

Sun StorEdge Volume Manager Administration

ES-310

Student Guide



Sun Microsystems, Inc.
MS BRM01-209
500 Eldorado Boulevard
Broomfield, Colorado 80021
U.S.A.

Revision A, October 1999

Copyright 2000 Sun Microsystems, Inc., 901 San Antonio Road, Palo Alto, California 94303, U.S.A. All rights reserved.

This product or document is protected by copyright and distributed under licenses restricting its use, copying, distribution, and decompilation. No part of this product or document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any.

Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the U.S. and other countries, exclusively licensed through X/Open Company, Ltd.

Sun, Sun Microsystems, the Sun Logo, Solaris, StorEdge Volume Manager, Ultra, Answerbook, Java, NFS, Solstice DiskSuite, and OpenBoot are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and other countries.

All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

UNIX is a registered trademark in the U.S. and other countries, exclusively licensed through X/Open Company, Ltd.

The OPEN LOOK and Sun Graphical User Interface was developed by Sun Microsystems, Inc. for its users and licensees. Sun acknowledges the pioneering efforts of Xerox in researching and developing the concept of visual or graphical user interfaces for the computer industry. Sun holds a non-exclusive license from Xerox to the Xerox Graphical User Interface, which license also covers Sun's licensees who implement OPEN LOOK GUIs and otherwise comply with Sun's written license agreements.

U.S. Government approval required when exporting the product.

RESTRICTED RIGHTS: Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g) (2)(6/87) and FAR 52.227-19(6/87), or DFAR 252.227-7015 (b)(6/95) and DFAR 227.7202-3(a).

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS, AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.



Please
Recycle



Adobe PostScript

Contents

About This Course	xiii
Course Overview	xiv
Course Map.....	xv
Module-by-Module Overview	xvi
Course Objectives.....	xix
Skills Gained by Module.....	xx
Guidelines for Module Pacing	xxi
Topics Not Covered.....	xxii
How Prepared Are You?.....	xxiii
Introductions	xxiv
How to Use Course Materials	xxv
Course Icons and Typographical Conventions	xxvii
Icons	xxvii
Typographical Conventions	xxviii
Sun Storage Introduction	1-1
Relevance.....	1-2
Disk Storage Administration.....	1-3
SSVM Software Installation	1-3
RAID Volume Design.....	1-4
RAID Volume Creation.....	1-5
RAID Volume Administration	1-5
Disk Storage Concepts.....	1-6
Multi-Host Access.....	1-6
Host-based RAID (Software RAID Technology).....	1-9
Controller-based RAID (Hardware RAID Technology)....	1-10
Redundant Dual Active Controller Driver.....	1-11
Dynamic Multi-Path Driver.....	1-12
Hot Swapping.....	1-13
SPARCstorage Array 100	1-14
SPARCstorage Array 100 Features.....	1-14
SPARCstorage Array 100 Addressing	1-15

RSM Storage Array	1-16
RSM Storage Array Features	1-16
RSM Storage Array Addressing.....	1-17
SPARCstorage Array 214/219.....	1-18
SPARCstorage Array 214/219 Features	1-18
SPARCstorage Array 214 Addressing	1-19
Sun StorEdge A3000 (RSM Array 2000).....	1-20
Sun StorEdge A3000 Features	1-20
Sun StorEdge A3000 Addressing.....	1-21
Sun StorEdge A1000/D1000.....	1-23
Sun StorEdge A1000/D1000 Features.....	1-23
Sun StorEdge A1000 Differences	1-24
Sun StorEdge A1000 Addressing.....	1-24
Sun StorEdge D1000 Differences	1-25
Sun StorEdge D1000 Addressing.....	1-25
Sun StorEdge A3500	1-26
Sun StorEdge A3500 Features	1-26
Sun StorEdge A3500 Addressing.....	1-28
Sun StorEdge A5000	1-29
Sun StorEdge A5000 Features	1-29
Sun StorEdge A5000 Addressing.....	1-31
Sun StorEdge A7000	1-34
Sun StorEdge A7000 Enclosure.....	1-34
Sun StorEdge A7000 Functional Elements	1-36
Sun StorEdge A7000 Addressing.....	1-38
Combining SSVM and A7000 Devices.....	1-39
SPARCstorage MultiPack	1-40
SPARCstorage MultiPack Features	1-41
SPARCstorage MultiPack Addressing.....	1-41
Check Your Progress	1-42
Think Beyond	1-43
Sun StorEdge Volume Manager Installation.....	2-1
Relevance.....	2-2
Installation Process	2-3
Pre-Installation Planning	2-3
Current System Checkpoint	2-6
Installation and Testing of New Configuration.....	2-6
SSVM Software Installation.....	2-7
Software Package Installation	2-8
Software Distribution	2-9
Software Installation.....	2-9
Option Support Packages	2-10
Initializing the Sun StorEdge Volume Manager.....	2-11
The vxinstall Program	2-12

SSVM Disk Management	2-20
Physical Disk Layout	2-20
Private Region Usage	2-22
Disk Header	2-22
Configuration Database	2-23
Kernel Log	2-23
Overriding Default Values	2-24
SSVM Environment	2-25
SSVM System Startup Files	2-25
System Startup Messages	2-27
System Startup Processes	2-28
System and User Executable Files	2-29
Exercise: Configuring the Sun StorEdge Volume Manager	2-31
Preparation	2-31
Task – Installing the SSVM Software	2-32
Task – Initializing the SSVM Software	2-33
Task – Verifying the SSVM Startup	2-37
Task – Verifying the SSVM System Processes	2-38
Task – Verifying the SSVM System Files	2-39
Exercise Summary	2-40
Check Your Progress	2-41
Think Beyond	2-42
Introduction to Managing Data	3-1
Objectives	3-1
Relevance	3-2
Virtual Disk Management	3-3
Data Availability	3-3
Performance	3-4
Scalability	3-4
Maintainability	3-4
RAID Technology Overview	3-5
RAID Standards	3-6
Concatenation – RAID 0	3-7
Limitations	3-9
Striping – RAID 0	3-10
Advantages	3-11
Limitations	3-12
Guidelines for Choosing an Optimized Stripe Unit Size	3-12
Mirroring – RAID 1	3-13
Advantages	3-15
Limitations	3-15
Striping and Mirroring – RAID 0+1	3-16
Advantages	3-17
Limitations	3-17

Mirroring and Striping – RAID 1+0	3-18
Advantages	3-19
Limitations	3-19
Striping With Distributed Parity – RAID 5	3-20
Advantages	3-22
Limitations	3-22
Performance Factors	3-23
Guidelines for Optimizing Stripe Width	3-26
Check Your Progress	3-27
Think Beyond	3-28
Volume Manager Storage Administrator (VMSA) Software	4-1
Objectives	4-1
Relevance.....	4-2
Volume Manager Storage Administrator Software	4-3
Server / Client Software Installation	4-4
VMSA Server Software Startup	4-5
VMSA Client Software Startup	4-5
Client Software Startup	4-6
VMSA Initialization Display	4-7
VMSA Client Display	4-8
VMSA Client Software Features	4-9
Tool Bar	4-10
VMSA Menu Bar	4-12
VMSA Object Tree	4-13
VMSA Command Launcher	4-15
Docking Windows	4-16
Using the Create Menu	4-18
Exercise: Using the VMSA Client Software.....	4-23
Preparation.....	4-23
Task – Setting up the Environment	4-24
Task – Installing the VMSA Client Software.....	4-25
Task – Starting VMSA Client Software.....	4-26
Task – Setting up the VMSA Client Display	4-27
Task – Determining VMSA Client Command Functions..	4-28
Task – Defining VMSA Client Object Tree Functions	4-29
Exercise Summary.....	4-30
Check Your Progress	4-31
Think Beyond	4-32
Sun StorEdge Volume Manager Basic Operations.....	5-1
Objectives	5-1
Relevance.....	5-2
SSVM Initialization Review	5-3
Initialization	5-3
Encapsulation	5-4

Private and Public Region Format.....	5-5
Initialized Disk Types.....	5-5
Storage Configuration	5-6
Identifying Storage Devices.....	5-6
Identifying Controller Configurations.....	5-9
SSVM Objects.....	5-10
Sun StorEdge Volume Manager Disks.....	5-10
Disk Groups	5-11
Subdisks.....	5-12
Plexes	5-13
Volumes.....	5-14
Command-Line Status.....	5-15
Using vxprint.....	5-15
Using vxdisk.....	5-17
Using vxdg	5-18
Exercise: Performing SSVM Disk Drive Operations.....	5-19
Preparation.....	5-19
Task – Verifying Initial Disk Status.....	5-20
Task – Creating the First Disk Group	5-20
Task – Verifying Free Disk Space	5-26
Task – Renaming Disk Drives	5-28
Task – Removing Disks From a Disk Group.....	5-29
Task – Finishing Up	5-29
Exercise Summary.....	5-30
Check Your Progress	5-31
Think Beyond	5-32
Sun StorEdge Volume Manager Volume Operations.....	6-1
Objectives	6-1
Relevance.....	6-2
Disk Group Review	6-3
Primary Functions of a Disk Group	6-3
Disk Group Requirements	6-5
Movement of SSVM Disks Between Disk Groups	6-5
SSVM Volume Definition.....	6-6
Selecting a Disk Group.....	6-6
Using Volume Naming Conventions.....	6-9
Determining Volume Size.....	6-9
Identifying Volume Types.....	6-12
Volume Creation Using VMSA.....	6-14
The New Volume Form.....	6-15
Volume Creation Using the Command Line	6-17
The vxassist Command Format.....	6-17

Adding a UFS File System	6-19
Using the VMSA New File System Form	6-20
Adding a File System From the Command Line.....	6-21
Dirty Region Logging.....	6-23
DRL Overview	6-23
DRL Space Requirements	6-24
RAID-5 Logging	6-25
RAID-5 Log Overview.....	6-25
Log Placement	6-27
Planning for Logs.....	6-28
Exercise: Creating a Volume and a File System	6-29
Preparation.....	6-29
Task – Creating a Simple Concatenation.....	6-30
Task – Adding a Mirror.....	6-31
Task – Creating a RAID-5 Volume	6-34
Task – Displaying Volume Layout Details.....	6-36
Task – Performing Volume to Disk Mapping.....	6-38
Task – Removing a Volume.....	6-40
Task – Adding a File System	6-42
Task – Resizing a Volume or File System	6-44
Task – Adding a Dirty Region Log.....	6-45
Exercise Summary.....	6-46
Check Your Progress	6-47
Think Beyond	6-48
Sun StorEdge Volume Manager Advanced Operations.....	7-1
Relevance.....	7-2
Evacuating a Disk	7-3
Evacuation Conflicts.....	7-4
Evacuation Preparation.....	7-4
Performing an Evacuation	7-5
Moving Disks Without Preserving Data.....	7-6
Moving a Disk Using the Command Line.....	7-6
Moving a Disk From VMSA	7-8
Determining Which Disks Are Involved.....	7-10
Saving the Configuration.....	7-11
Moving the Disks to a New Disk Group	7-11
Reloading the Volume Configuration.....	7-13
Moving Disk Groups	7-14
Disk Group Ownership.....	7-15
Disk Group States	7-15
Preparation for Deporting a Disk Group.....	7-16
Deporting Options	7-16
Importing Disk Groups.....	7-17
Importing rootdg After a Crash.....	7-18

Hot Devices	7-19
Hot Spare Overview	7-19
Hot Relocation Overview	7-20
Failed Subdisk Detection	7-21
Hot-Relocation Failures.....	7-22
Enabling the Hot-Spare Feature.....	7-23
Snapshot Operations	7-24
Snapshot Prerequisites	7-24
Online Volume Relayout	7-26
Volume Relayout Prerequisites.....	7-27
Layered Volumes	7-28
Striped Pro Volume Structure.....	7-29
Exercise: Performing Advanced Operations.....	7-30
Preparation.....	7-30
Task – Moving a Populated Volume to Another Disk Group	7-31
Task – Moving a Disk Group Between Systems (Optional)	7-33
Task – Adding and Disabling a Hot Spare.....	7-34
Task – Performing a Snapshot Backup	7-35
Task – Creating a Striped Pro Volume.....	7-36
Exercise Summary.....	7-37
Check Your Progress	7-38
Think Beyond	7-39
Sun StorEdge Volume Manager Performance Management	8-1
Relevance.....	8-2
Performance Guidelines.....	8-3
Data Assignment.....	8-3
Bandwidth Improvement	8-6
Performance Monitoring.....	8-10
Gathering Statistical Information	8-10
Displaying Statistics Using the vxstat Command	8-11
Displaying Statistics Using the vxtrace Command	8-12
Performance Analysis.....	8-13
Preparation.....	8-14
Volume Statistics.....	8-15
Disk Statistics.....	8-15
Trace Information	8-16
RAID-5 Write Performance	8-17
Read-Modify-Write Operations.....	8-17
Full-Stripe Write Operations	8-20
Exercise Summary.....	8-23
Check Your Progress	8-24
Think Beyond	8-25

RAID Manager Architecture	9-1
Relevance.....	9-2
RAID Manager Components and Features.....	9-3
RAID Manager Components.....	9-4
RAID Manager Features	9-5
Definitions.....	9-6
RAID Module	9-7
Drive Group.....	9-8
Logical Unit (LUN).....	9-10
Drive Group Numbering	9-12
Hot Spare Drive.....	9-14
RAID Reconstruction.....	9-16
Degraded Mode.....	9-17
Reconstruction	9-18
Hot Spares	9-18
RAID 1 (Mirroring) LUN Difference.....	9-19
Cache Memory	9-20
Controller Cache	9-20
Performance	9-20
Write Cache Mirroring	9-21
Cache Without Batteries.....	9-21
RAID Manager Applications.....	9-23
Configuration	9-24
Status.....	9-24
Recovery Guru.....	9-25
Maintenance / Tuning	9-25
About	9-25
Command-Line Interface	9-26
Device Naming Conventions	9-28
Standard Device Names.....	9-28
Exercise: Reviewing RAID Manager Architecture.....	9-30
Task	9-30
Exercise Summary.....	9-33
Check Your Progress	9-34
Think Beyond	9-35
 Sun StorEdge Volume Manager Recovery Procedures	A-1
Summary	A-2
Detecting Failed Physical Disks.....	A-3
Plex States.....	A-4
Volume States	A-6
RAID-5 Volume States.....	A-7

Moving Data From a Failing Disk	A-8
Preparing for an Evacuation.....	A-8
Performing an Evacuation	A-8
Recovering a Volume	A-9
Recovering a RAID-5 Volume (A5000).....	A-10
Recovering a RAID-5 Volume (SPARCstorage Array).....	A-11
Recovering a Mirror (A5000).....	A-12
Recovering a Mirror (SPARCstorage Array)	A-13
Replacing a Failed SSVM Disk (A5000).....	A-14
Replacing a Failed SSVM Disk (SPARCstorage Array).....	A-15
Booting After a Failure – Booting From a Mirror.....	A-17
Replacing a Failed Boot Disk.....	A-18
Moving a Storage Array to Another Host	A-19
Sun StorEdge Volume Manager Boot Disk Encapsulation	B-1
Summary	B-2
Boot Disk Encapsulation Overview	B-3
Preferred Boot Disk Configuration	B-3
Prerequisites for Boot Disk Encapsulation	B-4
Primary and Mirror Configuration Differences	B-4
Encapsulating the Boot Disk Using VMSA.....	B-5
Encapsulation Files	B-11
Files in the /etc/vx Directory	B-11
The /etc/vfstab File	B-13
Boot PROM Changes	B-13
Un-Encapsulating the Boot Disk.....	B-14
Sun StorEdge Volume Manager and RAID Manager.....	C-1
Summary	C-2
SSVM and RAID Manager.....	C-3
Determining What Is Seen by the System	C-3
Installing Sun StorEdge Volume Manager.....	C-4
Using Sun StorEdge Volume Manager With RAID Manager.....	C-5
Determining Supported Configurations	C-6
Determining Unsupported Configurations	C-7
Using SSVM Hot Relocation and RAID Manager Hot Sparing.....	C-8
The Veritas VxFS File System	D-1
Summary	D-2
Introduction to VxFS	D-3
Fast File System Recovery.....	D-4
Online System Administration	D-5
Defragmentation	D-5
Resizing	D-5
Online Backup	D-6

Enhanced File System Performance	D-7
Extent-based Allocation	D-8
Disk Layout.....	D-9
Superblock.....	D-10
Object-Location Table.....	D-10
Intent Log	D-11
Allocation Unit	D-12
RAID Manager Procedures	E-1
Relevance.....	E-2
Starting RM6.....	E-3
Creating a Drive Group.....	E-6
Adding LUNs to an Existing Drive Group	E-19
Creating a Hot Spares Pool.....	E-26
Deleting a LUN.....	E-33
Recovering Failures	E-40
Exercise: Using RAID Manager Procedures	E-56
Task – Creating a Drive Group	E-57
Task – Adding LUNs To An Existing Drive Group.....	E-58
Task – Creating a Hot Spares Pool	E-59
Task – Deleting a LUN	E-60
Task – Recovering Failures	E-61
Exercise Summary.....	E-62
Check Your Progress	E-63
Think Beyond	E-64

About This Course

Course Goal

The goal of this course is to train you to install, configure, and manage the Volume Manager 3.0 utility on a wide range of Sun™ disk storage arrays.



Sun Educational Services

Course Overview

- Introduce Sun storage arrays
- Install and initialize Volume Manager
- Introduce RAID and data management
- Perform Volume Manager disk, volume, and file system operations
- Create and manipulate hot spare devices
- Perform Volume Manager recovery operations
- Use Volume Manager for performance tuning
- Introduce RAID Manager software

Course Overview

This course provides the essential information and skills to manage disks on a variety of Sun disk storage arrays using Sun StorEdge Volume Manager™ (SSVM) software.

You will be introduced to the Volume Manager (VM) installation process including issues relating to boot disk encapsulation. Practical information about the creation and uses of all redundant array of inexpensive disks (RAID) volume structures will be presented along with basic performance issues.

An important feature of this course is that disk replacement procedures for a variety of Sun storage arrays are presented. There are critical differences in the replacement process for some storage arrays.

Course Map

The following course map enables you to see what you have accomplished and where you are going in reference to the course goal:

Overview

Sun Storage
Introduction

Installation

Sun StorEdge
Volume Manager
Installation

Volume Manager Introduction

Introduction
to
Managing
Data

Volume Manager
Storage
Administrator
(VMSA) Software

Operations

Sun StorEdge
Volume
Manager
Basic
Operations

Sun StorEdge
Volume
Manager
Volume
Operations

Sun StorEdge
Volume
Manager
Advanced
Operations

Tuning

Sun StorEdge Volume
Manager Performance
Management

RAID Manager

RAID Manager
Architecture



Module-by-Module Overview

- Module 1 – “Sun Storage Introduction”
- Module 2 – “Sun StorEdge Volume Manager Installation”
- Module 3 – “Introduction to Managing Data”
- Module 4 – “Volume Manager Storage Administrator (VMSA) Software”
- Module 5 – “Sun StorEdge Volume Manager Basic Operations”
- Module 6 – “Sun StorEdge Volume Manager Volume Operations”

Module-by-Module Overview

This course contains the following modules:

- Module 1 – “Sun Storage Introduction”

General overviews of currently supported Sun disk storage arrays are presented. The primary focus of this module is to emphasize that several storage arrays have unique disk replacement procedures that must be followed.

Lab exercise – There are no lab exercises for this module.

- Module 2 – “Sun StorEdge Volume Manager Installation”

The goal of this module is to install and initialize the SSVM software.

Lab exercise – Install, initialize, and verify proper operation of the SSVM software.

Module-by-Module Overview

- Module 3 – “Introduction to Managing Data”

This module provides an introduction to data management concepts in the Solaris™ Operating Environment (“Solaris”).

Lab exercise – There are no lab exercises for this module.

- Module 4 – “Volume Manager Storage Administrator (VMSA) Software”

The focus of this module is the VMSA background and terminology that is necessary to successfully use the SSVM graphical administration interface.

Lab exercise – Install, connect, and become familiar with VMSA software.

- Module 5 – “Sun StorEdge Volume Manager Basic Operations”

The background and terminology that is necessary to perform all basic SSVM disk management operations are presented in this module.

Lab exercise – Perform basic SSVM disk operations such as displaying disk properties, setting up a disk for SSVM use, creating and modifying disk groups, and encoding the command-line equivalent of graphical user interface (GUI) operations.

- Module 6 – “Sun StorEdge Volume Manager Volume Operations”

This module provides the background and terminology that is necessary to create mirrored and RAID-5 volumes.

Lab exercise – Create and manipulate RAID volumes, resize volumes, and add file systems and logs to volumes.

Module-by-Module Overview

- Module 7 – “Sun StorEdge Volume Manager Advanced Operations”

This module provides the background and terminology that is necessary to perform advanced SSVM tasks.

Lab exercise – Move disks between disk groups, move disk groups between systems, perform a snapshot backup, and manipulate hot spares.

- Module 8 – “Sun StorEdge Volume Manager Performance Management”

The information necessary to obtain and use performance data to establish priorities that can improve overall system performance is presented in this module.

Lab exercise – Observe an instructor demonstration that illustrates the performance differences between three types of RAID-5 write operations.

- Module 9 – “RAID Manager Architecture”

This optional module introduces the RAID Manager software package. It is intended to be used as an information clearing house for the RAID Manager, which is used to configure controller-based RAID arrays prior to using the Volume Manager.

Lab exercise – Answer module review questions.

Course Objectives

Upon completion of this course, you should be able to:

- Install and initialize SSVM software
- Define SSVM objects
- Describe public and private regions
- Start and customize SSVM GUIs
- Perform operations using the command-line interface
- Perform disk and volume operations
- Create RAID-5 volumes and dirty region logs
- Perform common file system operations using the SSVM GUI
- Create and manipulate disks groups
- Remove and replace failed disk drives
- Create and manage hot spare pools
- Manage and disable the hot relocation feature
- Perform basic performance analysis
- Identify RAID Manager features

Skills Gained by Module

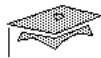
The skills for *Sun StorEdge Volume Manager Administration* are shown in column 1 of the matrix below. The black boxes indicate the main coverage for a topic; the gray boxes indicate the topic is briefly discussed.

Skills Gained	Module								
	1	2	3	4	5	6	7	8	9
Install and Initialize SSVM software		■		■					
Define SSVM objects	■								
Describe public and private regions			■						
Start and customize SSVM GUIs				■					
Perform operations using the command-line interface					■				
Perform disk and volume operations					■	■	■		
Create RAID-5 volumes and dirty region logs						■			
Perform common file system operations using the SSVM GUI						■			
Create and manipulate disks groups							■		
Remove and replace failed disk drives					■	■	■		
Create and manage hot spare pools							■		
Manage and disable the hot relocation feature							■		
Perform basic performance analysis								■	
Identify RAID Manager features									■

Guidelines for Module Pacing

The following table provides a rough estimate of pacing for this course:

Module	Day 1	Day 2	Day 3	Day 4
“About This Course”	A.M.			
“Sun Storage Introduction”	A.M.			
“Sun StorEdge Volume Manager Installation”	P.M.			
“Introduction to Managing Data”		A.M.		
“Volume Manager Storage Administrator (VMSA) Software”		P.M.		
“Sun StorEdge Volume Manager Basic Operations”			A.M.	
“Sun StorEdge Volume Manager Volume Operations”			P.M.	
“Sun StorEdge Volume Manager Advanced Operations”				A.M.
“Sun StorEdge Volume Manager Performance Management”				P.M.
“RAID Manager Architecture”				P.M.



Topics Not Covered

- Solaris™ Operating Environment ("Solaris") installation
– Covered in SA-237: *Solaris 7 System Administration I*
- Storage system maintenance – Covered in SM-250: *Sun Storage System Maintenance*

Topics Not Covered

This course does not cover the topics shown on the above overhead. Many of the topics listed on the overhead are covered in other courses offered by Sun Educational Services:

- Solaris operating system (OS) installation – Covered in SA-237: *Solaris 7 System Administration I*
- Storage system maintenance – Covered in SM-250: *Sun Storage System Maintenance* (releases in calendar year '99)

Refer to the Sun Educational Services catalog for specific information and registration.



How Prepared Are You?

- Can you edit files using one of the standard editors available with Solaris 7 OS?
- Can you perform simple command-line operations?
- Can you use the man pages?

How Prepared Are You?

To be sure you are prepared to take this course, can you answer yes to the questions shown on the above overhead?

- Can you edit files using one of the standard editors available with the Solaris 7 OS?
- Can you perform simple command-line operations?
- Can you use the man pages?



Sun Educational Services

Introductions

- Name
- Company affiliation
- Title, function, and job responsibility
- Volume Manager experience
- Reasons for enrolling in this course
- Expectations for this course

Introductions

Now that you have been introduced to the course, introduce yourself to each other and the instructor, addressing the items shown on the above overhead.



How to Use Course Materials

- Objectives
- Relevance
- Overhead image
- Lecture
- Exercise
- Check Your Progress
- Think Beyond

How to Use Course Materials

To enable you to succeed in this course, these course materials employ a learning model that is composed of the following components:

- **Course map** – An overview of the course content appears in the “About This Course” module so you can see how each module fits into the overall course goal.
- **Objectives** – What you should be able to accomplish after completing this module is listed here.
- **Relevance** – This section, which appears in every module, provides scenarios or questions that introduce you to the information contained in the module and provoke you to think about how the module content relates to the Sun StorEdge Volume Manager.
- **Overhead image** – Reduced overhead images for the course are included in the course materials to help you easily follow where the instructor is at any point in time. Overheads do not appear on every page.

How to Use Course Materials

- **Lecture** – The instructor will present information specific to the topic of the module. This information will help you learn the knowledge and skills necessary to succeed with the exercises.
- **Exercise** – Lab exercises will give you the opportunity to practice your skills and apply the concepts presented in the lecture.
- **Check your progress** – Module objectives are restated, sometimes in question format, so that before moving on to the next module you are sure that you can accomplish the objectives of the current module.
- **Think beyond** – Thought-provoking questions are posed to help you apply the content of the module or predict the content in the next module.

Course Icons and Typographical Conventions

The following icons and typographical conventions are used in this course to represent various training elements and alternative learning resources.

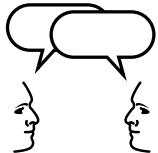
Icons



Additional resources – Indicates additional reference materials are available.



Demonstration – Indicates a demonstration of the current topic is recommended at this time.



Discussion – Indicates a small-group or class discussion on the current topic is recommended at this time.



Exercise objective – Indicates the objective for the lab exercises that follow. The exercises are appropriate for the material being discussed.

Course Icons and Typographical Conventions

Note – Additional important, reinforcing, interesting or special information.



Caution – A potential hazard to data or machinery.



Warning – Anything that poses personal danger or irreversible damage to data or the operating system.

Typographical Conventions

Courier is used for the names of command, files, and directories, as well as on-screen computer output. For example:

```
Use ls -al to list all files.  
system% You have mail.
```

Courier bold is used for characters and numbers that you type. For example:

```
system% su  
Password:
```

Courier italic is used for variables and command-line placeholders that are replaced with a real name or value. For example:

To delete a file, type `rm filename`.

Palatino italics is used for book titles, new words or terms, or words that are emphasized. For example:

Read Chapter 6 in *User's Guide*.
These are called *class* options
You *must* be root to do this.

Objectives

Upon completion of this module, you should be able to:

- Describe the major disk storage administration tasks
- List the disk storage concepts common to many storage arrays
- List the general features of current Sun disk storage models
- Describe the basic Sun StorEdge Volume Manager disk drive replacement process
- Describe a typical disk replacement process variation

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

- Is disk technology becoming simpler? Why is the discussion on hardware issues at the beginning of this course?
- I am only responsible for a small part of my company's SSVM administration program. Why do I need to understand so much about the hardware?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The online manual page for `luxadm`
- *Platform Notes: Using luxadm Software, Solaris 2.6 Hardware:5/98*
- *Sun Storage A5000 Hardware Configuration Guide*



Sun Educational Services

Disk Storage Administration

As a Sun StorEdge Volume Manager administrator you will be responsible for one or more of the following duties:

- SSVM software installation
- RAID volume design strategies
- RAID volume creation
- RAID volume administration

Disk Storage Administration

SSVM Software Installation

Installing the SSVM software is essentially the same as installing any Solaris application. It is done using the standard command-line or graphical package installation tools.

SSVM Initialization

The big difference when installing SSVM is that at least one disk drive must be specially initialized and brought under SSVM control using the `vxinstall` utility.

Disk Storage Administration

SSVM Software Installation

Required Hardware Knowledge

The SSVM installation process is the same regardless of the system platform or storage technology used but you must be able to identify array storage device addresses from other types of disk storage.

If you are not familiar with the device address strategy in your particular storage devices, you might accidentally initialize the wrong disk drives. This could destroy valuable data, including the operating system.

RAID Volume Design

Generally, virtual volume structures are designed with one or more of the following goals in mind:

- Cost savings
- Performance
- Availability
- Maintainability

In most cases, compromises are made when choosing between cost savings, performance, and availability.

Required Hardware Knowledge

A thorough understanding of interface types, addressing schemes, and internal hardware structure is required to achieve design goals. It is possible to design virtual volume structures without this background knowledge, but the result will probably perform poorly and might not have the reliability that is required for your application.

Disk Storage Administration

RAID Volume Creation

Creating RAID volume structures using SSVM can be done using a graphical interface or command-line utilities. The command-line utilities are most commonly used when volume creation must be automated using script files.

The graphical interface can be configured to display command-line equivalents for each operation.

Required Hardware Knowledge

Even though you might not be responsible for the design of your SSVM volume structures, you must still be familiar with most aspects of your particular storage devices.

RAID Volume Administration

In larger installations, the most common SSVM administrative task is identifying and replacing failed disk drives. At the simplest level, this involves the use of a single SSVM utility, `vxdiskadm`. For certain storage platforms, another utility, `luxadm`, must also be used during the disk replacement process.

Most SSVM administrative tasks require analyzing error messages. This is done using SSVM utilities such as `vxprint` and `vxdisk` along with some basic Solaris OS commands.

Required Hardware Knowledge

Administering RAID volumes requires a number of hardware related skills including:

- Decoding device error messages
- Relating device addresses to physical devices
- Following hardware removal procedures that are appropriate for each particular disk storage technology



Sun Educational Services

Disk Storage Concepts

In order to better comprehend the balance of this course, the following disk storage concepts must be understood:

- Multi-host access
- Host-based RAID
- Controller-based RAID
- Redundant dual active controllers
- Dynamic multi-pathing
- Hot swapping

Disk Storage Concepts

Multi-Host Access

In the past, this feature was referred to as dual-porting. With the advent of current technology such as the Sun StorEdge A5000, as many as four different hosts can be connected to the same storage device.

Multi-Initiated SCSI

Sun MultiPack storage devices support physical small computer system interface (SCSI) interface connections from two different host systems. The SCSI interface on each of the systems must have a different initiator identifier (ID) setting. This is a system firmware configuration known as the `scsi-initiator-id`.

Disk Storage Concepts

Multi-Host Access

Multi-Initiated SCSI

As shown in Figure 1-1, the `scsi-initiator-id` on one of the host systems must be changed to eliminate the addressing conflict between the two host systems.

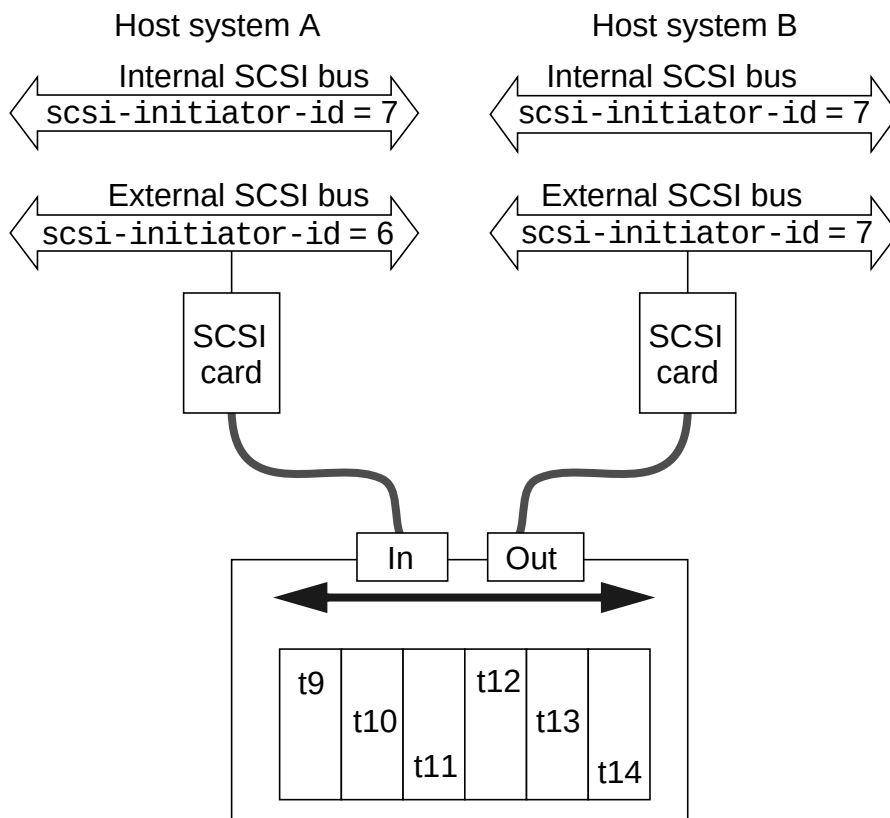


Figure 1-1 SCSI Initiator Configuration

The SCSI initiator values are changed using complex system firmware commands. The process varies with system hardware platforms.

Disk Storage Concepts

Multi-Host Access

Multi-Host Fiber Optic Interface

Two different fiber-optic interface storage arrays support multiple host connections. The SPARCstorage Array™ 100 allows up to two host systems to connect to a single storage array. The Sun StorEdge A5000 array allows up to four host system connections.

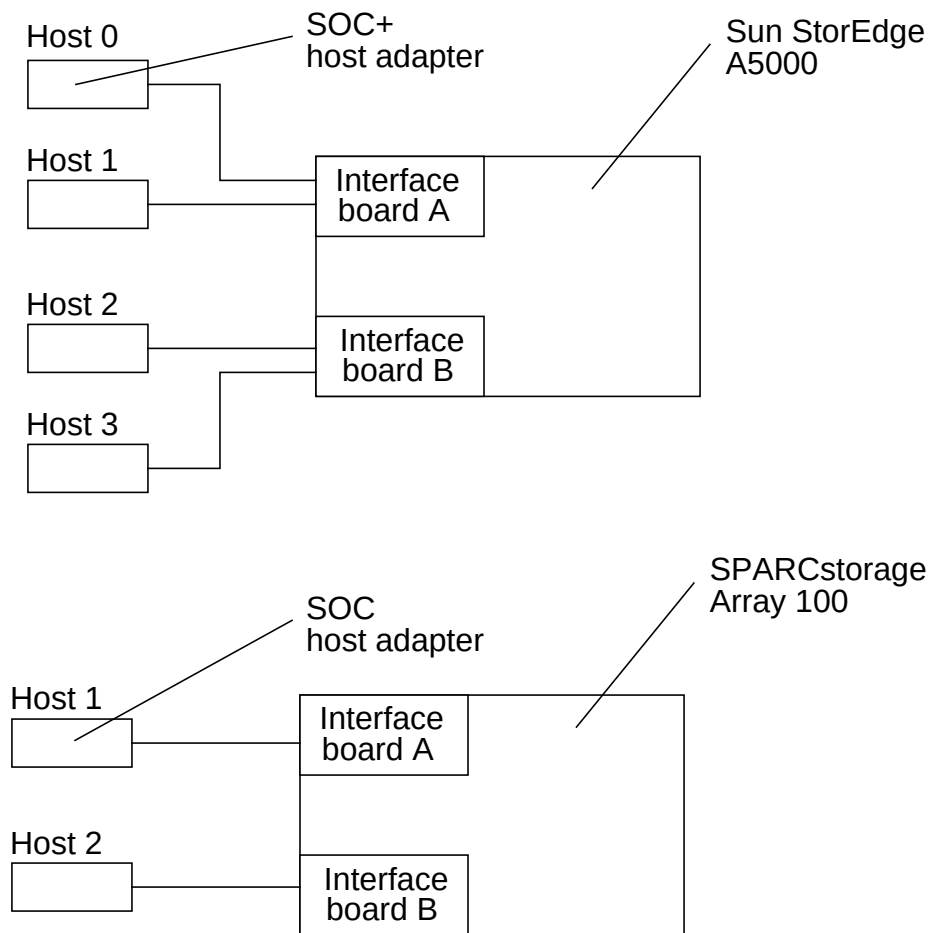


Figure 1-2 Fiber-Optic Multiple Host Connections

Disk Storage Concepts

Host-based RAID (Software RAID Technology)

The SSVM is a good example of software RAID technology. As shown in Figure 1-3, user applications access a virtual structure through a single path that is actually composed of three separate disk drives.

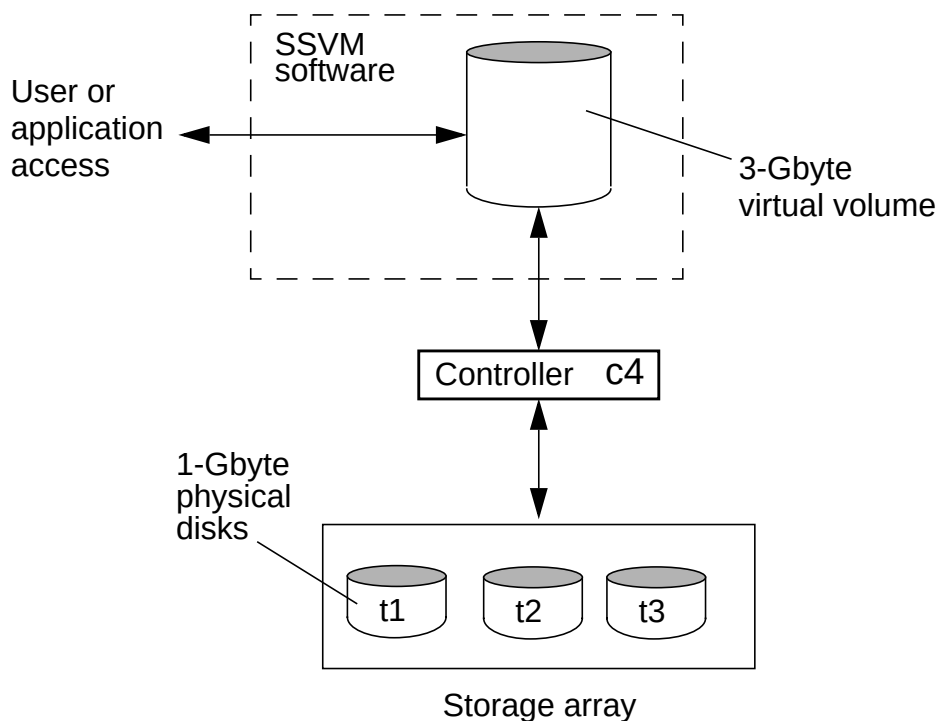


Figure 1-3 Host-based RAID Technology

A typical virtual volume pathname would be similar to

```
/dev/vx/dsk/dga/volume-01
```

Even though the physical paths to the three disk drives in Figure 1-3 still exist, they are not accessed directly by users or applications. Only the virtual volume paths are referenced by users.

The virtual structures are created and managed by software that runs on the host system.

Disk Storage Concepts

Controller-based RAID (Hardware RAID Technology)

Controller-based RAID solutions use firmware that runs in external controller boards to maintain virtual structures that are composed of one or more physical disk drives. As shown in Figure 1-4, RAID Manager software running on the host system is used to configure virtual structures in the external controller board. After initial configuration, the controller board firmware manages the virtual structures.

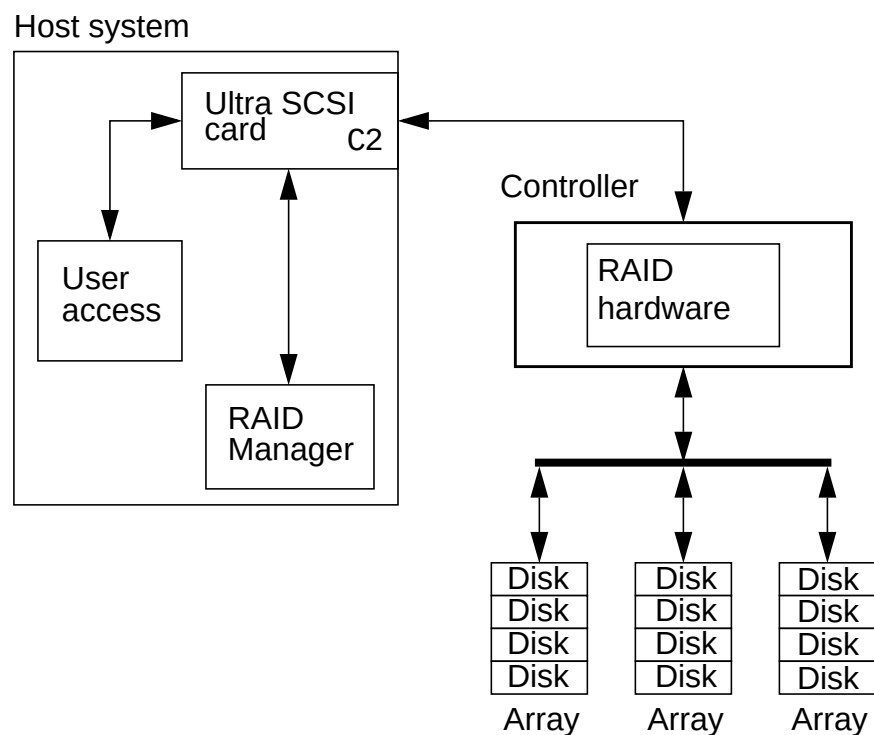


Figure 1-4 Controller-based RAID Technology

A typical hardware RAID device appears to be the same as any physical path, such as `/dev/dsk/c0t5d0s0`. Applications are unaware of the underlying RAID structures.

Hardware RAID solutions typically offer much better performance for some types of RAID structures because RAID overhead calculations are performed at very high speed by the controller-resident hardware instead of on the host system as in host-based RAID.

Disk Storage Concepts

Redundant Dual Active Controller Driver

Some Sun storage devices allow dual connections to a storage array from a single host system. One host adapter can be configured as a backup if the primary access path fails or the two adapters can be used in a load balancing configuration.

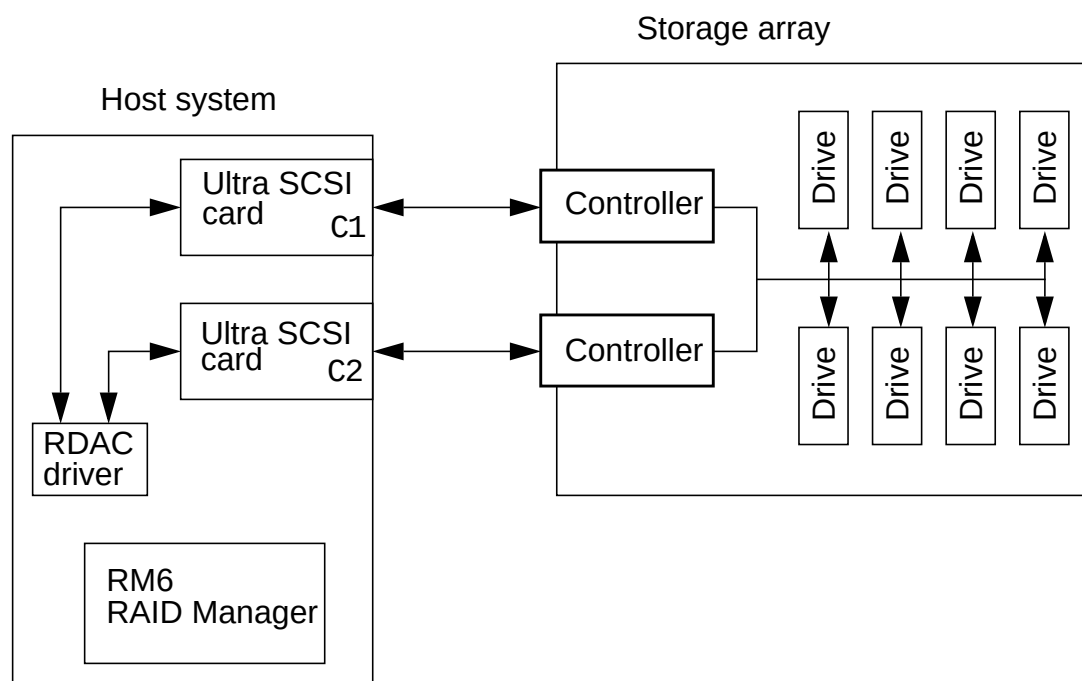


Figure 1-5 Redundant Dual Active Controller Driver

The redundant dual active controller (RDAC) driver is a special purpose driver that manages dual interface connections. This driver is available with some of the Sun hardware RAID storage arrays, which include the A3000 and A3500 models.

Applications interface with the RDAC driver and are unaware of interface failure. If one of the dual-controller paths fails, the RDAC driver automatically directs input/output (I/O) to the functioning path.

The controller-based RAID solution is used only on SCSI hardware interfaces.

Disk Storage Concepts

Dynamic Multi-Path Driver

The dynamic multi-path driver (DMP) is unique to the SSVM product. It is used only with fiber-optic interface storage arrays. As shown in Figure 1-6, the DMP driver can access the same storage array through more than one path. The DMP driver will automatically configure multiple paths to the storage array. Depending on the storage array model, the paths will either be used for load-balancing in a primary mode of operation, or in a backup mode of operation.

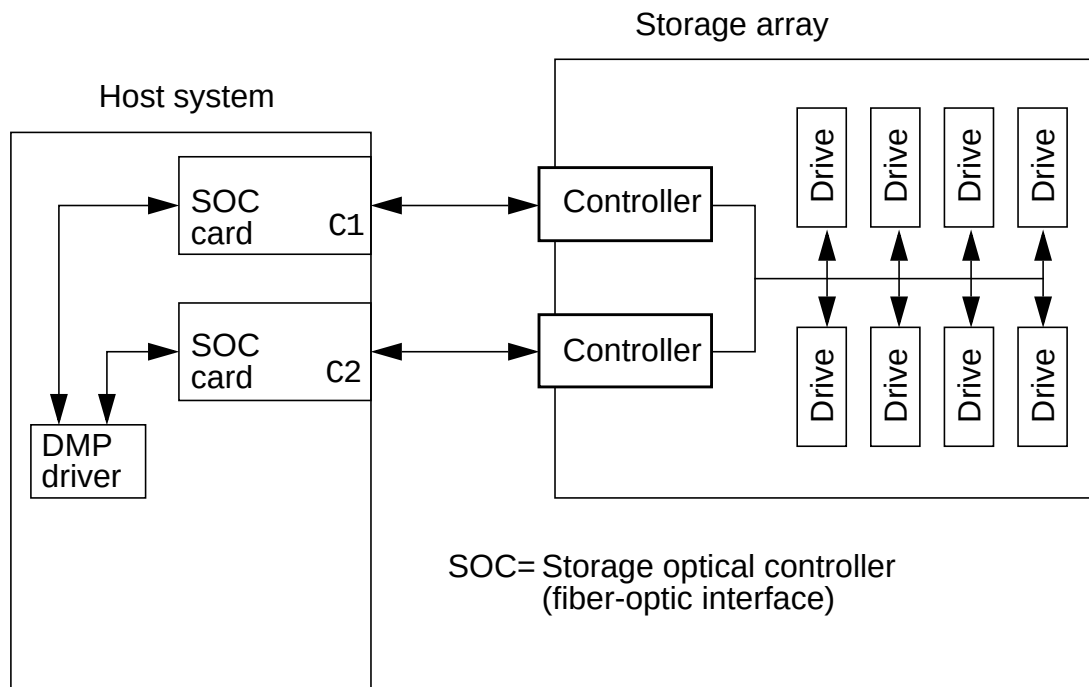


Figure 1-6 Dynamic Multi-Path Driver

The paths can be enabled and disabled with the SSVM `vxddmpadm` command.

