

IBM TotalStorage SAN Volume Controller



Planning Guide

Version 1.2.1

IBM TotalStorage SAN Volume Controller



Planning Guide

Version 1.2.1

Fourth Edition (October 2004)

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

© Copyright International Business Machines Corporation 2003, 2004. All rights reserved.

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	v
Tables	vii
About this guide	ix
Who should use this guide?	ix
Summary of Changes	ix
Summary of Changes for GA22-1052-03 SAN Volume Controller Planning Guide	ix
Emphasis	x
SAN Volume Controller library and related publications.	x
Related Web sites	xii
How to order IBM publications	xii
How to send your comments.	xiii
Chapter 1. Virtualization and the IBM TotalStorage SAN Volume Controller	1
Virtualization	1
The need for virtualization	3
Fabric level virtualization models	3
Symmetric virtualization	4
SAN Volume Controller	5
Uninterruptible power supply overview.	8
Master console	11
Overview of backup functions	12
Cluster configuration backup	13
FlashCopy	13
Remote Copy	14
Chapter 2. Installation planning	17
Preparing your SAN Volume Controller environment	17
Preparing your uninterruptible power supply environment	18
Preparing your master console environment	19
Ports and connections	20
Chapter 3. Preparing the physical configuration	23
Completing the hardware location chart	23
Hardware location guidelines.	24
Hardware location chart	25
Completing the cable connection table	26
Cable connection table	27
Completing the configuration data table	28
Configuration data table	29
Chapter 4. Planning guidelines for using your SAN Volume Controller in a SAN environment.	31
Storage Area Network	31
Switch zoning for the SAN Volume Controller.	32
Zoning considerations for Remote Copy.	35
Switch operations over long distances	36
Performance of fibre-channel extenders.	37
Nodes	37
Clusters	38
Cluster state	39

Cluster operation and quorum disks	39
I/O groups and uninterruptible power supply	40
Uninterruptible power supply and power domains	41
Disk controllers	42
Data migration	43
Image mode virtual disk migration	44
Copy Services	44
FlashCopy	45
FlashCopy mappings.	45
FlashCopy consistency groups	48
Remote Copy	49
Synchronous Remote Copy	50
Remote Copy consistency groups	50
Chapter 5. Object descriptions	53
Storage subsystems	54
Managed disks	56
Managed disk groups	58
Virtual disks	60
Virtual disk-to-host mapping	62
Host objects	64
Chapter 6. Planning for configuring the SAN Volume Controller	67
Maximum configuration	68
Configuration rules and requirements.	70
Configuration rules	71
Storage subsystems	71
Host bus adapters.	75
Nodes	76
Power requirements	77
Fibre-channel switches	77
Configuration requirements	79
Chapter 7. SAN Volume Controller supported environment	83
Supported host attachments	83
Supported storage subsystems	83
Supported fibre-channel host bus adapters	83
Supported switches	84
Supported fibre-channel extenders.	84
Accessibility	85
Notices	87
Trademarks	88
Definitions of notices.	88
Glossary	91
Index	99

Figures

1. Levels of virtualization	2
2. Symmetrical virtualization	4
3. A SAN Volume Controller node	6
4. Example of a SAN Volume Controller in a fabric	7
5. Uninterruptible power supply	9
6. I/O groups and uninterruptible power supply relationship	11
7. Cluster, nodes, and cluster state.	39
8. I/O group and uninterruptible power supply	41
9. Relationship between I/O groups and uninterruptible power supply units	42
10. Objects in a virtualized system.	54
11. Controllers and MDisks	57
12. MDisk group	60
13. Managed disk groups and VDIs	61
14. Hosts, WWPNs, and VDIs	64
15. Hosts, WWPNs, VDIs and SCSI mappings	64
16. Disk controller system shared between SAN Volume Controller and a host	73
17. ESS LUs accessed directly with a SAN Volume Controller	74
18. FASTT direct connection with a SAN Volume Controller on one host	75
19. Fabric with Inter-Switch Links between nodes in a cluster	79
20. Fabric with Inter-Switch Links in a redundant configuration	79

Tables

1.	Emphasis descriptions	x
2.	Publications in the SAN Volume Controller library	xi
3.	Other IBM publications	xii
4.	Web sites	xii
5.	17
6.	Sample of completed hardware location chart	24
7.	Hardware location chart	25
8.	Cable connection table	27
9.	Example of cable connection table	28
10.	Four hosts and their ports	33
11.	Six hosts and their ports	34
12.	Node state	38
13.	Required uninterruptible power supply (UPS) units	42
14.	Managed disk status	57
15.	Managed disk group status	59
16.	Capacities of the cluster given extent size	60
17.	Virtual disk status	62
18.	SAN Volume Controller maximum configuration values	68

About this guide

This publication introduces the IBM® TotalStorage® SAN Volume Controller, its components and its features.

It also provides planning guidelines for installing and configuring the SAN Volume Controller.

Related tasks

Chapter 2, “Installation planning,” on page 17

Before the service representative can start to set up your SAN Volume Controller, verify that the prerequisite conditions for the SAN Volume Controller and uninterruptible power supply installation are met.

Chapter 3, “Preparing the physical configuration,” on page 23

Before the service representative installs the SAN Volume Controller, uninterruptible power supply unit, and master console, you must plan the physical configuration and the initial settings for the system.

Chapter 4, “Planning guidelines for using your SAN Volume Controller in a SAN environment,” on page 31

Follow these planning steps to set up your SAN Volume Controller environment.

Related reference

“Accessibility” on page 85

Accessibility features help a user who has a physical disability, such as restricted mobility or limited vision, to use software products successfully.

Chapter 7, “SAN Volume Controller supported environment,” on page 83

The IBM Web site provides up-to-date information about the supported environment for the SAN Volume Controller.

Chapter 6, “Planning for configuring the SAN Volume Controller,” on page 67

Before you configure the SAN Volume Controller, you must complete these planning tasks.

Who should use this guide?

This publication is intended for anyone who is planning to install and configure an IBM TotalStorage SAN Volume Controller.

Summary of Changes

This document contains terminology, maintenance, and editorial changes.

Technical changes or additions to the text and illustrations are indicated by a vertical line to the left of the change. This summary of changes describes new functions that have been added to this release.

Summary of Changes for GA22-1052-03 SAN Volume Controller Planning Guide

The Summary of Changes provides a list of new, modified, and changed information since the last version of the guide.

New information

This topic describes the changes to this guide since the previous edition, GA22-1052-02.

This version includes the following new information:

- Clusters can contain from one to four pairs of nodes.
- Software which is installed on the master console is listed.
- A cluster must have two to four uninterruptible power supply units depending on the number of nodes.
- Except when migrating between groups, a VDisk must be associated with just one MDisk group.
- An MDisk can be associated with just one MDisk group.
- There is a new procedure for installing and configuring switches to create zones.

Changed information

No information has been changed in this version.

Deleted information

No deletions were made in this version.

Emphasis

Different typefaces are used in this guide to show emphasis.

The following typefaces are used to show emphasis:

Table 1. Emphasis descriptions

Boldface	Text in boldface represents menu items and command names.
<i>Italics</i>	Text in <i>italics</i> is used to emphasize a word. In command syntax, it is used for variables for which you supply actual values, such as a default directory or the name of a cluster.
Monospace	Text in monospace identifies the data or commands that you type, samples of command output, examples of program code or messages from the system, or names of command flags, parameters, arguments, and name-value pairs.

SAN Volume Controller library and related publications

A list of other publications that are related to this product are provided to you for your reference.

The tables in this section list and describe the following publications:

- The publications that make up the library for the IBM TotalStorage SAN Volume Controller
- Other IBM publications that relate to the SAN Volume Controller

SAN Volume Controller library

Table 2 on page xi lists and describes the publications that make up the SAN Volume Controller library. Unless otherwise noted, these publications are available in Adobe portable document format (PDF) on a compact disc (CD) that comes with

the SAN Volume Controller. If you need additional copies of this CD, the order number is SK2T-8811. These publications are also available as PDF files from the following Web site:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Table 2. Publications in the SAN Volume Controller library

Title	Description	Order number
<i>IBM TotalStorage SAN Volume Controller: CIM Agent Developer's Reference</i>	This reference guide describes the objects and classes in a Common Information Model (CIM) environment.	SC26-7590
<i>IBM TotalStorage SAN Volume Controller: Command-Line Interface User's Guide</i>	This guide describes the commands that you can use from the SAN Volume Controller command-line interface (CLI).	SC26-7544
<i>IBM TotalStorage SAN Volume Controller: Configuration Guide</i>	This guide provides guidelines for configuring your SAN Volume Controller.	SC26-7543
<i>IBM TotalStorage SAN Volume Controller: Host Attachment Guide</i>	This guide provides guidelines for attaching the SAN Volume Controller to your host system.	SC26-7575
<i>IBM TotalStorage SAN Volume Controller: Installation Guide</i>	This guide includes the instructions the service representative uses to install the SAN Volume Controller.	SC26-7541
<i>IBM TotalStorage SAN Volume Controller: Planning Guide</i>	This guide introduces the SAN Volume Controller and lists the features you can order. It also provides guidelines for planning the installation and configuration of the SAN Volume Controller.	GA22-1052
<i>IBM TotalStorage SAN Volume Controller: Service Guide</i>	This guide includes the instructions the service representative uses to service the SAN Volume Controller.	SC26-7542
<i>IBM TotalStorage SAN Volume Controller: Translated Safety Notices</i>	This guide contains the danger and caution notices for the SAN Volume Controller. The notices are shown in English and in numerous other languages.	SC26-7577

Other IBM publications

Table 3 on page xii lists and describes other IBM publications that contain additional information related to the SAN Volume Controller.

Table 3. Other IBM publications

Title	Description	Order number
<i>IBM TotalStorage Enterprise Storage Server, IBM TotalStorage SAN Volume Controller, IBM TotalStorage SAN Volume Controller for Cisco MDS 9000, Subsystem Device Driver: User's Guide</i>	This guide describes the IBM Subsystem Device Driver Version 1.5 for TotalStorage Products and how to use it with the SAN Volume Controller. This publication is referred to as the <i>IBM TotalStorage Subsystem Device Driver: User's Guide</i> .	SC26-7608

Related Web sites

Table 4 lists Web sites that have information about SAN Volume Controller or related products or technologies.

Table 4. Web sites

Type of information	Web site
SAN Volume Controller support	http://www-1.ibm.com/servers/storage/support/virtual/2145.html
Technical support for IBM storage products	http://www.ibm.com/storage/support/

How to order IBM publications

The publications center is a worldwide central repository for IBM product publications and marketing material.

The IBM publications center

The IBM publications center offers customized search functions to help you find the publications that you need. Some publications are available for you to view or download free of charge. You can also order publications. The publications center displays prices in your local currency. You can access the IBM publications center through the following Web site:

www.ibm.com/shop/publications/order/

Publications notification system

The IBM publications center Web site offers you a notification system for IBM publications. Register and you can create your own profile of publications that interest you. The publications notification system sends you a daily e-mail that contains information about new or revised publications that are based on your profile.

If you want to subscribe, you can access the publications notification system from the IBM publications center at the following Web site:

www.ibm.com/shop/publications/order/

How to send your comments

Your feedback is important to help us provide the highest quality information. If you have any comments about this book or any other documentation, you can submit them in one of the following ways:

- e-mail

Submit your comments electronically to the following e-mail address:

starpubs@us.ibm.com

Be sure to include the name and order number of the book and, if applicable, the specific location of the text you are commenting on, such as a page number or table number.

- Mail

Fill out the Readers' Comments form (RCF) at the back of this book. If the RCF has been removed, you can address your comments to:

International Business Machines Corporation
RCF Processing Department
Department 61C
9032 South Rita Road
Tucson, Arizona 85775-4401
U.S.A.

Chapter 1. Virtualization and the IBM TotalStorage SAN Volume Controller

Before you begin planning the installation, understand why virtualization is needed, what virtualization is, and the IBM TotalStorage SAN Volume Controller system.

Virtualization

Virtualization is a concept that applies to many areas of the information technology industry.

Where data storage is concerned, virtualization includes the creation of a pool of storage that contains several disk subsystems. These subsystems can be from various vendors. The pool can be split into virtual disks that are visible to the host systems that use them. Therefore, virtual disks can use mixed back-end storage and provide a common way to manage a storage area network (SAN).

Historically, the term *virtual storage* has described the virtual memory techniques that have been used in operating systems. The term *storage virtualization*, however, describes the shift from managing physical volumes of data to logical volumes of data. This shift can be made on several levels of the components of storage networks. Virtualization separates the representation of storage between the operating system and its users from the actual physical storage components. This technique has been used in mainframe computers for many years through methods such as system-managed storage and products like the IBM Data Facility Storage Management Subsystem (DFSMS). Virtualization can be applied at four main levels:

- Virtualization at the *server* level is performed by managing volumes on the operating systems servers. An increase in the amount of logical storage over physical storage is suitable for environments that do not have storage networks.
- Virtualization at the *storage device* level is in common use. Striping, mirroring, and redundant arrays of independent disks (RAIDs) are used by almost all disk subsystems. This type of virtualization can range from simple RAID controllers to advanced volume management such as that provided by the IBM TotalStorage Enterprise Storage Server (ESS) or by Log Structured Arrays (LSA). The Virtual Tape Server (VTS) is another example of virtualization at the device level.
- Virtualization at the *fabric* level enables storage pools to be independent of the servers and the physical components that make up the storage pools. One management interface can be used to manage different storage systems without affecting the servers. The SAN Volume Controller is used to perform virtualization at the fabric level.
- Virtualization at the *file system* level provides the highest benefit because data is shared, allocated, and protected, not volumes.

Virtualization is a radical departure from traditional storage management. In traditional storage management, storage is attached directly to a host system, which controls storage management. SANs introduced the principle of networks of storage, but storage is still primarily created and maintained at the RAID subsystem level. Multiple RAID controllers of different types require knowledge of, and software that is specific to, the given hardware. Virtualization brings a central point of control for disk creation and maintenance. It brings new ways of handling storage maintenance.

Where storage is concerned, one problematic area that virtualization addresses is that of unused capacity. Rather than individual storage systems remaining islands unto themselves, allowing excess storage capacity to be wasted when jobs do not require it, storage is pooled so that jobs needing the highest storage capacity can use it when they need it. Regulating the amount of storage available becomes easier to orchestrate without computing resource or storage resource having to be turned off and on.

Types of virtualization

Virtualization can be performed either asymmetrically or symmetrically. See Figure 1 for more information.

Asymmetric

A virtualization engine is outside the data path and performs a metadata style service.

Symmetric

A virtualization engine sits in the data path, presenting disks to the hosts but hiding the physical storage from the hosts. Advanced functions, such as cache and Copy Services, can therefore be implemented in the engine itself.

Virtualization at any level provides benefits. When several levels are combined, however, the benefits of those levels can also be combined. An example of how you can gain the highest benefits is if you attach a low cost RAID controller to a virtualization engine that provides virtual volumes for use by a virtual file system.

Note: The SAN Volume Controller implements fabric-level *virtualization*. Within the context of the SAN Volume Controller and throughout this document, *virtualization* refers to symmetric fabric-level virtualization.

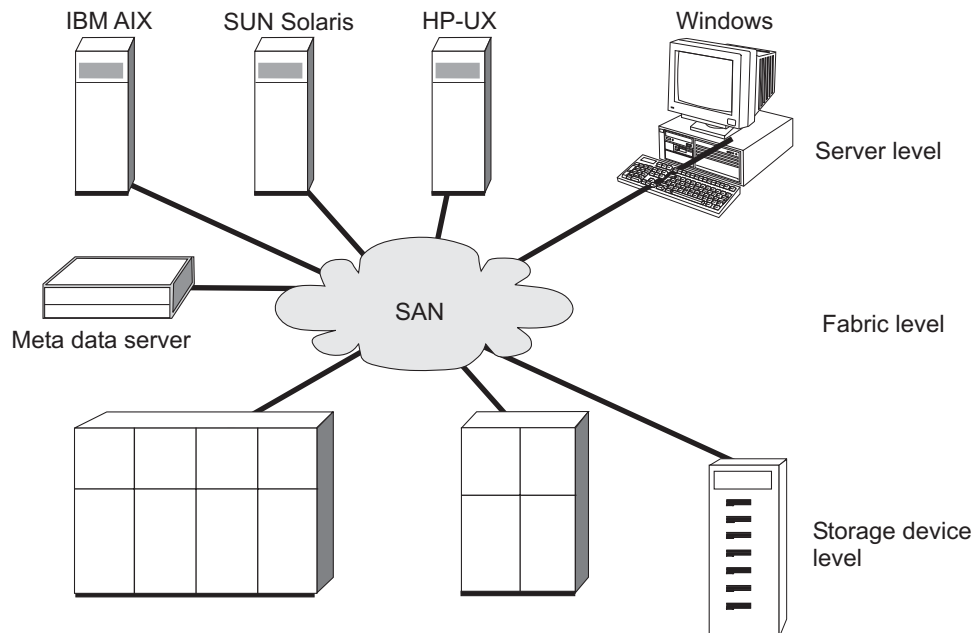


Figure 1. Levels of virtualization

Related concepts

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

The need for virtualization

Storage is a facility that computer users want to access at any time, from any location, with a minimum amount of management.

Users expect the storage devices to provide enough capacity and to be reliable. The amount of storage that users require, however, is increasing quickly. Internet users use large amounts of storage daily. Many users are mobile, access patterns cannot be predicted, and the content of the data becomes more and more interactive. Because the amount of data that is handled is large, it can no longer be managed manually. Automatic management is required, as are new levels of bandwidth and load balancing. Also, it is important that all this data can be shared between different types of computer platforms, because the communication networks cannot handle the large replication, download, and copying operations that are required.

Storage area networks (SANs) are high speed switched networks that let multiple computers share access to many storage devices. SANs allow for the use of advanced software that automatically manages the storage of data. With such advanced software, the computers that are connected to a particular network can, therefore, access storage wherever that storage is available in the network. The user is no longer aware of, and no longer needs to know, which physical devices contain which data. The storage has become virtualized. In a similar way to how virtual memory has solved the problems of the management of a limited resource in application programs, the virtualization of storage has given users a more intuitive use of storage, while software quietly manages the storage network in the background.

Fabric level virtualization models

In traditional storage management, storage devices are connected directly to host systems and are maintained locally by those host systems.

Although storage area networks (SANs) have introduced the principle of networks, storage devices are still mainly assigned to individual host systems and storage is still mainly created and maintained at the RAID subsystem level. Therefore, RAID controllers of different types need knowledge of, and software that is specific to, the hardware that is used.

Virtualization provides a complete change from the traditional storage management. It provides a central point of control for disk creation and management, and therefore requires changes to the way in which storage management is done.

Fabric level virtualization is the principle in which a pool of storage is created from more than one disk subsystem. This pool is then used to set up virtual disks that are made visible to the host systems. These virtual disks use whatever storage is available and permit a common way to manage SAN storage.

Fabric level virtualization can be done in either of two ways: asymmetric or symmetric.

With asymmetric virtualization, the virtualization engine is outside the data path. It provides a metadata server that contains all the mapping and the locking tables. The storage devices contain only data.

Because the flow of control is separated from the flow of data, input/output (I/O) operations can use the full bandwidth of the SAN. A separate network or SAN link is used for control.

However, there are disadvantages to asymmetric virtualization:

- Data is at risk to increased security exposures, and the control network must be protected with a firewall.
- Metadata can become very complicated when files are distributed across several devices.
- Each host that accesses the SAN must know how to access and interpret the metadata. Specific device driver or agent software must therefore be running on each of these hosts.
- The metadata server cannot run advanced functions, such as caching or Copy Services, because it only has access to the metadata, not the data itself.

Symmetric virtualization

The SAN Volume Controller provides symmetric virtualization.

Virtualization splits the physical storage Redundant Array of Independent Disks (RAID) arrays into smaller chunks of storage that are known as extents. These extents are then concatenated together, using various policies, to make virtual disks. With symmetric virtualization, host systems can be isolated from the physical storage. Advanced functions, such as data migration, can run without the need to re-configure the host. With symmetric virtualization, the virtualization engine is the central configuration point for the SAN.

In symmetric virtual storage networks (see Figure 2), data and control both flow over the same path. Because the separation of the control from the data occurs in the data path, the storage can be pooled under the control of the virtualization engine. The virtualization engine performs the logical-to-physical mapping.

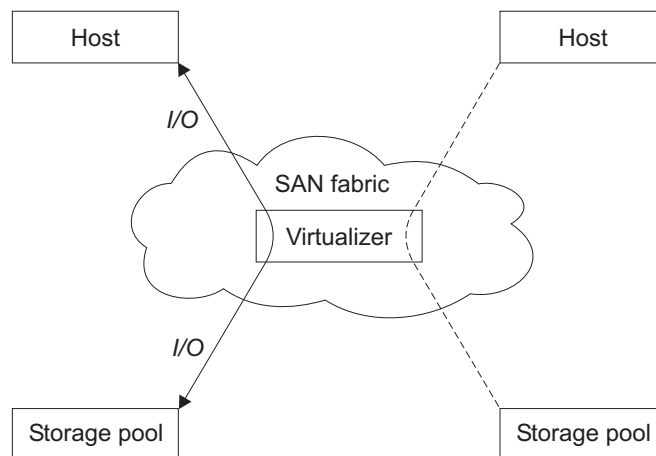


Figure 2. Symmetrical virtualization

The virtualization engine directly controls access to the storage and to the data that is written to the storage. As a result, locking functions that provide data integrity and

advanced functions, such as cache and copy services, can be run in the virtualization engine itself. The virtualization engine is, therefore, a central point of control for device and advanced function management. Symmetric virtualization also allows you to build a kind of firewall in the storage network. Only the virtualization engine can give access through the firewall. Symmetric virtualization does, however, cause some problems.

The main problem that is associated with symmetric virtualization is related to poor performance, because all I/O must flow through the virtualization engine. This problem is one of scalability. You can use an n-way cluster of virtualization engines that has failover capacity to solve this problem. You can scale the additional processor power, cache memory, and adapter bandwidth to get the level of performance that you want. The memory and processing power can be used to run the advanced functions, such as copy services and caching.

The IBM TotalStorage SAN Volume Controller uses symmetric virtualization. Single virtualization engines, which are known as nodes, are combined to create clusters. Each cluster can contain between two and eight nodes.

Related concepts

“Virtualization” on page 1

Virtualization is a concept that applies to many areas of the information technology industry.

SAN Volume Controller

The SAN Volume Controller is a SAN appliance that attaches open-systems storage devices to supported open-systems hosts.

The IBM TotalStorage SAN Volume Controller provides symmetric virtualization by creating a pool of managed disks from the attached storage subsystems, which are then mapped to a set of virtual disks for use by attached host computer systems. System administrators can view and access a common pool of storage on the SAN, which enables them to use storage resources more efficiently and provides a common base for advanced functions.

The SAN Volume Controller is analogous to a logical volume manager (LVM) on a SAN. It performs the following functions for the SAN storage that it is controlling:

- Creates a single pool of storage
- Manages logical volumes
- Provides advanced functions for the SAN, such as:
 - Large scalable cache
 - Copy services
 - Point-in-time Copy
 - FlashCopy® (point-in-time copy)
 - Remote Copy (synchronous copy)
 - Data migration
 - Space management
 - Mapping that is based on desired performance characteristics
 - Quality of service metering

A *node* is a single storage engine. See Figure 3 on page 6 for a visual of a node. The storage engines are always installed in pairs with one to four pairs of nodes

constituting a *cluster*. Each node in a pair is configured to back up the other. Each pair of nodes is known as an *I/O group*. All I/O operations that are managed by the nodes in an I/O group are cached on both nodes for resilience. Each virtual volume is defined to an I/O group. To avoid any single point of failure, the nodes of an I/O group are protected by independent uninterruptible power supply units.

