targets** tab as shown in Figure 22-16.



To discover and log on to a target using a basic connection, type the IP address or DNS name of the target and then click Quick Connect.

Target:

Name	Status
iqn.1986-03.com.ibm:sn.d356e979366311e3832412347...	Inactive

To connect using advanced options, select a target and then click Connect.

To completely disconnect a target, select the target and then click Disconnect.

For target properties, including configuration of sessions, select the target and click Properties.

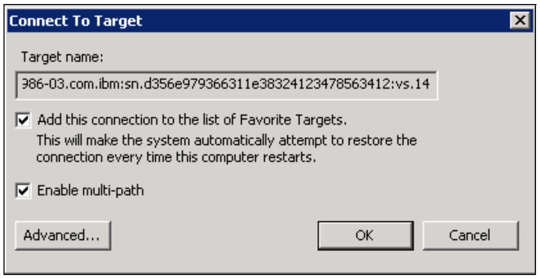
For configuration of devices associated with a target, select the target and then click Devices.

More about basic iSCSI connections and targets

Figure 22-16 Targets tab after specifying Target Portal Ip address

8. In order to proceed the next step, verify that the discovered target appears in the list and click **Connect**.

   After clicking the **Connect** button, you can see the pop-up window shown in Figure 22-17.



Target name:

986-03.com.ibm:sn.d356e979366311e38324123478563412:vs.14

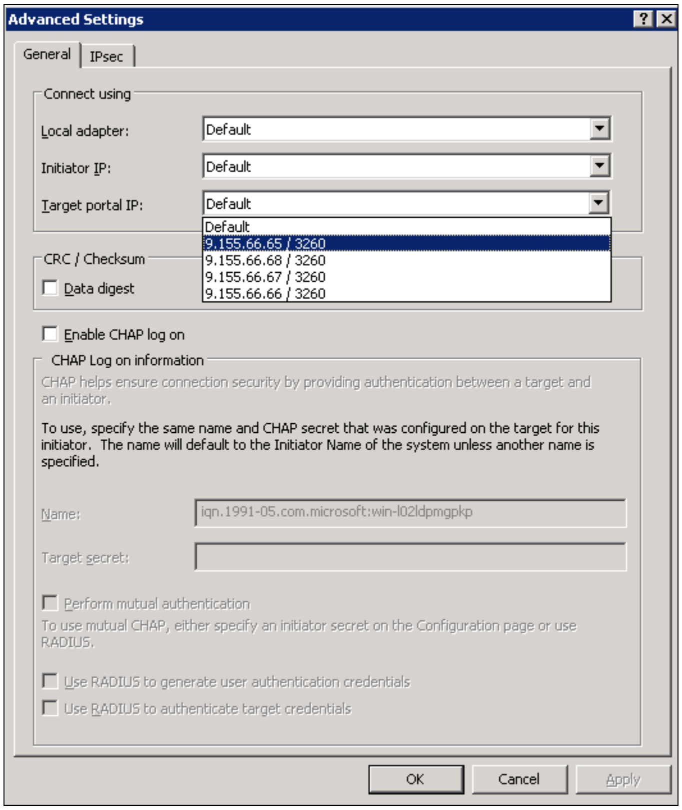
☑ Add this connection to the list of Favorite Targets.
This will make the system automatically attempt to restore the connection every time this computer restarts.

☑ Enable multi-path

Figure 22-17 Enabling multi-path settings

9. In this dialog box, select **Enable multi-path** and click **Advanced**.

10..When you see the Advanced Settings dialog box in Figure 22-18, select the *lowest* target portal IP address, and click **OK** from the **Target portal IP** list.



Connect using

Local adapter: Default

Initiator IP: Default

Target portal IP: Default

Default
9.155.66.65 / 3260
9.155.66.68 / 3260
9.155.66.67 / 3260
9.155.66.66 / 3260

CRC / Checksum

□ Data digest

□ Enable CHAP log on

CHAP Log on information

CHAP helps ensure connection security by providing authentication between a target and an initiator.

To use, specify the same name and CHAP secret that was configured on the target for this initiator. The name will default to the Initiator Name of the system unless another name is specified.

Name: iqn.1991-05.com.microsoft:win-l02ldpmgpkp

Target secret:

□ Perform mutual authentication

To use mutual CHAP, either specify an initiator secret on the Configuration page or use RADIUS.

□ Use RADIUS to generate user authentication credentials

□ Use RADIUS to authenticate target credentials

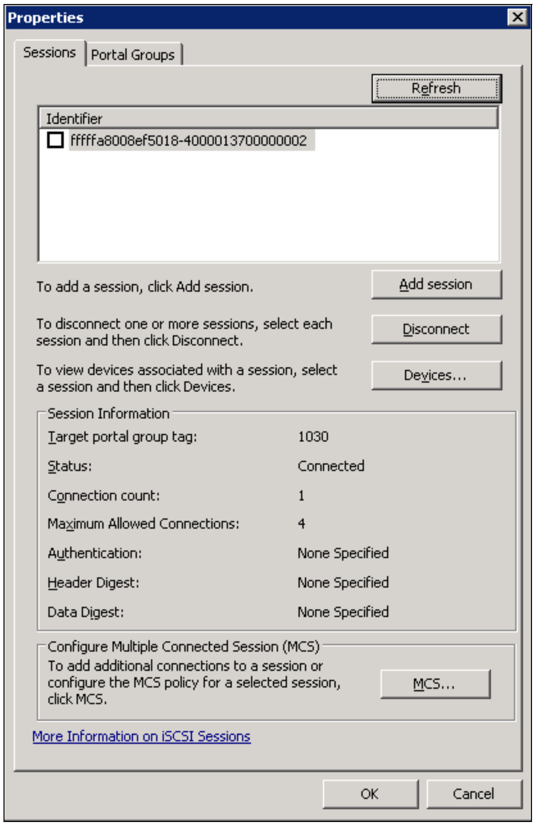
Figure 22-18 Setting Target portal IP

11.Then click **OK** to close the Connect to Target dialog box and start a new iSCSI session between the initiator and target.

12.Check the current status of iSCSI on the **Targets** tab to see the link status change from *inactive* to *connected*.

You have just made one iSCSI connection between a Clustered Data ONTAP system and a Windows 2008 host. Now you need to create the remaining iSCSI paths between them.

13. Click **Properties** to begin creating additional sessions with all of the iSCSI LIFs within the port set. See Figure 22-19.



Identifier
□ ffffa8008ef5018-4000013700000002

To add a session, click Add session.

To disconnect one or more sessions, select each session and then click Disconnect.

To view devices associated with a session, select a session and then click Devices.

Session Information

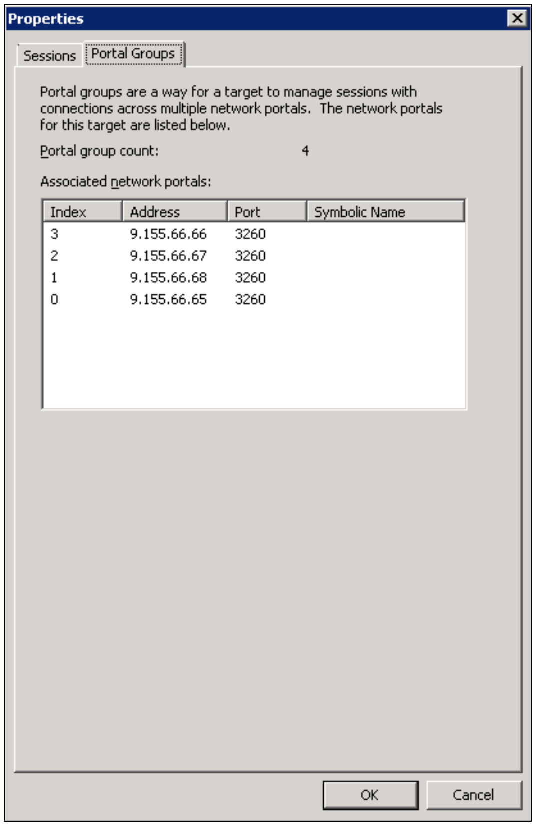
Target portal group tag:	1030
Status:	Connected
Connection count:	1
Maximum Allowed Connections:	4
Authentication:	None Specified
Header Digest:	None Specified
Data Digest:	None Specified

Configure Multiple Connected Session (MCS)
To add additional connections to a session or configure the MCS policy for a selected session, click MCS.

Figure 22-19 Properties dialog box

In the Properties dialog box, ensure that there is only one current session on the Session tab.

14. Click the **Portal Groups** tab and review all of the IPs, including the remaining three IPs that are currently available for sessions. See Figure 22-20.



Properties

Sessions | Portal Groups

Portal groups are a way for a target to manage sessions with connections across multiple network portals. The network portals for this target are listed below.

Portal group count: 4

Associated network portals:

Index	Address	Port	Symbolic Name
3	9.155.66.66	3260	
2	9.155.66.67	3260	
1	9.155.66.68	3260	
0	9.155.66.65	3260	

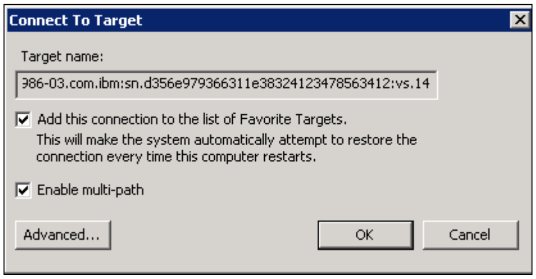
Figure 22-20 iSCSI IP addresses in Portal Groups tab

15. After verifying all of the iSCSI IP addresses, click the **Sessions** tab again to add more sessions.

    You can repeat the following steps three times to create four iSCSI sessions.

    a. Click **Add session**.

       In the Connect To Target dialog box, select Enable multi-path and click **Advanced**. See Figure 22-21.



Connect To Target

Target name:

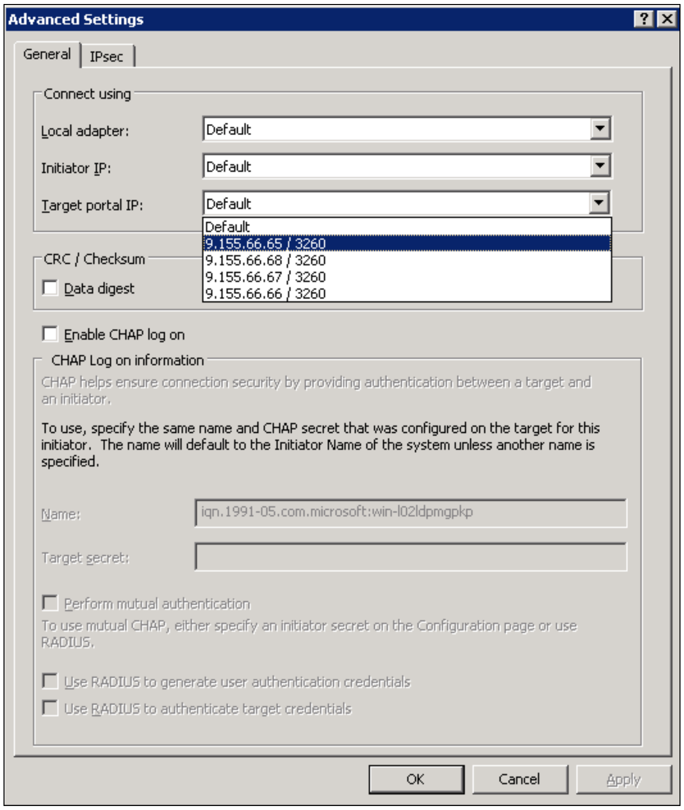
986-03.com.ibm:sn.d356e979366311e38324123478563412:vs.14

☑ Add this connection to the list of Favorite Targets.
This will make the system automatically attempt to restore the connection every time this computer restarts.

☑ Enable multi-path

Figure 22-21 Connect to Target

b. In the Advanced Settings dialog box, from the Target portal IP list, select the target portal IP address of one of the iSCSI LIFs that you have not assigned yet. See Figure 22-22.



Advanced Settings

Connect using

Local adapter: Default

Initiator IP: Default

Target portal IP: Default

Default
9.155.66.65 / 3260
9.155.66.68 / 3260
9.155.66.67 / 3260
9.155.66.66 / 3260

CRC / Checksum

□ Data digest

□ Enable CHAP log on

CHAP Log on information

CHAP helps ensure connection security by providing authentication between a target and an initiator.

To use, specify the same name and CHAP secret that was configured on the target for this initiator. The name will default to the Initiator Name of the system unless another name is specified.

Name: iqn.1991-05.com.microsoft:win-l02ldpmgpkp

Target secret:

□ Perform mutual authentication

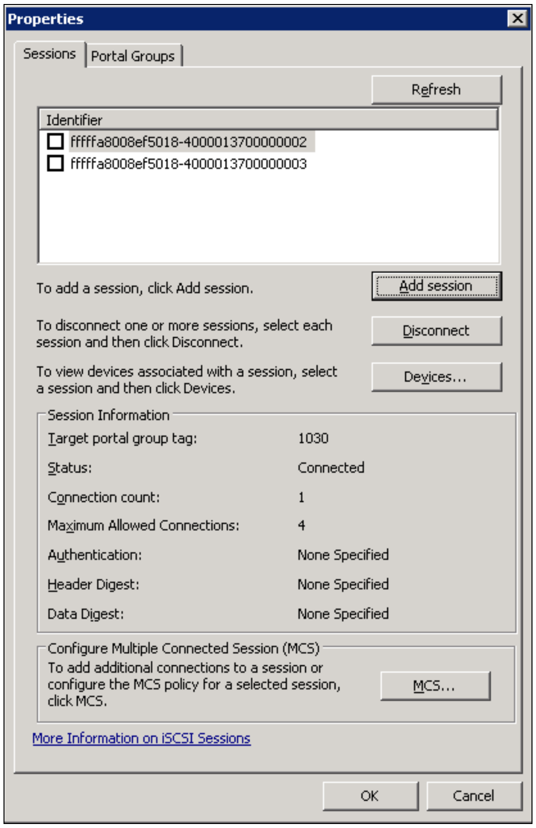
To use mutual CHAP, either specify an initiator secret on the Configuration page or use RADIUS.

□ Use RADIUS to generate user authentication credentials

□ Use RADIUS to authenticate target credentials

Figure 22-22 Advanced Settings

c. In the Properties dialog box, on the Sessions tab, verify that a new session has been created. See Figure 22-23.



□ fffffa8008ef5018-4000013700000002
□ fffffa8008ef5018-4000013700000003

To add a session, click Add session.

To disconnect one or more sessions, select each session and then click Disconnect.

To view devices associated with a session, select a session and then click Devices.

Session Information

Target portal group tag:	1030
Status:	Connected
Connection count:	1
Maximum Allowed Connections:	4
Authentication:	None Specified
Header Digest:	None Specified
Data Digest:	None Specified

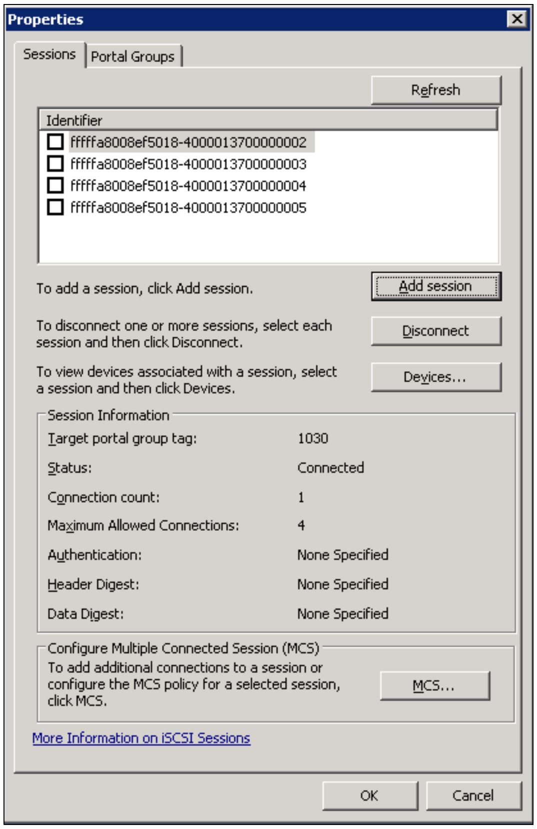
Configure Multiple Connected Session (MCS)

To add additional connections to a session or configure the MCS policy for a selected session, click MCS.

More Information on iSCSI Sessions

Figure 22-23 Sessions properties

16. Repeat Steps 1-3 to create two more sessions, for a total four sessions, each with an iSCSI LIF in the port set of the target. See Figure 22-24.



Identifier

□ fffffa8008ef5018-4000013700000002
□ fffffa8008ef5018-4000013700000003
□ fffffa8008ef5018-4000013700000004
□ fffffa8008ef5018-4000013700000005

To add a session, click Add session.

To disconnect one or more sessions, select each session and then click Disconnect.

To view devices associated with a session, select a session and then click Devices.

Session Information

Target portal group tag:	1030
Status:	Connected
Connection count:	1
Maximum Allowed Connections:	4
Authentication:	None Specified
Header Digest:	None Specified
Data Digest:	None Specified

Configure Multiple Connected Session (MCS)

To add additional connections to a session or configure the MCS policy for a selected session, click MCS.

Figure 22-24 Properties - iSCSI Sessions

You have now completed all of the iSCSI configuration. Your Windows 2008 machine is ready to use any iSCSI LUNs through the iSCSI connections.

22.2.3 Creating an iSCSI LUN

In the following steps, we will make a 100 MB LUN for Windows 2008 host. You can create different types volumes instead of this kind of normal LUN which we are making here.

Before creating any LUN, keep in mind that when you create a volume, Data ONTAP automatically performs the following actions:

- Reserves 5% of the space for Snapshot copies
- Schedules Snapshot copies

Note: Use the following guidelines to create volumes that contain LUNs:

- Do not create any LUNs in the system’s root volume. Data ONTAP uses this volume to administer the storage system.
- Use a SAN volume to contain the LUN.
- Ensure that no other files or directories exist in the volume that contains the LUN.

In the OnCommand System Manager, proceed as follows:

1. Click several icons to make an iSCSI LUN. In our lab environment, the path is **Vservers** → **Cluster Name(cdot-cluster01)** → **iSCSI SVM(vs_iSCSI_01)** → **Storage** → ***LUNs***. See Figure 22-25.

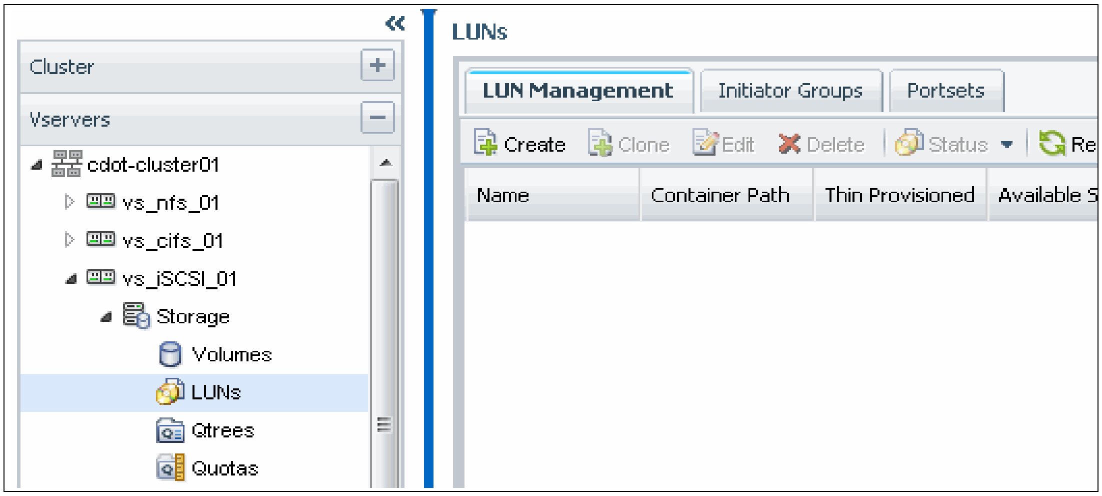
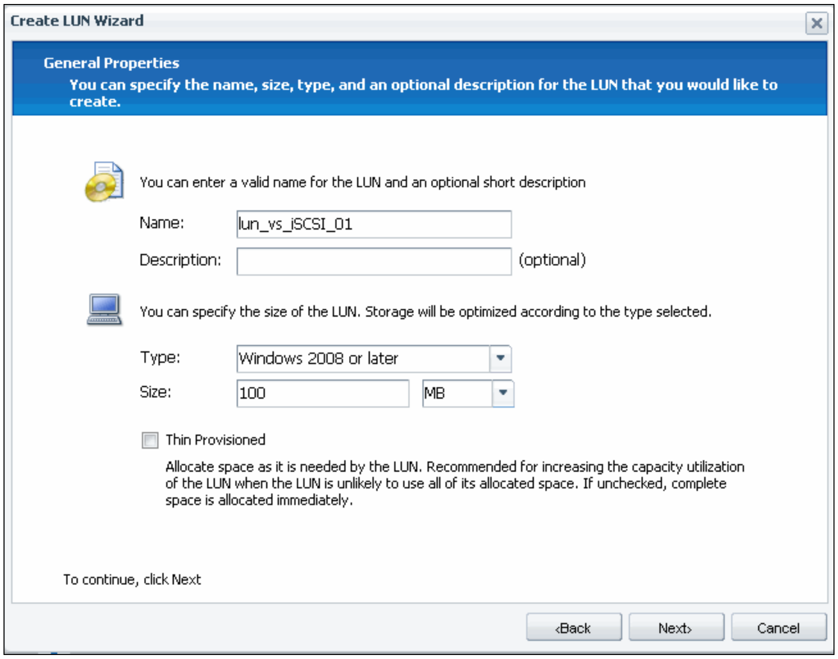


Figure 22-25 iSCSI SVM LUNs

2. Click the **Create** button to open the Create LUN Wizard. You will see the next window after you click the **Next** button. See Figure 22-26.



General Properties

You can specify the name, size, type, and an optional description for the LUN that you would like to create.

You can enter a valid name for the LUN and an optional short description

Name: lun_vs_iSCSI_01

Description: (optional)

You can specify the size of the LUN. Storage will be optimized according to the type selected.