Note – The DMP feature of SSVM is not compatible with the alternate pathing software of the operating system. During installation, SSVM checks to see if alternate pathing (AP 2.x) is configured and if so, it does not install the DMP (driver) software.

Disk Storage Concepts

Hot Swapping

Most Sun disk storage arrays are engineered so that a failed disk drive can be replaced without interrupting customer applications. The disk replacement process also includes one or more software operations that can vary with each disk storage platform.

Standard SSVM Disk Replacement Procedure

In its simplest form, the process to replace a failed disk drive that is under SSVM control is as follows:

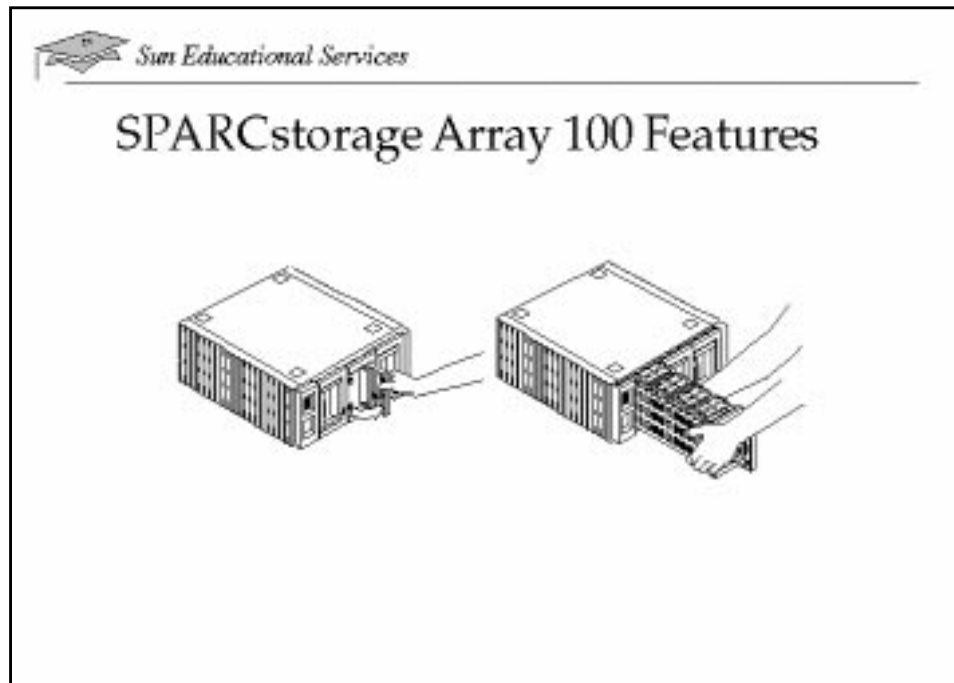
1. Use the SSVM `vxdiskadm` utility to logically remove the disk.
2. Hot swap in a new disk drive.
3. Use the SSVM `vxdiskadm` utility to logically install the new disk.

Disk Replacement Variations

The basic SSVM disk replacement process is more complex for some storage arrays such as the StorEdge A5000. The A5000 procedure is as follows:

1. Use the SSVM `vxdiskadm` utility options 4 and 11 to logically remove and offline the disk.
2. Use the `luxadm` utility to remove the physical disk drive path.
3. Hot swap in a new disk drive.
4. Use the `luxadm` utility to build a new physical disk drive path.
5. Use the SSVM `vxdiskadm` utility to logically install the new disk.

Note – There are other variations on the basic disk replacement process. You must be familiar with the exact process for your particular disk storage devices.

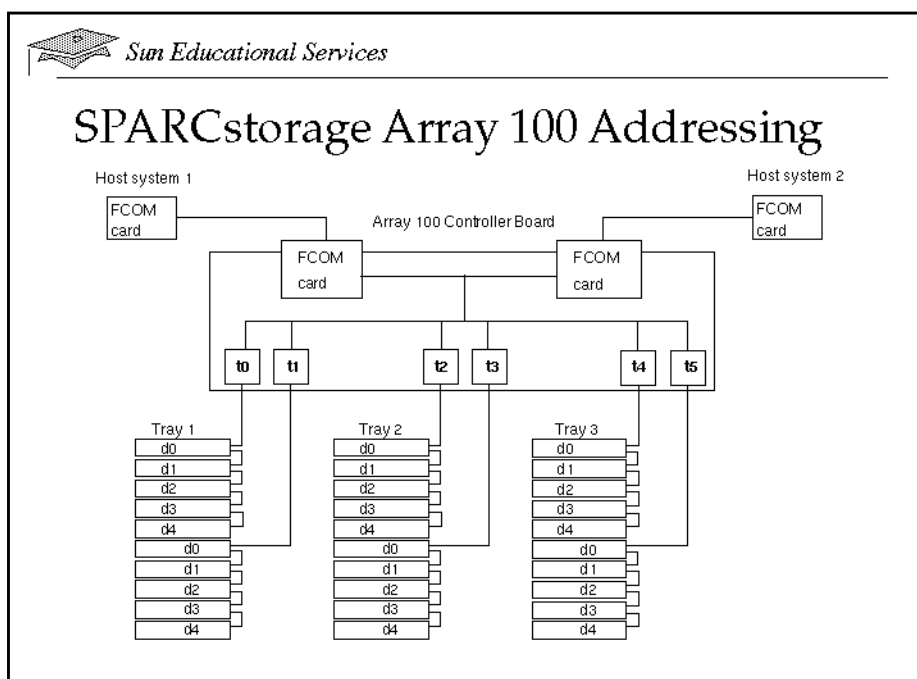


SPARCstorage Array 100

SPARCstorage Array 100 Features

The SPARCstorage Array 100 (SSA100) has the following features:

- Thirty disk drives
 - Ten disk drives per removable tray
 - Dual-ported, fiber-optic interfaces
 - Six internal SCSI target addresses
 - Hot-pluggable disk trays with restrictions and cautions
- ▼ All the disks in an SSA100 drive tray must put into a quiescent state before you can pull the drive tray.



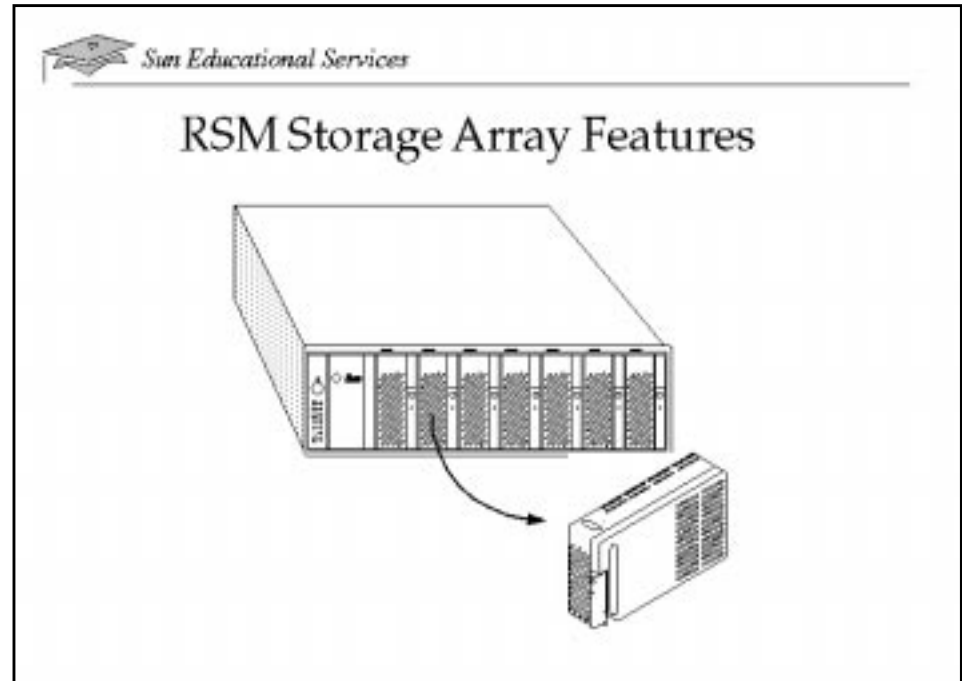
SPARCstorage Array 100

SPARCstorage Array 100 Addressing

Typical Address Paths

These are:

- c0t3d0s2
- c4t2d4s4

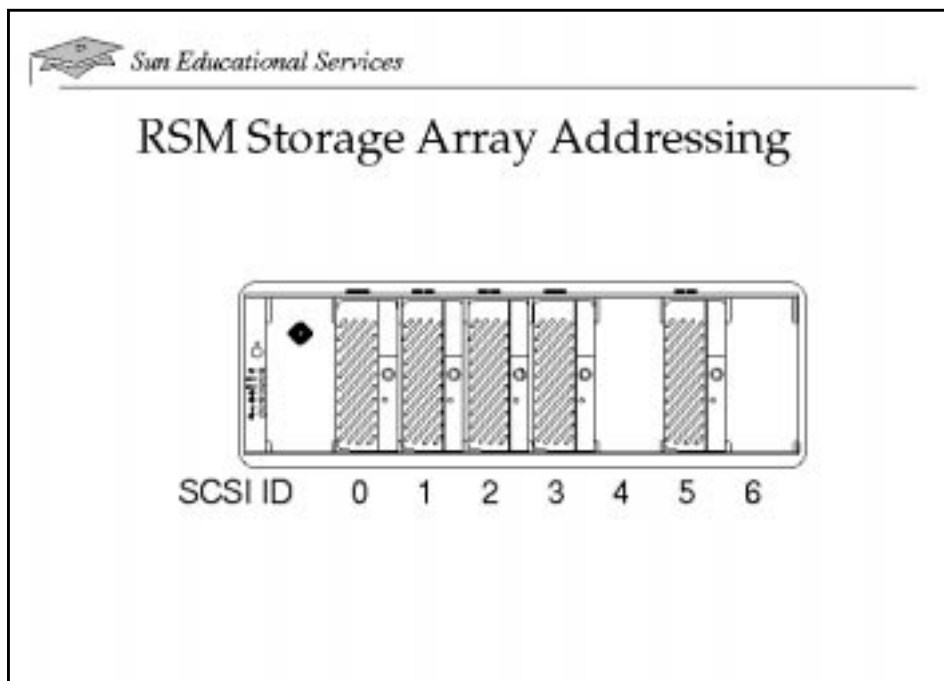


RSM Storage Array

RSM Storage Array Features

The RSM storage tray can be used as a standalone unit attached to a single differential SCSI SBus card or it can be rack mounted and used with a special dual-ported controller assembly. It has:

- Seven disk drives in each array
- Disks which are hot-pluggable
- Redundant power modules
- Redundant cooling modules
- Drives which are individually removable



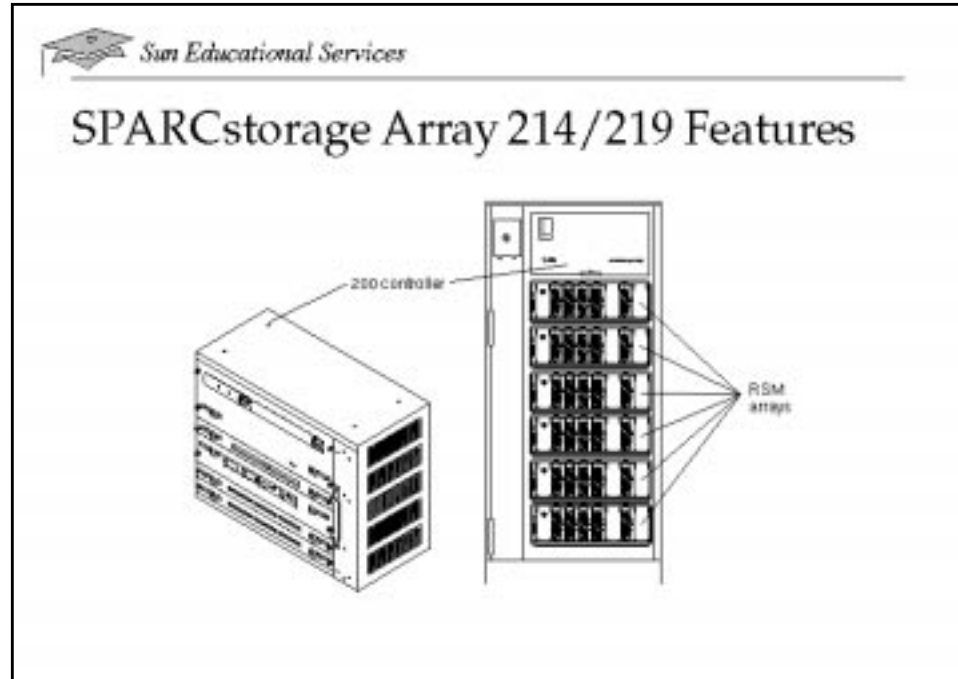
RSM Storage Array

RSM Storage Array Addressing

If the RSM storage tray is attached to a differential-wide SCSI interface, the SCSI target ID corresponds to the slot number in the tray. Typical physical addresses would be:

- c2t0d0s2
- c4t2d0s4
- c2t5d0s3

The device number will always be zero.

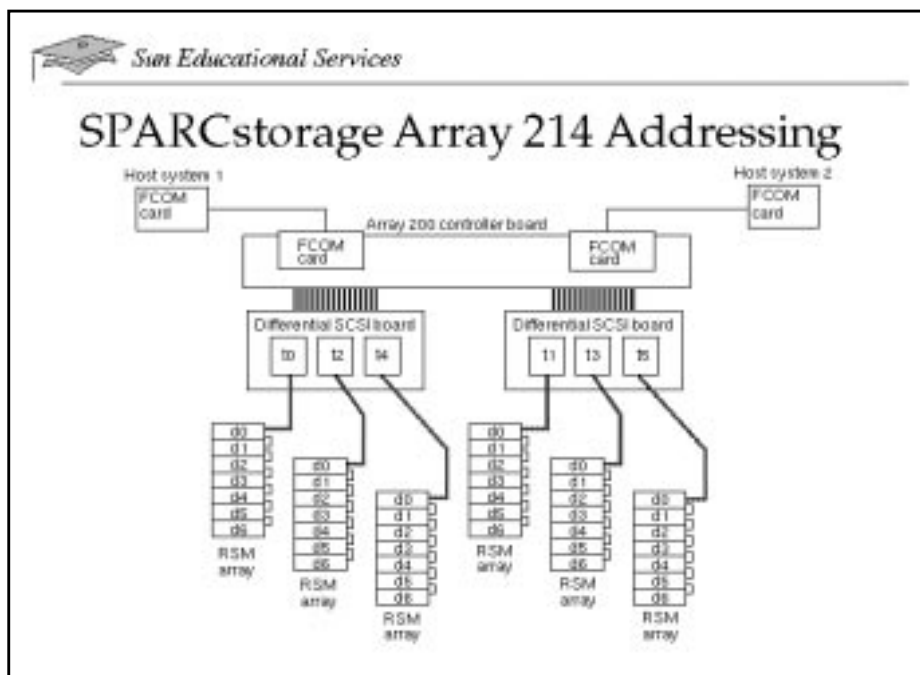


SPARCstorage Array 214/219

SPARCstorage Array 214/219 Features

The SPARCstorage Array Model 214/219 combines a SPARCstorage Array 200 disk controller with up to six removable storage module (RSM) differential SCSI disk trays. This array:

- Is rack mounted in a 56-inch expansion cabinet
- Has dual port fiber-optic interface
- Has six differential SCSI outputs
- Is typically connected to RSM array trays (with a six tray maximum)
- Has seven devices per RSM tray, which are either 4 or 9 Gbytes each
- Has individual devices which can be removed from the tray without affecting other drives in the tray



SPARCStorage Array 214

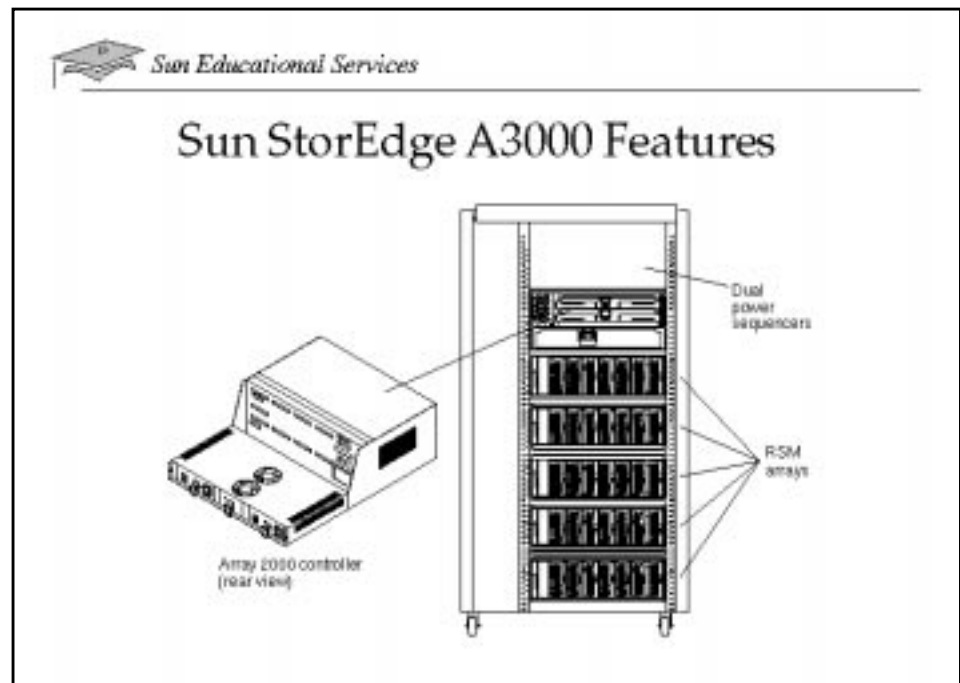
SPARCstorage Array 214 Addressing

Typical Address Paths

These are:

- c0t3d0s2
- c4t2d5s4

In this configuration, the SCSI device number corresponds to the slot number in each RSM tray.

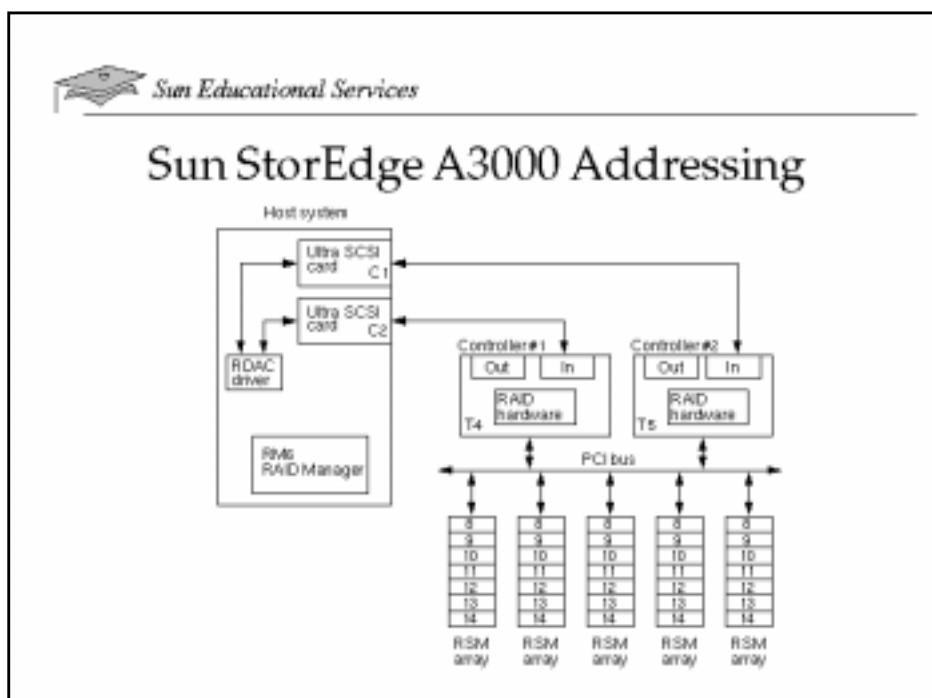


Sun StorEdge A3000 (RSM Array 2000)

Sun StorEdge A3000 Features

The Sun StorEdge A3000 controller is a compact unit that provides hardware-based RAID technology. Two SCSI controllers manage up to five RMS storage arrays. These include:

- Redundant hot plug RAID controllers
- Redundant power and cooling
- Data cache with battery back-up
- Support for RAID 0, 1, 0+1, 1+0, 3, and 5
- Dual Ultra™ SCSI host interface (40 Mbytes/second)
- Hot-plug controllers, power supplies, and cooling



Sun StorEdge A3000 (RSM Array 2000)

Sun StorEdge A3000 Addressing

The Sun StorEdge A3000 is not directly addressed. A RAID Manager GUI called RM6 is used to configure hardware RAID devices consisting of groups of RSM disks. The RDAC driver enables automatic failover to the backup access path through the second Ultra SCSI interface. Different hardware RAID volume access can be directed through each interface for load balancing. There are several utilities associated with managing the hardware RAID devices created with RM6.

Note – Once created, the RM6 “devices” can be referenced by other system utilities such as virtual volume managers.

Sun StorEdge A3000 (RSM Array 2000)

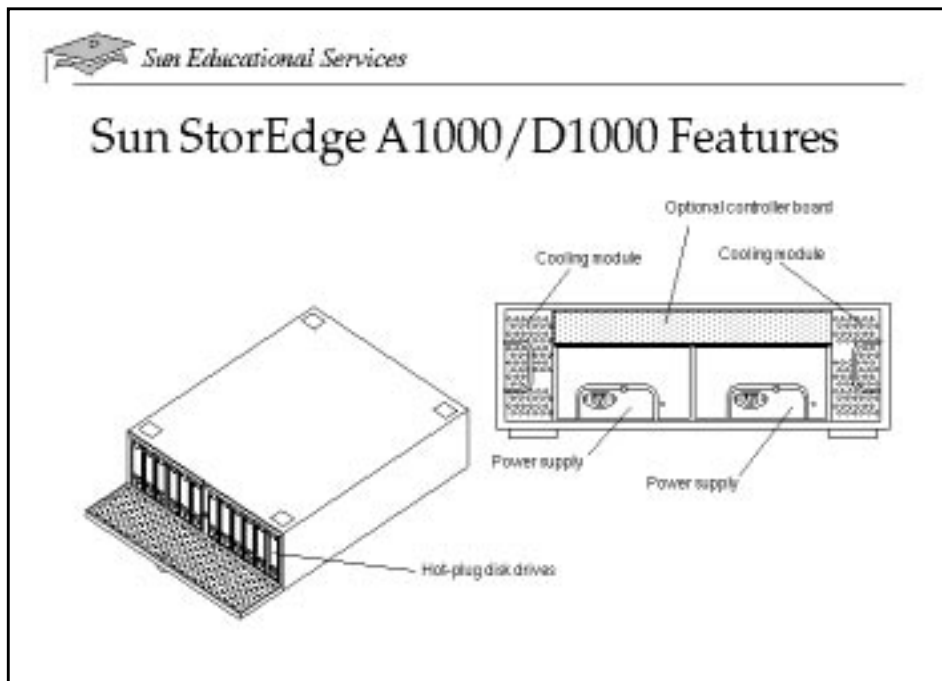
Sun StorEdge A3000 Addressing

The RM6 RAID Manager software can take one or more physical disk drives in the storage trays, and configure them as a single logical unit (LUN).

This LUN can have a hardware RAID structure.

Once configured, a RM6 LUN appears to be a regular physical address such as `c2t3d0s2`. The underlying configuration is hidden from applications.

There are potential problems with configuring a SSVM software RAID-5 device on top of a RM6 hardware RAID-5 device. You must read the array documentation carefully.



Sun StorEdge A1000/D1000

Sun StorEdge A1000/D1000 Features

Except for controller boards, the Sun StorEdge A1000 and D1000 models have the following features in common:

- Eight 1.6-inch or twelve 1-inch Ultra SCSI disk drives
- Dual power supplies
- Dual cooling modules

Note – The disk drives, power supplies, and cooling modules are all hot pluggable.

Sun StorEdge A1000/D1000

Sun StorEdge A1000 Differences

The A1000 is often referred to as the “desktop” hardware RAID solution. It is a standalone, hardware RAID device and is programmed by the RM6 RAID Manager software in exactly the same manner as the Sun StorEdge A3000.

As shown in Figure 1-7, the Sun StorEdge A1000 controllers has two SCSI ports. Usually one port is connected to the host system through a Ultra Differential Fast/ Wide Intelligent SCSI (UDWIS) adapter. The other port is terminated.

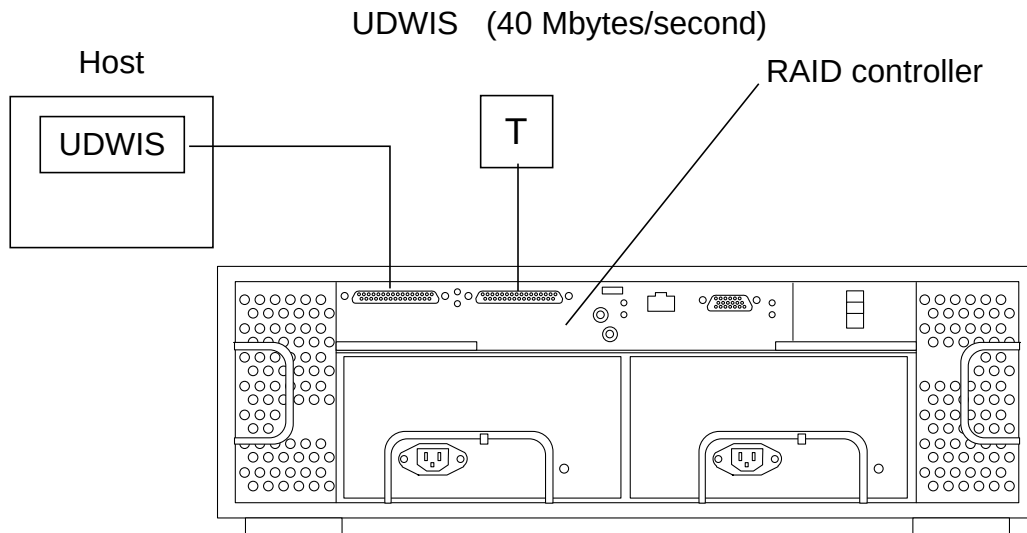


Figure 1-7 Sun StorEdge A1000 Connection

Sun StorEdge A1000 Addressing

The addressing scheme is identical to that used by the Sun StorEdge A3000 unit. The RM6 RAID Manager software takes one or more physical disk drives in the storage tray and configures them as a single LUN, which appears to be a regular physical address such as c2t3d0s2. The underlying configuration is hidden from applications.

There are potential problems with configuring a SSVM software RAID-5 device on top of a RM6 hardware RAID-5 device. You must read the array documentation carefully.

Sun StorEdge A1000/D1000

Sun StorEdge D1000 Differences

As shown in Figure 1-8, the Sun StorEdge D1000 controller has four SCSI ports. The controller can be configured so that half of the disks are connected to one pair of ports and half to the other pair. They can also be configured so that all the disks are available through a single connection. Each pair of ports provides a UDWIS connection.

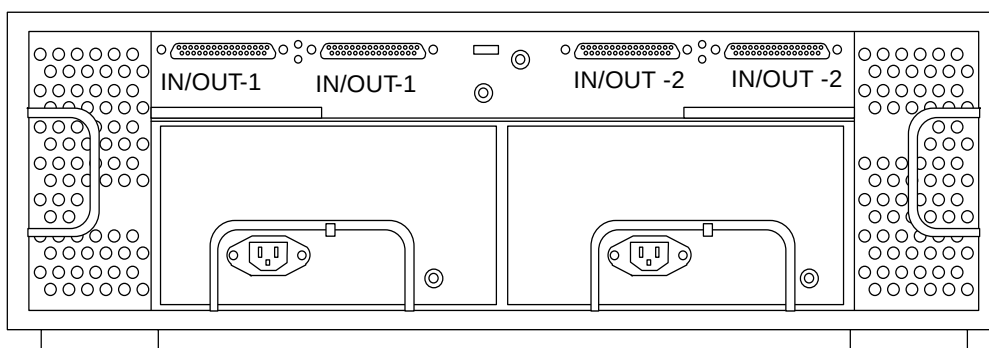
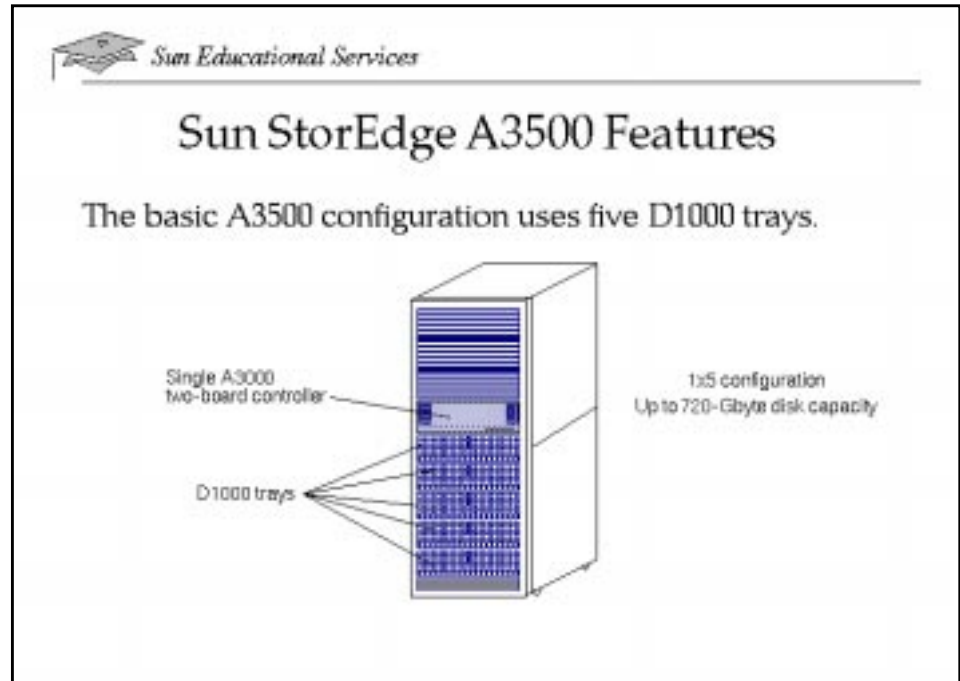


Figure 1-8 Sun StorEdge A1000 Connection

Sun StorEdge D1000 Addressing

The Sun StorEdge D1000 trays are used in exactly the same way the RSM trays are used in the Sun StorEdge A3000 and with the same hardware RAID controller boards.

The addressing scheme is identical to that used by the Sun StorEdge A3000 unit. The RM6 RAID Manager software takes one or more physical disk drives in the storage tray and configures them as a single LUN, which appears to be a regular physical address such as c2t3d0s2. The underlying configuration is hidden from applications.



Sun StorEdge A3500

The Sun StorEdge A3500 unit uses the Sun StorEdge D1000 trays the same way the Sun StorEdge A3000 uses the RSM trays. They are connected using the same two-board controller that is used in the Sun StorEdge A3000. The main difference is the cabinet size.

Sun StorEdge A3500 Features

Depending on its configuration, a Sun StorEdge A3500 system can have up to 2.16-Tbyte disk capacity. The main features are:

- Hardware RAID controller(s)
- Scalable configuration
- A 72-inch rack used to hold up to seven D1000 trays

Sun StorEdge A3500

Sun StorEdge A3500 Features

As shown in Figure 1-9, the Sun StorEdge A3500 array has two additional configurations that can be purchased:

- The 2x7 configuration with two dual-board controllers and seven D1000 trays
- The 3x15 configuration with three dual-board controllers and fifteen D1000 trays

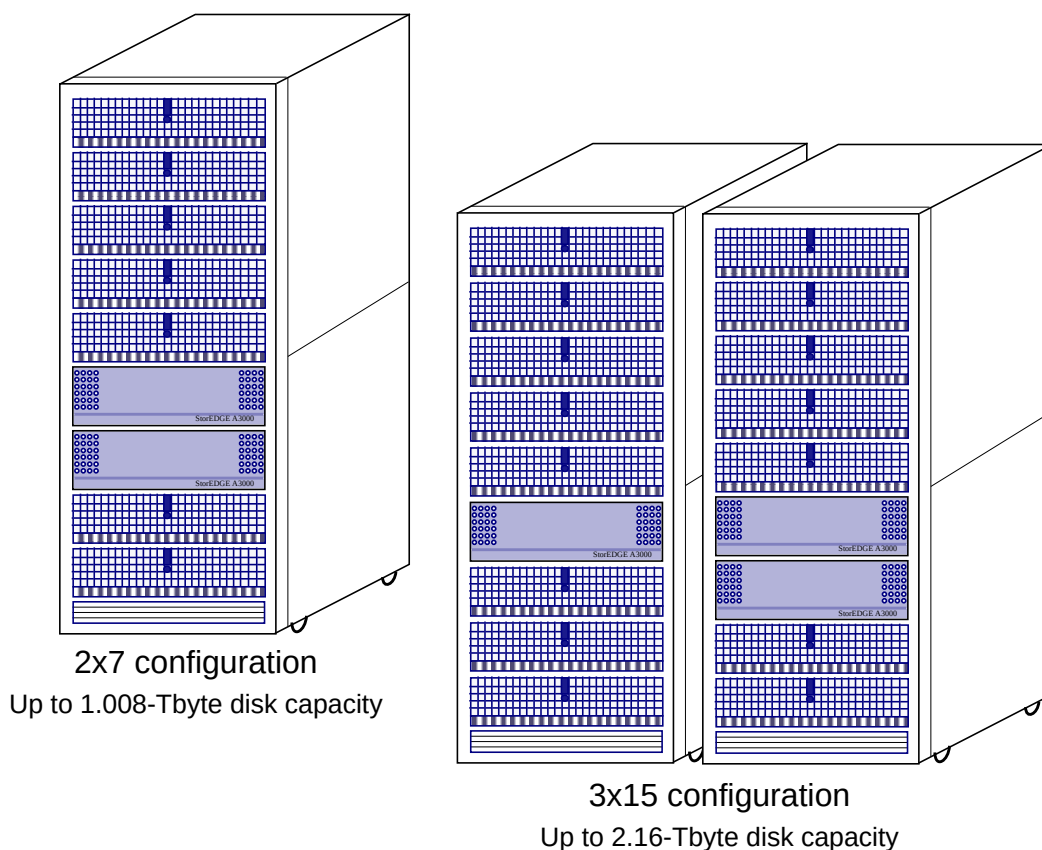
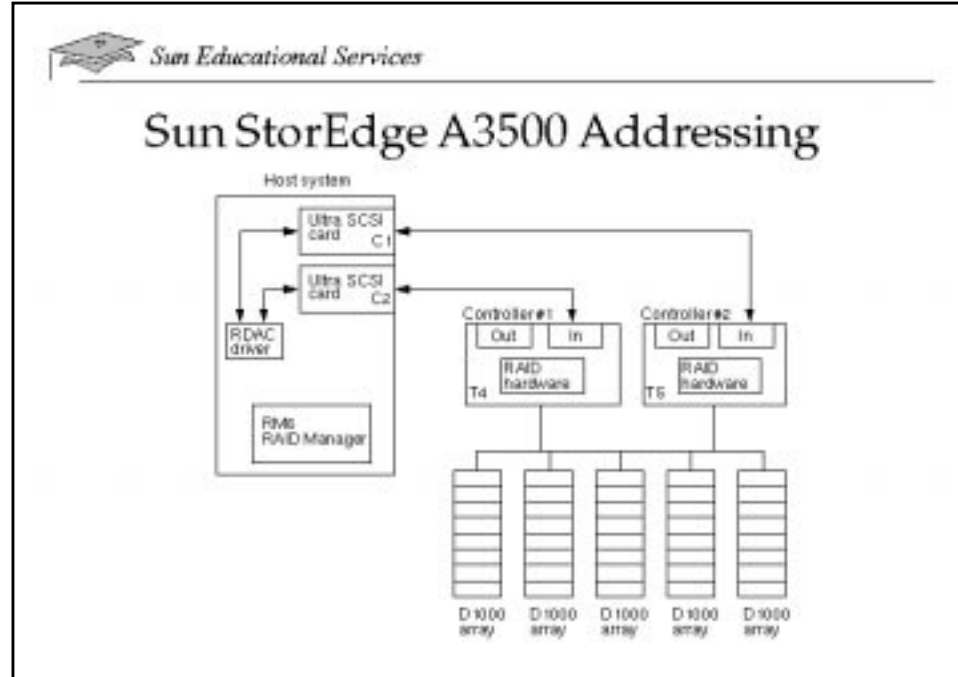


Figure 1-9 Sun StorEdge A3500 Scalability



Sun StorEdge A3500

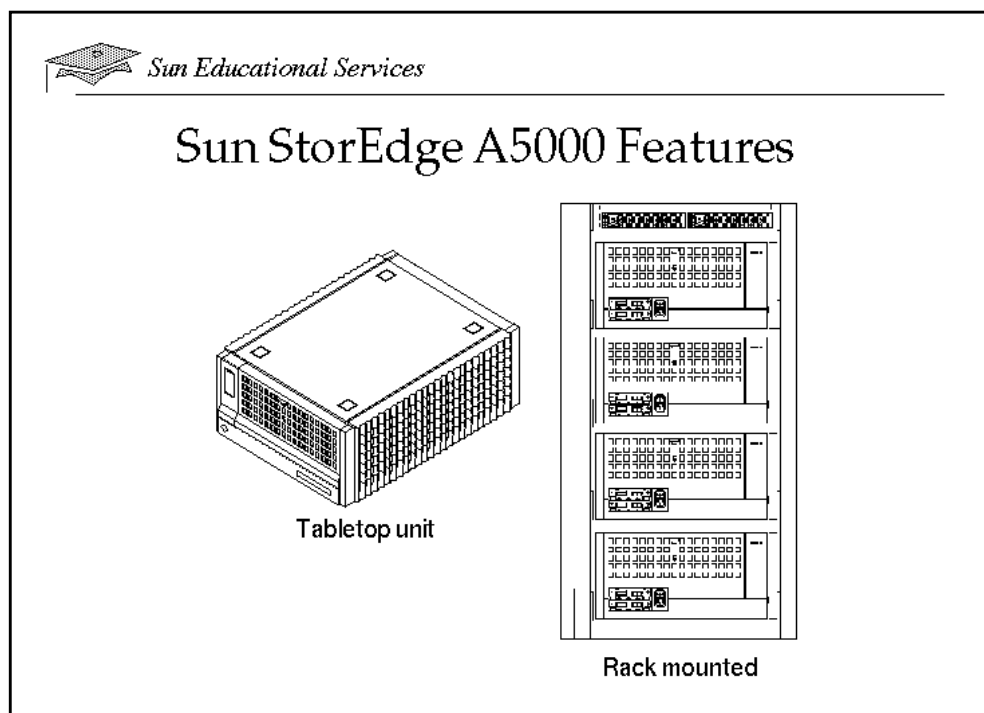
Sun StorEdge A3500 Addressing

The RM6 RAID Manager software takes *one* or more physical disk drives in the storage trays, and configures them as a LUN. This LUN can have a hardware RAID structure.

Once configured, a RM6 LUN appears to be a regular physical address such as c2t4d0s2. The underlying configuration is hidden from applications.



Caution – There are potential problems with configuring a SSVM software RAID-5 device on top of a RM6 hardware RAID-5 device. There will be redundant parity calculations that can cause extremely poor performance.



Sun StorEdge A5000

The Sun StorEdge A5000 is a highly available, mass storage subsystem. The A5000 is the building block for high-performance and high availability configurations with fully redundant, hot-swappable, active components. The A5000 has the highest reliability, availability, and serviceability (RAS) features of any Sun storage array yet.

Sun StorEdge A5000 Features

These include:

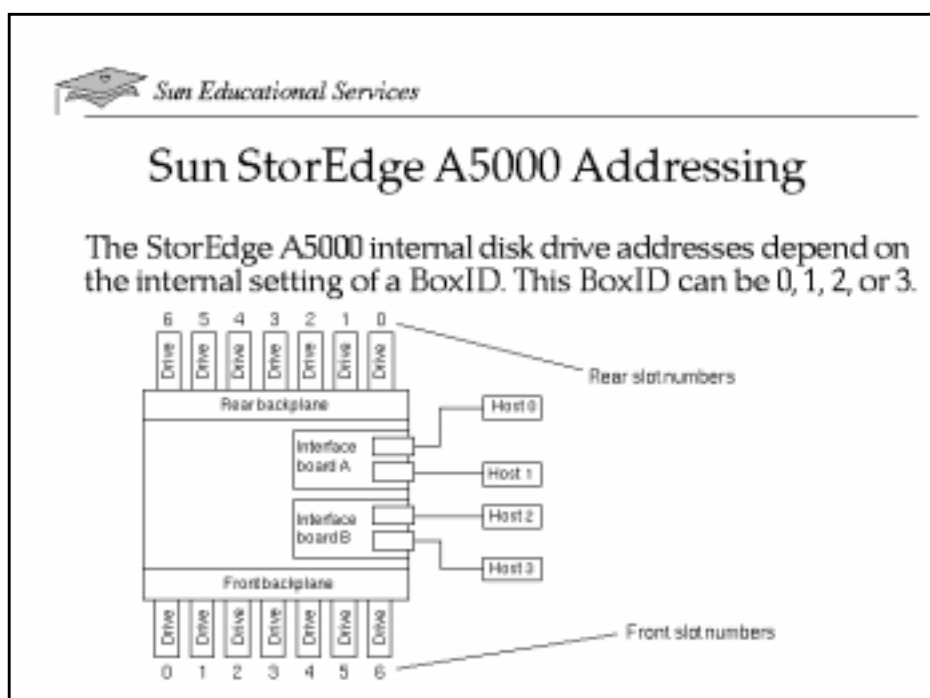
- A Sun second generation Fibre Channel storage subsystem.
- Up to four tabletop units which can be mounted in a 56-inch rack, and up to six tabletop units which can be mounted in a 72-inch rack.
- ▼ Each rack includes two hubs.

Sun StorEdge A5000

Sun StorEdge A5000 Features

- A new way of storing data that is:
 - ▼ Extremely fast (100 Mbytes/second)
 - ▼ Highly available; best RAS features
 - ▼ Scalable in capacity, bandwidth, and I/O rate
- Up to 14 half-height (HH – 1.6 inch) or 22 low profile (LP – 1 inch) hot-pluggable, dual-ported, and Fibre Channel-arbitrated loop (FC-AL) disk drives.
- Two interface boards with GigaBit Interface Converters (GBICs), which provide dual-path capability to the dual-ported disk drives. Two hosts can be attached to each path.
- A Front Panel Module (FPM) which allows the configuration and status of the enclosure to be displayed and modified.
- Active components in the disk enclosure that are redundant and can be replaced while the subsystem is operating.
- An automatic reconfiguration which bypasses whole failed components, or portions thereof.
- An enclosure designed for tabletop use or up to six arrays can be mounted in a standard Sun rack.
- A 123.75-Gbyte *usable raw* formatted capacity in each unit, which supports over 495 Gbytes per loop. (There is a maximum of four units per loop.)

Note – The Sun Enterprise Network Array connects to the host node using the SOC+ FC-AL interface card or built-in FC-AL interfaces in some Sun Enterprise Server I/O boards.



Sun StorEdge A5000

Sun StorEdge A5000 Addressing

The A5000 storage supports a configuration of either 22 disk drives or 14 disk drives. The physical locations are described in terms of front slots 0–10 and rear slots 0–10 (22-drive configuration) or front slots 0–6 and rear slots 0–6 (14-drive configuration). Each box can be assigned a box identifier from 0–3. Each identifier determines a preconfigured address range for the box. Each address is directly related to a SCSI target number (assuming the 14-drive configuration).

The addresses are as follows:

- Box ID 0 addressing

Rear drives: t22 t21 t20 t19 t18 t17 t16

Front drives: t0 t1 t2 t3 t4 t5 t6

Sun StorEdge A5000

Sun StorEdge A5000 Addressing

- Box ID 1 addressing

Rear drives: t54 t53 t52 t51 t50 t49 t48

Front drives: t32 t33 t34 t35 t36 t37 t38

- Box ID 2 addressing

Rear drives: t86 t85 t84 t83 t82 t81 t80

Front drives: t64 t65 t66 t67 t68 t69 t70

- Box ID 3 addressing

Rear drives: t118 t117 t116 t115 t114 t113 t112

Front drives: t96 t97 t98 t99 t100 t101 t102

Physical Addresses

The Box ID addresses create a range of target addresses from 0 to 122 so that up to four A5000 storage arrays can be daisy-chained on a single controller without any SCSI address conflicts. Typical addresses for A5000 storage array devices are:

- c0t3d0s2
- c4t67d0s3
- c3t98d0s2
- c1t6d0s4
- c5t113d0s0
- c2t83d0s4

Note – The target address identifier within the storage array device address is the same as the target number within the daisy-chained array.

Sun StorEdge A5000

Sun StorEdge A5000 Internal Addressing

Although the A5000 target addresses range from 0–128, all the addresses are not used. Configurations that use only the 14-slot arrays will only use fifty-six of the available target IDs for the array devices. Configurations that use only the 22-slot arrays will use as many as eighty-eight of the target IDs for the array devices. Table 1-1 lists the target address assignments for each box ID in a daisy-chained A5000 storage array configuration. It also provides the formula used to compute each target address.