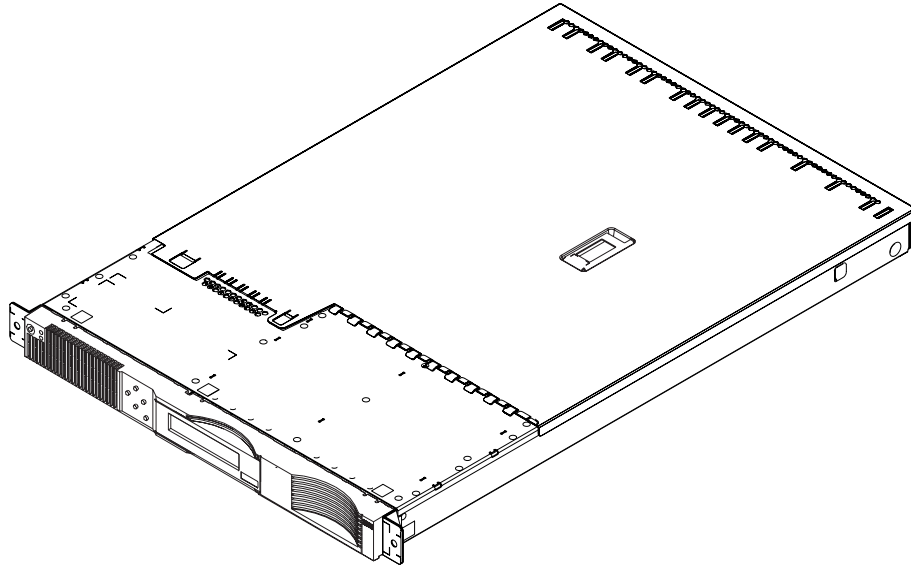


Figure 3. A SAN Volume Controller node

The SAN Volume Controller I/O groups see the storage presented to the SAN by the backend controllers as a number of disks known as *managed disks*. The application services do not see these managed disks. Instead they see a number of logical disks, known as *virtual disks*, that are presented to the SAN by the SAN Volume Controller. Each node must only be in one I/O group and provide access to the virtual disks in the I/O group.

The SAN Volume Controller helps to provide continuous operations and can also optimize the data path to ensure performance levels are maintained. Ensure that you use IBM TotalStorage Multiple Device Manager performance manager to analyze the performance statistics. See *IBM TotalStorage Multiple Device Manager Configuration and Installation Guide* and *IBM TotalStorage Multiple Device Manager CLI Guide* for more information.

The fabric contains two distinct zones: a host zone and a disk zone. In the host zone, the host systems can identify and address the nodes. You can have more than one host zone. Generally, you will create one host zone per operating system type. In the disk zone, the nodes can identify the disk drives. Host systems cannot operate on the disk drives directly; all data transfer occurs through the nodes. As shown in Figure 4 on page 7, several host systems can be connected to a SAN fabric. A cluster of SAN Volume Controllers is connected to the same fabric and presents virtual disks to the host systems. You configure these virtual disks using the disks located on the RAID controllers.

Note: You can have more than one host zone. Generally you create one host zone per operating system type because some operating systems will not tolerate other operating systems in the same zone.

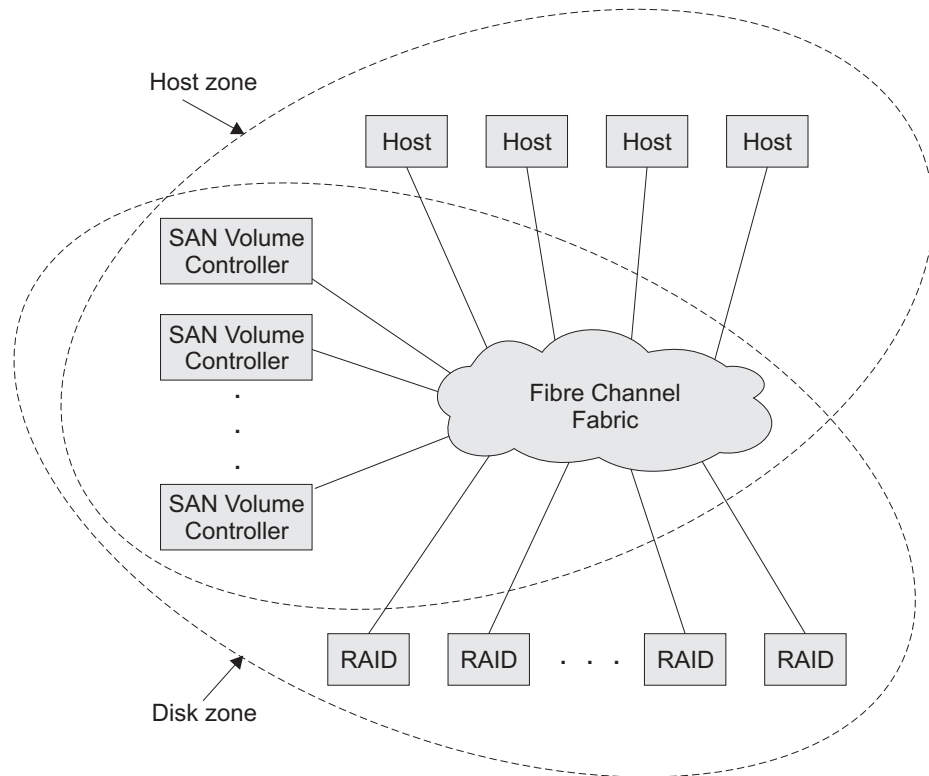


Figure 4. Example of a SAN Volume Controller in a fabric

You can remove one node in each I/O group from a cluster when hardware service or maintenance is required. After you remove the node, you can replace the field replaceable units (FRUs) in the node. All disk drive communication and communication between nodes is performed through the SAN. All SAN Volume Controller configuration and service commands are sent to the cluster through an Ethernet network.

Each node contains its own vital product data (VPD). Each cluster contains VPD that is common to all the nodes on the cluster, and any system connected to the Ethernet network can access this VPD.

Cluster configuration information is stored on every node that is in the cluster to allow concurrent replacement of FRUs. An example of this information might be information that is displayed on the menu screen of the SAN Volume Controller. When a new FRU is installed and when the node is added back into the cluster, configuration information that is required by that node is read from other nodes in the cluster.

SAN Volume Controller operating environment

- Minimum of one pair of SAN Volume Controller nodes
- Minimum two uninterruptible power supplies
- One master console is required per SAN installation for configuration

Features of a SAN Volume Controller node

- 19-inch rack mounted enclosure
- 4 fibre channel ports
- 2 fibre channel adapters

- 4 GB cache memory

Supported hosts

For a list of supported operating systems, see the IBM TotalStorage SAN Volume Controller Web site at:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Multipathing software

- IBM Subsystem Device Driver (SDD)
- Redundant Dual Active Controller (RDAC)

Note: Direct attach hosts sharing a back end storage controller with a SAN Volume Controller can run multipath drivers SDD and RDAC. There is no support for the co-existence of native multipath drivers with SDD on the same host.

Check the following Web site for the latest support and coexistence information:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

User interfaces

The SAN Volume Controller provides the following user interfaces:

- IBM TotalStorage SAN Volume Controller Console, a Web-accessible graphical user interface (GUI) that supports flexible and rapid access to storage management information
- A command-line interface (CLI) using Secure Shell (SSH)

Application programming interfaces

The SAN Volume Controller provides the following application programming interface:

- IBM TotalStorage Common Information Model (CIM) Agent for the SAN Volume Controller, which supports the Storage Management Initiative Specification of the Storage Network Industry Association.

Related concepts

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Related reference

“Supported host attachments” on page 83

The IBM Web site provides up-to-date information about the supported host attachment operating systems.

Uninterruptible power supply overview

The uninterruptible power supply provides the SAN Volume Controller with a secondary power source to be used if you lose power from your primary power source due to power failures, power sags, power surges, or line noise.

If a power outage occurs, the uninterruptible power supply will maintain power long enough to save any configuration and cache data contained in the dynamic random

access memory (DRAM). The data will be saved to the SAN Volume Controller internal disk. Figure 5 provides a visual of the uninterruptible power supply.

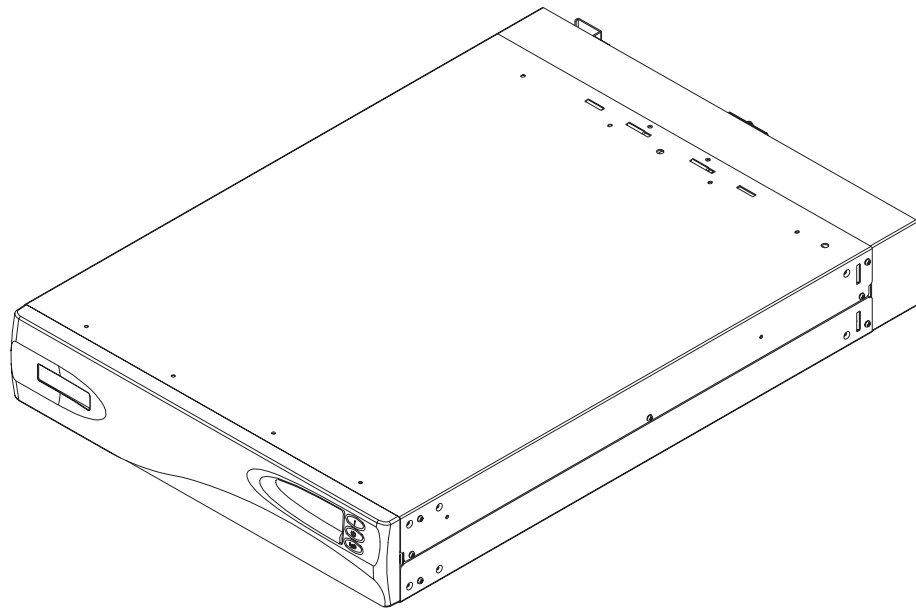


Figure 5. Uninterruptible power supply

Note: The SAN Volume Controller uninterruptible power supply is an integral part of the SAN Volume Controller solution, and maintains continuous SAN Volume Controller specific communications with its attached SAN Volume Controller nodes. The SAN Volume Controller will not operate without the uninterruptible power supply. The SAN Volume Controller uninterruptible power supply must be used in accordance with documented guidelines and procedures and must not power any equipment other than SAN Volume Controller nodes.

To provide full redundancy and concurrent maintenance, the SAN Volume Controller must be installed in pairs. Each SAN Volume Controller of a pair must be connected to a different uninterruptible power supply. Each uninterruptible power supply can support up to two SAN Volume Controller nodes. It is also recommended that you connect the two uninterruptible power supply units for the pair to different independent electrical power sources. This reduces the chance of an input power failure at both uninterruptible power supply units.

The uninterruptible power supply must be in the same rack as the nodes. When using 6 or 8 node support, ensure that 4 uninterruptible power supplies are used. Ensure that you are following the uninterruptible power supply support guidelines as described below:

Number of nodes	Number of uninterruptible power supplies
2	2
4	2
6	4
8	4

Attention:

1. Do not connect the uninterruptible power supplies to an input power source that does not conform to standards. Review the requirements for uninterruptible power supplies listed under "Related reference" at the end of this topic.
2. Each uninterruptible power supply pair must power only one SAN Volume Controller cluster.

Each uninterruptible power supply includes power (line) cords that will connect the uninterruptible power supply to either a rack power distribution unit (PDU), if one exists, or to an external power source. Each uninterruptible power supply power input requires the protection of a UL approved (or equivalent) 250 volt, 15 amp circuit breaker.

The uninterruptible power supply is connected to the SAN Volume Controllers with a power cable and a signal cable. To avoid the possibility of power and signal cables being connected to different uninterruptible power supply units, these cables are wrapped together and supplied as a single field replaceable unit. The signal cables enable the SAN Volume Controllers to read status and identification information from the uninterruptible power supply.

Each SAN Volume Controller monitors the operational state of the uninterruptible power supply to which it is attached. If the uninterruptible power supply reports a loss of input power, the SAN Volume Controller stops all I/O operations and dumps the contents of its DRAM to the internal disk drive. When input power to the uninterruptible power supply is restored, the SAN Volume Controllers restart and restore the original contents of the DRAM from the data saved on the disk drive.

A SAN Volume Controller is not fully operational until the uninterruptible power supply battery charge state indicates that it has sufficient capacity to power the SAN Volume Controller for long enough to permit it to save all its memory to the disk drive in the event of a power loss. The uninterruptible power supply has sufficient capacity to save all the data on the SAN Volume Controller at least twice. For a fully-charged uninterruptible power supply, even after battery capacity has been used to power the SAN Volume Controllers while they save DRAM data, sufficient battery capacity will remain to let the SAN Volume Controllers become fully operational as soon as input power is restored.

Note: Under normal circumstances, if input power is disconnected from the uninterruptible power supply, the SAN Volume Controller(s) connected to that uninterruptible power supply will perform a power down sequence. This operation, which saves the configuration and cache data to an internal disk in the SAN Volume Controller, typically takes about three minutes, at which time power is removed from the output of the uninterruptible power supply. In the event of a delay in the completion of the power down sequence, the uninterruptible power supply output power will be removed five minutes after the time that power was disconnected to the uninterruptible power supply. Since this operation is controlled by the SAN Volume Controller, an uninterruptible power supply that is not connected to an active SAN Volume Controller will not shut off within the five-minute required period. In the case of an emergency, you will need to manually shut down the uninterruptible power supply by pushing the uninterruptible power supply power off button.

Attention: Data integrity could be compromised by pushing the uninterruptible power supply power off button. Never shut down an uninterruptible power supply without first shutting down the SAN Volume Controller nodes that it supports.

It is very important that the two nodes in the I/O group are connected to different uninterruptible power supplies. This configuration ensures that cache and cluster state information is protected in the event of a failure of the uninterruptible power supply or mainline power source.

When nodes are added to the cluster, you must specify the I/O group they will join. The configuration interfaces will also check the uninterruptible power supply units and ensure that the two nodes in the I/O group are not connected to the same uninterruptible power supply units.

Figure 6 shows a cluster of four nodes, with two I/O groups and two uninterruptible power supply units.

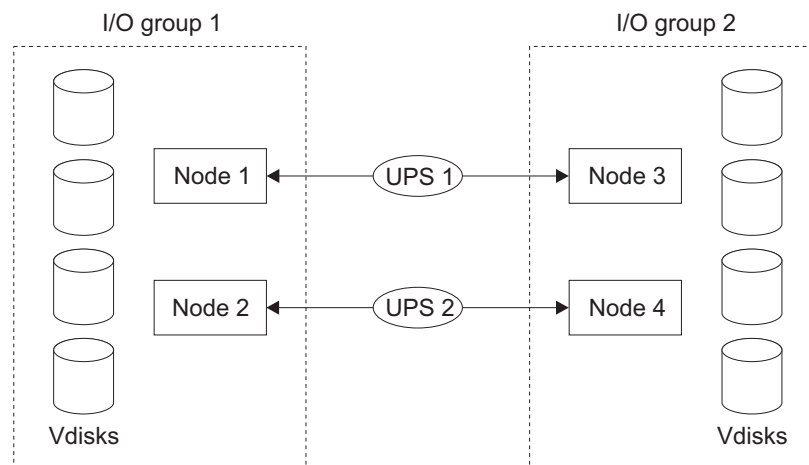


Figure 6. I/O groups and uninterruptible power supply relationship

Master console

The SAN Volume Controller provides a master console that can be used as a single platform to configure, manage, and service the SAN Volume Controller.

The master console allows system administrators to integrate rapidly the SAN Volume Controller into their environment. The master console monitors the configuration of the whole system and all of the internal components. It offers a standard and central location for all aspects of the operation, including SAN topology rendering, SNMP trap management, Call Home (Service Alert) and Remote Service facilities, as well as all the configuration and diagnostic utilities for the components.

Note: VPN connection is required for Remote Service facilities.

The master console provides the following functions:

- Browser support for:
 - SAN Volume Controller Console
 - Fibre-channel switch
- CLI configuration support using Secure Shell (SSH)
- SAN Topology rendering using Tivoli[®] SAN Manager

- Remote Service capability through VPN
- IBM Director
 - SNMP Trap management
 - Call Home (Service Alert) capability
 - E-mail notification to the customer, for example, to the system administrator

Master console components

These lists describe the hardware and installed software that are included with the master console.

- 19-inch 1U rack-mounted server
- 19-inch 1U flat panel monitor and keyboard

Attention: If more than one power distribution bus is available, the two power connectors, one supplying the master console and the other supplying the master console monitor, should be connected to the same power distribution bus.

The following software is included with and installed on the system:

- Microsoft® Windows® 2003 Standard Server Edition with the latest service pack
- Tivoli Storage Area Network Manager
- FASiT Storage Manager
- QLogic 2342 fibre-channel host bus adapter driver
- PuTTY, a client for Telnet and Secure Shell (SSH) protocol communications
 - Putty.exe, the client software
 - Puttygen.exe, a utility for generating encryption keys
 - Plink.exe, the command-line interface to the PuTTY client software
- IBM Director Server, a client/server workgroup manager
- SAN Volume Controller Console
- Adobe Acrobat Reader
- IBM Connection Manager virtual private network (VPN)

See the following Web site for the current list of supported software versions:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

You must configure the software that is provided on the master console to meet your requirements.

Overview of backup functions

The SAN Volume Controller includes functions that help you to back up cluster configuration settings and business data.

To enable routine maintenance of the SAN Volume Controller clusters, the configuration settings for each cluster are stored on each node. If power fails on a cluster or if a node in a cluster is replaced, the cluster configuration settings will be automatically restored when the repaired node is added back to the cluster. To restore the cluster configuration in the event of a disaster (if all nodes in a cluster are lost simultaneously), plan to back up the cluster configuration settings to tertiary storage. You use the configuration backup functions to back up the cluster configuration.

For complete disaster recovery, regularly back up the business data that is stored on virtual disk at the application server level or the host level. The SAN Volume Controller provides the following Copy Services functions that you can use to back up data: Remote Copy and FlashCopy.

Cluster configuration backup

Configuration backup is the process of extracting configuration data from a cluster and writing it to disk.

Backing up the cluster configuration enables you to restore it in the event that configuration data is lost. The data that is backed up is the metadata that describes the cluster configuration, not the data that your enterprise uses to run its business.

The backup configuration files can be saved on the master console or the configuration node.

Objects included in the backup

Configuration data is information about a cluster and the objects that are defined in it. The following objects are copied:

- Storage subsystem
- Hosts
- I/O groups
- Managed disks (MDisks)
- MDisk groups
- Nodes
- Virtual disks (VDisks)
- VDisk-to-host mappings
- SSH key
- FlashCopy mappings
- FlashCopy consistency groups
- Remote Copy relationships
- Remote Copy consistency groups

Related concepts

“Clusters” on page 38

All configuration and service is performed at the cluster level.

FlashCopy

FlashCopy is a copy service available with the SAN Volume Controller.

It copies the contents of a source virtual disk (VDisk) to a target VDisk. Any data that existed on the target disk is lost and is replaced by the copied data. After the copy operation has been completed, the target virtual disks contain the contents of the source virtual disks as they existed at a single point in time unless target writes have been performed. Although the copy operation takes some time to complete, the resulting data on the target is presented in such a way that the copy appears to have occurred immediately. FlashCopy is sometimes described as an instance of a time-zero copy (T 0) or point-in-time copy technology. Although the FlashCopy operation takes some time, this time is several orders of magnitude less than the time which would be required to copy the data using conventional techniques.

It is difficult to make a consistent copy of a data set that is being constantly updated. Point-in-time copy techniques are used to help solve the problem. If a copy of a data set is taken using a technology that does not provide point in time techniques and the data set changes during the copy operation, then the resulting copy may contain data which is not consistent. For example, if a reference to an object is copied earlier than the object itself and the object is moved before it is itself copied then the copy will contain the referenced object at its new location but the reference will point to the old location.

Source VDisks and target VDisks must meet the following requirements:

- They must be the same size.
- The same cluster must manage them.

Related concepts

“FlashCopy consistency groups” on page 48

A consistency group is a container for mappings. You can add many mappings to a consistency group.

“FlashCopy mappings” on page 45

A FlashCopy mapping defines the relationship between a source VDisk and a target VDisk.

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Remote Copy

Remote Copy enables you to set up a relationship between two virtual disks, so that updates that are made by an application to one virtual disk are mirrored on the other virtual disk.

Although the application only writes to a single virtual disk, the SAN Volume Controller maintains two copies of the data. If the copies are separated by a significant distance, then the remote copy can be used as a backup for disaster recovery. A prerequisite for the SAN Volume Controller Remote Copy operations between two clusters is that the SAN fabric to which they are attached provides adequate bandwidth between the clusters.

One VDisk is designated the primary and the other VDisk is designated the secondary. Host applications write data to the primary VDisk, and updates to the primary VDisk are copied to the secondary VDisk. Normally, host applications do not perform input or output operations to the secondary VDisk. When a host writes to the primary VDisk, it will not receive confirmation of I/O completion until the write operation has completed for the copy on the secondary disk as well as on the primary.

Remote Copy supports the following features:

- Intracluster copying of a VDisk, in which both VDisks belong to the same cluster and I/O group within the cluster.
- Intercluster copying of a VDisk, in which one VDisk belongs to a cluster and the other VDisk belongs to a different cluster

Note: A cluster can only participate in active Remote Copy relationships with itself and one other cluster.

- Intercluster and intracluster Remote Copy can be used concurrently within a cluster.

- The intercluster link is bidirectional. Meaning, it can support copying of data from clusterA to clusterB for one pair of VDisks while copying data from clusterB to clusterA for a different pair of VDisks.
- The copy direction can be reversed for a consistent relationship by issuing a simple **switch** command. See *IBM TotalStorage SAN Volume Controller: Command-Line Interface User's Guide*.
- Remote Copy consistency groups are supported for ease of managing a group of relationships that need to be kept in sync for the same application. This also simplifies administration, as a single command issued to the consistency group will be applied to all the relationships in that group.

Related concepts

“Synchronous Remote Copy” on page 50

In the synchronous mode, Remote Copy provides a *consistent copy*, which means that the primary VDisk is always the exact match of the secondary VDisk.

“Remote Copy consistency groups” on page 50

Remote Copy provides the facility to group a number of relationships into a Remote Copy consistency group so that they can be manipulated in unison.

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Chapter 2. Installation planning

Before the service representative can start to set up your SAN Volume Controller, verify that the prerequisite conditions for the SAN Volume Controller and uninterruptible power supply installation are met.

1. Does your physical site meet the environment requirements for the SAN Volume Controller, master console, and uninterruptible power supply?
2. Do you have adequate rack space for your hardware?
 - a. SAN Volume Controller: One Electrical Industries Association (EIA) unit high for each node.
 - b. Uninterruptible power supply: Two EIA units high for each uninterruptible power supply.
 - c. Master console: Two EIA units high.
3. Do you have power distribution units in the rack to provide power to the uninterruptible power supply units?

A clearly visible and accessible emergency power off switch is required.
4. Ensure that you provide appropriate connectivity.

Related reference

“Master console” on page 11

The SAN Volume Controller provides a master console that can be used as a single platform to configure, manage, and service the SAN Volume Controller.

Preparing your SAN Volume Controller environment

Before installing the SAN Volume Controller, prepare the physical environment.

Dimensions and weight

Height	Width	Depth	Maximum Weight
43 mm (1.7 in.)	440 mm (17.3 in.)	660 mm (26 in.)	12.7 kg (28 lb.)

Additional space requirements

Location	Additional Space Required	Reason
Left and right sides	50 mm (2 in.)	Cooling air flow
Back	minimum: 100 mm (4 in.)	Cable exit

AC input-voltage requirements

Table 5.

Power Supply Assembly Type	Voltage	Frequency
200 to 240V	88 to 264 V ac	50 to 60 Hz

Environment

Environment	Temperature	Altitude	Relative Humidity	Maximum Wet Bulb Temperature
Operating in Lower Altitudes	10°C to 35°C (50°F to 95°F)	0 to 914 m (0 to 2998 ft.)	8% to 80% noncondensing	23°C (74°F)
Operating in Higher Altitudes	10°C to 32°C (50°F to 88°F)	914 to 2133 m (2998 to 6988 ft.)	8% to 80% noncondensing	23°C (74°F)
Powered Off	10°C to 43°C (50°F to 110°F)	–	8% to 80% noncondensing	27°C (81°F)
Storing	1°C to 60°C (34°F to 140°F)	0 to 2133 m (0 to 6988 ft.)	5% to 80% noncondensing	29°C (84°F)
Shipping	-20°C to 60°C (-4°F to 140°F)	0 to 10668 m (0 to 34991 ft.)	5% to 100% condensing, but no precipitation	29°C (84°F)

Heat output (maximum)

350 watts (1195 Btu per hour)

Preparing your uninterruptible power supply environment

Ensure that your physical site meets the installation requirements for the uninterruptible power supply.

Your uninterruptible power supply should be configured with the following considerations:

- Each uninterruptible power supply should be connected to a separate branch circuit.
- A UL listed 15 A circuit breaker must be installed in each branch circuit that supplies power to the uninterruptible power supply.
- The voltage supplied to the uninterruptible power supply must be 200–240 V single phase.
- The frequency supplied must be between 50 and 60 Hz.