Type: Windows 2008 or later

Size: 100 MB

□ Thin Provisioned

Allocate space as it is needed by the LUN. Recommended for increasing the capacity utilization of the LUN when the LUN is unlikely to use all of its allocated space. If unchecked, complete space is allocated immediately.

To continue, click Next

Figure 22-26 General Properties

3. You can create LUNs for an existing aggregate, volume, or qtree when there is available free space. Also, you can create a LUN in an existing volume or create a new FlexVol volume for the LUN.

   Here are some values you can enter in this panel:

   – Name: Specify the name of the iSCSI LUN which you want to make.

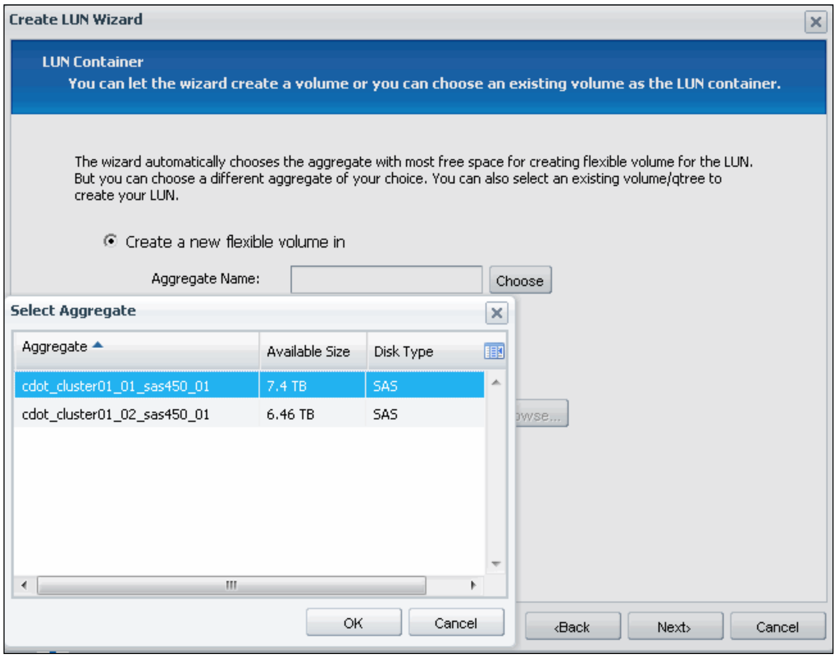
   – Type: The LUN multiprotocol type, or operating system type, specifies the operating system of the host accessing the LUN. It also determines the layout of data on the LUN, the geometry used to access that data, and the minimum and maximum size of the LUN.

     > **Note:** Not all Data ONTAP versions support all LUN multiprotocol types. To get the most up-to-date information, you should consult the Interoperability Matrix.

     In this step, we can select one of three Windows host types. There is Windows 2003 MBR, Windows 2003 GTP, and Windows 2008 or later. If your host operating system is Windows Server 2008 or later, both MBR and GPT partitioning methods are supported. But if you are using Windows Server 2003, check which partitioning method is supported.

   – Thin Provisioned: Specify whether thin provisioning is enabled.

4. In the next window, you can select an aggregate for this LUN. See Figure 22-27.



Create LUN Wizard

LUN Container

You can let the wizard create a volume or you can choose an existing volume as the LUN container.

The wizard automatically chooses the aggregate with most free space for creating flexible volume for the LUN. But you can choose a different aggregate of your choice. You can also select an existing volume/qtree to create your LUN.

● Create a new flexible volume in

Aggregate Name:

Select Aggregate

Aggregate ▲	Available Size	Disk Type
cdot_cluster01_01_sas450_01	7.4 TB	SAS
cdot_cluster01_02_sas450_01	6.46 TB	SAS

Figure 22-27 Selecting a aggregate for an iSCSI LUN

5. In this step, you can select an aggregate for the iSCSI LUN. You can assign any aggregate included in Clustered Data ONTAP. That means several Volumes and LUNs are located in each aggregate. To increase the performance of the specific LUNs such as this iSCSI LUN, you can select one of aggregates with low I/O workload.

   You can also specify the volume or accept the default volume name. And click **Next**.

6. On the Initiators Mapping page, you will define at least one more initiator group to make a connection between the iSCSI LUN and iSCSI Host. See Figure 22-28.

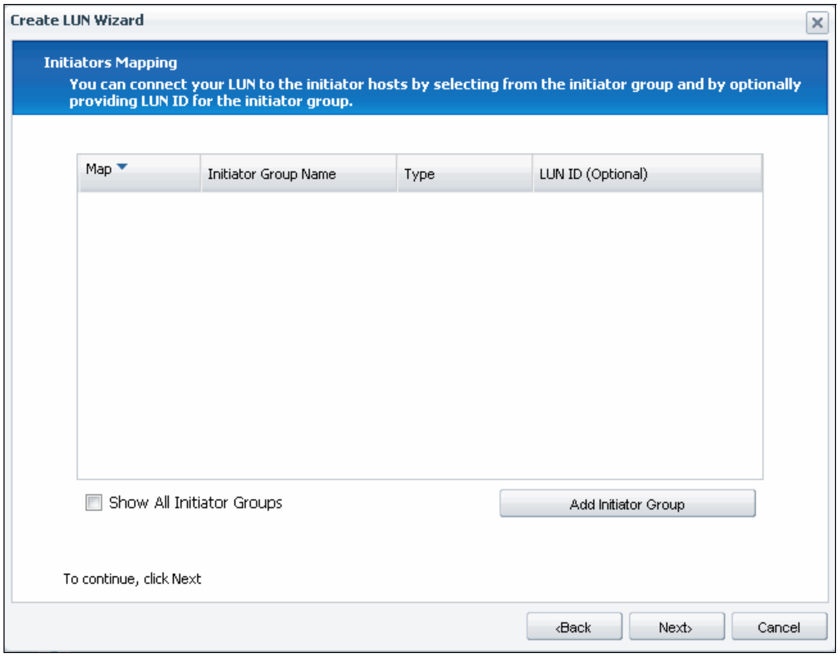
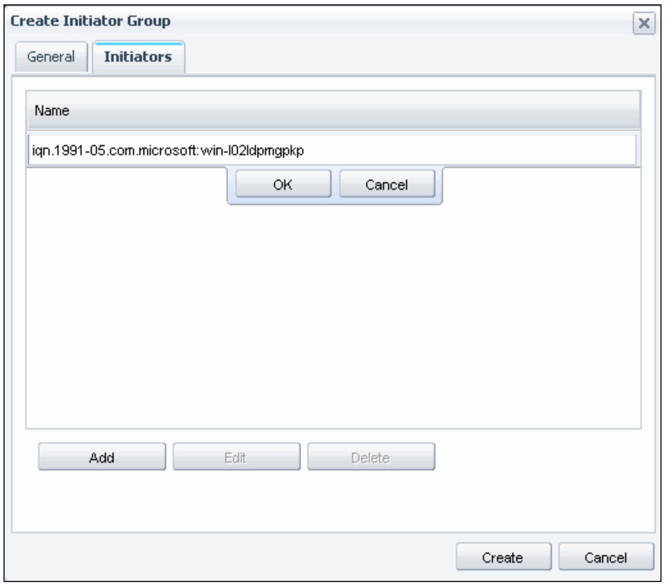

Create LUN Wizard
Initiators Mapping
You can connect your LUN to the initiator hosts by selecting from the initiator group and by optionally providing LUN ID for the initiator group.
Map
Initiator Group Name
Type
LUN ID (Optional)
□ Show All Initiator Groups
Add Initiator Group
To continue, click Next
‹Back
Next›
Cancel

Figure 22-28 Initiators Mapping

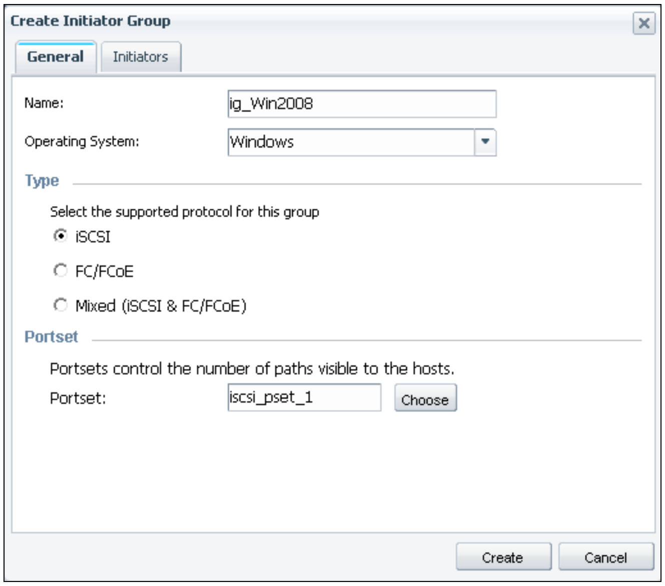
7. The next step is to create an iSCSI initiator group with the name of IQN in Windows 2008. Click **Add Initiator Group** button. You must specify the name of IQN which was made before in Windows 2008. See Figure 22-29.



iqn.1991-05.com.microsoft:win-l02ldpmgpkp

Figure 22-29 Adding initiators with the IQN name

8. Going back to the **General** tab, specify the rest values for the iSCSI initiator group. See Figure 22-30.



Name: ig_Win2008

Operating System: Windows

Type

Select the supported protocol for this group

● iSCSI

○ FC/FCoE

○ Mixed (iSCSI & FC/FCoE)

Portset

Portsets control the number of paths visible to the hosts.

Portset: iscsi_pset_1

Figure 22-30 General tab in a Initiator group setting

Here are the values that you will need to enter:

– Name: Specify an initiator group name you want.

– Operating system: Specify an operating system type.

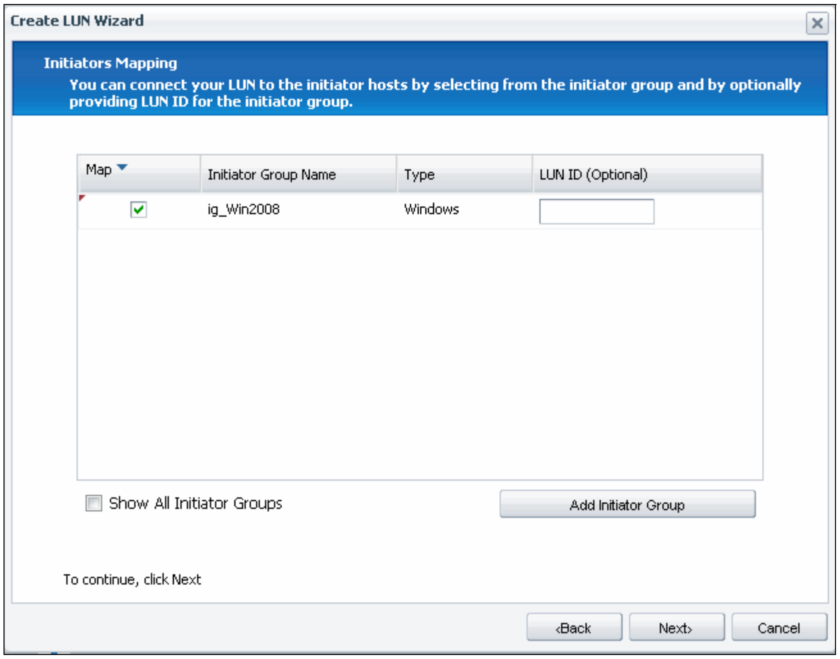
> **Note:** The operating system types is different than the previously selected OS type. This configuration is only for an iSCSI initiator group.

– Type: Specify the supported protocol for the group.

– Portset: Select the portset for the iSCSI connections.

> **Note:** A port set is a collection of LIFs. If you do not bind those igroups to a port set, any initiator within an igroup can access any LUN mapped to that igroup through any LIF.
>
> The way an initiator connects to a LUN is through an igroup. In the following example, initiator1 can access LUN1 through LIF1 and LIF2. Without a port set, initiator1 can access LUN1 over any LIF.

9. When you have completed specifying all of these values, click the **Create** button. Back at the Initiators Mapping page, verify that the new igroup has been added to the list. Also select the **Map** check box to the left of the igroup and click **Next**. See Figure 22-31.



Create LUN Wizard

Initiators Mapping

You can connect your LUN to the initiator hosts by selecting from the initiator group and by optionally providing LUN ID for the initiator group.

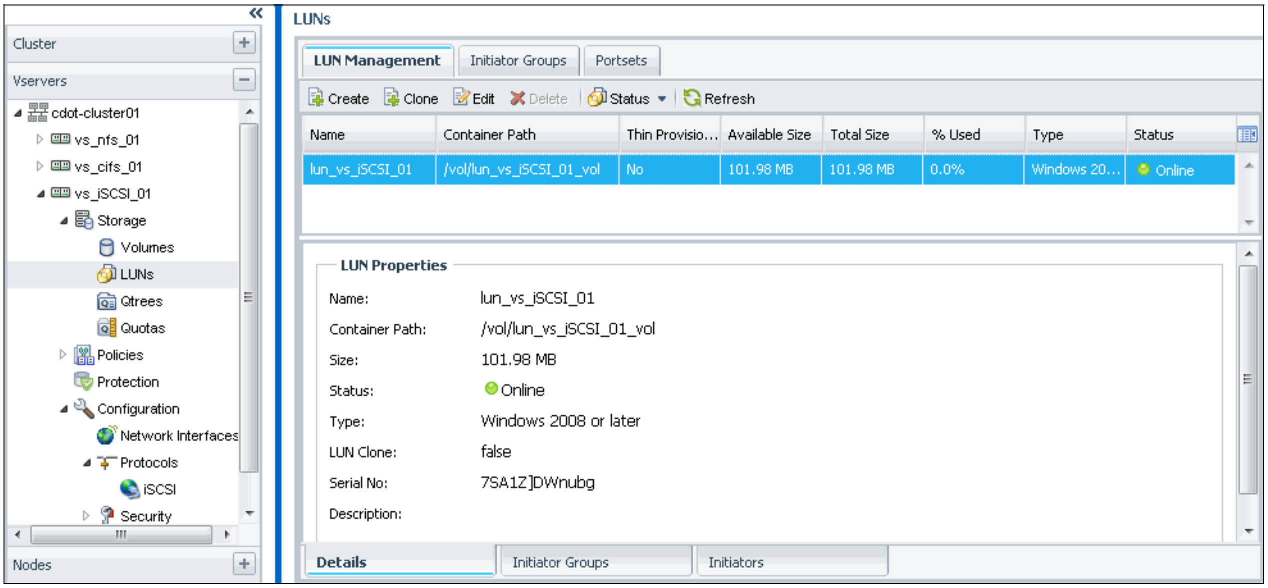
Map	Initiator Group Name	Type	LUN ID (Optional)
☑	ig_Win2008	Windows	

□ Show All Initiator Groups

To continue, click Next

Figure 22-31 Initiators Mapping finish

10.When the iSCSI LUN creation is done, you can see the following panel on the OnCommand System Manager. See Figure 22-32.



Name	Container Path	Thin Provisio...	Available Size	Total Size	% Used	Type	Status
lun_vs_iSCSI_01	/vol/lun_vs_iSCSI_01_vol	No	101.98 MB	101.98 MB	0.0%	Windows 20...	Online

LUN Properties

Name: lun_vs_iSCSI_01
Container Path: /vol/lun_vs_iSCSI_01_vol
Size: 101.98 MB
Status: Online
Type: Windows 2008 or later
LUN Clone: false
Serial No: 7SA1Z]DWnubg
Description:

Figure 22-32 Checking iSCSI LUN status

All of iSCSI related activities have been done. Now we will check that the Windows 2008 host can see the iSCSI LUN correctly.

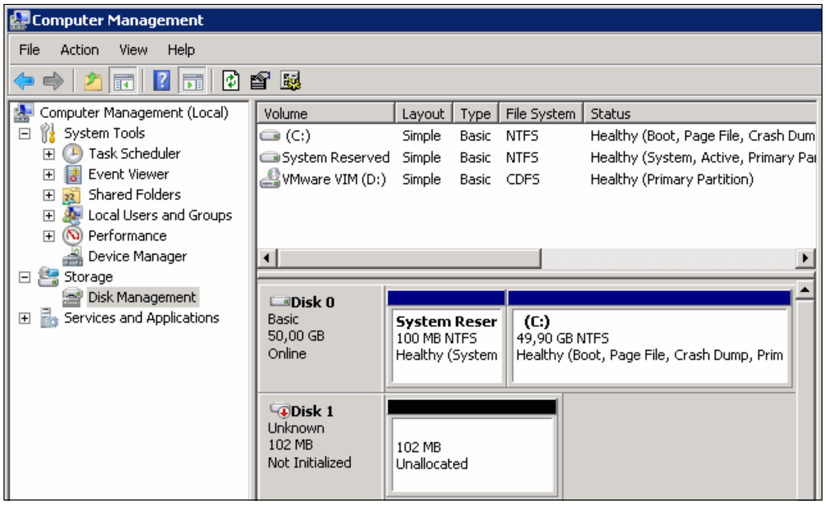
22.3 Accessing the iSCSI LUN on the Windows 2008 Host

Here are the final steps needed to access an iSCSI LUN from Windows 2008. This is a simple way to configure iSCSI mapping on the Windows host, so we will describe it briefly.

Checking whether an iSCSI LUN was recognized or not

We can easily check that a Windows host can access the iSCSI LUN:

1. Open **Computer Management MMC** and check the output of **Disk Management**. See Figure 22-33.



Volume	Layout	Type	File System	Status
(C:)	Simple	Basic	NTFS	Healthy (Boot, Page File, Crash Dum
System Reserved	Simple	Basic	NTFS	Healthy (System, Active, Primary Pa
VMware VIM (D:)	Simple	Basic	CDFS	Healthy (Primary Partition)

Disk 0
Basic
50,00 GB
Online

System Reser
100 MB NTFS
Healthy (System

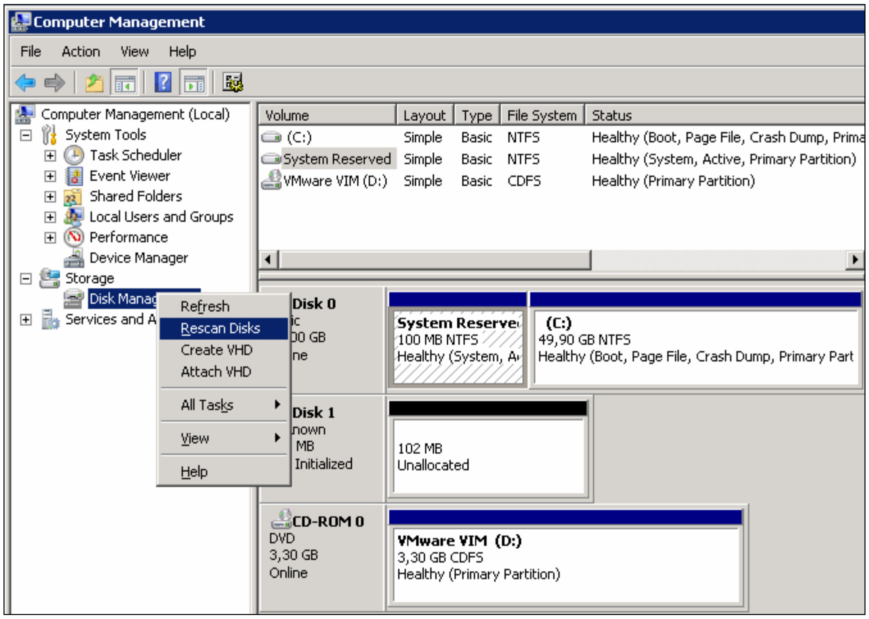
(C:)
49,90 GB NTFS
Healthy (Boot, Page File, Crash Dump, Prim

Disk 1
Unknown
102 MB
Not Initialized

102 MB
Unallocated

Figure 22-33 Checking the output of Disk Management

2. If you do not see the LUN disk in the bottom section of the center pane, right-click the Disk Management node in the left pane and select **Rescan Disks**. See Figure 22-34.



Volume	Layout	Type	File System	Status
(C:)	Simple	Basic	NTFS	Healthy (Boot, Page File, Crash Dump, Prima
System Reserved	Simple	Basic	NTFS	Healthy (System, Active, Primary Partition)
VMware VIM (D:)	Simple	Basic	CDFS	Healthy (Primary Partition)

Disk 0
ic
00 GB
ne

System Reserve
100 MB NTFS
Healthy (System, A

(C:)
49,90 GB NTFS
Healthy (Boot, Page File, Crash Dump, Primary Part

Disk 1
nown
MB
Initialized

102 MB
Unallocated

CD-ROM 0
DVD
3,30 GB
Online

VMware VIM (D:)
3,30 GB CDFS
Healthy (Primary Partition)

Figure 22-34 Rescan Disks

3. If the disk is offline, select **online**. Then you can select **Initialize Disk** as in Figure 22-35.

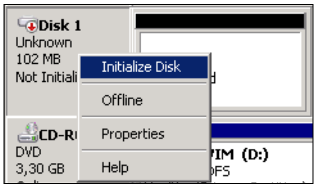
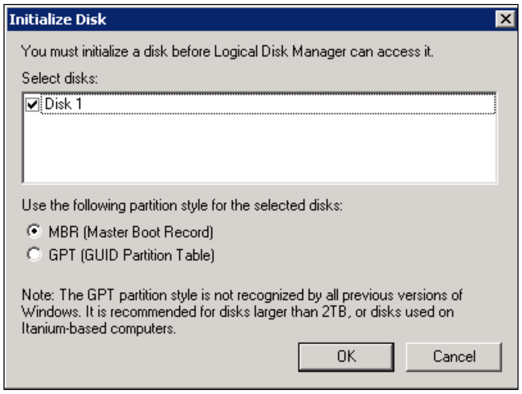


Figure 22-35 Initialize Disk

4. Now you will see the panel in Figure 22-36 where you can select a partition style.



You must initialize a disk before Logical Disk Manager can access it.