Table 1-1 A5000 SCSI Target Addresses

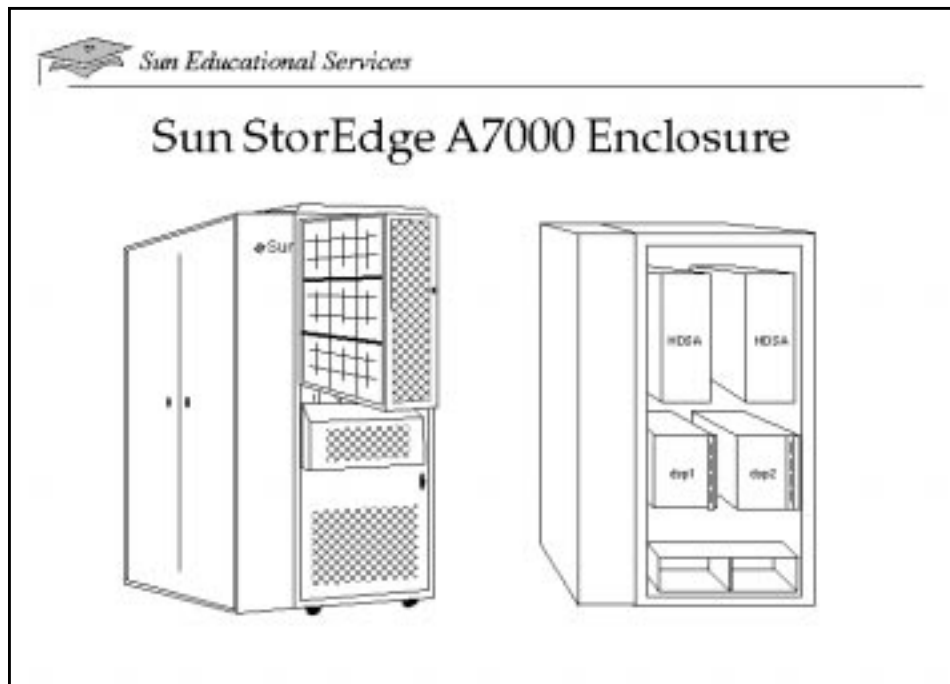
A5000_ID	14-Slot Configuration		22-Slot Configuration	
	Front Drives	Rear Drives	Front Drives	Rear Drives
0	t0–t6	t16–t22	t0–t10	t16–t26
1	t32–t38	t48–t54	t32–t42	t48–t58
2	t64–t70	t80–t86	t64–t74	t80–t90
3	t96–t102	t112–t118	t96–t106	t112–t122

SCSI target# = (A5000_ID × 32) + (16 × backplane#) + slot#

A5000_ID – Is programmed using the FPM

Backplane# – Is 0 for the front and 1 for the rear

Slot# – Is 0–6 and is read from the drive guide below the drive



Sun StorEdge A7000

The Sun StorEdge A7000 intelligent storage servers is a mainframe class subsystem designed to address the storage needs of UNIX[®] and NT[™] hosts as well as IBM and plug-compatible mainframes on a single versatile platform.

Sun StorEdge A7000 Enclosure

In addition to fully redundant hardware, including controllers, cache, hot-pluggable disks, fans, power, and power cords, the Sun StorEdge A7000 enclosure contains two high-density storage array (HDSA) units and two data storage processor (DSP) units.

Sun StorEdge A7000

Sun StorEdge A7000 Enclosure

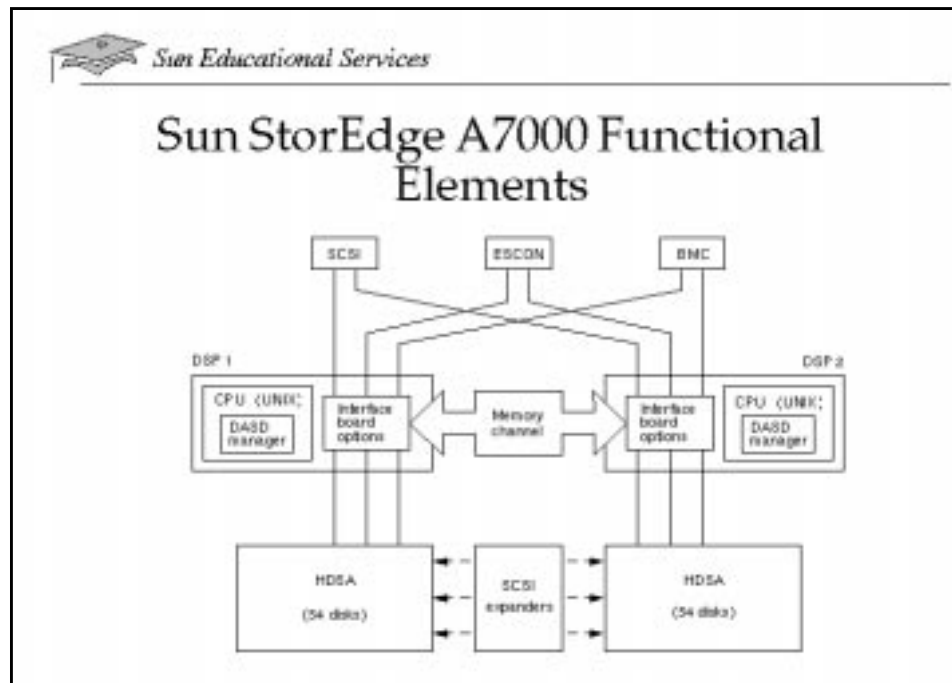
High-Density Storage Array Units

Each HDSA unit can hold up to fifty-four 9.1-Gbyte disk drives. They are housed in removable carriers and, together with software redundancy options, provide hot-swappable disks. They are arranged in six-packs and plug into an intelligent backplane that automatically sets the SCSI ID of the device according to its position in the six-pack. Capacity for the A7000 can be expanded from 24 to 324 disk drives (217-Gbytes to 2.93-Tbytes of total storage) by adding a expansion cabinet containing four additional HDSA units.

Data Storage Processor Units

Each of the DSP unit operates independently and controls one of the HDSA units. Each DSP unit has the following features:

- A 14-slot chassis backplane
- Multiple host system adapters
 - ▼ Quad block multiplexor channel (BMC) adapter
 - ▼ Dual-channel enterprise system connection (ESCON) adapter
 - ▼ SCSI target emulation (STE) adapters
- UNIX Systems laboratories' UNIX System V operating system



Sun StorEdge A7000

Sun StorEdge A7000 Functional Elements

Host Adapter Options

Each DSP unit has five slots available for any mix of SCSI, ESCON, or BMC adapter boards. Simultaneous connections can be made from any of the supported host types.

Sun StorEdge A7000

Sun StorEdge A7000 Functional Elements

Memory Channel Adapter

The two DSP units are connected by a high-speed memory bus. Each DSP unit has up to four memory channel board slots. The memory channel interconnect allows each DSP subsystem to keep the other informed of its state of operation including the presence of any unwritten data. In the event of a DSP failure, the partner DSP unit can take over operation and maintain data integrity.

Direct Access Storage Device Manager

The DASD manager is a GUI tool that enables service personnel to configure the storage on an A7000. The configuration information is stored in the master configuration database (MCD) in each of the DSP units.

The DASD manager can be used to create and manage the following storage configurations:

- Linear partitions
- RAID spare devices
- RAID 5
- RAID 1
- RAID 0
- RAID 1+0

SCSI Expanders

The SCSI expanders allow each DSP unit to access the other's disk storage. This path is used only if one of the DSP units fail.

Sun StorEdge A7000

Sun StorEdge A7000 Addressing

Using the DASD manager, HDSA disks can be configured in a variety of ways. Each type of configuration has associated special device files that can be referenced by SSVM commands and used to build software RAID devices on top of A7000 RAID devices.

A7000 device types and their associated device names are listed in Table 1-2.

Table 1-2 Sun StorEdge A7000 Device Addresses

Special Device	Description
/dev/rdisk/cd4	Linear partitions which function as normal disk partitions. The last segment of the address is determined by the disks physical location in the HDSA.
/dev/rdisk/0r3 /dev/rdisk/0r5	RAID-5 devices that are composed of multiple linear partitions.
/dev/rdisk/mp0	RAID-1 devices which are termed mirrored partitions by the A7000 documentation. They are composed of multiple linear partitions.
/dev/rdisk/vp0	RAID-0 devices which are termed either striped virtual partitions or concatenated partitions by the A7000 documentation. They are composed of multiple linear partitions.
/dev/rdisk/vp0	RAID-1+0 devices.

Sun StorEdge A7000

Combining SSVM and A7000 Devices

Probably the most compelling use for combined host-based and control unit-based RAID is the attainment of very high sequential throughput, such as for large decision-support systems. These systems are usually limited by the bandwidth of the connection between the host and the storage subsystem. As shown in Figure 1-10, a useful configuration takes suitable volumes implemented with A7000-based RAID 5 (or 1/0) and stripes them together with host-based RAID 0.

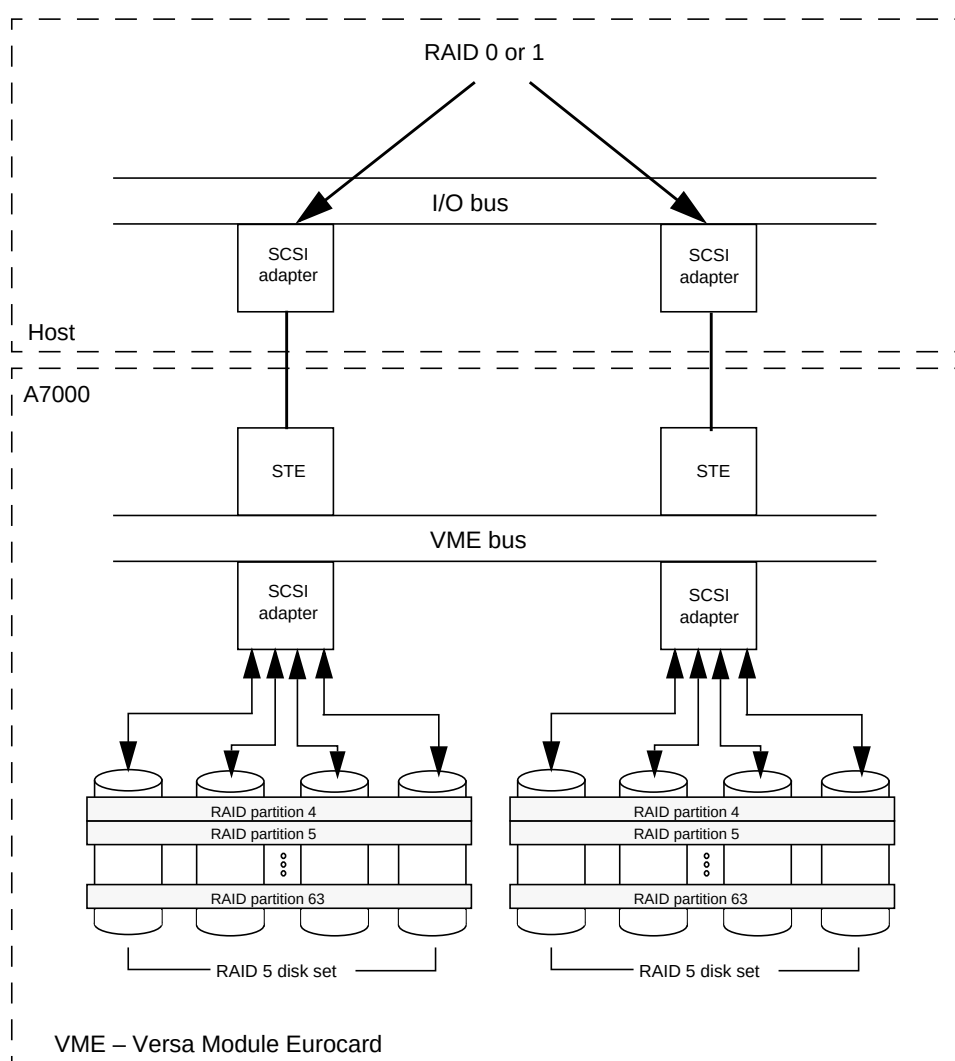
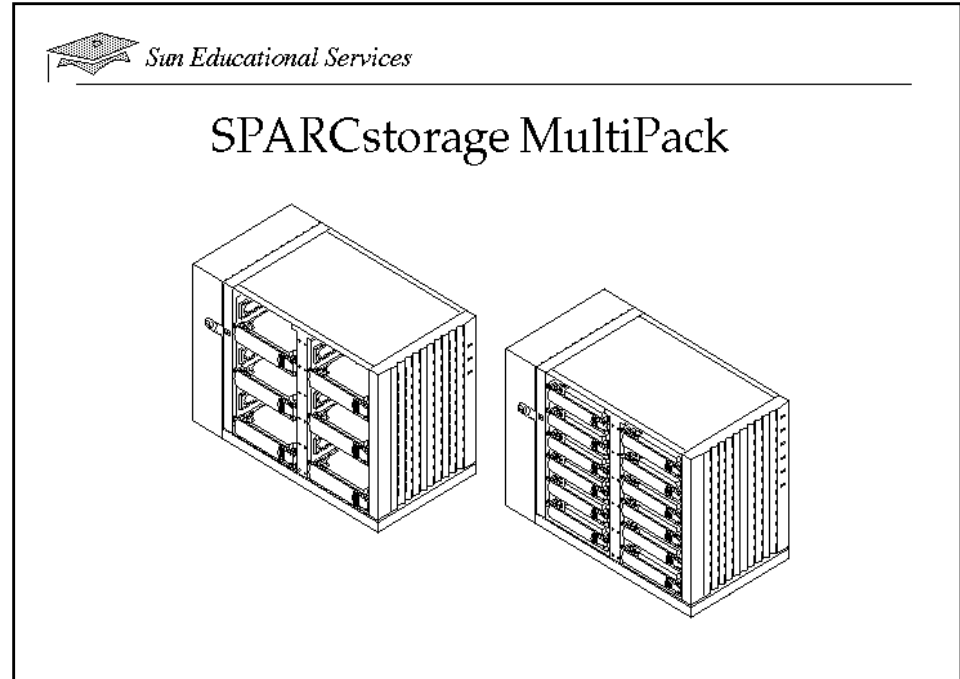


Figure 1-10 Combining Host-based RAID 0 or 1 and A7000-based RAID 5



SPARCstorage MultiPack

The SPARCstorage MultiPack enclosure is a multiple-disk storage device equipped with a fast wide SCSI interface. The Multipack-2 provides an ultra wide SCSI interface. There are two versions of the device:

- SPARCstorage MultiPack unit that supports up to six 1.6-inch high, single-connector disk drives
- SPARCstorage MultiPack unit that supports up to twelve 1-inch high, single-connector disk drives

The Multipack enclosure is 9 inches high.

You can use SPARCstorage MultiPack in a multi-initiated SCSI configuration.

Note – If you do not have SPARCstorage Arrays attached to your system, you will need a special license to use SSVM in a MultiPack only configuration.

SPARCstorage MultiPack

SPARCstorage MultiPack Features

These include:

- A 68-pin fast wide, or ultra wide, SCSI interface
- Drive addresses determined by position (hardwired)
- Six drive units which can be used on a standard 50-pin (narrow) SCSI bus
- A twelve-drive unit only for use on 68-pin (wide) SCSI bus
- Twelve 1.0-inch, 7200 revolutions per minute disks (2.1 or 4.2 Gbytes)
- Six 1.6-inch, 5400 rpm disks (9.1 or 18 Gbytes)

SPARCstorage MultiPack Addressing

The SPARCstorage MultiPack addressing is determined automatically based on the type and physical position of the disks used. The address range is selectable with the six-drive model. The address ranges are as follows:

- The six-drive model addresses are switch selectable and are 1–6 or 9–14
- The twelve-drive model addresses are designed so that addresses 6 and 7 are not used to eliminate `scsi-initiator-id` conflicts.
 - ▼ Addresses 2–5
 - ▼ Addresses 8–15

The addresses directly relate to target numbers. A typical device address path would be `/dev/dsk/c0t8d0s2`.

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Describe the major disk storage administration tasks
- ☐ List the disk storage concepts common to many storage arrays
- ☐ List the general features of current Sun disk storage models
- ☐ Describe the basic Sun StorEdge Volume Manager disk drive replacement process
- ☐ Describe a typical disk replacement process variation

Think Beyond

How much does the physical architecture of each storage array affect RAID design and implementation?

How explicit do you think system error messages are about disk drive failures?

Sun StorEdge Volume Manager Installation

2 

Objectives

Upon completion of this module, you should be able to:

- Describe how the SSVM utilizes disk space
- Install the Sun StorEdge Volume Manager 3.x software
- Explain the difference between the SSVM `vxinstall` Quick Installation and Custom Installation options
- Initialize the SSVM installation with `vxinstall`

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

- I have installed many applications and the process is much the same for all of them. Why are we taking so much care with this installation?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- *Sun StorEdge Volume Manager 3.0 Installation Guide*
- Online man pages for vxinstall, pkginfo, pkgchk, pkgrm, and vxunroot



Installation Process

The Sun StorEdge Volume Manager installation process can be complex and might involve all of the following:

- Pre-installation planning
- Current system checkpoint
- The installation and testing of new configurations

Installation Process

Sun StorEdge Volume Manager installations vary in size from a small desktop system to large servers with terabytes (Tbytes) of data storage. Regardless of the system size, the basic installation process is the same.

Pre-Installation Planning

The software installation process can be a very challenging. The process and method chosen varies with each organization's current configuration and requirements but one variable remains constant: The installation should be carefully planned in advance.

System Downtime

During a new installation or an upgrade, some system downtime is always required. This is usually scheduled during off-peak system usage. Thorough pre-installation planning will usually minimize the system downtime.

Installation Process

Pre-Installation Planning

Disk Space and Swap Space Requirements

It is advisable to determine the amount of space and partition layout which will be required for the new operating system. You may need to consider increasing swap space, or allocating enough space for the application which loads files into the general `sbin` and `usr` directories.

Disk Encapsulation Policy

You must decide which disks will be managed by the Sun StorEdge Volume Manager. As you further configure SSVM, you need to determine the logical organization or grouping of the disks. For example, 10 of the disks will be allocated for accounting's use and 20 disks will be allocated for marketing's use.

You have the option of not placing certain disks under SSVM control. This is useful if you have applications that are currently using file systems or partitions and you do not want to update the application's reference to these file systems or partitions.

You may also want to put your system disk under SSVM control so that all disks can be accessed using a single, easy-to-use disk administration system. This reason, in and of itself, does not warrant this procedure. The primary reason for putting the system disk under SSVM control is so it can be mirrored. Mirroring provides the redundancy which ensures access to the data in case of disk failure. If the root disk is within SSVM control and it is mirrored, the system is still usable after a disk failure.

One major disadvantage to placing your system disk under SSVM control is that recovery (in the event of a failed root disk) is much more complex.

Existing data on disk drives can be *encapsulated*. When data is encapsulated, the partition is added to SSVM's control, leaving the partition intact and maintaining the integrity of the data. This can be done at installation or at a later time.

Installation Process

Pre-Installation Planning

New Hardware Configuration

In addition to having a clear plan for the use of new disk storage devices, you might also need to plan for increased system loads by adding more memory and larger backup tape systems.

Installation Method

With the Custom Installation option, the user manually chooses between encapsulating and initializing a disk or leaving the disk alone. With Quick Installation, all disks with active partitions (including the system disk) are automatically encapsulated, while all other disks are initialized. It is best to use the Custom Installation option.

Upgrade Resources

One of the most frustrating issues can be finding you are missing a CD-ROM, or do not have the needed patches, and in addition, have misplaced the install documents. Having all of the required CD-ROMs, documentation, and patches on the appropriate media, will definitely minimize your frustration. Not only should you have documentation (for example, release notes and installation procedures), you should read it. This is the only way to be assured you have all the needed patches.

Licensing

SSVM uses license keys to control access. If you have a SPARCstorage Array Controller or an A5000 attached to your system, SSVM will grant you a limited-use license automatically. The Array license grants you unrestricted use of disks attached to a SPARCstorage Array Controller or an A5000 interface, but disallows striping and RAID 5 on non-array disks (either SSA or A5000). You may, however, mirror and concatenate non-array drives connected to the same host. Further functionality will require an additional licensing purchase.

Installation Process

Current System Checkpoint

When installing a new and complex application such as the Sun StorEdge Volume Manager, you must always be prepared to return your system to its original state.

Backups

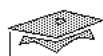
Not only must you have backups, but you must verify them. Should there be a hardware failure or not enough space to facilitate the upgrade, you must be able to recover or back-out the software. Perform a complete backup immediately prior to the installation process.

Documentation of System/Application Configuration

It is critical that you be able to define and possibly reconstruct your configuration. You need to know such items as your printer configuration, routing tables, Domain Name System (DNS) configuration, application-specific files and their contents, and disk and swap configuration. With an upgrade or install of a new operating system, you may be asked to provide the previous functionality. As a result, you need to document not only the general system configuration, but application-specific changes and their associated files.

Installation and Testing of New Configuration

If this is a new install, prior to going into production mode, test it. All Sun's products are extensively tested prior to shipment, however, due to environmental changes and unknown factors in shipping, if time permits, a testing period should be utilized. It would be ideal to test all the components, including the storage subsystem. During this time, any issues related to patches and firmware can be resolved.



Sun Educational Services

Software Installation Process

In order to fully understand a Sun StorEdge Volume Manager installation, you must be familiar with the following aspects:

- SSVM distribution package names and functions
- SSVM package installation method
- SSVM package distribution during installation
- Option-support packages

SSVM Software Installation

With each release of the SSVM product and the Solaris operating system, there can be notable changes so the supporting documentation should be referenced.

With every release of Sun StorEdge Volume Manager, there is an installation guide which accompanies the software. This guide is a comprehensive document which defines the various scenarios for installing and upgrading SSVM as well as the Solaris operating system. These procedures should be followed explicitly.

SSVM Software Installation

Software Package Installation

There are five packages on the CD-ROM. These packages are listed on the next page. They include online manual pages, online documentation (AnswerBook™), and the Sun StorEdge Volume Manager. To use them:

1. Install the Solaris operating environment.
2. Install any Solaris patches.
3. Install the Sun StorEdge Volume Manager 3.0 packages.
4. Install any necessary SSVM patches.
5. If any patches were installed, reboot the system. Otherwise continue with initializing the software.
6. Go through the `vxinstall` procedure to initialize the SSVM software.

Note – These installation instructions apply to a fresh install. If you are upgrading from a previous installation, refer to the installation guide and release notes.

SSVM Software Installation

Software Distribution

The SSVM software distribution CD-ROM contains the following packages:

Package	Title
VRTSvmdev	VERITAS Volume Manager, Header and Library Files
VRTSvmdoc	VERITAS Volume Manager (user documentation)
VRTSvmman	VERITAS Volume Manager, Manual Pages
VRTSvmsa	VERITAS Volume Manager Storage Administrator
VRTSvxvm	VERITAS Volume Manager, Binaries

Software Installation

The SSVM software packages are installed using the `pkgadd` command as follows:

```
# pkgadd -d pwd
```

The following packages are available:

1	VRTSvmdev	VERITAS Volume Manager, Header and Library Files (sparc) 3.0.2,REV=08.30.1999.15.56
2	VRTSvmdoc	VERITAS Volume Manager (user documentation) (sparc) 3.0.2,REV=08.26.1999.22.51
3	VRTSvmman	VERITAS Volume Manager, Manual Pages (sparc) 3.0.2,REV=08.30.1999.15.55
4	VRTSvmsa	VERITAS Volume Manager Storage Administrator (sparc) 3.0.3,REV=08.27.1999.13.55
5	VRTSvxvm	VERITAS Volume Manager, Binaries (sparc) 3.0.2,REV=08.30.1999.15.56

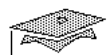
Select package(s) you wish to process (or 'all' to process all packages). (default: all) [?,??,q]: **all**

SSVM Software Installation

Option Support Packages

Depending on your system's hardware configuration, additional packages might be needed to support special options. Read all installation documentation thoroughly before starting an installation.

Generally, special support software comes bundled with the related product and is already installed. A typical example is the special software drive support required for some of the newer storage arrays.



Sun Educational Services

Initializing the Sun StorEdge Volume Manager

The SSVM software is initialized by placing at least one disk drive under SSVM control using the SSVM `vxinstall` program. The process has the following phases:

- Automatic disk drive exclusion
- Option selection
- Boot disk encapsulation
- Initialization selection
- System reconfiguration

Initializing the Sun StorEdge Volume Manager

The SSVM software will not start correctly at system boot time until at least one disk drive is placed under SSVM control. The initialization is performed using the `vxinstall` utility. It is important that you understand this process, especially during a software upgrade.

Note – The `vxinstall` utility does not make any modifications until you examine and approve a summary of your selections near the end of the process.

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

The `vxinstall` program first searches for attached controllers on the system and then prompts you for an installation option – Quick Installation or Custom Installation. You must understand the implications of each option before proceeding.



Warning – `vxinstall` should only be run once! If it is run multiple times, SSVM is likely to start behaving strangely. If you forget to add a disk during `vxinstall`, do not run the utility again; add the disk later using the GUI or command-line interpreter (CLI).

Exclusion of Disks and Controllers

You can create the `/etc/vx/disks.exclude` file and list the disks that you want to exclude from SSVM control. Each disk entry must be on a separate line. You can also create an `/etc/vx/cntrls.exclude` file to exclude controllers from installation.

Note – You cannot exclude all disks: at least one disk must be added during the `vxinstall` process. Any excluded disks can be added later, using either the SSVM GUI or CLI.

The vxinstall Startup Sequence

The `vxinstall` utility scans the system and attempts to identify all disk controller interfaces. The controllers listed can include your system boot disk and any other disks that might have mounted file systems.

You must be careful how you proceed during the initialization process. Before starting, understand your system hardware configuration.

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

The vxinstall Startup Sequence

```
# vxinstall
Generating list of attached controllers....
-/\-/\-/\-/\-/\-/\
Volume Manager Installation
Menu: Volume Manager/Install
```

The Volume Manager names disks on your system using the controller and disk number of the disk, substituting them into the following pattern:

```
c<controller> t<disk> d<disk>
```

If the Multipathing driver is installed on the system then for the disk devices with multiple access paths, the controller number represents a multipath pseudo controller number. For example, if a disk has 2 paths from controllers c0 and c1, then the Volume Manager displays only one of them such as c0 to represent both the controllers.

Some examples would be:

```
c0t0d0 - first controller, first target, first disk
c1t0d0 - second controller, first target, first disk
c1t1d0 - second controller, second target, first disk
```

The Volume Manager has detected the following controllers on your system:

```
c0:
c2:
c4:
```

Hit RETURN to continue.

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

The vxinstall Option Selection

After identifying all available disk controllers, vxinstall prompts you for an installation option.

Volume Manager Installation
Menu: Volume Manager/Install

You will now be asked if you wish to use Quick Installation or Custom Installation. Custom Installation allows you to select how the Volume Manager will handle the installation of each disk attached to your system.

Quick Installation examines each disk attached to your system and attempts to create volumes to cover all disk partitions that might be used for file systems or for other similar purposes.

If you do not wish to use some disks with the Volume Manager, or if you wish to reinitialize some disks, use the Custom Installation option. Otherwise, we suggest that you use the Quick Installation option.

Hit RETURN to continue.

Volume Manager Installation Options
Menu: Volume Manager/Install

- 1 Quick Installation
- 2 Custom Installation

- ? Display help about menu
- ?? Display help about the menuing system
- q Exit from menus

Select an operation to perform: **2**

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

The vxinstall Option Selection

The Quick Installation is not recommended, as the default action is to encapsulate all disks, including the root disk. This option includes the following features:

- Examines all disks connected to the system, and either encapsulates existing partitions (placing these partitions under SSVM's control, leaving them intact and maintaining the integrity of the data), or initializes disks that do not have existing partitions.
- Adds all disks to the default disk group, rootdg.
- Updates `/etc/vfstab` to ensure that file systems previously mounted on disk partitions will be mounted on volumes instead.

The Custom Installation enables control over which disks are placed under SSVM control and how they are added (encapsulated or initialized).

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

Boot Disk Encapsulation

During the Custom Installation, vxinstall is aware of disks that contain functional data.

Volume Manager Custom Installation
Menu: Volume Manager/Install/Custom

The c0t0d0 disk is your Boot Disk. You can not add it as a new disk. If you encapsulate it, you will make your root filesystem and other system areas on the Boot Disk into volumes. This is required if you wish to mirror your root filesystem or system swap area.

Encapsulate Boot Disk [y,n,q,?] (default: n) **n**

Volume Manager Custom Installation
Menu: Volume Manager/Install/Custom/c0
Generating list of attached disks on c0....

<excluding root disk c0t0d0>
No disks were found attached to controller c0 !
Hit **RETURN** to continue.

Note – The encapsulation process is more complex than simple initialization and will be discussed in a later module.

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

Selective Initialization

If you are not sure which disks to initialize, choose the option that prompts for one disk at a time.

The Volume Manager has detected the following disks on controller c2:

c2t33d0 c2t35d0 c2t37d0 c2t50d0 c2t52d0

Installation options for controller c2

Menu: Volume Manager/Install/Custom/c2

- 1 Install all disks as pre-existing disks(encapsulate)
- 2 Install all disks as new disks. (discards data on disks!)
- 3 Install one disk at a time.
- 4 Leave these disks alone.

Select an operation to perform: **3**

Installation options for disk c2t33d0

Menu: Volume Manager/Install/Custom/c2/c2t33d0

- 1 Install as a pre-existing disk. (encapsulate)
- 2 Install as a new disk. (discards data on disk!)
- 3 Leave this disk alone.

Select an operation to perform: **2**

Are you sure (destroys data on c2t33d0) [y,n,q,?] (default: n) **y**

Enter disk name for c2t33d0 [<name>,q,?] (default: disk01) **newroot**

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

Selective Initialization

You are presented with the following options for each controller:

- Install as pre-existing disks (encapsulate) – If you choose this option, a volume will be created, which encapsulates any partitions on this disk. The `/etc/vfstab` file will be updated to ensure that file systems previously mounted on disk partitions will be mounted on volumes instead.
- Install as new disks (discard data) – All disks on this controller will be re-initialized. This destroys all data on the disks and makes the disks available as free space for allocating new volumes or as mirrors of existing volumes.
- Install one disk at a time – You will be prompted for each disk, whether to encapsulate it, install as a new disk, or leave it alone.
- Leave alone – These disks will not be brought under SSVM control. If there are applications that use this disk that you do not want to upgrade to use SSVM, use this option to ensure that the applications can continue to use the disk without modification.

By default, all disks are added to the disk group `rootdg` during the `vxinstall` process. If you want to create additional disk groups, it is easiest to use Custom Installation and choose to leave some disks alone. Then, later use the GUI or CLI to add the excluded disks to a different disk group. Note, however, that you must add at least one disk to `rootdg` during the `vxinstall` process.

Initializing the Sun StorEdge Volume Manager

The vxinstall Program

Completion

The vxinstall program does not initialize or alter any disks until the selection process is complete. You can choose to quit at any time until the very end of the process.

The following is a summary of your choices.

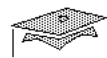
c2t33d0 New Disk

Is this correct [y,n,q,?] (default: y) **y**

```
The Volume Manager is now reconfiguring (partition phase)...
Volume Manager: Partitioning c2t33d0 as a new disk.
The Volume Manager is now reconfiguring (initialization phase)...
Volume Manager: Adding newroot(c2t33d0) as a new disk.
The Volume Daemon has been enabled for transactions.
```

The system now must be shut down and rebooted in order to continue the reconfiguration.

Shutdown and reboot now [y,n,q,?] (default: n) **y**



SSVM Disk Management

When the Sun StorEdge Volume Management brings disk drives under its control, it creates two regions on the disks:

- Private region
- Public region

SSVM Disk Management

Physical Disk Layout

As shown in Figure 2-1, a physical disk drive that has been initialized by SSVM is divided into two sections called the *private* region and the *public* region.

- The private region is used for configuration information.
- The public region is used for data storage

SSVM Disk Management

Physical Disk Layout

By default, SSVM uses partitions 3 and 4 for the private and public regions.

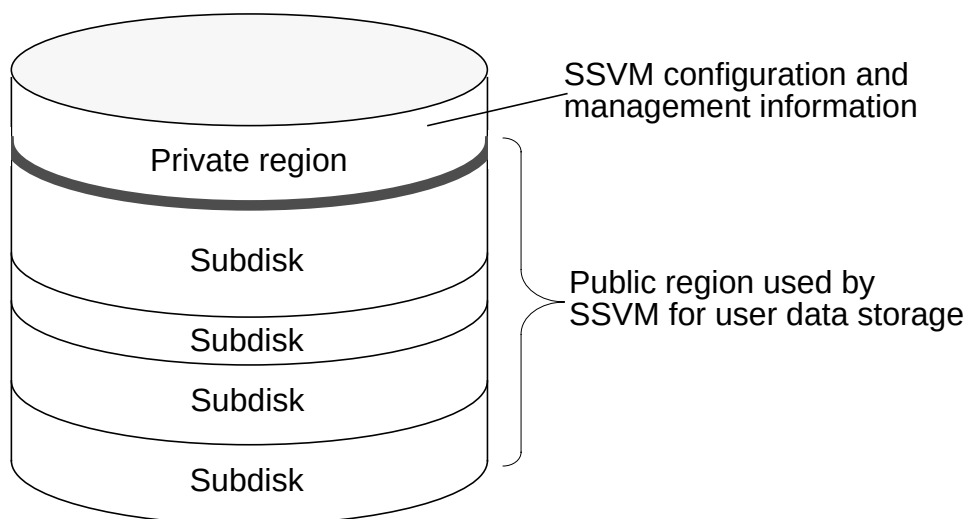
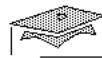


Figure 2-1 SSVM Physical Disk Layout

SSVM will take one cylinder for the private region which will vary in size depending on the geometry of the drive. On the larger drives, one cylinder can store more than a megabyte (Mbyte).

The public region is configured to be the rest of the physical disk.



Private Region Usage

The private region of a SSVM initialized disk drive contains the following management information:

- Disk header
- Configuration database
- Kernel log

Private Region Usage

Disk Header

The disk header is a block stored in the private region of a disk that defines the following important properties of the disk:

- Current host ownership of the disk drive

When a disk is part of a disk group that is in active use by a particular host, the disk is stamped with that host's host ID (hostname).

If another system attempts to access the disk, it will detect that the disk has a nonmatching host ID (hostname) and will disallow access until the first system discontinues use of the disk.

- Disk identifier

A 64-byte unique identifier is assigned to a physical disk when its private region is initialized.

Private Region Usage

Configuration Database

The configuration database contains information about the configuration of a particular disk group. By default, SSVM keeps four copies of the configuration database per disk group to avoid any possibility of losing your disk group information.

Each copy of the configuration database contains the following information:

- *dgname* – The name of the disk group that is assigned by the administrator.
- *dgid* – A 64-byte universally unique identifier that is assigned to a disk group when the disk group is created. This identifier is in addition to the disk group name, which is assigned by the administrator. The disk group ID is used to check for disk groups that have the same administrator-assigned name but are actually distinct.
- Records – One record for each SSVM object (volume, plex, subdisk, and so on).

Note – Not all of the private regions have a copy of the configuration database.

Kernel Log

The kernel log is kept in the private region on the disk and is written by the SSVM kernel. The log contains records describing certain types of actions such as transaction commits, plex detaches resulting from I/O failures, dirty region log failures, first write to a volume, and volume close. It is used after a crash or clean reboot to recover the state of the disk group just prior to the crash or reboot.

Private Region Usage

Overriding Default Values

Occasionally, SSVM administrators feel they must set up a disk group using non-standard values such as:

- A private region larger than the default 1024 sectors
- A greater number of configuration databases per disk group
- A greater number of kernel log records per disk group

Preparing for Large configdb Records

When disks are first initialized for SSVM use, the size of the private region can be made larger than the 1024 sector default. Do this if you anticipate more than 2000 SSVM objects in a disk group; for example, if you plan to create a large number of small mirrored volumes with several subdisks in each plex. The command-line format is as follows:

```
# vxdisksetup -i c2t3d0 privlen=10080s
```

Specifying configdb and log Records

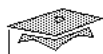
When a new disk group is first created and the first disk is added to it, you can specify the total number of configdb and log records that will be distributed throughout the disks that are added later. The command line to initialize a new disk group and add the first disk to it is as follows:

```
# vxdg init group1 p001=c2t0d0 nconfig=20 nlog=20
```

In this example, a disk name, p001, has been given to the disk and the number of configdb and log record copies has been set to 20. The disk name can be used later with many SSVM commands as a substitute for the physical path name (c2t0d0).

The configdb and log record copies will be created as necessary when more disks are added to the newly created disk group. The command-line format to add disks to an existing disk group is:

```
# vxdg -g group1 adddisk p002=c2t0d1
```

Sun Educational Services

SSVM Environment

During the SSVM installation, software is distributed to several locations. You should be familiar with the following:

- Location and function of SSVM startup files
- SSVM system boot messages and processes
- Location of important binaries and scripts

SSVM Environment

Once the SSVM software is installed and initialized, you must be familiar with the general environment if you are to be an effective administrator.

SSVM System Startup Files

During SSVM installation, the following changes are made to the `/etc/system` file and SSVM startup files are added to the `/etc/rcX.d` directories.

/etc/system File Changes

Entries are added to the `/etc/system` file to force load the vx device drivers (`vxio`, `vxspec`, and `vxdmp`).

SSVM Environment

SSVM System Startup Files

/etc/rcX.d Script File Additions

A number of SSVM script files are added to the `/etc/rcX.d` directories to start the SSVM software at boot time.

- `/etc/rcS.d/S25vxvm-sysboot` – This file runs early in the boot sequence to configure the `/` and `/usr` volumes. This file also contains configurable debugging parameters.
- `/etc/rcS.d/S35vxvm-startup1` – This file runs after `/` and `/usr` are available, and makes other volumes available that are needed early in the Solaris boot sequence.
- `/etc/rcS.d/S85vxvm-startup2` – This file starts I/O daemons, rebuilds the `/dev/vx/dsk` and `/dev/vx/rdisk` directories, imports all disk groups, and starts all volumes that were not started earlier in the boot sequence.
- `/etc/rcS.d/S86vxvm-reconfig` – This file contains commands to execute `fsck` on the root partition before anything else on the system executes.
- `/etc/rc2.d/S95vxvm-recover` – This file attaches plexes and starts SSVM watch daemons.
- `/etc/rc2.d/S96vmsa-server` – This file starts the new SSVM command server that responds to the remote client software.

SSVM Environment

System Startup Messages

When the Solaris operating system is booted and the SSVM startup files execute, several important boot messages are displayed:

```
Rebooting with command: boot
Boot device: disk File and args:
SunOS Release 5.7 Version Generic 64-bit [UNIX(R) System V Release 4.0]
Copyright (c) 1983-1998, Sun Microsystems, Inc.
VxVM starting in boot mode...
configuring network interfaces: hme0.
Hostname: devsys1
VxVM general startup...
The system is coming up. Please wait.
checking ufs filesystems
/dev/rdisk/c0t0d0s3: is clean.
starting routing daemon.
starting rpc services: rpcbind keyserv
keyserv done.
Setting default interface for multicast: add net 224.0.0.0: gateway
devsys1
syslog service starting.
Print services started.
volume management starting.
Starting RMI Registry
Starting Sun StorEdge VM Command Server
Starting Sun StorEdge VM Server
The system is ready.
```

SSVM Environment

System Startup Processes

`vxconfigd`

The volume configuration daemon (`vxconfigd`) is started by the `S25vxvm-sysboot` script early in the boot process. It needs to be running in order for the SSVM software to function. The default disk group, `rootdg`, must be configured in order for this daemon to start.

`vxrelocd` or `vxsparecheck`

One or the other will be started by the `S95vxvm-recover` script during the boot process.

`vxnotify`

This daemon is started by either the `vxrelocd` or the `vxsparecheck` script file.

`vxrecover`

This daemon can be run by either the `S35vxvm-startup1`, `S85vxvm-startup2`, or `S95vxvm-recover` scripts during a system boot, depending on the need for volume repair.

`vmsa_server`

The `S96vmsa-server` script starts the `/opt/SUNWvmsa/bin/vmsa_server` file in the background. The `vmsa_server` script starts two Java™ (`jre`) processes and one `cmdserver` process.

SSVM Environment

System and User Executable Files

SSVM Software in /opt

These include:

- /opt/SUNWvxvm – Header files and man pages
- /opt/SUNWvmsa – SSVM server software

SSVM Software in /usr/sbin

These include:

vxassist	vxctl	vxdg	vxdisk	vxdiskadd
vxdiskadm	vxedit	vxinfo	vxinstall	vxiod
vxlicense	vxmake	vxmend	vxnotify	vxplex
vxprint	vxrecover	vxrelayout	vxsd	vxserial
vxstat	vxtask	vxtrace	vxvol	

SSVM Environment

System and User Executable Files

SSVM Software in /etc/vx/bin

These include:

egettxt	strtovoff	vxapslice	vxbootsetup
vxcap-part	vxcap-vol	vxcheckda	vxchksundev
vxckdiskrm	vxcntrllist	vxdevlist	vxdiskrm
vxdisksetup	vxdiskunsetup	vxdmpadm	vxedvtoc
vxeeprom	vxencap	vxevac	vxmirror
vxmkboot	vxmksdpart	vxnewdmname	vxparms
vxpartadd	vxpartinfo	vxpartrm	vxpartrmall
vxprvtoc	vxr5check	vxr5vrfy	vxreattach
vxrelocd	vxresize	vxroot	vxrootmir
vx slicer	vxspare	vxsparecheck	vxswapreloc
vxtaginfo	vxunroot		

Note – At start-up time, the volume configuration daemon, /sbin/vxconfigd, is started.

Exercise: Configuring the Sun StorEdge Volume Manager



Exercise objective – In this exercise you will:

- Install the SSVM software packages
- Initialize the SSVM installation
- Verify proper SSVM startup at boot time
- Verify the appropriate SSVM processes are running

Preparation

This exercise is to be performed as a group. Ask your instructor to furnish the following information:

- A diagram of your classroom configuration.
- The location of the SSVM software. It might be on a CD-ROM or it can be NFS™ mounted.

SSVM location: _____

SSVM location: _____

- The physical address of the disk or disks to be initialized.

Disk address: _____

Disk address: _____

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Installing the SSVM Software

Complete the following steps:

1. Log in as user `root` on the system attached to the storage arrays.
2. Either insert the Sun StorEdge Volume Manager 3.0 CD-ROM into the CD-ROM drive or change directory to a location furnished by your instructor.
3. If you are working from the SSVM 3.0 CD-ROM, change to the `/cdrom/sun_ssvm_3_0_sparc/Product` directory.
4. Verify you are in the correct location.

```
# ls
```

```
VRTSvmdev VRTSvmman VRTSvxvm
VRTSvmdoc VRTSvmsa
```

5. Install all the packages in this directory.

```
# pkgadd -d `pwd`
```

The following packages are available:

1	VRTSvmdev	VERITAS Volume Manager, Header and Library Files (sparc) 3.0.2,REV=08.30.1999.15.56
2	VRTSvmdoc	VERITAS Volume Manager (user documentation) (sparc) 3.0.2,REV=08.26.1999.22.51
3	VRTSvmman	VERITAS Volume Manager, Manual Pages (sparc) 3.0.2,REV=08.30.1999.15.55
4	VRTSvmsa	VERITAS Volume Manager Storage Administrator (sparc) 3.0.3,REV=08.27.1999.13.55
5	VRTSvxvm	VERITAS Volume Manager, Binaries (sparc) 3.0.2,REV=08.30.1999.15.56

Select package(s) you wish to process (or 'all' to process all packages).
(default: all) `[?,??,q]: all`

After the installation of all the SSVM packages, you are ready to initialize the SSVM software.

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Initializing the SSVM Software

Use the following steps:

1. Start the initialization process by running the `vxinstall` program.

```
# vxinstall
```

Note – See your instructor for additional information on how to initialize disk for your lab environment.

2. Choose the Custom Installation option.

Volume Manager Installation Options

Menu: Volume Manager/Install

- 1 Quick Installation
- 2 Custom Installation

- ? Display help about menu
- ?? Display help about the menuing system
- q Exit from menus

Select an operation to perform: **2**

Volume Manager Custom Installation

Menu: VolumeManager/Install/Custom

The c0t0d0 disk is your Boot Disk. You can not add it as a new disk. If you encapsulate it, you will make your root file system and other system areas on the Boot Disk into volumes. This is required if you wish to mirror your root file system or system swap area.

Encapsulate Boot Disk [y,n,q,?] (default: n) n

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Initializing the SSVM Software



Warning – Do not encapsulate or initialize the system disk. Choose the option to “leave this disk alone.”

Volume Manager Custom Installation

Menu: VolumeManager/Install/Custom/c0

Generating list of attached disks on c0....

-<excluding root disk c0t0d0>

/|\

The Volume Manager has detected the following disks on controller c0:

c0t1d0 c0t2d0 c0t3d0

Hit RETURN to continue.

Installation options for controller c0

Menu: VolumeManager/Install/Custom/c0

1Install all disks as pre-existing disks. (encapsulate)

2Install all disks as new disks. (discards data on disks!)

3Install one disk at a time.

4Leave these disks alone.

?Display help about menu

??Display help about the menuing system

qExit from menus

Select an operation to perform: 3

Note – Selecting menu option 3 enables you to answer initialization questions on a disk-by-disk basis.

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Initializing the SSVM Software

Installation options for disk c0t1d0

Menu: VolumeManager/Install/Custom/c0/c0t1d0

- 1Install as a pre-existing disk. (encapsulate)
- 2Install as a new disk. (discards data on disk!)
- 3Leave this disk alone.

?Display help about menu

??Display help about the menuing system

qExit from menus

Select an operation to perform: 2



Caution – Select option 3 for the remaining disks to ensure only one disk is initialized.

Enter disk name for c0t1d0 [<name>,q,?] (default: disk01)

Installation options for disk c0t2d0

Menu: VolumeManager/Install/Custom/c0/c0t2d0

- 1Install as a pre-existing disk. (encapsulate)
- 2Install as a new disk. (discards data on disk!)
- 3Leave this disk alone.

?Display help about menu

??Display help about the menuing system

qExit from menus

Select an operation to perform: 3

Note – Continue the initialization according to the guidelines defined by your instructor.

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Initializing the SSVM Software



3. Proceed through the disk selection process until you get to the selection summary.

Volume Manager Custom Installation
Menu: Volume Manager/Install/Custom

The following is a summary of your choices.

c2t33d0 New Disk

Note – This message will vary according to the devices that were configured by you during the previous initialization exercise.

Is this correct [y,n,q,?] (default: y)

4. Stop at this point.

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Initializing the SSVM Software

5. Before proceeding with the initialization, verify that the only disk selected for initialization meets the following criteria:
 - ▼ It is not the system boot disk
 - ▼ Only a single disk is selected
 - ▼ It has the physical address furnished by your instructor
6. If the selected disk drive initializes without errors, reply **yes** to the rebooting prompt.

The Volume Daemon has been enabled for transactions.

The system now must be shut down and rebooted in order to continue the reconfiguration.