Attention: Ensure that you comply with the following requirements for uninterruptible power supplies.

Note: If the uninterruptible power supply is cascaded from another uninterruptible power supply, the source uninterruptible power supply must have at least three times the capacity per phase and the total harmonic distortion must be less than 5% with any single harmonic being less than 1%. The uninterruptible power supply must also have input voltage capture that has a slew rate faster than 3 Hz per second and 1 msec glitch rejection.

Dimensions and weight

Height	Width	Depth	Maximum weight
89 mm (3.5 in.)	483 mm (19 in.)	622 mm (24.5 in.)	37 kg (84 lb.)

AC input-voltage requirements

Power Supply Assembly Type	Voltage	Frequency
200 to 240 V	160 to 288 V ac	50 to 60 Hz

Environment

	Operating Environment	Non-operating Environment	Storing Environment	Shipping Environment
Air Temperature	0°C to 40°C (32°F to 104°F)	0°C to 40°C (32°F to 104°F)	0°C to 25°C (32°F to 77°F)	-25°C to 55°C (-13°F to 131°F)
Relative Humidity	5% to 95% non-condensing	5% to 95% non-condensing	5% to 95% non-condensing	5% to 95% non-condensing

Altitude

	Operating Environment	Non-operating Environment	Storing Environment	Shipping Environment
Altitude (from sea level)	0 to 2000 m (0 to 6560 ft.)	0 to 2000 m (0 to 6560 ft.)	0 to 2000 m (0 to 6560 ft.)	0 to 15 000 m (0 to 49212 ft.)

Heat output (maximum)

142 watts (485 Btu per hour) during normal operation.

553 watts (1887 Btu per hour) when power has failed and the uninterruptible power supply is supplying power to the nodes of the SAN Volume Controller.

Preparing your master console environment

Ensure that your physical site meets the installation requirements for the master console server and console monitor kit.

Server dimensions and weight

Height	Width	Depth	Maximum Weight
43 mm (1.7 in.)	430 mm (16.69 in.)	424 mm (16.69 in.)	12.7 kg (28 lb.)

Note: The above dimensions are for a 1U monitor and keyboard assembly.

Server AC and input-voltage requirements

Power Supply	Electrical Input
203 watt (110 or 220 V ac auto-sensing)	Sine-wave input (47–63 Hz) required Input voltage low range: Minimum: 100 V ac Maximum: 127 V ac Input voltage high range: Minimum: 200 V ac Maximum: 240 V ac Input kilovolt-amperes (kVA), approximately: Minimum: 0.0870 kVA Maximum: 0.150 kVA

Server environment

Environment	Temperature	Altitude	Relative Humidity
Server On	10° to 35°C (50°F to 95°F)	0 to 914 m (2998.0 ft.)	8% to 80%
Server Off	-40°C to 60°C (-104°F to 140°F)	Maximum: 2133 m (6998.0 ft.)	8% to 80%

Server heat output

Approximate heat output in British thermal units (BTU) per hour:

- Minimum configuration: 87 watts (297 BTU)
- Maximum configuration: 150 watts (512 BTU)

Monitor console kit dimensions and weight

Height	Width	Depth	Maximum Weight
43 mm (1.7 in.)	483 mm (19.0 in.)	483 mm (19.0 in.)	17.0 kg (37.0 lb.)

Ports and connections

Each SAN Volume Controller requires the following ports and connections:

- Each SAN Volume Controller node requires one Ethernet cable to connect it to an Ethernet switch or hub. A 10/100 Mb Ethernet connection is required.
- Two TCP/IP addresses are normally required for a SAN Volume Controller cluster, a cluster address and a service address.
- Each SAN Volume Controller node has four fibre-channel ports, which are supplied fitted with LC-style optical Small Form-factor Pluggable (SFP) GBICs for connection to a fibre-channel switch.

Each uninterruptible power supply requires the following:

- Serial cables that connect the uninterruptible power supply to the SAN Volume Controller nodes. Ensure that for each node, the serial and power cables come from the same uninterruptible power supply.

The master console requires the following connections:

- Two Ethernet cables:
 - One from the master console, Ethernet port 1, to DMZ or firewall pass-through. This will be used for a VPN connection for remote support.
 - One from the master console, Ethernet port 2, to Ethernet switch or hub.

An IP address must be set up for each Ethernet port. The connections must be 10/100 Mb Ethernet connections.

- The master console has two FC ports for connection to fibre-channel switches.

Chapter 3. Preparing the physical configuration

Before the service representative installs the SAN Volume Controller, uninterruptible power supply unit, and master console, you must plan the physical configuration and the initial settings for the system.

To plan the configuration, print or photocopy the blank charts and tables in this publication and use a pencil or pen to plan the system configuration. Before you begin writing in the charts and tables, make copies of the blank charts and tables so that you can later revise the configuration or create a new one if needed.

1. Use the hardware location chart to record the physical configuration of your system.
2. Use the cable connection table to record how your SAN Volume Controller, the uninterruptible power supply unit, and the master console are to be connected.
3. Use the configuration data table to record the data that you and the service representative need before the initial installation.

When you have completed these tasks you are ready to perform the physical installation.

Completing the hardware location chart

The hardware location chart represents the rack into which the SAN Volume Controller is to be installed. Each row of the chart represents one Electrical Industries Association (EIA) 19-inch rack space.

- Uninterruptible power supply units are heavy and should always be installed as near the bottom of the rack as possible. IBM recommends that you place them within the range of row 1 through row 8.
- The maximum power rating of the rack and input power supply must not be exceeded.
- The SAN Volume Controller should be positioned such that information on the display screen can be easily viewed and the controls used to navigate the display menu can be easily reached. The recommended range is EIA 11-38.
- To enable easy access to the connectors on the rear of the master console, the console, keyboard and monitor unit should be positioned adjacent to each other. To permit easy access to the CD drive, the master console should be located above the keyboard and monitor unit. The recommended range is EIA 17-24.
- A SAN Volume Controller is one EIA unit high. Therefore, for each SAN Volume Controller that is to be installed, fill in the row that represents the position that the SAN Volume Controller is to occupy.
- An uninterruptible power supply is two EIA units high. Therefore, for each uninterruptible power supply, fill in two rows.
- The master console is two EIA units high: one EIA unit for the server and one EIA unit for the keyboard and monitor.
- If there are any hardware devices already contained in the rack, record this information on the chart.
- Fill in rows for all other units that will be present in the rack, including Ethernet hubs and fibre-channel switches. Hubs and switches are usually one EIA unit high, but check with your supplier. The uninterruptible power supply units must be installed at the bottom of the rack so it might be necessary to relocate some other devices before the SAN Volume Controller installation is started.

Hardware location guidelines

When you fill in the hardware location chart, follow these basic guidelines.

- The SAN Volume Controller must be installed in pairs to provide redundancy and concurrent maintenance.
- A cluster can contain no more than four SAN Volume Controllers.
- Each SAN Volume Controller of a pair must be connected to a different uninterruptible power supply.
- Each uninterruptible power supply pair can support one SAN Volume Controller cluster.
- To reduce the chance of a simultaneous input power failure at both uninterruptible power supply units, each uninterruptible power supply should be connected to a separate electrical power source on a separate branch circuit.
- Because uninterruptible power supply units are heavy, they must be installed into the lowest available positions of the rack. If necessary, move any lighter units that are already in the rack to higher positions.
- IBM does not install the Ethernet hub or the fibre-channel switches. You must arrange for either the suppliers or someone in your organization to install those items. Provide the installer with a copy of the completed hardware location chart.

In the following example, assume that the rack is empty and you want to create a system that contains the following components:

- Four SAN Volume Controllers named SVC1, SVC2, SVC3 and SVC4.
- One master console.
- Two uninterruptible power supply units named uninterruptible power supply 1 and uninterruptible power supply 2.
- One Ethernet hub named Ethernet hub 1. For this example, it is assumed that the hub is one EIA unit high.
- Two fibre-channel switches named FC switch 1 and FC switch 2. In this example, each switch is one EIA unit high.
- RAID controllers named RAID controller 1, RAID controller 2, RAID controller 3, RAID controller 4.

Your completed chart might look like Table 6:

Table 6. Sample of completed hardware location chart

Rack row	Component
EIA 36	Blank
EIA 35	Ethernet Hub 1
EIA 34	Blank
EIA 33	Blank
EIA 32	Blank
EIA 31	Blank
EIA 30	Blank
EIA 29	Blank
EIA 28	FC Switch 1
EIA 27	FC Switch 2
EIA 26	Blank

Table 6. Sample of completed hardware location chart (continued)

Rack row	Component
EIA 25	Blank
EIA 24	Blank
EIA 23	Blank
EIA 22	SAN Volume Controller 4
EIA 21	SAN Volume Controller 3
EIA 20	SAN Volume Controller 2
EIA 19	SAN Volume Controller 1
EIA 18	Master console
EIA 17	Master console keyboard and monitor
EIA 16	RAID Controller 4
EIA 15	
EIA 14	
EIA 13	RAID Controller 3
EIA 12	
EIA 11	
EIA 10	RAID Controller 2
EIA 9	
EIA 8	
EIA 7	RAID Controller 1
EIA 6	
EIA 5	
EIA 4	uninterruptible power supply 2
EIA 3	
EIA 2	uninterruptible power supply 1
EIA 1	

You might want to put the switches between the SAN Volume Controller nodes. Remember, however, that the uninterruptible power supply units must be in the lowest positions of the rack.

Hardware location chart

The hardware location chart helps you to plan the location of the hardware.

Each row of the chart in Table 7 represents one EIA unit.

Table 7. Hardware location chart

Rack row	Component
EIA 36	
EIA 35	
EIA 34	
EIA 33	

Table 7. Hardware location chart (continued)

Rack row	Component
EIA 32	
EIA 31	
EIA 30	
EIA 29	
EIA 28	
EIA 27	
EIA 26	
EIA 25	
EIA 24	
EIA 23	
EIA 22	
EIA 21	
EIA 20	
EIA 19	
EIA 18	
EIA 17	
EIA 16	
EIA 15	
EIA 14	
EIA 13	
EIA 12	
EIA 11	
EIA 10	
EIA 9	
EIA 8	
EIA 7	
EIA 6	
EIA 5	
EIA 4	
EIA 3	
EIA 2	
EIA 1	

Completing the cable connection table

The cable connection table helps you to plan how to connect the units that will be placed in the rack.

- Node number. The nominal number (name) of the SAN Volume Controller.
- Uninterruptible power supply. The uninterruptible power supply to which the SAN Volume Controller is connected.

- Ethernet. The Ethernet hub or switch to which the SAN Volume Controller is connected.
- FC Ports 1 through 4. The fibre-channel switch ports to which the four SAN Volume Controller fibre-channel ports are connected. When viewed from the back of the SAN Volume Controller, the ports are numbered 1 through 4, from left to right. Ignore the markings on the back of the SAN Volume Controller.

For the master console, complete the cable connection table as follows:

- Ethernet Port 1. Ethernet port 1 is used for your VPN connection. This port is required if you configure your master console to enable remote support. A remote support connection can only be enabled when this port has access to an external internet connection. For added security, you can disconnect this port when a remote support connection is not being used.
- Ethernet Port 2. Ethernet port 2 is used to connect the SAN Volume Controller to the network.
- FC Ports 1 and 2. FC Ports 1 and 2 are the fibre-channel switch ports to which the master console fibre-channel ports are connected. Connect one FC port to each of the SAN Volume Controller fabrics.

Cable connection table

Complete the cable connection table to plan the connections of the units in the rack.

Table 8. Cable connection table

SAN Volume Controller	Uninterruptible power supply	Ethernet hub or switch	FC port-1	FC port-2	FC port-3	FC port-4

Master console	Ethernet		FC port-1	FC port-2
	Public network	VPN		

Example of a completed cable connection table:

For this example, assume you are completing the cabling details for the system. Remember that SAN Volume Controllers are configured in pairs, and that the two SAN Volume Controllers of a pair must *not* be connected to the same uninterruptible power supply. Also, the two uninterruptible power supply units of a pair should not be connected to the same power source to reduce the chance of input power failure at both uninterruptible power supply units. For this example, assume that the pairs of SAN Volume Controllers are: node 1 and node 2, and node 3 and node 4, and that the two power sources provided by the uninterruptible power supply units are A and B.

Note: The uninterruptible power supply requires two dedicated branch circuits that meet the following specifications:

- 15 amp circuit breaker in each branch circuit that supplies the power to an uninterruptible power supply
- Single-phase
- 50 – 60 Hz
- 220 volt

For the Ethernet connection, you must use Ethernet port 1 of the SAN Volume Controller. Do not use any other Ethernet port, because the software is configured for Ethernet port 1 only.

Note: All SAN Volume Controller nodes that are part of the same cluster must be connected to the same Ethernet subnet, otherwise TCP/IP address failover will not work.

Table 9 illustrates this example.

Table 9. Example of cable connection table

SAN Volume Controller	Uninterruptible power supply	Ethernet hub or switch	FC Port-1	FC Port-2	FC Port-3	FC Port-4
Node 1	Uninterruptible power supply A	Hub or switch 1, Port 1	FC switch 1, Port 1	FC switch 2, Port 1	FC switch 1, Port 2	FC switch 2, Port 2
Node 2	Uninterruptible power supply B	Hub or switch 1, Port 2	FC switch 1, Port 3	FC switch 2, Port 3	FC switch 1, Port 4	FC switch 2, Port 4
Node 3	Uninterruptible power supply A	Hub or switch 1, Port 3	FC switch 1, Port 5	FC switch 2, Port 5	FC switch 1, Port 6	FC switch 2, Port 6
Node 4	Uninterruptible power supply B	Hub or switch 1, Port 4	FC switch 1, Port 7	FC switch 2, Port 7	FC switch 1, Port 8	FC switch 2, Port 8

Master console	Ethernet		FC switch 1, Port-9	FC switch 2, Port-9
	Public network	VPN		
Master console	Ethernet hub 1, Port 5	Ethernet hub 1, Port 6	FC Port-1 FC switch 1, Port 9	FC Port-2 FC switch 2, Port 9

Completing the configuration data table

The configuration data table helps you to plan the initial settings for the cluster configuration.

Include the following initial settings for the cluster:

- Language. The national language in which you want the messages displayed on the front panel. This option applies only to service messages. The default is English.
- Cluster IP address. The address that will be used for all typical configuration and service access to the cluster.
- Service IP address. The address that will be used for emergency access to the cluster.
- Gateway IP address. The IP address for the default local gateway for the cluster.
- Subnet mask. The subnet mask of the cluster.

- Fibre-channel switch speed. The fibre-channel switch speed can be either 1 Gb or 2 Gb.

Include the following information for the master console:

- Machine name. The name you want the master console to be known as. This must be a fully qualified DNS name. The default setting is *mannode* (not fully qualified).
- Master console IP addresses. The addresses that will be used for access to the master console. The default settings are:
 - Port 1 = 192.168.1.11
 - Port 2 = 192.168.1.2
- Master console gateway IP address. The IP address for the local gateway for the master console. The default setting is: 192.168.1.1.
- Master console subnet mask. The default subnet mask for the master console is 255.255.255.0.

Configuration data table

Use the configuration data table to plan the initial settings for the cluster configuration.

Cluster		
Language		
Cluster IP address		
Service IP address		
Gateway IP address		
Subnet mask		
Fibre-channel switch speed		
Master console		
Machine name		
	Ethernet Port 1	Ethernet Port 2
Master console IP address		
Master console gateway IP address		
Master console subnet mask		

Chapter 4. Planning guidelines for using your SAN Volume Controller in a SAN environment

Follow these planning steps to set up your SAN Volume Controller environment.

1. Plan your configuration.
2. Plan your SAN environment.
3. Plan your fabric setup.
4. Create the RAID resources that you intend to virtualize.
5. Determine if you have a RAID array that contains data that you want to merge into the cluster.
6. Determine if you will migrate data into the cluster or keep them as image-mode VDisks.
7. Determine if will use Copy Services. These services are provided for all supported hosts that are connected to the SAN Volume Controller that enables you to copy VDisks.

Related concepts

“Image mode virtual disk migration” on page 44

Image mode virtual disks (VDisks) have the special property that the last extent in the VDisk can be a partial extent.

Storage Area Network

A storage area network (SAN) is a high-speed dedicated network for sharing storage resources.

A SAN allows the establishment of direct connections between storage devices and servers. It offers simplified storage management, scalability, flexibility, availability, and improved data access, movement, and backup.

A SAN storage system consists of two to eight SAN Volume Controller nodes that are arranged in a cluster. These will appear as part of the SAN fabric, along with the host systems, the RAID controllers, and the storage devices, all connected together to create the SAN. Other devices such as fabric switches may be required to complete the SAN.

It is important to note these two types of SAN: redundant and counterpart. A *redundant* SAN consists of a fault tolerant arrangement of two counterpart SANs. A redundant SAN configuration provides two independent paths for each device attached to the SAN. A *counterpart* SAN is a non-redundant portion of a redundant SAN and provides all the connectivity of the redundant SAN, but without the redundancy. Each counterpart SAN provides an alternate path for each device attached to the SAN.

Note: IBM highly recommends that a redundant SAN be used with the SAN Volume Controller, however a non-redundant SAN is supported.

To install a SAN Volume Controller into an existing SAN that will be in use during installation, you must first ensure that the switch zoning is set to isolate the new SAN Volume Controller connections from the active part of the SAN.

See the following Web site for specific firmware levels and the latest supported hardware:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

- Consider the design of the SAN according to your requirement for high availability.
- Identify the operating system for each host system that will be connected to the SAN Volume Controller, ensuring compatibility and suitability.
 1. Specify the host bus adapters (HBAs) for each host
 2. Define the performance requirements
 3. Determine the total storage capacity
 4. Determine the storage capacity per host
 5. Determine the host LUN sizes
 6. Determine the total number of ports and the bandwidth needed between the host to the SAN Volume Controller
 7. Determine if your SAN has enough ports to connect all hosts and backend storage
 8. Determine if your SAN provides enough ports to connect your backend storage
- Ensure that the existing SAN components meet the requirements for the SAN Volume Controller:
 1. Determine the host system versions
 2. Ensure that the HBAs, switches, and controllers are at or above the minimum requirements
 3. Identify any components that must be upgraded

Switch zoning for the SAN Volume Controller

Consider these constraints when zoning a switch.

Overview

The number of virtual paths to each virtual disk is limited. Implementation of the following rules will help you achieve the correct number of virtual paths.

- Each host (or partition of a host) can have between one and four fibre-channel ports.
- Switch zoning should be used to ensure that each host fibre-channel port is zoned to exactly one fibre-channel port for each SAN Volume Controller node in a cluster.
- To obtain the best performance from a host with multiple fibre-channel ports, the zoning should ensure that each fibre-channel port of a host is zoned with a different group of SAN Volume Controller ports.
- To obtain the best overall performance of the subsystem, the workload for each SAN Volume Controller port should be equal. This will typically involve zoning roughly the same number of host fibre channel ports to each SAN Volume Controller fibre-channel port.

IBM recommends that you manually set the domain IDs prior to building the multiswitch fabric and prior to zoning for the following reasons:

- When two switches are joined while active, they will determine if the domain ID is already in use as before, but if there is a conflict it cannot be changed in an active switch. This conflict will cause the fabric merging process to fail.
- The domain ID is used to identify switch ports when zoning is implemented using the domain and switch port number. If domain IDs are negotiated at every fabric

start up, there is no guarantee that the same switch will have the same ID the next time. Therefore, zoning definitions can become invalid.

- If the domain ID is changed after a SAN is set up, some host systems may have difficulty logging back in with the switch, and it may be necessary to reconfigure the host in order to detect devices on the switch again.

The maximum number of paths from the SAN Volume Controller nodes to a host is eight. The maximum number of host bus adapter (HBA) ports is four (for example, no more than two two-port HBAs or four one-port HBAs).

In the following example, consider the following SAN environment:

- Two SAN Volume Controller nodes, nodes A and B
- Nodes A and B have four ports each
 1. Node A has ports A0, A1, A2, and A3
 2. Node B has ports B0, B1, B2, and B3
- Four hosts called P, Q, R, and S
- Each of the four hosts has four ports, as described in Table 10.

Table 10. Four hosts and their ports

P	Q	R	S
P0	Q0	R0	S0
P1	Q1	R1	S1
P2	Q2	R2	S2
P3	Q3	R3	S3

- Two switches called X and Y
- One storage controller
- The storage controller has four ports on it called I0, I1, I2, and I3

An example configuration would be the following:

1. Attach ports 1 (A0, B0, P0, Q0, R0, and S0) and 2 (A1, B1, P1, Q1, R1, and S1) of each node and host to switch X.
2. Attach ports 3 (A2, B2, P2, Q2, R2, and S2) and 4 (A3, B3, P3, Q3, R3, and S3) of each node and host to switch Y.
3. Attach ports 1 and 2 (I0 and I1) of the storage controller to switch X.
4. Attach ports 3 and 4 (I2 and I3) of the storage controller to switch Y.

On switch X we would create the following host zones:

5. Create a host zone containing ports 1 (A0, B0, P0, Q0, R0, and S0) of each node and host.
6. Create a host zone containing ports 2 (A1, B1, P1, Q1, R1, and S1) of each node and host.

Similarly, on switch Y we would create the following host zones:

7. Create a host zone on switch Y containing ports 3 (A2, B2, P2, Q2, R2, and S2) of each node and host.
8. Create a host zone on switch Y containing ports 4 (A3, B3, P3, Q3, R3, and S3) of each node and host.

Last, we would create the following storage zone:

9. Create a storage zone that is configured on each switch. Each storage zone contains all the SAN Volume Controller and storage ports on that switch.

In the following example, the SAN environment is similar to the first example, with two additional hosts with two ports each.

- Two SAN Volume Controller nodes called A and B
- Nodes A and B have four ports each
 1. Node A has ports A0, A1, A2, and A3
 2. Node B has ports B0, B1, B2, and B3
- Six hosts called P, Q, R, S, T and U
- Four hosts have four ports each, and two hosts have two ports each as described in Table 11.

Table 11. Six hosts and their ports

P	Q	R	S	T	U
P0	Q0	R0	S0	T0	U0
P1	Q1	R1	S1	T1	U1
P2	Q2	R2	S2	—	—
P3	Q3	R3	S3	—	—

- Two switches called X and Y
- One storage controller
- The storage controller has four ports on it called I0, I1, I2, and I3

An example configuration would be the following:

1. Attach ports 1 (A0, B0, P0, Q0, R0, S0 and T0) and 2 (A1, B1, P1, Q1, R1, S1 and T1) of each node and host to switch X.
2. Attach ports 3 (A2, B2, P2, Q2, R2, S2 and T2) and 4 (A3, B3, P3, Q3, R3, S3 and T3) of each node and host to switch Y.
3. Attach ports 1 and 2 (I0 and I1) of the storage controller to switch X.
4. Attach ports 3 and 4 (I2 and I3) of the storage controller to switch Y.

Attention: Hosts T and U (T0 and U0) and (T1 and U1) are zoned to different SAN Volume Controller ports so that each SAN Volume Controller port is zoned to the same number of host ports.

On switch X we would create the following host zones:

5. Create a host zone containing ports 1 (A0, B0, P0, Q0, R0, S0 and T0) of each node and host.
6. Create a host zone containing ports 2 (A1, B1, P1, Q1, R1, S1 and U0) of each node and host.

Similarly, on switch Y we would create the following host zones:

7. Create a host zone on switch Y containing ports 3 (A2, B2, P2, Q2, R2, S2 and T1) of each node and host.
8. Create a host zone on switch Y containing ports 4 (A3, B3, P3, Q3, R3, S3 and U1) of each node and host.

Last, we would create the following storage zone:

9. Create a storage zone configured on each switch. Each storage zone contains all the SAN Volume Controller and storage ports on that switch.

Related reference

“Fibre-channel switches” on page 77

Follow these guidelines for configuring the fibre-channel switches that are supported on the SAN.

Zoning considerations for Remote Copy

Consider these constraints when zoning a switch to support the Remote Copy service.

SAN configurations that use the Remote Copy feature between two clusters need additional switch zoning considerations. These considerations include:

- Additional zones for remote copy. For Remote Copy operations involving two clusters, these clusters must be zoned so that the nodes in each cluster can see the ports of the nodes in the other cluster.
- Use of extended fabric settings in a switched fabric.
- Use of Inter Switch Link (ISL) trunking in a switched fabric.
- Use of redundant fabrics.