Select disks:

☑ Disk 1

Use the following partition style for the selected disks:

● MBR (Master Boot Record)

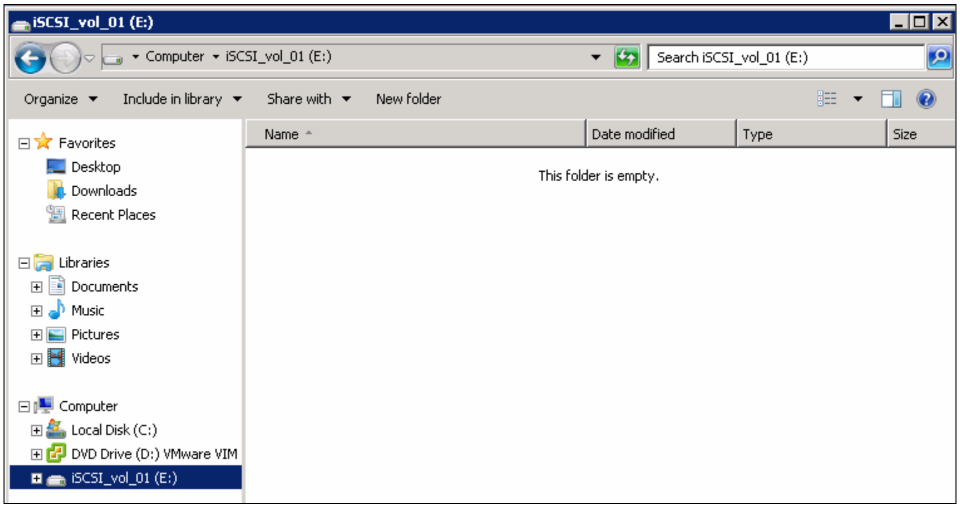
○ GPT (GUID Partition Table)

Note: The GPT partition style is not recognized by all previous versions of Windows. It is recommended for disks larger than 2TB, or disks used on Itanium-based computers.

Figure 22-36 Initialize Disk-Selecting Partition Style

5. Then normally you can assign the iSCSI LUN as a simple volume and format the drive with one of file system formats you want to use.

6. After that, you can see the iSCSI LUN through Windows Explorer. See Figure 22-37.



This folder is empty.

Figure 22-37 Accessing the iSCSI LUN

22.4 Further information

More details on iSCSI implementation can be found in the following IBM support documents:

- *OnCommand System Manager 3.0 Help for Use with Clustered Data ONTAP Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

- *Windows Host Utilities 6.0.2 Release Notes*, located at this website:

  http://www.ibm.com/support/docview.wss?uid=ssg1S7003280

- *Fibre Channel and iSCSI Configuration Guide for the Data ONTAP 8.0 Release Family*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

23

CIFS storage

This chapter shows a realistic example for Common Internet File System (CIFS) storage that consists of a Clustered Data ONTAP system and a Windows 2008 server having several Ethernet ports.

We show you how to configure a storage virtual machine (SVM) for a Windows Host. Also, you will learn to configure several features such as creating a home directory, monitoring SMB statistics, and so on.

In the Data ONTAP 8.2 operating system, CIFS is currently supported in clusters of up to 24 nodes.

The following topics are covered:

- Planning and checking the CIFS environment
- Configuring CIFS for both Windows 2008 and Clustered Data ONTAP
- Monitoring SMB statistics
- Further information

23.1 Planning and checking the CIFS environment

In this chapter, you will create an SVM for Windows 2008 CIFS through an OnCommand System Manager GUI. According to the official procedure, you will check the interoperability and configure both a Windows 2008 host and a Clustered Data ONTAP system.

A cluster is a scalable system that allows you to group pairs of controllers, share physical resources, and distribute workloads across the systems while consolidating management for administrators into one unified interface. The Clustered Data ONTAP 8.2 storage system has the capability to join up to 24 individual nodes or 12 pairs of nodes.

23.1.1 Checking your environment before configuration of CIFS

Before configuring CIFS, you will need to perform several checks as described next.

Checking the interoperability

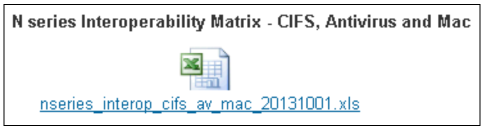
Here is the link to the N series interoperability matrix:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003897

In this case, we will check the interoperability between Windows 2008 and Clustered Data ONTAP. You can check CIFS interoperability with the following link, which is included in the prior link:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003769

At the time of writing this chapter, the latest interoperability matrix was made on 01 October 2013, as follows. See Figure 23-1.



N series Interoperability Matrix - CIFS, Antivirus and Mac

nseries_interop_cifs_av_mac_20131001.xls

Figure 23-1 N series Interoperability Matrix - CIFS, Antivirus and Mac

The support information is organized into the following topics in this interoperability matrix:

- Certified Windows Domain Controllers versions
- Certified Windows Hosts versions
- Supported CPU types
- Supported Off-box Anti-virus versions
- Information and alerts:

  Refer to Information and alerts which include important information regarding OS patches, considerations and so on.

> **Note:** It is crucial to check the interoperability before the implementation and the planning phase. If you cannot find any certified solution, you must submit a SCORE/RPQ(Request for Price Quotations). To submit a SCORE/RPQ, contact your IBM Representative.

23.2 Configuring CIFS for both Windows 2008 and Clustered Data ONTAP

In a Clustered Data ONTAP environment, you must create an SVM to provide CIFS services to the host. We assume that you already know how to configure the basic cluster system.

23.2.1 Creating an SVM for CIFS

Before creating an SVM on Clustered Data ONTAP, you must check the interoperability, as you have already checked the interoperability matrix. Normally there will be some important information in the interoperability matrix Excel file. See Example 23-1.

Example 23-1 CIFS Interoperability matrix

```
Info 6868 IPV6 support is not available for Windows XP and Windows 2003 clients
```

As you can see from this information, if you have any clients using Windows XP and Windows 2003 clients, you must configure the SVM as an IPv4 network.

Prerequisites

Before proceeding the next step, check the following items:

- The CIFS license must be installed on your storage system.
- While configuring CIFS in the Active Directory domain, you must ensure that the following requirements are met:
  - DNS must be enabled and configured correctly.
  - The storage system must be able to communicate with the domain controller using the fully qualified domain name (FQDN).
  - The time differences (clock skew) between the storage system time and the domain time must not be more than the skew time that is configured in Data ONTAP.
- If CIFS is the only protocol configured on the SVM, you must ensure that the following requirements are met:
  a. The root volume security style must be NTFS.

  > **Note:** By default, System Manager sets the security style as UNIX.

  b. Superuser access must be set to Any for CIFS protocol.

Preferred Domain Controllers (Optional)

The storage server automatically discovers Active Directory services during server setup and afterward based on various criteria such as the setup on the storage server, the Active Directory site setup, the distance from the storage server to the targeting services, and so on.

The storage server administrator can explicitly set up a list of Windows DCs to connect to as its top preference for various CIFS operations. To set up the preferred DCs, use the command shown in Example 23-2.

Example 23-2 Setting up the preferred DCs

```
cdot-cluster01::> vserver cifs domain preferred-dc create -vserver <vserver>
-domain <Fully Qualified Domain Name> -preferred-dc <InetAddress>, ...
```

Creating an SVM for CIFS

To provide CIFS services to specific hosts, we must make at least one CIFS SVM for CIFS services. In this case, we will make one SVM for one Windows 2008 server.

To create an SVM, proceed as follows:

1. Click **Vservers** in the System Manager navigation frame.
2. Click the **Create** button on the System Manager Vservers page as shown in Figure 23-2.

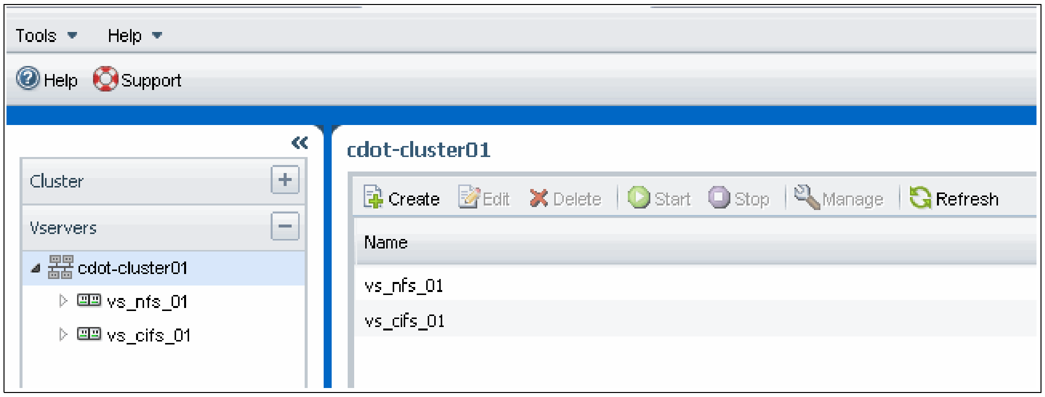


Figure 23-2 Create an SVM for CIFS

3. Next you see the Vserver Setup wizard as shown in Figure 23-3. You need to specify a Vserver name and several options in this window.

   Here we list some values to be entered:

   – Vserver Name: Define a name for the SVM that you want to name.

   – Data Protocols: Each SVM can be a server that provides multiple protocols or at least one protocol. In this chapter, select CIFS only.

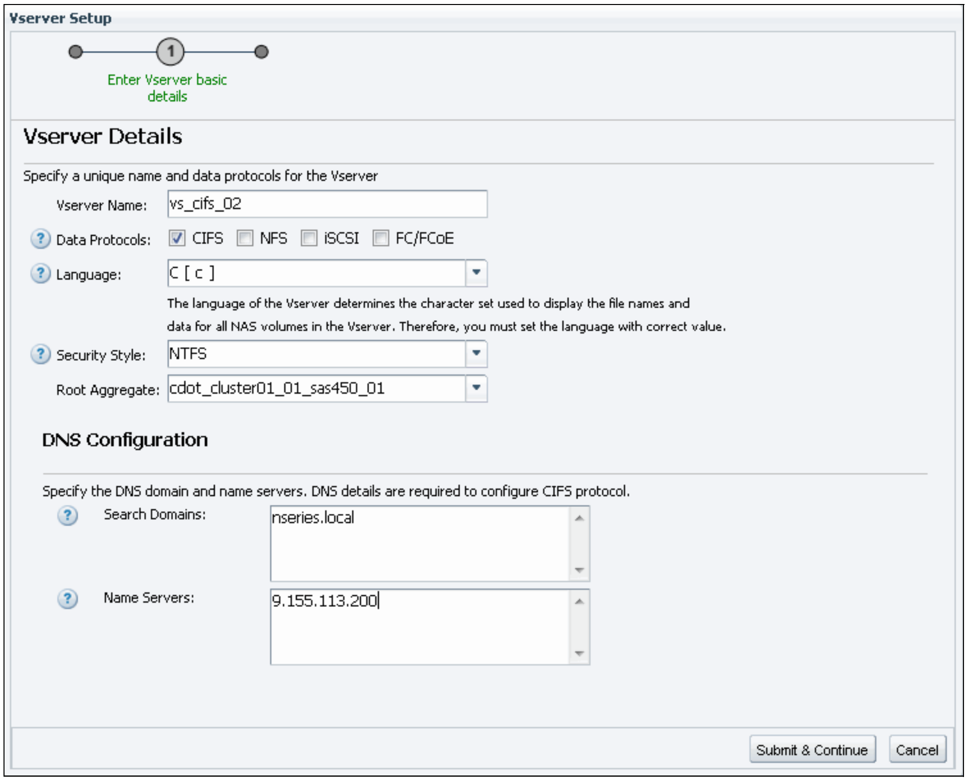
   – Language: This parameter depends on the specific characters that are used by your language environment. The default language is set to C.UTF-8. If your country uses two-byte characters, select a correct one.

   > **Note:** The language of an SVM with FlexVol volumes can be modified after the SVM is created.

   – Security style: There are three security styles. UNIX, NTFS, and Mixed are the options. You can decide what security style to use on a volume, and you should consider two factors. The primary factor is the type of administrator that manages the file system. The secondary factor is the type of user or service that accesses the data on the volume.

   – Root Aggregate: Select a root aggregate from one of the node root aggregates.

4. When you have made your selections, click the **Submit & Continue** button.



Vserver Details

Specify a unique name and data protocols for the Vserver

Vserver Name: vs_cifs_02

Data Protocols: ☑ CIFS ☐ NFS ☐ iSCSI ☐ FC/FCoE

Language: C [c]

The language of the Vserver determines the character set used to display the file names and data for all NAS volumes in the Vserver. Therefore, you must set the language with correct value.

Security Style: NTFS

Root Aggregate: cdot_cluster01_01_sas450_01

DNS Configuration

Specify the DNS domain and name servers. DNS details are required to configure CIFS protocol.

Search Domains: nseries.local

Name Servers: 9.155.113.200

Figure 23-3 Vserver Details

5. If you want to make and configure your SVM through the shell command, see the following command:

   vserver create -vserver *vserver_name* **-aggregate** *aggregate_name* **- rootvolume** *root_volume_name* **-rootvolume-security-style** {*unix*|*ntfs*|*mixed*} **-ns-switch** {*nis*|*file*|*ldap*},... **[-nm-switch** {*file*|*ldap*},...**] [-language** *language* **[-snapshot-policy** *snapshot_policy_name*] **[-quota-policy** *quota_policy_name*] **-comment comment]**

   – **-ns-switch** specifies which directory stores to use for UNIX user and group information and the order in which they are searched.

   – **-nm-switch** specifies which directory stores to use for name mapping information and the order in which they are searched.

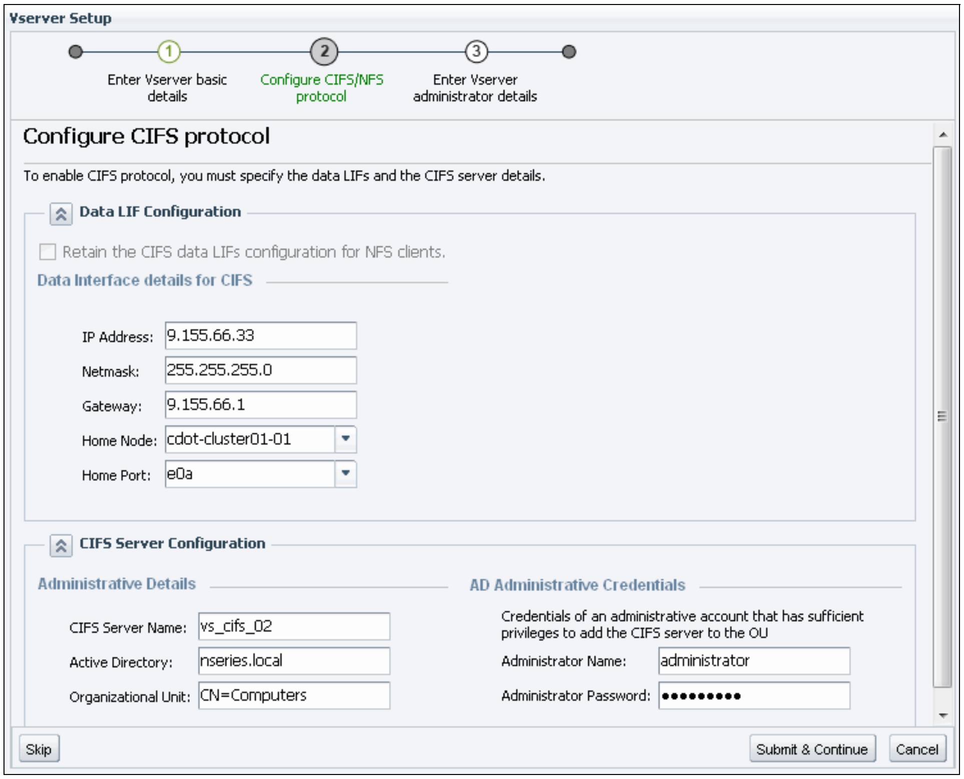
   Example 23-3 shows a typical use of these commands.

Example 23-3 Create an SVM for CIFS

```
cdot-cluster01::> vserver create -vserver vs_cifs_02 -rootvolume vs_cifs_02_root
-aggregate cdot_cluster01_01_sas450_01 -rootvolume-security-style ntfs -ns-switch
file -nm-switch file
cdot-cluster01::> vserver show -vserver vs_cifs_02

                                    Vserver: vs_cifs_02
                               Vserver Type: data
                               Vserver UUID: 1af2ab91-3a49-11e3-ba64-123478563412
                                Root Volume: vs_cifs_02_root
                                  Aggregate: cdot_cluster01_01_sas450_01
                        Name Service Switch: file
                        Name Mapping Switch: file
```

```
                              NIS Domain: -
                Root Volume Security Style: ntfs
                               LDAP Client: -
              Default Volume Language Code: C
                           Snapshot Policy: default
                                   Comment:
                Antivirus On-Access Policy: default
                              Quota Policy: default
               List of Aggregates Assigned: cdot_cluster01_01_sas450_01,
                                            cdot_cluster01_02_sas450_01
Limit on Maximum Number of Volumes allowed: unlimited
                       Vserver Admin State: running
                         Allowed Protocols: cifs
                      Disallowed Protocols: nfs, fcp, iscsi, ndmp
            Is Vserver with Infinite Volume: false
                          QoS Policy Group: -
```

6. Next, you see the panel shown in Figure 23-4.


Vserver Setup
1
Enter Vserver basic details
2
Configure CIFS/NFS protocol
3
Enter Vserver administrator details
Skip
Submit & Continue
Cancel

Configure CIFS protocol

To enable CIFS protocol, you must specify the data LIFs and the CIFS server details.

Data LIF Configuration

☐ Retain the CIFS data LIFs configuration for NFS clients.

Data Interface details for CIFS

IP Address: 9.155.66.33

Netmask: 255.255.255.0

Gateway: 9.155.66.1

Home Node: cdot-cluster01-01

Home Port: e0a

CIFS Server Configuration

Administrative Details

CIFS Server Name: vs_cifs_02

Active Directory: nseries.local

Organizational Unit: CN=Computers

AD Administrative Credentials

Credentials of an administrative account that has sufficient privileges to add the CIFS server to the OU

Administrator Name: administrator

Administrator Password: ●●●●●●●●●

Figure 23-4 Configure CIFS protocol

7. In this step, you can configure a CIFS protocol on the SVM to provide file-level data access for NAS clients.To enable CIFS protocol, you must create the data logical interfaces (LIFs) and the CIFS server.

   Before you begin, observe these considerations:

   – Protocols that you want to configure or allow on the SVM must be licensed.
   If the protocol is not allowed on the SVM, you can use the Edit Vserver window to enable the protocol for the SVM.

- You must have the Active Directory, Organizational unit, and administrative account credentials for configuring CIFS protocol.
- You must have the IP address, netmask, and gateway information to create the IP network interface on the SVM.

Here we explain some values in the boxes:

- Retain the CIFS data LIFs configuration for NFS clients check box:

  This option specifies that the data LIF supports both CIFS and NFS sessions. You can either retain the same data LIF configuration for both CIFS and NFS or configure a new LIF for each protocol.

- Data interface details for CIFS:

  This option specifies the IP addresses, netmask, and gateway. Also, you can specify the Home Node and Home port for this SVM.

- CIFS Server Configuration:

  Specify the following information to create a CIFS server:

  - CIFS server name
  - Active Directory to associate with the CIFS server
  - Organizational unit (OU) within the Active Directory domain to associate with the CIFS server. By default, this parameter is set to CN=Computers.
  - Credentials of an administrative account that has sufficient privileges to add the CIFS server to the OU

  Note: (Optional) You can also specify the IP addresses of the NIS servers and NIS domain name to configure NIS services on the SVM.

- AD Administrative Credentials:

  Specify the active directory administrator ID and password to register this SVM to your active directory environment.

8. Then create the CIFS SVM with the following command:

   vserver cifs create -vserver *vserver_name* **-domain** *FQDN* **[-ou** *organizational_unit***]**

   Example 23-4 illustrates these commands.

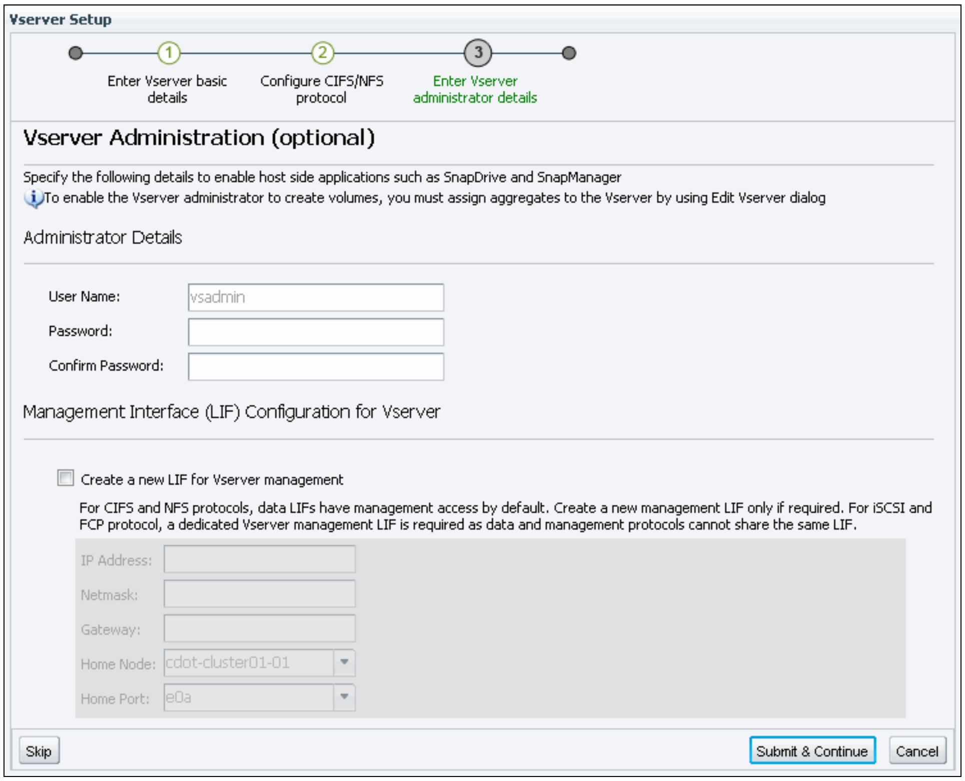
Example 23-4 Create a CIFS SVM

```
cdot-cluster01::> vserver cifs create -vserver vs_cifs_02 -domain NSERIES.LOCAL
-cifs-server VS_CIFS_02
cdot-cluster01::> vserver cifs show -vserver vs_cifs_02

                                          Vserver: vs_cifs_02
                         CIFS Server NetBIOS Name: VS_CIFS_02
                    NetBIOS Domain/Workgroup Name: NSERIES
                      Fully Qualified Domain Name: NSERIES.LOCAL
Default Site Used by LIFs Without Site Membership:
                             Authentication Style: domain
                CIFS Server Administrative Status: up
```

9. After completing these steps, click **Submit & Close**.

10. The next step is for configuring any administration information for each SVM. See Figure 23-5.

You can add an SVM user account and specify a login user method to access the storage system. This phase is optional.