Shutdown and reboot now [y,n,q,?] (default: n) **y**

Task – Verifying the SSVM Startup

Complete these steps:

1. When the system reboots, verify that the following messages display during the reboot operation:

VxVM starting in boot mode
VxVM general startup
Starting RMI Registry
Starting Sun StorEdge VM Command Server
Starting Sun StorEdge VM Server

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Verifying the SSVM System Processes

Complete the following steps:

1. Log in as user `root` and use the `ps -e` command to verify that the following processes are present:
 - ▼ `vxconfigd`
 - ▼ `vxrelocd` or `vxsparecheck`
 - ▼ `vxnotify`
 - ▼ `vmsa_server`
 - ▼ `cmdserver`
 - ▼ `jre`

Exercise: Configuring the Sun StorEdge Volume Manager

Task – Verifying the SSVM System Files

Use the following steps:

1. Verify that the `SUNWvxvm` and `SUNWvmsa` directories are present in `/opt`.
2. Verify that the SSVM software is present in `/usr/sbin`.

cd /usr/sbin; ls vx*

<code>vxassist</code>	<code>vxdisk</code>	<code>vxinfo</code>	<code>vxmend</code>	<code>vxrelayout</code>	<code>vxtrace</code>
<code>vxclust</code>	<code>vxdiskadd</code>	<code>vxinstall</code>	<code>vxnotify</code>	<code>vxsd</code>	<code>vxvol</code>
<code>vxconfigd</code>	<code>vxdiskadm</code>	<code>vxiod</code>	<code>vxplex</code>	<code>vxserial</code>	
<code>vxctl</code>	<code>vxdkpdm</code>	<code>vxlicense</code>	<code>vxprint</code>	<code>vxstat</code>	
<code>vxdg</code>	<code>vxedit</code>	<code>vxmake</code>	<code>vxrecover</code>	<code>vxtask</code>	

3. Verify the SSVM software is present in `/etc/vx/bin`.

cd /etc/vx/bin; ls vx*

<code>vxapslice</code>	<code>vxcntrllist</code>	<code>vxencap</code>	<code>vxpartinfo</code>	<code>vxroot</code>
<code>vxbadxcld</code>	<code>vxxcld</code>	<code>vxevac</code>	<code>vxpartrm</code>	<code>vxrootmir</code>
<code>vxbadxcld</code>	<code>vxdevlist</code>	<code>vxliccheck</code>	<code>vxpartrmall</code>	<code>vxlicer</code>
<code>vxbootsetup</code>	<code>vxdiskrm</code>	<code>vxmirror</code>	<code>vxprtvtoc</code>	<code>vxspare</code>
<code>vxcap-part</code>	<code>vxdisksetup</code>	<code>vxmkboot</code>	<code>vxr5check</code>	<code>vxsparecheck</code>
<code>vxcap-vol</code>	<code>vxdiskunsetup</code>	<code>vxmkbdpart</code>	<code>vxr5vrfy</code>	<code>vxswapreloc</code>
<code>vxcheckda</code>	<code>vxdxcl</code>	<code>vxnewdmname</code>	<code>vxreattach</code>	<code>vxtaginfo</code>
<code>vxchksundev</code>	<code>vxedvtoc</code>	<code>vxparms</code>	<code>vxrelocd</code>	<code>vxunroot</code>
<code>vxckdiskrm</code>	<code>vxeeprom</code>	<code>vxpartadd</code>	<code>vxresize</code>	

Exercise: Configuring the Sun StorEdge Volume Manager

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences
- Interpretations
- Conclusions
- Applications

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Describe how the SSVM utilizes disk space
- ☐ Install the Sun StorEdge Volume Manager 3.x software
- ☐ Explain the difference between the SSVM `vxinstall` Quick Installation and Custom Installation options
- ☐ Initialize the SSVM installation with `vxinstall`

Think Beyond

What if I decide to place additional disk drives under SSVM control at a later time? Why not just run `vxinstall` again?

How much preparation is necessary before configuring a large SSVM installation?

Objectives

Upon completion of this module, you should be able to:

- Describe problems associated with managing large numbers of disks
- List requirements and techniques for managing large amounts of data
- Describe commonly implemented RAID levels
- Describe a performance or reliability consideration relevant to each RAID implementation
- List guidelines for choosing an optimized stripe width for sequential and random I/O

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

- If you use RAID-5 structures to conserve disk usage, how can you improve write performance?
- What is the least expensive way to improve data reliability?
- What RAID configuration provides the highest level of protection against data loss?
- What is the relationship between data availability and data redundancy?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The RAID Advisory Board. 1996. *The RAID Book*, Lino Lakes, MN.
- “Sun Performance Tuning Overview,” December, 1993, Part Number 801-4872-07.
- Wong, Brian. *Configuration and Capacity Planning for Solaris Servers*. ISBN 0-13-349952-9.
- Chen, Lee, Gibson, Katz, and Patterson. October, 1993. *RAID: High Performance, Reliable Secondary Storage*.



Sun Educational Services

Virtual Disk Management

With the size of systems today and the demands made upon them, the Sun StorEdge Volume Manager provides valuable assistance in the following areas:

- Availability
- Performance
- Scalability
- Maintainability

Virtual Disk Management

Data Availability

Servers today are required to maintain very high levels of data availability. SSVM provides improvements in this area in the following ways:

- Prevent failed disks from making data unavailable

The probability of a single disk failure increases with the number of disks on a system. Data redundancy techniques prevent failed disks from making data unavailable.

- Allow file systems to grow while they are in use

Allowing file systems to grow while they are in use reduces the system downtime and eases the system administration burden.

Virtual Disk Management

Data Availability

- Allow multiple-host configurations

In a dual-host configuration, one host can take over disk management for another failed host. This prevents a failed host from making data unavailable.

Performance

Many applications today require very high data throughput levels. The SSVM products can assist in this area by more efficiently balancing the I/O load across disks.

Note – Several SSVM performance techniques will be discussed in detail in a later module.

Scalability

Traditionally, file system size has been limited to the size of a single disk. Using SSVM techniques, you can create file systems that consist of many disk drives. This is transparent to all applications. The size limit of file systems is increased to the UNIX limit of 1 Tbyte.

Maintainability

Administration is made easier by a GUI. Administering large numbers of disks and file systems is complex; an intuitive GUI makes administration easier.



RAID Technology Overview

RAID is an acronym for redundant array of inexpensive disks. The SSVM application supports the following RAID levels:

- RAID 0: Striping/ Concatenation
- RAID 1: Mirroring
- RAID 0+1: Striping plus Mirroring
- RAID 5: Striping with distributed parity

RAID Technology Overview

RAID is an acronym for redundant array of inexpensive disks or, more recently, redundant array of independent disks.

The RAID concept was introduced at the University of California at Berkeley in 1987 by David Patterson, Garth Gibson, and Randy Katz. Their goal was to show that a RAID could be made to achieve performance comparable to or higher than available single large expensive disks of the day.

During the development phase of the project, it was determined that it was necessary to provide redundancy to avoid data loss due to frequent disk failure. This aspect of the project became of great importance to the future of RAID.

RAID Technology Overview

RAID Standards

Many RAID levels are technologically possible but are not commonly used. The complete list of RAID levels includes:

- RAID 0: Striping or concatenation
- RAID 1: Mirroring
- RAID 0+1: Striping plus mirroring
- RAID 1+0: Mirroring plus striping
- RAID 2: Hamming code correction
- RAID 3: Striping with dedicated parity
- RAID 4: Independent reads and writes
- RAID 5: Striping with distributed parity

Note – RAID levels 2, 3, and 4 are not available with SSVM. They are not commonly implemented in commercial applications.



Sun Educational Services

Concatenation – RAID 0

Concatenation combines multiple physical disks into a single virtual disk and is organized as follows:

- Address space is contiguous.
- There is no data redundancy.
- Chunks can be thought of as physical disks or contiguous regions of disk space.

Concatenation – RAID 0

The primary reason for employing this technique is to create a virtual disk that is larger than one physical disk device. It is used to obtain more storage capacity by logically combining portions of two or more physical disks. Concatenation also enables you to grow a virtual disk by concatenating additional physical disk devices to it. This technique does not restrict the mix of different size drives – member drives can be of any size, therefore no storage space is lost.

Concatenation – RAID 0

The example in Figure 3-1 shows the concatenation of three physical disk devices.

The array management software is responsible for taking the three physical disk devices and combining them into one virtual disk that is presented to the application as a contiguous storage area.

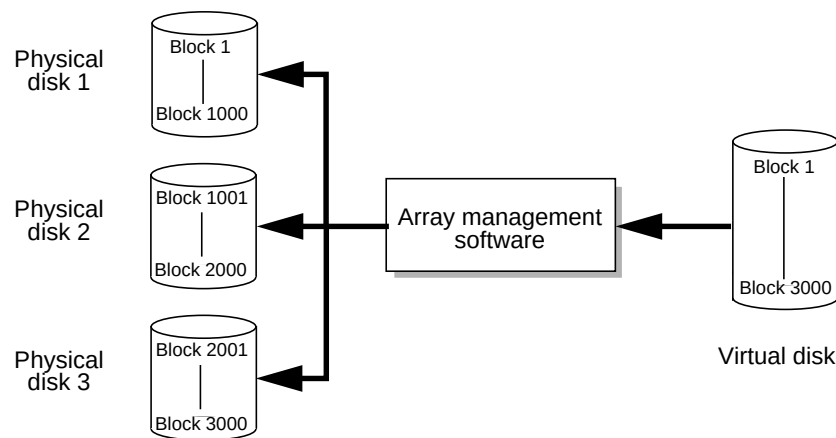


Figure 3-1 RAID-0 Concatenated Structure

The term *block* represents a disk block or sector of data

Concatenation – RAID 0

Advantages

The following advantages can be gained by using a RAID-0 concatenated structure:

- Concatenation can improve performance for random I/O as the data is spread over multiple disks.
- Write performance is the same; read performance may be improved if the reads are random.
- One hundred percent of the disk capacity is available for user data.

Limitations

These include:

- Using only concatenation, there is no redundancy. Concatenated volumes can be mirrored to achieve redundancy.
- Concatenation is less reliable, as the loss of one disk results in the loss of data on all disks.
- When the disks are full, the data is spread throughout all the members, however, when the disks are not full, the last disks are unused thereby lowering the utilization of all the drives.



Sun Educational Services

Striping – RAID 0

Striping combines multiple physical disks into a single virtual disk and is organized as follows:

- Address space is segmented.
- The I/O stream is switched from disk to disk.
- There is no data redundancy.
- There is significant performance increase

Striping – RAID 0

The primary reason for employing this technique is to improve I/O per second (IOPS) performance. The performance increase comes from accessing the data in parallel. Parallel access increases I/O throughput because all disks in the virtual device are busy most of the time servicing I/O requests.

The array management software is responsible for making the array look like a single virtual disk. It takes portions of multiple physical disks and combines them into one virtual disk that is presented to the application.

Striping – RAID 0

As shown in Figure 3-2, the I/O stream is divided into segments called stripe units which are mapped across two or more physical disks, forming one logical storage unit. The stripe units are interleaved so that the combined space is made alternately from each slice, in effect, shuffled like a deck of cards or analogous to the lanes of a freeway.

There is no data protection in this scheme and, in fact, because of the way that striping is implemented, loss of one disk results in loss of data on all striped disks. Therefore, while this implementation improves performance, it degrades reliability.

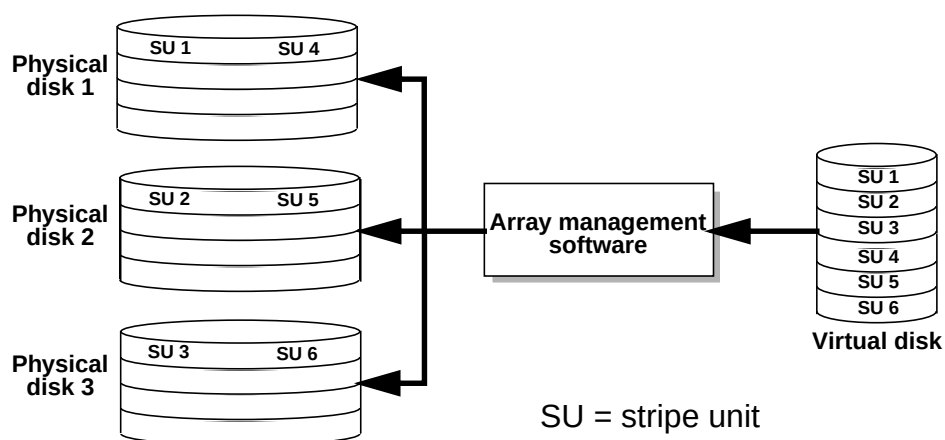


Figure 3-2 RAID 0 Striped Structure

Advantages

The following advantages can be gained by using a RAID-0 concatenated structure:

- Performance is improved for large sequential I/O requests and for random I/O. Stripe unit size can be optimized for sequential or random access.
- One hundred percent of the disk capacity is available for user data.

Striping – RAID 0

Limitations

Some of these are:

- There is no redundancy.
- Striping is less reliable, as the loss of one disk results in the loss of data on all striped disks.

Guidelines for Choosing an Optimized Stripe Unit Size

The guidelines for optimizing the stripe unit size of a striped RAID-0 structure are dependent on the type of volume access.

Sequential Access Environment

In a sequential environment, striping improves performance when the request impacts all member drives in the stripe width. For example, with an I/O request for 128 Kbytes where the stripe will include four disks, configure the stripe unit size to 32 Kbytes.

Random Access Environment

In a random access environment, striping can improve performance. Random access is dominated by seeks and rotation times of the drives. Random I/O tends to be much smaller than sequential I/O; usually ranging from 2 Kbytes to 8 Kbytes.

Performance is optimized when the stripe unit size is configured to be larger than the size of the request. For example, for an I/O request of 8 Kbytes, configure the stripe unit size to be at least 16 Kbytes.

Note – The default stripe unit size in Volume Manager is 128 sectors or 64 Kbytes.



Sun Educational Services

Mirroring – RAID 1

Mirroring provides maximum data availability and has the following features:

- Fully redundant copy of all data
- Improved read performance
- Transparent disk drive failures

Mirroring – RAID 1

The primary reason for employing this technique is to provide a high level of availability or reliability.

Mirroring (RAID 1), provides data redundancy by recording data multiple times on independent spindles. The mirrored disks appear as one virtual disk to the application. In the event of a physical disk failure, the mirror on the failed disk becomes unavailable, but the system continues to operate using the unaffected mirrors.

Mirroring – RAID 1

The array management software takes duplicate copies of the data located on multiple physical disks and presents one virtual disk to the application (Figure 3-3).

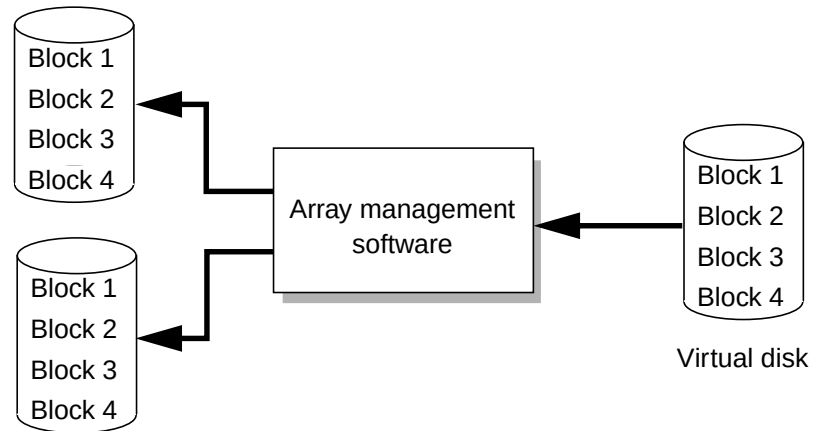


Figure 3-3 RAID 1 Mirror Structure

In Volume Manager, the mirror is seen as a single logical address, block 0 to n blocks in length. Because of this, the Volume Manager does not concern itself with the format of each individual mirror. One side of the mirror can be striped and the other side of the mirror can be concatenated. The need to do this can be due to lack of enough physical disks, or it may be implemented for performance testing. In any case, Volume Manager can guarantee consistent data across both sides of the mirror, no matter what the format, because it writes to a given logical block address.

Mirroring – RAID 1

Advantages

The following advantages can be gained by using a RAID 1 mirrored structure:

- There is a fully redundant copy of the data on one or more disks.

If the mirror resides in a storage array that is attached to a different interface board, a very high level of availability can be achieved.

- All drives can be used for reads to improve performance.

Mirroring improves read performance only in a multiuser or multitasking situation where more than one disk member can satisfy read requests. Conversely, if there is just a single thread reading from the volume, performance will not improve.

- You can set up three-way mirroring, but there is a performance penalty.

Write performance can suffer up to 44 percent with a three-way mirror.

Limitations

Mirroring uses twice as many disk drives, which essentially doubles the cost per Mbyte of storage space

Mirroring degrades write performance by about 15 percent. This is substantially less than the typical RAID 5 write penalty (which can be as much as 70 percent).

 *Sun Educational Services*

Striping and Mirroring – RAID 0+1

Combining striping with mirroring can provide multiple gains including:

- Significant performance increases
- Complete data redundancy
- Transparent disk drive failures

Striping and Mirroring – RAID 0+1

The primary reason for using striping and mirroring in combination is to gain the performance offered by RAID 0 and the availability offered by RAID 1.

This can be a relatively high-cost installation, but many customers consider it a worthwhile investment.

Striping and Mirroring – RAID 0+1

As shown in Figure 3-4, two drives are first striped and then mirrored. The reliability is as high as with mirroring. Since the technique of striping is also employed, performance here is much better than using just mirroring.

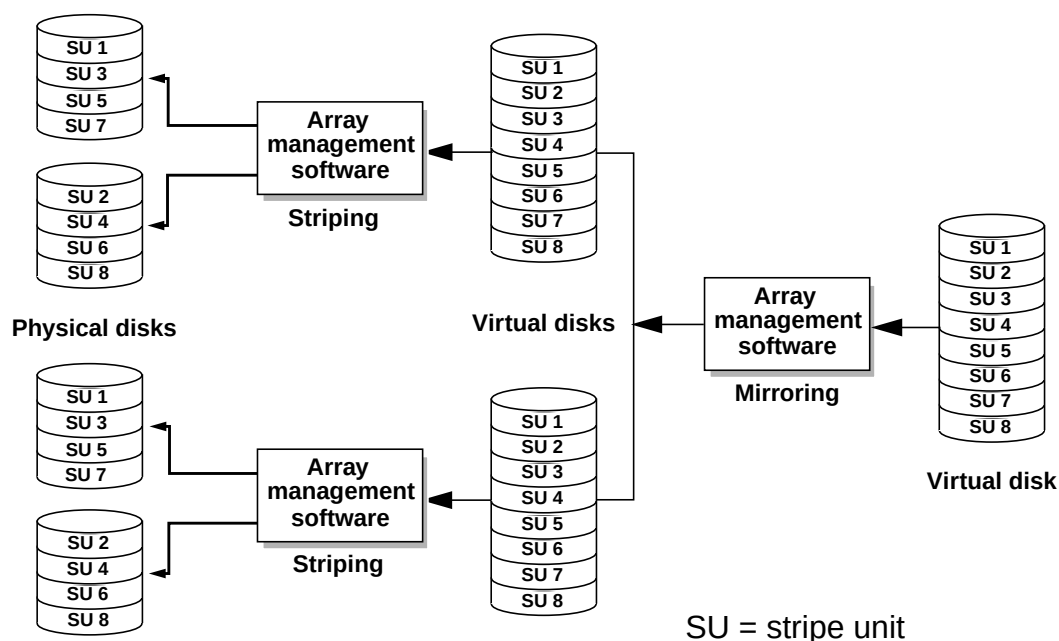


Figure 3-4 RAID 0+1 Structure

Advantages

One advantage it offers is the benefit of spreading data across a disk (improved I/O per second) while providing added redundancy of the data.

Limitations

RAID 0+1 systems suffer the high cost of mirrored systems, requiring twice the disk space of fully independent spindles.



Sun Educational Services

Mirroring and Striping – RAID 1+0

Combining mirroring with striping can provide multiple gains including:

- Significant performance increases
- Complete data redundancy
- Transparent disk drive failures
- Higher disk failure tolerance than RAID 0+1

Mirroring and Striping – RAID 1+0

RAID 1+0 has all of the performance and reliability advantages of RAID 0+1, but can tolerate a higher percentage of disk drive failures without data loss.

RAID 1+0 is sometimes referred to as mirrored stripes as opposed to RAID 0+1, which is considered to be striped mirrors.

Mirroring and Striping – RAID 1+0

As shown in Figure 3-5, the concept of RAID 1+0 is fundamentally different from RAID 0+1. In a RAID-1+0 configuration, each stripe is mirrored separately.

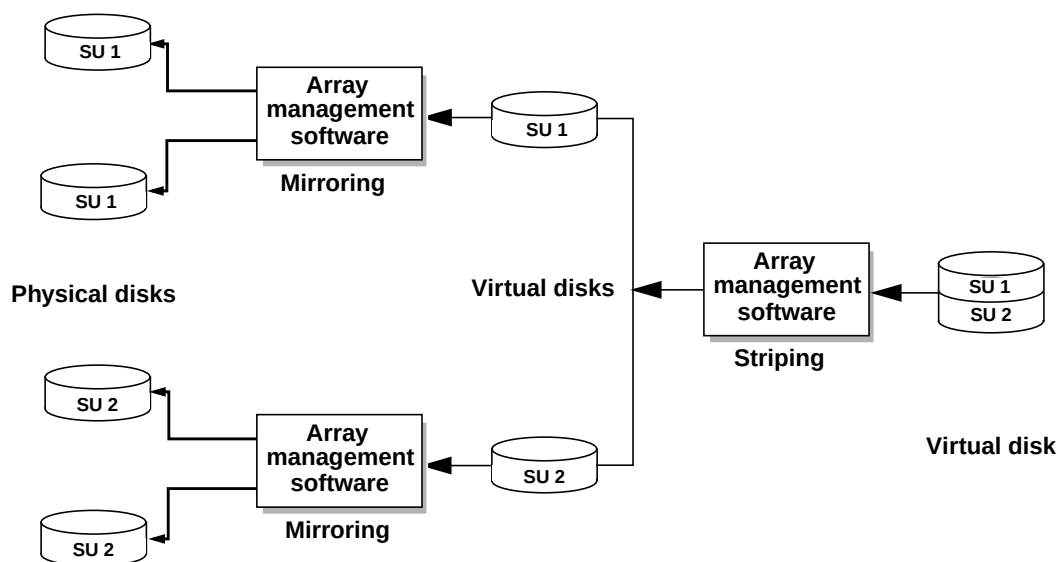


Figure 3-5 RAID-1+0 Structure

Advantages

Since each stripe is mirrored separately, a larger number of disk failures can be tolerated without disabling the volume. Availability increases exponentially with disk (stripe component) count

This configuration has the performance benefits of RAID 0+1.

Limitations

RAID 1+0 systems suffer the high cost of mirrored systems, requiring twice the disk space of fully independent spindles.

 *Sun Educational Services*

Striping With Distributed Parity – RAID 5

A RAID-5 volume structure can be an attractive alternative to mirroring because:

- Striped construction has good read performance.
- Reliability is increased but it is not as costly as mirroring.

Striping With Distributed Parity – RAID 5

RAID 5 configurations can be an attractive choice for read-intensive applications that require increased data protection.

Three of the RAID levels introduced by the Berkeley Group have been referred to as parity RAID since they employ a common data protection mechanism. RAID 3, 4, and 5 all use the concept of bit-by-bit parity to protect against data loss.

Striping With Distributed Parity – RAID 5

RAID 3, 4 and 5 all implement the *Boolean Exclusive OR* (XOR) function to compute parity. It is implemented bit-by-bit to corresponding stripe units of member drives and the result is written to a corresponding parity disk found in RAID 3 and 4. This parity is distributed throughout all the member drives in RAID 5.

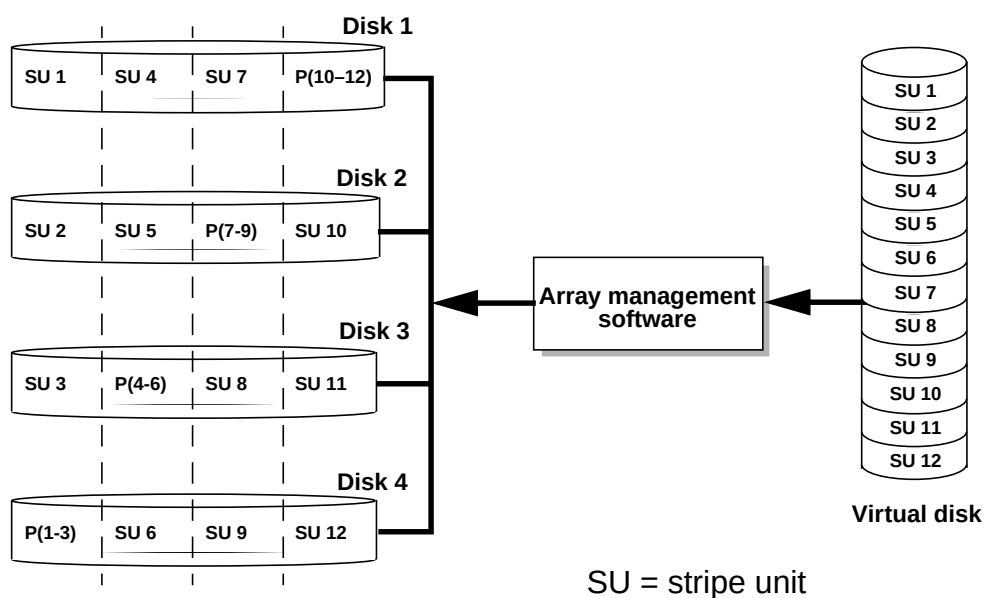


Figure 3-6 RAID-5 Structure

Additional features include:

- Independent access is available to individual drives.
- Data and parity are both striped across spindles.
- Reads per second can reach disk rate times number of disks.
- Overall random I/O performance is dependent on percentage of writes. If there are more than 20 percent writes, an alternative RAID option, such as RAID 0+1, should be considered.

Striping With Distributed Parity – RAID 5

Advantages

Some advantages are:

- Parity protects against single disk failure.
- RAID 5 requires only one additional drive beyond those used for data.

Limitations

Some limitations are:

- A minimum of three disks are required to implement RAID 5 in Volume Manager.
- RAID 5 cannot be mirrored, redundancy is provided through the parity information.
- Write-intensive performance can be poor.

If there are more than 20 percent writes, an alternative RAID option, such as RAID 0+1, should be considered.

- There can be severe performance degradation with a failed disk in a write-intensive environment.

Striping With Distributed Parity – RAID 5

Performance Factors

Data can be accessed with a failed drive, with some performance penalties:

- To read data from a surviving drive – No change.
- To read data from a failed drive – Corresponding stripe units from surviving drives in the stripe are read and linked together with XOR to derive the data.
- To write to a surviving drive – If the failed drive holds the parity data, the write proceeds normally without calculating parity. If the failed drive holds data, then a read-modify-write sequence is required.
- To write to a failed drive – All the data from the surviving data drives are linked with the new data using XOR, and the result is written to the parity drive.
- To recover from a single disk failure – Data from the remaining stripe units in the stripe must be read, linked together with XOR, and the end result written to the replacement drive, given there is an available spare drive in the configuration.

The write modifies the entire stripe width. (Preservation of any existing data is not necessary.)

- All the new data stripe units are linked together using XOR, generating a new parity value.
- Data and parity are written to the log.
- The new parity is written to the parity stripe unit. The data is written to the data stripe units. All stripe units are written in a single write.

Striping With Distributed Parity – RAID 5

Performance Factors

If a write modifies more than one disk but less than an entire stripe width (the least desirable scenario), it is referred to as a *read-modify-write* sequence. Performance can be negatively impacted by up to 80 percent. The sequence of steps are as follows:

1. The data stripe units being updated with the new write data are accessed and read into internal buffers.
2. An XOR is performed on the new data to generate the new parity stripe unit..
3. The new data and the parity are written to a log.
4. The new parity is written to the parity stripe unit, and the data is written to the data stripe units. All writes are done in parallel.

Striping With Distributed Parity – RAID 5

Performance Factors

If the write modifies more than half of the data disks, but less than a full stripe, it saves more I/O time than the read-modify-write because it does not require a read of the parity region, and only requires a read of the unaffected data (which amounts to less than half of the stripe units in the stripe).

- Unaffected data is read from the unchanged data stripe unit(s) into internal buffers.
- An XOR is performed on the new data with the old, unaffected data to generate the new parity stripe unit.
- The new data and resulting parity are logged (if logging is enabled).
- The new parity is written to the parity stripe unit. The new data is written to the data stripe units. All stripe units are written to a single write.

Note – Full stripe writes that are the exact width of the stripe can be performed without the read-modify-write sequence. In this manner, some RAID-5 implementations can deliver high performance for large sequential transfers.

Striping With Distributed Parity – RAID 5

Guidelines for Optimizing Stripe Width

These guidelines provide a good rule of thumb for the stripe unit size to avoid serious performance penalties. I/O request sizes can vary significantly in size. The application vendor or the software developer can provide this information. If absolutely unsure of the projected I/O size, it is safe to accept the defaults provided by Volume Manager.

Sequential Access Environment

In a sequential environment, striping improves performance since the

- Performance is best addressed if the request impacts all member drives in the stripe width.
- I/O request divided by the number of columns equals the stripe unit size

For example, if the I/O request is 128 Kbytes, and the stripe includes four disks, configure the stripe unit size to 32 Kbytes.

Random Access Environment

In a random access environment, striping can also improve performance. Random access is dominated by seeks and the rotation time of the drives. Random I/O also tends to be much smaller than sequential, usually ranging from 2 Kbytes to 8 Kbytes. Performance is best if the stripe unit size is configured large, relative to the size of the request.

For example, if the I/O request is 8 Kbytes, configure the stripe unit size to be at least 16 Kbytes.

Note – The default stripe unit size in Volume Manager for RAID-5 volumes is 32 sectors, or 16 Kbytes.

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Describe problems associated with managing large numbers of disks
- ☐ List requirements and techniques for managing large amounts of data
- ☐ Describe commonly implemented RAID levels
- ☐ Describe a performance or reliability consideration relevant to each RAID implementation
- ☐ List guidelines for choosing an optimized stripe width for sequential and random I/O

Think Beyond

How large a configuration can be easily managed using the Sun StorEdge Volume Manager software?

What other types of training might be necessary in order to maintain a typical installation?

Volume Manager Storage Administrator (VMSA) Software

4 

Objectives

Upon completion of this module, you should be able to:

- Describe the VMSA server/client relationship
- Verify the VMSA server software is running
- Start the VMSA client software
- Use the main VMSA features
- Use the Options menu to customize the behavior of VMSA
- Describe two important uses for the Task Request Monitor

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

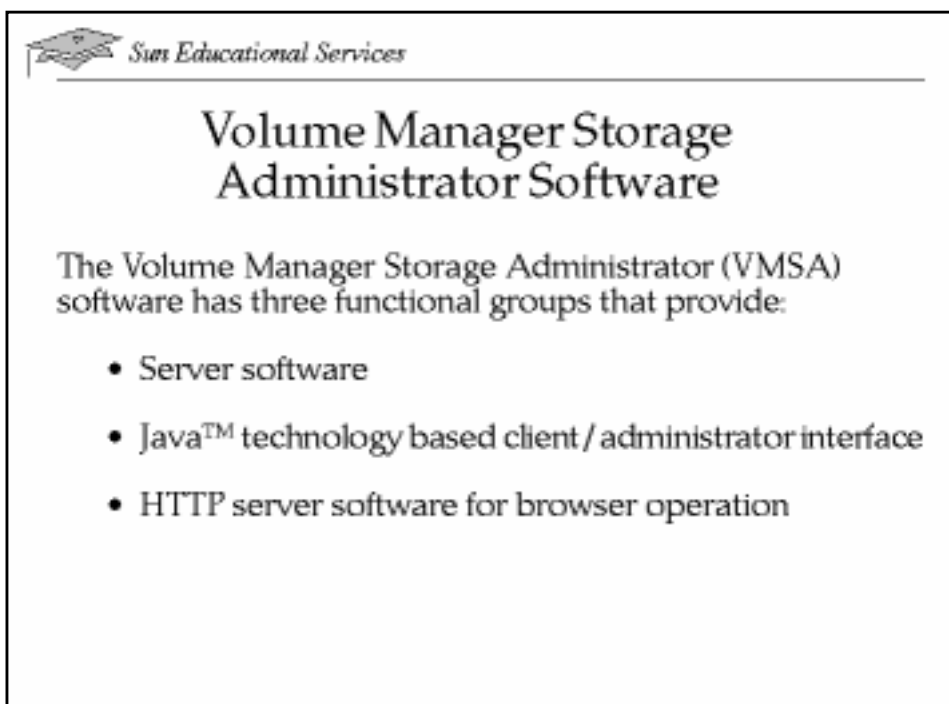
- A graphical interface is very useful but how can I learn to perform equivalent operations from the command line?
- I use Windows NT systems at my company, how can I use them to administer the Sun storage servers?
- Can the VMSA GUI effectively manage very large SSVM installations with thousands of disk drives?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The RAID Advisory Board. 1996. *The RAID Book*, Lino Lakes, MN.
- “Sun Performance Tuning Overview,” December, 1993, Part Number 801-4872-07.
- Wong, Brian. *Configuration and Capacity Planning for Solaris Servers*. ISBN 0-13-349952-9.
- Chen, Lee, Gibson, Katz, and Patterson. October, 1993. *RAID: High Performance, Reliable Secondary Storage*.



Volume Manager Storage Administrator Software

The VMSA software is a new generation of disk storage administration software that allows greater flexibility for SSVM administrators.

Core administrative software that runs on the SSVM system is designed to interact directly with either a Web browser or a graphical administration application, VMSA.

The VMSA tool is a Java application that can be run locally on the SSVM server or remotely on any networked system. It will run on any Java 1.1 Runtime Environment (including Solaris, HP-UX, Windows NT, or Windows 95).

Hypertext Transfer Protocol (HTTP) server software can be activated on the SSVM server which enables administrative access through several different Web browsers.

Volume Manager Storage Administrator Software

Server/Client Software Installation

During the installation of the VRTSvmsa software, you are asked if you want Web server software installed and then if you want the SSVM Server software installed.

Processing package instance <VRTSvmsa> from </SSVM>

Sun StorEdge Volume Manager

Where should this package be installed? (default: /opt]) [?,q]

Should the Apache HTTPD (Web Server) included in this package be installed? (default: n) [y,n,?,q]

Should the StorEdge Volume Manager Server be installed on this system?

(The StorEdge Volume Manager Client will be installed regardless)

(default y) [y,n,?,q]

The client portion of the VRTSvmsa package is always installed.

Volume Manager Storage Administrator Software

VMSA Server Software Startup

If installed, the server portion of the VMSA software is automatically started at boot time by the `/etc/rc2.d/S96vmsa-server` script. The `/opt/VRTSvmsa/bin/vmsa_server` script file is started in the background.

You can manually stop and start the server portion of the VMSA software on the SSVM server using the following options:

- `vmsa_server -V` prints the version
- `vmsa_server -q` verifies the server software is running
- `vmsa_server &` uses the normal startup
- `vmmsa_server -k` kills the VMSA server software

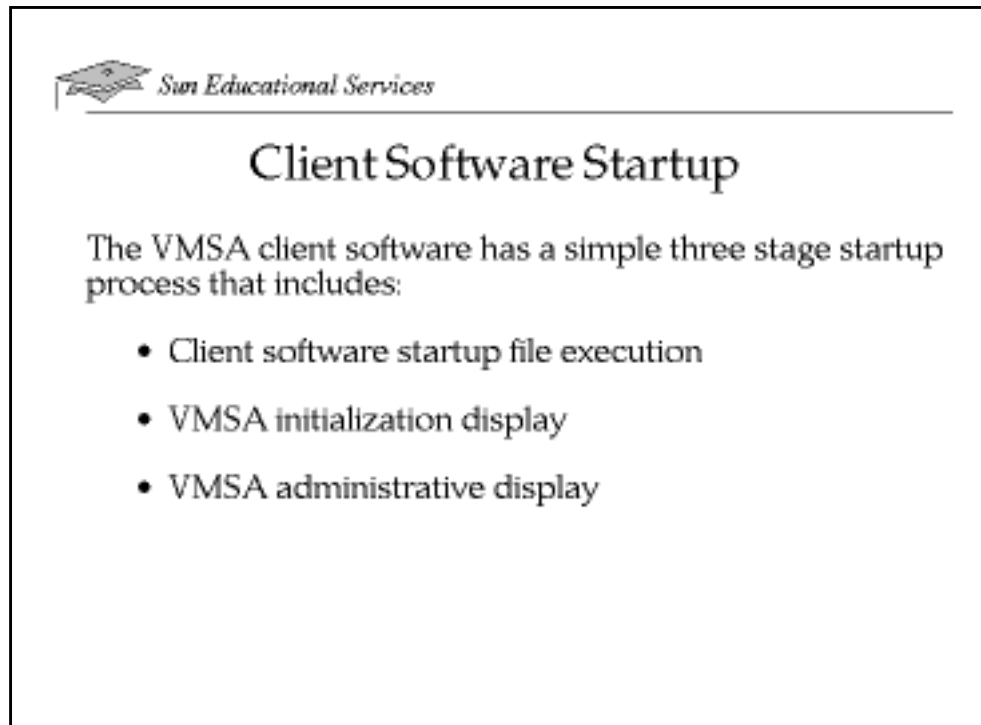
You can also run the server software in read-only mode by editing the `/opt/VRTSvmsa/vmsa/properties` file and changing the value of the variable `vrts.server.readonly=false` to `true`. You must restart the server software for this change to take effect.

VMSA Client Software Startup

The client graphical interface is started by manually running the `/opt/VRTSvmsa/bin/vmsa` script file. The VMSA client software can be started and displayed on the server, started on the server and remotely displayed on another system, or loaded and started on a remote system.

The name of the server to be monitored can be included in the startup command as follows:

```
# /opt/VRTSvmsa/bin/vmsa server_name &
```



Volume Manager Storage Administrator Client

Client Software Startup

The VMSA client application is a pure Java technology based application and is started from the `/opt/VRTSvmsa/bin/vmsa` script file by executing the Java interpreter and assigning the following values to properties:

- `vrts.packageName=VRTS`
- `vrts.codebase=$VMSA_HOME/vmsa/java`
- `vrts.iconbase=$VMSA_HOME/vmsa/java`
- `vrts.server.host=$HOST`
- `vrts.localHost='hostname' Vmsa`

The classname argument used is `Vmsa`.

Volume Manager Storage Administrator Client

VMSA Initialization Display

Before the VMSA Client application is started, a session initialization window is displayed. (See Figure 4-1.)

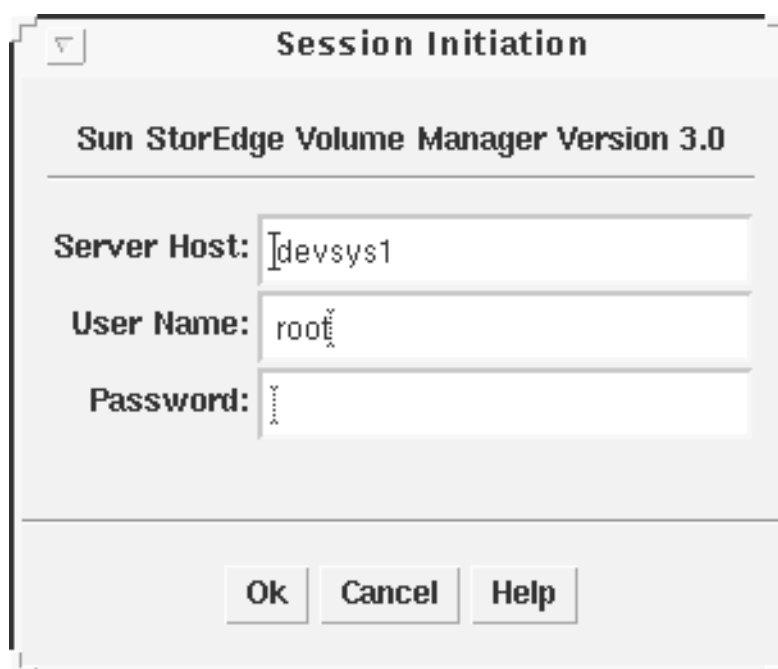


Figure 4-1 VMSA Initialization Display

You can start the VMSA Client application and connect to any system that is running the VMSA Server software.

Note – Even if you furnish the server name when you start the `vmsa` script, the session initiation window is still displayed.

Volume Manager Storage Administrator Client

VMSA Client Display

The VMSA initial Client Display has distinct functional areas as shown in Figure 4-2.

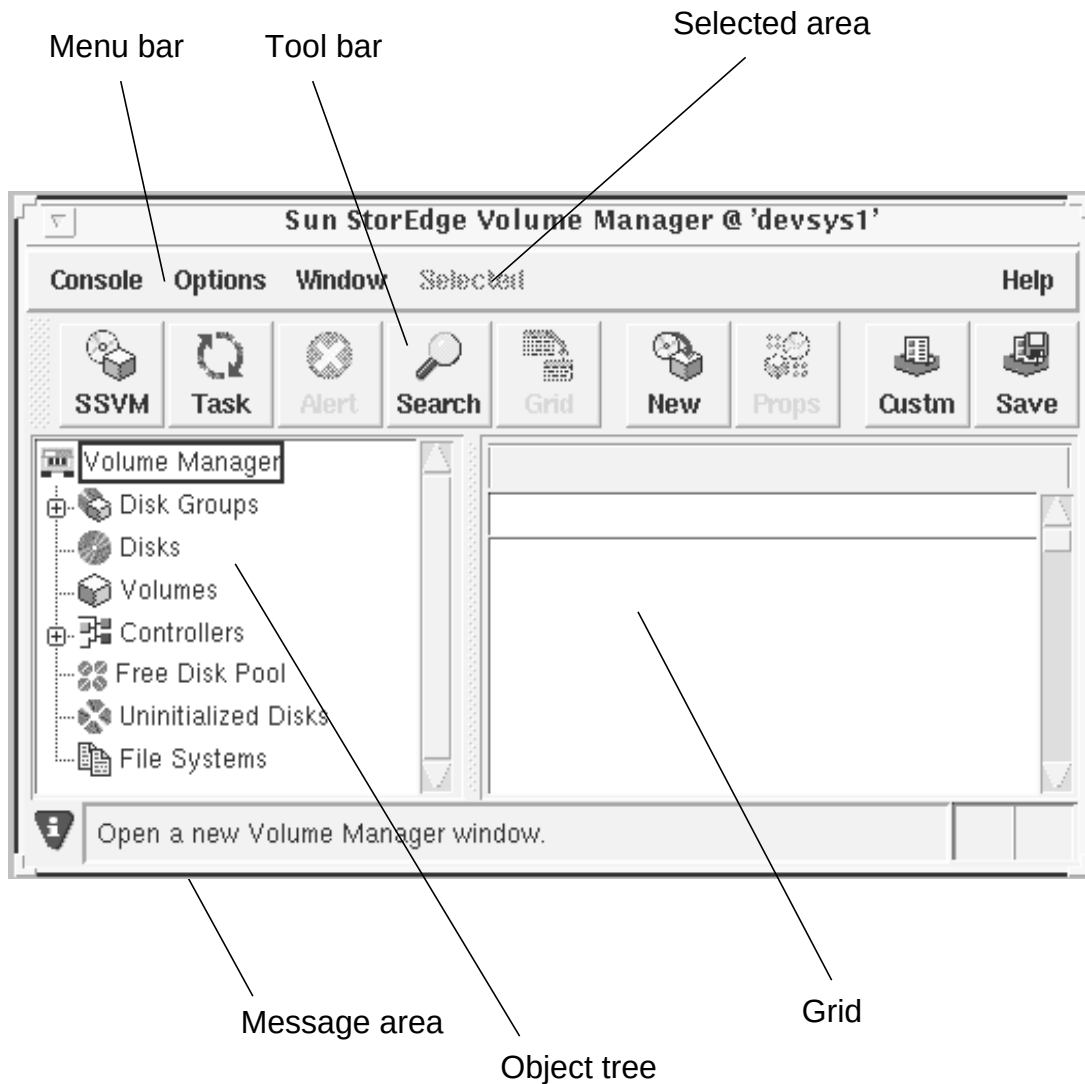


Figure 4-2 VMSA Client Administrative Display

Note – The Selected menu changes name according to the type of objects being selected in the grid window.



Sun Educational Services

VMSA Client Software Features

The VMSA client software has the following major functional elements:

- Tool bar
- Menu bar functions
- Object Tree display
- Grid display
- Command Launcher

VMSA Client Software Features

The VMSA client software has the following major functional elements:

- Tool bar
- Menu bar functions
- Object Tree display
- Grid display
- Command Launcher

VMSA Client Software Features

Tool Bar

The tool bar shown in Figure 4-3 provides quick access to general VMSA functions.

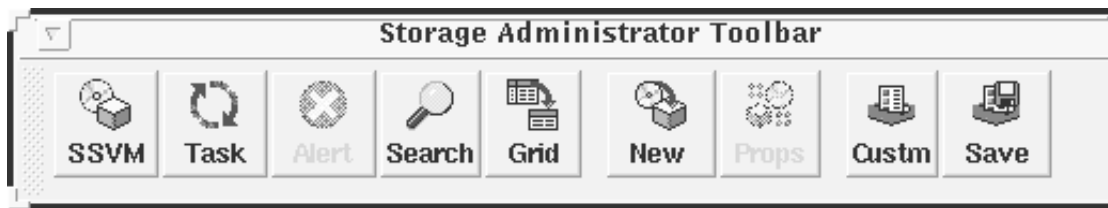


Figure 4-3 VMSA Tool Bar

The tool bar provides direct access to a number of complex functions. All of the functions are available elsewhere in menus but the tool bar offers a convenient way to access them.

VMSA Client Software Features

Tool Bar

The functions available from the Tool Bar icons are listed in Table 4-1.

Table 4-1 Tool Bar Icon Functions

Tool Bar Icon	Function
SSVM	Open a New Volume Manager Window
TASK	Open the task monitor
ALERT	Open the alert monitor
SEARCH	Open a new search window
GRID	Open a new window that contains a copy of the main grid
NEW	Create a new volume
PROPS	Open a selected object properties window
CUSTM	Customize Volume Manager
SAVE	Save customization settings

Note – Some of these features will not be covered until you perform the practical exercises at the end of this module.

VMSA Client Software Features

VMSA Menu Bar

The VMSA Client menu bar has the functions shown in Figure 4-4.

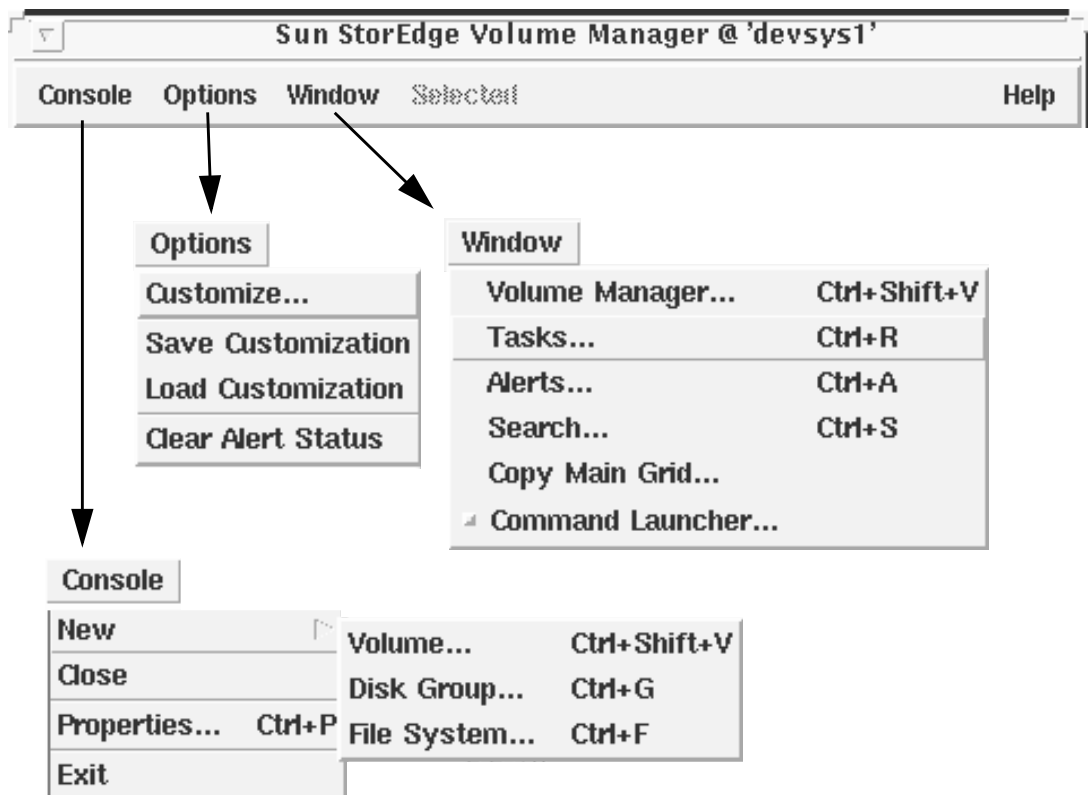


Figure 4-4 VMSA Menu Bar Functions

The Selected area of the menu bar changes depending on what type of object is selected in the grid area. For instance, if you select a disk in the grid window, a Disks menu will be displayed in the Selected area. This menu usually has more advanced commands, many of which are not available in other similar menus.