Note: These considerations do not apply if the simpler, intracluster mode of Remote Copy operation is in use, when only a single cluster is needed.

For intracluster Remote Copy relationships, no additional switch zones are required. For intercluster Remote Copy relationships, you must:

1. Form a SAN that contains both clusters that are to be used in the Remote Copy relationships. If cluster A is in SAN A originally, and cluster B is in SAN B originally, this means that there must be at least one fibre-channel connection between SAN A and SAN B. This connection will be one or more inter-switch links. The fibre-channel switch ports associated with these inter-switch ports should not appear in any zone.
2. A single SAN can only be formed out of combining SAN A and SAN B if the domain numbers of the switches in each SAN are different, prior to the connection of the two SANs. You should ensure that each switch has a different domain ID before connecting the two SANs.
3. Once the switches in SAN A and SAN B are connected, they should be configured to operate as a single group of switches. Each cluster should retain the same set of zones that were required to operate in the original single SAN configuration.
4. A new zone must be added that contains all the switch ports that are connected to SAN Volume Controller ports. This will contain switch ports that were originally in SAN A and in SAN B.
5. You can adjust the switch zoning so that the hosts originally in SAN A can see cluster B. This allows a host to examine data in both the local and remote cluster if required. This view of both clusters is purely optional and in some cases may complicate the way you operate the overall system, therefore, unless specifically needed, it should not be implemented.
6. You should verify that the switch zoning is such that cluster A cannot see any of the back-end storage owned by cluster B. Two clusters may not share the same back-end storage devices.

The following zones would therefore be needed in a typical intercluster Remote Copy configuration:

1. A zone in the local cluster that contains all the ports in the SAN Volume Controller nodes in that local cluster and the ports on the backend storage associated with that local cluster. These zones would be required whether or not Remote Copy is in use.

2. A zone in the remote cluster that contains all the ports in the SAN Volume Controller nodes in that remote cluster and the ports on the back-end storage associated with that remote cluster. These zones would be required whether or not Remote Copy is in use.
3. A zone that contains all the ports in the SAN Volume Controller nodes in both the local and remote cluster. This zone is required for intercluster communication and is specifically required by Remote Copy.
4. Additional zones that contain ports in host HBAs and selected ports on the SAN Volume Controller nodes in a particular cluster. These are the zones that allow a host to see VDisks presented by an I/O group in a particular cluster. These zones would be required whether or not Remote Copy were in use.

Note:

1. While it is normal to zone a server connection so that it is only visible to the local or remote cluster, it is also possible to zone the server so that the host HBA can see nodes in both the local and remote cluster at the same time.
2. Intracluster Remote Copy operation does not require any additional zones, over and above those needed to run the cluster itself.

Switch operations over long distances

Some SAN switch products provide features that allow the users to tune the performance of I/O traffic in the fabric in a way that can affect Remote Copy performance.

The two most significant features are ISL trunking and extended fabric.

ISL trunking	<p>Trunking enables the switch to use two links in parallel and still maintain frame ordering. It does this by routing all traffic for a given destination over the same route even when there may be more than one route available. Often trunking is limited to certain ports or port groups within a switch. For example, in the IBM 2109-F16 switch, trunking can only be enabled between ports in the same quad (for example, same group of four ports). For more information on trunking with the MDS, refer to "Configuring Trunking" on the Cisco Systems Web site.</p> <p>Some switch types may impose limitations on concurrent use of trunking and extended fabric operation. For example, with the IBM 2109-F16 switch, it is not possible to enable extended fabric for two ports in the same quad. Thus, extended fabric and trunking are effectively mutually exclusive. (Although it is possible, to enable extended fabric operation one link of a trunked pair this does not offer any performance advantages and adds complexity to the configuration setup. This mixed mode of operation is therefore not recommended.)</p>
---------------------	---

Extended fabric	<p>Extended fabric operation allocates extra buffer credits to a port. This is important over long links usually found in inter-cluster remote copy operation because, due to the time it takes for a frame to traverse the link, it is possible to have more frames in transmission at any instant in time than would be possible over a short link. The additional buffering is required to allow for the extra frames.</p> <p>For example, the default license for the IBM 2109-F16 switch has two extended fabric options, Normal and Extended Normal.</p> <ul style="list-style-type: none"> • Normal is suitable for short links and Extended Normal is suitable for links up to 10km long. (With the additional Extended fabric license the user gets two extra options, Medium, up to 10-50km and Long, 50-100km.) • The Extended Normal setting gives significantly better performance for the links up to 10 km long. Medium and Long settings are not recommended for use in the inter-cluster remote copy links currently supported.
------------------------	--

Performance of fibre-channel extenders

When planning to use fibre-channel extenders, it is important to be aware that the performance of the link to the remote location decreases as the distance to the remote location increases.

For fibre-channel IP extenders, throughput is limited by latency and bit error rates. Typical I/O latency can be expected to be 10 microseconds per kilometer. Bit error rates will vary depending on the quality of the circuit provided.

You should review the total throughput rates that might be expected for your planned configuration with the vendor of your fibre-channel extender and your network provider.

Related reference

“Supported fibre-channel extenders” on page 84

The SAN Volume Controller supports the CNT UltraNet Edge Storage Router to support synchronous copy services.

Nodes

A SAN Volume Controller node is a single processing unit within a SAN Volume Controller cluster.

For redundancy, nodes are deployed in pairs to make up a cluster. A cluster can have one to four pairs of nodes in it. Each pair of nodes is known as an I/O group. Each node can be in *only* one I/O group. A maximum of four I/O groups each containing two nodes is supported.

At any one time, a single node in the cluster is used to manage configuration activity. This configuration node manages a cache of the configuration information that describes the cluster configuration and provides a focal point for configuration commands. If the configuration node fails, another node in the cluster will take over its responsibilities.

Table 12 describes the operational states of a node.

Table 12. Node state

State	Description
Adding	The node was added to the cluster but is not yet synchronized with the cluster state (see Note).
Deleting	The node is in the process of being deleted from the cluster.
Online	The node is operational, assigned to a cluster, and has access to the fibre-channel SAN fabric.
Offline	The node is not operational. The node was assigned to a cluster but is not available on the fibre-channel SAN fabric. Run the Directed Maintenance Procedures to determine the problem.
Pending	The node is transitioning between states and, in a few seconds, will move to one of the other states.
Note: It is possible that a node can stay in the Adding state for a long time. If this is the case, delete the node and then re-add it. However, you should wait for at least 30 minutes before doing this. If the node that has been added is at a lower code level than the rest of the cluster, the node will be upgraded to the cluster code level, which can take up to 20 minutes. During this time the node will be shown as adding.	

Clusters

All configuration and service is performed at the cluster level.

A cluster can consist of two nodes, with a maximum of eight nodes. Therefore, you can assign up to eight SAN Volume Controller nodes to one cluster.

Some service actions can be performed at node level, but all configuration is replicated across all nodes in the cluster. Because configuration is performed at the cluster level, an IP address is assigned to the cluster instead of each node.

All your configuration and service actions are performed at the cluster level. Therefore, after configuring your cluster, you can take advantage of the virtualization and the advanced features of the SAN Volume Controller.

Cluster state and the configuration node

The cluster state holds all configuration and internal cluster data for the cluster. This cluster state information is held in nonvolatile memory. If the mainline power fails, the two uninterruptible power supplies maintain the internal power long enough for the cluster state information to be stored on the internal disk drive of each node. The read and write cache information is also held in nonvolatile memory. Similarly, if the power fails to a node, configuration and cache data for that node will be lost and the partner node attempts to flush the cache. The cluster state is still maintained by the other nodes on the cluster.

Figure 7 on page 39 shows an example cluster containing four nodes. The cluster state shown in the grey box does not actually exist, instead each node holds a copy of the entire cluster state.

The cluster contains a single node that is elected as the configuration node. The configuration node can be thought of as the node that controls the updating of

cluster state. For example, a user request is made (item 1), that results in a change being made to the configuration. The configuration node controls updates to the cluster (item 2). The configuration node then forwards the change to all nodes (including Node 1), and they all make the state-change at the same point in time (item 3). Using this state-driven model of clustering ensures that all nodes in the cluster know the exact cluster state at any one time.

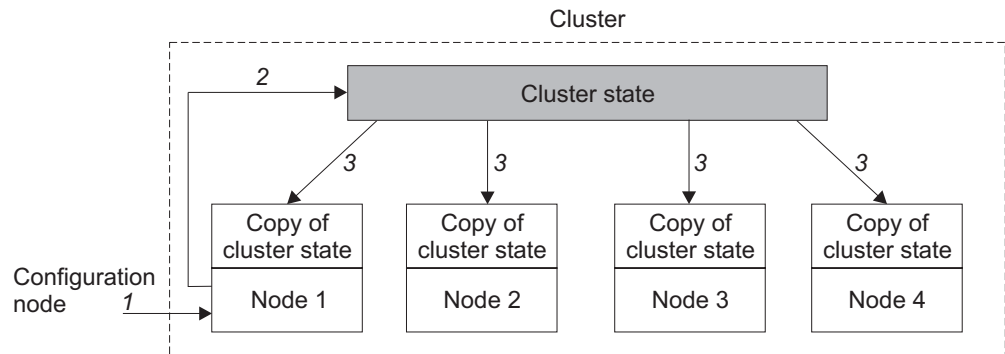


Figure 7. Cluster, nodes, and cluster state.

Cluster state

The cluster state holds all configuration and internal cluster data for the cluster.

The cluster state information is held in non-volatile memory. If the mainline power fails, the two uninterruptible power supply units maintain the internal power long enough for the cluster state information to be stored on the internal SCSI disk drive of each node. The read and write cache information, which is also held in memory, is stored on the internal SCSI disk drives of the nodes in the I/O group that are using that information.

Each node in the cluster maintains an identical copy of the cluster state. When a change is made to the configuration or internal cluster data, then the same change is applied to all nodes. For example, a user configuration request is made to the configuration node. This node forwards the request to all nodes in the cluster and they all make the change to the cluster state at the same point in time. This ensures that all nodes are aware of the configuration change. If the configuration node fails, then the cluster can elect a new node to take over its responsibilities.

Cluster operation and quorum disks

The cluster must contain at least half of its nodes to function.

Nodes are deployed in pairs known as I/O groups, and one to four I/O groups comprise a cluster. In order to function, one node in each I/O group must be operational. If both of the nodes in an I/O group are not operational, access is lost to the VDisks that are managed by the I/O group.

Note that the cluster can survive more than half the nodes failing as long as the cluster has stabilized between failures.

A tie-break situation can occur if exactly half the nodes in a cluster fail at the same time, or if the cluster is divided so that exactly half the nodes in the cluster cannot

communicate with the other half. For example, in a cluster of four nodes, if any two nodes fail at the same time or any two cannot communicate with the other two, a tie-break exists and must be resolved.

The cluster automatically chooses three managed disks to be *quorum disks* and assigns them quorum indexes of 0, 1, and 2. One of these disks is used to settle a tie-break condition.

If a tie-break occurs, the first half of the cluster to access the quorum disk after the split has occurred locks the disk and continues to operate. The other side stops. This action prevents both sides from becoming inconsistent with each other.

You can change the assignment of quorum disks at any time by issuing the following command:

```
svctask setquorum
```

I/O groups and uninterruptible power supply

Each pair of nodes is known as an **I/O group**.

Each node can only be in one I/O group. The I/O groups are connected to the SAN so that all backend storage and all application servers are visible to all of the I/O groups. Each pair has the responsibility to serve I/O on a particular virtual disk.

Virtual disks are logical disks that are presented to the SAN by SAN Volume Controller nodes. Virtual disks are also associated with an I/O group. The SAN Volume Controller does not contain any internal battery backup units and therefore must be connected to an uninterruptible power supply to provide data integrity in the event of a cluster wide power failure.

When an application server performs I/O to a virtual disk, it has the choice of accessing the virtual disk with either of the nodes in the I/O group. A virtual disk can have a preferred node specified when it is created. This is specified once the virtual disk is created. This is the node through which a virtual disk should normally be accessed. As each I/O group only has two nodes, the distributed cache in the SAN Volume Controller need only be two-way. When I/O is performed to a virtual disk, the node that processes the I/O duplicates the data onto the partner node that is in the I/O group.

I/O traffic for a particular virtual disk is, at any one time, handled exclusively by the nodes in a single I/O group. Thus, although a cluster may have many nodes within it, the nodes handle I/O in independent pairs. This means that the I/O capability of the SAN Volume Controller scales well, since additional throughput can be obtained by adding additional I/O groups.

Figure 8 on page 41 shows an example I/O group. A write operation from a host is shown (item 1), that is targeted for virtual disk A. This write is targeted at the preferred node, Node 1 (item 2). The write is cached and a copy of the data is made in the partner node, Node 2's cache (item 3). The write is now complete so far as the host is concerned. At some time later the data is written, or destaged, to storage (item 4). The figure also shows two uninterruptible power supply units (1 and 2) correctly configured so that each node is in a different power domain.

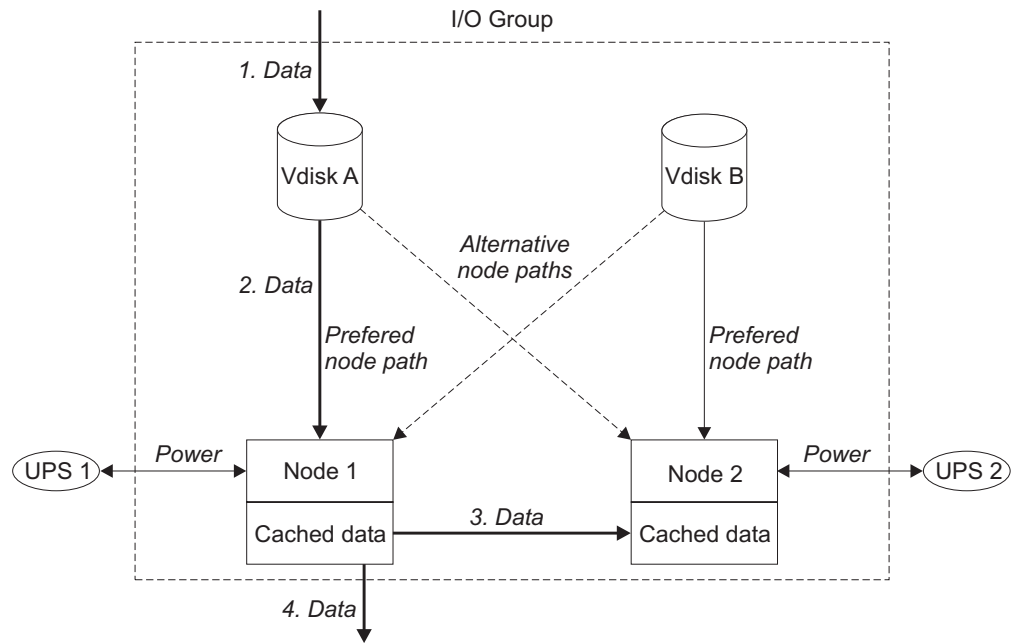


Figure 8. I/O group and uninterruptible power supply

When a node fails within an I/O group, the other node in the I/O group will take over the I/O responsibilities of the failed node. Data loss during a node failure is prevented by mirroring the I/O read/write data cache between the two nodes in an I/O group.

If only one node is assigned to an I/O group, or a node has failed in an I/O group, the cache goes into write-through mode. Therefore, any writes for the virtual disks that are assigned to this I/O group is not cached, it is sent directly to the storage device. If both nodes in an I/O group go offline, the virtual disks that are assigned to the I/O group cannot be accessed.

When a virtual disk is created, the I/O group that will provide access to the virtual disk must be specified. However, virtual disks can be created and added to I/O groups that contain offline nodes. I/O access will not be possible until at least one of the nodes in the I/O group is online.

The cluster also provides a **recovery I/O group**. This is used when both nodes in the I/O group have multiple failures. This allows you to move the virtual disks to the recovery I/O group and then into a working I/O group. I/O access is not possible when virtual disks are assigned to the recovery I/O group.

Uninterruptible power supply and power domains

An uninterruptible power supply protects the cluster against power failures.

If the mainline power fails to one or more nodes in the cluster, the uninterruptible power supply maintains the internal power long enough for the cluster state information to be stored on the internal SCSI disk drive of each node.

A cluster must have two or four uninterruptible power supply units. It is required that each node in the cluster be connected to an uninterruptible power supply. This allows the cluster to continue to work in degraded mode if one uninterruptible power supply fails.

It is very important that the two nodes in an I/O group are not both connected to the same power domain. Each SAN Volume Controller of an I/O group must be connected to a different uninterruptible power supply. This configuration ensures that the cache and cluster state information is protected against the failure of the uninterruptible power supply or of the mainline power source. If possible, each uninterruptible power supply should be connected to a different power source. Otherwise, a power source failure will result in the I/O group being taken offline. Table 13 shows the required number of uninterruptible power supply units for the number of nodes in a cluster.

Table 13. Required uninterruptible power supply (UPS) units

Number of nodes	Number of required UPS units
2 nodes	2 UPS units
4 nodes	2 UPS units
6 nodes	4 UPS units
8 nodes	4 UPS units

When nodes are added to the cluster, the I/O group they will join must be specified. The configuration interfaces will also check the uninterruptible power supply units and ensure that the two nodes in the I/O group are not connected to the same uninterruptible power supply units.

Figure 9 shows a cluster of four nodes, with two I/O groups and two uninterruptible power supply units.

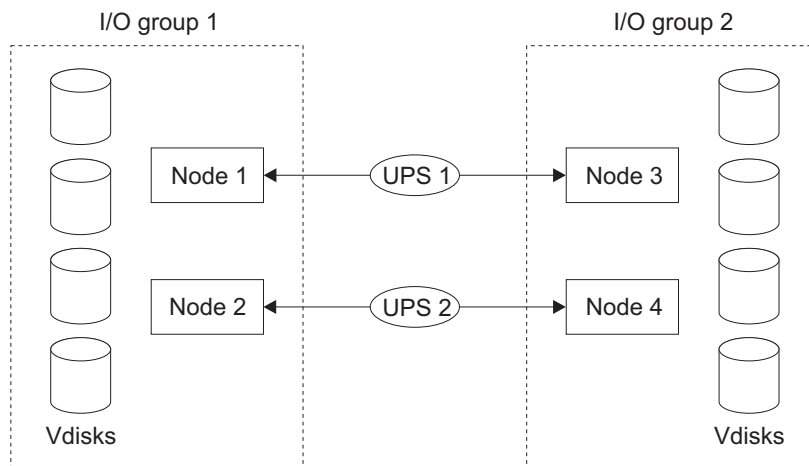


Figure 9. Relationship between I/O groups and uninterruptible power supply units

Attention: Do not connect two clusters to the same pair of uninterruptible power supply units. Both clusters will be lost in the event that a power failure occurs on both of the units.

Disk controllers

A disk controller is a device that coordinates and controls the operation of one or more disk drives and synchronizes the operation of the drives with the operation of the system as a whole.

Disk controllers provide the storage that the cluster detects as managed disks (MDisks).

When configuring disk controllers, ensure that you configure and manage the disk controllers and its devices for optimal performance.

The supported RAID controllers are detected by the cluster and reported by the user interfaces. The cluster can also determine which MDisks each controller has and can provide a view of MDisks filtered by controller. This view enables you to associate the MDisks with the RAID arrays that the controller presents.

Note: The SAN Volume Controller supports RAID controllers, but it is possible to configure a controller as a non-RAID controller. RAID controllers provide redundancy at the disk level. Therefore, a single physical disk failure does not cause an MDisk failure, an MDisk group failure, or a failure in the virtual disks (VDisks) that were created from the MDisk group.

The controller may have a local name for the RAID arrays or single disks that it is providing. However it is not possible for the nodes in the cluster to determine this name as the namespace is local to the controller. The controller will surface these disks with a unique ID, the controller LUN or LU number. This ID, along with the controller serial number or numbers (there may be more than one controller), can be used to associate the managed disks in the cluster with the RAID arrays presented by the controllers.

To prevent loss of data, virtualize only those RAID arrays that provide some form of redundancy, that is, RAID 1, RAID 10, RAID 0+1 or RAID 5. Do not use RAID 0 because a single physical disk failure might cause the failure of many VDIs.

Unsupported disk controller systems (generic controllers)

When a disk controller system is detected on the SAN, the SAN Volume Controller attempts to recognize it using its Inquiry data. If the disk controller system is recognized as one of the explicitly supported storage models, then the SAN Volume Controller uses error recovery programs that can be tailored to the known needs of the disk controller system. If the storage controller is not recognized, then the SAN Volume Controller configures the disk controller system as a generic controller. A generic controller may or may not function correctly when addressed by a SAN Volume Controller. The SAN Volume Controller does not regard accessing a generic controller as an error condition and, consequently, does not log an error. MDisks presented by generic controllers are not eligible to be used as quorum disks.

Related concepts

“Managed disks” on page 56

A managed disk (MDisk) is a logical disk (typically a RAID array or partition thereof) that a storage subsystem has exported to the SAN fabric to which the nodes in the cluster are attached.

Data migration

Data migration affects the mapping of the extents for a virtual disk (VDisk) to the extents for a managed disk (MDisk).

The host can access the VDisk during the data migration process.

Applications for data migration

There are several applications for data migration.

- Redistribution of workload within a cluster across managed disks:
 - Moving workload onto newly installed storage
 - Moving workload from old or failing storage, prior to replacing it
 - Moving workload to re-balance workload that has changed
- Migrating data from legacy disks to disks that are managed by the SAN Volume Controller.

Image mode virtual disk migration

Image mode virtual disks (VDisks) have the special property that the last extent in the VDisk can be a partial extent.

Managed mode disks do not have this property.

Once data is migrated off of a partial extent, you cannot migrate data back onto the partial extent.

Copy Services

The SAN Volume Controller provides Copy Services that enable you to copy virtual disks (VDisks).

These Copy Services are available for all supported hosts that are connected to the SAN Volume Controller.

FlashCopy

Makes an instant, point-in-time copy from a source VDisk to a target VDisk.

Remote Copy

Provides a consistent copy of a source VDisk on a target VDisk. Data is written to the target VDisk synchronously after it is written to the source VDisk, so the copy is continuously updated.

Applications for FlashCopy

You can use FlashCopy to back up data that changes frequently. After creating the point-in-time copy, it can be backed up to tertiary storage such as tape.

Another use of FlashCopy is for application testing. It is often important and useful to test a new version of an application on real business data before you move the application into production. This reduces the risk that the new application will fail because it is not compatible with actual business data.

You can also use FlashCopy to create copies for your auditing and data mining purposes.

In the scientific and technical arena, FlashCopy can create restart points for long-running batch jobs. Therefore, if a batch job fails many days into its run, you might be able to restart the job from a saved copy of its data. This is preferable to rerunning the entire multi-day job.

Applications for Remote Copy

Disaster recovery is the primary application for Remote Copy. Because an exact copy of your business data can be maintained at a remote location, you can use your remote location as a recovery site in the event of a local disaster.

FlashCopy

FlashCopy is a copy service available with the SAN Volume Controller.

It copies the contents of a source virtual disk (VDisk) to a target VDisk. Any data that existed on the target disk is lost and is replaced by the copied data. After the copy operation has been completed, the target virtual disks contain the contents of the source virtual disks as they existed at a single point in time unless target writes have been performed. Although the copy operation takes some time to complete, the resulting data on the target is presented in such a way that the copy appears to have occurred immediately. FlashCopy is sometimes described as an instance of a time-zero copy (T 0) or point-in-time copy technology. Although the FlashCopy operation takes some time, this time is several orders of magnitude less than the time which would be required to copy the data using conventional techniques.

It is difficult to make a consistent copy of a data set that is being constantly updated. Point-in-time copy techniques are used to help solve the problem. If a copy of a data set is taken using a technology that does not provide point in time techniques and the data set changes during the copy operation, then the resulting copy may contain data which is not consistent. For example, if a reference to an object is copied earlier than the object itself and the object is moved before it is itself copied then the copy will contain the referenced object at its new location but the reference will point to the old location.

Source VDIsks and target VDIsks must meet the following requirements:

- They must be the same size.
- The same cluster must manage them.

Related concepts

“FlashCopy consistency groups” on page 48

A consistency group is a container for mappings. You can add many mappings to a consistency group.

“FlashCopy mappings”

A FlashCopy mapping defines the relationship between a source VDisk and a target VDisk.

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

FlashCopy mappings

A FlashCopy mapping defines the relationship between a source VDisk and a target VDisk.