Vserver Setup
1
Enter Vserver basic details
2
Configure CIFS/NFS protocol
3
Enter Vserver administrator details

Vserver Administration (optional)

Specify the following details to enable host side applications such as SnapDrive and SnapManager

To enable the Vserver administrator to create volumes, you must assign aggregates to the Vserver by using Edit Vserver dialog

Administrator Details

User Name: vsadmin

Password:

Confirm Password:

Management Interface (LIF) Configuration for Vserver

□ Create a new LIF for Vserver management

For CIFS and NFS protocols, data LIFs have management access by default. Create a new management LIF only if required. For iSCSI and FCP protocol, a dedicated Vserver management LIF is required as data and management protocols cannot share the same LIF.

IP Address:

Netmask:

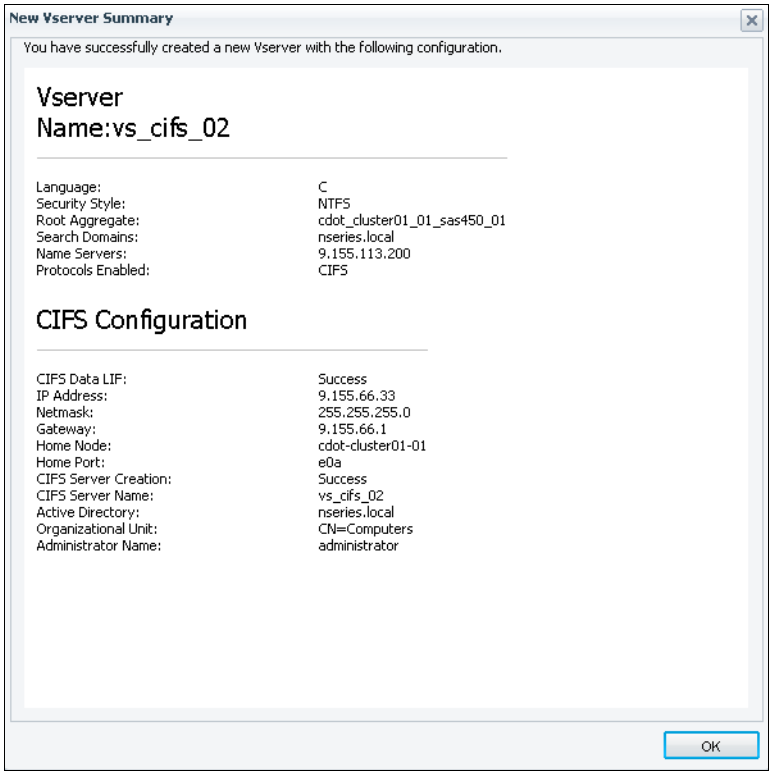
Gateway:

Home Node: cdot-cluster01-01

Home Port: e0a

Figure 23-5 Vserver Administration

11. After completing the CIFS SVM creation step, you can see the result panel in Figure 23-6.



You have successfully created a new Vserver with the following configuration.

Vserver Name:vs_cifs_02

Language:	C
Security Style:	NTFS
Root Aggregate:	cdot_cluster01_01_sas450_01
Search Domains:	nseries.local
Name Servers:	9.155.113.200
Protocols Enabled:	CIFS

CIFS Configuration

CIFS Data LIF:	Success
IP Address:	9.155.66.33
Netmask:	255.255.255.0
Gateway:	9.155.66.1
Home Node:	cdot-cluster01-01
Home Port:	e0a
CIFS Server Creation:	Success
CIFS Server Name:	vs_cifs_02
Active Directory:	nseries.local
Organizational Unit:	CN=Computers
Administrator Name:	administrator

Figure 23-6 CIFS SVM result panel

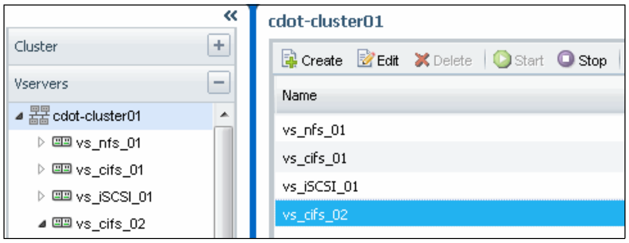
Configuring the CIFS SVM

After completion of the CIFS SVM creation, you can edit some default values to configure the CIFS SVM.

If you do not use an LDAP service as a name mapping switch, you can disable it through the following panel. See discussion. Also, you can assign any specific aggregates as CIFS data volumes.

To edit the CIFS SVM property, proceed as follows:

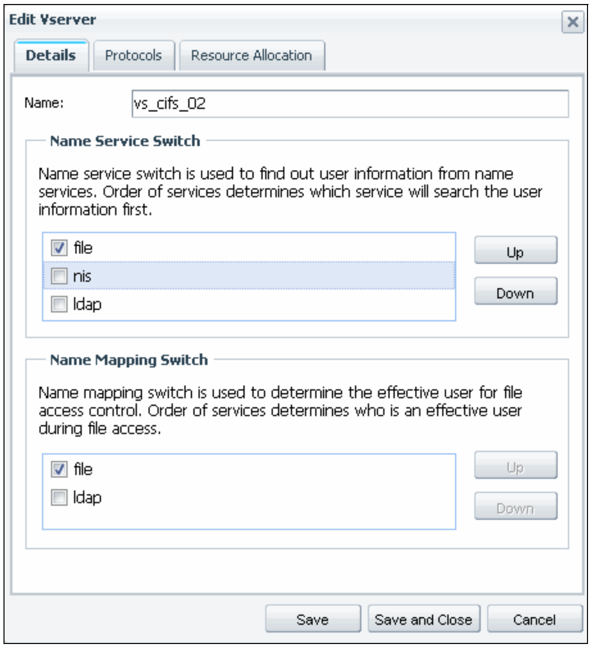
1. Click Vserver and select the SVM you made.
2. Then click the **Edit** button. See Figure 23-7.



Name
vs_nfs_01
vs_cifs_01
vs_iSCSI_01
vs_cifs_02

Figure 23-7 Editing a CIFS SVM

3. If you do not have an LDAP or NIS environment, you can clear any other check boxes except for the **file** check boxes. See Figure 23-8.



Name: vs_cifs_02

Name Service Switch

Name service switch is used to find out user information from name services. Order of services determines which service will search the user information first.

☑ file
☐ nis
☐ ldap

Name Mapping Switch

Name mapping switch is used to determine the effective user for file access control. Order of services determines who is an effective user during file access.

☑ file
☐ ldap

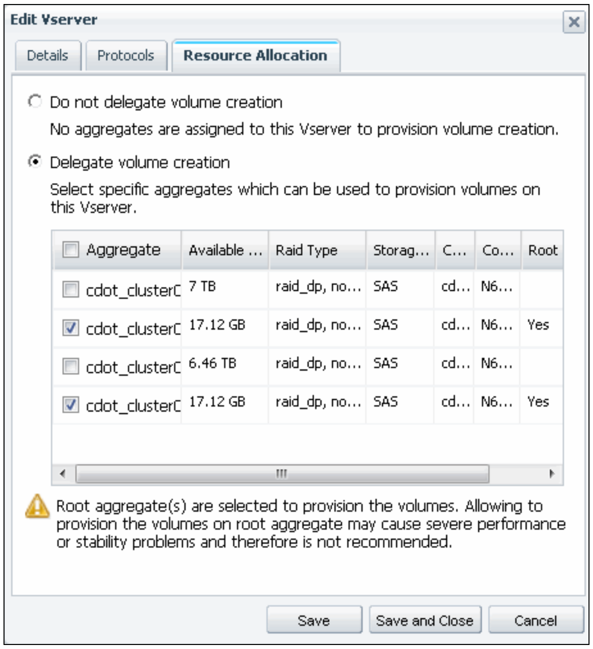
Figure 23-8 Name Server Switch and Name Mapping Switch panel

4. Click the **Resource Allocation** tab as shown in Figure 23-9:

   a. In this window, you can select any specific aggregate for the volumes for CIFS. Through this option, you can assign the CIFS volumes on the specific aggregates for your purpose.

   b. If you select root aggregates, the GUI will warn you because this may cause severe performance or stability problems. Therefore, you need to avoid assigning aggregates that include any root aggregate.

   See Figure 23-9 for resource allocation.



○ Do not delegate volume creation

No aggregates are assigned to this Vserver to provision volume creation.

● Delegate volume creation

Select specific aggregates which can be used to provision volumes on this Vserver.

☐ Aggregate	Available ...	Raid Type	Storag...	C...	Co...	Root
☐ cdot_clusterC	7 TB	raid_dp, no...	SAS	cd...	N6...	
☑ cdot_clusterC	17.12 GB	raid_dp, no...	SAS	cd...	N6...	Yes
☐ cdot_clusterC	6.46 TB	raid_dp, no...	SAS	cd...	N6...	
☑ cdot_clusterC	17.12 GB	raid_dp, no...	SAS	cd...	N6...	Yes

⚠ Root aggregate(s) are selected to provision the volumes. Allowing to provision the volumes on root aggregate may cause severe performance or stability problems and therefore is not recommended.

Figure 23-9 Resource allocation

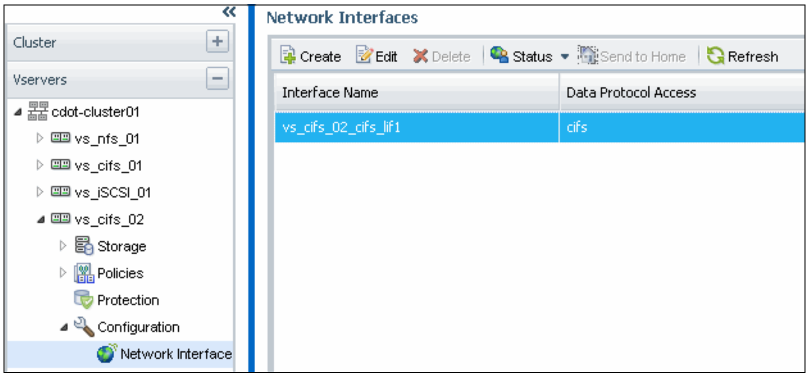
5. Click **Save and Close**.

23.2.2 Creating a NAS data LIF

In these steps, you will create a second CIFS data LIF with some other options. As you did in the previous task, the SVM setup wizard configured a data LIF with some default values.

To create additional data LIFs, proceed as follows:

1. Select the **Network Interfaces**.
2. Click the **Create** button. You can find that through the following path. See Figure 23-10.

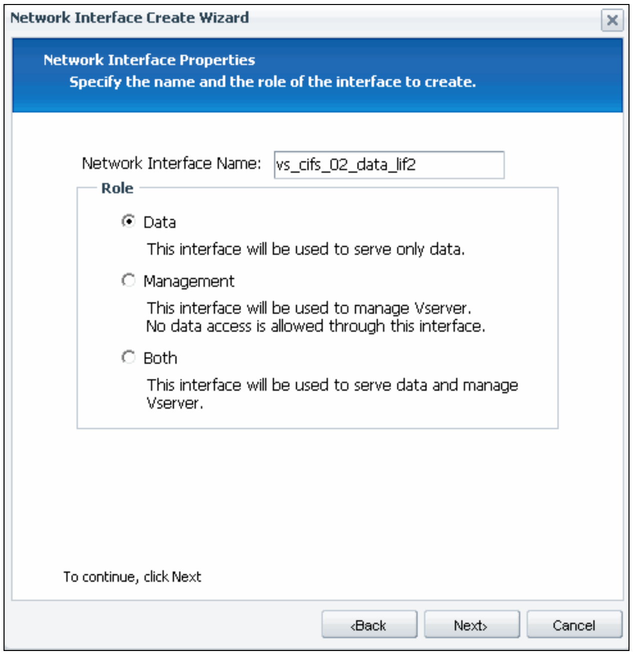


Interface Name	Data Protocol Access
vs_cifs_02_cifs_lif1	cifs

Figure 23-10 Create additional data LIF

3. After clicking the **Next** button on the welcome panel, you can see the panel in Figure 23-11.

4. In this step, you can specify the network interface name and select one of the data types for the CIFS SVM. The role can be data, management, or both of them. Click the **Next** button.



Network Interface Create Wizard

Network Interface Properties
Specify the name and the role of the interface to create.

Network Interface Name: vs_cifs_02_data_lif2

Role

● Data
This interface will be used to serve only data.

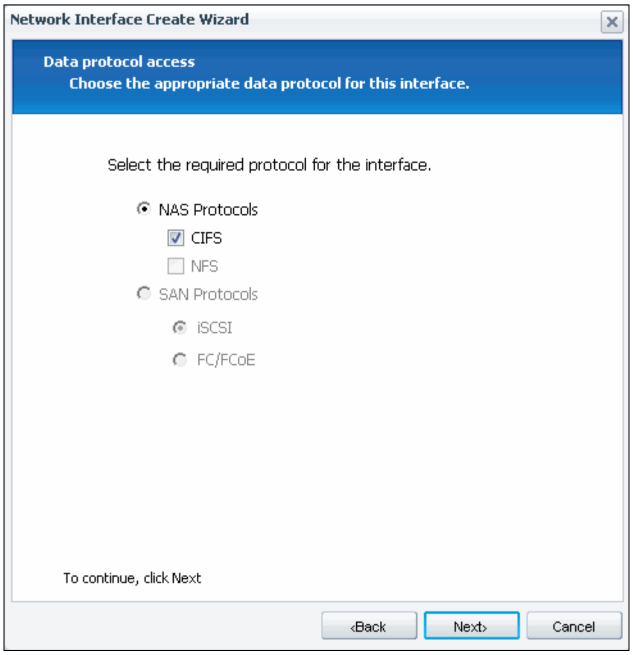
○ Management
This interface will be used to manage Vserver.
No data access is allowed through this interface.

○ Both
This interface will be used to serve data and manage Vserver.

To continue, click Next

Figure 23-11 Network Interface Create Wizard - Role

5. In this step, you can select which NAS protocols will be provided by this SVM. If you configured this SVM for both CIFS and NFS, you can select both protocols or either one of them. See Figure 23-12. Click the **Next** button.



Network Interface Create Wizard

Data protocol access
Choose the appropriate data protocol for this interface.

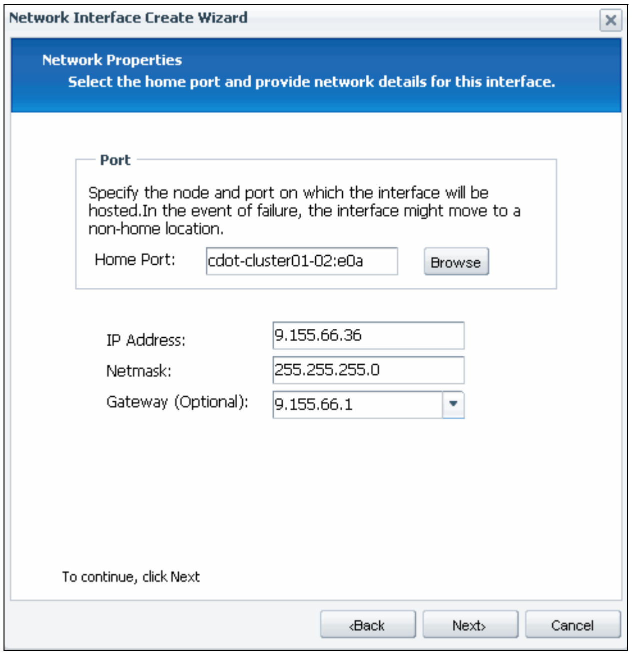
Select the required protocol for the interface.

● NAS Protocols
☑ CIFS
☐ NFS
○ SAN Protocols
● iSCSI
○ FC/FCoE

To continue, click Next

Figure 23-12 Data protocol access - NAS protocols

6. In this step, you can specify the home port and additional IP address information including IP address, netmask, and Gateway (optional). See Figure 23-13. Click the **Next** button.



Network Interface Create Wizard

Network Properties

Select the home port and provide network details for this interface.

Port

Specify the node and port on which the interface will be hosted.In the event of failure, the interface might move to a non-home location.

Home Port: cdot-cluster01-02:e0a

IP Address: 9.155.66.36

Netmask: 255.255.255.0

Gateway (Optional): 9.155.66.1

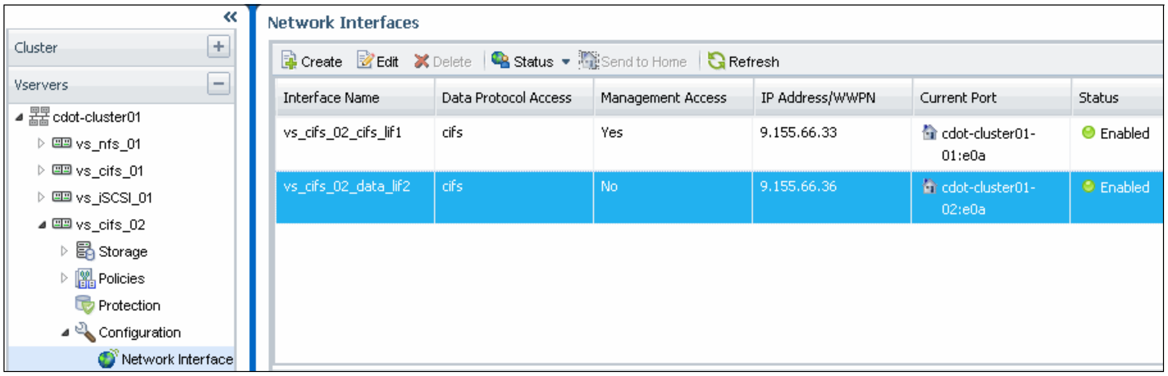
To continue, click Next

Figure 23-13 Network properties

Note: The combination of home node and home port is what determines which physical port on which physical node will be the home of this LIF. The home concept exists because data LIFs can migrate to other ports on the same node or to any other node in the cluster.

7. After completing these steps, review the summary. If you need to make any changes from the previous steps, click the **Back** button. If all values are correct, click the **Next** button to activate the additional data LIF and then click **Finish**.

8. You can verify the creation of the second LIF as shown in Figure 23-14.



Network Interfaces

Interface Name	Data Protocol Access	Management Access	IP Address/WWPN	Current Port	Status
vs_cifs_02_cifs_lif1	cifs	Yes	9.155.66.33	cdot-cluster01-01:e0a	Enabled
vs_cifs_02_data_lif2	cifs	No	9.155.66.36	cdot-cluster01-02:e0a	Enabled

Figure 23-14 Network Interfaces in the CIFS SVM

23.2.3 Creating an export policy

When a volume is created, it automatically inherits the default export policy of the root volume of the SVM. You can change the default export policy associated with the volume to redefine the client access to data. Or you can create an additional export policy as well.

To create an export policy, proceed as follows:

1. Select **Export** policies as shown in Figure 23-15.
2. Click the **Create Policy** button.

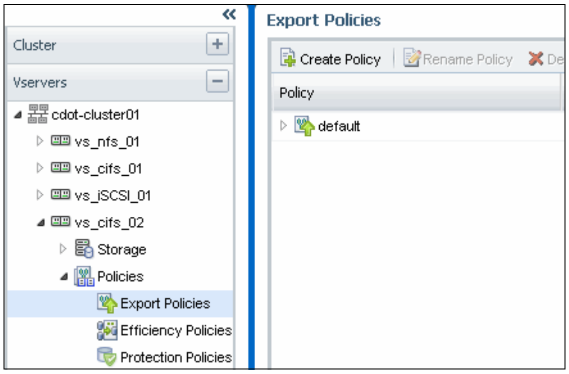
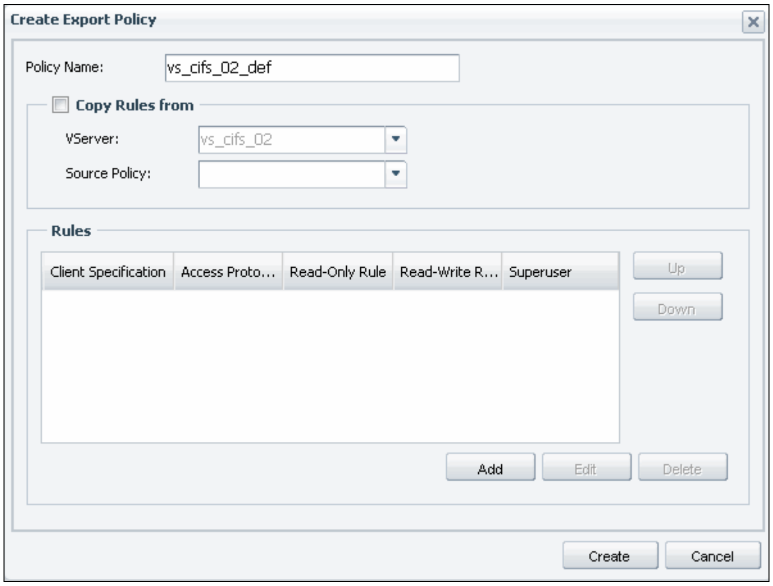


Figure 23-15 Create an export policy - Export Policies

3. In this step, you can specify the policy name and add a rule. In Figure 23-16, click the **Add** button to specify any rule for CIFS clients.