Note – Most of the menu functions will not be discussed until later in the course. Other more basic aspects of VMSA be understood first.

Volume Manager Storage Administrator Client

VMSA Object Tree

The VMSA Object Tree window, shown in Figure 4-5, has an icon for every type of VMSA object that exists or can be created.

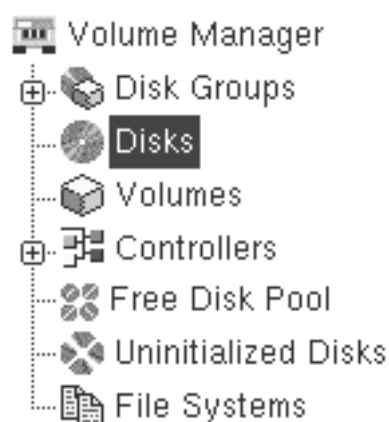


Figure 4-5 VMSA Object Tree

When an object is selected with the left mouse button, expanded configuration information about that object is displayed in the grid area as shown in Figure 4-6.

Volume Manager/Disks/						
Device	Size	Unused	State	Spare	Reserve	
c0t0d0s2	0	-	Not Setup	-	-	
c2t33d0s2	8.4 G	8.4 G	Imported	No	No	
c2t35d0s2	8.4 G	8.4 G	Imported	No	No	
c2t37d0s2	8.4 G	-	Free	-	-	
c2t50d0s2	8.4 G	-	Free	-	-	
c2t52d0s2	0	-	Not Setup	-	-	
c3t18d0s2	8.4 G	8.4 G	Imported	No	No	
c3t1d0s2	8.4 G	-	Free	-	-	
c3t20d0s2	0	-	Not Setup	-	-	
c3t3d0s2	0	-	Not Setup	-	-	
c3t5d0s2	0	-	Not Setup	-	-	

Figure 4-6 VMSA Grid Display

Volume Manager Storage Administrator Client

VMSA Object Tree

Some branches on the object tree have a small node that contains a plus (+) sign. If you select these nodes, they will expand the display to deeper levels as shown in Figure 4-7.

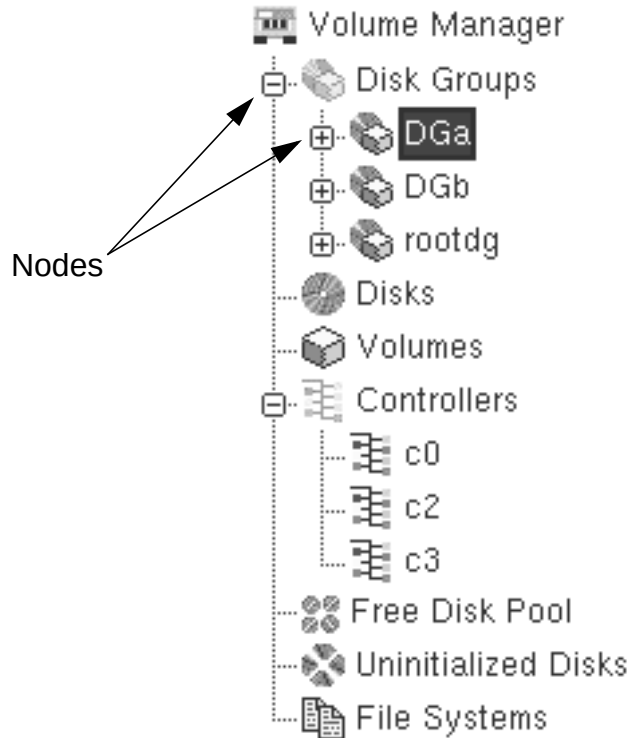


Figure 4-7 VMSA Object Tree Expansion

In Figure 4-7, the Disk Groups and Controllers branches have been expanded. These branches can be expanded even further. When a node cannot be expanded further, it contains a minus (-) sign. Selecting a node again will reverse the expansion.

Volume Manager Storage Administrator Client

VMSA Command Launcher

The Command Launcher window, shown in Figure 4-8, provides a scrollbar menu of all object-related commands that can be selected. Many of the command selections require additional information. A separate form will be displayed.

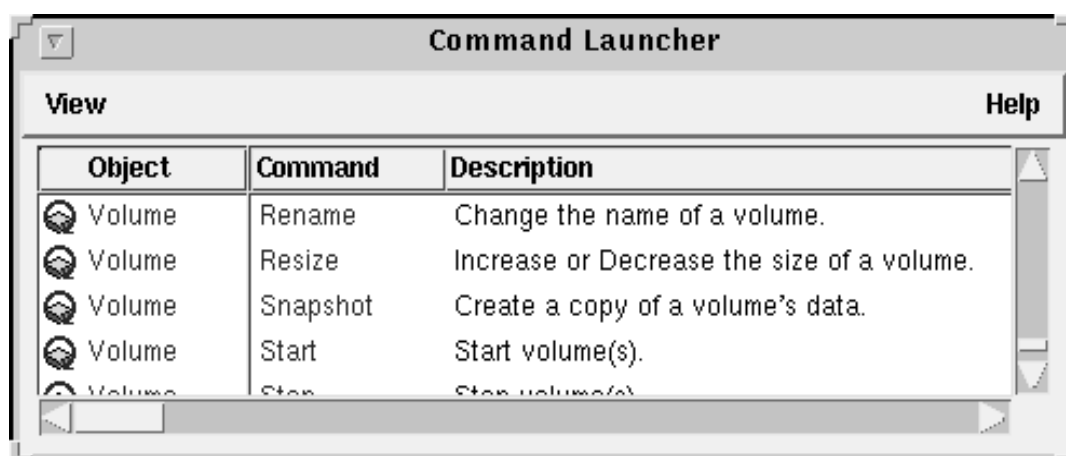


Figure 4-8 VMSA Command Launcher

Command Summary

The Command Launch window has many subcommands under each of the following object categories:

- Disk (10 subcommands)
- Disk Group (6 subcommands)
- File System (6 subcommands)
- Log (2 subcommands)
- Mirror (4 subcommands)
- Subdisk (3 subcommands)
- Volume (17 subcommands)

Volume Manager Storage Administrator Client

Docking Windows

The VMSA docking feature enables you to split the tool into separate windows for ease of use as shown in Figure 4-9

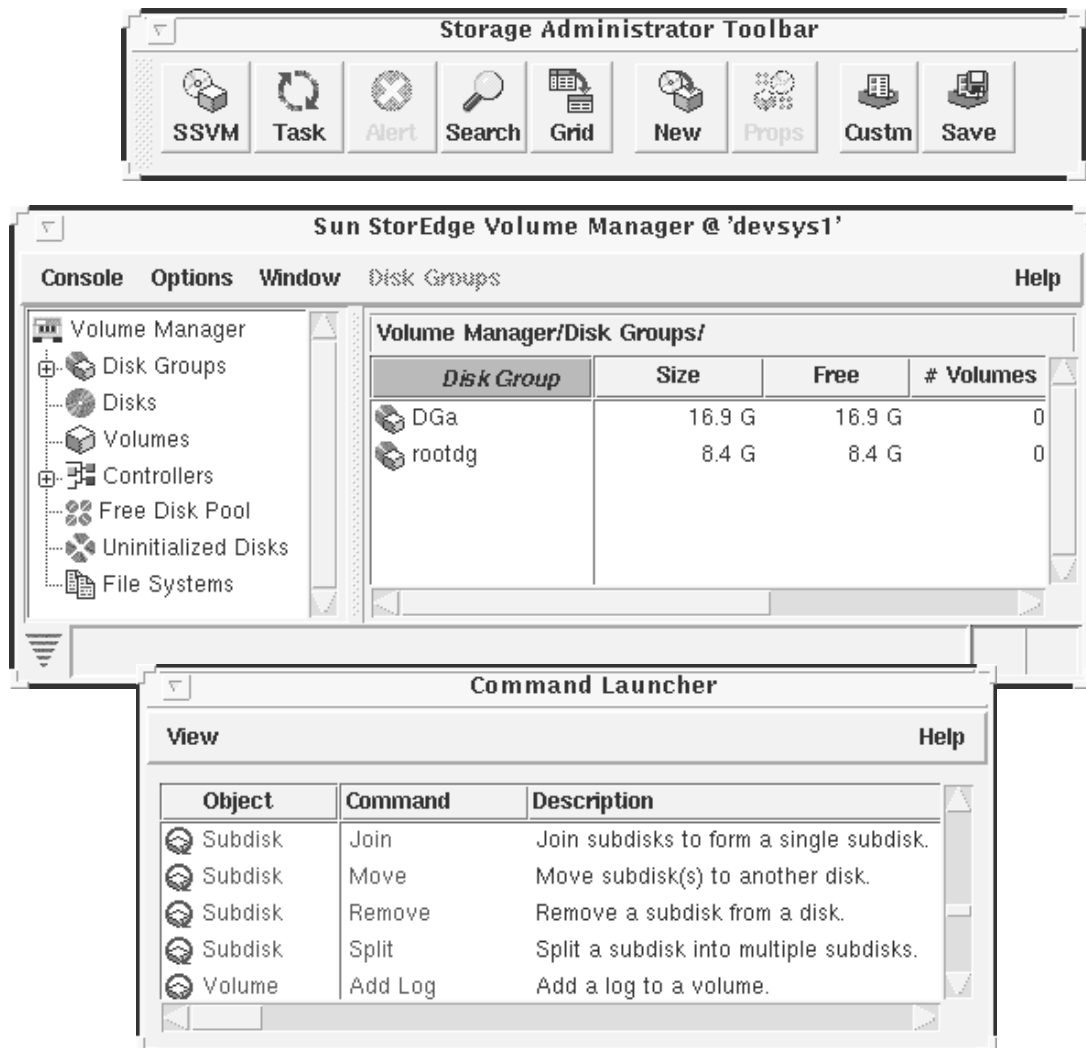


Figure 4-9 VMSA Docking Feature

Note – The Custom button on the tool bar displays forms that enable you to customize many VMSA features.



Sun Educational Services

VMSA Tasks

A typical VMSA administrative task, such as creating a new disk group, involves the following steps:

1. Select the New Disk Group menu function.
2. Complete the Disk Group form.
3. Select Encapsulate or Initialize.
4. Review the Task Request Monitor window.
5. Review the command-line operations in the Task Properties window.

VMSA Tasks

It can be very informative to see a step-by-step example of a typical VMSA task. The task shown in the following examples is that of creating a new disk group containing two disk drives.

This process is similar to many VMSA tasks you might perform.

VMSA Tasks

Using the Create Menu

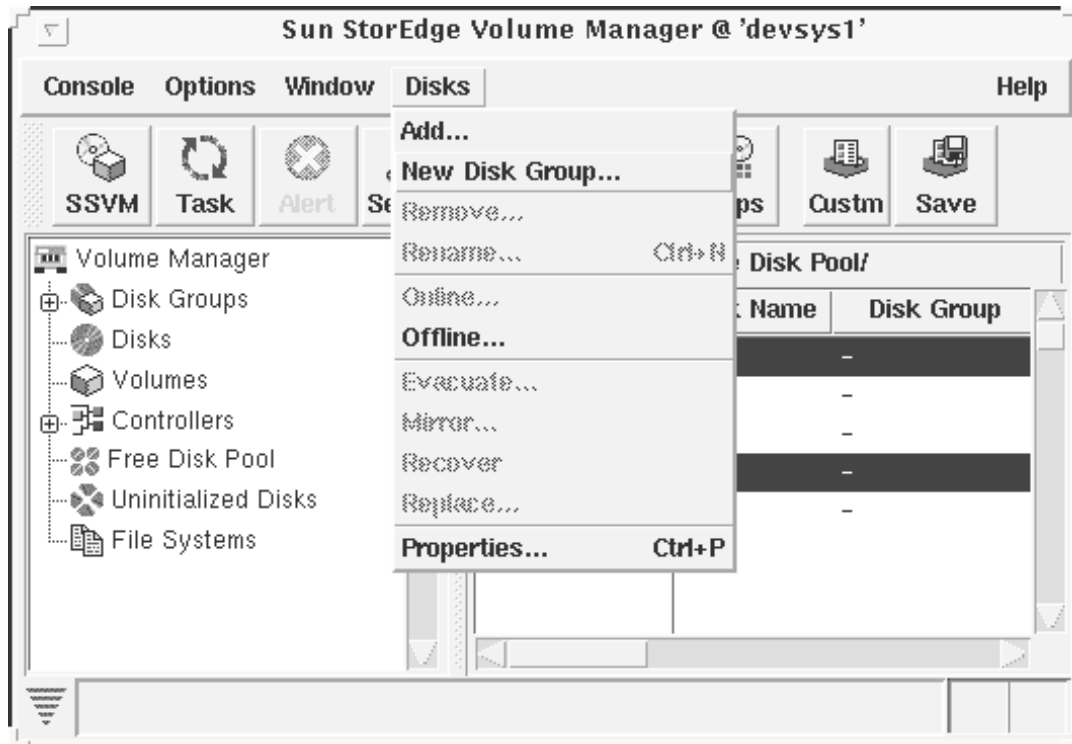


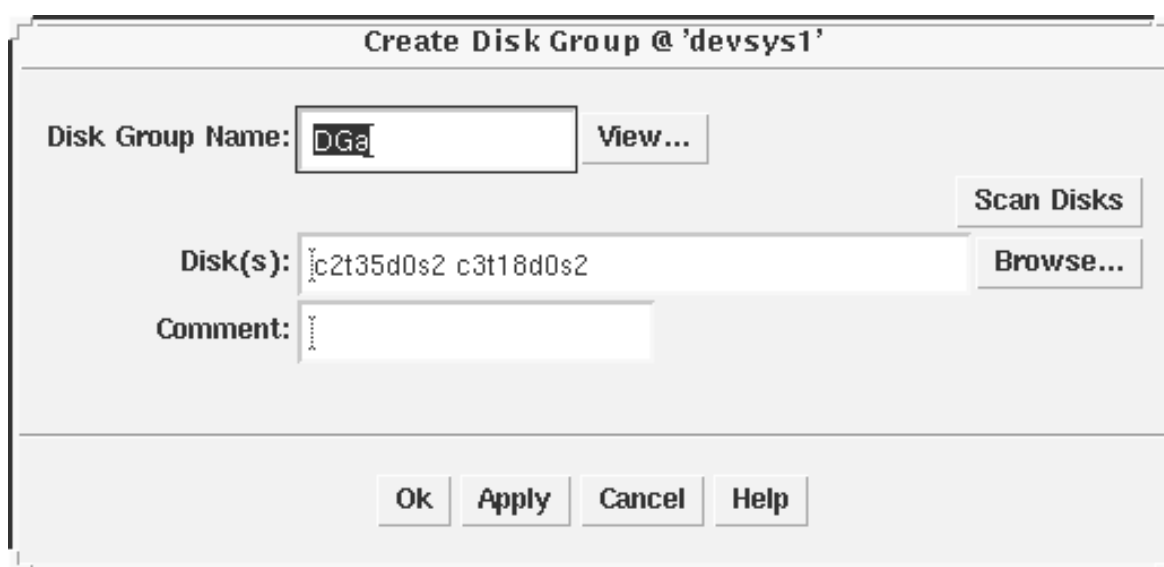
Figure 4-10 Volume Manager Disks Menu

Note – Selecting the disk drives in advance as shown can save time in later steps.

VMSA Tasks

Using the Disk Group Form

As shown in Figure 4-11, the Disk(s) section is already filled and you only need to furnish the name of the disk group. This happens if you have highlighted the desired disk drives in advance in the Grid window.



The screenshot shows a dialog box titled "Create Disk Group @ 'devsys1'". It contains the following fields and buttons:

- Disk Group Name:** A text field containing "DGa". To its right is a "View..." button.
- Disk(s):** A text field containing "c2t35d0s2 c3t18d0s2". To its right is a "Browse..." button.
- Comment:** An empty text field.
- Buttons:** "Scan Disks" (top right), "Ok", "Apply", "Cancel", and "Help" (bottom).

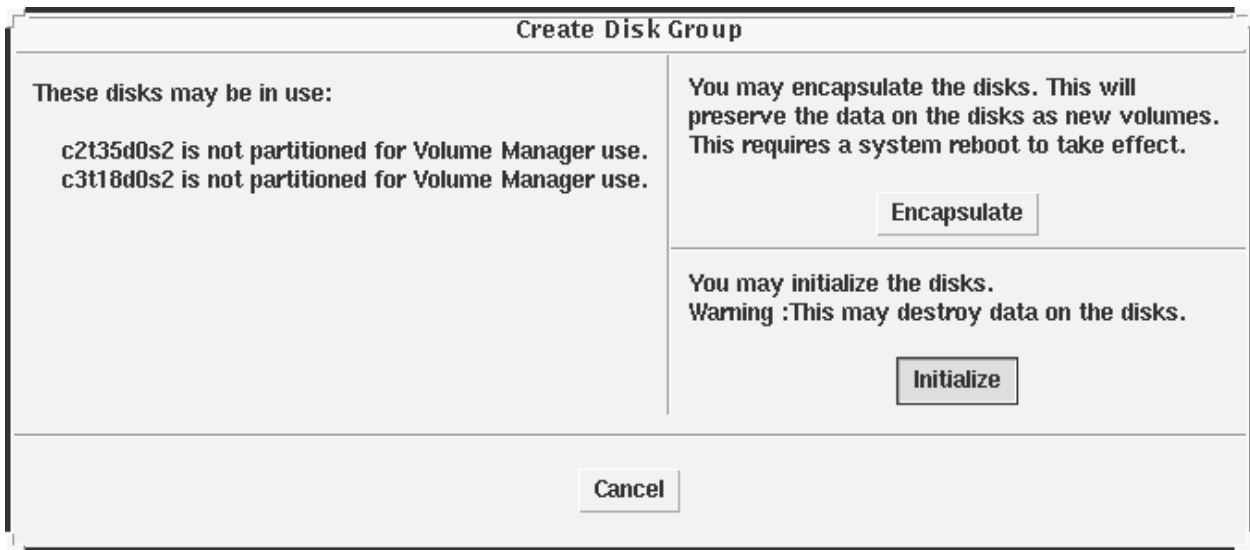
Figure 4-11 VMSA Disk Group Form

Note – If you do not select the target disk drives in advance, you can use the Browse button shown in Figure 4-11. This will display a small version of the Object Tree so you can select appropriate disk drives.

VMSA Tasks

Selecting Encapsulation or Initialization

As mentioned in a previous module, SSVM initializes disk drives by repartitioning them into slices 3 and 4. When you are performing disk operations with VMSA, it “assumes” there might be data on a disk that is not partitioned this way.



You must be very careful when initializing disk drives. VMSA determines as best it can whether or not there is a risk of accidentally losing data on a disk that is about to be initialized.

VMSA Tasks

Verifying Task Completion

After a task has apparently completed, verify the success by checking the Task Request Monitor as shown in Figure 4-12.

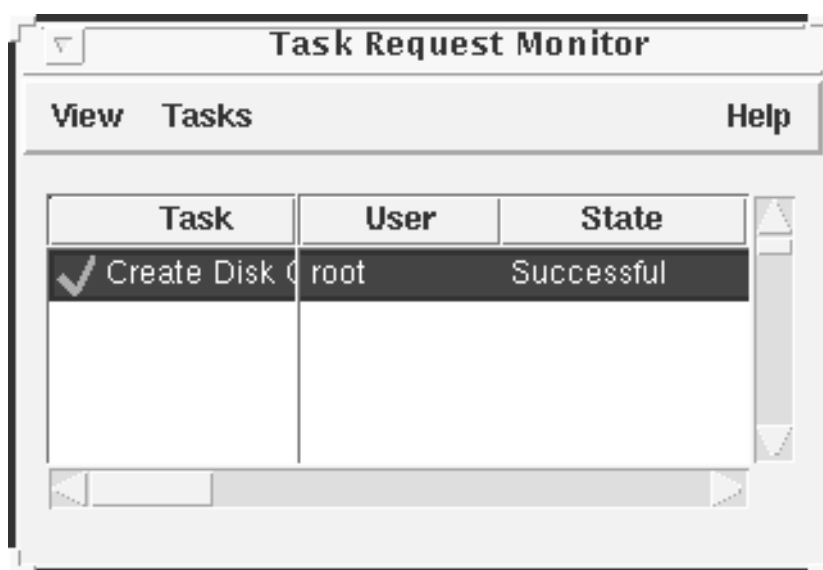


Figure 4-12 Task Request Monitor

Tasks that complete successfully are preceded by a check mark. Tasks that fail are preceded by the international symbol for No (⊘).

If you double-click on a task entry you can view detailed information about the commands used to perform that task.

Note – The Task Request Monitor window shown in Figure 4-12 has additional start and stop time fields that can be viewed either by using the scroll bar at the bottom of the window or by expanding the width of the window.

VMSA Tasks

Viewing Task Properties

If you double-click on a task entry in the Task Monitor window, detailed task information is displayed in the Task Properties window (Figure 4-13).



Figure 4-13 Task Properties

Note – You can cut and paste the command lines from the Task Properties display. This is a valuable tool for using SSVM command-line operations.

Exercise: Using the VMSA Client Software



Exercise objective – In this exercise you will:

- Install the VMSA Client software on a remote system
- Start the VMSA Client software
- Connect to the VMSA server
- Familiarize yourself with the basic VMSA Client software features and functionality
- Perform at least one simple task and record the resulting command-line operation

Preparation

Ask your instructor to furnish the following information:

- The location of the VRTSvmsa software package. It might be on a CD-ROM or it can be NFS mounted.

Note – If for some reason you cannot install the VMSA Client software on a remote workstation, it can be run on the server and remotely displayed. You will have to set the `DISPLAY` variable on the server and enable `xhost` access on the remote workstation. It is also possible to work directly on the server if it has a frame buffer.

Exercise: Using the VMSA Client Software

Task – Setting up the Environment

Complete the following steps:

1. On the SSVM server as user `root`, enter the `env` shell command.
2. Verify that the following environment exists on the server:

```
TERM = dtterm
PATH=/bin:/usr/bin:/usr/sbin:/opt/VRTSvmsa/bin:/etc/vx/bin:.
MANPATH=/usr/man:/opt/VRTSvmsa/man:/opt/VRTSvxvm/man
DISPLAY = remote workstation:0.0
```

3. Log out of the SSVM server if you are working from a remote workstation.
4. On the remote workstation, type the `env` shell command and verify that the following environment exists:

```
TERM = dtterm
PATH=/bin:/usr/bin:/usr/sbin:/opt/VRTSvmsa/bin:.
MANPATH=/usr/man:/opt/VRTSvmsa/man
```

5. Make sure the system you are going to work on has access control disabled. This can be done manually in the Console window with the `xhost +` command.

Note – The `xhost +` command can be placed in a `.xinitrc` file in the `login` directory.

Exercise: Using the VMSA Client Software

Task – Installing the VMSA Client Software

Use the following steps:

1. Log in as user `root` on the remote classroom system you have been assigned.
2. Obtain access to the `VRTSvmsa` software package. Your instructor should have given you instructions about this.
3. Install the VMSA Client software as follows:

```
# pkgadd ./VRTSvmsa
```

```
Processing package instance <VRTSvmsa> from </tmp>
```

```
Sun StorEdge Volume Manager
```

```
Where should this package be installed? (default: /opt]) [?,q] /opt
```

```
Should the Apache HTTPD (Web Server) included in this package be  
installed? (default: n) [y,n,?,q] n
```

```
Should the StorEdge Volume Manager Server be installed on this system?
```

```
(The StorEdge Volume Manager Client will be installed regardless)
```

```
(default y) [y,n,?,q] n
```

4. After the package installation is complete on the workstation, the VMSA client software requires that the remote system it is running on allows remote displays. Enable remote display with a `xhost +` command.

Note – You might want to put the `xhost +` in a `.xinitrc` file in the `root` login directory on the remote workstation.

Exercise: Using the VMSA Client Software

Task – Starting VMSA Client Software

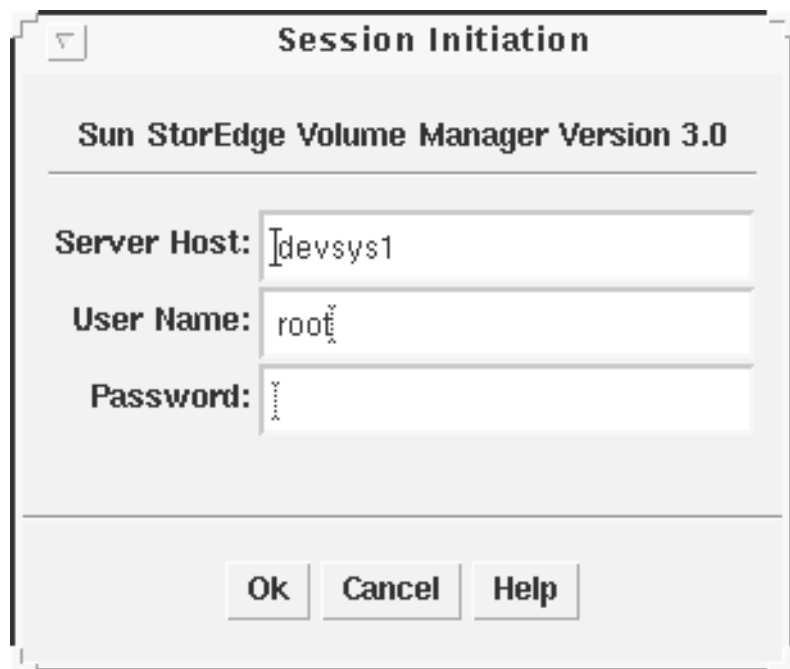
Complete these steps:

1. Start up the VMSA Client software on the remote workstation as follows:

```
# /opt/VRTSvmsa/bin/vmsa &
```

Note – You can also furnish the SSVM server name as an option; for example, `/opt/VRTSvmsa/bin/vmsa devsys1 &`.

2. Furnish the SSVM server system name, user name, and password in the Session Initiation window as shown. Your server host name will be different.



Note – Remember, this display originates from the SSVM server and requires that the remote workstation will allow remote displays.

3. The initial Client display window (titled Volume Manager Storage Administrator) will be displayed after a short time.

Exercise: Using the VMSA Client Software

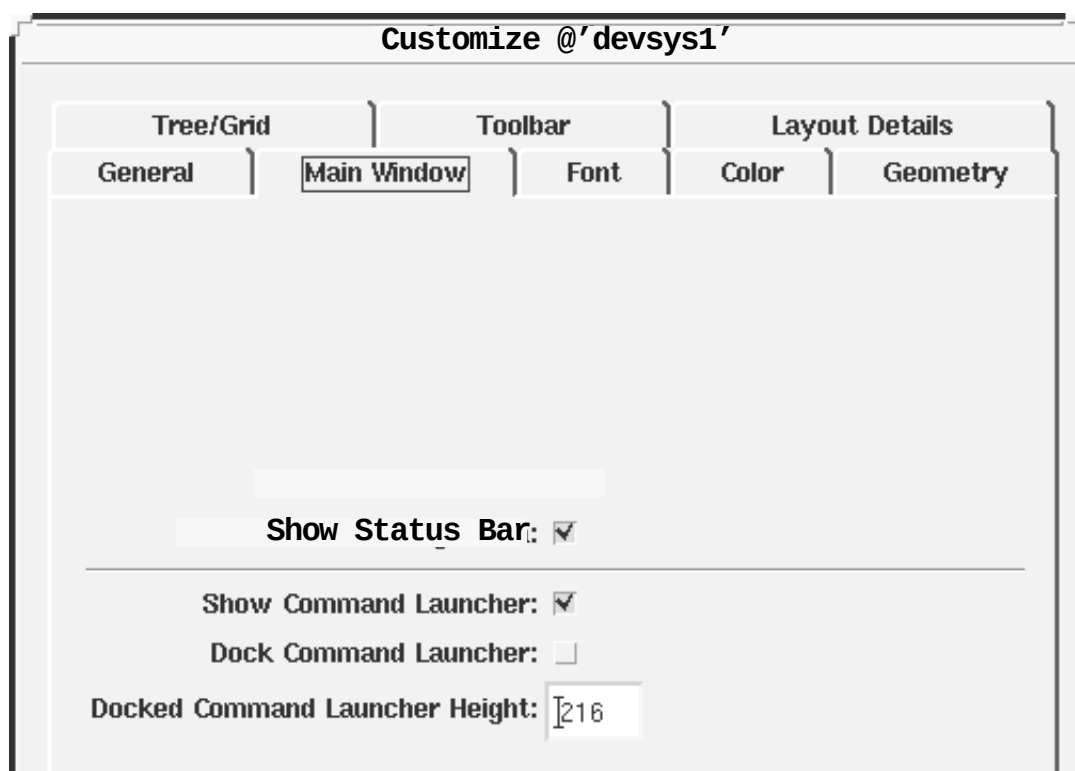
Task – Setting up the VMSA Client Display

Complete the following steps:

1. Select the Options/Customize menu entry in the menu bar.



2. Select the Main Window display in the Customize window and configure the Command Launcher as shown. Select Apply.



3. Display the Toolbar display in the Preferences window. Select the Show Toolbar box and deselect the Dock Toolbar box.
4. Select Apply and click on OK when done.

Exercise: Using the VMSA Client Software

Task – Determining VMSA Client Command Functions

Use these steps:

1. Access the Command Launcher (SSVM►Window►Command Launcher) and record the name of each mirror-related object.

Command: _____

Command: _____

Command: _____

Command: _____

2. In the Command Launcher window, select Disk►Scan.

3. Select the Task function in the Toolbar window.



4. In the Task Request window, double-click on the Disk►Scan entry to display the Task Properties window.

5. Record the executed command listings from the Task Properties window.

Command: _____

Command: _____

Command: _____

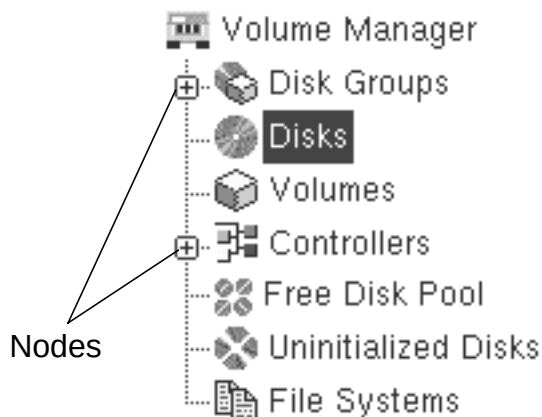
Note – All commands are also recorded on the SSVM server in the /var/opt/vmsa/logs/command file.

Exercise: Using the VMSA Client Software

Task – Defining VMSA Client Object Tree Functions

Complete the following steps:

1. Practice expanding and collapsing any nodes showing a plus or minus sign.



2. Highlight the Disk Groups portion of the object tree.
3. Select the Window ► Copy Main Grid entry. This will create a functional copy of the current grid display.

Note – You can create multiple copies of the grid and use them to display different sets of objects.

4. Close the grid copy you just created.
5. Highlight the Controllers portion of the Object Tree, and then click on one of the listed controllers.

Note – The item in the Menu bar that was previously labelled Selected has now changed to Controllers so you can now display a controller-related command menu.



Exercise: Using the VMSA Client Software

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences

- Interpretations

- Conclusions

- Applications

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Describe the VMSA server/client relationship
- ☐ Verify the VMSA server software is running
- ☐ Start the VMSA client software
- ☐ Use the main VMSA features
- ☐ Use the Options menu to customize the behavior of VMSA
- ☐ Describe two important uses for the Task Request Monitor

Think Beyond

When testing a large prototype system, you might need to destroy and re-create hundreds of volumes many times. Do you think the VMSA GUI will be effective in this kind of situation?

How can you easily manage a thousand mirrored volumes using the VMSA GUI?

Objectives

Upon completion of this module, you should be able to:

- Define the function and relationship of SSVM objects
- Display properties of SSVM objects
- Initialize a disk drive for SSVM use
- Create a disk group and add disks to it
- Rename a SSVM disk drive
- Remove a disk from a disk group
- Remove a disk from SSVM control
- Determine available free disk space
- Record the command line equivalent for any VMSA operation

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

- How do I identify available storage devices that I can use to create virtual volume structures?
- How can I tell if a disk is already under SSVM control?
- How can I determine the SSVM configuration when the VMSA GUI is not available?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The RAID Advisory Board. 1996. Lino Lakes, MN. *The RAID Book*.
- “Sun Performance Tuning Overview,” December, 1993, Part Number 801-4872-07.
- Wong, Brian. *Configuration and Capacity Planning for Solaris Servers*. ISBN 0-13-349952-9.
- Chen, Lee, Gibson, Katz, and Patterson. October, 1993. *RAID: High Performance, Reliable Secondary Storage*.



Sun Educational Services

SSVM Initialization Review

Before proceeding, review the following areas:

- The initialization or encapsulation process
- The private and public region format
- Initialized disk types

SSVM Initialization Review

When the SSVM software brings a disk under its control, it will examine the disk first and then determine how best to proceed. If data might be present on the existing disk, a variation of initialization can be performed.

Initialization

When the Sun StorEdge Volume Manager initializes a new disk, it creates two partitions: a small partition called the *private region*, and a large partition called the *public region* that covers the remainder of the disk.

Note – Throughout the rest of this module, the terms *block* and *sector* mean the same thing and are 512 bytes in size.

SSVM Initialization Review

Initialization

The public region is used for general space allocation.

The private region contains various administrative data for the Sun StorEdge Volume Manager, including the configuration database for all disks in a particular disk group.

Sun StorEdge Volume Manager uses tag 14 for the partition used for the public region and tag 15 for the private region partition. (The `prtvtoc` command displays information about a disk, including the tag information for each partition.)

Encapsulation

If you have existing data on the disk, you would not want to initialize the disk, as this destroys any data. Instead, you can choose to *encapsulate* the disk.

In order for Sun StorEdge Volume Manager to encapsulate the disk, there should be at least 1024 sectors in an unused slice at the beginning or end of the disk and two free partitions.

If a disk does not have 1024 sectors of space (one or two cylinders; depending on the geometry of the disk) and two free slices in the volume table of contents (VTOC), it can still be brought under Sun StorEdge Volume Manager control. It must, however, have a `nopriv` SSVM disk (see the following section) created for it. Because a `nopriv` SSVM disk does not contain a copy of the private region (which contains the configuration database for a disk group), a disk group cannot consist entirely of `nopriv` devices.

Encapsulation of the root disk is handled differently. It is preferable to give Sun StorEdge Volume Manager the space it needs for the private region. If, however, there is not enough space, it will take space from the end of `swap`.

SSVM Initialization Review

Private and Public Region Format

The private and public region format of an initialized SSVM disk can be verified with the `prtvtoc` command. In the following example, slice 2 is defined as the entire disk. Slice 3 has been assigned tag 15 and is 2016 sectors in size. Slice 4 has been assigned tag 14 and is the rest of the disk.

In this example, the private region is the first two cylinders on the disk. The disk is a 1.05-Gbyte disk and a single cylinder only has 1008 sectors or blocks, which does not meet the 1024 sector minimum size for the private region. This is calculated by using the `nhead=14` and `nsect=72` values for the disk found in the `/etc/format.dat` file.

```
# prtvtoc /dev/rdisk/c2t4d0s2
```

Partition	Tag	Flags	First Sector	Sector Count	Last Sector
2	5	01	0	2052288	2052287
3	15	01	0	2016	2015
4	14	01	2016	2050272	2052287

Initialized Disk Types

By default, SSVM initializes disk drives with the type Sliced. There are other possible variations. The three types of initialized disks are:

- Simple – Private and public regions are on the same partition.
- Sliced – Private and public regions are on different partitions (default).
- `nopriv` – There is no private region.

Note – The use of `nopriv` is strongly discouraged. It is normally used only for random access memory (RAM) disk storage on systems not built by Sun.

 *Sun Educational Services*

Storage Configuration

To build reliable virtual disk structures, you must be able to:

- Identify different types of storage devices
- Determine disk controller configurations

Storage Configuration

Identifying Storage Devices

The best way to identify the type and model of storage devices connected to your system is to read the product model tag and study the related technical manuals. Occasionally, you might be working with systems remotely and need to identify the hardware configuration using operating system commands and other tools.

Using the `luxadm` Command

The `luxadm` program is an administrative command that manages the SENA, RSM, and SPARCstorage Array subsystems. It can be used to find and report basic information about supported storage arrays as follows:

```
# luxadm probe
```

Unfortunately, the `probe` option only recognizes certain types of storage arrays. This is not comprehensive enough.

Storage Configuration

Identifying Storage Devices

Using the luxadm Command

The `luxadm` command can give very useful information if you know some basic controller addresses. It is still limited to certain storage models, and will give error messages if unsupported devices are examined. Some examples of `luxadm` output follow.

```
# luxadm disp c0
```

```
luxadm: Error opening /devices/io-  
unit@f,e0200000/sbi@0,0/dma@0,81000/esp@0,80000:ctlr  
No such file or directory
```

The `c0` controller is a standard SCSI interface, so `luxadm` cannot identify it.

```
# luxadm disp c1
```

```
SPARCstorage Array 110 Configuration
```

The `c1` controller is for a SPARCstorage Array 100 models which `luxadm` can identify.

```
# luxadm disp c3
```

```
luxadm: Error opening /devices/io-  
unit@f,e3200000/sbi@0,0/SUNW,socal@3,0/sf@1,0:ctlr  
No such file or directory
```

```
# luxadm probe
```

```
Found SENA Name:kestrel  
Node WWN:50800200000000878  
Logical Path:/dev/es/ses0  
Logical Path:/dev/es/ses1
```

The `c3` controller is a supported StorEdge A5000 array, but you must use a different `luxadm` option to see it. The `probe` option successfully discovers the array. To display the A5000 details, use the following command:

```
# luxadm display kestrel
```

Storage Configuration

Identifying Storage Devices

Using the format Utility

The Solaris `format` utility is the only reliable program for gathering basic storage configuration information. It is not the complete answer but it will report all storage devices, regardless of type or model. The following sample output shows three different types of storage devices:

AVAILABLE DISK SELECTIONS:

```
0. c0t0d0 <SUN1.05 cyl 2036 alt 2 hd 14 sec 72>
/io-unit@f,e0200000/sbi@0,0/dma@0,81000
/esp@0,80000/sd@0,0

1. c1t0d0 <SUN1.05 cyl 2036 alt 2 hd 14 sec 72>
/io-unit@f,e1200000/sbi@0,0/SUNW,
soc@3,0/SUNW,pln@a0000000,8023c7/ssd@0,0

2. c1t0d1 <SUN1.05 cyl 2036 alt 2 hd 14 sec 72>
/io-unit@f,e1200000/sbi@0,0/SUNW,
soc@3,0/SUNW,pln@a0000000,8023c7/ssd@0,1

3. c3t98d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
/io-unit@f,e3200000/sbi@0,0/SUNW,
socal@3,0/sf@1,0/ssd@62,0
```

From these examples, you can determine the following:

- The `esp` in the path name indicates device 0 is a standard SCSI interface.
- The `soc` in the path name indicates devices 1 and 2 are SPARCstorage Array 100 disks.
- The `socal` in the path name indicates device 3 is an FC-AL storage array.

Storage Configuration

Identifying Controller Configurations

The `format` utility can also be used to identify storage arrays that have multi-path controller connections.

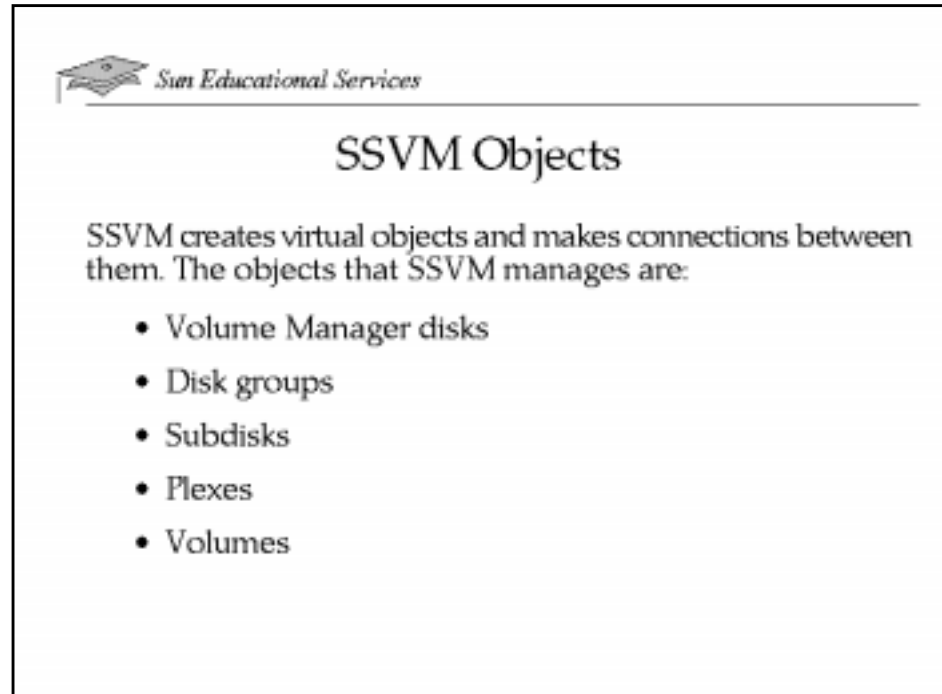
Identifying Dynamic Multi-Path Devices

DMP connections be identified using the `format` utility as follows:

AVAILABLE DISK SELECTIONS:

- 0. c0t0d0 <SUN2.1G cyl 2733 alt 2 hd 19 sec 80>
/sbus@3,0/SUNW,fas@3,88000000/sd@0,0
- 1. c2t33d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
/sbus@3,0/SUNW,socal@0,0/sf@0,0/ssd@w22000020370c0de8,0
- 2. c3t33d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
/sbus@3,0/SUNW,socal@0,0/sf@1,0/ssd@w21000020370c0de8,0

Notice that the device paths for devices 1 and 2 have the same disk drive identifier, 20370c0de8. Since the controller numbers are different, they are connected to two different controller interfaces in the same system.



SSVM Objects

The connection between physical objects and virtual objects is made when you place a disk drive under SSVM control.

SSVM creates virtual objects and makes logical connections between the objects. The virtual objects are then used by SSVM to do storage management tasks

Sun StorEdge Volume Manager Disks

There are two phases to bringing a physical disk drive under SSVM control. Sometimes, both operations are done in one step and you are unaware that the process is more complex.

SSVM Objects

Sun StorEdge Volume Manager Disks

Free Disk Pool

When you use the VMSA application to bring a disk drive under SSVM control, you can:

- Add it to an existing disk group
- Add it to a new disk group
- Add it to the free disk pool

The simplest operation is to add it to the free disk pool. The `vxdisksetup` command is used to repartition the disk into SSVM format and then a blank header is written to the disk.

If you add it to a disk group, the disk is assigned a unique name and associated with a disk group object. This information is then written into the blank header on the disk. Unless you intervene, the default names given to disks are `disk01`, `disk02`, and so on.

Disk groups will be discussed in more detail in the following section.

Disk Groups

A disk group is a collection of SSVM disks that share a common configuration. Typically the disk group contains volumes that are all related in some way such as file system volumes that belong to a particular department or database volumes that are all tables for a single database.

Each disk group is owned by a single host system. The current ownership is written into all configuration records. Many of the disks in the disk group have a copy of the configuration record.

A disk group and all of its components can be moved as a unit from one host system to another. Usually both host systems are connected to the same dual-ported storage arrays.

SSVM Objects

Subdisks

A subdisk is a set of contiguous disk blocks. A subdisk must reside entirely on a single physical disk. The public region of a disk in a disk group can be divided into one or more subdisks. The subdisk cannot overlap or share the same portions of a public region.

By default, subdisks are named based on the name of the disk on which they reside. This relationship is shown in Figure 5-1.

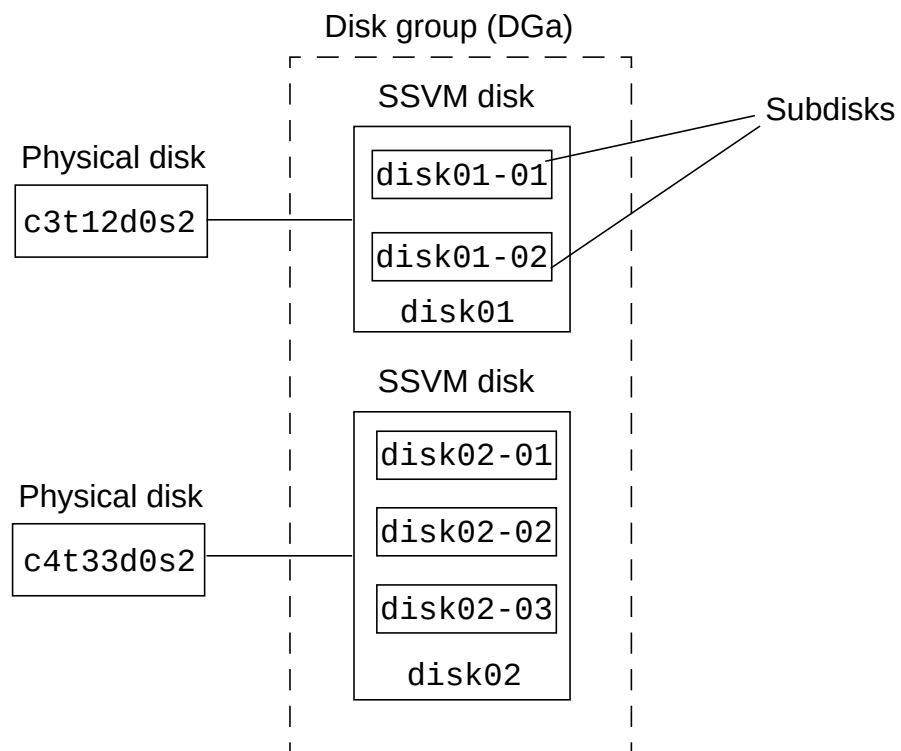


Figure 5-1 Subdisk Naming Conventions

Note – As shown in Figure 5-1, the disk drives are on different controllers indicating they are in different storage arrays. Disk groups can span storage arrays.

SSVM Objects

Plexes

The SSVM application uses subdisks to build virtual objects called *plexes*. A plex consists of one or more subdisks located on one or more physical disks. Figure 5-2 illustrates the relationship of subdisks to plexes.