Because FlashCopy copies one VDisk to another VDisk, the SAN Volume Controller Console needs to be aware of that relationship. A particular virtual disk can take part in only one mapping; that is, a virtual disk can be the source or target of only one mapping. You cannot, for example, make the target of one mapping the source of another mapping.

FlashCopy makes an instant copy of a virtual disk at the time it is started. To create a FlashCopy of a virtual disk, you must first create a mapping between the source virtual disk (the disk that is copied) and the target virtual disk (the disk that receives the copy). The source and target must be of equal size.

To copy a VDisk, it must be part of a FlashCopy mapping or of a consistency group.

A FlashCopy mapping can be created between any two virtual disks in a cluster. It is not necessary for the virtual disks to be in the same I/O group or managed disk group. When a FlashCopy operation is started, a checkpoint is made of the source virtual disk. No data is actually copied at the time a start occurs. Instead, the checkpoint creates a bitmap that indicates that no part of the source virtual disk has yet been copied. Each bit in the bitmap represents one region of the source virtual disk. Such a region is called a grain.

After a FlashCopy operation starts, read operations to the source virtual disk continue to occur. If new data is written to the source (or target) virtual disk, then the existing data on the source is copied to the target virtual disk before the new data is written to the source (or target) virtual disk. The bitmap is updated to mark that the grain of the source virtual disk has been copied so that later write operations to the same grain do not recopy the data.

Similarly, during a read operation to the target virtual disk the bitmap is used to determine whether or not the grain has been copied. If the grain has been copied, the data is read from the target virtual disk. If the grain has not been copied, the data is read from the source virtual disk.

When you create a mapping, you specify the background copy rate. This rate determines the priority that is given to the background copy process. If you want to end with a copy of the whole source at the target (so that the mapping can be deleted, but the copy can still be accessed at the target), you must copy to the target virtual disk all the data that is on the source virtual disk.

When a mapping is started and the background copy rate is greater than zero (or a value other than NOCOPY is selected in the SAN Volume Controller Console's Creating FlashCopy Mappings panel), the unchanged data is copied to the target, and the bitmap is updated to show that the copy has occurred. After a time, the length of which depends on the priority given and the size of the virtual disk, the whole virtual disk is copied to the target. The mapping returns to the idle/copied state. You can restart the mapping at any time to create a new copy at the target; the process copy starts again.

If the background copy rate is zero (or NOCOPY), only the data that changes on the source is copied to the target. The target never contains a copy of the whole source unless every extent is overwritten at the source. You can use this copy rate when you need only a temporary copy of the source.

You can stop the mapping at any time after it has been started. This action makes the target inconsistent and therefore the target virtual disk is taken offline. You must restart the mapping to correct the target.

FlashCopy mapping states

At any point in time, a FlashCopy mapping is in one of the following states:

Idle or copied

The source and target VDisks act as independent VDisks even if a FlashCopy mapping exists between the two. Read and write caching is enabled for both the source and the target.

Copying

The copy is in progress.

Prepared

The mapping is ready to start. While in this state, the target VDisk is offline.

Preparing

Any changed write data for the source VDisk is flushed from the cache. Any read or write data for the target VDisk is discarded from the cache.

Stopped

The mapping is stopped because either you issued a command or an input/output (I/O) error occurred. Preparing and starting the mapping again can restart the copy.

Suspended

The mapping started, but it did not complete. The source VDisk might be unavailable, or the copy bitmap might be offline. If the mapping does not return to the copying state, stop the mapping to reset the mapping.

Before you start the mapping, you must prepare it. By preparing the mapping, you ensure that the data in the cache is destaged to disk and that a consistent copy of the source exists on disk. At this time the cache goes into write-through mode. Data that is written to the source is not cached in the SAN Volume Controllers; it passes straight through to the managed disks. The prepare operation for the mapping might take you a few minutes; the actual length of time depends on the size of the source virtual disk. You must coordinate the prepare operation with the operating system. Depending on the type of data that is on the source virtual disk, the operating system or application software might also cache data write operations. You must flush, or synchronize, the file system and application program before you prepare for, and finally start, the mapping.

For customers who do not need the complexity of consistency groups, the SAN Volume Controller allows a FlashCopy mapping to be treated as an independent entity. In this case the FlashCopy mapping is known as a stand alone mapping. For FlashCopy mappings which have been configured in this way, the **Prepare** and **Start** commands are directed at the FlashCopy mapping name rather than the consistency group ID.

Veritas Volume Manager

For FlashCopy target VDisks, the SAN Volume Controller sets a bit in the inquiry data for those mapping states where the target VDisk could be an exact image of the source VDisk. Setting this bit enables the Veritas Volume Manager to distinguish between the source and target VDisks and thus provide independent access to both.

Related concepts

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

FlashCopy consistency groups

A consistency group is a container for mappings. You can add many mappings to a consistency group.

The consistency group is specified when the mapping is created. You can also change the consistency group later. When you use a consistency group, you prepare and trigger that group instead of the various mappings. This ensures that a consistent copy is made of all the source VDisks. Mappings that you want to control at an individual level instead of at a consistency group level, should not be put into a consistency group. These mappings are known as stand-alone mappings.

To copy a VDisk, it must be part of a FlashCopy mapping or of a consistency group.

When you copy data from one virtual disk (VDisk) to another, that data might not include all that you need to enable you to use the copy. Many applications have data that spans multiple VDisks and that include the requirement that data integrity is preserved across VDisks. For example, the logs for a particular database usually reside on a different VDisk than the VDisk that contains the data.

Consistency groups address the problem when applications have related data that spans multiple VDisks. In this situation, FlashCopy must be performed in a way that preserves data integrity across the multiple VDisks. One requirement for preserving the integrity of data being written is to ensure that dependent writes are run in the intended sequence of the application.

FlashCopy consistency group states

At any point in time, a FlashCopy consistency group is in one of the following states:

Idle or copied

The source and target VDisks act independently even if a FlashCopy consistency group exists. Read and write caching is enabled for the source VDisks and target VDisks.

Copying

The copy is in progress.

Prepared

The consistency group is ready to start. While in this state, the target VDisks are offline.

Preparing

Any changed write data for the source VDisks is flushed from the cache. Any read or write data for the target VDisks is discarded from the cache.

Stopped

The consistency group is stopped because either you issued a command or an input/output (I/O) error occurred. Preparing and starting the consistency group again can restart the copy.

Suspended

The consistency group was started, but it did not complete. The source VDisks might be unavailable, or the copy bitmap might be offline. If the consistency group does not return to the copying state, stop the consistency group to reset the consistency group.

Related concepts

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Remote Copy

Remote Copy enables you to set up a relationship between two virtual disks, so that updates that are made by an application to one virtual disk are mirrored on the other virtual disk.

Although the application only writes to a single virtual disk, the SAN Volume Controller maintains two copies of the data. If the copies are separated by a significant distance, then the remote copy can be used as a backup for disaster recovery. A prerequisite for the SAN Volume Controller Remote Copy operations between two clusters is that the SAN fabric to which they are attached provides adequate bandwidth between the clusters.

One *VDisk* is designated the primary and the other *VDisk* is designated the secondary. Host applications write data to the primary *VDisk*, and updates to the primary *VDisk* are copied to the secondary *VDisk*. Normally, host applications do not perform input or output operations to the secondary *VDisk*. When a host writes to the primary *VDisk*, it will not receive confirmation of I/O completion until the write operation has completed for the copy on the secondary disk as well as on the primary.

Remote Copy supports the following features:

- Intracluster copying of a *VDisk*, in which both *VDisks* belong to the same cluster and I/O group within the cluster.
- Intercluster copying of a *VDisk*, in which one *VDisk* belongs to a cluster and the other *VDisk* belongs to a different cluster

Note: A cluster can only participate in active Remote Copy relationships with itself and one other cluster.

- Intercluster and intracluster Remote Copy can be used concurrently within a cluster.
- The intercluster link is bidirectional. Meaning, it can support copying of data from clusterA to clusterB for one pair of *VDisks* while copying data from clusterB to clusterA for a different pair of *VDisks*.
- The copy direction can be reversed for a consistent relationship by issuing a simple **switch** command. See *IBM TotalStorage SAN Volume Controller: Command-Line Interface User's Guide*.
- Remote Copy consistency groups are supported for ease of managing a group of relationships that need to be kept in sync for the same application. This also simplifies administration, as a single command issued to the consistency group will be applied to all the relationships in that group.

Related concepts

“Synchronous Remote Copy” on page 50

In the synchronous mode, Remote Copy provides a *consistent copy*, which means that the primary *VDisk* is always the exact match of the secondary *VDisk*.

“Remote Copy consistency groups” on page 50

Remote Copy provides the facility to group a number of relationships into a Remote Copy consistency group so that they can be manipulated in unison.

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Synchronous Remote Copy

In the synchronous mode, Remote Copy provides a *consistent* copy, which means that the primary VDisk is always the exact match of the secondary VDisk.

The host application writes data to the primary VDisk but does not receive the final status on the write operation until the data is written to the secondary VDisk. For disaster recovery, this mode is the only practical mode of operation because a consistent copy of the data is maintained. However, synchronous mode is slower than asynchronous mode because of the latency time and bandwidth limitations imposed by the communication link to the secondary site.

Related concepts

“Remote Copy consistency groups”

Remote Copy provides the facility to group a number of relationships into a Remote Copy consistency group so that they can be manipulated in unison.

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Remote Copy consistency groups

Remote Copy provides the facility to group a number of relationships into a Remote Copy consistency group so that they can be manipulated in unison.

Certain uses of Remote Copy require the manipulation of more than one relationship. A command issued to the consistency group will be applied to all of the relationships in the group simultaneously.

For some uses it might be that the relationships share some loose association and that the grouping simply provides a convenience for the administrator. But a more significant use arises when the relationships contain VDisks that have a tighter association. One example is when the data for an application is spread across more than one VDisk. A more complex example is when multiple applications run on different host systems. Each application has data on different VDisks, and these applications exchange data with each other. Both these examples are cases in which specific rules exist as to how the relationships must be manipulated, in unison. This ensures that the set of secondary VDisks contains usable data. The key property is that these relationships be consistent. Hence, the groups are called consistency groups.

A relationship can be part of a single consistency group or not be part of a consistency group at all. Relationships that are not part of a consistency group are called stand-alone relationships. A consistency group can contain zero or more relationships. All the relationships in a consistency group must have matching master and auxiliary clusters. All relationships in a consistency group must also have the same copy direction and state.

Remote Copy consistency group states

Inconsistent (Stopped)

The primary VDisks are accessible for read and write input/output (I/O) operations but the secondary VDisks are not accessible for either. A copy process needs to be started to make the Secondary VDisks consistent.

Inconsistent (Copying)

The primary VDisks are accessible for read and write I/O operations but the secondary VDisk are not accessible for either. This state is entered after a **Start** command is issued to a consistency group in the InconsistentStopped state. This state is also entered when a **Start** command is issued, with the force option, to a consistency group in the Idling or ConsistentStopped state.

Consistent (Stopped)

The secondary VDisks contain a consistent image, but it might be out-of-date with respect to the primary VDisks. This state can happen when a relationship was in the ConsistentSynchronized state and experiences an error which forces a freeze of the consistency group. This state can also happen when a relationship is created with the CreateConsistentFlag set to TRUE.

Consistent (Synchronized)

The primary VDisks are accessible for read and write I/O operations. The secondary VDisks are accessible for read-only I/O operations.

Idling Master VDisks and Auxiliary VDisks are operating in the primary role. Consequently the VDisks are accessible for write I/O operations.

Idling (Disconnected)

The VDisks in this half of the consistency group are all operating in the primary role and can accept read or write I/O operations.

Inconsistent (Disconnected)

The VDisks in this half of the consistency group are all operating in the secondary role and will not accept read or write I/O operations.

Consistent (Disconnected)

The VDisks in this half of the consistency group are all operating in the secondary role and will accept read I/O operations but not write I/O operations

Empty

The consistency group contains no relationships.

Related concepts

“Virtual disks” on page 60

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Chapter 5. Object descriptions

The SAN Volume Controller is based on a group of virtualization concepts. Before setting up the system, you should understand the concepts and the objects in the system.

The smallest processing unit in a SAN Volume Controller is a single **node**. Nodes are deployed in pairs to make up a cluster. A cluster can consist of one to four pairs of nodes. Each pair of nodes is known as an **I/O group**. Each node may be in only one I/O group.

Virtual disks (Vdisks) are logical disks that are presented by clusters. Each virtual disk is associated with a particular I/O group. The nodes in the I/O group provide access to the virtual disks in the I/O group. When an application server performs I/O to a virtual disk, it has the choice of accessing the virtual disk with either of the nodes in the I/O group. As each I/O group only has two nodes, the distributed cache that the SAN Volume Controller provides is only two-way.

Each node does not contain any internal battery backup units and therefore must be connected to an **Uninterruptible power supply (UPS)** to provide data integrity in the event of a cluster wide power failure. In such situations, the UPS will maintain power to the nodes while the contents of the distributed cache are dumped to an internal drive.

The nodes in a cluster see the storage presented by backend **disk controllers** as a number of disks, known as **managed disks (MDisks)**. Because the SAN Volume Controller does not attempt to provide recovery from physical disk failures within the backend disk controllers, a managed disk is usually, but not necessarily, a RAID array.

Each managed disk is divided up into a number of **extents** (default size is 16 MB) which are numbered, from 0, sequentially from the start to the end of the managed disk.

Managed disks are collected into groups, known as **managed disk groups (MDisk group)**. Virtual disks are created from the extents contained by a managed disk group. The managed disks that constitute a particular virtual disk must all come from the same managed disk group.

At any one time, a single node in the cluster is used to manage configuration activity. This **configuration node** manages a cache of the information that describes the cluster configuration and provides a focal point for configuration.

The SAN Volume Controller detects the fibre-channel ports that are connected to the SAN. These correspond to the World Wide Port Names (WWPNs) of the host bus adapter (HBA) fibre-channels that are present in the application servers. The SAN Volume Controller allows you to create logical host objects that group together WWPNs belonging to a single application server or a set of them.

Application servers can only access virtual disks that have been allocated to them. Virtual disks can be mapped to a host object. Mapping a virtual disk to a host object makes the virtual disk accessible to the WWPNs in that host object, and hence the application server itself.

The SAN Volume Controller provides block-level aggregation and volume management for disk storage within the SAN. In simpler terms, this means that the SAN Volume Controller manages a number of back-end storage controllers and maps the physical storage within those controllers into logical disk images that can be seen by application servers and workstations in the SAN. The SAN is configured in such a way that the application servers cannot see the back-end physical storage. This prevents any possible conflict between the SAN Volume Controller and the application servers both trying to manage the back-end storage.

Figure 10 illustrates the objects that are described in this section and their logical positioning in a virtualized system. To simplify the example, virtual disk to host mappings are not shown.

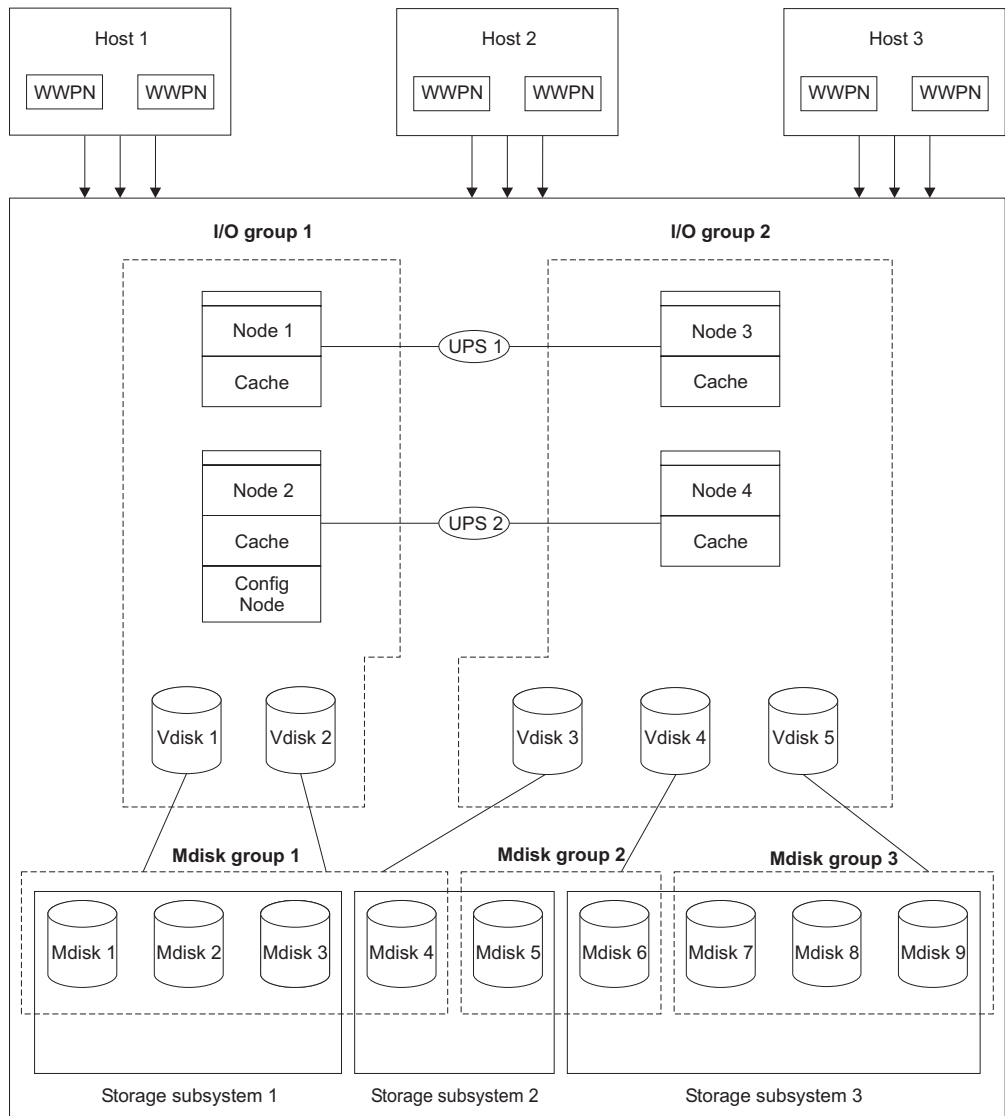


Figure 10. Objects in a virtualized system

Storage subsystems

A storage subsystem is a device that coordinates and controls the operation of one or more disk drives and synchronizes the operation of the drives with the operation of the system as a whole.

Storage subsystems attached to the SAN fabric provide the physical storage devices that the cluster detects as managed disks. These are usually RAID arrays, because the SAN Volume Controller does not attempt to provide recovery from physical disk failures within the storage subsystem. The nodes in the cluster are connected to one or more fibre-channel SAN fabrics.

The exported storage devices are detected by the cluster and reported by the user interfaces. The cluster can also determine which managed disks each storage subsystem is presenting, and can provide a view of managed disks filtered by the storage subsystem. This allows you to associate the managed disks with the RAID arrays that the subsystem exports.

The storage subsystem may have a local name for the RAID arrays or single disks that it is providing. However it is not possible for the nodes in the cluster to determine this name, because the namespace is local to the storage subsystem. The storage subsystem will surface these storage devices with a unique ID, the logical unit number (LUN). This ID, along with the storage subsystem serial number or numbers (there may be more than one controller in a storage subsystem), can be used to associate the managed disks in the cluster with the RAID arrays exported by the subsystem.

Storage subsystems export storage to other devices on the SAN. The physical storage associated with a subsystem is normally configured into RAID arrays that provide recovery from physical disk failures. Some subsystems also allow physical storage to be configured as RAID-0 arrays (striping) or as JBODs. However, this does not provide protection against a physical disk failure and, with virtualization, can lead to the failure of many virtual disks.

Many storage subsystems allow the storage provided by a RAID array to be divided up into many SCSI logical units (LUs) that are presented on the SAN. With the SAN Volume Controller it is recommended that storage subsystems are configured to present each RAID array as a single SCSI LU that will be recognized by the SAN Volume Controller as a single managed disk. The virtualization features of the SAN Volume Controller can then be used to divide up the storage into virtual disks.

Some storage subsystems allow the exported storage to be increased in size. The SAN Volume Controller will not use this extra capacity. Instead of increasing the size of an existing managed disk, a new managed disk should be added to the managed disk group and the extra capacity will be available for the SAN Volume Controller to use.

Attention: If you delete a RAID that is being used by the SAN Volume Controller, the MDisk group will go offline and the data in that group will be lost.

When configuring your storage subsystems, ensure that you configure and manage your subsystems and its devices for optimal performance.

The cluster detects and provides a view of the storage subsystems that the SAN Volume Controller supports. The cluster can also determine which MDisks each subsystem has and can provide a view of MDisks filtered by device. This view enables you to associate the MDisks with the RAID arrays that the subsystem presents.

Note: The SAN Volume Controller Console supports storage that is internally configured as a RAID array. However, it is possible to configure a storage subsystem as a non-RAID device. RAID provides redundancy at the disk

level. For RAID devices, a single physical disk failure does not cause an MDisk failure, an MDisk group failure, or a failure in the virtual disks (VDisks) that were created from the MDisk group.

Storage subsystems reside on the SAN fabric and are addressable by one or more fibre-channel ports (target ports). Each port has a unique name known as a worldwide port name (WWPN).

Managed disks

A managed disk (MDisk) is a logical disk (typically a RAID array or partition thereof) that a storage subsystem has exported to the SAN fabric to which the nodes in the cluster are attached.

A managed disk might, therefore, consist of multiple physical disks that are presented as a single logical disk to the SAN. A managed disk always provides usable blocks of physical storage to the cluster even if it does not have a one-to-one correspondence with a physical disk.

Each managed disk is divided into a number of *extents*, which are numbered, from 0, sequentially from the start to the end of the managed disk. The extent size is a property of managed disk groups. When an MDisk is added to an MDisk group, the size of the extents that the MDisk will be broken into depends on the attribute of the MDisk group to which it has been added.

Access modes

The access mode determines how the cluster uses the MDisk. The possible modes are:

Unmanaged

The MDisk is not used by the cluster.

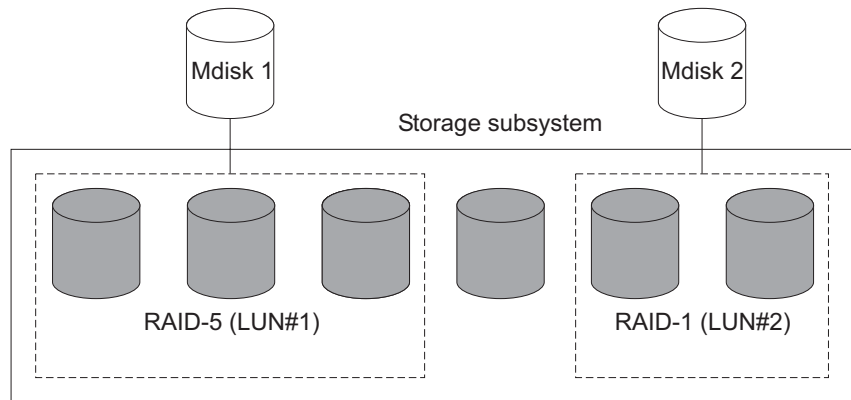
Managed

The MDisk is assigned to an MDisk group and is providing extents that virtual disks (VDisks) can use.

Image The MDisk is assigned directly to a VDisk with a one-to-one mapping of extents between the MDisk and the VDisk.

Attention: If you add a managed disk that contains existing data to a managed disk group, you will lose the data that it contains. The *image mode* is the only mode that will preserve this data.

Figure 11 on page 57 shows physical disks and managed disks.





Key:  = Physical disks  = Logical disks (managed disks as seen by the 2145)

Figure 11. Controllers and MDisks

Table 14 describes the operational states of a managed disk:

Table 14. Managed disk status

Status	Description
Online	The MDisk can be accessed by all online nodes. That is, all the nodes that are currently working members of the cluster can access this MDisk. The MDisk is online when the following conditions are met: <ul style="list-style-type: none"> • All timeout error recovery procedures complete and report the disk as online. • LUN inventory of the target ports correctly reported the MDisk. • Discovery of this LUN completed successfully. • All of the managed disk target ports report this LUN as available with no fault conditions.
Degraded	The MDisk cannot be accessed by all the online nodes. That is, one or more (but not all) of the nodes that are currently working members of the cluster cannot access this MDisk. The MDisk may be partially excluded; that is, some of the paths to the MDisk (but not all) have been excluded.
Excluded	The MDisk has been excluded from use by the cluster after repeated access errors. Run the Directed Maintenance Procedures to determine the problem. You can reset an MDisk and include it in the cluster again by running the svctask includemdisk command.
Offline	The MDisk cannot be accessed by any of the online nodes. That is, all of the nodes that are currently working members of the cluster cannot access this MDisk. This state can be caused by a failure in the SAN, the storage subsystem, or one or more physical disks connected to the storage subsystem. The MDisk will only be reported as offline if all paths to the disk fail.