Create Export Policy

Policy Name: vs_cifs_02_def

□ Copy Rules from

VServer: vs_cifs_02

Source Policy:

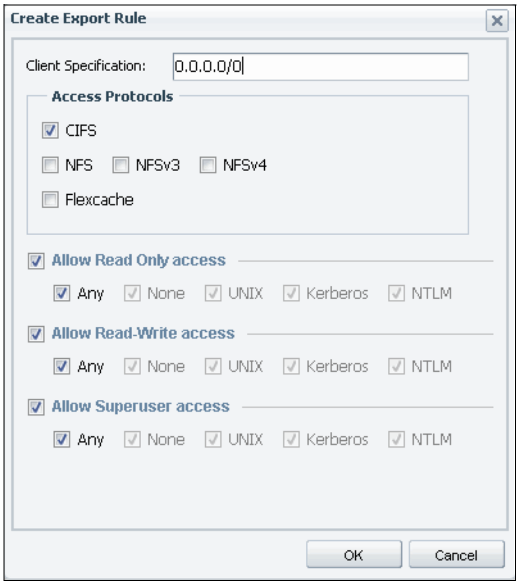
Rules

Client Specification	Access Proto...	Read-Only Rule	Read-Write R...	Superuser

Figure 23-16 Create Export Policy

4. In Figure 23-17, you can specify the match in any of the following formats:
   – As a host name; for example, host1
   – As an IPv4 address; for example, 10.1.12.24
   – As an IPv4 address with a network mask; for example, 10.1.16.0/255.255.255.0.
   – As a netgroup, with the netgroup name preceded by the @ character; for example, @netgroup
   – As a domain name preceded by the "." symbol; for example, .example.com

> **Note:** 0.0.0.0/0 means every client.



Create Export Rule

Client Specification: 0.0.0.0/0

Access Protocols

☑ CIFS

☐ NFS ☐ NFSv3 ☐ NFSv4

☐ Flexcache

☑ Allow Read Only access

☑ Any ☑ None ☑ UNIX ☑ Kerberos ☑ NTLM

☑ Allow Read-Write access

☑ Any ☑ None ☑ UNIX ☑ Kerberos ☑ NTLM

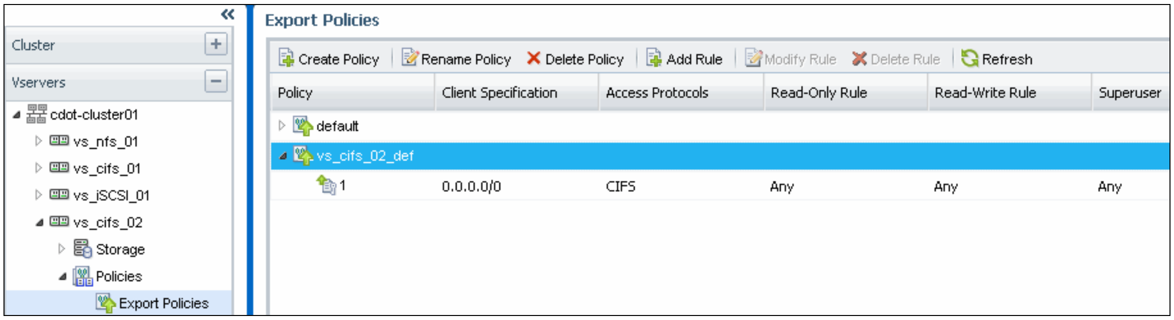
☑ Allow Superuser access

☑ Any ☑ None ☑ UNIX ☑ Kerberos ☑ NTLM

Figure 23-17 Client Specification

5. After completing all of these steps, click **OK** and then click **Create**.

   You can see the export policy that you made as shown in Figure 23-18.



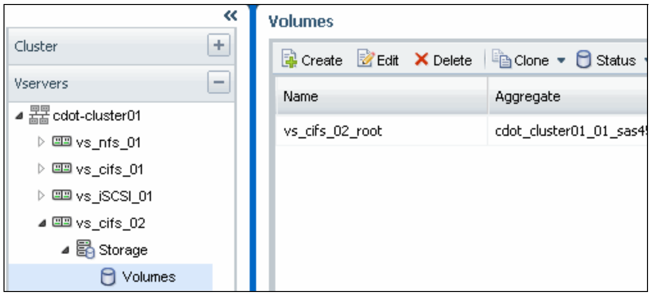
Policy	Client Specification	Access Protocols	Read-Only Rule	Read-Write Rule	Superuser
default					
vs_cifs_02_def					
1	0.0.0.0/0	CIFS	Any	Any	Any

Figure 23-18 Check the export policies made

23.2.4 Creating and exporting a volume

In these steps, you will make a CIFS volume with the CIFS SVM that you made previously. Proceed as follows:

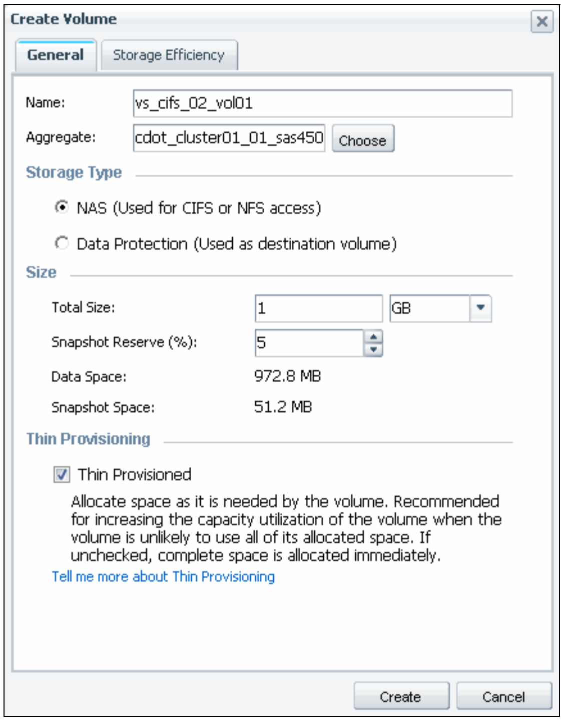
1. Select **Volumes** as in Figure 23-19 to make a CIFS volume.
2. Click the **Create** button.



Name	Aggregate
vs_cifs_02_root	cdot_cluster01_01_sas4

Figure 23-19 Create a CIFS volume

3. In this step, you can specify several parameters for a CIFS volume. See Figure 23-20.



Name: vs_cifs_02_vol01

Aggregate: cdot_cluster01_01_sas450

Storage Type

● NAS (Used for CIFS or NFS access)

○ Data Protection (Used as destination volume)

Size

Total Size: 1 GB

Snapshot Reserve (%): 5

Data Space: 972.8 MB

Snapshot Space: 51.2 MB

Thin Provisioning

☑ Thin Provisioned

Allocate space as it is needed by the volume. Recommended for increasing the capacity utilization of the volume when the volume is unlikely to use all of its allocated space. If unchecked, complete space is allocated immediately.

Tell me more about Thin Provisioning

Figure 23-20 Create volume

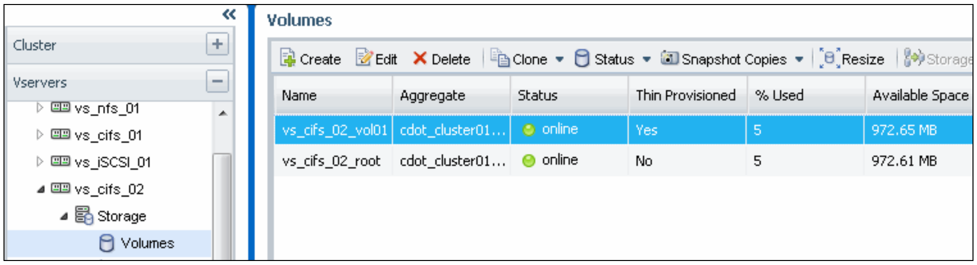
Here we explain some of the entries for this window:

- Name: Specify the name that you want.
- Aggregate: Specify the aggregate that provide the CIFS volume.
- Storage Type: In this case, NAS is for the CIFS service. If you create a snapmirror destination volume, you can select Data Protection type.
- Size: Specify the size and the percentage for the snapshot.

- Thin Provisioned: If you want to configure the thin provisioning volume, check this check box.

  With thin provisioning, when you create volumes and LUNs for different purposes in a given aggregate, you do not actually allocate any space for those volumes in advance. The space is allocated as data is written to the volumes.

4. After the wizard completes the volume creation, verify the new volume in the volume list. Figure 23-21 shows one of the one of examples.

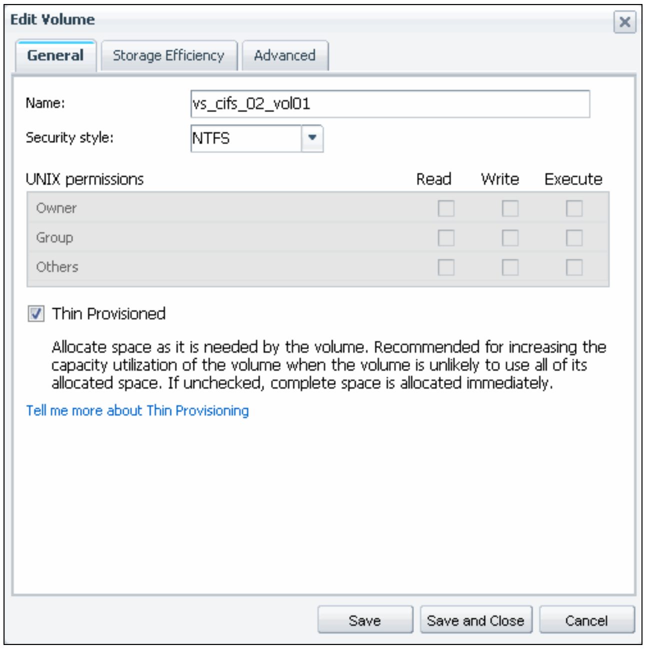


Name	Aggregate	Status	Thin Provisioned	% Used	Available Space
vs_cifs_02_vol01	cdot_cluster01…	online	Yes	5	972.65 MB
vs_cifs_02_root	cdot_cluster01…	online	No	5	972.61 MB

Figure 23-21   Check the CIFS volume

Editing and checking volume properties

After a volume creation completes, you can edit the CIFS volume properties to configure permissions or other properties. See Figure 23-22.



Name: vs_cifs_02_vol01

Security style: NTFS

UNIX permissions	Read	Write	Execute
Owner	□	□	□
Group	□	□	□
Others	□	□	□

☑ Thin Provisioned

Allocate space as it is needed by the volume. Recommended for increasing the capacity utilization of the volume when the volume is unlikely to use all of its allocated space. If unchecked, complete space is allocated immediately.

Tell me more about Thin Provisioning

Figure 23-22   Edit Volume

In this chapter, we are configuring the CIFS only SVM. Because of that, you can select the NTFS security style. Also, you can clear the **Thin Provisioned** check box here.

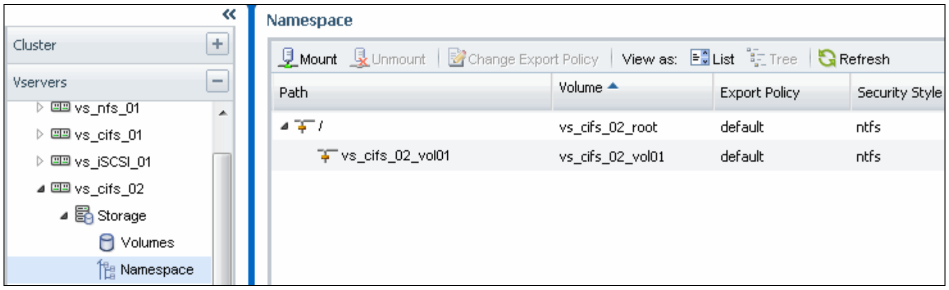
> **Note:** You can configure a UNIX security style for a CIFS SVM, but every Windows host can the CIFS volume as FAS file system only. For more information, see the Redbooks publications such as the *IBM System Storage N series Software Guide*, SG24-7129.

Click **Save and Close** to activate your selections.

Editing and checking the namespace for the CIFS volume

After the CIFS volume creation is completed, the namespace is created automatically. See Figure 23-23. The new volume has been mounted in the namespace.

Also, this figure indicates that the volume is accessed by clients as `/vs_cifs_02_vol01` and that the default export policy has been assigned to the volume.

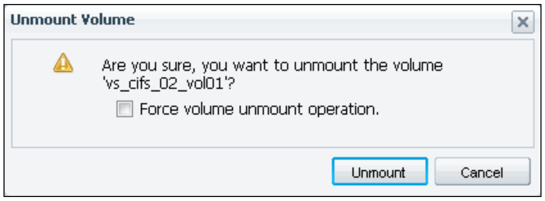


Path	Volume ▲	Export Policy	Security Style
/	vs_cifs_02_root	default	ntfs
vs_cifs_02_vol01	vs_cifs_02_vol01	default	ntfs

Figure 23-23 Namespace

You can reassign the namespace for the volume with a different junction name. In order to reassign the namespace, you need to unmount the volume and remount it:

1. To remount the volume, select the volume on the **Namespace** page and click **Unmount**, and leaving the “Force volume unmount operation” check box *unselected*, click **Unmount**. See Figure 23-24.

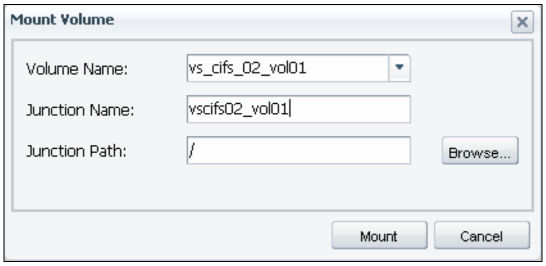


Are you sure, you want to unmount the volume 'vs_cifs_02_vol01'?

□ Force volume unmount operation.

Figure 23-24 Unmount volume

2. Then click the **Mount** button. See Figure 23-25.



Volume Name: vs_cifs_02_vol01

Junction Name: vscifs02_vol01

Junction Path: /

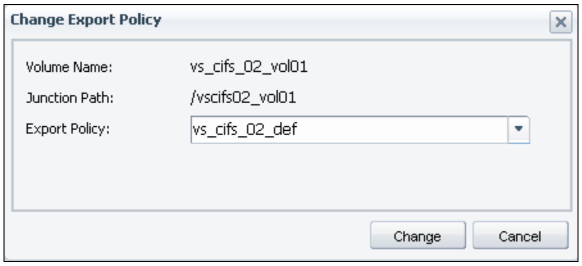
Figure 23-25 Mount volume

Here, the junction name is slightly different from the volume name. It is not necessary for the name to be the same. The volume name is used to reference the volume within the cluster. The junction name is used to reference the root of the volume in the namespace.

Changing the export policy

In this section, you will change the default export policy associated with the volume to redefine the client access to data:

1. Click the **Change Export Policy button**. See Figure 23-26. You can select the export policy that you want to change.
2. In this case, select the export policy that you already made.



Change Export Policy

Volume Name: vs_cifs_02_vol01

Junction Path: /vscifs02_vol01

Export Policy: vs_cifs_02_def

Figure 23-26 Change Export Policy

3. Verify that the export policy column in the **Namespace** window displays the export policy that you applied to the volume.

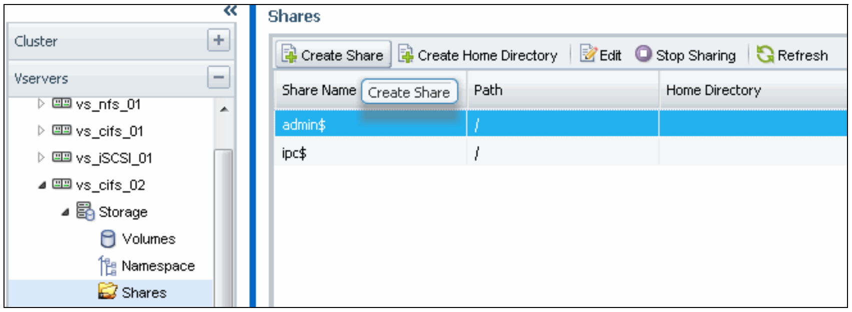
23.2.5 Creating CIFS shares

In this section, you will create two CIFS shares. One is a normal CIFS share and the other is a CIFS home directory.

Creating a normal CIFS share

To create a normal CIFS share for the new volume, proceed as follows:

1. Select the **Shares** and click the **Create Share** button. See Figure 23-27.

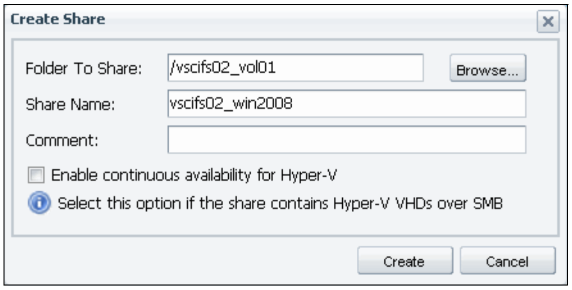


Share Name	Path	Home Directory
admin$	/	
ipc$	/	

Figure 23-27 Create CIFS Shares

2. When the Create Share window appears, click the **Browse** button to select the folder, qtree, or volume that should be shared. Specify a name for the new CIFS share.

3. You can select the **Enable continuous availability for Hyper-V** option to permit SMB 3.0 and later clients that support it to open files persistently during nondestructive operations. Files opened using this option are protected from disruptive events, such as failover, giveback, and LIF migration. This option is available only for clusters running Data ONTAP 8.2 or later. See Figure 23-28.



Folder To Share: /vscifs02_vol01

Share Name: vscifs02_win2008

Comment:

□ Enable continuous availability for Hyper-V

Select this option if the share contains Hyper-V VHDs over SMB

Figure 23-28 Create Share

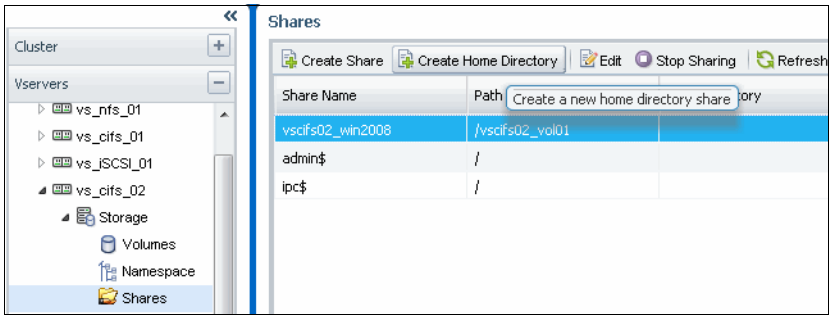
> **Note:** You can share a subdirectory within a volume or any auction path.

4. After reviewing a description for the share, click the **Create** button to create a CIFS share volume.

Creating a CIFS home directory

You can create a CIFS home directory with a method similar to a normal CIFS share:

1. To create a CIFS home directory, click the **Create Home Directory** button as in Figure 23-29.



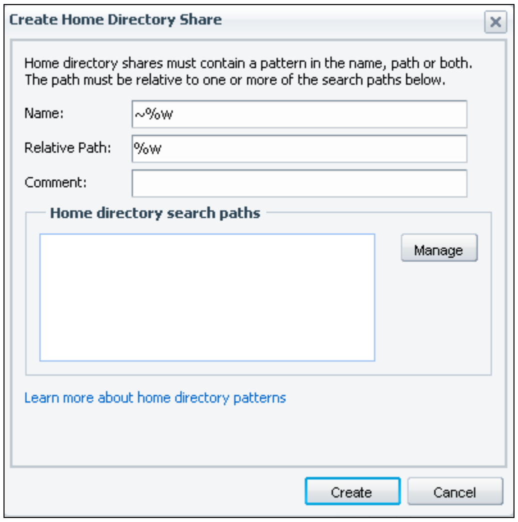
Share Name	Path	ory
vscifs02_win2008	/vscifs02_vol01	
admin$	/	
ipc$	/	

Figure 23-29 Create CIFS Home Directory

2. When the Create Home Directory window appears, enter the suitable parameters:

   – Name: ~%w
   – Relative Path: %w

   For an example, see Figure 23-30.


Create Home Directory Share

Home directory shares must contain a pattern in the name, path or both.
The path must be relative to one or more of the search paths below.

Name: ~%w

Relative Path: %w

Comment:

Home directory search paths
Manage
Learn more about home directory patterns
Create
Cancel

Figure 23-30 Create Home Directory Share

Data ONTAP CIFS home directories enable you to configure a share that maps to different directories based on the user that connects to it and a set of variables. Instead of having to create separate shares for each user, you can configure a single share with a few home directory parameters to define a user's relationship between an entry point (the share) and their home directory (a directory on the SVM).

There are four variables that determine how a user is mapped to a directory:

– Share name:

  This is the name of the share that you create that the user connects to. It can be static (for example, home), dynamic (for example, %w), or a combination of the two. You must set the home directory property for this share.

  The share name can use the following dynamic names:
  %w (the user's Windows user name)
  %d (the user's Windows domain name)
  %u (the user's mapped UNIX user name)

– Share path:

  This is the relative path, defined by the share and therefore associated with one of the share names, that is appended to each search path to generate the user's entire home directory path from the root of the SVM. It can be static (for example, home), dynamic (for example, %w), or a combination of the two (for example, eng/%w).

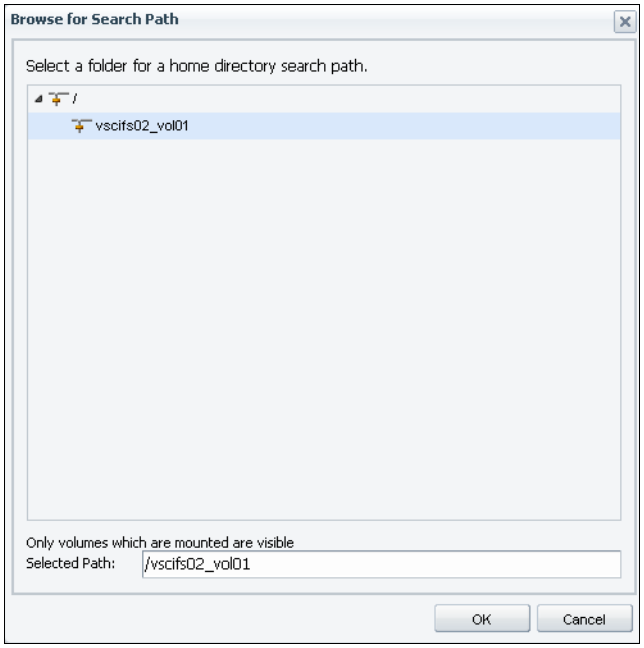
– Search paths:

  This is the set of absolute paths from the root of an SVM that you specify that directs the Data ONTAP search for home directories. You specify one or more search paths by using the SVM cifs home-directory search-path add command. If you specify multiple search paths, Data ONTAP tries them in the order specified until it finds a valid path.