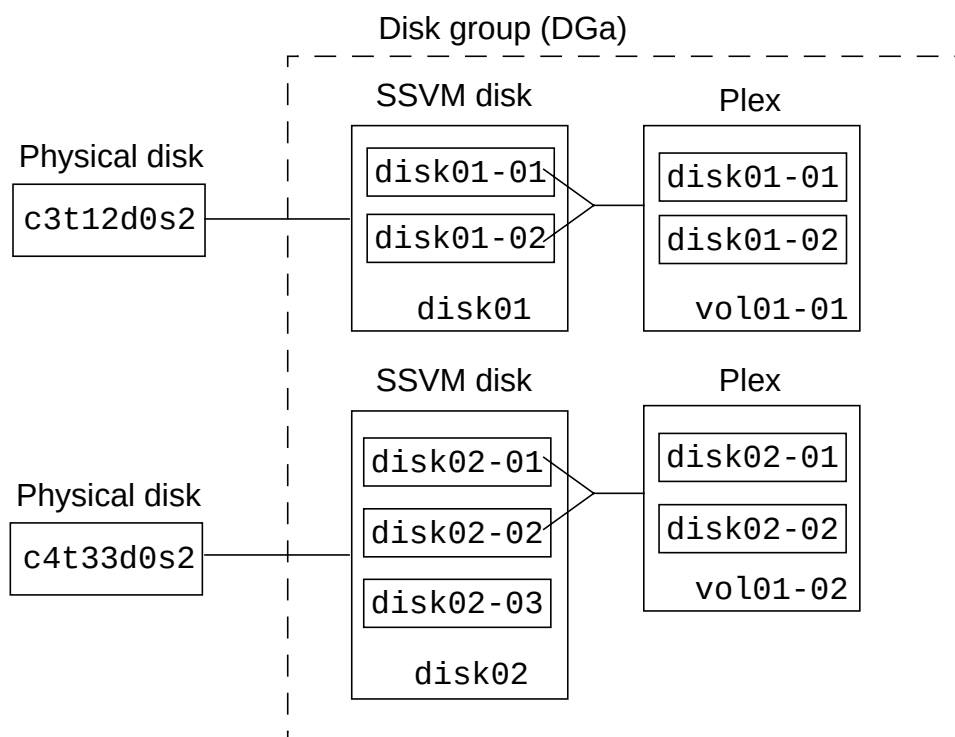


Figure 5-2 Plex Configurations

The data to be stored on the subdisks of a plex can be organized by using any of the following methods:

- Concatenation
- Striping
- Striping with parity (RAID 5)

SSVM Objects

Volumes

A volume consists of one or more plexes. By definition, a volume with two plexes is mirrored. Figure 5-3 illustrates the relationship of plexes in a mirrored volume.

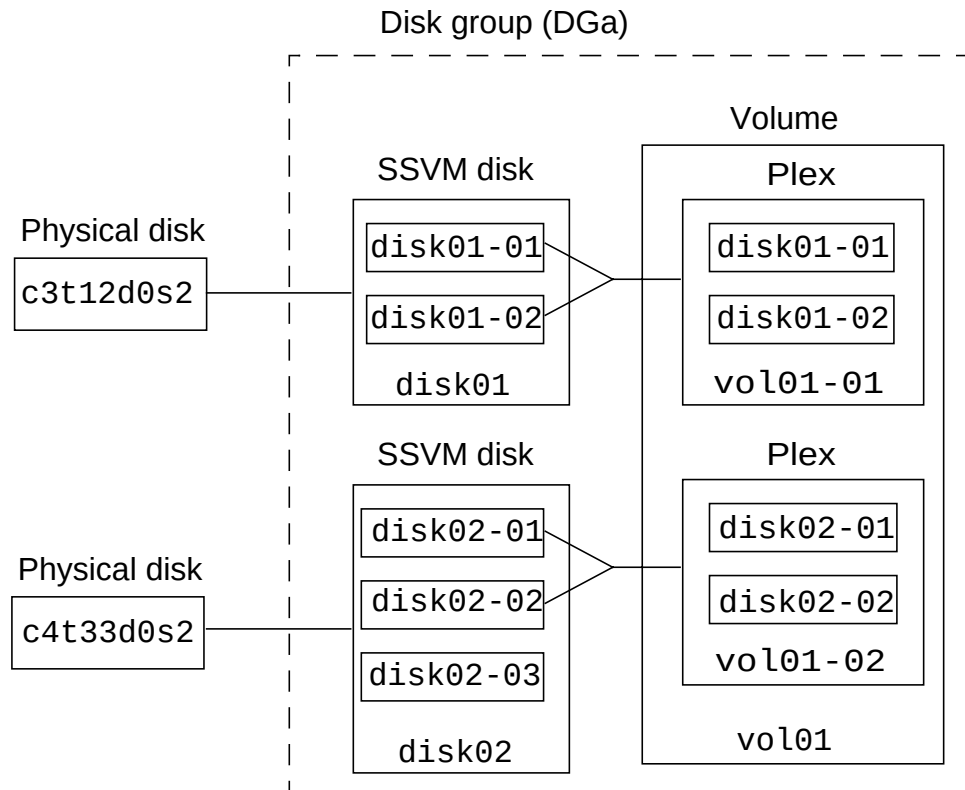


Figure 5-3 Mirrored Volume Structure

Although there are many important points about volumes, the basic points you should understand now are that:

- Volumes can have more than two mirrors.
- RAID-5 volumes cannot be mirrored.
- A plex can also be a log structure, which is not used for data storage



Sun Educational Services

Command-Line Status

As early as possible, learn how to use three very important command-line tools to gather SSVM status:

- `vxprint`
- `vxdisk`
- `vxdbg`

Command-Line Status

Using `vxprint`

When analyzing volume problems, the `vxprint` utility is an essential tool. Unlike the VMSA graphical interface, `vxprint` displays all information by using the terms *disk group*, *plex*, *volume*, and *subdisk*. The following sample illustrates how a typical concatenated volume would appear in the `vxprint` output.

Command-Line Status

Using vxprint

```
# vxprint -g skyvols -ht
DG NAME   NCONFIG  NLOG      MINORS    GROUP-ID
DM NAME   DEVICE   TYPE      PRIVLEN   PUBLLEN   STATE
V  NAME   USETYPE  KSTATE    STATE     LENGTH    READPOL   PREFPLEX
PL NAME   VOLUME   KSTATE    STATE     LENGTH    LAYOUT    NCOL/WID
SD NAME   PLEXDISK DISK       OFFS      LENGTH    [COL/]OFF DEVICE

dg DGa     default  default  87000    skydome

dm c1t0d3  c1t0d3s2 sliced   2015     2050272  -
dm c1t0d4  c1t0d4s2 sliced   2015     2050272  -

v  vol01   fsgen    ENABLED   ACTIVE    4096512   ROUND     -
pl plex01  vol001   ENABLED   ACTIVE    4096512   CONCAT    -
sd sd01    plex01   disk01    0         2048256   0         c1t0d3
sd sd02    plex01   disk02    0         2048256   2048256   c1t0d4
```

You can determine the following details from this sample output:

- The disk group is called **DGa**, and it contains a single volume.
- The volume is called **vol01**, and it has a single plex, **plex01**.
- The plex is concatenated, and it has two subdisks created from the SSVM disks, **disk01** and **disk02**.
- Each subdisk is approximately 1.02 Mbytes in size.
- The device paths to the two disks are **c1t0d3** and **c1t0d4**.
- The volume and plex are enabled and active.

Note – Usually, SSVM utilities display size values in disk blocks (sectors). Divide the numbers by 2000 to convert to Mbytes.

Command-Line Status

Using vxdisk

The `vxdisk` command has many task related options, but the most commonly used option is `list`. The `vxdisk list` command displays a quick summary of the state and ownership of all disks attached to the system. A typical `vxdisk list` output appears as follows:

```
# vxdisk list
DEVICE      TYPE      DISK      GROUP      STATUS
c0t0d0s2    sliced    -         -          error
c2t33d0s2    sliced    droot     rootdg     online
c2t35d0s2    sliced    dga01     DGa        online
c2t37d0s2    sliced    -         -          online
c2t50d0s2    sliced    -         -          online
c2t52d0s2    sliced    -         -          error
c3t1d0s2     sliced    -         -          online
c3t3d0s2     sliced    -         -          error
c3t5d0s2     sliced    -         -          error
c3t18d0s2    sliced    dga02     DGa        online
c3t20d0s2    sliced    -         -          error
```

By examining this sample `vxdisk` output, you can determine the following:

- All devices with error status are not under SSVM control.
- All devices with an online status have been initialized at some level.

If devices do not have a disk name and are not part of a disk group, they are not yet fully initialized. They have only been repartitioned for SSVM and have a blank disk header.

Note – If a disk shows a status of `failed was c0t0d0`, it means there has been a major failure and SSVM cannot access the physical disk but it knows what the address was before the failure. You will find that the `failed was` address will not be displayed in a `vxprint -ht` output.

Command-Line Status

Using vx dg

The `vx dg` command can be used to create and manipulate disk groups. It can also provide very important information about the amount of free disk space left in a disk group. The following `list` option output can be valuable when preparing to add new volumes to an existing disk group:

```
# vx dg -g DGa free
DISK    DEVICE    TAG      OFFSET    LENGTH    FLAGS
disk01   c3t1d0s2  c3t1d0   1843647   206625    -
disk02   c3t1d1s2  c3t1d1   2046240   4032      -
disk03   c4t2d0s2  c4t2d0    0         2050272   -
```

The `LENGTH` field shows the amount of free space left on each of the disks in the disk group. The values are shown in disk blocks or sectors. You can convert them to Mbytes by dividing the length by 2000.

By examining this sample `vx dg` output, you can determine the following:

- The `DGa` disk group has three physical disks.
- One of the disks, `disk03`, is on a different controller.
- The total amount of free space available is about 1.1 Gbytes.

$$(206625 + 4032 + 2050272) \div 2000 = 1130.46 \text{ Mbytes}$$
- You can create a 100-Mbyte concatenated / mirrored volume.
- You can create a 4-Mbyte RAID-5 volume.

Exercise: Performing SSVM Disk Drive Operations



Exercise objective – In this exercise you will:

- Display properties of SSVM objects
- Initialize a disk drive for SSVM use
- Create a disk group and add disks to it
- Rename a SSVM disk drive
- Remove a disk from a disk group
- Remove a disk from SSVM control
- Determine available free disk space
- Record the command line equivalent for a VMSA operation

Preparation

Ask your instructor to furnish the following information:

- Two code letters for your work group (a and b, c and d, e and f, and so on).

Work group code letters: _____

- The physical paths to six disk drives for your work group.

Disk: _____ Disk: _____

Disk: _____ Disk: _____

Disk: _____ Disk: _____

Exercise: Performing SSVM Disk Drive Operations

Task – Verifying Initial Disk Status

Complete the following steps:

1. Log in either directly or remotely to the SSVM server and enter the `vxdisk list` command.
2. Check the output carefully and verify that each of the disks that were assigned to your work group show a status of **error**.

If any of your assigned disks show a non-error status, check with your instructor. If your instructor thinks it is appropriate, use the `vxdiskunsetup` command as shown to remove them from SSVM control.

```
# /usr/lib/vxvm/bin/vxdiskunsetup -C c0t22d0
```

Note – You must substitute the physical path to your disk drives.

Do not proceed until you are sure your assigned disk drives are *not* under SSVM control.

Task – Creating the First Disk Group

You are going to create a disk group with two of your assigned disk drives in it. You will name the disk group according to your first work group letter. If your work group letters are a and b, then this first disk group will be named **DGa**.

1. Start the VMSA client software in the background if it is not already running.

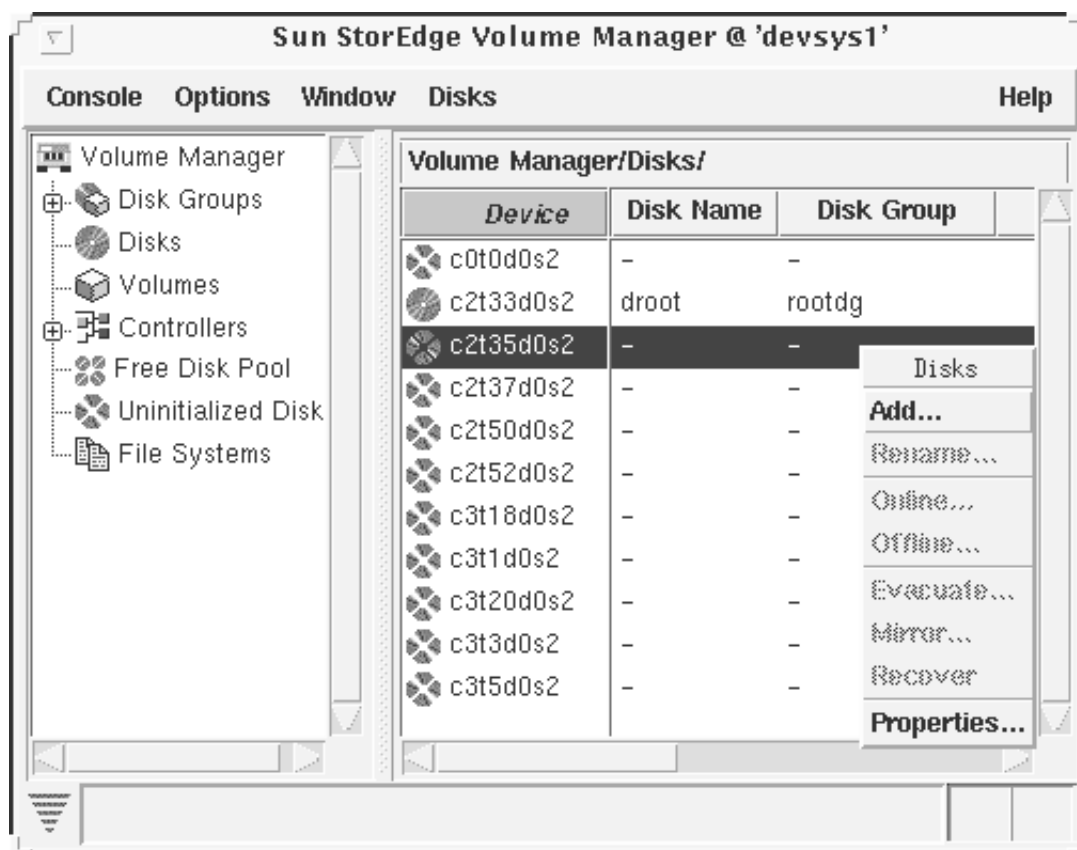
```
# /opt/VRTSvmsa/bin/vmsa
```

2. Select the Disks icon in the Object Tree window.

Exercise: Performing SSVM Disk Drive Operations

Task – Creating the First Disk Group

3. Select one of your assigned disks in the VMSA Grid window and use the pop-up menu (which you display by clicking the right mouse button) to select the Add function.



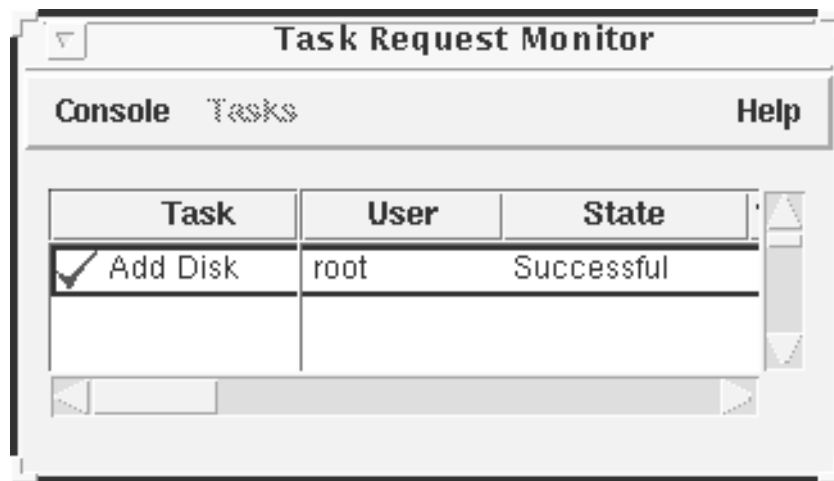
4. Verify that the physical path in the Add Disks form is correct and the Free Disk Pool is selected.
5. Select OK in the Add Disk form.
6. Carefully read the information in the second Add Disk form.

Note – The warnings are important and can indicate that you are about to initialize a disk with data on it or that is already in use by SSVM. The software tries to anticipate possible mistakes.

Exercise: Performing SSVM Disk Drive Operations

Task – Creating the First Disk Group

7. Select Initialize on the second Add Disk form.
8. Open the Task Request Monitor window from the tool bar and verify the operation completed successfully.



9. Double-click on the Add Disk task entry so that the Task Properties window is displayed.
10. Record the Executed Commands section of the Task Properties window.
 Commands: _____
11. Cancel (Exit) the Task Properties window.
12. Select two more disks in the VMSA Grid window and add them to the free disk pool. If possible, select disks that are on different controllers than the first one you initialized.
13. Click on the Uninitialized Disks icon in the Object Tree window and verify that your initialized disks are no longer displayed there.
14. Click on the Free Disk Pool icon in the Object Tree window and verify that the two disks you just initialized are displayed as members of the free disk pool.

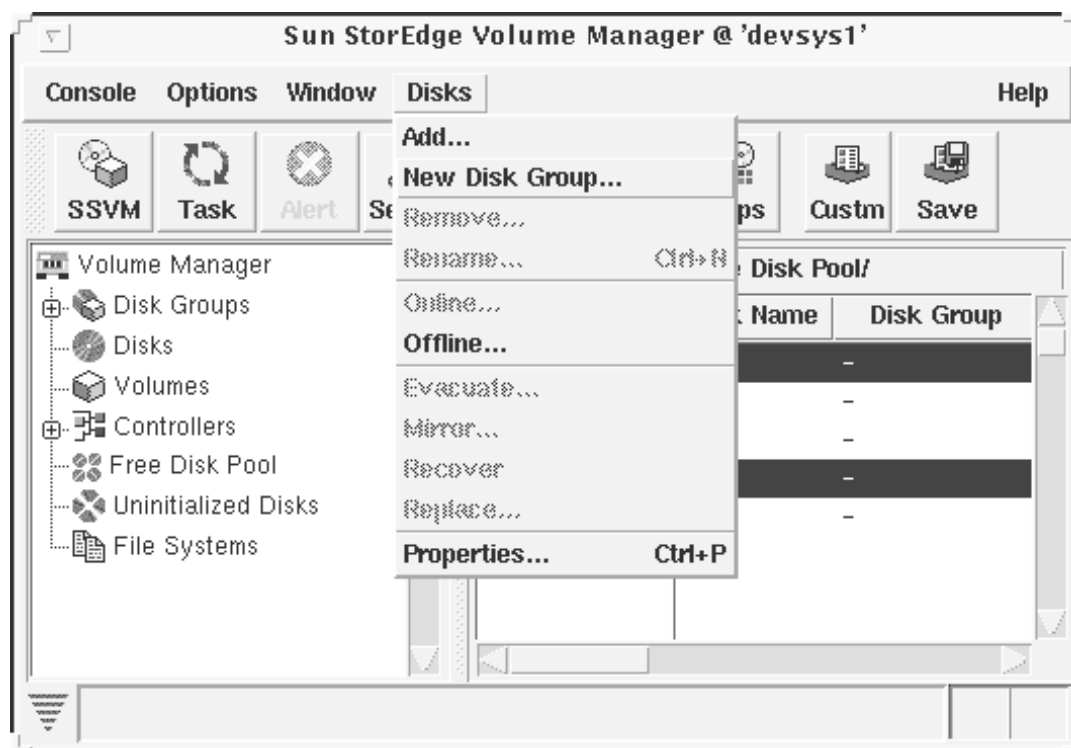
Exercise: Performing SSVM Disk Drive Operations

Task – Creating the First Disk Group

15. In the VMSA Grid window, select the three disks you have just initialized.

Note – Use the Control key with the left mouse button to select the second disk.

16. In the menu bar, click on the Disks menu and then select New Disk Group.



Note – Remember, the Selected area in the menu bar changes when different types of objects are displayed in the Grid window.

Exercise: Performing SSVM Disk Drive Operations

Task – Creating the First Disk Group

17. Use your first work group code letter as part of the disk group name and disk drive names as shown. This example uses the letter “a”.

New Disk Group @ 'devsys1'

Disk Group Name:

Disk Device(s):

Options:

☒ Create Shared

Disk Name(s):

Comment:

18. When you have finished configuring the New Disk Group form, select OK to perform the operation.
19. In the Task Request Monitor window, double-click on the Add Dsk Grp task and record the Executed Commands section of the Task Properties window.
Commands: _____
20. Cancel (Exit) the Task Properties window.
21. Expand the Disk Groups node in the Object Tree window and examine your new disk group.

Exercise: Performing SSVM Disk Drive Operations

Task – Creating the Second Disk Group

Complete these steps:

1. Display the Free Disk Pool in the Grid window and select the three additional disks you were assigned.
2. In the menu bar under the Disks menu, select New Disk Group.
3. Use your second work group code letter as part of the disk group name and disk drive names. If your second work group code letter is “b”, then configure the New Disk Group form as follows:

Disk group name: DGb

Disk name(s): dgb01 dgb02 dgb03
4. Select OK on the New Disk Group form.
5. Double-click on the Disk Group task in the Task Request Monitor window and review the Executed Commands section.
6. Cancel (Exit) the Task Properties window.
7. Verify the status of both of your new disk groups in the VMSA Grid window.
8. On the SSVM server, verify the status of your new disk group with the `vxprint` and `vxdisk list` commands.

Exercise: Performing SSVM Disk Drive Operations

Task – Verifying Free Disk Space

When you get ready to build volumes on disk drives in a disk group, create a volume as large as possible that uses all of the available disk space. This requires some care. There are two methods of determining available disk space:

- Using the VMSA Client interface
- Using SSVM command-line options

VMSA Unused Disk Space

Complete the following steps:

1. Expand your first disk group in the object tree and display the disks in the grid. This should be the disk group DGx (for example, DGa).
2. Use the scroll bar on the VMSA main window to display the unused space in the disk group. Record the results.

dgx01 unused: _____

dgx02 unused: _____

dgx03 unused: _____

Command-Line Unused Disk Space

Although you might think you can create a mirrored volume equal to the size of dgx01 or dgx02, use a command-line utility and verify this amount before trying to build a volume.

3. On the SSVM server, use the `vx dg` command to verify the available space on the disks in your first disk group.

`vx dg -g DGa free`

Note – The LENGTH column is the free space available in blocks. Divide this amount by 2000 (or 2048) to convert to Mbytes.

Exercise: Performing SSVM Disk Drive Operations

Task – Verifying Free Disk Space

Command-Line Unused Disk Space

4. On the SSVM server, use the `vxassist` command to verify the available space for creating a maximum-size, mirrored volume on two disks in your first disk group.

```
# vxassist -g DGA maxsize layout=mirror dga01 dga02
```

Note – Substitute the name of your disk group and disks.

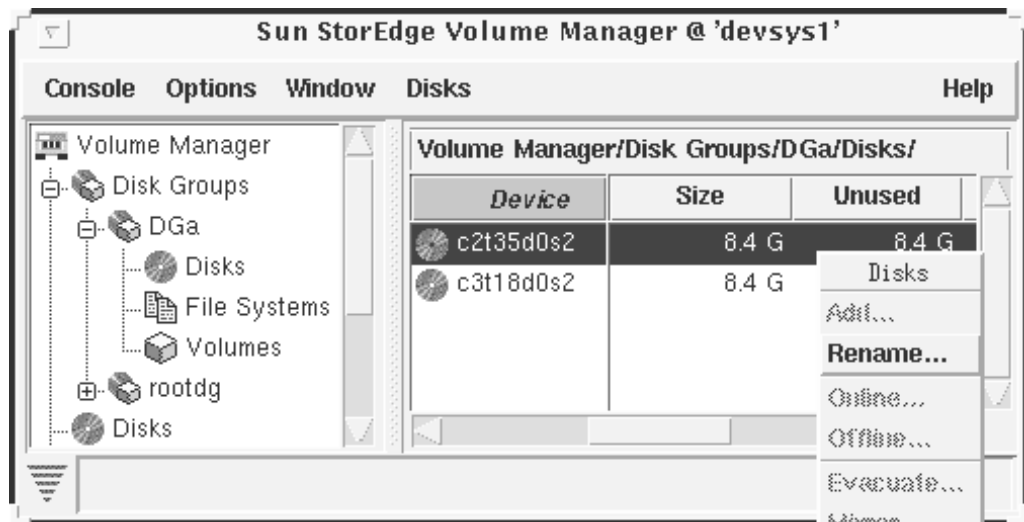
5. Discuss the results of the `vxdg` and `vxassist` commands with your instructor.

Exercise: Performing SSVM Disk Drive Operations

Task – Renaming Disk Drives

Complete the following steps:

1. Display one of your disk groups and the VMSA Grid areas. Select the Rename function from the pop-up menu.



2. Rename the disk to be xyzzy.
3. Rename the disk again and restore its original name.

Exercise: Performing SSVM Disk Drive Operations

Task – Removing Disks From a Disk Group

There are two levels of removing a disk drive. They are:

- Remove from a disk group, return to free disk pool

Note – Remove the check mark next to *evacuate the disk* if using this method.

- Remove from SSVM control, return to uninitialized state

For example:

```
# /usr/sbin/vxdg -g DGx rmdisk dgx02
```

Complete the following step:

1. Practice removing and then adding a disk drive in one of your disk groups.

Note – Discuss this with your instructor if you have any questions.

Task – Finishing Up

Make sure both of your disk groups are still complete and meet the following guidelines:

- You have two disk groups name DGx and DGy.
- The disk group DGx has three disks in it, and they are named dgx01, dgx02, and dgx03.
- The disk group DGy has three disks in it, and they are named dgy01, dgy02, and dgy03.

Note – Substitute your work group codes for the x and y in DGx and DGy.

Exercise: Performing SSVM Disk Drive Operations

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences

- Interpretations

- Conclusions

- Applications

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Define the function and relationship of SSVM objects
- ☐ Display properties of SSVM objects
- ☐ Initialize a disk drive for SSVM use
- ☐ Create a disk group and add disks to it
- ☐ Rename a SSVM disk drive
- ☐ Remove a disk from a disk group
- ☐ Remove a disk from SSVM control
- ☐ Determine available free disk space
- ☐ Record the command-line equivalent for a VMSA operation

Think Beyond

Are there other advantages to creating multiple disk groups besides general administrative organization?

What advantage is there to limiting the number of disks that are in the rootdg disk group?

Objectives

Upon completion of this module, you should be able to:

- Create simple, striped, and RAID-5 volumes
- Remove a volume
- Add a mirror to a volume
- Remove a mirror from a volume
- Resize a volume (make it larger)
- Display properties of a volume
- Display volume mapping
- Add a file system to a volume
- Add a dirty region log to a mirrored volume
- Add a log to a volume

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

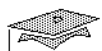
- When would it be appropriate to use a very simple volume structure with no data redundancy?
- Are some of the command-line programs more important than others?
- How can I be sure I have used every bit of available disk space?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The RAID Advisory Board. 1996. Lino Lakes, MN. *The RAID Book*.
- “Sun Performance Tuning Overview,” December, 1993, Part Number 801-4872-07.
- Wong, Brian. *Configuration and Capacity Planning for Solaris Servers*. ISBN 0-13-349952-9.
- Chen, Lee, Gibson, Katz, and Patterson. October, 1993. *RAID: High Performance, Reliable Secondary Storage*.



Sun Educational Services

Disk Group Review

When designing and administering disk groups, you should be familiar with the following:

- Disk groups have two primary functions
- The `rootdg` disk group should be small
- Disk groups should have unique names
- A disk group should have at least two disks
- Moving populated disks requires care

Disk Group Review

A disk group is a collection of SSVM disks that share a common configuration. The default disk group is `rootdg`. Volumes are created within a disk group using the SSVM drives which exist in that group.

Primary Functions of a Disk Group

Disk groups have two primary functions:

- Assist administrative management
- Provide higher data availability

Disk Group Review

Primary Functions of a Disk Group

Administrative Advantage

Disk groups enable the system administrator to group disks into logical collections for administrative convenience, grouping according to department or application; for example, sales, finance, and development.

Increased Data Availability

A disk group and its components can be moved as a unit from one host machine to another. This feature provides a higher availability to the data in the following ways:

- The first system *deports* the disk group.

Deporting a disk group disables access to that disk group by that host. Another host can then import the disk group and start accessing all disks in the disk group.

- The second system *imports* the disk group and starts accessing it.

A host can only import disk groups with unique names. Therefore, all disk groups on all systems (with the exception of `rootdg`, which is required) should be given unique names.

If one system fails, another system running Sun Enterprise Volume Manager can import its non-`rootdg` disk groups and provide access to them.

Disk Groups Review

Disk Group Requirements

These include:

- All disk groups across all systems should have unique names.

This makes it easier to move them between hosts and to differentiate their functionality. They can be renamed during the process of importation.
- All disk groups which reside on one host *must* have unique names.

This is an application restriction.
- Each system must have a disk group named `rootdg`

This is an application restriction.
- In general, the `rootdg` disk group should be kept small.

The `rootdg` disk group has a special relationship with the SSVM software and is therefore more difficult to deport or import to another system during system failures. It must be renamed because the backup system also has a disk group named `rootdg`.
- All disk groups must contain at least one disk.

However, at least two disks per disk group are required so that copies of the disk group configuration can be stored on multiple disks for redundancy reasons.

Movement of SSVM Disks Between Disk Groups

It is easy to move an entire disk group between hosts. It is also easy to move an empty SSVM disk (one that does not contain any SSVM objects) between hosts.

However, it is more complex to move one or more populated SSVM disks from one disk group to another. When SSVM disks are removed from a disk group, the configuration information is not saved. Care should be taken when moving disks between disk groups.



SSVM Volume Definition

Before constructing a volume, the following must be defined:

- Volume disk group
- Volume disk(s)
- Volume name
- Volume size
- Volume type (structure)

SSVM Volume Definition

Creating volume structures is easy to do. The tools seem simple. It is also easy to make mistakes unless you understand each aspect of the volume creation process.

Selecting a Disk Group

A common mistake is to place all of the disk drives in the default `rootdg` disk group. The configuration records for a disk group cannot contain information for more than 2048 objects. Each volume, plex, subdisk, and disk drive is considered an object and requires 256 bytes of private region space. The default private region length is 1024 blocks. There is a maximum of 2048 objects in a disk group.

Another reason for breaking disks into separate groups is that you might want to deport a disk group and import it to another connected host systems. This can be part of a disaster recovery plan or a load balancing measure. You cannot deport the `rootdg` disk group.

SSVM Volume Definition

Selecting a Disk Group

A disk group can be designed so that it is better for particular tasks. Each disk group shown in Figure 6-1 has three disks and each disk is in a different storage array.

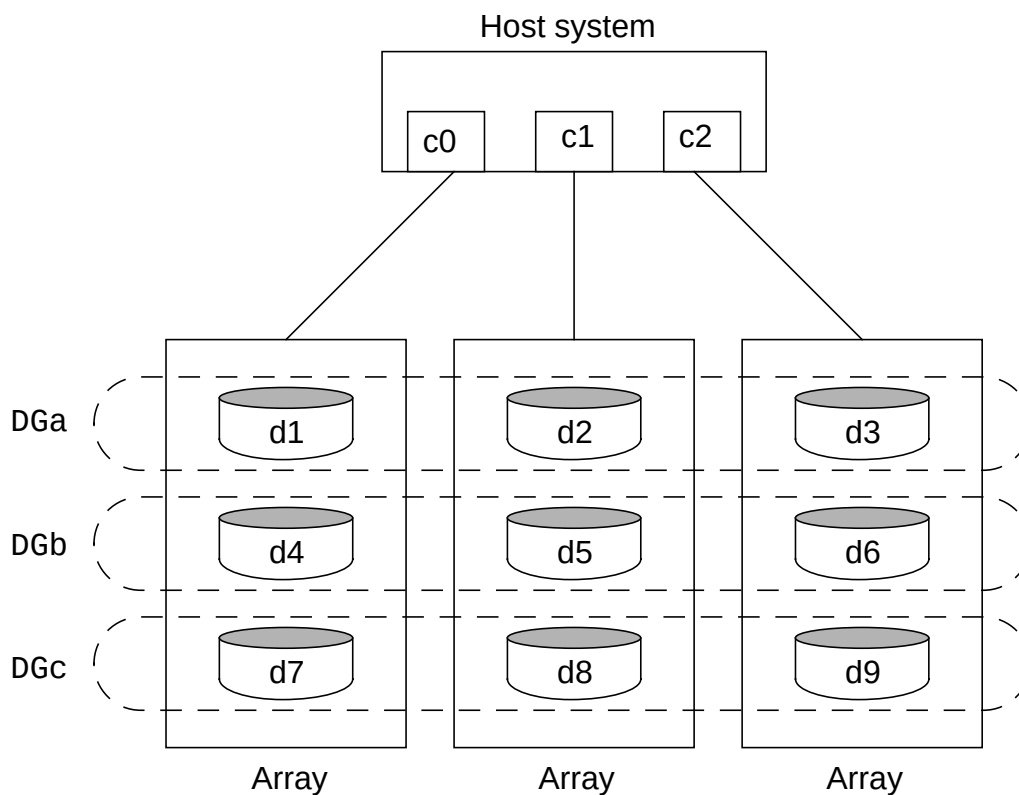


Figure 6-1 Disk Groups for Striped Volumes

Disk groups organized in this manner would be very good for creating striped volumes types such as RAID 5 and for mirrored volumes. The most important feature is that each disk in the disk group is in a separate box and on a different controller.

Note – Care must be taken with disk groups that span storage arrays. You must be sure that the loss of an entire array will not disrupt both mirrors in a volume or more than one column in a RAID-5 volume.

SSVM Volume Definition

Selecting a Disk Group

Another disk group structure, such as the one shown in Figure 6-2, would be better utilized with straight concatenated volumes.

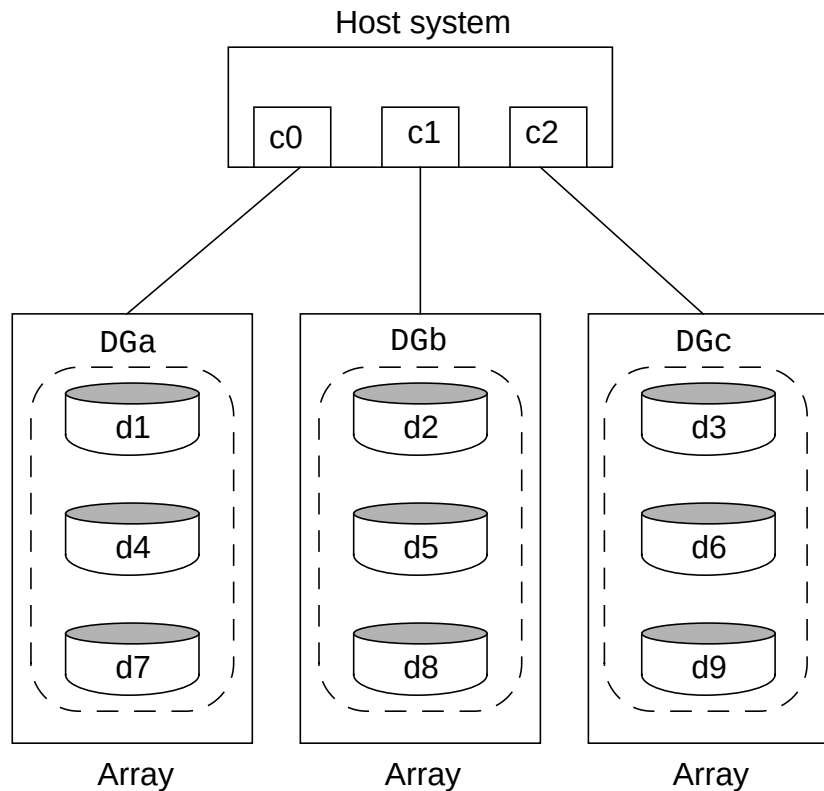


Figure 6-2 Disk Groups for Concatenated Volumes

Perhaps the volumes are large, static, read-only structures that only need a periodic backup to tape. They do not need any higher level of reliability or availability.

SSVM Volume Definition

Using Volume Naming Conventions

Unless you override the default values, the SSVM software will automatically furnish a name for each new volume created. The name will be systematic such as vol01, vol02, vol03, and so on. The problem with this is that each of the volumes may have very different features that are not reflected in the name. Typical naming conventions reflect volume attributes such as:

- The volume structure
- Which department uses them
- Which database they are associated with
- Special purposes within a work group

Although naming conventions do not seem to be of much importance, they can help establish priorities during emergency situations such as major power outages.

Determining Volume Size

Although choosing a general size for a volume is frequently dictated by the application, administrators frequently want to use as much space as is practical on a set of disk drives. There are many way to get space maximum space for a volume. Among them are:

- Let the SSVM software automatically find the space.
- Limit the search for space to selected disks in a group.
- Research available space with command-line programs.

Automatic Space Allocation

If you do not specify anything more than a disk group name, the SSVM software can find pieces of unused disk space and assemble them into a volume. This can lead to a very disorganized structure and create very poor performance for some volume types.

SSVM Volume Definition

Determining Volume Size

Restricted Space Allocation

Rather than letting SSVM find space anywhere within a disk group, it is better to define several disks you want to use. Direct SSVM to find the maximum space available, but you choose the disk drives that are better suited for the type of volume you want. The illustration in Figure 6-3 demonstrates the point very well.

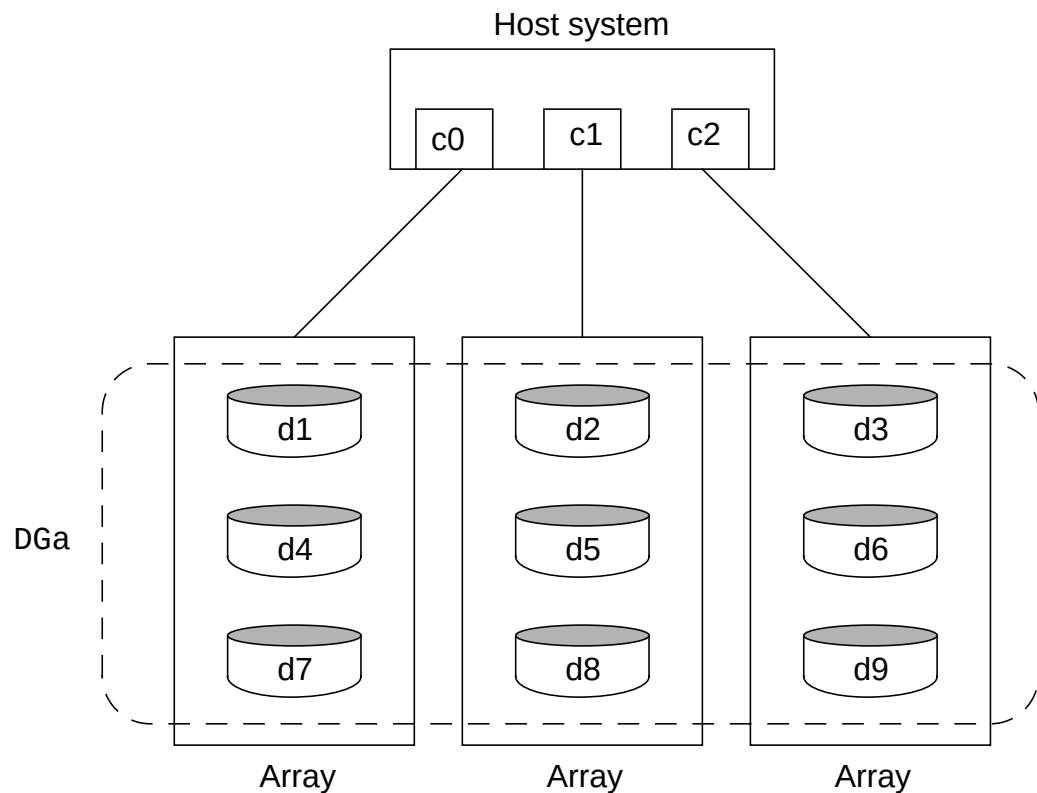


Figure 6-3 Selecting Disks for a Volume

If you wanted to created a RAID 5 volume, you might select disks d1, d2, and d3. For a concatenated volume, you might use disks d1, d4, and d7. For a mirrored and concatenated volume, you might use disks d1, d4, and d7 for one mirror and disks d3, d6, and d9 for the other mirror.

SSVM Volume Definition

Determining Volume Size

Researched Space Allocation

It is frequently better to spend some time analyzing free disk space before creating a volume. Look for patterns of free space that fit your needs. Examples of some commands that can be used to research free space in a disk group are:

```
# vxdg free
# vxassist maxsize
```

The following examples demonstrate the use of these commands:

```
# vxdg -g DGa free
DISK      DEVICE    TAG      OFFSET    LENGTH    FLAGS
disk01    c3t1d0s2  c3t1d0   1843647   206625    -
disk02    c3t1d1s2  c3t1d1   2046240   4032      -
disk03    c3t2d0s2  c3t2d0   0         2050272   -
```

```
# vxassist -g DGa maxsize \
layout=nomirror,nostripe disk01 disk02 disk03
```

Maximum volume size: 2258944 (1103Mb)

```
# vxassist -g DGa maxsize \
layout=raid5,nolog disk01 disk02 disk03
```

Maximum volume size: 6144 (3Mb)

SSVM Volume Definition

Identifying Volume Types

The SSVM application supports the following general types of volume structures:

- Simple concatenation
- Simple striped
- Mirrored (concatenated or striped)
- RAID 5 (striped with parity)

Simple Concatenation

This involves:

- Efficient use of storage space
- Simpler hardware requirements

Simple Striping

This structure provides:

- Better read and write performance

Mirroring

Some benefits of this structure are:

- Improved reliability with both concatenation and striping
- Fully redundant copy of the data

SSVM Volume Definition

Identifying Volume Types

RAID 5

One advantage of this structure is:

- Somewhat improved reliability



Sun Educational Services

Volume Creation Using VMSA

All volumes are created in VMSA using a single form that has the following features:

- Everything except the disk group is optional.
- There is browser-assisted assignment of disk group and disks.
- Special SSVM layered volumes are available.
- File systems can also be configured.

Volume Creation Using VMSA

The volume creation process can be initiated in VMSA using the following:

- The Toolbar New button
- The Menu bar: Console►New►Volume entry
- The Command Launcher: Volume►Create entry

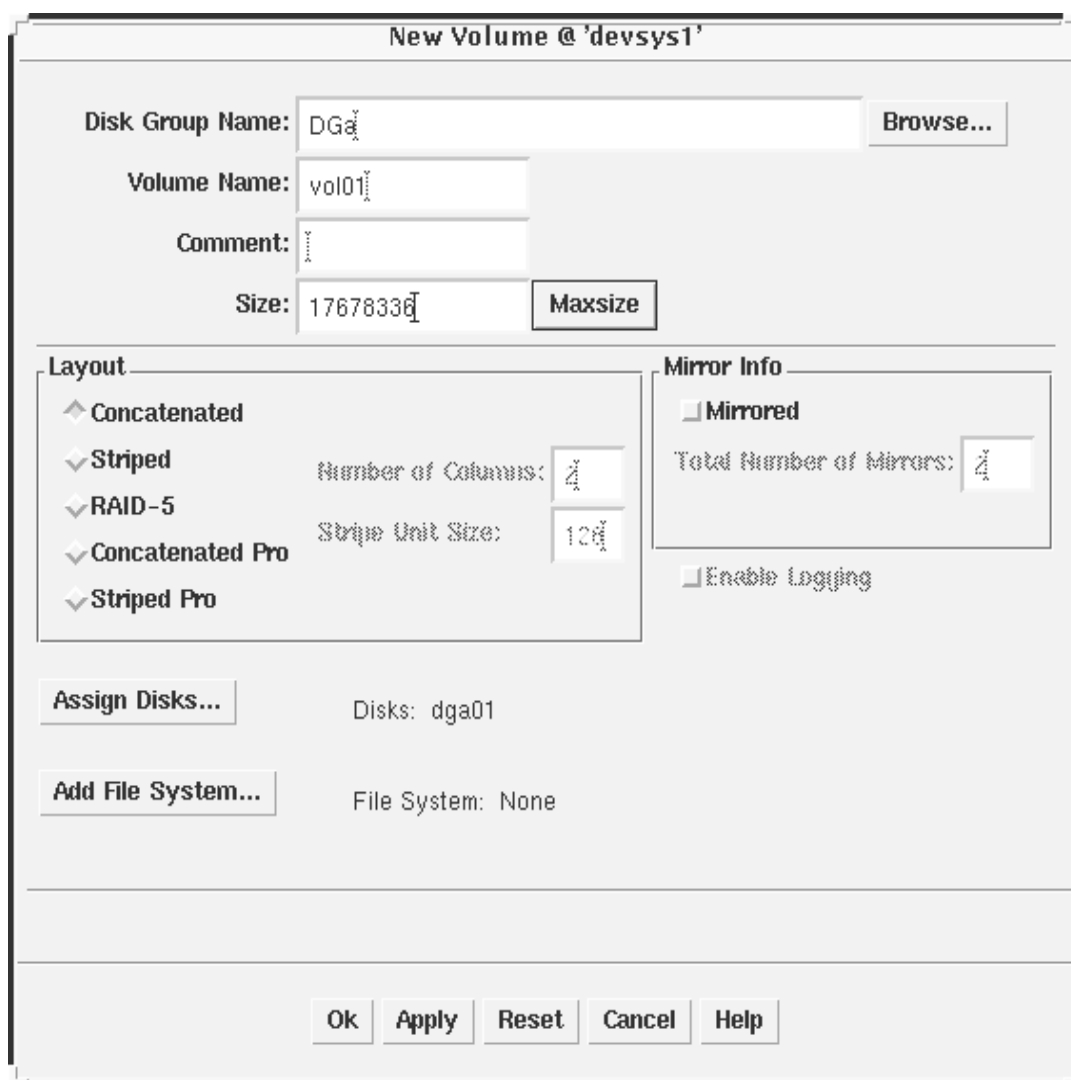
Regardless of how you initiate a volume creation session, the same New Volume form is displayed.

Much of the information in the form does not need to be furnished. VMSA will substitute default values, but it is not a good idea to use them. You might get a volume that does not meet your needs.

Volume Creation Using VMSA

The New Volume Form

All new volume creation is done using the New Volume form (Figure 6-4).



New Volume @ 'devsys1'

Disk Group Name:

Volume Name:

Comment:

Size:

Layout

- ☒ Concatenated
- ☒ Striped Number of Columns:
- ☒ RAID-5 Stripe Unit Size:
- ☒ Concatenated Pro
- ☒ Striped Pro

Mirror Info

☐ Mirrored

Total Number of Mirrors:

☐ Enable Logging

Disks: dga01

File System: None

Figure 6-4 VMSA New Volume Form

The VMSA software will automatically select default values for many of the values on the form.

Volume Creation Using VMSA

The New Volume Form

Consider the following points when configuring volumes using the VMSA New Volume form:

- Default volume names might not be clear enough for administrative purposes.
- Using the Maxsize button can result in a volume composed of many small subdisks that are randomly located, which can create some very bad performance bottlenecks. You could create huge volumes with file systems that take a very long time to complete the `newfs` and mirror synchronization phases.
- The Number of Columns value only applies when creating striped and RAID-5 layouts. It will be entered automatically if you pre-selected disk drives by highlighting them in the Grid window.
- The Assign Disks browser can be used if you did not pre-assign disk drives.
- The Add File System button displays an additional form that enables you to configure a fully operational file system on the new volume.