Extents

Each MDisk is divided into chunks of equal size called *extents*. Extents are a unit of mapping the data between MDisks and virtual disks (VDisks).

Attention: If your fabric is undergoing transient link breaks or you have been replacing cables or connections in your fabric, you might see one or more MDisks change to the degraded status. If an I/O operation was attempted during the link breaks and the same I/O failed several times, the MDisk will be partially excluded and will change to a status of degraded. You should include the MDisk to resolve the problem. You can include the MDisk by either selecting the Include MDisk task from the Work with Managed Disks - Managed Disk panel in the SAN Volume Controller Console, or issue the following command:

```
svctask includemdisk <mdiskname/id>
```

Managed disk path Each managed disk will have an online path count, which is the number of nodes that have access to that managed disk; this represents a summary of the I/O path status between the cluster nodes and the particular storage device. The maximum path count is the maximum number of paths that have been detected by the cluster at any point in the past. Thus if the current path count is not equal to the maximum path count then the particular managed disk may be degraded. That is, one or more nodes may not see the managed disk on the fabric.

Managed disk groups

An *MDisk group* is a collection of MDisks that jointly contain all the data for a specified set of virtual disks (VDisks).

All MDisks in a group are split into extents of the same size. VDIsks are created from the extents that are available in the group. You can add MDisks to an MDisk group at any time. This way you increase the number of extents that are available for new VDIsks or to expand existing VDIsks.

Note: RAID array partitions on HP StorageWorks subsystems controllers are only supported in single-port attach mode. MDisk groups that consist of single-port attached subsystems and other storage subsystems are not supported.

You can add MDisks to an MDisk group at any time either to increase the number of extents that are available for new VDIsks or to expand existing VDIsks. You can add only MDisks that are in unmanaged mode. When MDisks are added to a group, their mode changes from unmanaged to managed.

You can delete MDisks from a group under the following conditions:

- VDIsks are not using any of the extents that are on the MDisk.
- Enough free extents are available elsewhere in the group to move any extents that are in use from this MDisk.

Attention: If you delete an MDisk group, you destroy all the VDIsks that are made from the extents that are in the group. If the group is deleted, you cannot recover the mapping that existed between extents that are in the group and the extents that VDIsks use. The MDisks that were in the group are returned to unmanaged mode and can be added to other groups. Because the deletion of a group can cause a loss of data, you must force the deletion if VDIsks are associated with it.

Table 15 describes the operational states of an MDisk group.

Table 15. Managed disk group status

Status	Description
Online	The MDisk group is online and available. All the MDisks in the group are available.
Degraded	The MDisk group is available; however, one or more nodes cannot access all the MDisks in the group.
Offline	The MDisk group is offline and unavailable. No nodes in the cluster can access the MDisks. The most likely cause is that one or more MDisks are offline or excluded.

Attention: If a single MDisk in an MDisk group is offline and therefore cannot be seen by any of the online nodes in the cluster, then the MDisk group of which this MDisk is a member goes offline. This causes *all* the VDIs that are being presented by this MDisk group to go offline. Care should be taken when creating MDisk groups to ensure an optimal configuration.

Consider the following guidelines when you create MDisk groups:

- If you are creating image-mode VDIs, do not put all of these VDIs into one MDisk group because a single MDisk failure results in all of these VDIs going offline. Allocate your image-mode VDIs between your MDisk groups.
- Ensure that all MDisks that are allocated to a single MDisk group are the same RAID type. This ensures that a single failure of a physical disk in the storage subsystem does not take the entire group offline. For example, if you have three RAID-5 arrays in one group and add a non-RAID disk to this group, then you lose access to all the data striped across the group if the non-RAID disk fails. Similarly, for performance reasons you should not mix RAID types. The performance of all VDIs will be reduced to the lowest performer in the group.
- If you intend to keep the virtual disk allocation within the storage exported by a storage subsystem, ensure that the MDisk group that corresponds with a single subsystem is presented by that subsystem. This also enables nondisruptive migration of data from one subsystem to another subsystem and simplifies the decommissioning process if you want to decommission a controller at a later time.
- Except when migrating between groups, a VDI must be associated with just one MDisk group.
- An MDisk can be associated with just one MDisk group.

Extent

To track the space that is available, the SAN Volume Controller divides each MDisk in an MDisk group into chunks of equal size. These chunks are called *extents*, and are indexed internally. Extent sizes can be 16, 32, 64, 128, 256, or 512 MB.

You must specify the extent size when you create a new MDisk group. You cannot change the extent size later; it must remain constant throughout the lifetime of the MDisk group. MDisk groups can have different extent sizes. However, different extent sizes can place restrictions on the use of data migration. The choice of extent size affects the total amount of storage that can be managed by a SAN Volume Controller cluster. Table 16 on page 60 shows the maximum amount of

storage that can be managed by a cluster for each extent size. Because the SAN Volume Controller allocates a whole number of extents to each virtual disk that is created, using a larger extent size can increase the amount of wasted storage at the end of each virtual disk. Larger extent sizes also reduce the ability of the SAN Volume Controller to distribute sequential I/O workloads across many managed disks. Therefore, larger extent sizes might reduce the performance benefits of virtualization.

Table 16. Capacities of the cluster given extent size

Extent size	Maximum storage capacity of cluster
16 MB	64 TB
32 MB	128 TB
64 MB	256 TB
128 MB	512 TB
256 MB	1 PB
512 MB	2 PB

Figure 12 shows an MDisk group containing four MDisks.

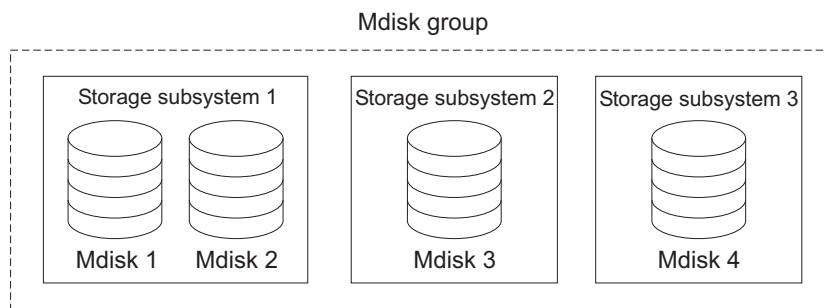


Figure 12. MDisk group

Virtual disks

A *VDisk* is a logical disk that the cluster presents to the storage area network (SAN).

Application servers on the SAN access *VDisks*, not managed disks (MDisks). *VDisks* are created from a set of extents in an MDisk group. There are three types of *VDisks*: striped, sequential, and image.

Types

You can create the following types of *VDisks*:

Striped

The striping is at extent level. One extent is allocated, in turn, from each managed disk that is in the group. For example, a managed disk group that has 10 MDisks takes one extent from each managed disk. The 11th extent is taken from the first managed disk, and so on. This procedure, known as a round-robin, is similar to RAID-0 striping.

You can also supply a list of MDisks to use as the stripe set. This list can contain two or more MDisks from the managed disk group. The round-robin procedure is used across the specified stripe set.

Attention: Care should be taken when specifying a stripe set if your MDisk group contains MDisks of unequal size. By default striped VDIs are striped across all MDisks in the group. If some of the MDisks are smaller than others, the extents on the smaller MDisks will be used up before the larger MDisks run out of extents. Manually specifying the stripe set in this case might result in the VDisk not being created.

If you are unsure about whether there is sufficient free space to create a striped VDisk select one of the following options:

- Check the free space on each MDisk in the group, using the **svcinfolsfreeextents** command
- Let the system automatically create the VDisk, by not supplying a specific stripe set.

Figure 13 shows an example of a managed disk group containing three MDisks. This figure also shows a striped virtual disk created from the extents available in the group.

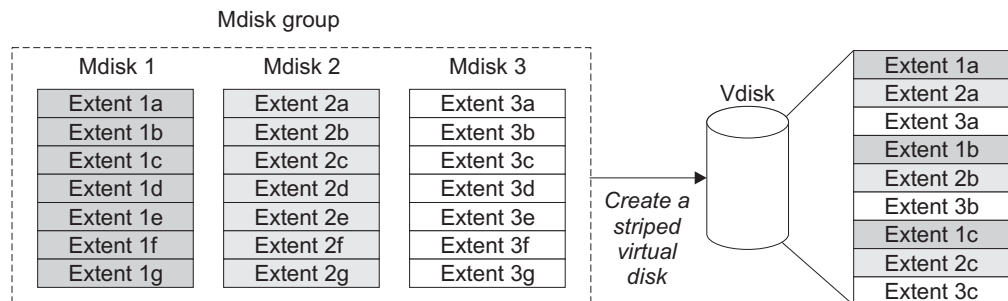


Figure 13. Managed disk groups and VDIs

Sequential

When selected, extents are allocated sequentially on one managed disk to create the virtual disk if enough consecutive free extents are available on the chosen managed disk.

Image

Image-mode VDIs are special VDIs that have a direct relationship with one managed disk. If you have a managed disk that contains data that you want to merge into the cluster, you can create an image-mode virtual disk. When you create an image-mode virtual disk, a direct mapping is made between extents that are on the managed disk and extents that are on the virtual disk. The managed disk is not virtualized. In other words, the logical block address (LBA) x on the managed disk is the same as LBA x on the virtual disk.

When you create an image-mode VDisk, you must assign it to a managed disk group. An image-mode VDisk must be at least one extent in size. In other words, the minimum size of an image-mode VDisk is the extent size of the MDisk group to which it is assigned.

The extents are managed in the same way as other VDIs. When the extents have been created, you can move the data onto other MDisks that are in the group without losing access to the data. After you move one or

more extents, the virtual disk becomes a real virtualized disk, and the mode of the managed disk changes from image to managed.

Attention: If you add an MDisk to an MDisk group as a managed disk, any data on the MDisk will be lost. Ensure that you create image-mode VDIs from the MDisks that contain data before you start adding any MDisks to groups.

MDisks that contain existing data have an initial mode of unmanaged, and the cluster cannot determine whether it contains partitions or data.

A virtual disk can have one of three states. Table 17 describes the different states of a virtual disk:

Table 17. Virtual disk status

Status	Description
Online	The virtual disk is online and available if both nodes in the I/O group can access the virtual disk. A single node will only be able to access a VDisk if it can access all the MDisks in the MDisk group associated with the VDisk.
Offline	The VDisk is offline and unavailable if both nodes in the I/O group are missing or none of the nodes in the I/O group that are present can access the VDisk.
Degraded	The status of the virtual disk is degraded if one node in the I/O group is online and the other node is either missing or cannot access the virtual disk.

You can also use more sophisticated extent allocation policies to create VDIs. When you create a striped virtual disk, you can specify the same managed disk more than once in the list of MDisks that are used as the stripe set. This is useful if you have a managed disk group in which not all the MDisks are of the same capacity. For example, if you have a managed disk group that has two 18 GB MDisks and two 36 GB MDisks, you can create a striped virtual disk by specifying each of the 36 GB MDisks twice in the stripe set so that two thirds of the storage is allocated from the 36 GB disks.

If you delete a virtual disk, you destroy access to the data that is on the virtual disk. The extents that were used in the virtual disk are returned to the pool of free extents that is in the managed disk group. The deletion might fail if the virtual disk is still mapped to hosts. The deletion might also fail if the virtual disk is still part of a FlashCopy or a Remote Copy mapping. If the deletion fails, you can specify the force-delete flag to delete both the virtual disk and the associated mappings to hosts. Forcing the deletion will also delete the copy services relationship and mappings.

Related concepts

“Virtualization” on page 1

Virtualization is a concept that applies to many areas of the information technology industry.

Virtual disk-to-host mapping

Virtual disk-to-host mapping is the process of controlling which hosts have access to specific virtual disks (VDIs) within the SAN Volume Controller.

Virtual disk-to-host mapping is similar in concept to logical unit number (LUN) mapping or masking. LUN mapping is the process of controlling which hosts have access to specific logical units (LUs) within the disk controllers. LUN mapping is typically done at the disk controller level. Virtual disk-to-host mapping is done at the SAN Volume Controller level.

Application servers can only access VDisks that have been made accessible to them. The SAN Volume Controller detects the fibre-channel ports that are connected to the SAN. These correspond to the host bus adapter (HBA) worldwide port names (WWPNs) that are present in the application servers. The SAN Volume Controller enables you to create logical hosts that group together WWPNs belonging to a single application server. VDisks can then be mapped to a host. The act of mapping a virtual disk to a host makes the virtual disk accessible to the WWPNs in that host, and hence the application server itself.

VDisks and host mappings

The SAN concept known as LUN masking usually requires device driver software in each host. The device driver software masks the LUNs as instructed by the user. After the masking has been done, only some disks are visible to the operating system. The SAN Volume Controller performs a similar function, but, by default, it presents to the host only those VDisks that are mapped to that host. You must therefore map the VDisks to the hosts that are to access those VDisks.

Each host mapping associates a virtual disk with a host object and allows all HBA ports in the host object to access the virtual disk. You can map a virtual disk to multiple host objects. When a mapping is created, multiple paths might exist across the SAN fabric from the hosts to the SAN Volume Controllers that are presenting the virtual disk. Most operating systems present each path to a virtual disk as a separate storage device. The SAN Volume Controller, therefore, needs the IBM Subsystem Device Driver (SDD) software to be running on the host. This software handles the many paths that are available to the virtual disk and presents a single storage device to the operating system.

When you map a virtual disk to a host, you can optionally specify a SCSI ID for the virtual disk. This ID controls the sequence in which the VDisks are presented to the host. Take care when you specify a SCSI ID, because some device drivers stop looking for disks if they find an empty slot. For example, if you present three VDisks to the host, and those VDisks have SCSI IDs of 0, 1, and 3, the virtual disk that has an ID of 3 might not be found because no disk is mapped with an ID of 2. The cluster automatically assigns the next available SCSI ID if none is entered.

Figure 14 on page 64 and Figure 15 on page 64 show two VDisks, and the mappings that exist between the host objects and these VDisks.

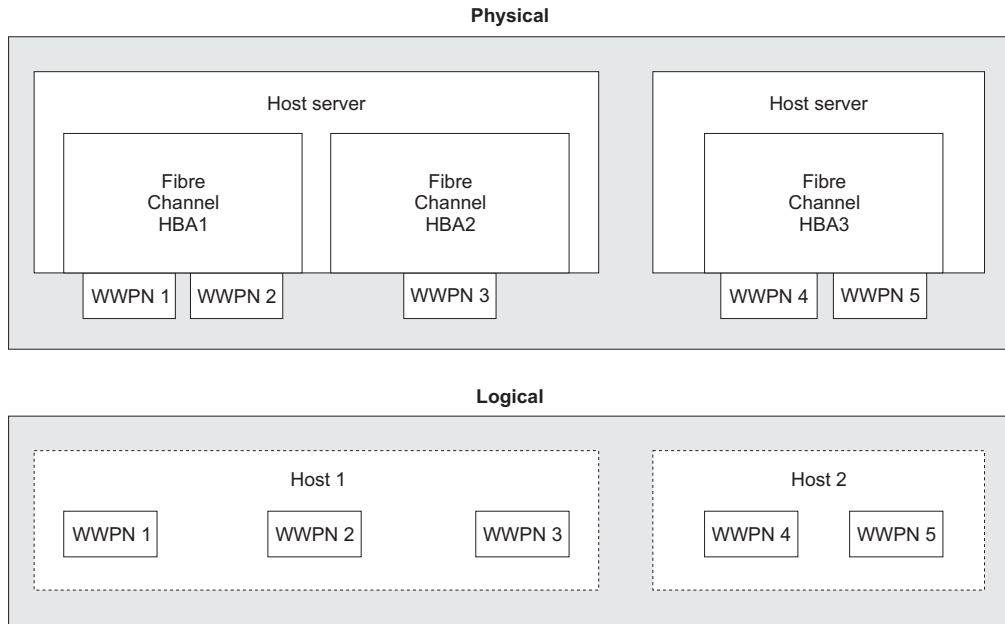


Figure 14. Hosts, WWPNs, and VDisks

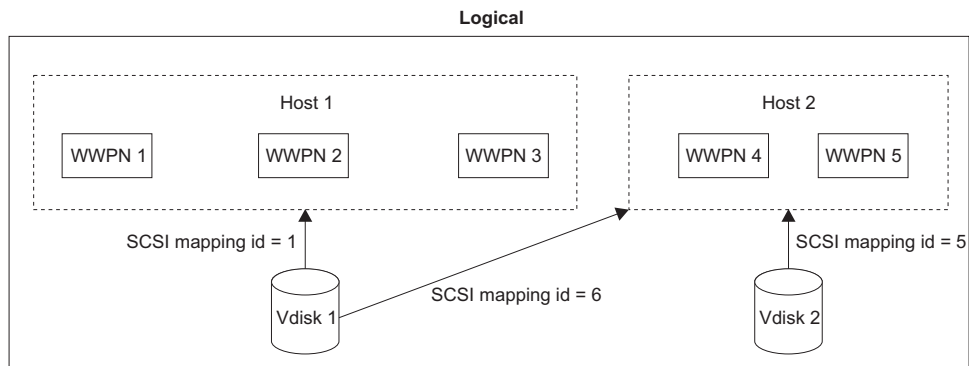


Figure 15. Hosts, WWPNs, VDisks and SCSI mappings

Host objects

A host system is an open-systems computer that is connected to the switch through a fibre-channel interface.

Creating a host in a cluster results in the creation of a logical host object. A logical host object has one or more worldwide port names (WWPNs) assigned to it. Generally, a logical host object is associated with a physical host system. However, a single logical host object can have WWPNs from multiple physical host systems that are assigned to it.

A host object is a logical object that groups one or more worldwide port names (WWPNs) of the host bus adapters (HBAs) that the cluster has detected on the SAN. A typical configuration has one host object for each host that is attached to the SAN. If, however, a cluster of hosts is going to access the same storage, you can add HBA ports from several hosts into the one host object to make a simpler configuration.

The cluster does not automatically present VDIs on the fibre. You must map each virtual disk to a particular set of ports to enable the virtual disk to be accessed through those ports. The mapping is made between a host object and a virtual disk.

When you create a new host object, the configuration interfaces provide a list of unconfigured WWPNs. These WWPNs represent the fibre channel ports that the cluster has detected.

The cluster can detect only ports that are logged into the fabric. Some HBA device drivers do not let the ports remain logged in if no disks are visible on the fabric. This condition causes a problem when you want to create a host because, at this time, no VDIs are mapped to the host. The configuration interface provides a method by which you can manually enter port names under this condition.

Attention: You must not include a node port in a host object.

A port can be added to only one host object. When a port has been added to a host object, that port becomes a configured WWPN, and is not included in the list of ports that are available to be added to other hosts.

Node login counts

The number of nodes that can see each port is reported on a per node basis and is known as the node login count. If the count is less than the number of nodes in the cluster, then there is a fabric problem, and not all nodes can see the port.

Chapter 6. Planning for configuring the SAN Volume Controller

Before you configure the SAN Volume Controller, you must complete these planning tasks.

Planning the clusters

Before you create a SAN Volume Controller cluster:

- Determine the number of clusters and the number of pairs of nodes. Each pair of nodes (the I/O group) is the container for one or more VDIs.
- Determine the number of hosts that will be used with the SAN Volume Controller. Hosts should be grouped by operating system and by type of host bus adapter (HBA).
- Determine the number of I/Os per second between the hosts and SAN Volume Controller nodes.

Planning the host groups

Host systems have access to specific logical units (LUs) within the disk controllers as a result of LUN masking. To plan a host group, gather the following information:

- List all of the worldwide port names (WWPNs) of the fibre-channel host bus adapter ports in the hosts.
- Determine the name to assign to the host or host group.
- Determine the VDIs to assign to the host.

Planning the managed disks

To plan the managed disks (MDisks), determine the logical or physical disks (logical units) in the backend storage.

Planning the managed disk groups

Before you create managed disk (MDisk) groups, determine the following factors:

- Determine the types of backend controllers in the system.
- If you want to create VDIs with the sequential policy, plan to create a separate MDisk group for these VDIs or ensure that you create these VDIs before creating VDIs with the striped policy.
- Plan to create MDisk groups for the backend controllers that provide the same level of performance or reliability, or both. For example, you can group all of the managed disks that are RAID 10 in one MDisk group and all of the MDisks that are RAID 5 in another group.

Planning the virtual disks

An individual virtual disk is a member of one managed disk group and one I/O group. The managed disk group defines which managed disks provide the backend storage that makes up the virtual disk. The I/O group defines which SAN Volume Controller nodes provide I/O access to the virtual disk. Determine the following information before creating a virtual disk:

- The name to assign to the virtual disk.
- The I/O group to which the virtual disk will be assigned.

- The managed disk group to which the virtual disk will be assigned.
- The capacity of the virtual disk.

Maximum configuration

Ensure that you are familiar with the maximum configurations of the SAN Volume Controller.

Table 18 shows the maximum configuration values to consider when planning the SAN Volume Controller installation.

Table 18. SAN Volume Controller maximum configuration values

Objects	Maximum number	Comments
Cluster Properties		
Nodes	8	Arranged as four I/O groups.
I/O groups	4	Each containing two nodes.
MDisk group	128	--
MDisks	4096	Represents an average of 64 per controller.
Object MDisks per MDisk group	128	--
MDisk size	2 TB	Defined by 32-bit LBA limit.
Addressability	2.1 PB	Maximum extent size 512 MB, arbitrary limit of 2^{22} extents in map.
LU size	2 TB	Defined by 32-bit LBA limit.
Concurrent SCSI tasks (commands) per node	2500	--
Concurrent commands per node	2500	Assumes a backend latency of 100 ms.
Concurrent commands per FC port	2048	--
SDD	512 SAN Volume Controller vpaths per host	<p>One vpath is created for each VDisk mapped to a host. Although the SAN Volume Controller only permits 512 VDIs to be mapped to a host, the SDD limit can be exceeded by either:</p> <ul style="list-style-type: none"> • Creating two (or more) host objects for one physical host and mapping more than 512 VDIs to the host using the multiple host objects. • Creating two (or more) clusters and mapping more than 512 VDIs to the host using the multiple clusters. <p>Note: Both of these operations are unsupported.</p>
VDIs per MDisk Group		Cluster limit applies.
Front-end Properties		
SAN ports	256	Maximum size of fabric, including all SAN Volume Controller nodes.

Table 18. SAN Volume Controller maximum configuration values (continued)

Objects	Maximum number	Comments
Fabrics	2	Dual fabric configurations.
Host IDs per cluster	64	A host ID is associated with a map table that associates SCSI LUNs with VDisks. It is also associated with one or more host worldwide port names.
Host ports per cluster	128	Up to 128 distinct host worldwide port names are recognized.
Host LUN size	2 TB	Defined by 32-bit LBA limit.
Virtual disks (VDisks)	4096	Includes managed-mode VDisks and image-mode VDisks.
VDisks per I/O group	1024	-- --
VDisks per host ID	512	The limit may be different based on host operating system.
VDisks-to-host mappings	20 000	-- --
Maximum persistent reservation keys	132 000	-- --
Back-end Properties		
Managed Disks (MDisks)	4096	Represents an average of 64 per world wide node name.
Back-end Storage WWNNs	64	Maximum number of device fabric world wide node name.
Back-end Storage WWPNS	256	16 ports per controller
LUs per back-end WWNN	4096	Maximum of 512 LUs presented for each world wide node name.
WWNNs per subsystem	4	-- --
WWPNs per WWNN	16	The maximum number of ports per world wide node name.
Preferred ports per subsystem	4	
Copy Services Properties		
Remote Copy relationships	4096	-- --
Remote Copy consistency groups	256	-- --

Table 18. SAN Volume Controller maximum configuration values (continued)

Objects	Maximum number	Comments
Remote Copy VDisk per I/O group	16 TB	-- --
FlashCopy mappings	2048 (See Note.)	-- --
FlashCopy consistency groups	128	-- --
FlashCopy VDisk per I/O group	16 TB	-- --
Note: SAN Volume Controller supports up to 512 FlashCopy mappings per consistency group.		