– Directory:

  This is the user's home directory that you create for the user. It is usually the user's name. You must create it in one of the directories defined by the search paths.

3. Click the **Manage** button to add one or more search paths for a directory name match. You can select one of the junction paths (see Figure 23-31). Click the **Add** button.



Select a folder for a home directory search path.

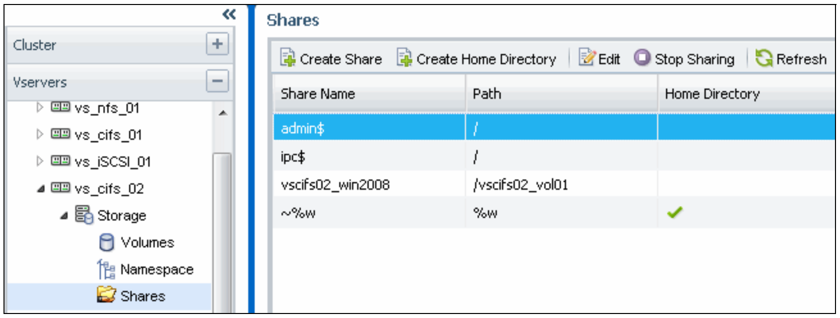
Only volumes which are mounted are visible

Selected Path: /vscifs02_vol01

Figure 23-31 Browse for Search Path

4. Then click the **Add** button and **Save & Close** button continuously. Review the parameters that were selected and click the **Create** button.

If you followed the foregoing procedure correctly, you can see the panel in Figure 23-32.



Share Name	Path	Home Directory
admin$	/	
ipc$	/	
vscifs02_win2008	/vscifs02_vol01	
~%w	%w	✓

Figure 23-32 Check two CIFS shares

23.2.6 Mapping CIFS shares in a Windows 2008 host

You can map CIFS shares through Windows Explorer or a command prompt. In this section, you will use a Windows command prompt to map CIFS shares.

Accessing a normal CIFS share

To access a normal CIFS share, proceed as follows:

1. From the Windows command line, enter the following command:

   `net view` *`[the name of CIFS SVM host]`*

   See Example 23-5.

Example 23-5 Checking CIFS shares from the CIFS SVM

```
C:\Users\administrator.NSERIES>net view \\vs_cifs_02
Shared resources at \\vs_cifs_02

(null)

Share name        Type  Used as  Comment

-------------------------------------------------------------------------------
vscifs02_win2008  Disk
The command completed successfully.
```

> **Note:** If you encounter Error 53 “The Network Path was not found,” attempt to identify the problem by performing one or more of the following actions:
>
> - Verify that the export policy allows CIFS access.
> - Verify that CIFS access is enabled for the SVM.
> - Review the data LIF setup to ensure that the LIF has the proper routing group and that you can ping the IP address form the Windows client.
> - Verify that can you ping the CIFS server by name.
> - If you cannot ping the CIFS server by name (the DNS is not set up to resolve the CIFS server), you can attempt to access the CIFS server with the IP address of a data LIF.

2. To map a network drive to the CIFS share, use the command in Example 23-6:

   `net use` *`[drive name] [\\cifs server\volume name]`*

Example 23-6 Map a network drive

```
C:\Users\administrator.NSERIES>net use * \\vs_cifs_02\vscifs02_win2008
Drive Z: is now connected to \\vs_cifs_02\vscifs02_win2008.

The command completed successfully.

C:\Users\administrator.NSERIES>dir z:
 Volume in drive Z has no label.
 Volume Serial Number is 8000-0418

 Directory of Z:\

21.10.2013  15:32    <DIR>          .
21.10.2013  15:32    <DIR>          ..
               0 File(s)              0 bytes
               2 Dir(s)   1.019.875.328 bytes free
```

Accessing a CIFS Home Directory

To access a CIFS Home Directory manually, follow these action items:

1. Create a directory called a log-in user name like administrator in your z: drive.
2. Map a network drive.

Example 23-7 shows how this can be done in our environment.

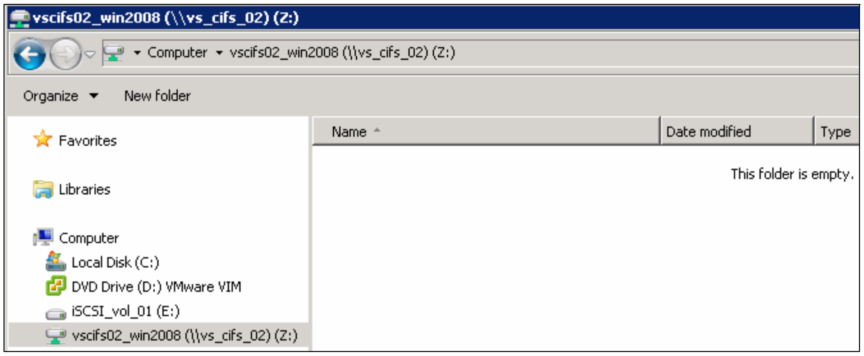

vscifs02_win2008 (\\vs_cifs_02) (Z:)
Computer ▸ vscifs02_win2008 (\\vs_cifs_02) (Z:)
Organize
New folder
Favorites
Libraries
Computer
Local Disk (C:)
DVD Drive (D:) VMware VIM
iSCSI_vol_01 (E:)
vscifs02_win2008 (\\vs_cifs_02) (Z:)
Name
Date modified
Type
This folder is empty.

Figure 23-33 Create a directory called a log-in user name

Example 23-7 Map and check a CIFS home directory

```
C:\Users\administrator.NSERIES>net use * \\vs_cifs_02\~administrator
Drive Y: is now connected to \\vs_cifs_02\~administrator.

The command completed successfully.

C:\Users\administrator.NSERIES>dir y:
 Volume in drive Y has no label.
 Volume Serial Number is 8000-0418

 Directory of Y:\

21.10.2013  17:30    <DIR>          .
21.10.2013  17:30    <DIR>          ..
21.10.2013  17:30                 0 adminfile.txt
               1 File(s)              0 bytes
               2 Dir(s)   1.019.863.040 bytes free
```

23.3 Monitoring SMB statistics

You can monitor the system with SMB statistics. In this section, we introduce several commands to monitor the CIFS performance.

23.3.1 Statistics with summary

Here we provide examples of clusterwide statistics and SVM-wide statistics.

Clusterwide statistics

You can check the clusterwide statistics with the following command:

`statistics show-periodic -node cluster:summary`

This starts the collection of clusterwide statistics data. See Example 23-8.

Example 23-8 Clusterwide statistics

```
cdot-cluster01::> statistics show-periodic -node cluster:summary
cluster:summary: cluster.cluster: 10/24/2013 16:21:43
  cpu    total                 data     data     data cluster  cluster  cluster
disk     disk
 busy      ops  nfs-ops cifs-ops busy     recv     sent    busy     recv     sent
read    write
 ---- -------- -------- -------- ---- -------- -------- ------- -------- --------
-------- --------
   2%        0        0        0   0%   2.59KB      57B      0%   97.4KB   98.4KB
3.86KB       0B
   2%        0        0        0   0%     253B       0B      0%   33.0KB   33.0KB
15.8KB   27.6KB
   5%        1        1        0   0%   1.13KB     553B      0%   76.5KB   76.1KB
553KB    926KB
  2%        0        0        0   0%     262B     89B      0%   17.9KB   18.3KB
0B       0B
   2%        0        0        0   0%     126B       0B      0%   12.9KB   12.8KB
786KB    962KB
```

SVM-wide statistics

You can check the performance collection data with the following set of commands.

- **`statistics start -object cifs:vserver -vserver`** *[the name of SVM]*
- **`statistics stop`**
- **`statistics show -sample-id`** *[sample id]*

This starts the collection of performance data. See Example 23-9.

Example 23-9 SVM-wide statistics

```
cdot-cluster01::> statistics start -object cifs:vserver -vserver vs_cifs_02
Statistics collection is being started for Sample-id: sample_4
Statistics collection is being started for Sample-id: sample_4

cdot-cluster01::> statistics stop
Statistics collection is being stopped for Sample-id: sample_4

cdot-cluster01::> statistics show -sample-id sample_4

Object: cifs:vserver
Instance: vs_cifs_02
Start-time: 10/24/2013 16:29:50
End-time: 10/24/2013 16:30:51
Cluster: cdot-cluster01

    Counter                                                     Value
    -------------------------------- --------------------------------
    active_searches                                                 0
    change_notifications_outstanding                                0
    cifs_latency                                                   -
    cifs_ops                                                        0
    cifs_read_ops                                                   0
    cifs_write_ops                                                  0
    commands_outstanding                                            0
    connected_shares                                                0
    connections                                                     0
    established_sessions                                            0
    instance_name                                          vs_cifs_02
    instance_uuid                                                  15
    open_files                                                      0
    signed_sessions                                                 0
14 entries were displayed.
```

23.3.2 Statistics at intervals

With the command shown in Example 23-10, Clustered Data ONTAP continuously displays performance statistics at regular intervals:

statistics show-periodic -node *[node name]* **-object cifs -instance** *[SVM name]* **-interval** *[second(s)]*

Example 23-10 Statistics at Intervals

```
cdot-cluster01::> statistics show-periodic -node cdot-cluster01-01 -object cifs
-instance vs_cifs_02 -interval 1
cdot-cluster01: cifs.vs_cifs_02: 10/24/2013 17:13:41
          auth_reject        change                      cifs     cifs
   active         too notifications     cifs             read    write
commands connected            established instance     node     node     open
process   signed  vserver  vserver
 searches        many   outstanding  latency cifs_ops      ops      ops
outstanding   shares connections    sessions     name     name     uuid    files
name sessions       id     name
 -------- ----------- ------------- -------- -------- -------- --------
----------- --------- ----------- ----------- -------- -------- -------- --------
-------- -------- -------- --------
        0           0             0     0us        0        0        0
0         0           0           0 vs_cifs_02 cdot-cluster01-01
c41e66ef-3193-11e3-bc39-d5fef0172dde 0 - 0 15 vs_cifs_02
        0           0             0     0us        0        0        0
0         0           0           0 vs_cifs_02 cdot-cluster01-01
c41e66ef-3193-11e3-bc39-d5fef0172dde 0 - 0 15 vs_cifs_02
        0           0             0     0us        0        0        0
0         0           0           0 vs_cifs_02 cdot-cluster01-01
c41e66ef-3193-11e3-bc39-d5fef0172dde 0 - 0 15 vs_cifs_02
        0           0             0     0us        0        0        0
0         0           0           0 vs_cifs_02 cdot-cluster01-01
c41e66ef-3193-11e3-bc39-d5fef0172dde 0 - 0 15 vs_cifs_02
        0           0             0     0us        0        0        0
0         0           0           0 vs_cifs_02 cdot-cluster01-01
c41e66ef-3193-11e3-bc39-d5fef0172dde 0 - 0 15 vs_cifs_02
```

23.4 Further information

More details on CIFS implementation can be found in the following IBM support documents:

- *OnCommand System Manager 3.0 Help for Use with Clustered Data ONTAP Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

- *Clustered Data ONTAP 8.2 File Access and Protocols Management Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

24

NFS storage

This chapter shows a realistic example for an NFS storage that consists of a Clustered Data ONTAP system and a Ubuntu Linux server having several Ethernet ports.

We show you how to configure a storage virtual machine (SVM) for a Linux Host. Also, you will learn how to configure several features such as creating an SVM, creating an export policy, and so on. We explain some changes such as the **`showmount`** command in a Clustered ONTAP environment.

In the Data ONTAP 8.2 operating system, NFS is supported in clusters of up to 24 nodes.

The following topics are covered:

- Planning and checking the NFS environment
- Configuring NFS for both Ubuntu Linux and Clustered Data ONTAP
- Further information

24.1 Planning and checking the NFS environment

Through this chapter, you will create an SVM for Ubuntu Linux through an OnCommand System Manager GUI or a Shell command. According to the official procedure, you must check the interoperability and configure both a Ubuntu Linux host and a Clustered Data ONTAP system before your planning or implementation.

A cluster is a scalable system that allows you to group pairs of controllers, share physical resources, and distribute workloads across the systems while consolidating management for administrators into one unified interface. The Clustered Data ONTAP 8.2 storage system has the capability to join up to 24 individual nodes or 12 pairs of nodes.

24.1.1 Checking your environment before configuration of NFS

Before configuring NFS, you will need to perform several checks as described next.

Checking the interoperability

Here is the link to the N series interoperability matrix:

http://www.ibm.com/support/docview.wss?uid=ssg1S7003897

In this case, you will check the interoperability between any Linux host and the Clustered Data ONTAP. You can check the NFS interoperability matrix with the following link which is included the above link.

http://www-01.ibm.com/support/docview.wss?uid=ssg1S7003770

At the time of writing this chapter, the latest interoperability matrix was made on 01 October 2013, as follows. See Figure 24-1.

The support information is organized into the following topics in this interoperability matrix:

- Support Policies
- NFS for Client OS
- NFS for Virtual Servers: ESX, ESXi, RHEL KVM and XenServer
- Alerts and Info
- Historical DOT Client OS
- Historical Kerberos
- Disclaimer

N series Interoperability Matrix - NFS

nseries_interop_nfs_20131001.xls

Figure 24-1 NFS Interoperability matrix

We can review the certified version and some check lists from the Excel file.

Here are some examples for NFS:

- Certified UNIX versions for IBM AIX, Red Hat Linux, and so on
- Certified KDC types and servers
- Certified NFS versions
- Supported NFS features such as ACLs, delegations, and so on
- Information and alerts

See Information and alerts, which include important information regarding OS patches, considerations, and so on.

Note: It is crucial to check the interoperability before the implementation and the planning phase. If you cannot find any certified solution, you must submit a SCORE/RPQ (Request for Price Quotations). To submit a SCORE/RPQ, contact your IBM Representative.

24.2 Configuring NFS for both Ubuntu Linux and Clustered Data ONTAP

In a Clustered Data ONTAP environment, you must create an SVM to provide NFS services to the host. In this chapter, we assume that you already know how to configure the basic cluster system.

24.2.1 Creating an SVM for NFS

Before creating an SVM on Clustered Data ONTAP, you must check the interoperability, because you have already checked the interoperability matrix. Normally there is some important information in the interoperability matrix Excel file. See Example 24-1.

Example 24-1 NFS Interoperability matrix

```
Name: 20121220-230142918
Features: ACLs; Exports; kerberos; Locking; Referrals; Sessions and Callbacks;
pNFS
Client OS: RHEL Server 6.2 32-bit; RHEL Server 6.2 64-bit; RHEL Server 6.3 32-bit;
RHEL Server 6.3 64-bit
```

At the time of writing this chapter, pNFS with clustered ONTAP is certified with only one Linux OS. Red Hat Linux 6.2 or above is only supported officially. Before configuring or planning to use Clustered Data ONTAP, make sure you are trying to configure the system as a certified solution. This is really crucial to prevent any future issues.

Normally the procedure for creating an SVM for NFS is quite similar to an SVM for CIFS. So we may skip a procedure if it is the same as a procedure of CIFS. But we let you know where you can find the procedure in other chapters.

Creating an SVM of NFS

In these steps, you will create an SVM of NFS for UNIX hosts. As you know, we can create an SVM to support both CIFS and NFS protocols at the same time. But you will create an SVM of NFS only and review some special options for NFS.

To create an SVM of NFS, proceed as follows:

1. Open OnCommand System Manager and find Vservers as in Example 24-2.
2. Click the **Create** button.

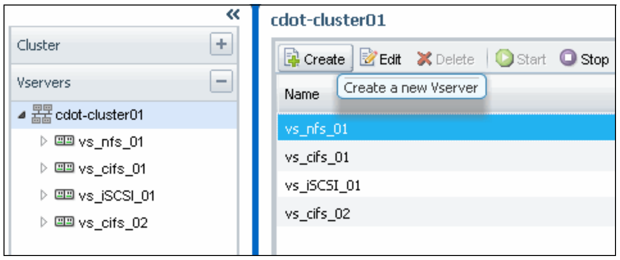


Figure 24-2 Create an SVM of NFS

3. You will see the Vserver Setup wizard as in Figure 24-3. Enter an SVM name and several options in this window.

   Here we list some values to be entered:

   – Vserver Name: Define a name of SVM that you want to name.

   – Data Protocols: Each SVM can be a server that provides multiple protocols or at least one protocol. Here, we select NFS only.

   – Language: This parameter depends on the specific characters that are used by your language environment. The default language is set to C.UTF-8. If your country uses two-byte characters, select a correct one.

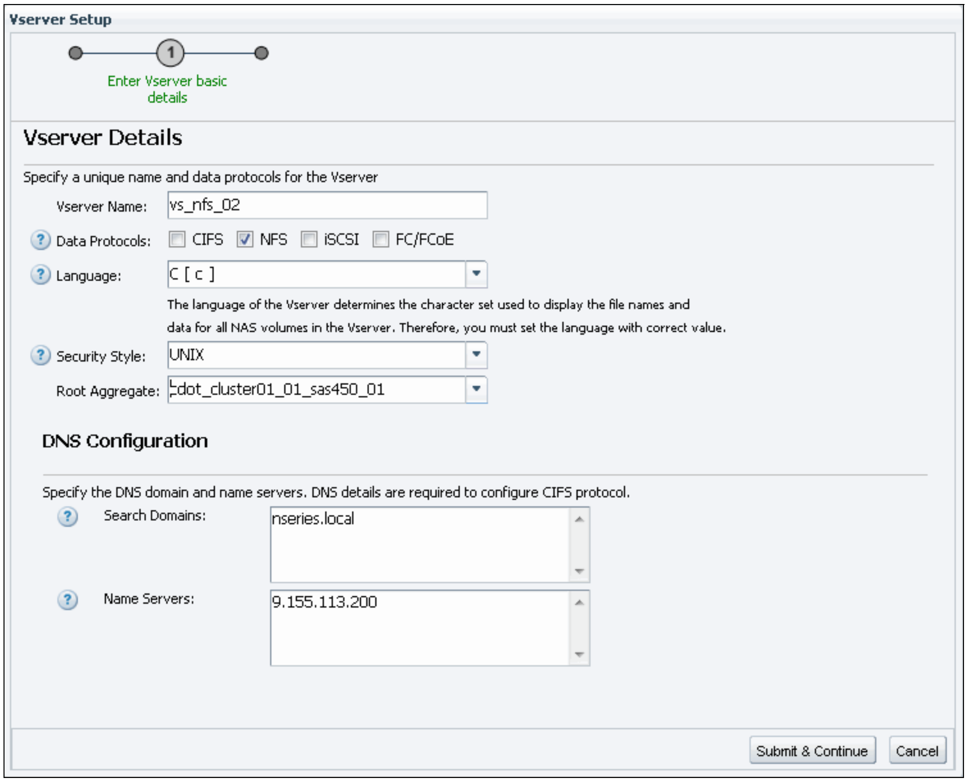
   > **Note:** The language of an SVM with FlexVol volumes can be modified after the SVM is created.

   – Security style: Choose from one of the following options: UNIX, NTFS, and Mixed.

   You can decide what security style to use on a volume, and you should consider two factors. The primary factor is the type of administrator that manages the file system. The secondary factor is the type of user or service that accesses the data on the volume.

   – Root Aggregate: Select a root aggregate from one of the node root aggregates.

4. Then click the **Submit & Continue** button.



Vserver Details

Specify a unique name and data protocols for the Vserver

Vserver Name: vs_nfs_02

Data Protocols: □ CIFS ☑ NFS □ iSCSI □ FC/FCoE

Language: C [c]

The language of the Vserver determines the character set used to display the file names and data for all NAS volumes in the Vserver. Therefore, you must set the language with correct value.

Security Style: UNIX

Root Aggregate: cdot_cluster01_01_sas450_01

DNS Configuration

Specify the DNS domain and name servers. DNS details are required to configure CIFS protocol.

Search Domains: nseries.local

Name Servers: 9.155.113.200

Figure 24-3 Vserver Details

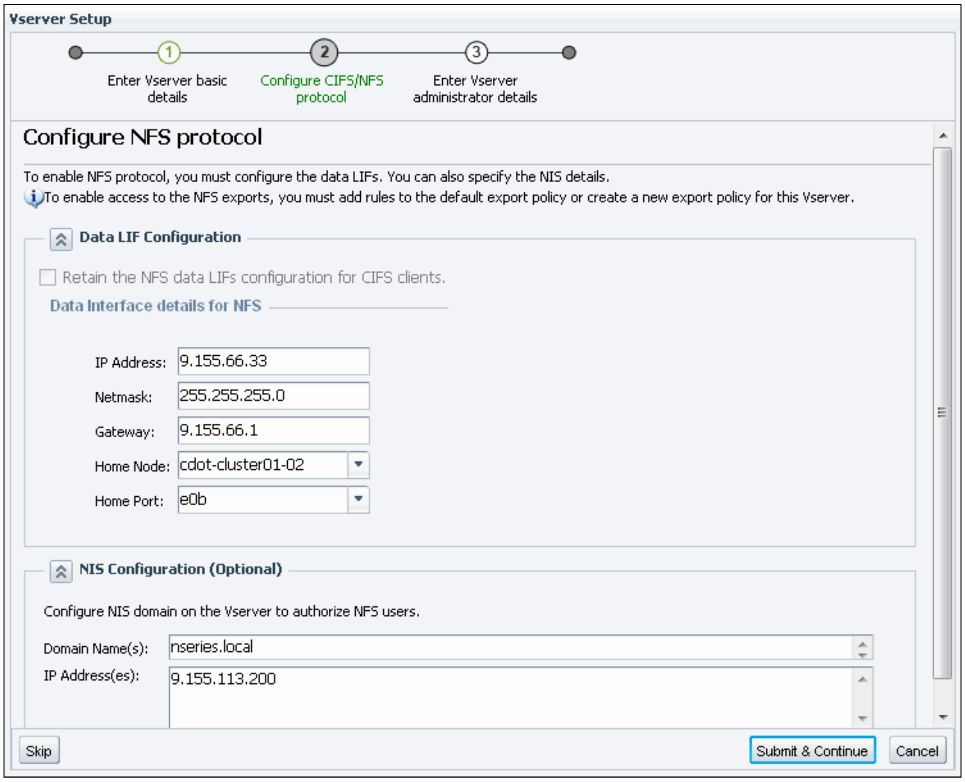
> **Note:** For the Data Protocols selection check boxes, you can select at least one data protocol. If the customer has several Windows and UNIX hosts and needs special volumes that are shared by both protocols at the same time, you can select both CIFS and NFS in a shot and configure the SVM. Also, the SVM can share the network paths for both protocols.