You can have all aspects of a new file system created automatically including:

- ▼ The mount point
- ▼ The `/etc/vfstab` entry
- ▼ The `newfs` and `mkfs` operations
- ▼ First time mounting
- ▼ Volume ownership and protection
- ▼ File system type (UNIX file system [UFS] or Veritas file system [VxFS])

Note – The VxFS file system software is ordered and licensed separately from the basic SSVM software.



Volume Creation Using the Command Line

You can create volumes using only the `vxassist` command. It can be used to:

- Create volumes of all types
- Mirror existing volumes
- Grow or shrink volume sizes
- Assist in preparing for volume backups

Volume Creation Using the Command Line

The vxassist Command Format

The `vxassist` command has many options. Most of the options have default values if not explicitly entered.

The simplest form of the command used to create a volume is:

```
# vxassist make vol02 50m
```

The problem with this simple format is that it assumes the following:

- The disk group is `rootdg`.
- The volume type is a simple concatenation with no log.
- It can use any disk drives that have available space.

Without options, the `vxassist` command will probably not give you what you need and can create serious performance issues.

Volume Creation Using the Command Line

The vxassist Command Format

Using vxassist Command Options

If you furnish even a few options with the `vxassist` command, the outcome is more clearly defined. A typical command using limited options is:

```
# vxassist -g dg2 make newvol 2000m layout=raid5,nolog disk01 \  
disk02 disk03
```

This form of the `vxassist` command is more explicit and guarantees that the following will be true:

- The disk group that will be used is `dg2`.
- The name of the volume will be `newvol`.
- The amount of available data storage will be 2 Gbytes.
- This will be a RAID-5 volume without a log and with three columns.
- All disk space will come from `disk01`, `disk02`, and `disk03`.

Other examples of using the `vxassist` command are:

```
# vxassist -g dg3 make newvol 20m layout=stripe disk01 \  
disk02 disk03  
# vxassist -g dg3 make newvol 20m layout=stripe stripeunit=32k \  
disk01 disk02 disk03
```




Sun Educational Services

Adding a UFS File System

A UFS file system can be added to an existing volume using either of the following methods:

- The VMSA New File System form
- Solaris operating system commands

Adding a UFS File System

Adding a UFS file system to a volume is very easy. You can fill out a simple form in VMSA or create one from the command line using standard Solaris OS commands. Both methods will be discussed in this section.

Note – Both UFS and VxFS are supported by SSVM. This course does not cover VxFS; however, Appendix D contains some information on this product. For detailed information, refer to the *Veritas File System Administrator's Guide*.

Adding a UFS File System

Using the VMSA New File System Form

Adding a new file system to an existing volume is very simple if you use the VMSA New File System form. The form, as shown in Figure 6-6, contains all of the information necessary to proceed.

Figure 6-5 VMSA New File System Form

If you previously selected the volume in the VMSA Grid area, the New File System form automatically displays the volume name and the proposed mount point. You can change any of this information. You can also enter any valid `mkfs` options with the `Mkfs Details` button.

If you select the `Mount at Boot` option, the following operations are performed automatically:

- The mount information is recorded in the `/etc/vfstab` file.
- The file system is initialized using the `mkfs` command.
- The mount point is created.
- The finished file system is mounted using the `mount` command.

Adding a UFS File System

Adding a File System From the Command Line

When a new file system is initialized from the command line, three important file system parameters can be adjusted to make more efficient use of available space. They are:

- File system free space
- Number of bytes per inode
- File system cluster size

File System Free Space (minfree)

`minfree` is the amount of file system which is deliberately left unused during initialization. It can act as an emergency overflow.

Prior to the Solaris 2.6 OS, the `minfree` value defaulted to 10 percent. With the Solaris 2.6 OS, the default parameter of `newfs` was changed. Using `newfs`, `minfree` is calculated based on the size of the file system. It scales by $(64 \text{ Mbytes} \div \text{partition size}) \times 100$, rounded down to the nearest integer. It is limited to between 1 percent and 10 percent.

Since `mkfs` still has the `minfree` default at 10 percent, it is much more efficient to create a file system using `newfs`, particularly for large file systems. In very large file systems you can safely set `minfree` to 1 percent.

Bytes per Inode

The default bytes per inode is 2048 (2 Kbytes). Unless the file system consists of many small files, this can be safely increased to 8192 (8 Kbytes). For example:

```
# newfs -i 8192 /dev/vx/rdisk/rootdg/vol01
```

Adding a UFS File System

Adding a File System From the Command Line

File System Cluster Size

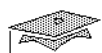
You can set the `maxcontig` parameter for a file system to control the file system I/O cluster size. This parameter specifies the number of 8-Kbyte blocks that will be clustered together on a write to the disk. The default is 7, which equals 56 Kbytes.

To optimize for sequential performance, set `maxcontig` to (number of spindles in the stripe $\times$ the stripe unit size) $\div$ the file system block size (8 Kbytes). If you are optimizing for random performance, set it to 1.

Performance may be improved if the file system I/O cluster size is some integral of the stripe width. For example, setting the `maxcontig` parameter to 16 results in 128-Kbyte clusters (16 blocks $\times$ 8-Kbyte file system block size). For best sequential access, the file system cluster size should match some integer multiple of the stripe width as follows:

- Four disks in stripe and stripe unit size = 32 Kbytes
(32-Kbyte stripe unit size $\times$ 4 disks = 128-Kbyte stripe width)
- `maxcontig` = 16
(16 $\times$ 8-Kbyte blocks = 128-Kbyte clusters)

Note – The VMSA New File System form has a Mkfs Details button that enables you to configure any valid `mkfs` option. You can also set mount options such as ownership, protection, `suid`, `rw`, `ro`, and `largefiles`.



Sun Educational Services

Dirty Region Logging

A Dirty Region Log (DRL) keeps a map of data regions that have been modified in mirrored volumes and provides the following advantages:

- Highly accelerated mirror resynchronization after a crash
- Protection against a second crash during mirror resynchronization

Dirty Region Logging

Dirty region logging (DRL) is a SSVM log file that tracks data changes made to mirrored volumes. The DRL is used to speed recovery time when a failed mirror needs to be synchronized with a surviving mirror.

DRL Overview

A DRL is a small, special-purpose plex attached to a mirrored volume which has the following features:

- It is a log which keeps track of the regions within volumes that have changed as a result of writes to a plex by maintaining a bitmap and storing this information in a log subdisk.
- After a system failure, only the regions marked as dirty in the dirty region log will be recovered.

Dirty Region Logging

DRL Space Requirements

A DRL has a single recovery map and an active map for the host system.

The log size is one block per map for every 2 Gbytes of volume size. For a 2-Gbyte volume, the DRL would be 2 blocks in size.

For a 10-Gbyte volume, the DRL log size would be 10 blocks.

Note – The maximum DRL size is 5 Kbytes. For larger volumes, SSVM changes the log granularity to accommodate the larger volume.



Sun Educational Services

RAID-5 Logging

A RAID-5 log tracks the most recent data or parity information written to disk and provides the following advantage:

- Prevents post-crash recovery corruption

It does not speed up the recovery process as does the dirty region log used with mirrored volumes.

RAID-5 Logging

RAID-5 logs help prevent data corruption in case of a system crash mid-write. Without logging, if a system fails during a write, there is no way to tell if the data and parity were both written to disk. This could result in corrupted data.

RAID-5 Log Overview

When RAID-5 logging is used, a copy of the data and parity are written to the RAID-5 log before being written to disk.

RAID-5 logging is optional, but is strongly recommended to prevent data corruption in the event of a system panic or reboot.

By default, RAID-5 logs are created. You should always run a system with RAID-5 logs to ensure data integrity.

RAID-5 Logging

RAID-5 Log Space Requirements

A RAID-5 log is displayed as a second plex in the output of the `vxprint` command. The length of the log is 2109 blocks or slightly over 1 Mbyte. For example:

vxprint apps-v1

v	apps-v1	raid5	ENABLED	1024032	-	ACTIVE
pl	apps-v1-01	apps-v1	ENABLED	1025088	-	ACTIVE
sd	apps-d01-01	apps-v1-01	ENABLED	341715	0	-
sd	apps-d03-01	apps-v1-01	ENABLED	341715	0	-
sd	apps-d02-01	apps-v1-01	ENABLED	341715	0	-
pl	apps-v1-02	apps-v1	ENABLED	2109	-	LOG
sd	apps-d05-01	apps-v1-02	ENABLED	2109	0	-

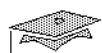
The size of RAID-5 logs is automatically set by SSVM. It is dependent on the stripe width of the volume. The larger the stripe width (not volume), the larger the RAID-5 log. It is intended to hold several full-stripe writes simultaneously.

The default log size for a RAID-5 volume is four times the full stripe width (the stripe unit size $\times$ the number of stripe columns).

The default stripe unit size for RAID-5 volumes is 16 Kbytes (16,384 bytes).

Therefore, the log size for a RAID-5 volume with six disks would be calculated as follows:

$$4 \times 6 \times 16.384 \text{ Kbytes} = 393.22 \text{ Kbytes} = 768 \text{ blocks}$$



Sun Educational Services

Log Placement

If not properly placed, log volumes can create performance problems. The most important rule about log volumes is:

- Log volumes should not reside on the same disk as the volumes they manage.

Log Placement

Logs can be very beneficial to volume recovery after a system crash. If they are not properly managed, however, they can create I/O bottlenecks that negatively impact system performance.

Logs for both RAID-5 and mirrored volumes should be planned for in advance. They do not take much space but can cause problems.

Special care must be taken with RAID-5 logs because all data written to all RAID-5 stripe-units must also be written to the log. This is discussed in more detail in a later module.

Log Placement

Planning for Logs

As shown in Figure 6-6, leaving a small amount of free space at the end of all disks ensures you will always have alternate locations to move logs.

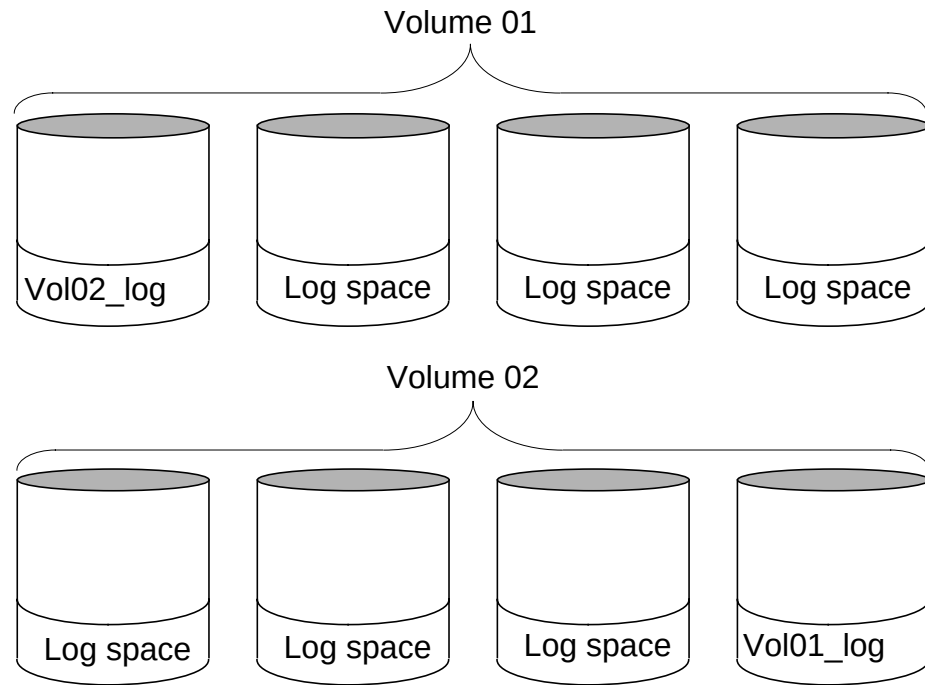


Figure 6-6 Log Space Allocation

If possible, a log should not reside on the same disks as its related volume.

Exercise: Creating a Volume and a File System



Exercise objective – In this exercise you will:

- Create simple, striped, and RAID 5 volumes
- Remove a volume
- Add a mirror to a volume
- Remove a mirror from a volume
- Resize a volume (make it larger)
- Display properties of a volume
- Display volume mapping
- Add a file system to a volume
- Add a dirty region log to a mirrored volume
- Add a log to a RAID-5 volume

Preparation

Ask your instructor for the following information:

- A set of unique names for the volumes you will be creating during this exercise

Concat/mirror volume name: _____

RAID-5 volume name: _____

- A unique mount point name for the file system you will be creating during this exercise

Mount point name: _____

Exercise: Creating a Volume and a File System

Task – Creating a Simple Concatenation

Complete the following steps:

1. If necessary, undock the Toolbar and the Command Launcher (use Options►Customize►Main Window).
2. In the Main window, display the disks in your first disk group (DGa) and select only one of the disks.
3. Select New in the tool bar. A New Volume form will be displayed.
4. Configure the New Volume form as follows:
 - ▼ Enter your assigned concatenated volume name
 - ▼ Set Layout to concatenated
 - ▼ Do not select Mirrored or Add File System
 - ▼ Select Maxsize
5. Select OK when done.
6. Check the status of the new volume with the vxprint command.

```
devsys1# vxprint -g DGa
```

TY	NAME	ASSOC	KSTATE	LENGTH	PLOFFS	STATE
dg	DGa	DGa	-	-	-	-
dm	dga01	c2t35d0s2	-	17678493	-	-
dm	dga02	c3t18d0s2	-	17678493	-	-
v	vol01	fsgen	ENABLED	17678336	-	ACTIVE
pl	vol01-01	vol01	ENABLED	17678493	-	ACTIVE
sd	dga01-01	vol01-01	ENABLED	17678493	0	-

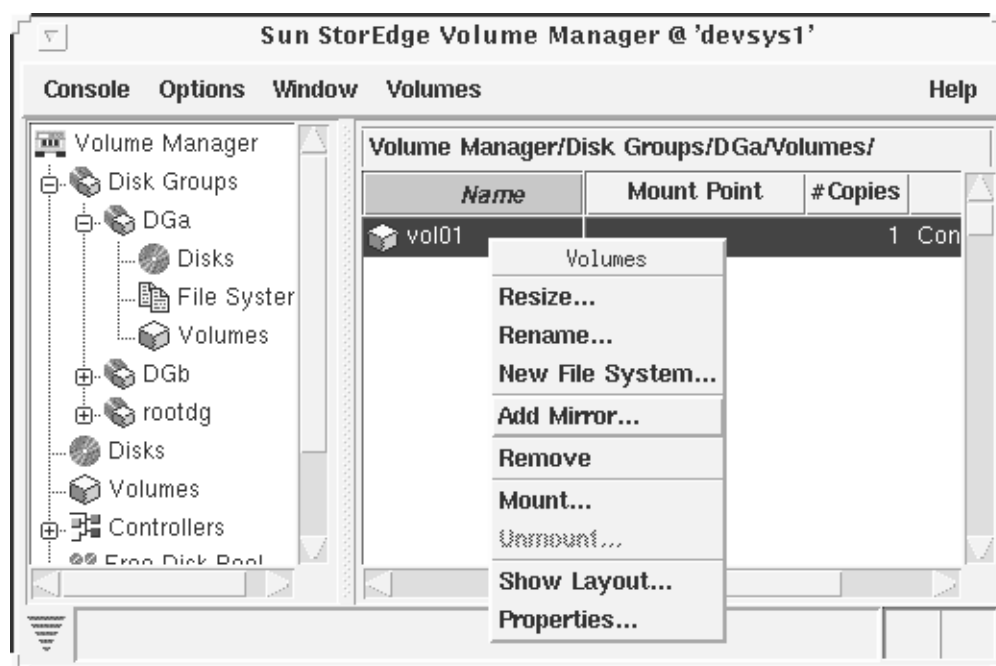
7. Verify that your new volume has a single plex with one subdisk and that the volume and plex are ENABLED and ACTIVE.
8. Check that your new volume is displayed in the Main window.
9. Select the new volume and use Control-p (or the pop-up menu) to view the volume properties.
10. Cancel the volume properties.

Exercise: Creating a Volume and a File System

Task – Adding a Mirror

Complete these steps:

1. Select the volume again if necessary and select Add Mirror in the pop-up menu.



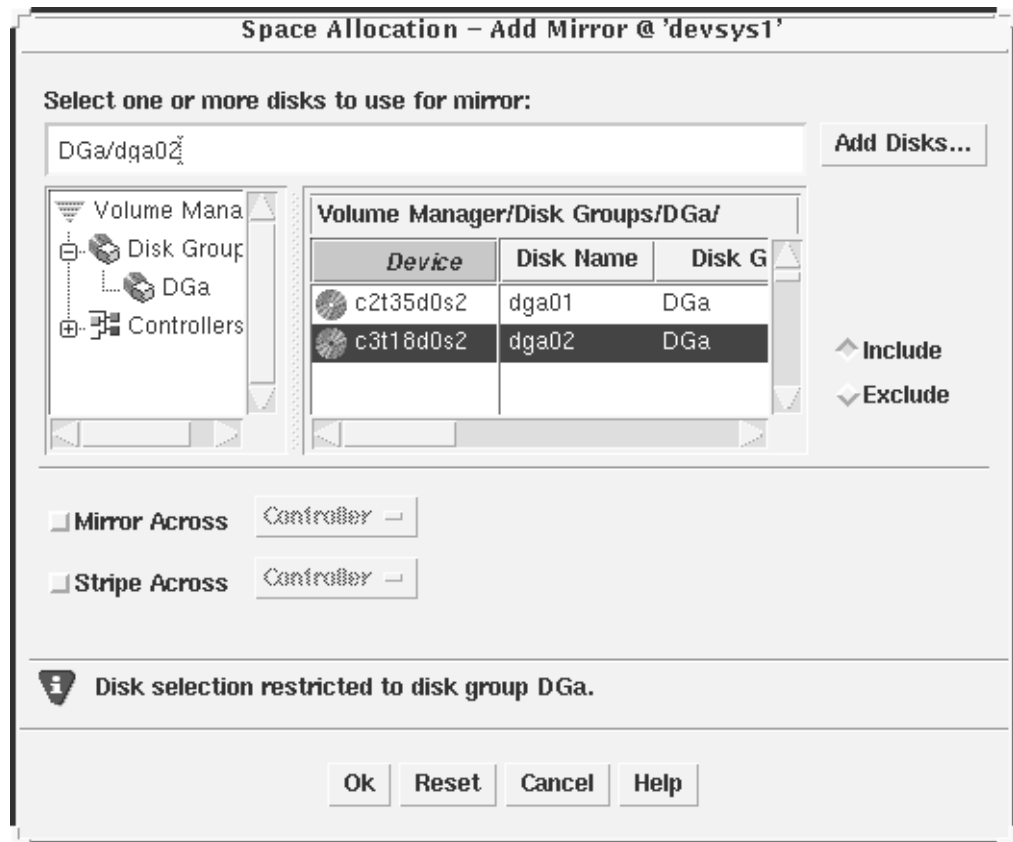
2. Select Assign Disks in the Add Mirror form.



Exercise: Creating a Volume and a File System

Task – Adding a Mirror

3. Highlight the second disk in your disk group. Select OK in the Space Allocation form.



4. Select OK in the Add Mirror form.
5. Select Yes in the Add Mirror warning message.



Exercise: Creating a Volume and a File System

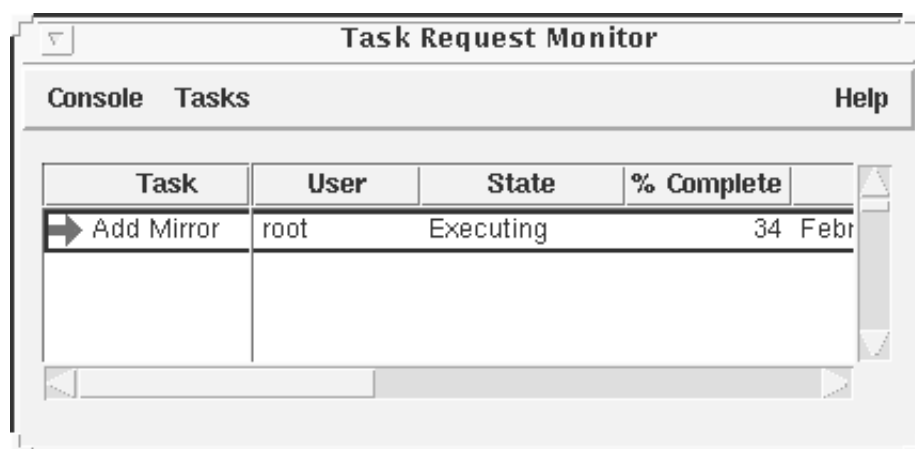
Task – Adding a Mirror

6. On the SSVM server, verify the state of your new mirror with the `vxprint` command.

You should now see two plexes in your volume. If the mirror you configured is very large, it will take a while to synchronize it with the existing mirror.

Note – Until the resynchronization is complete, the related plex will be in a `TEMPRMSD` state.

7. View the Task Request Monitor window, you will probably find the command is still executing. A new mirror can take a long time to synchronize when it is first created.



8. On the SSVM server, use the `man vxinfo` command and find out what the state `TEMPRMSD` means when applied to a plex.

Note – You might find the explanation quite misleading. Discuss this with your instructor.

Exercise: Creating a Volume and a File System

Task – Creating a RAID-5 Volume

Complete the following steps:

1. On the SSVM server, calculate the available disk space for the three disks in your second disk group for building a RAID-5 volume with no log.

```
# vxassist -g DGb maxsize layout=raid5,nolog dgb01 dgb02 dgb03  
Maximum volume size: 35356672 (17264Mb)
```

2. In the Main window, display the disks in your second disk group and select all three of them with the Control key and the left mouse button.
3. Select New in the tool bar.
4. Configure the New Volume form as follows:
 - ▼ Enter your assigned RAID-5 volume name
 - ▼ Set Layout to RAID-5
 - ▼ Disable Logging
 - ▼ Leave the default Stripe Unit Size (32)
 - ▼ *Do not* select Add File System
 - ▼ Enter 3m in the Size field (3 Mbytes)
5. Select OK when done.

Exercise: Creating a Volume and a File System

Task – Creating a RAID-5 Volume

6. On the SSVM server, check the status of the new RAID-5 volume with the vxprint command.

```
# vxprint -g DGb
```

TY	NAME	ASSOC	KSTATE	LENGTH	PLOFFS	STATE
dg	DGb	DGb	-	-	-	-
dm	dgb01	c3t1d0s2	-	17678493	-	-
dm	dgb02	c2t50d0s2	-	17678493	-	-
dm	dgb03	c2t37d0s2	-	17678493	-	-
v	raid5vol	raid5	ENABLED	6144	-	ACTIVE
pl	raid5vol-01	raid5vol	ENABLED	7168	-	ACTIVE
sd	dgb03-01	raid5vol-01	ENABLED	3591	0	-
sd	dgb02-01	raid5vol-01	ENABLED	3591	0	-
sd	dgb01-01	raid5vol-01	ENABLED	3591	0	-

7. In the Main window, display the Volumes. You should see your new volume.
8. Select the volume and use Control-p to view the volume properties.
9. Cancel the volume properties.
10. Highlight the RAID volume task in the Task Request Monitor window and view the properties with Control-p.

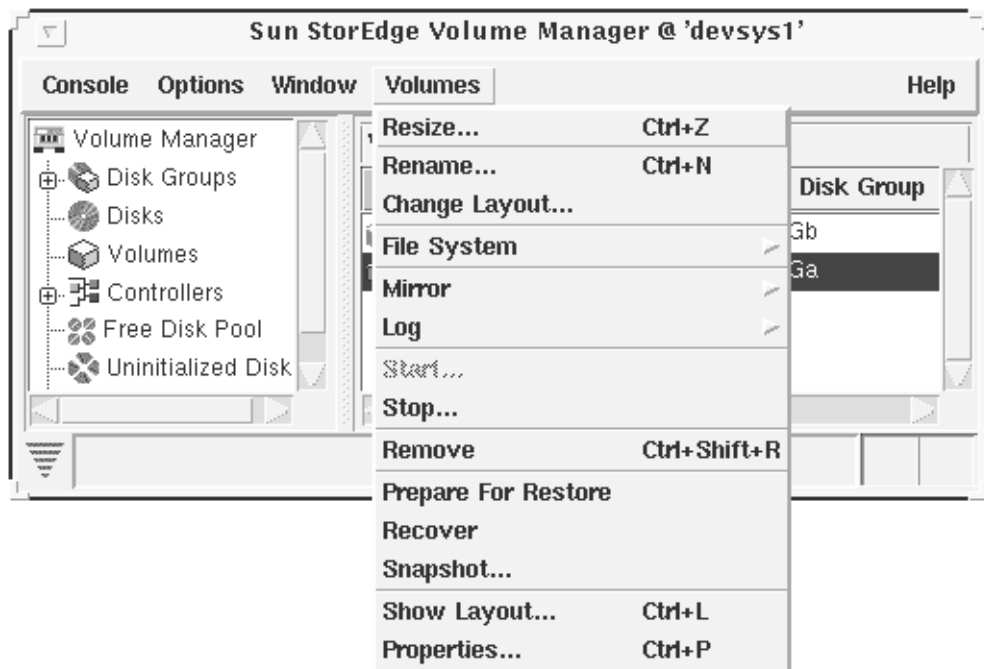
Note – The vxassist command line is quite long but it demonstrates how you can perform very complex tasks from the command line.

Exercise: Creating a Volume and a File System

Task – Displaying Volume Layout Details

Complete the following:

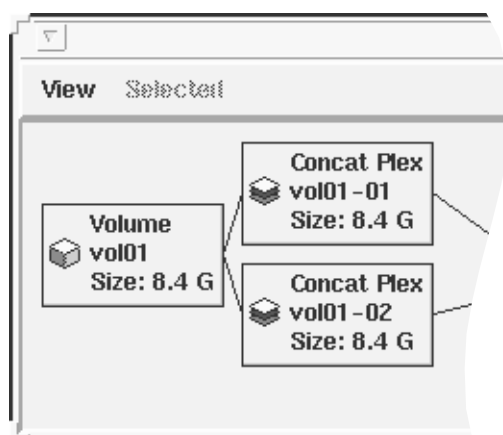
1. Highlight your mirrored volume in the Grid window and select the Show Layout entry from the pop-up window.



Exercise: Creating a Volume and a File System

Task – Displaying Volume Layout Details

2. Look at the menus that are available on the different volume components.
3. Click on Volume box.
4. Click on a RAID level box.
5. Click on a specific subdisk.



Volume "vol01" Layout Details @ 'devsys1'

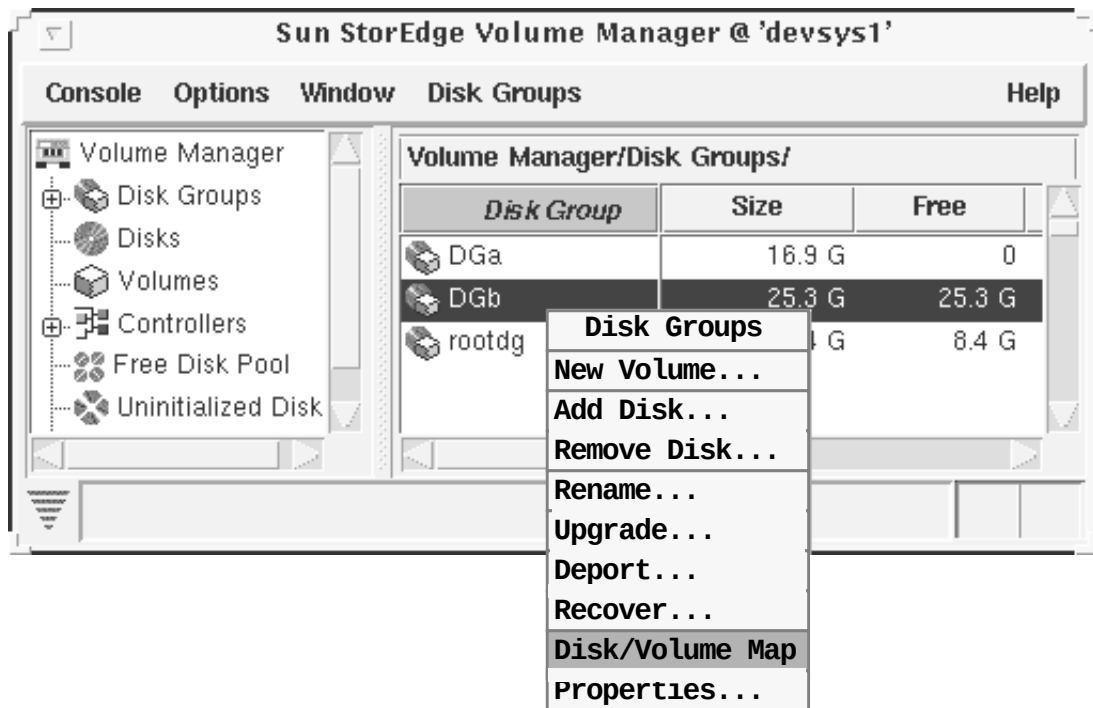
Subdisk	Disk	Plex	Length	Offset	Column	Start
dga01-01	dga01	vol01-01	8.4 G	0		0
dga02-01	dga02	vol01-02	8.4 G	0		0

Exercise: Creating a Volume and a File System

Task – Performing Volume to Disk Mapping

Use the following steps:

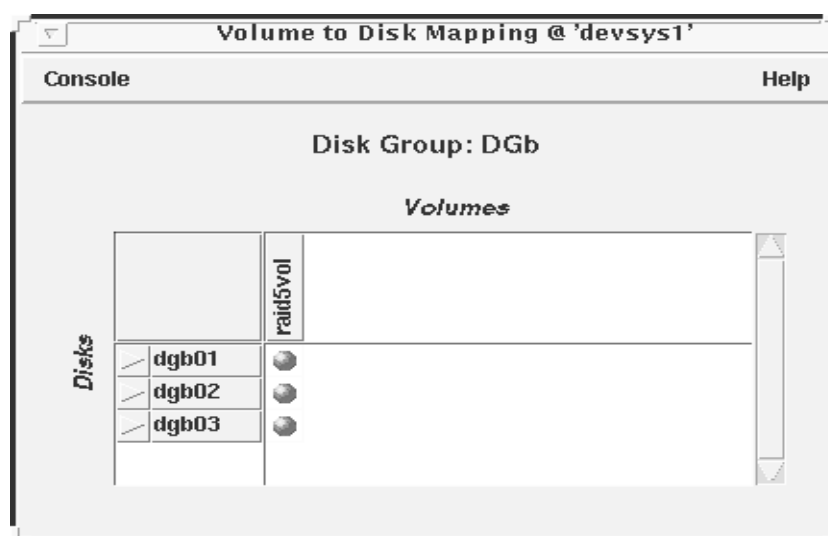
1. Display the Disk Groups in the Grid window. Click on your second disk group. Select the Disk / Volume Map entry in the pop-up menu.



Exercise: Creating a Volume and a File System

Task – Performing Volume to Disk Mapping

You should see the map of disks to volume names.



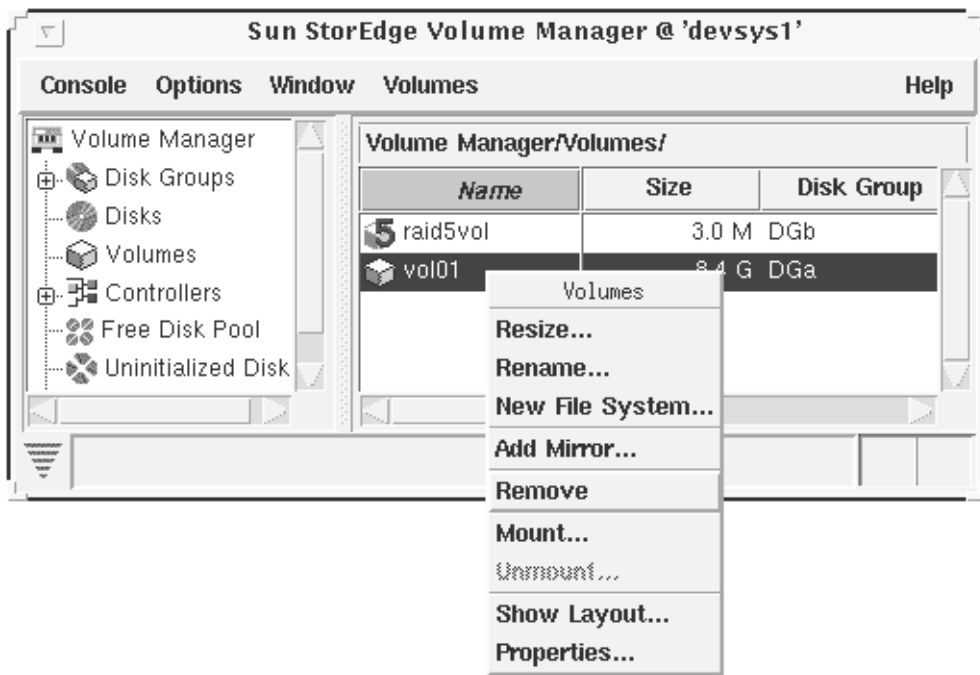
2. Close the Volume map.

Exercise: Creating a Volume and a File System

Task – Removing a Volume

Complete these steps:

1. Display all volumes in the Grid window.
2. Using the pop-up menu, remove the mirrored volume.



3. Rebuild the concatenated / mirrored volume to be the same as it was but reduce the size.
4. In the Main window, display the disks in your first disk group (DGa) and select two of them.
5. Select New in the tool bar. A New Volume form will be displayed.

Exercise: Creating a Volume and a File System

Task – Removing a Volume

6. Configure the New Volume form as follows:
 - ▼ Enter your assigned Concatenated volume name
 - ▼ Enter 3m in the Size field
 - ▼ Set Layout to Concatenated
 - ▼ Select Mirrored (2 mirrors)
 - ▼ Disable logging
 - ▼ *Do not* select Add File System
7. Select OK when done.

Exercise: Creating a Volume and a File System

Task – Adding a File System

Use the following steps:

1. Select your mirrored volume in the Grid area and select NewFile System from the pop-up menu.

The screenshot shows a window titled "New File System @ 'devsys1'". Inside the window, there are two text input fields: "Volume Name" containing "DGa/newvola" and "Mount Point" containing "/newvola". To the right of the "Volume Name" field is a "Browse..." button. Below the "Mount Point" field is a checkbox labeled "Mount at Boot" which is checked. On the right side of the window, there is a section titled "FS Type" with two radio buttons: "ufs" (which is selected) and "vxfs". Below the "FS Type" section are two buttons: "Mount Details..." and "Mkfs Details...". At the bottom of the window, there is a row of five buttons: "Ok", "Apply", "Reset", "Cancel", and "Help".

2. Configure the New File System form as follows:

- ▼ Enter your assigned mount name
- ▼ Make sure that the file system type is set to ufs
- ▼ Make sure Mount at Boot is selected

Review the Mount Details and Mkfs Details information but do not change anything.

3. Select OK when done.

Exercise: Creating a Volume and a File System

Task – Adding a File System

4. On the SSVM server, verify that the following are true:
 - ▼ The mount point is present in root.
 - ▼ Your file system is mounted.
 - ▼ The mount entry is in the `/etc/vfstab` file.
 - ▼ The `df -k` output seems appropriate.
5. Copy some test data into the volume.

```
# cp -r /usr/kernel /mount_point (about 750 kbytes)
```

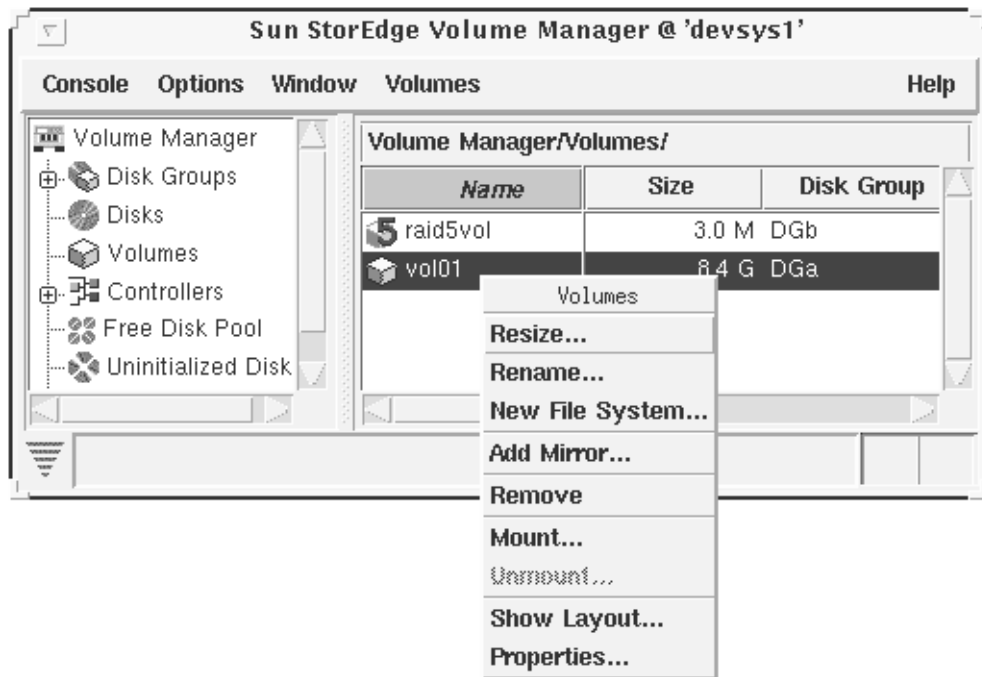
Exercise: Creating a Volume and a File System

Task – Resizing a Volume or File System

If a volume has a file system, you can resize both at the same time by selecting either Volume ► Resize or FileSystem ► Resize. You can select these in several place including the Command Launcher.

Note – You cannot shrink a volume with a file system unless the file system is a VxFS type.

1. Display your volumes in the Grid area and choose Volume Resize from the pop-up menu.



2. Use the Browse button in the Resize Volume form to select your mirrored volume.
3. Enter 2m in the Add By field.
4. Select OK when ready to begin.
5. After the task has completed, verify the results.

Exercise: Creating a Volume and a File System

Task – Adding a Dirty Region Log

In this section you will add a DRL to your mirrored volume.

Note – Prior to adding this DRL, you must ensure there is a disk available for the DRL within the same disk group.

1. Display your volumes in the VMSA Grid area.
2. Highlight the mirrored volume.
3. Select Log in the Volume (Selected) pull-down menu and click on Log➤Add.
4. In the Add Log window, either enter the disk name or click on Browse to select a disk using the GUI.
5. After clicking on OK in the Add Log window, return to the command line on the SSVM server and use `vxprint` to verify the following:
 - ▼ The mirrored volume now has a log plex.
 - ▼ The log is not on the same disks as the mirrored volume.

Note – Look at the subdisk entries to determine log placement.

6. Delete the logs from your mirrored volume as follows:
 - a. Highlight the mirror volume.
 - b. Using the Volume (Selected) pull-down menu, select Log➤Remove.
 - c. Identify either the log disk or log name.
 - d. Click on OK.
7. Use the SSVM `vxprint` command to verify the log has been removed.

Exercise: Creating a Volume and a File System

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences
- Interpretations
- Conclusions
- Applications

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Create simple, striped, and RAID-5 volumes
- ☐ Remove a volume
- ☐ Add a mirror to a volume
- ☐ Remove a mirror from a volume
- ☐ Resize a volume (make it larger)
- ☐ Display properties of a volume
- ☐ Display volume mapping
- ☐ Add a file system to a volume
- ☐ Add a dirty region log to a mirrored volume
- ☐ Add a log to a volume

Think Beyond

What methods can be used to correct existing volume configuration errors without having to destroy and rebuild the volumes?

Objectives

Upon completion of this module, you should be able to:

- Move an empty disk to a different disk group
- Move a populated disk to a new disk group
- Perform a snapshot backup
- Move a disk group between systems
- Assign and remove hot spares
- Enable and disable hot relocation
- Create a striped pro volume with a file system

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

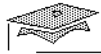
- Are there special SSVM features to assist in making file system backups?
- Are there any volume structures that can provide an unusually high level of reliability?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The RAID Advisory Board 1996. *The RAID Book* Lino Lakes, MN.
- “Sun Performance Tuning Overview,” December, 1993, part number #801-4872-07.
- Wong, Brian. *Configuration and Capacity Planning for Solaris Servers* ISBN 0-13-349952-9.
- Chen, Lee, Gibson, Katz, and Patterson October, 1993. *RAID: High Performance, Reliable Secondary Storage*.



Evacuating a Disk

If a disk begins to fail, its contents can be easily moved to another disk providing the following is true:

- The new disk is in the same disk group
- The new disk does not contain conflicting objects
- The new disk has sufficient free space

Evacuating a Disk

The volume structures on a disk drive that is starting to experience recoverable data errors can be evacuated to a different disk before the disk fails entirely. This can reduce the risk of data loss by minimizing the time a volume might be operating without a mirror.

This procedure can also be used to reduce or eliminate performance bottlenecks that have been identified.

Evacuation can only be performed on disks within the same group.

Evacuating a Disk

Evacuation Conflicts

Before you proceed with disk evacuation, carefully investigate the configuration of both the failing and the new disk drive.

Verify that the evacuation process is not going to create any of the following conflicts:

- Both volume mirrors on the same physical disk drive.
- More than one stripe column of a striped or RAID-5 volume is on the same disk drive.

Evacuation Preparation

Before starting the evacuation process, you must:

- Find out what volume the failing plex is associated with and the name of the disks that are associated with it.
- Find out the disk group associated with the failing disk drive.
- Determine if any other volumes are associated with the failing disk drive.
- Find a new disk with enough free space to perform the evacuation.
- Check for any volume conflicts associated with the new disk.

Evacuating a Disk

Evacuation Preparation

The following example illustrates how you would prepare for disk evacuation if the failing plex was named `plex002`:

```
# vxprint -ht |grep plex002
pl plex002 vol002  ENABLED  ACTIVE  2048256  CONCAT
sd sd01      plex002  disk01   0        2048256  0 c1t1d0

# vxdisk list |grep disk01
c1t1d0s2      sliced    disk01    skyvols  online

# vxprint -g skyvols -ht |grep disk01
dm disk01    c1t1d0s2  sliced    2015    2050272  -
sd sd01      plex002  disk01    0        2048256  0 c1t1d0

# vxdg -g skyvols free
DISK      DEVICE      TAG      OFFSET    LENGTH    FLAGS
disk03    c1t0d3s2    c1t0d3    2048256    2016      -
disk04    c1t0d4s2    c1t0d4     0         2048256    -
disk01    c1t1d0s2    c1t1d0    2048256    2016      -
disk02    c1t1d1s2    c1t1d1    2048256    2016      -

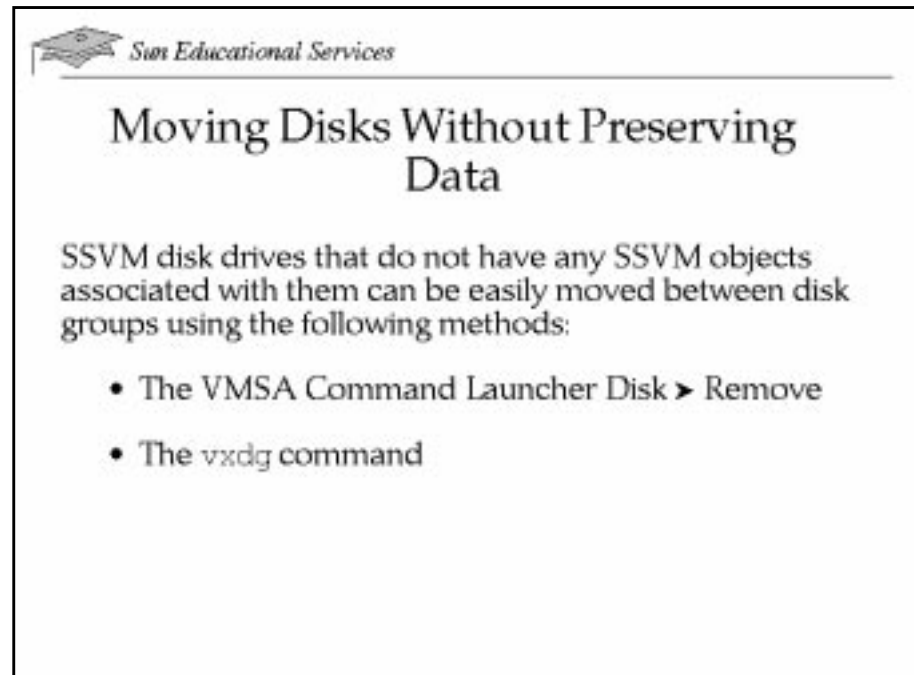
# vxprint -g skyvols -ht |grep disk04
```

Performing an Evacuation

The evacuation process can be performed from the VMSA application as follows:

1. Select the disk that contains the objects and data to be moved.
2. Choose Disks ► Evacuate from the Selected pop-up menu.
3. Enter the destination disk in the Evacuate Disk dialogue box.

Note – You can also use `vxdiskadm` option 7 or the `vxevac` command directly to perform a disk evacuation.



Moving Disks Without Preserving Data

You might want to move a SSVM disk to a different disk group because the destination disk group needs the additional disk space. As long as the disk does not contain any functional data that you need to preserve, the process is fairly simple.

Moving a Disk Using the Command Line

If the disk you want to move contains an active volume and you *do not care if the data is lost*, you must:

1. Unmount a related file system and disable any related applications.
2. Stop the volume.
3. Delete the volume configuration.
4. Remove the disk from the disk group.
5. Add the disk to a different disk group.

Moving Disks Without Preserving Data

Moving a Disk Using the Command Line

Stopping the Volume

You can stop volumes using the `vxvol` command as follows:

```
# vxvol stop volume_name
```

Delete the Volume Configuration

You can recursively delete all objects in a volume as follows:

```
# vxedit -r rm volume_name
```

Remove the Disk From the Disk Group

You can use the `vx dg` command to remove a disk from a disk group as follows:

```
# vx dg rmdisk disk_name
```

Note – Even after the `vx dg rmdisk` operation, the disk will still be initialized for SSVM use. Only the `vx diskunsetup` command can completely remove a disk from SSVM control.

Add the Disk to a New Disk Group

The `vx dg` command is used to add the disk to a different disk group as follows:

```
# vx dg -g new_dg adddisk new02=c1t3d0
```

Note – In the previous steps, the disk group by default was `rootdg`. You must be specific about the new disk group, the new disk name, and the physical path to the disk drive.

Moving Disks Without Preserving Data

Moving a Disk From VMSA

Moving a disk drive to a new disk group is easy using the VMSA interface. After selecting the volume you want to remove, use the Stop and Remove menu entries as shown in Figure 7-1.