Configuration rules and requirements

Ensure that you understand the rules and requirements when configuring the SAN Volume Controller.

The following terms and definitions will guide you in understanding the rules and requirements.

- ISL hop. A hop on an Inter-Switch Link (ISL). With reference to all pairs of N-ports or end-nodes that are in a fabric, the number of ISL hops is the number of links that are crossed on the shortest route between the node pair whose nodes are farthest apart from each other. The distance is measured only in terms of the ISL links that are in the fabric.
- Oversubscription. The ratio of the sum of the traffic that is on the initiator N-node connections to the traffic that is on the most heavily-loaded ISLs, where more than one ISL is in parallel between these switches. This definition assumes a symmetrical network and a specific workload that is applied equally from all initiators and sent equally to all targets. A symmetrical network means that all initiators are connected at the same level and all the controllers are connected at the same level. The SAN Volume Controller makes this calculation difficult, because it puts its back-end traffic onto the same network, and this back-end traffic varies by workload. Therefore, the oversubscription that a 100% read hit gives is different from the oversubscription that 100% write-miss gives. If you have an oversubscription of 1 or less, the network is nonblocking.
- Virtual SAN (VSAN). A VSAN is a virtual storage area network (SAN).
- Redundant SAN. A SAN configuration in which if any one component fails, connectivity between the devices that are in the SAN is maintained, possibly with degraded performance. The way to make a redundant SAN is to split the SAN into two independent counterpart SANs.
- Counterpart SAN. A non-redundant portion of a redundant SAN. A counterpart SAN provides all the connectivity of the redundant SAN, but without the redundancy. The SAN Volume Controller is typically connected to a redundant SAN that is made out of two counterpart SANs.
- Local fabric. The fabric that consists of those SAN components (switches and cables) that connect the components (nodes, hosts, and switches) of the local

cluster. Because the SAN Volume Controller supports Remote Copy, significant distances might exist between the components of the local cluster and those of the remote cluster.

- Remote fabric. The fabric that consists of those SAN components (switches and cables) that connect the components (nodes, hosts, and switches) of the remote cluster. Because the SAN Volume Controller supports remote copy, significant distances might exist between the components of the local cluster and those of the remote cluster.
- Local/remote fabric interconnect. The SAN components that connect the local fabrics to the remote fabrics. These components might be single-mode optical fibres that are driven by Gigabit Interface Converters (GBICs), or they might be other, more advanced components, such as channel extenders.
- SAN Volume Controller fibre-channel port fan in. The number of hosts that can see any one port. Some controllers recommend that the number of hosts using each port be limited to prevent excessive queuing at that port. If the port fails, or the path to that port fails, the host might failover to another port, and the fan in requirements might be exceeded in this degraded mode.
- Invalid configuration. In an invalid configuration, an attempted operation will fail and will generate an error code to indicate what caused it to become invalid.
- Unsupported configuration. A configuration that might operate successfully, but for which IBM does not guarantee to be able to solve problems that might occur. Usually this type of configuration does not create an error log entry.
- Valid configuration. A configuration that is neither invalid nor unsupported.
- Degraded. A valid configuration that has had a failure, but continues to be neither invalid nor unsupported. Typically, a repair action is required to restore the degraded configuration to a valid configuration.

Configuration rules

SAN configurations that contain SAN Volume Controller clusters can be set up in various ways.

Some configurations do not work and are known as *invalid*. You can avoid creating invalid configurations if you follow the rules that are given in this section.

A SAN configuration that contains SAN Volume Controllers is valid if it observes *all* of the following rules. These rules are discussed in the following section.

Storage subsystems

Follow these rules when planning the configuration of storage subsystems in the SAN fabric.

All SAN Volume Controller nodes of a cluster must be able to see the same set of storage subsystem ports on each device. Any operation that is in this mode in which two nodes do not see the same set of ports on the same device is degraded, and the system logs errors that request a repair action. This rule can have important effects on storage subsystem such as FASTT, which has exclusion rules that determine to which host bus adapter (HBA) WWNNs a storage partition can be mapped.

A configuration in which a SAN Volume Controller bridges a separate host device and a RAID array is not supported. Typical compatibility matrixes are shown in a document titled *Supported Hardware List* on the following Web page:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

The SAN Volume Controller clusters must not share its storage subsystem devices with hosts. A device can be shared with a host under certain conditions as described in this topic.

Two SAN Volume Controller clusters must not share the same storage subsystem. That is, one device cannot present LUs to two different SAN Volume Controller clusters. This configuration is not supported.

The SAN Volume Controller must be configured to manage only LUNs that are presented by supported disk controller systems. Operation with other devices is not supported.

Unsupported storage subsystem (generic device)

When a storage subsystem is detected on the SAN, the SAN Volume Controller attempts to recognize it using its Inquiry data. If the device is recognized as one of the explicitly supported storage models, then the SAN Volume Controller uses error recovery programs that are potentially tailored to the known needs of the storage subsystem. If the device is not recognized, then the SAN Volume Controller configures the device as a generic device. A generic device may or may not function correctly when addressed by a SAN Volume Controller. In any event, the SAN Volume Controller does not regard accessing a generic device as an error condition and, consequently, does not log an error. MDisks presented by generic devices are not eligible to be used as quorum disks.

Split controller configurations

| If a single RAID controller presents multiple LUs, either by having multiple RAID
| arrays configured or by partitioning one or more RAID arrays into multiple LUs, then
| each LU can be owned by either SAN Volume Controller or a direct attached host.
| Suitable LUN masking must be in place to ensure that LUs are not shared between
| SAN Volume Controllers and direct attached hosts.

In a split controller configuration, a RAID array presents LUs to both a SAN Volume Controller (which treats the LU as an MDisk) and to another host. The SAN Volume Controller presents VDIs created from the MDisk to another host. There is no requirement for the pathing driver in the two hosts to be the same (although, if the RAID controller were an ESS, both hosts would use SDD). Figure 16 on page 73 shows that the RAID controller is a FASTT, with RDAC used for pathing on the directly attached host, and SDD used on the host that is attached with the SAN Volume Controller. Hosts can simultaneously access LUs, that are provided by the SAN Volume Controller and directly by the device.

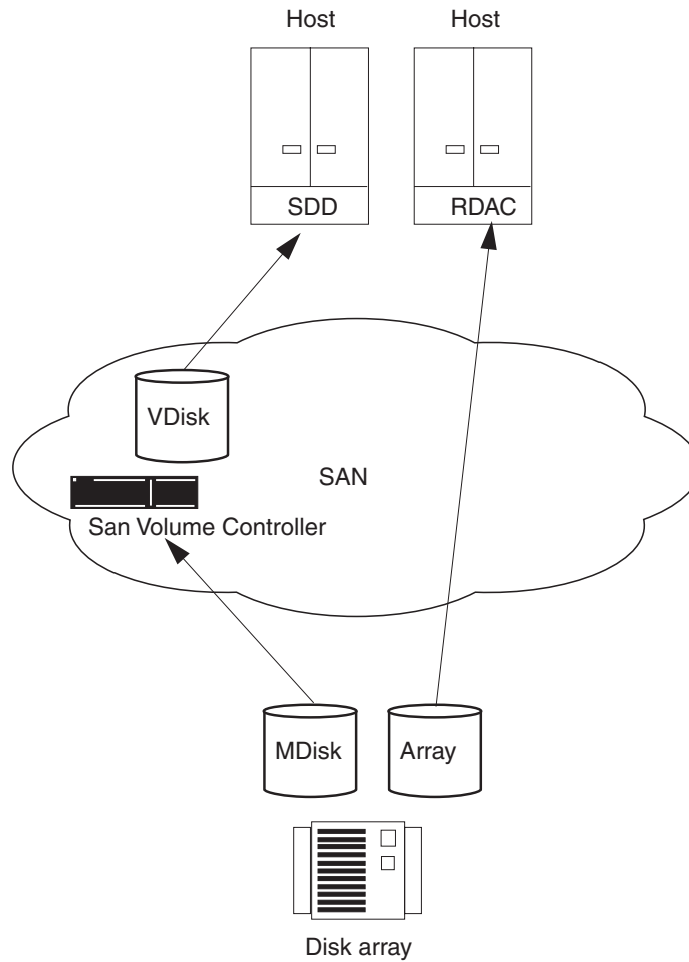


Figure 16. Disk controller system shared between SAN Volume Controller and a host

In the case where the RAID controller is an ESS, the pathing driver in the host would be IBM Subsystem Device Driver (SDD) for the ESS and SDD for the SAN Volume Controller LUs. Figure 17 on page 74 shows a supported configuration because the same pathing driver is used for both direct and virtual disks.

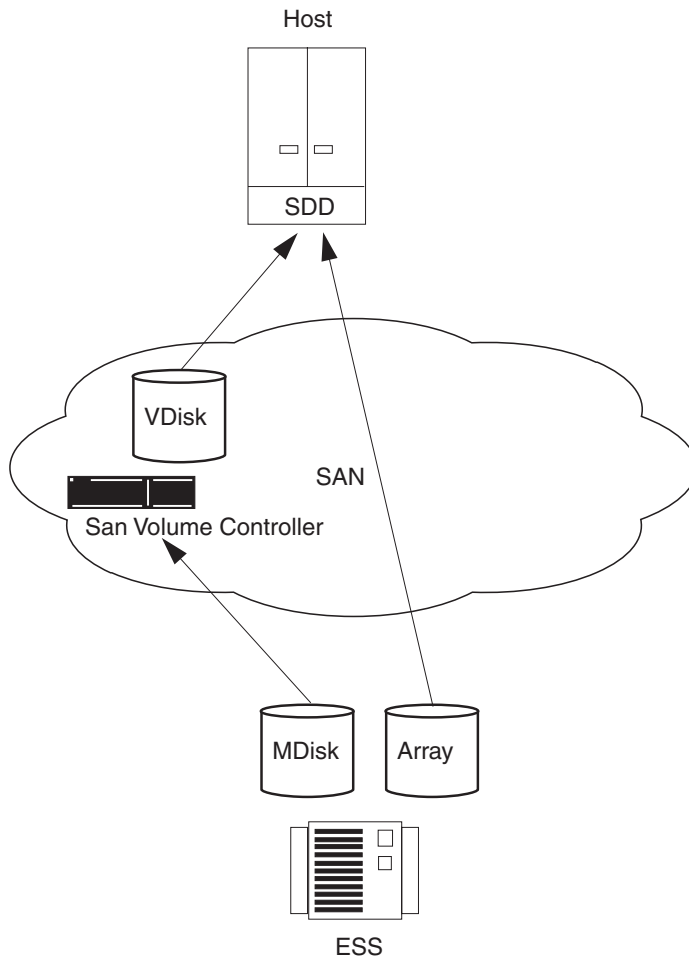


Figure 17. ESS LUs accessed directly with a SAN Volume Controller

Figure 18 on page 75 illustrates another configuration.

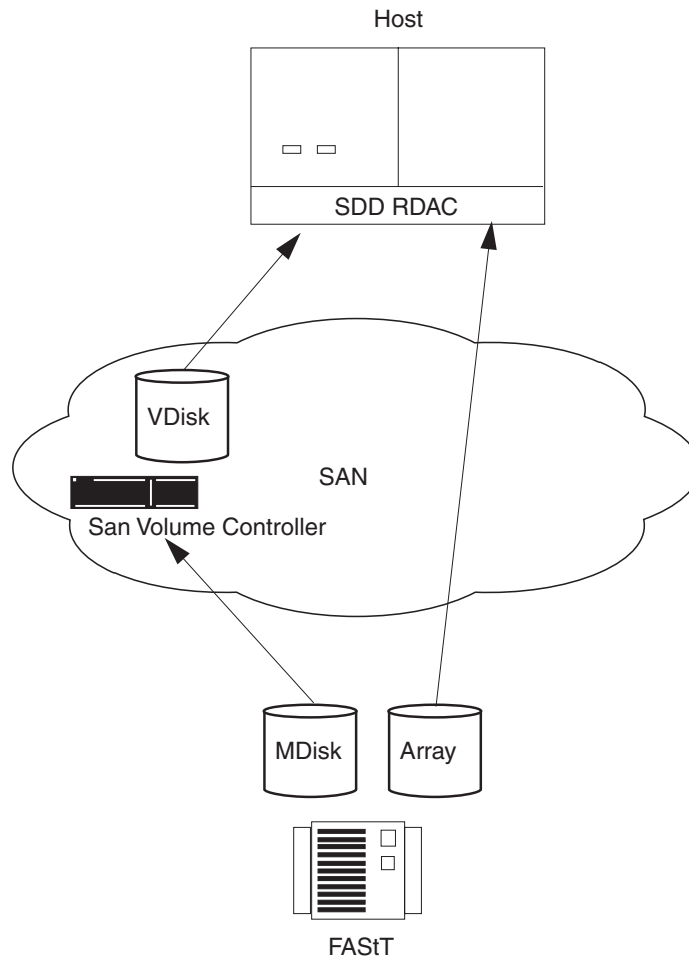


Figure 18. FAST direct connection with a SAN Volume Controller on one host

Related concepts

“Cluster operation and quorum disks” on page 39

The cluster must contain at least half of its nodes to function.

“Managed disks” on page 56

A managed disk (MDisk) is a logical disk (typically a RAID array or partition thereof) that a storage subsystem has exported to the SAN fabric to which the nodes in the cluster are attached.

Host bus adapters

Follow these configuration rules for host bus adapters (HBAs).

SAN Volume Controller nodes always contain two host bus adapters. Each HBA must present two ports. If an HBA fails, the configuration is still valid, and the node operates in degraded mode. If an HBA is physically removed from a SAN Volume Controller node, the configuration is not supported.

HBAs that are in dissimilar hosts or dissimilar HBAs that are in the same host must be in separate zones. For example, if you have an HP/UX® host and a Windows 2000 server host, those hosts must be in separate zones. Here, *dissimilar* means that the hosts are running different operating systems or that they are different hardware platforms. Different levels of the same operating system are considered to

be similar. This requirement ensures that different SANs can operate with each other. A configuration that breaks this requirement is not supported.

The SAN Volume Controller must be configured to export virtual disks only to host fibre-channel ports that are on the supported HBAs. See the following Web site for specific firmware levels and the latest supported hardware:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Operation with other HBAs is not supported.

The number of paths from the SAN Volume Controller nodes to a host must not exceed eight. The maximum number of HBA ports must not exceed four (for example, no more than two 2-port HBAs or four 1-port HBAs). Each SAN Volume Controller node in an I/O group presents four images of a virtual disk (VDisk) onto the SAN, and each host SAN attachment has up to four HBA ports. Therefore, with more simplified zoning, the number of paths can equal up to 32: 4 SAN Volume Controller ports x 2 nodes per I/O group x 4 HBA ports. If you want to restrict the number of paths to a host, the switches should be zoned so that each HBA port is zoned with one SAN Volume Controller port for each node in the cluster. If a host has multiple HBA ports, each port should be zoned to a different set of SAN Volume Controller ports to maximize performance and redundancy.

Nodes

Follow these configuration rules for nodes.

The SAN Volume Controller nodes must always be deployed in pairs. If a node fails or is removed from the configuration, the remaining node operates in a degraded mode, but the configuration is still valid.

The uninterruptible power supply must be in the same rack as the nodes. When using 6 or 8 node support, ensure that 4 uninterruptible power supplies are used. Ensure that you are following the uninterruptible power supply support guidelines as described below:

Number of nodes	Number of uninterruptible power supplies
2	2
4	2
6	4
8	4

Support for optical connections is based on the fabric rules that the manufacturers impose for the following connection methods:

- Node to a switch
- Host to a switch
- Backend to a switch
- Switch to an Inter-Switch Link

For the SAN Volume Controller, the following optical connections are supported:

- Shortwave optical fibre
- Longwave optical fibre up to 10 KM

High-power Gigabit Interface Converters (GBICs) and longwave fibre connections beyond 10 KM are not supported.

To ensure cluster failover operations, all nodes in a cluster must be connected to the same IP subnet.

Power requirements

Note the power requirements for the SAN Volume Controller.

The uninterruptible power supply must be in the same rack that contains the SAN Volume Controller nodes that it supplies. The combination power and signal cable for connection between SAN Volume Controller and uninterruptible power supply units is two meters long. The SAN Volume Controller and uninterruptible power supply must connect with both the power and the signal cable to function correctly.

Fibre-channel switches

Follow these guidelines for configuring the fibre-channel switches that are supported on the SAN.

The SAN must contain only supported switches. The SAN Volume Controller supports specific IBM 2109, McData, and InRange switch models and the Cisco MDS 9000 switch and switches supported by the Cisco MDS 9000.

See the following Web site for specific firmware levels and the latest supported hardware:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Operation with other switches is not supported.

Different vendor switches cannot be intermixed in the same counterpart SAN. A redundant SAN, made up of more than one counterpart SAN can contain different vendor switches, provided the same vendor is used within each counterpart SAN.

The SAN must consist of two independent switches (or networks of switches) so that the SAN includes a redundant fabric, and has no single point of failure. If one SAN fabric fails, the configuration is in a degraded mode, but it is still valid. If the SAN contains only one fabric, it is still a valid configuration, but a failure of the fabric might cause a loss of access to data. Such a SAN, therefore, is seen as a single point of failure.

Configurations with more than two SANs are not supported.

On the fibre-channel SAN, the SAN Volume Controller nodes must always and only be connected to SAN switches. Each node must be connected to each of the counterpart SANs that are in the redundant fabric. Any operation that uses direct connections between host and node, or controller and node, is not supported.

On the fibre-channel SAN, back-end storage must always and only be connected to SAN switches. Multiple connections are permitted from the redundant controllers of the back-end storage, to improve data bandwidth performance. It is not necessary to have a connection between each redundant disk controller system of the back-end storage and each counterpart SAN. For example, in a FAStT configuration in which the FAStT contains two redundant controllers, only two controller minihubs are usually used. Controller A of the FAStT is, therefore, connected to counterpart

SAN A, and controller B of the of the FASiT is connected to counterpart SAN B. Any operation that uses a direct connection between the host and the controller is not supported.

The connections between the switches and the SAN Volume Controllers can operate at 1 Gbps or at 2 Gbps. All the ports for the SAN Volume Controllers that are in a single cluster, however, must run at one speed. Any operation that runs different speeds on the node-to-switch connections that are in a single cluster is not valid.

Attention: The default transfer rate in the SAN Volume Controller is 2 Gbps. If your environment is set up to use 1 Gbps switches, the switch rate must be set at the transfer rate.

Mixed speeds are permitted in the fabric. Lower speeds can be used to extend distances or to make use of 1 Gbps legacy components.

The switch configuration of a SAN Volume Controller SAN must observe the switch manufacturer's configuration rules. These rules might put restrictions on the switch configuration; for example, the switch manufacturer might not permit other manufacturer's switches to be in the SAN. Any operations that run outside the manufacturer's rules is not supported.

The switch must be configured so that the SAN Volume Controller nodes can see the back-end storage and the front-end HBAs. However, the front-end HBAs and the back-end storage must not be in the same zone. Any operation that runs outside these zoning rules is not supported.

Because each SAN Volume Controller has four ports, the switches can be zoned so that a particular SAN Volume Controller port is used only for internode communication, for communication to the host, or for communication to back-end storage. Whatever the configuration, each SAN Volume Controller node must remain connected to the full SAN fabric. Zoning must not be used to split the SAN into two parts.

With Remote Copy, additional zones are required that contain only the local nodes and the remote nodes. It is valid for the local hosts to see the remote nodes, or for the remote hosts to see the local nodes. Any zone that contains the local and the remote back-end storage and local nodes or remote nodes, or both, is not valid.

Fibre-channel switches and Inter-Switch Links

The local or remote fabric must not contain more than three Inter-Switch Links (ISLs) hops in each fabric. Any operation that uses more than three ISLs is not supported. When a local fabric is connected to a remote fabric for Remote Copy purposes, the ISL count between a local node and a remote node must not exceed seven. Therefore, some ISLs can be used in a cascaded switch link between local and remote clusters if the internal ISL count of the local or remote cluster is less than three.

The local and remote fabric interconnections must be only one ISL hop between a switch that is in the local fabric and a switch that is in a remote fabric. That is, it must be a single-mode fibre up to 10 KM (32 810 ft.) long. Any operation that uses other local or remote fabric interconnections is not supported.

When ISLs are used, each ISL oversubscription must not exceed six. Any operation that uses higher values is not supported.

With Inter-Switch Links between nodes in the same cluster, the ISLs are considered a single point of failure. This is illustrated in Figure 19.

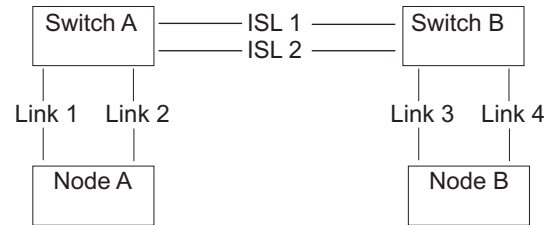


Figure 19. Fabric with Inter-Switch Links between nodes in a cluster

If Link 1 or Link 2 fails, the cluster communication does not fail.

If Link 3 or Link 4 fails, the cluster communication does not fail.

If ISL 1 or ISL 2 fails, the communication between Node A and Node B will fail for a period of time, and the node will not be recognized, even though there is still a connection between the nodes.

To ensure that a fibre-channel link failure does not cause nodes to fail when there are ISLs between nodes, it is necessary to use a redundant configuration. This is illustrated in Figure 20.

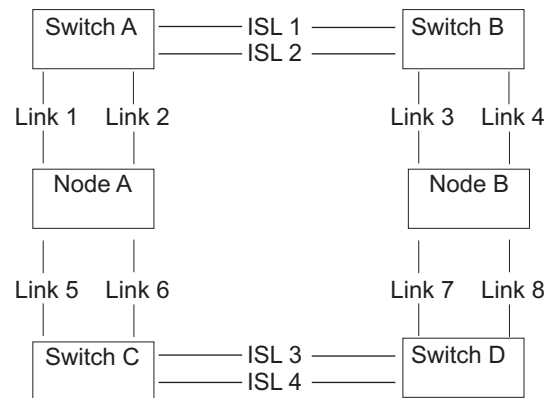


Figure 20. Fabric with Inter-Switch Links in a redundant configuration

With a redundant configuration, if any one of the links fails, then communication on the cluster will not fail.

Configuration requirements

You *must* perform these steps before you configure the SAN Volume Controller.

1. Your IBM service representative must have installed the SAN Volume Controller.
2. Install and configure your disk controller systems and create the RAID resources that you intend to virtualize. To prevent loss of data, virtualize only those RAID that provide some kind of redundancy, that is, RAID 1, RAID 10, RAID 0+1, or RAID 5. Do *not* use RAID 0 because a single physical disk failure might cause the failure of many virtual disks. RAID 0, like other types of RAID

offers cost-effective performance by using available capacity through data striping. However, RAID 0 does not provide a parity disk drive for redundancy (RAID 5) or mirroring (RAID 10).

When creating RAID with parity protection (for example, RAID 5), consider how many component disks to use in each array. The more disks you use, the fewer disks you need to provide availability for the same total capacity (one per array). However, if you use more disks, it will take longer to rebuild a replacement disk after a disk failure. If a second disk failure occurs during the rebuild period, all data on the array is lost. More data is affected by a disk failure for a larger number of member disks resulting in reduced performance while rebuilding onto a hot spare and more data being exposed if a second disk fails before the rebuild has completed. The smaller the number of disks, the more likely it is that write operations span an entire stripe (strip size x number of members minus 1). In this case, write performance is improved because the disk write operations do not have to be preceded by disk reads. The number of disk drives required to provide availability might be unacceptable if the arrays are too small.

When in doubt, create arrays with between six and eight member disks.

If reasonably small RAID arrays are used, it is easier to extend an MDisk group by adding a new RAID array of the same type. Construct multiple RAID devices of the same type, when possible.

When creating RAID with mirroring, the number of component disks in each array does not affect redundancy or performance.

Most back-end disk controller systems enable RAID to be divided up into more than one SCSI logical unit (LU). When configuring new storage for use with the SAN Volume Controller, you do not need to divide up the array. New storage should be presented as one SCSI LU. This will give a one-to-one relationship between MDisks and RAID.

Attention: Losing an array in an MDisk group can result in the loss of access to *all* MDisks in that group.