Example 24-2 shows the shell command to create this SVM.

Example 24-2 Create NFS SVM

```
cdot-cluster01::>vserver create -vserver vs_nfs_02 -rootvolume vs_nfs_02_root
-aggregate cdot_cluster01_01_sas450_01 -rootvolume-security-style unix -ns-switch
file -nm-switch file
cdot-cluster01::>vserver nfs create -vserver vs_nfs_02
```

5. Figure 24-4 will be displayed to configure data logical interfaces (LIFs) for NFS hosts.


Vserver Setup
1 Enter Vserver basic details
2 Configure CIFS/NFS protocol
3 Enter Vserver administrator details

Configure NFS protocol

To enable NFS protocol, you must configure the data LIFs. You can also specify the NIS details.

To enable access to the NFS exports, you must add rules to the default export policy or create a new export policy for this Vserver.

Data LIF Configuration

□ Retain the NFS data LIFs configuration for CIFS clients.

Data Interface details for NFS

IP Address: 9.155.66.33
Netmask: 255.255.255.0
Gateway: 9.155.66.1
Home Node: cdot-cluster01-02
Home Port: e0b

NIS Configuration (Optional)

Configure NIS domain on the Vserver to authorize NFS users.

Domain Name(s): nseries.local
IP Address(es): 9.155.113.200

Figure 24-4 Configure NFS protocol

In these steps. you can configure a NFS protocol on the SVM to provide file-level data access for NAS clients.To enable NFS protocol, you must create the data LIFs and the NFS server.

Here we explain some of the values in the boxes:

– Retain the CIFS data LIFs configuration for NFS clients check box:

  You can select this check box if you are the SVM for both NFS and CIFS protocols and the SVM will share the network paths for both protocols at the same time.

– Data Interface details for NFS

  - IP Address: Specify the IP address for the NFS service.
  - Netmask: Specify the subnet mask.
  - Gateway: Specify the default gateway IP address.
  - Home Node: Specify the home node for the NFS services. The home node is the node to which the logical interface returns when the LIF is reverted to its home port.
  - Home Port: Specify the home port for the NFS services. The home port is the port to which the logical interface returns when the LIF is reverted to its home port.

– NIS Configurations (Optional):

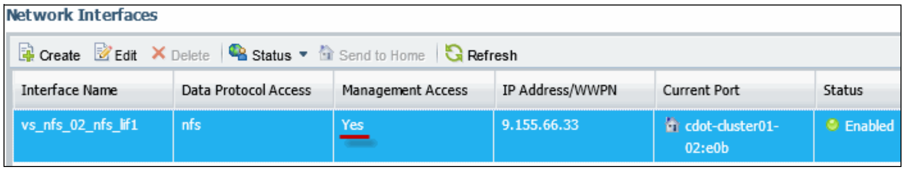
  A Network Information Service (NIS) domain provides a directory of hostnames and IP addresses in a network. An SVM administrator can manage NIS domains by creating, modifying, deleting, or displaying information about them. NIS cannot be configured for the cluster management server. You can configure multiple NIS domains for a given SVM, but only one NIS domain can be active on an SVM at any given time. You can also configure an NIS domain with more than one SVM.

6. Example 24-3 shows the shell commands for OnCommand System Manager.

Example 24-3 Create the LIF for the SVM

```
cdot-cluster01::>network interface create -vserver vs_nfs_02 -lif
vs_nfs_02_nfs_lif1 -role data -data-protocol nfs -home-node cdot-cluster01-02
-home-port e0b -address 9.155.66.33 -netmask 255.255.255.0
cdot-cluster01::>network routing-groups route create -server vs_nfs_02
-routing-group d9.155.16.0/24 -gateway 9.155.66.1
```

> **Note:** If you create an LIF with this wizard, the management access is allowed automatically. See Figure 24-5. If you do not want to permit the management access through this IP address, you can modify the properties with the following command.
>
> **network interface modify -vserver** *<SVM name>* **-lif** *<LIF name>* **-firewall-policy {***data | mgmt***}**
>
> An example is provided in Example 24-4.



Interface Name	Data Protocol Access	Management Access	IP Address/WWPN	Current Port	Status
vs_nfs_02_nfs_lif1	nfs	Yes	9.155.66.33	cdot-cluster01-02:e0b	Enabled

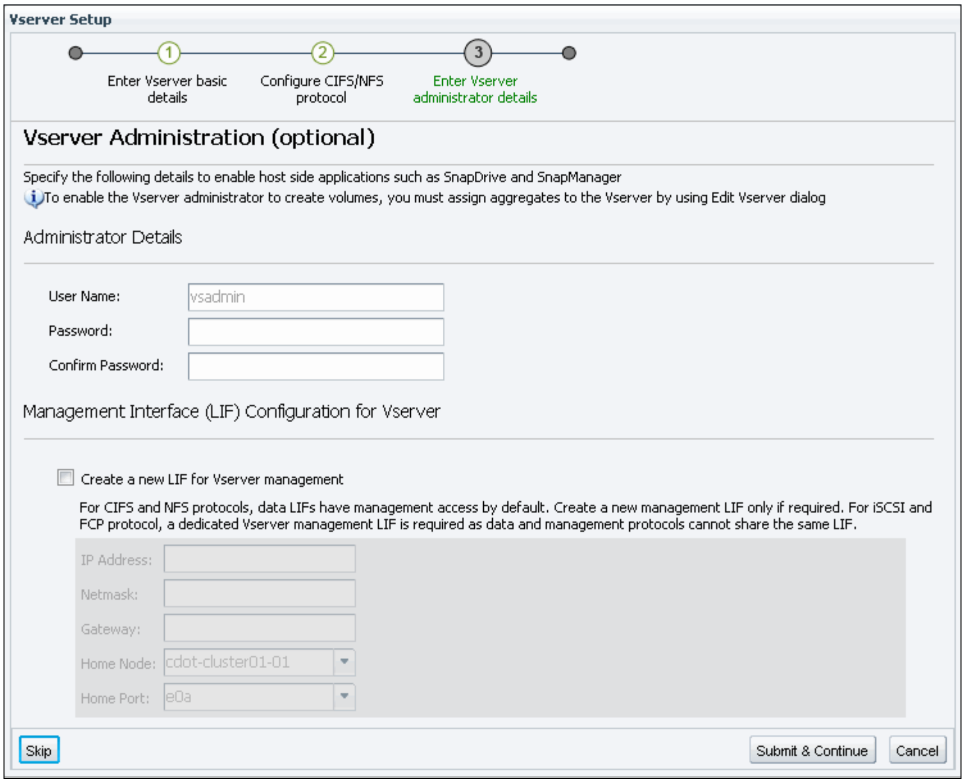
Figure 24-5 Management access granted automatically

Example 24-4 Check and Modify the permission of the management access

```
cdot-cluster01::> net int show -vserver vs_nfs_02 -lif vs_nfs_02_nfs_lif1
  (network interface show)

                    Vserver Name: vs_nfs_02
          Logical Interface Name: vs_nfs_02_nfs_lif1
                            Role: data
                   Data Protocol: nfs
                       Home Node: cdot-cluster01-02
                       Home Port: e0b
                    Current Node: cdot-cluster01-02
                    Current Port: e0b
              Operational Status: up
                 Extended Status: -
                         Is Home: true
                 Network Address: 9.155.66.33
                         Netmask: 255.255.255.0
             Bits in the Netmask: 24
                 IPv4 Link Local: -
              Routing Group Name: d9.155.66.0/24
           Administrative Status: up
                 Failover Policy: nextavail
                 Firewall Policy: mgmt
                     Auto Revert: false
   Fully Qualified DNS Zone Name: none
         DNS Query Listen Enable: false
             Failover Group Name: system-defined
                        FCP WWPN: -
```

```
                  Address family: ipv4
                         Comment: -
cdot-cluster01::> net int modify -vserver vs_nfs_02 -lif vs_nfs_02_nfs_lif1
-firewall-policy data
  (network interface modify)
```

7. After completing this step, click **`Submit & Close`**.
8. The next step is for configuring any administration information for each SVM. See Figure 24-6.



Vserver Setup
1 Enter Vserver basic details
2 Configure CIFS/NFS protocol
3 Enter Vserver administrator details

Vserver Administration (optional)

Specify the following details to enable host side applications such as SnapDrive and SnapManager

To enable the Vserver administrator to create volumes, you must assign aggregates to the Vserver by using Edit Vserver dialog

Administrator Details

User Name: vsadmin
Password:
Confirm Password:

Management Interface (LIF) Configuration for Vserver

□ Create a new LIF for Vserver management

For CIFS and NFS protocols, data LIFs have management access by default. Create a new management LIF only if required. For iSCSI and FCP protocol, a dedicated Vserver management LIF is required as data and management protocols cannot share the same LIF.

IP Address:
Netmask:
Gateway:
Home Node: cdot-cluster01-01
Home Port: e0a

Figure 24-6 Vserver Administration (Optional)

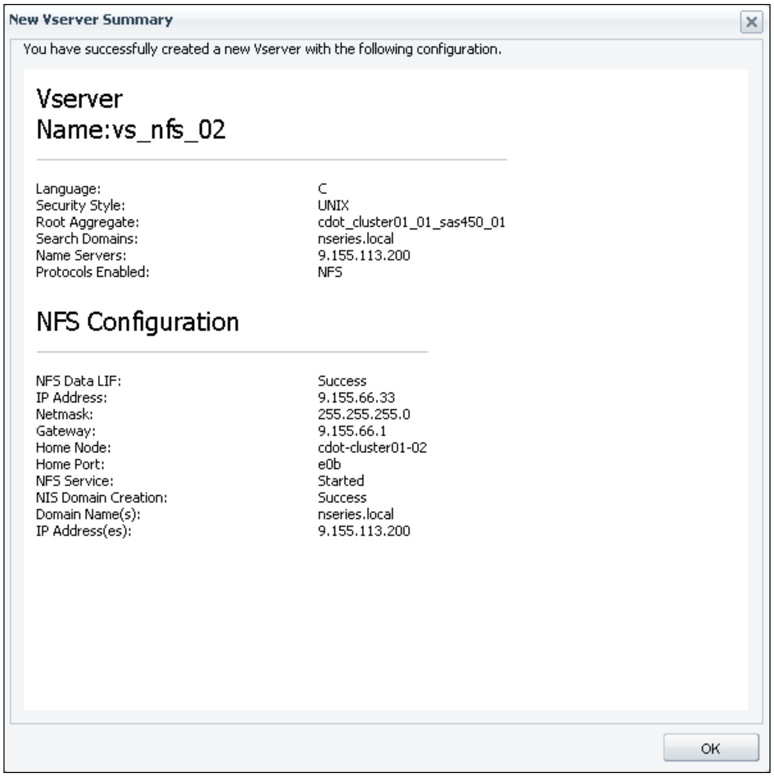
A Vserver administrator can administer an SVM and its resources, such as volumes, protocols, and services, depending on the capabilities assigned by the cluster administrator. A Vserver administrator cannot create, modify, or delete an SVM.

> **Note:** SVM administrators cannot log in to System Manager.

For more information about SVM administrator capabilities, see the *Clustered Data ONTAP System Administration Guide for SVM Administrators*.

9. After completing the NFS SVM creation step, you can see the following result panel (Figure 24-7).

> **Note:** This summary panel shows the result only. You cannot cancel the NFS SVM procedure itself. If you want to remove and recreate the SVM, you should remove the SVM that you made and recreate it manually.



You have successfully created a new Vserver with the following configuration.

Vserver Name:vs_nfs_02

Language:	C
Security Style:	UNIX
Root Aggregate:	cdot_cluster01_01_sas450_01
Search Domains:	nseries.local
Name Servers:	9.155.113.200
Protocols Enabled:	NFS

NFS Configuration

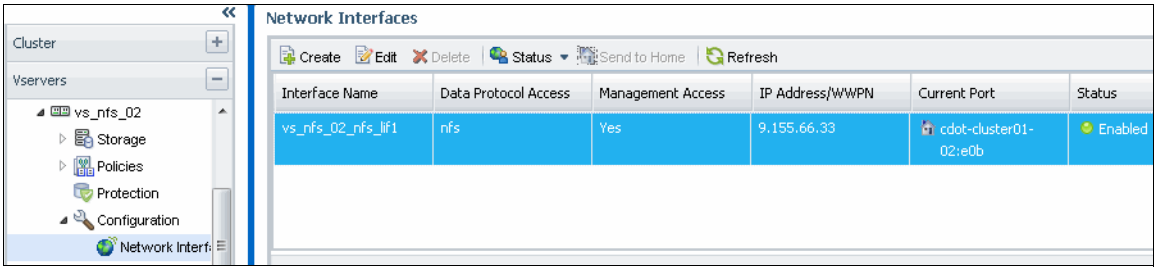
NFS Data LIF:	Success
IP Address:	9.155.66.33
Netmask:	255.255.255.0
Gateway:	9.155.66.1
Home Node:	cdot-cluster01-02
Home Port:	e0b
NFS Service:	Started
NIS Domain Creation:	Success
Domain Name(s):	nseries.local
IP Address(es):	9.155.113.200

Figure 24-7 Vserver Summary

Checking the result after the SVM creation wizard completion

You have just created the SVM of NFS. You can check the result via OnCommand System Manager or cluster shell commands. Proceed as follows:

1. You can check what were configured automatically and can be edited by you manually. See Figure 24-8. Through this panel, you can see the network state that you made through an SVM creation wizard.



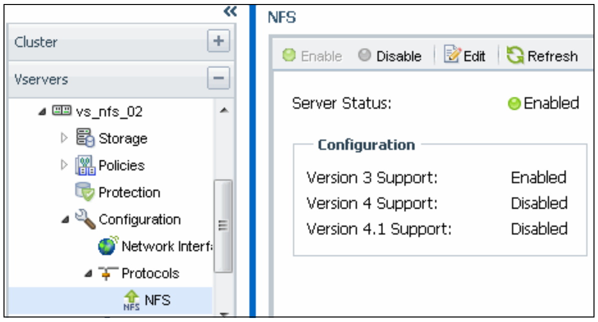
Interface Name	Data Protocol Access	Management Access	IP Address/WWPN	Current Port	Status
vs_nfs_02_nfs_lif1	nfs	Yes	9.155.66.33	cdot-cluster01-02:e0b	Enabled

Figure 24-8 Check Network Interfaces for NFS

As you can see in Figure 24-8, the NFS SVM creation wizard created the LIF that can be used for two purposes. One is data access and the other is management access. That means the customer can use NFS services and access the SVM through the previous IP address.

2. If you want to add more data LIFs for specific purposes, see 23.2.2, “Creating a NAS data LIF” on page 393.

3. After checking the network state, you can check which NFS protocol versions were configured by default. See Figure 24-9. Click **Vservers** → **Configuration** → **NFS**.



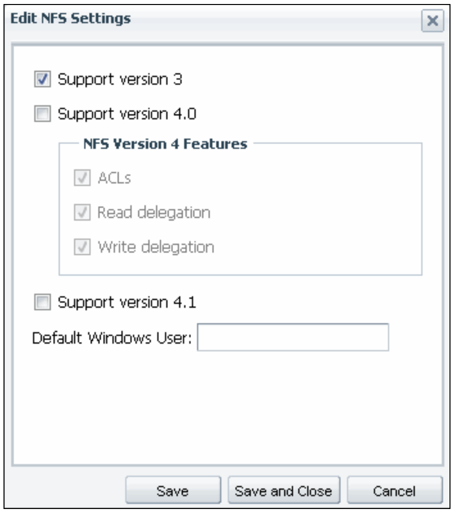
Server Status: Enabled

Configuration

Version 3 Support: Enabled
Version 4 Support: Disabled
Version 4.1 Support: Disabled

Figure 24-9 NFS configuration status

4. As you can see in Figure 24-9, the default NFS protocol is v3. If you want to use other NFS protocols and features, you can edit the NFS settings as follows. Click the **Edit** button. See Figure 24-10.



☑ Support version 3

☐ Support version 4.0

NFS Version 4 Features

☑ ACLs

☑ Read delegation

☑ Write delegation

☐ Support version 4.1

Default Windows User:

Figure 24-10 Edit NFS Settings

5. Before changing any settings for NFS, review to make sure that the protocols and features are supported officially. You can check it with the interoperability matrix as I mentioned earlier in this chapter. See "Checking the interoperability" on page 412.

   Here we explain the NFS settings:

   – Support Version 4.0:

     - NFS Version 4 features:

       ACLs: The NFSv4 protocol can provide access control in the form of NFSv4 Access Control Lists (ACLs), which are similar in concept to those found in CIFS. An NFSv4 ACL consists of individual Access Control Entries (ACEs), each of which provides an access control directive to the server. Clustered Data ONTAP 8.2 supports a maximum of 1,024 ACEs.

       Read delegation / Write delegation: Delegations can be used to improve the read and write performance of certain applications.

– Support Version 4.1:

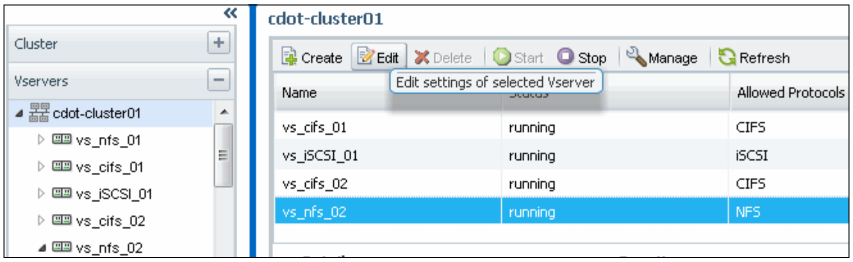
NFSv4.1 is considered a minor version of NFSv4 and also includes the feature of pNFS.

You can change any options for NFS in this panel if you have a supported environment.

24.2.2 Changing the default SVM properties

After creating an SVM with the wizard, we advise you to change some parameters as shown in the following examples. To prevent any future issues related to some volume assignments to root aggregates, it is best to change the default parameters. Proceed as follows:

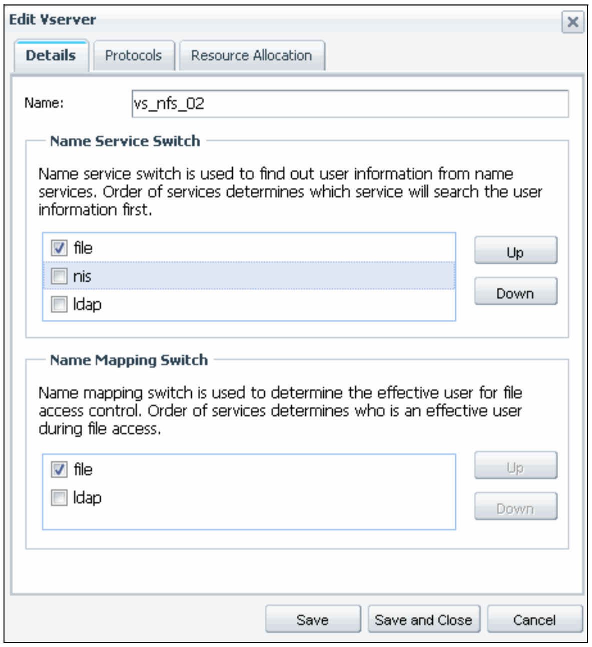
1. Open the properties of the SVM that you made previously. See Figure 24-11. Click the **Edit** button to open the properties window.



Name	Status	Allowed Protocols
vs_cifs_01	running	CIFS
vs_iSCSI_01	running	iSCSI
vs_cifs_02	running	CIFS
vs_nfs_02	running	NFS

Figure 24-11 Edit SVM Properties

2. After the SVM properties window appears, you can see the following pop-up window (Figure 24-12). If your SVM is using either NIS or LDAP, select the correct services. If not, clear all of the check boxes except for **file**.



Name: vs_nfs_02

Name Service Switch

Name service switch is used to find out user information from name services. Order of services determines which service will search the user information first.

☑ file
☐ nis
☐ ldap

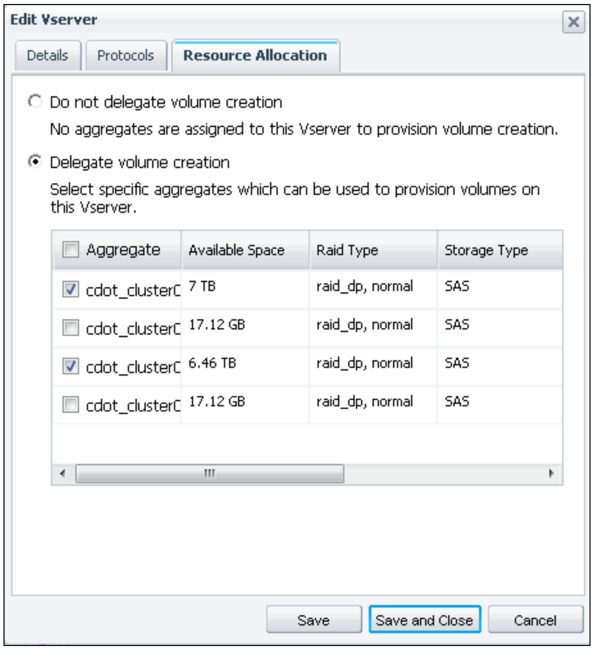
Name Mapping Switch

Name mapping switch is used to determine the effective user for file access control. Order of services determines who is an effective user during file access.

☑ file
☐ ldap

Figure 24-12 Vserver Details

3. Then click the **Resource Allocation** tab to change the default aggregates for the NFS volumes that will be used for this SVM. See Figure 24-13.



○ Do not delegate volume creation

No aggregates are assigned to this Vserver to provision volume creation.

● Delegate volume creation

Select specific aggregates which can be used to provision volumes on this Vserver.

□ Aggregate	Available Space	Raid Type	Storage Type
☑ cdot_clusterC	7 TB	raid_dp, normal	SAS
□ cdot_clusterC	17.12 GB	raid_dp, normal	SAS
☑ cdot_clusterC	6.46 TB	raid_dp, normal	SAS
□ cdot_clusterC	17.12 GB	raid_dp, normal	SAS

Figure 24-13 Resource Allocation

4. If you select one of the root aggregates, you can see the warning from the OnCommand System Manager, which is shown in Figure 23-9.

   The warning message suggests that you never store data volumes on a node's aggr0 aggregate.

   Note: If you do not set the specific aggregates in this window, whenever you will try to create an NFS volume in the SVM, you can see all of the aggregates including aggr0. We advise you to change the default value not to use root aggregates.

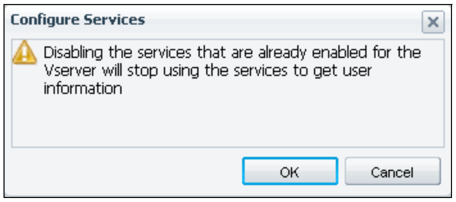
5. Here is the command to change the resource allocation properties as in this window. See Example 24-5.

```
vserver modify -vserver <SVM name> -aggr-list <aggregate name1, name2, ...>
```

Example 24-5 Modify the resource allocation

```
cdot-cluster01::> vserver modify -vserver vs_nfs_02 -aggr-list
cdot_cluster01_01_sas450_01,cdot_cluster01_02_sas450_01
```

6. After changing the default value, click the **Save and Close** button, to activate your changes. When you try to apply your changes, you will see the panel shown in Figure 24-14.



Disabling the services that are already enabled for the Vserver will stop using the services to get user information

Figure 24-14 Configure Services

24.2.3 Creating an export policy

In this section, you will create an export policy for a Ubuntu Linux Host using NFS v3 protocol. After creating the export policy you will apply the port policy and then assign the policy to a NFS volume.

When a volume is created, it automatically inherits the default export policy of the root volume of the SVM. You can change the default export policy associated with the volume to redefine the client access to data. Or you can create an additional export policy as well.

To create an export policy, proceed as follows:

1. Click the **Create Policy** button to create a new export policy. See Figure 24-15.

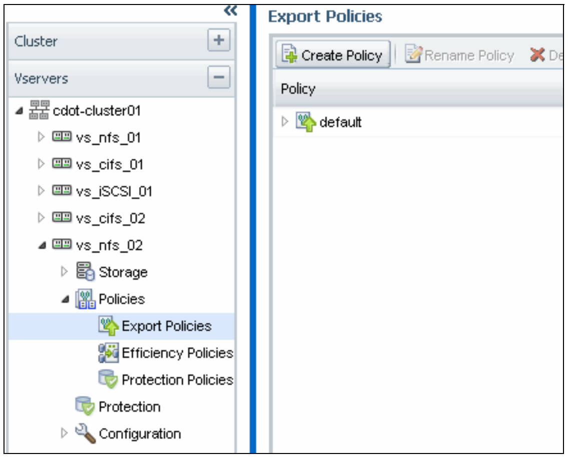
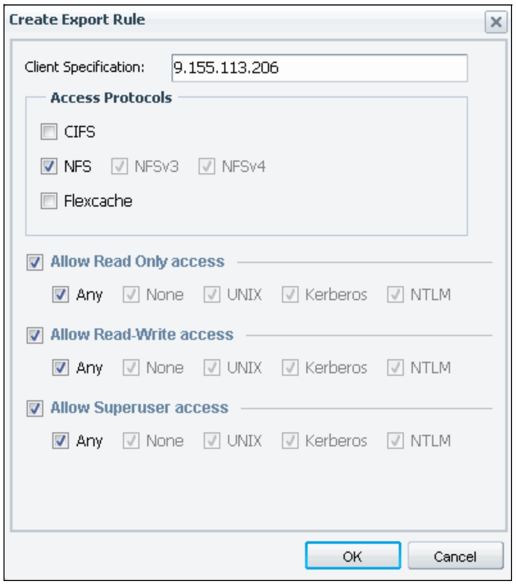


Figure 24-15 Create an Export Policy

2. In this step, you will input some values as follows. When the create export policy window appears, click the **Add** button. See Figure 24-16.



Client Specification: 9.155.113.206

Access Protocols

☐ CIFS

☑ NFS ☑ NFSv3 ☑ NFSv4

☐ Flexcache

☑ Allow Read Only access

☑ Any ☑ None ☑ UNIX ☑ Kerberos ☑ NTLM

☑ Allow Read-Write access

☑ Any ☑ None ☑ UNIX ☑ Kerberos ☑ NTLM

☑ Allow Superuser access

☑ Any ☑ None ☑ UNIX ☑ Kerberos ☑ NTLM

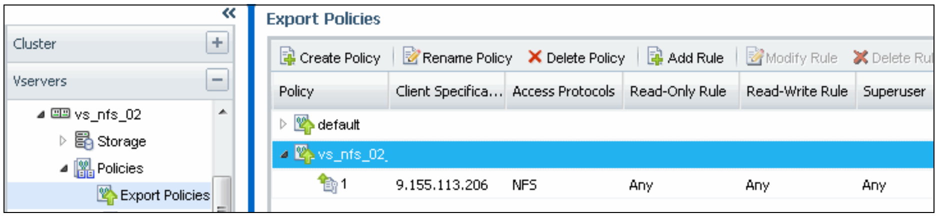
Figure 24-16 Create Export Rule

You can specify the match in any of the following formats:

- As a host name; for example, host1As an IPv4 address; for example, 10.1.12.24
- As an IPv4 address with a network mask; for example, 10.1.16.0/255.255.255.0.
- As a netgroup, with the netgroup name preceded by the @ character; for example, @netgroup
- As a domain name preceded by the "." symbol; for example, .example.com

> **Note:** In this example, we specify the Linux host IP only. That means this export policy is only for host having the specific IP address.

3. After you complete the creation of the export policy, you can see the panel in Figure 24-17.



Policy	Client Specifica...	Access Protocols	Read-Only Rule	Read-Write Rule	Superuser
default					
vs_nfs_02					
1	9.155.113.206	NFS	Any	Any	Any

Figure 24-17 Check the export policy

Here is the set of commands to create an export policy like the above OnCommand System Manager window. See Example 24-6.

```
vserver export-policy create -vserver <SVM name> <export-policy name>
vserver export-policy rule create -vserver <SVM name> -policyname
<export-policy name> -clientmatch <host> -rorule {any|none|...} -rwrule
{any|none|...} -superuser {any|none|...} -protocol <protocol>
```

Example 24-6 Creating a export policy

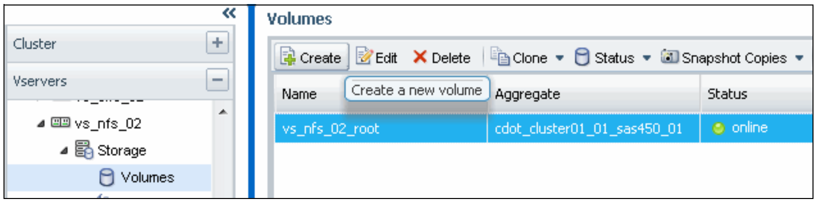
```
cdot-cluster01::> vserver export-policy rule show
             Policy          Rule    Access   Client                RO
Vserver      Name            Index   Protocol Match                 Rule
------------ --------------- ------  -------- --------------------- ---------
vs_cifs_02   vs_cifs_02_def  1       cifs     0.0.0.0/0             any
vs_nfs_01    default         1       any      0.0.0.0/0             any
vs_nfs_01    vmware_cluster_01
                             1       nfs      9.155.0.0/16          any
vs_nfs_02    vs_nfs_02_def   1       nfs      9.155.113.206         any
4 entries were displayed.

cdot-cluster01::> vserver export-policy create -vserver vs_nfs_02 vs_nfs_02_def1

cdot-cluster01::> vserver export-policy show
Vserver          Policy Name
---------------  -------------------
vs_cifs_01       default
vs_cifs_02       default
vs_cifs_02       vs_cifs_02_def
vs_iSCSI_01      default
vs_nfs_01        default
vs_nfs_01        vmware_cluster_01
```

```
vs_nfs_02         default
vs_nfs_02         vs_nfs_02_def
vs_nfs_02         vs_nfs_02_def1
9 entries were displayed.

cdot-cluster01::> vserver export-policy rule create -vserver vs_nfs_02 -policyname
vs_nfs_02_def1 -clientmatch 9.155.113.206 -rorule any -rwrule any -superuser any
-protocol nfs
```

24.2.4 Creating an NFS volume

In these steps, you will create a volume for NFS services. Proceed as follows:

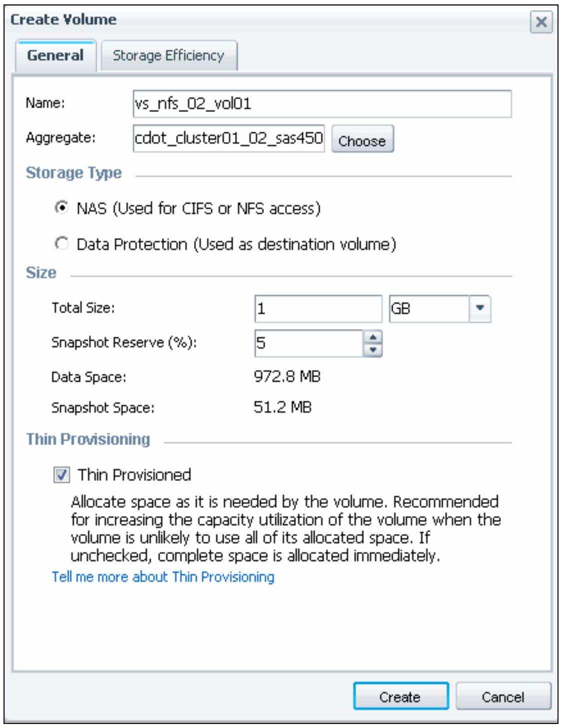
1. Click **Vservers** → **Storage** → **Volumes**. See Figure 24-18. Click the **Create** button.



Name	Aggregate	Status
vs_nfs_02_root	cdot_cluster01_01_sas450_01	online

Figure 24-18 Create a NFS volume

2. After you click the **Create** button, you will see the information shown in Figure 24-19.



Name: vs_nfs_02_vol01

Aggregate: cdot_cluster01_02_sas450

Storage Type

● NAS (Used for CIFS or NFS access)

○ Data Protection (Used as destination volume)

Size

Total Size: 1 GB

Snapshot Reserve (%): 5

Data Space: 972.8 MB

Snapshot Space: 51.2 MB

Thin Provisioning

☑ Thin Provisioned

Allocate space as it is needed by the volume. Recommended for increasing the capacity utilization of the volume when the volume is unlikely to use all of its allocated space. If unchecked, complete space is allocated immediately.

Tell me more about Thin Provisioning

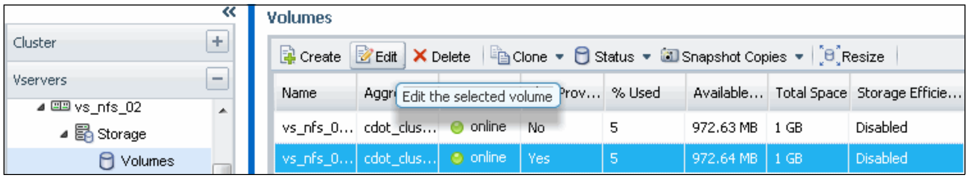
Figure 24-19 Create Volume

Here we explain some of the fields for this window:

- Name: Specify the name that you want.
- Aggregate: Specify the aggregate that provide the CIFS volume.
- Storage Type: In this case, NAS is for the NFS services. If you create a snapmirror destination volume, you can select a Data Protection type.
- Size: Specify the size and the percentage for the snapshot.
- Thin Provisioned: If you want to configure the thin provisioning volume, check this check box.

  With thin provisioning, when you create volumes and LUNs for different purposes in a given aggregate, you do not actually allocate any space for those volumes in advance. The space is allocated as data is written to the volumes.

3. After the wizard completes the volume creation, verify the new volume in the volume list. Figure 24-20 shows one example.



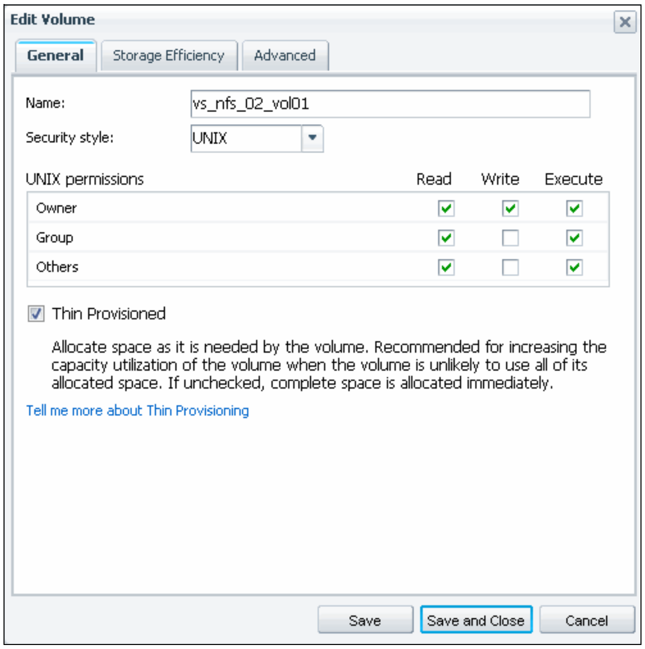
Name	Aggr		Prov…	% Used	Available…	Total Space	Storage Efficie…
vs_nfs_0…	cdot_clus…	online	No	5	972.63 MB	1 GB	Disabled
vs_nfs_0…	cdot_clus…	online	Yes	5	972.64 MB	1 GB	Disabled

Figure 24-20 Check the volume made

Editing NFS volume properties

After you complete the creation of NFS volume, you can check and change the default properties made by the wizard. Click the **Edit** button. See Figure 24-20.

When the Edit Volume window appears, you can see the panel in Figure 24-21.



Name: vs_nfs_02_vol01

Security style: UNIX

UNIX permissions	Read	Write	Execute
Owner	☑	☑	☑
Group	☑	□	☑
Others	☑	□	☑

☑ Thin Provisioned

Allocate space as it is needed by the volume. Recommended for increasing the capacity utilization of the volume when the volume is unlikely to use all of its allocated space. If unchecked, complete space is allocated immediately.

Tell me more about Thin Provisioning

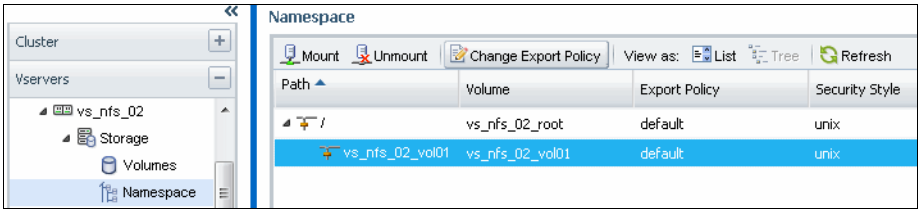
Figure 24-21 Edit Volume

In this window, you can change the permissions for each UNIX Owner, Group, and Others.

Note: The default values are that the write permissions for Group and Others are not granted.

Assigning NFS volumes to NFS hosts

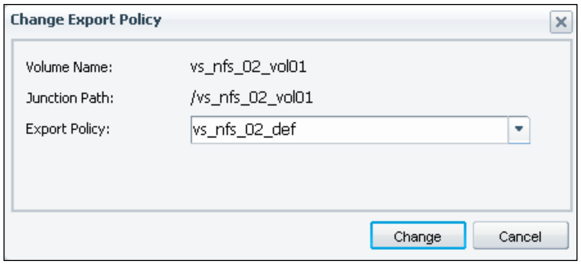
When we create a CIFS share, we need to create additional shares for CIFS users. But in the case of NFS protocols, you do not need to create any special shares like CIFS. The only thing is setting the export policy to export NFS volumes. See Figure 24-22.



Path	Volume	Export Policy	Security Style
/	vs_nfs_02_root	default	unix
vs_nfs_02_vol01	vs_nfs_02_vol01	default	unix

Figure 24-22 Namespace for NFS

To change the export policy, click the **Change Export Policy** button. The default export policy is default. You can change the export policy with your own purpose. See Figure 24-23.



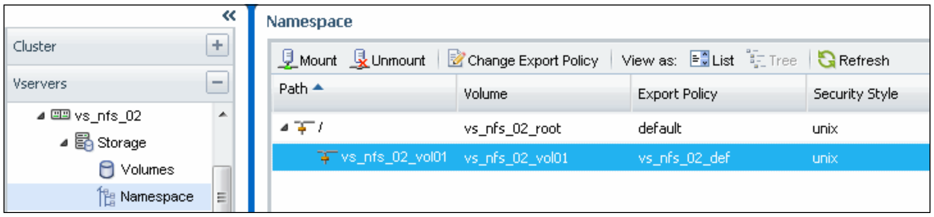
Volume Name: vs_nfs_02_vol01

Junction Path: /vs_nfs_02_vol01

Export Policy: vs_nfs_02_def

Figure 24-23 Change Export policy

See Figure 24-24 for the results after you complete the setting of the export policy.



Path	Volume	Export Policy	Security Style
/	vs_nfs_02_root	default	unix
vs_nfs_02_vol01	vs_nfs_02_vol01	vs_nfs_02_def	unix

Figure 24-24 Namespace after changing the export policy

Through this chapter, we created the SVM only for NFS and assign the export policy for only one Ubuntu Linux host. We can access the NFS volumes at the next step.

24.2.5 Mounting volumes on a UNIX host

You can mount NFS volumes on any UNIX host if the host has been defined by the export rule. In your case, the host should have the IP address of 9.155.133.206.

NFS connection test between a UNIX host and the SVM

You can use lots of methods to test the connection between the UNIX host and the SVM, for example, ping, showmount, port scan, and so on. In this phase, we show you the difference between 7-Mode and clustered mode Data ONTAP using a **`showmount`** command.

We use the command **`showmount -e`** to check the list of the exported volumes from an NFS server. Example 24-7 shows the actual output of this command.

Example 24-7 Showmount in the Clustered Data ONTAP

```
root@Ubuntu:/# showmount -e 9.155.66.33
Export list for 9.155.66.33:
/ (everyone)
root@Ubuntu:/# showmount -a 9.155.66.33
All mount points on 9.155.66.33:
root@Ubuntu:/#
```

> **Note:** The command **`showmount -a`** or **`showmount -e`** to a Cluster-Mode SVM will not display mount points because **`rmtab`** is not used.

If you are using the commands with 7-Mode or lower Data ONTAP, you can expect to see the following results, as shown in Example 24-8.

Example 24-8 Showmount in the 7-Mode or lower version Data ONTAP

```
root@Ubuntu:/# showmount -e 9.155.66.33
Export list for 9.155.66.33:
/ (everyone)
/vs_nfs_02_vol01 (everyone)
root@Ubuntu:/# showmount -a 9.155.66.33
All mount points on 9.155.66.33:
9.155.113.206:/
9.155.113.206:/vs_nfs_02_vol01
root@Ubuntu:/#
```

In a Clustered Data ONTAP environment, you can use the commands of **`showmount`** to test the NFS connection between the UNIX host and the SVM in the Clustered Data ONTAP. See this change.

Mounting NFS volumes on a Linux host

Before mounting NFS volumes, you have to create directories to mount the NFS volumes. In this case, we assumed that you already made the following folders. See Example 24-9.

Example 24-9 Linux directories

```
root@Ubuntu:/mnt# ls
vs-nas  vs-nas-sub
```

After creating directories, you can mount NFS volumes with the following commands.
See Example 24-10.

NFS -t nfs *<NFS server ip address>*:/<*volume(s) path(s)> <local path>*

Example 24-10 Mount NFS volumes and check the result

```
root@Ubuntu:/mnt# mount -t nfs 9.155.66.33:/ /mnt/vs-nas
root@Ubuntu:/mnt# mount -t nfs 9.155.66.33:/vs_nfs_02_vol01 /mnt/vs-nas-sub
root@Ubuntu:/mnt# ls
hgfs  vs-nas  vs-nas-sub
```

When you mount the NFS volumes, you can select other options such as TCP/UDP, NFS versions, and so on. From a performance point of view, you can change the transfer size as well. But before making some changes on NFS settings, considering your environment, we advise you to contact your IBM representative.

24.3 Further information

More details on CIFS implementation can be found in the following Redbooks publications:

- *OnCommand System Manager 3.0 Help for Use with Clustered Data ONTAP Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

- *Clustered Data ONTAP 8.2 File Access and Protocols Management Guide*, located at this website:

  http://www.ibm.com/support/entry/portal/documentation_expanded_list?rebuildGLPortletViewBean=Data%2520ONTAP

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 and IBM Redpaper publications

The following IBM Redbooks publications and IBM Redpaper 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 Software Guide*, SG24-7129
- *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 MetroCluster,* REDP-4259
- *IBM System Storage N series with FlexShare*, REDP-4291
- *IBM System Storage N series with VMware vSphere 4.1 using Virtual Storage Console 2*, REDP-4863

You can search for, view, download or order these documents and other Redbooks publications, Redpaper publications, Web Docs, draft 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

- VMware documentation:

  http://www.vmware.com/support/pubs/

- VMware vSphere 5 documentation:

  http://www.vmware.com/support/pubs/vsphere-esxi-vcenter-server-pubs.html
  http://pubs.vmware.com/vsphere-50/index.jsp

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 Clustered Data ONTAP

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

IBM System Storage N series Clustered Data ONTAP

Understand Clustered Data ONTAP benefits for dynamic storage solution

Learn about Clustered Data ONTAP features and functions

Design scaleable NAS solutions using N series

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 with Clustered Data ONTAP Version 8.2, which is the first version available on the IBM System Storage N series, and as of October 2013, is also the most current version available. Clustered Data ONTAP is different from previous ONTAP versions by the fact that it offers a storage solution that operates as a cluster with flexible scaling capabilities. Clustered Data ONTAP configurations allow clients to build a scale-out architecture, protecting their investment and allowing horizontal scaling of their environment.

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.

SG24-8200-00 ISBN 0738439304