3. Install and configure your switches to create the zones that the SAN Volume Controller requires. One zone must contain all the disk controller systems and the SAN Volume Controller nodes. For hosts with more than one port, use switch zoning to ensure that each host fibre-channel port is zoned to exactly one fibre-channel port of each SAN Volume Controller node in the cluster. Set up a zone on each fibre-channel switch that includes the master console and all of the SAN Volume Controller ports that are connected to that switch.
4. If you want the SAN Volume Controller to export redundant paths to disks, you must install the Subsystem Device Driver (SDD) on all of the hosts that are connected to the SAN Volume Controller. Otherwise, you will not be able to use the redundancy inherent in the configuration. Install the SDD from the following Web site:
<http://www-1.ibm.com/server/storage/support/software/sdd.html>
Version 1.4.x.x or later is required.
5. Install and configure the SAN Volume Controller master console. The communication between the master console and the SAN Volume Controller runs under a client-server network application called Secure Shell (SSH). Each SAN Volume Controller cluster is equipped with SSH Server software and the master console comes to you equipped with the SSH Client software called PuTTY. You will need to configure the SSH client key pair using PuTTY on the master console. Once you have installed your master console, you can configure and administer the SAN Volume Controller using a graphical interface or a command-line interface.

- a. You can configure the SAN Volume Controller using the SAN Volume Controller Console Web-based application that is preinstalled on the master console.

Note: You can also install the master console on another machine (which you provide) using the CD-ROM provided with the master console.

- b. You can configure the SAN Volume Controller using the command-line interface (CLI) commands.
- c. You can install an SSH client if you only want to use the CLI commands. If you want to use the CLI from a host other than the master console, ensure that the host has an SSH client installed on it.

Note:

- 1) AIX comes with an installed SSH client.
- 2) Linux comes with an installed SSH client.
- 3) PuTTY is recommended for Windows.

When you and the IBM service representative have completed the initial preparation steps, you must perform the following steps:

1. Add nodes to the cluster and set up the cluster properties.
2. Create managed disk groups from the managed disks to make pools of storage from which you can create virtual disks.
3. Create host objects from the HBA fibre-channel ports to which you can map virtual disks.
4. Create virtual disks from the capacity that is available in your managed disk groups.
5. Map the virtual disks to the host objects to make the disks available to the hosts, as required.
6. Optionally, create Copy Services (FlashCopy and Remote Copy) objects as required.

Related concepts

“Managed disk groups” on page 58

An *MDisk group* is a collection of MDisks that jointly contain all the data for a specified set of virtual disks (VDisks).

Related reference

“Fibre-channel switches” on page 77

Follow these guidelines for configuring the fibre-channel switches that are supported on the SAN.

Chapter 7. SAN Volume Controller supported environment

The IBM Web site provides up-to-date information about the supported environment for the SAN Volume Controller.

This includes:

- Host attachments
- Physical disk storage systems
- Host bus adapters
- Switches

See the following Web site for specific firmware levels and the latest supported hardware:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Supported host attachments

The IBM Web site provides up-to-date information about the supported host attachment operating systems.

For a list of supported host attachment operating systems, see the SAN Volume Controller Web site at:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

The SAN Volume Controller provides heterogeneous host attachments so that you can consolidate storage capacity and workloads for open systems hosts. The SAN Volume Controller supports a maximum of 64 separate hosts and a maximum of 128 host fibre-channel ports, identified by their worldwide port numbers (WWPNs).

Hosts are attached to the SAN Volume Controller using a switched, fibre-channel fabric.

Supported storage subsystems

The IBM Web site provides up-to-date information about the supported physical disk storage systems.

For a list of supported storage systems, see the SAN Volume Controller Web site:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>.

Supported fibre-channel host bus adapters

The IBM Web site provides up-to-date information about the supported host bus adapters.

Ensure that host bus adapters (HBAs) are at or above the minimum requirements.

For a list of supported HBAs, see the following Web site for specific firmware levels and the latest supported hardware:

Supported switches

The IBM Web site provides up-to-date information about the supported fibre-channel switches.

Ensure that switches are at or above the minimum requirements.

The SAN must contain only supported switches.

See the following Web site for the latest models and firmware levels:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Operation with other switches is not supported.

Supported fibre-channel extenders

The SAN Volume Controller supports the CNT UltraNet Edge Storage Router to support synchronous copy services.

The maximum one-way latency supported is 10 microseconds when using Brocade fabric, and 34 microseconds when using McData fabric. The relationship between latency and distance is dependent on the network and number of hops. The distance is approximately 100-150 kilometers per microsecond.

Note: Performance of copy services degrades as the distance increases.

See the following Web site for the latest supported hardware:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Accessibility

Accessibility features help a user who has a physical disability, such as restricted mobility or limited vision, to use software products successfully.

Features

These are the major accessibility features in the SAN Volume Controller master console:

- You can use screen-reader software and a digital speech synthesizer to hear what is displayed on the screen. The following screen readers have been tested: JAWS v4.5 and IBM Home Page Reader v3.0.
- You can operate all features using the keyboard instead of the mouse.

Navigating by keyboard

You can use keys or key combinations to perform operations and initiate many menu actions that can also be done through mouse actions. You can navigate the SAN Volume Controller Console and help system from the keyboard by using the following key combinations:

- To traverse to the next link, button, or topic, press Tab inside a frame (page).
- To expand or collapse a tree node, press → or ←, respectively.
- To move to the next topic node, press V or Tab.
- To move to the previous topic node, press ^ or Shift+Tab.
- To scroll all the way up or down, press Home or End, respectively.
- To go back, press Alt+←.
- To go forward, press Alt+→.
- To go to the next frame, press Ctrl+Tab.
- To move to the previous frame, press Shift+Ctrl+Tab.
- To print the current page or active frame, press Ctrl+P.
- To select, press Enter.

Accessing the publications

You can view the publications for the SAN Volume Controller in Adobe Portable Document Format (PDF) using the Adobe Acrobat Reader. The PDFs are provided on a CD that is packaged with the product or you can access them at the following Web site:

<http://www-1.ibm.com/servers/storage/support/virtual/2145.html>

Related reference

“SAN Volume Controller library and related publications” on page x
A list of other publications that are related to this product are provided to you for your reference.

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 give 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 PUBLICATIONS "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 Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites 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 measurement 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.

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

This information is for planning purposes only. The information herein is subject to change before the products described become available.

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.

Trademarks

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

- AIX
- e (logo)
- Enterprise Storage Server
- FlashCopy
- IBM
- Tivoli
- TotalStorage
- xSeries

Intel and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

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

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

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

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

Definitions of notices

Ensure that you understand the typographic conventions that are used in this publication to indicate special notices.

The following notices are used throughout this library to convey the following specific meanings:

Note: These notices provide important tips, guidance, or advice.

Attention: These notices indicate possible damage to programs, devices, or data. An attention notice appears before the instruction or situation in which damage could occur.

CAUTION:

These notices indicate situations that can be potentially hazardous to you. A caution notice precedes the description of a potentially hazardous procedural step or situation.

DANGER

These notices indicate situations that can be potentially lethal or extremely hazardous to you. A danger notice precedes the description of a potentially lethal or extremely hazardous procedural step or situation.

Glossary

Ensure you are familiar with the list of terms and their definitions used in this guide.

A

asymmetric virtualization

A virtualization technique in which the virtualization engine is outside the data path and performs a metadata-style service. The metadata server contains all the mapping and locking tables while the storage devices contain only data. See also *symmetric virtualization*

auxiliary virtual disk

The virtual disk that contains a backup copy of the data and that is used in disaster recovery scenarios. See also *master virtual disk*.

B

blade One component in a system that is designed to accept some number of components (blades). Blades could be individual servers that plug into a multiprocessing system or individual port cards that add connectivity to a switch. A blade is typically a hot-swappable hardware device.

block A unit of data storage on a disk drive.

block virtualization

The act of applying virtualization to one or more block-based (storage) services for the purpose of providing a new aggregated, higher-level, richer, simpler, or secure block service to clients. Block virtualization functions can be nested. A disk drive, RAID system, or volume manager all perform some form of block-address to (different) block-address mapping or aggregation. See also *virtualization*.

C

cache A high-speed memory or storage device used to reduce the effective time required to read data from or write data to lower-speed memory or a device. Read cache holds data in anticipation that it will be requested by a client. Write cache holds data written by a client until it can be safely stored on more permanent storage media such as disk or tape.

cascading

The process of connecting two or more fibre-channel hubs or switches together to increase the number of ports or extend distances.

cluster

In SAN Volume Controller, a pair of nodes that provides a single configuration and service interface.

CIM See *Common Information Model*.

CLI See *command line interface*.

command line-interface (CLI)

A type of computer interface in which the input command is a string of text characters.

Common Information Model (CIM)

A set of standards developed by the Distributed Management Task Force (DMTF). CIM provides a conceptual framework for storage management

and an open approach to the design and implementation of storage systems, applications, databases, networks, and devices.

connected

In a Remote Copy relationship, pertaining to the status condition that occurs when two clusters can communicate.

consistency group

A group of copy relationships between virtual disks that are managed as a single entity.

consistent copy

In a Remote Copy relationship, a copy of a secondary virtual disk (VDisk) that is identical to the primary VDisk from the viewpoint of a host system, even if a power failure occurred while I/O activity was in progress.

copied

In a FlashCopy relationship, a state that indicates that a copy has been started after the copy relationship was created. The copy process is complete and the target disk has no further dependence on the source disk.

Copy Services

In the SAN Volume Controller, the two services that enable you to copy virtual disks (VDisks): FlashCopy and Remote Copy.

copying

A status condition that describes the state of a pair of virtual disks (VDisks) that have a copy relationship. The copy process has been started but the two virtual disks are not yet synchronized.

counterpart SAN

A nonredundant portion of a redundant storage area network (SAN). A counterpart SAN provides all the connectivity of the redundant SAN but without the redundancy. Each counterpart SANs provides an alternate path for each SAN-attached device. See also *redundant SAN*.

cross-volume consistency

In SAN Volume Controller, a consistency group property that guarantees consistency between virtual disks when an application issues dependent write operations that span multiple virtual disks.

D

data migration

The movement of data from one physical location to another without disrupting I/O operations.

destage

A write command initiated by the cache to flush data to disk storage.

disk controller

A device that coordinates and controls the operation of one or more disk drives and synchronizes the operation of the drives with the operation of the system as a whole. Disk controllers provide the storage that the cluster detects as managed disks (MDisks).

E

error code

A value that identifies an error condition.

excluded

In SAN Volume Controller, the status of a managed disk that the cluster has removed from use after repeated access errors.

extent A unit of data that manages the mapping of data between managed disks and virtual disks.

F

fabric In fibre-channel technology, a routing structure, such as a switch, that receives addressed information and routes it to the appropriate destination. A fabric can consist of more than one switch. When multiple fibre-channel switches are interconnected, they are described as cascading. See also *cascading*.

failover

In SAN Volume Controller, the function that occurs when one redundant part of the system takes over the workload of another part of the system that has failed.

fibre channel

A technology for transmitting data between computer devices at a data rate of up to 4 Gbps. It is especially suited for attaching computer servers to shared storage devices and for interconnecting storage controllers and drives.

FlashCopy mapping

A relationship between two virtual disks.

FlashCopy relationship

See *FlashCopy mapping*.

FlashCopy service

In SAN Volume Controller, a copy service that duplicates the contents of a source virtual disk (VDisk) to a target VDisk. In the process, the original contents of the target VDisk are lost. See also *point-in-time copy*.

H

HBA See *host bus adapter*.

host An open-systems computer that is connected to the SAN Volume Controller through a fibre-channel interface.

host bus adapter (HBA)

In SAN Volume Controller, an interface card that connects a host bus, such as a peripheral component interconnect (PCI) bus, to the storage area network.

host ID

In SAN Volume Controller, a numeric identifier assigned to a group of host fibre-channel ports for the purpose of logical unit number (LUN) mapping. For each host ID, there is a separate mapping of Small Computer System Interface (SCSI) IDs to virtual disks (VDisks).

hub A communications infrastructure device to which nodes on a multi-point bus or loop are physically connected. Commonly used in Ethernet and fibre-channel networks to improve the manageability of physical cables. Hubs maintain the logical loop topology of the network of which they are a part, while creating a “hub and spoke” physical star layout. Unlike switches,

hubs do not aggregate bandwidth. Hubs typically support the addition or removal of nodes from the bus while it is operating. (S) Contrast with *switch*.

I

IBM Subsystem Device Driver (SDD)

An IBM pseudo device driver designed to support the multipath configuration environments in IBM products.

idling The status of a pair of virtual disks (VDisks) that have a defined copy relationship for which no copy activity has yet been started.

image mode

An access mode that establishes a one-to-one mapping of extents in the managed disk (MDisk) with the extents in the virtual disk (VDisk). See also *managed space mode* and *unconfigured mode*.

I/O group

A collection of virtual disks (VDisks) and node relationships that present a common interface to host systems.

inconsistent

In a Remote Copy relationship, pertaining to a secondary virtual disk (VDisk) that is being synchronized with the primary VDisk.

input/output (I/O)

Pertaining to a functional unit or communication path involved in an input process, an output process, or both, concurrently or not, and to the data involved in such a process.

Internet Protocol (IP)

In the Internet suite of protocols, a connectionless protocol that routes data through a network or interconnected networks and acts as an intermediary between the higher protocol layers and the physical network.

interoperability

The capability to communicate, run programs, or transfer data among various functional units in a way that requires the user to have little or no knowledge of the unique characteristics of those units.

Inter-Switch Link (ISL)

A protocol for interconnecting multiple routers and switches in a storage area network.

ISL See *Inter-Switch Link*.

ISL hop

Considering all pairs of node ports (N-ports) in a fabric and measuring distance only in terms of Inter-Switch Links (ISLs) in the fabric, the number of ISLs traversed is the number of ISL hops on the shortest route between the pair of nodes that are farthest apart in the fabric.

IP See *Internet Protocol*.

I/O See *input/output*.

J

JBOD (just a bunch of disks)

IBM definition: See *non-RAID*. HP definition: A group of single-device logical units not configured into any other container type.

L

line card

See *blade*.

local fabric

In SAN Volume Controller, those storage area network (SAN) components (such as switches and cables) that connect the components (nodes, hosts, switches) of the local cluster together.

logical unit (LU)

An entity to which Small Computer System Interface (SCSI) commands are addressed, such as a virtual disk (VDisk) or managed disk (MDisk).

logical unit number (LUN)

The SCSI identifier of a logical unit within a target. (S)

LU See *logical unit*.

LUN See *logical unit number*.

M

managed disk (MDisk)

A Small Computer System Interface (SCSI) logical unit that a redundant array of independent disks (RAID) controller provides and a cluster manages. The MDisk is not visible to host systems on the storage area network (SAN).

managed disk group

A collection of managed disks (MDisks) that, as a unit, contain all the data for a specified set of virtual disks (VDisks).

managed space mode

An access mode that enables virtualization functions to be performed. See also *image mode* and *unconfigured mode*.

Management Information Base (MIB)

Simple Network Management Protocol (SNMP) units of managed information that specifically describe an aspect of a system, such as the system name, hardware number, or communications configuration. A collection of related MIB objects is defined as a MIB.

mapping

See *FlashCopy mapping*.

master virtual disk

The virtual disk (VDisk) that contains a production copy of the data and that an application accesses. See also *auxiliary virtual disk*.

MDisk See *managed disk*.

migration

See *data migration*.

N

node One SAN Volume Controller. Each node provides virtualization, cache, and Copy Services to the storage area network (SAN).

node rescue

In SAN Volume Controller, the process by which a node that has no valid software installed on its hard disk drive can copy the software from another node connected to the same fibre-channel fabric.

non-RAID

Disks that are not in a redundant array of independent disks (RAID). IBM definition: Disks that are not in a redundant array of independent disks (RAID). HP definition: See *JBOD*.

O

offline Pertaining to the operation of a functional unit or device that is not under the continual control of the system or of a host.

online Pertaining to the operation of a functional unit or device that is under the continual control of the system or of a host.

P**point-in-time copy**

The instantaneous copy that the FlashCopy service makes of the source virtual disk (VDisk). In some contexts, this copy is known as a *T₀ copy*.

port ID

An identifier associated with a port.

primary virtual disk

In a Remote Copy relationship, the target of write operations issued by the host application.

PuTTY

A free implementation of Telnet and SSH for Windows 32-bit platforms

Q**quorum disk**

A managed disk (MDisk) that contains quorum data and that a cluster uses to break a tie and achieve a quorum.

R

rack A free-standing framework that holds the devices and card enclosure.

RAID See *redundant array of independent disks*.

redundant array of independent disks

A collection of two or more disk drives that present the image of a single disk drive to the system. In the event of a single device failure, the data can be read or regenerated from the other disk drives in the array.

redundant SAN

A storage area network (SAN) configuration in which any one single component might fail, but connectivity between the devices within the SAN is maintained, possibly with degraded performance. This configuration is normally achieved by splitting the SAN into two, independent, counterpart SANs. See also *counterpart SAN*.

relationship

In Remote Copy, the association between a master virtual disk (VDisk) and an auxiliary VDisk. These VDIsks also have the attributes of a primary or secondary VDisk. See also *auxiliary virtual disk*, *master virtual disk*, *primary virtual disk*, and *secondary virtual disk*.

Remote Copy

In SAN Volume Controller, a copy service that enables host data on a particular source virtual disk (VDisk) to be copied to the target VDisk designated in the relationship.

remote fabric

In Remote Copy, the storage area network (SAN) components (switches and cables) that connect the components (nodes, hosts, and switches) of the remote cluster.

roles Authorization is based on roles that map to the administrator and service roles in an installation. The switch translates these roles into SAN Volume Controller administrator and service user IDs when a connection is made to the node for the SAN Volume Controller.

S

SAN See *storage area network*.

SDD See *IBM Subsystem Device Driver*.

secondary virtual disk

In Remote Copy, the virtual disk (VDisk) in a relationship that contains a copy of data written by the host application to the primary VDisk.

Simple Network Management Protocol (SNMP)

In the Internet suite of protocols, a network management protocol that is used to monitor routers and attached networks. SNMP is an application-layer protocol. Information on devices managed is defined and stored in the application's Management Information Base (MIB).

SNMP See *Simple Network Management Protocol*.

storage area network (SAN)

A network whose primary purpose is the transfer of data between computer systems and storage elements and among storage elements. A SAN consists of a communication infrastructure, which provides physical connections, and a management layer, which organizes the connections, storage elements, and computer systems so that data transfer is secure and robust. (S)

switch

A network infrastructure component to which multiple nodes attach. Unlike hubs, switches typically have internal bandwidth that is a multiple of link bandwidth, and the ability to rapidly switch node connections from one to another. A typical switch can accommodate several simultaneous full link bandwidth transmissions between different pairs of nodes. (S) Contrast with *hub*.

symmetric virtualization

A virtualization technique in which the physical storage in the form of Redundant Array of Independent Disks (RAID) is split into smaller chunks of storage known as *extents*. These extents are then concatenated, using various policies, to make virtual disks (VDisks). See also *asymmetric virtualization*.

synchronized

In Remote Copy, the status condition that exists when both virtual disks (VDisks) of a pair that has a copy relationship contain the same data.

U

uninterruptible power supply

A device connected between a computer and its power source that protects the computer against blackouts, brownouts, and power surges. The uninterruptible power supply contains a power sensor to monitor the supply and a battery to provide power until an orderly shutdown of the system can be performed.

unconfigured mode

A mode in which I/O operations cannot be performed. See also *image mode* and *managed space mode*.

V

valid configuration

A configuration that is supported.

VDisk See *virtual disk*.

virtual disk (VDisk)

In SAN Volume Controller, a device that host systems attached to the storage area network (SAN) recognize as a Small Computer System Interface (SCSI) disk.

virtual storage area network (VSAN)

A fabric within the SAN.

virtualization

In the storage industry, a concept in which a pool of storage is created that contains several disk subsystems. The subsystems can be from various vendors. The pool can be split into virtual disks that are visible to the host systems that use them.

virtualized storage

Physical storage that has virtualization techniques applied to it by a virtualization engine.

VLUN See *virtual disk*.

VSAN See *virtual storage area network*.

W

worldwide node name (WWNN)

An identifier for an object that is globally unique. WWNNs are used by Fibre Channel and other standards.

WWNN

See *worldwide node name*.

WWPN

See *worldwide port name*.

worldwide port name (WWPN)

A unique 64-bit identifier associated with a fibre-channel adapter port. The WWPN is assigned in an implementation- and protocol-independent manner.

Index

A

- accessibility
 - keyboard 85
 - shortcut keys 85
- adapters
 - fibre channel 83

C

- cable connection table
 - example 27
- charts and tables 23
 - cable connection table 26, 27
 - configuration data table 28, 29
 - hardware location chart 23, 24, 25
- Cisco Systems
 - MDS 9000 Caching Services Module 5
 - MDS 9000 switch 5
- cluster state 39
- clusters
 - operation 39
 - overview 38
- configuration
 - maximum sizes 68
 - rules 70
- configuration requirements 79
- configuring
 - switches 77
- connections 20
- consistency group, Remote Copy 50
- consistency groups, FlashCopy 48
- console
 - master
 - overview 11
 - physical characteristics 19
- conventions
 - emphasis in text x
- copy services
 - overview 44

D

- disk controllers
 - overview 55

E

- emphasis in text x

F

- FlashCopy
 - consistency groups 48
 - mappings 45
 - overview 13, 45

H

- HBAs (host bus adapters)
 - configuration 75
- host bus adapters (HBAs)
 - configuration 75
- hosts 83
 - overview 64

I

- I/O groups 40
- information
 - center x
- installation
 - planning 17, 23, 31

K

- keyboard 85
 - shortcut keys 85

M

- migration 43

N

- node status 37
- nodes
 - configuration 76
- notices 89
 - legal 87

O

- object descriptions 53
- operating over long distances 36
- ordering publications xii
- overview
 - disk controllers 43
 - zoning 32

P

- physical characteristics
 - master console 19
 - uninterruptible power supply 18
- planning
 - configuration 67
 - installation 17, 23, 31
- ports 20
- power
 - SAN Volume Controller
 - requirements 17
 - power requirements 77

publications
ordering xii

R

related information x
Remote Copy
 overview 14, 49, 50
 zoning considerations 35
requirements
 ac voltage 17
 electrical 17
 power 17

S

safety
 caution notices 89
 danger notices 89
SAN Volume Controller
 air temperature 17
 dimensions and weight 17, 19
 heat output 17
 humidity 17
 overview 5
 product characteristics 17
 specifications 17
 weight and dimensions 17
SANs (storage area networks) 31
shortcut keys 85
site requirements
 connections 20
 ports 20
specifications
 SAN Volume Controller 19
status
 of cluster 39
 of node 37
storage
 devices
 supported 83
storage area networks (SAN) 31
support
 Web sites xii
switches
 operating over long distances 36
 supported 84
synchronous copy
 overview 50

T

text emphasis x
trademarks 88

U

uninterruptible power supply
 environment 18
 overview 8, 41

V

virtual disk-to-host mapping
 description 63
virtual disks (VDisks)
 modes 44
virtualization
 asymmetric 3
 overview 1
 symmetric 4

W

Web sites xii

Z

zoning
 considerations for Remote Copy 35
 overview 32

Readers' Comments — We'd Like to Hear from You

**IBM TotalStorage SAN Volume Controller
Planning Guide
Version 1.2.1**

Publication No. GA22-1052-03

We appreciate your comments about this publication. Feel free to comment on specific errors or omissions, accuracy, organisation, subject matter, or completeness of this book. The comments you send should pertain to only the information in this manual and the way in which the information is presented.

For technical questions and information about products and prices, please contact your IBM branch office, your IBM business partner, or your authorized remarketer.

For general questions, please call "Hello IBM" (phone number 01803/313233).

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you.

Comments:

Thank you for your support.

To submit your comments:

- Send your comments to the address on the reverse side of this form.

If you would like a response from IBM, please fill in the following information:

Name

Address

Company or Organization

Phone No.

E-mail address



Fold and Tape

Please do not staple

Fold and Tape



NO POSTAGE
NECESSARY
IF MAILED IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST-CLASS MAIL PERMIT NO. 40 ARMONK, NEW YORK

POSTAGE WILL BE PAID BY ADDRESSEE

International Business Machines Corporation
Information Development
Department 61C
9032 South Rita Road
Tucson, Arizona
USA 85775-4401



Fold and Tape

Please do not staple

Fold and Tape



Printed in USA

GA22-1052-03



Spine information:



IBM TotalStorage SAN Volume
Controller

SAN Volume Controller Planning Guide

Version 1.2.1