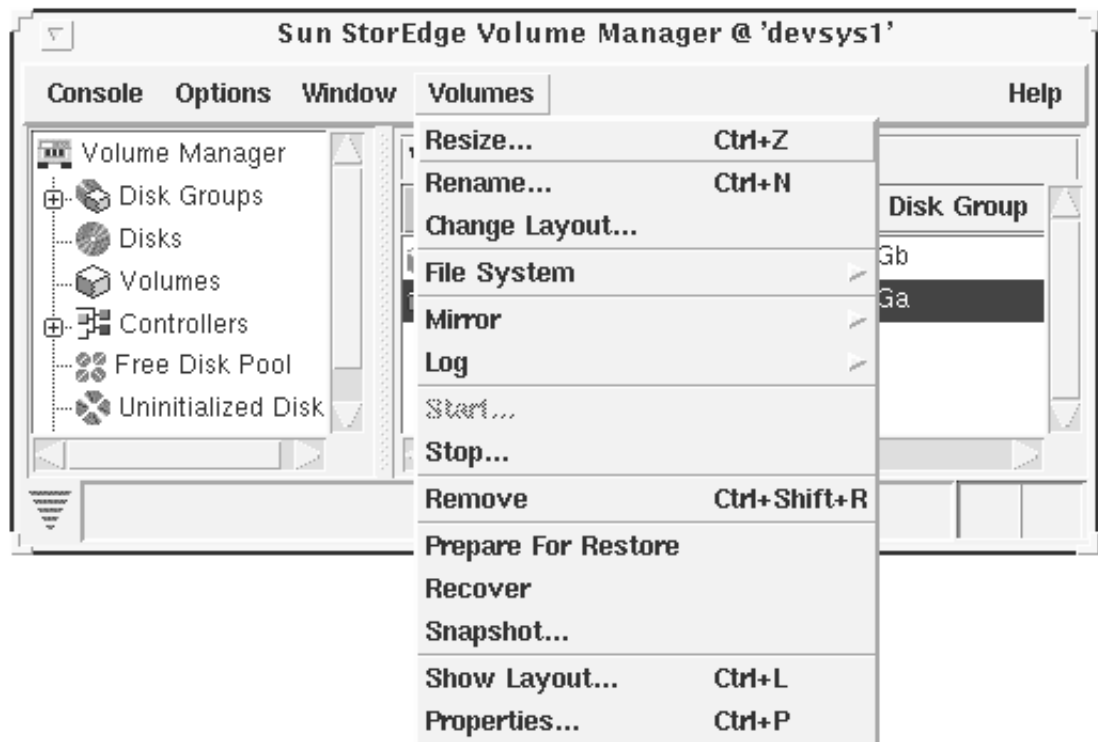


Figure 7-1 VMSA Volume Removal

The disk is now returned to the Free Disk Pool and can be selected in the Free Disk Pool display and then added to a different disk group using the Disk ➤ Add function in the Command Launcher window.

 *Sun Educational Services*

Moving Populated Disks to a New Disk Group

Moving a disk to a new disk group and preserving its data is complex and involves the following operations:

- Saving the disk configuration with `vxprint`
- Stopping the volume with `vxvol`
- Removing the volume with `vxedit`
- Moving the disk to the new group with `vx dg`
- Rebuilding the original configuration with `vxmake`

Moving Populated Disks to a New Disk Group

Moving populated SVM disks to a new or different disk group is a technique which may be used occasionally. One reason to use this technique is if you have mistakenly created all of your volumes in the `rootdg` disk group and now want to correct the mistake.

It is important to understand the concepts since many of the commands used can be used for other purposes, such as recovering a configuration.



Caution – This operation should not be done on a production system without first backing up data on all volumes. If this process fails, there is no way to recover without backup tapes.

Note – In this section, a volume called `vol01` is going to be moved from a disk group named `olddg` to a new disk group named `newdg`.

Moving Populated Disks to a New Disk Group

Determining Which Disks Are Involved

Before you take any action, you must determine which physical disks are part of your target volume. You must also make sure that the disks are not being used by other volumes.

If you use `vxprint` command with a `-ht` option, you will see a complete volume hierarchy displayed.

```
# vxprint -ht -g olddg
```

```
Disk group: olddg
```

DG	NAME	NCONFIG	NLOG	MINORS	GROUP-ID		
DM	NAME	DEVICE	TYPE	PRIVLEN	PUBLEN		STATE
V	NAME	USETYPE	KSTATE	STATE	LENGTH	READPOL	PREFPLEX
PL	NAME	VOLUME	KSTATE	STATE	LENGTH	LAYOUT	NCOL/WID
SD	NAME	PLEX	DISK	DISKOFFS	LENGTH	[COL/]OFF	DEVICE
dg	olddg	default	default	0	891019192.1025.baw	lmer	
dm	olddg01	c0t17d0s2	sliced	1519	4152640	-	
dm	olddg02	c0t18d0s2	sliced	1519	4152640	-	
dm	olddg03	c0t19d0s2	sliced	1519	4152640	-	
v	vol01	fsgen	ENABLED	ACTIVE	10240	SELECT	vol01-01
pl	vol01-01	vol01	ENABLED	ACTIVE	11015	STRIPE	3/128
sd	olddg01-01	vol01-01	olddg01	0	3591	1/0	c0t17d0
sd	olddg02-01	vol01-01	olddg02	0	3591	1/0	c0t18d0
sd	olddg03-01	vol01-01	olddg03	0	3591	2/0	c0t19d0

The Volume Hierarchy section lists an entry for the volume, followed by entries for its associated plexes and subdisks. In this example, volume `vol01` contains one plex (`vol01-01`). This plex is comprised of three subdisks (`olddg01-01`, `olddg02-01`, and `olddg03-01`), each of which is stored on a separate SSVM disk.

You can tell from this output that the three SSVM disks that need to be moved are `olddg01`, `olddg02`, and `olddg03`.

Moving Populated Disks to a New Disk Group

Saving the Configuration

To do this you would:

1. Use the `vxprint` command to save the volume configuration.

```
# vxprint -hmQq -g olddg vol01 > save_vol01
```

The `vxprint` command with the `-m` option is used to save the configuration in a format that can be used later by the `vxmake` utility.

In this case, you are saving the configuration for volume `vol01` in the file `save_vol01`.

The options used in this example are:

- h List complete hierarchies
- m Display information in a format that can be used as input to the `vxmake` utility
- Q Suppress the disk group header that separates each disk group
- q Suppress headers (in addition to disk group header)
- g Specify the disk group

Moving the Disks to a New Disk Group

Moving the disks to a new disk group requires several steps that you have seen earlier in this course. They are:

2. Unmount appropriate file systems, and / or stop any processes on the `vol01` volume.
3. Stop the volume.

```
# vxvol -g olddg stop vol01
```

Moving Populated Disks to a New Disk Group

Moving the Disks to a New Disk Group

4. Remove the definitions of the structures (volume, plexes, and subdisks) from the configuration database.

```
# vxedit -g olddg -r rm vol01
```

The `vxedit` command is used to remove the definitions of the volume, plexes, and subdisks from the configuration database for the old disk group, `olddg`.

Note – This does not affect the data; it only removes selected records from the configuration database. The `-r` option will recursively remove the volume and all associated plexes and subdisks.

5. Remove the disks from the original disk group.

```
# vxdg -g olddg rmdisk olddg01 olddg02 olddg03
```

6. If the new disk group, `newdg`, does not exist, initialize it using one of the disks to be moved (disk `olddg01`, in this example).

```
# vxdg init newdg olddg01=c0t17d0s2
```



Caution – The commands `vxdisk init` and `vxdg init` are similar, but perform very different operations: `vxdisk init` initializes a disk, destroying all existing data; `vxdg init` initializes a disk group, adding the specified disk to the new disk group.

Moving Populated Disks to a New Disk Group

Moving the Disks to a New Disk Group

7. Add the remaining disks to the new disk group.

```
# vxdg -g newdg adddisk olddg02=c0t18d0s2
# vxdg -g newdg adddisk olddg03=c0t19d0s2
```

8. Verify that the disks have been added to the new disk group.

```
# vxdisk list | grep newdg
c0t17d0s2    sliced    olddg01    newdg     online
c0t18d0s2    sliced    olddg02    newdg     online
c0t19d0s2    sliced    olddg03    newdg     online
```

Reloading the Volume Configuration

9. Use the `vxmake` command to reload the saved configuration for the volume `vol01`.

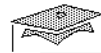
```
# vxmake -g newdg -d save_vol01
```

Recall that earlier the volume configuration was saved in the file `save_vol01`. The `-d` option is used to specify the description file to use for building subdisks, plexes, and volumes.

10. Use the `vxvol` command to bring the volumes back online.

```
# vxvol -g newdg init active vol01
```

Note – An alternative to this procedure is to create a new volume in another disk group and either dump a backup tape onto it or perform a direct copy from the old volume.



Moving Disk Groups

The process of deporting and importing disk groups requires an understanding of the following concepts:

- Why you might want to deport or import a disk group
- How disk groups are identified
- How to establish new disk group ownership
- Restrictions when deporting and importing

Moving Disk Groups

A disk group is associated with a particular host system. Sometimes the administrator might want to associate a disk group with another system. When done under an administrator's control, the process involves first deporting the disk group from one system and then importing it from another host.

The reasons for deporting a disk group are:

- Disable access to all volumes in that disk group
- Prepare for another host to import the disk group

This can be part of a planned maintenance outage or a load-balancing measure.

Note – If a system crash occurs, a disk group can be left in a state that is not cleanly deported. This possibility is discussed in this section.

Moving Disk Groups

Disk Group Ownership

When a disk group is created, it is assigned both a unique name and a unique group identifier. You can see both in the `vxprint` output in the following example:

```
dg    olddg    default    default    0    891019192.1025.bawlmer
```

In this example, the disk group name is `olddg` and the unique disk group ID is `891019192.1025.bawlmer`.

The `hostname` of that system is stored on all disks in the disk group.

Note – The SSVM documentation and many SSVM man pages incorrectly refer to the `hostname` as *hostid* or *host ID*.

Disk Group States

Disk groups that are deported by plan can be deported in several different states:

- The disk group name and identifier are unchanged, the `hostname` is cleared.

This is the typical state after a planned deport. The disk group will be imported again later by the same system.

- The disk group has been given a new name and assigned a new `hostname`.

This might be done to prepare the disk group for importation by a different host system during maintenance or to balance loads.

- The disk group name and identifier are unchanged and the `hostname` has not been cleared.

This is typically the state after a system crash. If a different system tries to import the disk group, there will be `hostname` conflicts and the import will fail unless extra steps are taken.

Moving Disk Groups

Preparation for Deporting a Disk Group

Before disk group is deported, the following actions must be taken:

- Stop all application or user access to file system or database volumes
- Unmount all file systems
- Stop all volumes

Note – If a system crashes, the volumes can be left in a state that requires volume recovery during a later import process.

Deporting Options

There are several variations available when deporting a disk group. The most common ones are:

- Normal deport operation, the `hostname` is cleared automatically.

vxdg deport *disk_group_name*

This is the normal deport. The same system will import the disk group again later.

- Deport the disk group and write a new `hostname` on the disks.

vxdg -h new_hostname deport *disk_group_name*

This might be done to prepare a disk group for importation by another host system. This will allow the second host to auto-import the disk group when it boots.

- Deport the disk group, give it a new name and a new `hostname`.

vxdg -n new_dgname -h new_hostname deport *dg_name*

This prepares the disk group for importation by another host that already has a disk group with the same name.

Moving Disk Groups

Importing Disk Groups

Depending on the state of a disk group, there are several variations of the import operation that might be useful.

- Doing a simple import of a clean disk group

```
# vxdg import disk_group_name
```

- Importing a disk group to another system after a crash

```
# vxdg -C import disk_group_name
```

The -C option is necessary to clear the old hostids that were left on the disks after the crash.

```
# vxdg -fC import disk_group_name
```



Warning – The -f option will force an import in the event all of the disks are not usable. This can be very dangerous and lead to multiple imports on dual-hosted storage arrays.

```
# vxrecover -g disk_group_name -sb
```

This should be done after a crash to start the volumes and perform a recovery process. This is done automatically during a reboot.

- Importing a disk group with a duplicate name

```
# vxdg -t -n new_disk_group import disk_group_name
```

The -t option makes the new disk group name temporary.

Moving Disk Groups

Importing rootdg After a Crash

After a crash it may be necessary to import rootdg to another system to perform repair operations. This is a little more complicated because you cannot have two rootdg disk groups on a system.

This requires the use of multiple options to:

- Assign a new temporary disk group name to rootdg
- Clear the original hostid ownership
- Use the unique rootdg group identifier

```
# vxdg -tC -n new_disk_group import group_id
```

The difficult part is that you must use the unique rootdg group identifier. This must be known in advance. You can determine the rootdg group identifier with the vxdisk command as follows:

```
# vxdisk -s list
Disk:    c0t2d0s2
type:    sliced
flags:   online ready private autoconfig autoimport imported
diskid:  791000525.1055.boulder
dgname:  rootdg
dgid:    791000499.1025.boulder
hostid:  boulder
On the importing host, the group will be renamed:
```




Sun Educational Services

Hot Devices

Automatic replacement of failed disk drives by assigned spares can be accomplished using either:

- An older hot spare daemon, `vxsparecheck`
- A new hot relocation daemon, `vxrelocd`

Hot Devices

Depending on how the `/etc/rc2.d/S95vxvm-recover` file is configured, either the older hot spare daemon, `vxsparecheck`, will start at boot time or the newer hot relocation daemon, `vxrelocd`, will start.

The functionality of the two daemons is very different. By default, hot relocation is enabled.

Hot Spare Overview

If it is enabled at boot time, the hot spare daemon, `vxsparecheck`, detects and reacts to total disk media failures by moving the entire contents of the failed disk to a pre-designated spare disk in the disk group. A disk is considered totally failed if SSVM cannot access one or more subdisks and also cannot access the private region on the disk.

Hot sparing is an older mode of operation but can still be enabled if desired.

Hot Devices

Hot Relocation Overview

The hot relocation daemon, `vxrelocd`, detects and reacts to partial disk media failures by moving the affected subdisk to free space on a different disk in the group. Free space can be found on disks that have been designated as hot relocation spares or SSM can find it randomly in a disk group's free space if there are no designated spares.

Hot relocation can only be performed for subdisks that are part of a redundant volume such as RAID 5 or a mirrored volume.

Hot relocation is enabled by default and goes into effect without system administrator intervention when a failure occurs.

Note – Although it is not advisable, hot relocation can be temporarily disabled by the system administrator at any time by stopping the `vxrelocd` daemon.

As shown in Figure 7-2, when a subdisk failure is detected, the contents of the subdisk are reconstructed on the designated hot spare. The volume continues to function with its original full redundancy.

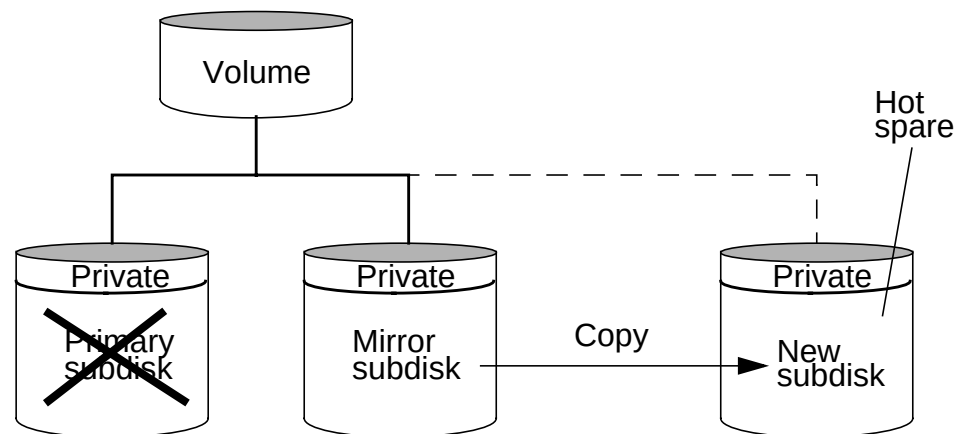


Figure 7-2 Subdisk Relocation

Hot-Relocation

Failed Subdisk Detection

The hot-relocation daemon, `vxrelocd`, detects and reacts to the following types of failures:

- Disk failure

This is first detected as an I/O failure from a SSVM object. An attempt to correct the error is made. If the error cannot be corrected, SSVM tries to access configuration information in the private region of the disk. If it cannot access the private regions, it considers the disk failed.

- Plex failure

This is detected as an uncorrectable I/O error in the plex. For mirrored volumes, the plex is detached.

- RAID-5 subdisk failure

This is detected as an uncorrectable I/O error in one of the RAID-5 subdisks. The subdisk is detached.

Hot-Relocation

Hot-Relocation Failures

If relocation is not possible, the system administrator is notified and no further action is taken. Relocation is not possible if:

- Subdisks do not belong to a mirrored or RAID-5 volume
- Not enough space is available on spare disks or free space
- The only available space for relocation is on a disk that contains any portion of the surviving mirror or RAID-5 volume
- A mirrored volume has a dirty region logging log subdisk as part of its data plex, subdisks belonging to that plex cannot be relocated
- The failure is a log plex, a new log plex is created so it is not actually relocated.

Note – Hot relocation can create a mirror of the boot disk if it is encapsulate and mirrored. There must be a spare in rootdg.

Hot-Relocation

Hot-Relocation Administration

Designating Hot Spare Disks

You can prepare for hot relocation by designating one or more disks per disk group as hot-relocation spares. To designate a disk as a hot spare for a disk group from the command line use:

```
# vxedit -g disk_group set spare=on diskname
```

You can verify the spare status of the disks with the `vxdisk list` command and disable the disks as spares with the `vxedit spare=off` command option.

Controlling Recovery Time

You can reduce the impact of recovery on system performance by instructing `vxrelocd` to increase the delay between the recovery of each region of a volume (`vxrelocd -o slow=500 &`). The value of `slow` is passed on to `vxrecover`. The default value is 250 milliseconds.

Monitoring

By default, the `vxrelocd` daemon sends email notification of errors to the server `root` account. You can modify the account name in the `vxrelocd root &` line in the `etc/rc2.d/S95vxvm-recover` file.

You can also examine system error logs for evidence of disk problems but the email notification to `root` is usually sufficient.

Enabling the Hot-Spare Feature

To enable the older hot-spare feature instead of the hot-relocation feature, edit the `/etc/rc2.d/S95vxvm-recover` file and comment out the `vxrelocd root &` line and uncomment the `#vxsparecheck root &` line.



Snapshot Backups

The SSVM snapshot feature automatically prepares a volume for backup by performing the following operations:

- Attaching a new write-only mirror to the volume
- Copying the data to the new mirror
- Detaching the mirror
- Creating a new volume from the new mirror

Snapshot Operations

When you need to backup the data on a volume such as a file system volume, you can use the SSVM snapshot function to create a copy of the volume. You then backup the new copy to tape without disrupting service to the original volume.

Snapshot Prerequisites

The following prerequisites must be satisfied before the snapshot process can be started:

- You must know the name of the volume to be backed up.
- You must furnish a name for the new snapshot copy.
- You can specify specific disk to use for the snapshot copy.
- You must have sufficient unused disk space for the snapshot.

Snapshot Operations

Snapshot Process

The general process for using the snapshot feature from the VMSA interface is as follows:

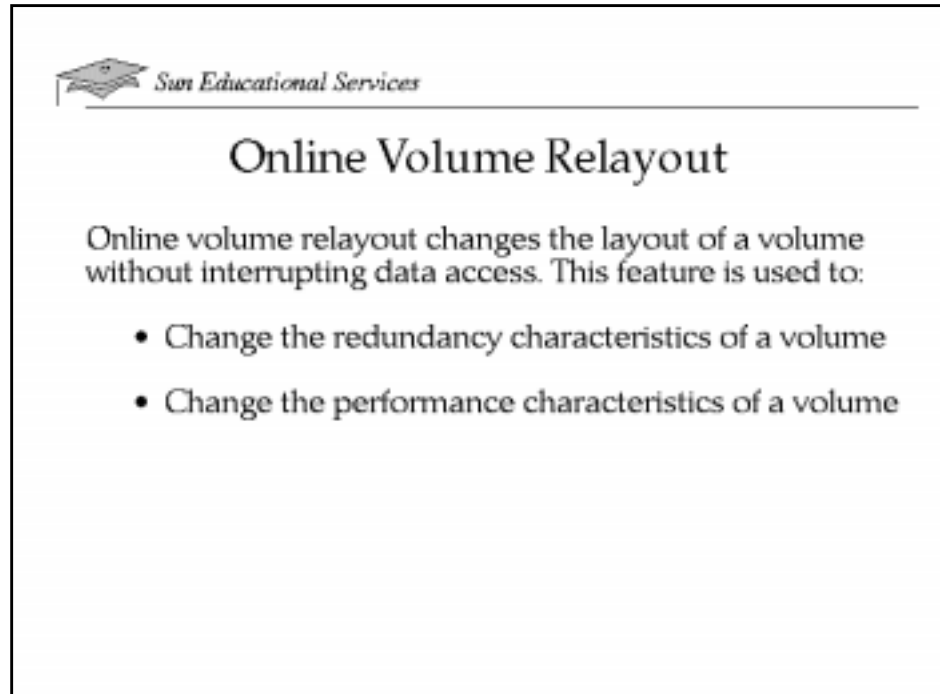
1. Select the volume to be copied to a snapshot.
2. Choose Volumes ► Snapshot (Selected) menu or Volume ► Snap from the Command Launcher.
3. Complete the Volume Snapshot dialogue box.
4. Click Snapshot in the dialogue box to start the snapshot process.

This may take quite a bit of time depending on the volume size.

5. Click Snapshot again when the mirror copy is complete.

This will detach the new mirror and create a separate volume from it.

6. Backup the new snapshot volume to tape.
7. Remove the snapshot volume.



Online Volume Relayout

Online volume relayout provides the administrator with a tool that can be used to correct configuration mistakes or enhance the configuration later when more disk resources might be available.

The relayout feature can be used to perform many operations such as:

- Adding more stripe columns to a RAID-5 volume
- Changing the stripe unit size of a volume
- Changing the type of volume from RAID 5 to mirrored or concatenated

Online Volume Relayout

Volume Relayout Prerequisites

You must provide the following information to the Relayout form before starting:

- Choose the new volume layout.

This includes Concatenated, Striped, RAID-5, Concatenated Pro, and Striped Pro.
- Specify additional disk space to be used for the new volume layout if needed, such as for RAID-5 parity space.
- Specify the temporary disk space to be used during the volume layout change.

Relayout Status Monitor

Once you fill out the Relayout form and start the relayout process, a Relayout Status window will be displayed. You can use the controls in the Relayout Status window to:

- Temporarily stop the relayout process (pause)
- Abort the relayout process
- Continue the process after a pause
- Undo the relayout changes (Reverse)

The percentage complete status is also displayed.

Note – The relayout task could fail if volumes are not created by VMSA or the vxassist command.



Sun Educational Services

Layered Volumes

SSVM has two custom volume structures called Concatenated Pro and Striped Pro. They provide the following:

- Increased tolerance to disk failures
- Improved data redundancy
- Recovery time can be quicker

Layered Volumes

A layered volume is built on one or more other volumes. The underlying volumes are typically mirrored.

The underlying volumes in a layered volume are used exclusively by SSVM and are not intended for user manipulation.

With SSVM 3.0 and above, you can create the following types of layered volumes:

- Concatenated pro volume

A concatenated pro volume is a layered concatenated volume that is mirrored.

- Striped pro volume

A striped pro volume is a layered striped volume that is mirrored.

Layered Volumes

Striped Pro Volume Structure

As shown in Figure 7-3, most of the underlying pro volume structure cannot be manipulated by the user. The user area components can be manipulated and operations such as changing the column width or adding another column can be performed.

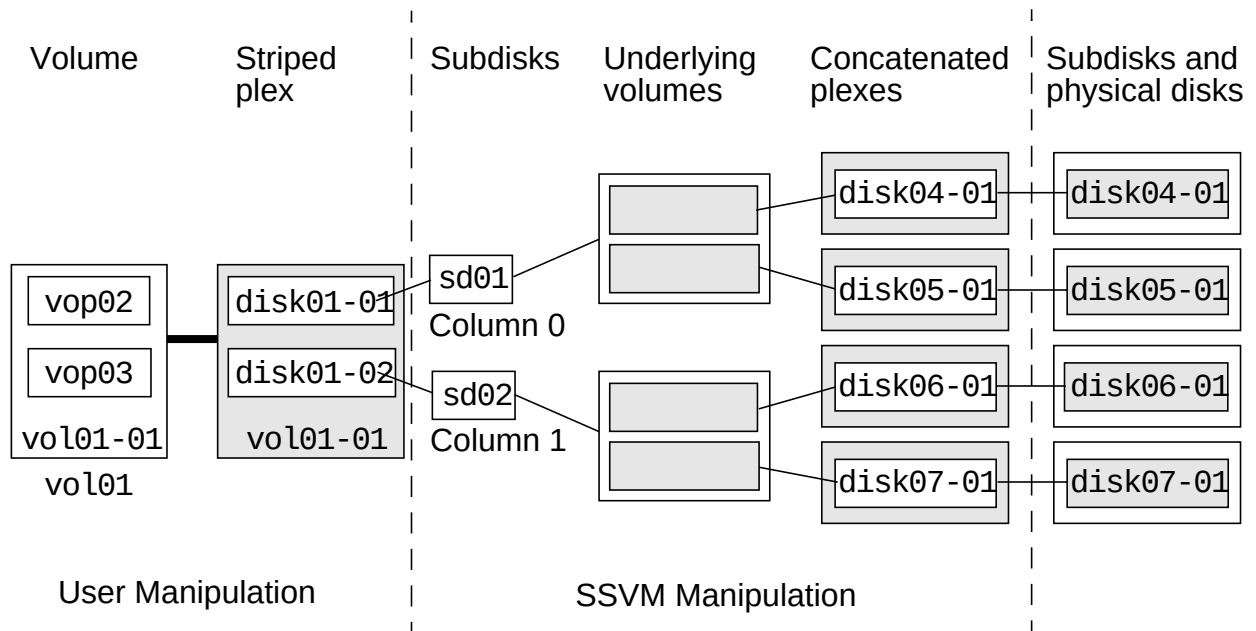


Figure 7-3 Striped Pro Volume Components

The lower levels of the layered volumes are ready-made configurations designed to provide the highest level of availability without increasing the administrative complexity.

Exercise: Performing Advanced Operations



Exercise objective – In this exercise you will:

- Move an empty disk to a different disk group
- Move a populated disk to a new disk group
- Perform a snapshot backup
- Move a disk group between systems
- Assign and remove hot spares
- Enable and disable hot relocation
- Create a striped pro volume with a file system

Preparation

There is no special preparation required for this exercise.

Exercise: Performing Advanced Operations

Task – Moving a Populated Volume to Another Disk Group

In this exercise, you are going to move your mirrored file system volume into the disk group that contains your RAID-5 volume.

1. On the SSVM server, use the `vxprint` command to determine the names of the two disks being used in your mirrored file system.

```
# vxprint -g DGa -ht
```

2. Record the names of the two disks being used in your mirrored volume.

First disk: _____ Second disk: _____

3. Use the `vxprint` command to save the volume configuration.

```
# vxprint -hmQq -g olddg vol01 > save_vol01
```

Note – The `save_vol01` file should not be located on the disk group that is being relocated.

4. Unmount appropriate file systems, and stop any processes on the `vol01` volume.

5. Stop the volume.

```
# vxvol -g olddg stop vol01
```

6. Remove the definitions of the structures (volume, plexes, and subdisks) from the configuration database.

```
# vxedit -g olddg -r rm vol01
```

7. Remove the disks from the original disk group.

Note – You need one additional disk in the disk group, other than the disks that you want to move, because at least one disk must remain for the disk group to continue to exist.

```
# vxdg -g olddg rmdisk olddg01 olddg02
```

Exercise: Performing Advanced Operations

Task – Moving a Populated Volume to Another Disk Group

8. Add the remaining disks to the new disk group.

```
# vxdg -g newdg adddisk olddg02=c0t18d0s2
# vxdg -g newdg adddisk olddg03=c0t19d0s2
```

9. Verify that the disks have been added to the new disk group.

```
# vxdisk list | grep newdg
c0t17d0s2    sliced    olddg01    newdg    online
c0t18d0s2    sliced    olddg02    newdg    online
c0t19d0s2    sliced    olddg03    newdg    online
```

10. Use the vxmake command to reload the saved configuration for the volume vol01.

```
# vxmake -g newdg -d save_vol01
```

Recall that earlier the volume configuration was saved in the file save_vol01. The -d option is used to specify the description file to use for building subdisks, plexes, and volumes.

11. Use the vxvol command to bring the volumes back online.

```
# vxvol -g newdg init active vol01
```

12. Mount the file system to return it to service.

Exercise: Performing Advanced Operations

Task – Moving a Disk Group Between Systems (Optional)

Before proceeding, ask your instructor if you have the proper hardware configuration to perform this section.

1. Prepare your disk group for deportation as follows:
 - ▼ Stop all application or user access to file system or database volumes
 - ▼ Unmount all file systems
 - ▼ Stop all volumes
2. Perform a normal deport operation, the hostid is cleared automatically.

```
# vxdg deport disk_group_name
```

3. Perform a simple import of a clean disk group.

```
# vxdg import disk_group_name
```

4. Use the `deport` and `import` subcommands to return the disk group to the original system.

Exercise: Performing Advanced Operations

Task – Adding and Disabling a Hot Spare

Do not perform this section unless you are sure there are enough spare disk drives to add one to your disk group.

1. Add a spare disk drive to your disk group.
2. On the SSVM server, designate the new disk as a hot spare for your disk group.

```
# vxedit -g disk_group set spare=on diskname
```

3. Verify the spare status of the disks with the command:

```
# vxdisk list
```

Look for the following output within the list:

```
c4t21d0s2    sliced    spare    DGa    online spare
```

Note – The device will be different, but the key to identifying the *in-use* spares is the word “spare” following the word “online”.

4. Disable the disk as a hot spare with the `vxedit spare=off` command option.

Note – The same operation sequence can be performed from the Disk Properties display.

Exercise: Performing Advanced Operations

Task – Performing a Snapshot Backup

Complete the following steps:

1. Select the volume to be copied to a snapshot.
2. Choose Volumes ► Snapshot (Selected) menu.
3. Complete the Volume Snap dialogue box.
4. Click Snapstart in the dialogue box to start the snapshot process.

This may take quite a bit of time depending on the volume size.

5. Click Snapshot when the mirror copy is complete.

This will detach the new mirror and create a separate volume from it.

6. Backup the new snapshot volume to tape (if possible).
7. Remove the snapshot volume.

Exercise: Performing Advanced Operations

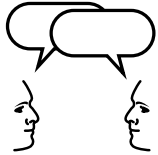
Task – Creating a Striped Pro Volume

Use the following steps:

1. Remove both the mirrored and RAID-5 volume structures in your disk group.
2. Display the disk in your disk group in the grid area and select four of them.
3. Select the New button in the tool bar.
4. Fill out the New Volume as follows:
 - ▼ Enter a Volume Name
 - ▼ Enter a Size of 4m
 - ▼ Select the Striped Pro layout
 - ▼ Assign a disk if it is not already shown
 - ▼ Add a file system that mounts at boot time.
5. Select OK on the Create Volume form.
6. Verify that your new Stripe Pro file system is operational.

Exercise: Performing Advanced Operations

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences
- Interpretations
- Conclusions
- Applications

Check Your Progress

Before continuing on to the next module, check that you are able to accomplish or answer the following:

- ☐ Move an empty disk to a different disk group
- ☐ Move a populated disk to a new disk group
- ☐ Perform a snapshot backup
- ☐ Move a disk group between systems
- ☐ Assign and remove hot spares
- ☐ Enable and disable hot relocation
- ☐ Create a stripe pro volume with a file system

Think Beyond

What do you think is the most important administrative duty you will regularly perform?

What will happen to remote file system users when a system crashes and the related disk group is imported to another host?

Sun StorEdge Volume Manager Performance Management

8 

Objectives

Upon completion of this module, you should be able to:

- Describe how data assignment planning can improve system performance
- List the volume configurations that can improve read and write performance
- List the SSVM commands that are used to gather performance information
- Describe the three types of RAID-5 write procedures
- List the three types of RAID-5 write procedures in order of performance efficiency

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

- Do you know how much data is written to a RAID-5 log?
- How can the read policy for mirrored volumes affect performance?
- How can the number of RAID-5 columns affect performance?

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- “Understanding Disk Arrays,” 1994, SunWin Token #11375
- *The RAID Book*, Lino Lakes, MN: The RAID Advisory Board, 1996.
- “Sun Performance Tuning Overview,” December, 1993, Part Number 801-4872-07.
- Wong, Brian. *Configuration and Capacity Planning for Solaris Servers* ISBN 0-13-349952-9.
- Chen, Lee, Gibson, Katz, and Patterson October, 1993. *RAID: High Performance, Reliable Secondary Storage*.



Sun Educational Services

Performance Guidelines

Several strategies can be used to optimize volume performance including:

- Assigning data to physical drives so that the I/O load is balanced among them
- Increasing the access bandwidth to data by using striping, mirroring, or RAID 5
- Increasing the access bandwidth to data by cabling across system boards and controllers

Performance Guidelines

Periodic reassessment of volume performance is necessary on any system. The access to any data structure can increase over time to the point of poor performance.

Data Assignment

When deciding where to locate file system, a system administrator usually attempts to balance I/O load among available disk drives. The success of this process is limited by the difficulty of anticipating future usage patterns.

In general, file systems that might have heavy I/O loading should not be placed on the same disk(s). Separate them into different storage arrays on various controllers.

Also, the placement of logs can be critical to performance. This is especially true of RAID-5 logs.

Performance Guidelines

Data Assignment

Figure 8-1 illustrates how data assignment mistakes can lead to a performance problem.

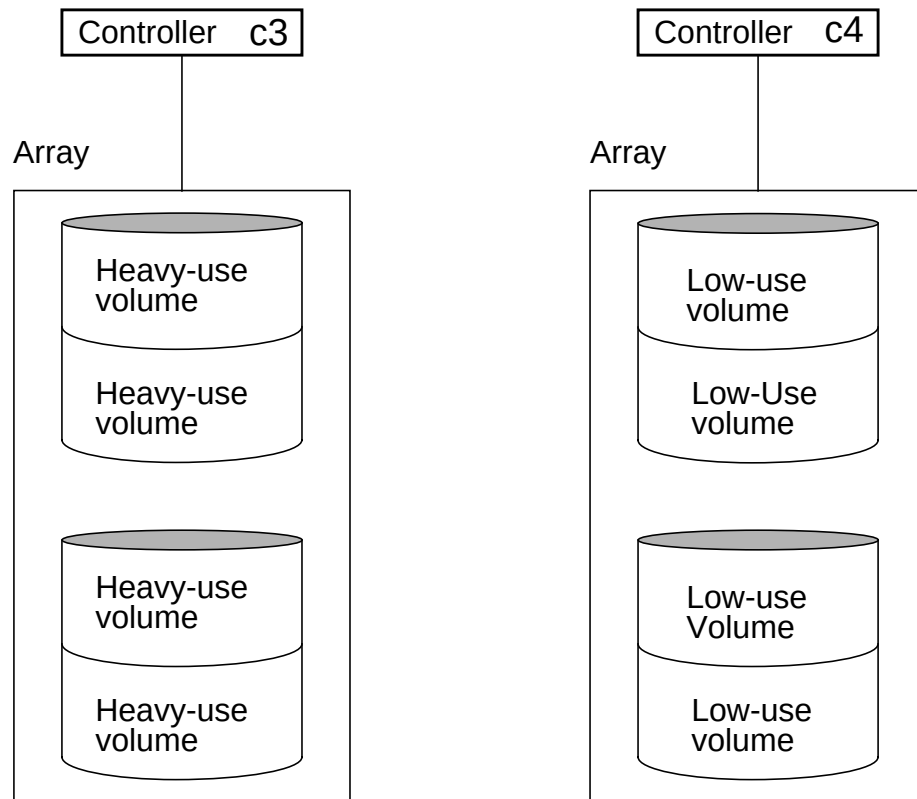


Figure 8-1 Data Assignment Bottleneck

The following solutions can be used to resolve the problem demonstrated in Figure 8-1:

- Swap some of the heavy-use volumes with the low-use volumes.
- Move one of the heavy use disks to a different storage array.

Note – Swapping volume locations is probably a better solution because it eliminates having two heavily used volumes on a single disk drive.

Performance Guidelines

Data Assignment

Another type of performance problem can occur when a log plex is placed on the same disk as its associated data plex. In the case of RAID-5 logs, you should always consider the following:

- The data written to all RAID-5 columns must also be written to the log.

In a six-column RAID-5 volume, this could increase the I/O rate of the log disk by as much as 600 percent.

As shown in Figure 8-2, leaving space at the end of all disks ensures you will always have alternate locations to move logs.

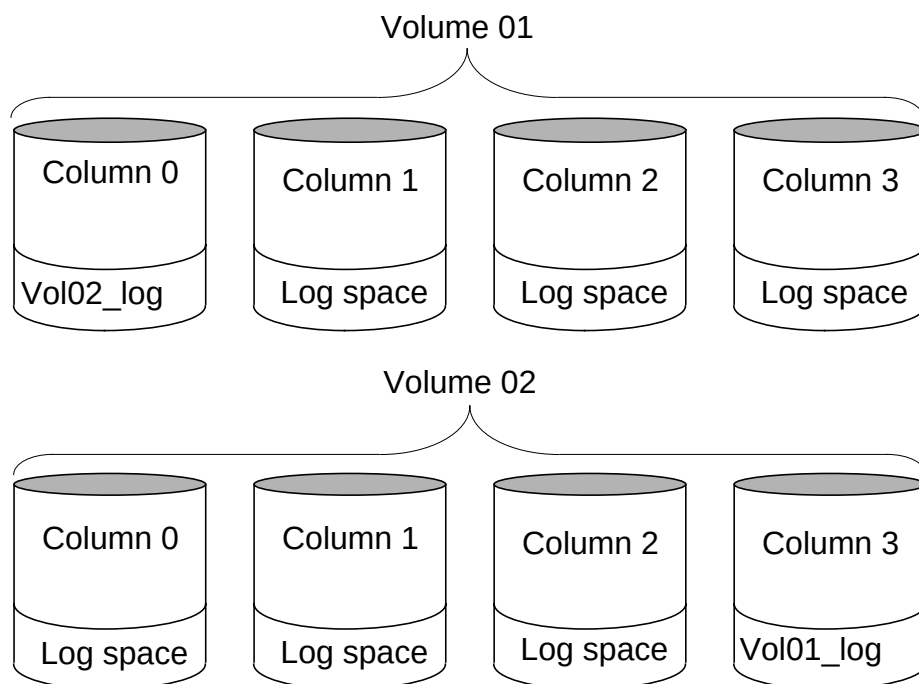


Figure 8-2 RAID-5 Log Placement

The log placement shown in Figure 8-2 would not work well if both volumes were heavily accessed. The configuration would work best if at least one of the volumes has low write activity.

Performance Guidelines

Bandwidth Improvement

Sometimes, performance problems are not due to physical volume locations and can be greatly reduced by reconfiguring the volume structures. In many cases, this can be accomplished using the SSVM Volume Relayout feature.

Striping

Striping distributes data across multiple devices to improve access performance. Striping improves performance for both read and write operations.

If the most heavily accessed volumes (containing file systems or databases) can be identified during the initial design stages, then performance bottlenecks can be eliminated by striping them across several devices. The example in Figure 8-3 shows a volume (Hot Vol) that was identified as being a data-access bottleneck. The volume is striped across four disks, leaving the remainder of those four disks free for use by less heavily used volumes.

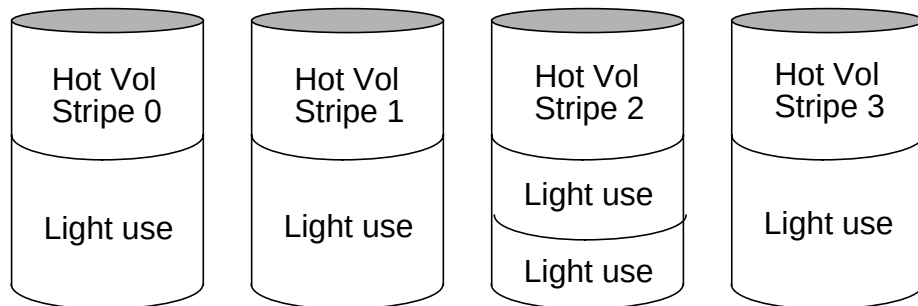


Figure 8-3 Using Striping for Performance

Performance Guidelines

Bandwidth Improvement

Mirroring

Mirroring stores multiple copies of data on a system. Mirroring is primarily used to protect against data loss due to physical media failure. It also improves the chance of data recovery in the event of a system crash.

In some cases, mirroring can also be used to improve system performance. Mirroring heavily accessed data not only protects the data from loss due to disk failure, but can also improve I/O performance. Unlike striping, however, performance gained through the use of mirroring depends on the read / write ratio of the disk accesses. If the system workload is primarily write-intensive (for example, greater than 30 percent writes), then mirroring can result in somewhat reduced performance.

To provide optimal read performance for different types of mirrored volumes, SSVM supports the following read policies:

- The *round-robin* read policy (round)

Read requests to a mirrored volume are satisfied in a round-robin manner from all plexes in the volume. This is selected when there is no significant performance advantage by using any particular mirror.

- The *preferred-plex* read policy (prefer)

Read requests to a mirrored volume are satisfied from one specific plex (presumably the plex with the highest performance), unless that plex has failed.

- The default read policy (select)

The appropriate read policy is automatically selected for the configuration; for example, selecting preferred-plex when there is only one striped plex associated with the volume and round-robin in most other cases.

Performance Guidelines

Bandwidth Improvement

Mirroring

In the example in Figure 8-4, the read policy of the volume labeled Hot Vol should be set to *prefer* for the striped plex labeled Plex 1. In this way, read requests are directed to the striped plex which has the best performance characteristics.

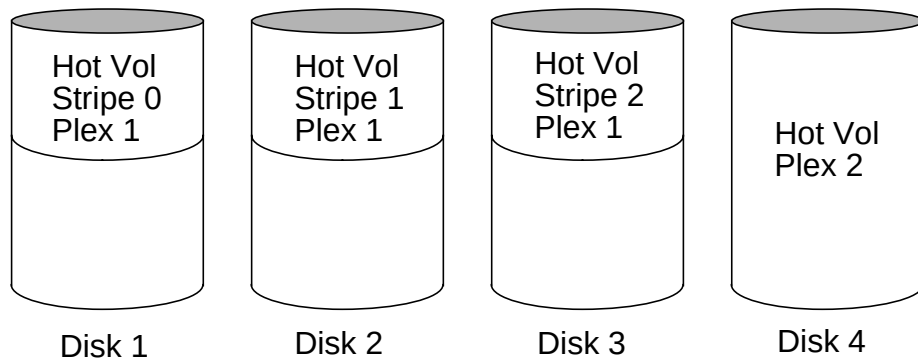


Figure 8-4 Preferred-Plex Read Policy

Mirroring and Striping

This is called RAID 0+1. It is sometimes referred to as striped mirrors. When used together, mirroring and striping provide the advantages of both spreading the data across multiple disks and providing redundancy of data.

Striping and Mirroring

This is a RAID 1+0 setup. It is sometimes referred to as mirrored stripes. The performance is usually the same as RAID 0+1, but this configuration can tolerate a higher percentage of disk failures.

Performance Guidelines

Bandwidth Improvement

RAID 5

RAID 5 provides the advantage of read performance similar to that of striping while also providing data protection via a distributed parity scheme. The disadvantage of RAID 5 is relatively slow write performance.

RAID 5 is not generally seen as a performance improvement mechanism except in cases of highly read-intensive applications.

Performance Cabling

For increased performance and/or availability, striping and mirroring should be done across system boards, controllers, and targets. The highest level of performance or reliability can be gained by striping or mirroring across system boards as shown in Figure 8-5.

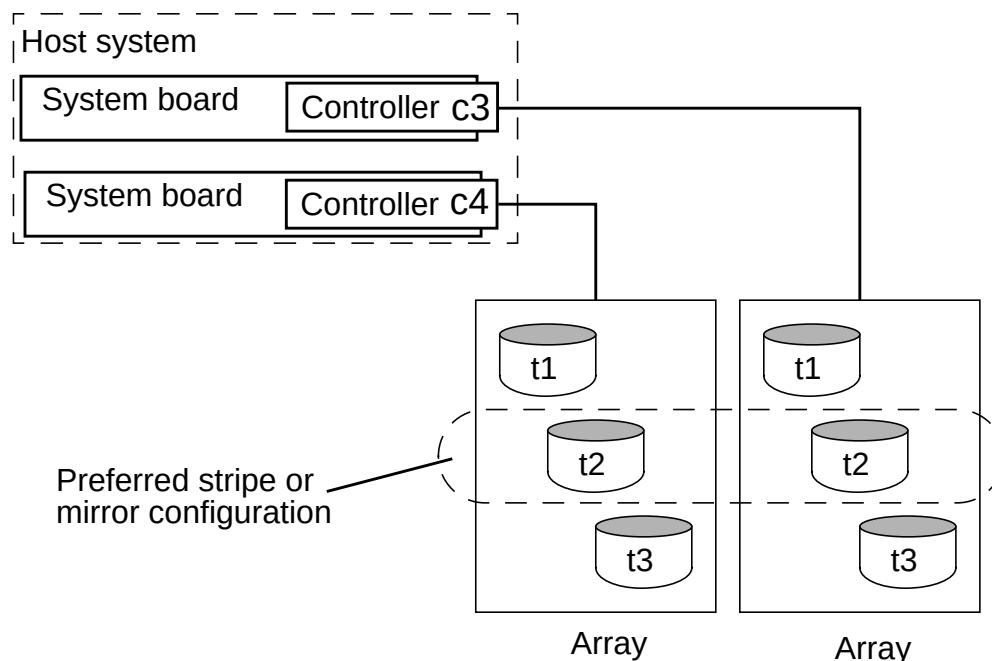


Figure 8-5 High Availability and Performance Cabling



Sun Educational Services

Performance Monitoring

The SSVM software continuously gathers statistical information which can be evaluated using:

- The `vxstat` command
- The `vxtrace` command

Performance Monitoring

Gathering Statistical Information

The SSVM software continuously gathers performance statistics about all devices and objects under its control. The types of information gathered include:

- A count of operations
- The number of blocks transferred (one operation can involve more than one block)
- The average operation time (which reflects the total time through the SSVM software and is not suitable for comparison against other statistics programs)

Performance Monitoring

Gathering Statistical Information

The statistics include reads, writes, atomic copies, verified reads, verified writes, plex reads, and plex writes for each volume. As a result, one write to a two-plex volume results in at least five operations: one for each plex, one for each subdisk, and one for the volume.

SSVM also maintains other statistical data such as read and write failures.

The statistics are continuously gathered starting with the system boot operation. Reset the statistics prior to a testing operation. This is done only to selected objects or when globally using the `vxstat` command.

Displaying Statistics Using the vxstat Command

The `vxstat` command is used to display the statistical information about different types of SSVM physical and logical objects. The following options can be used to control the display:

- **`vxstat -g disk_group`**

Display volume statistics for the specified disk group.

- **`vxstat -g disk_group vol01`**

Display statistics for the specified volume.

- **`vxstat -g disk_group -d`**

Display disk level statistics for the specified disk group.

- **`vxstat -g disk_group -d disk01`**

Display statistics for the specified disk.

Performance Monitoring

Displaying Statistics Using the vxtrace Command

The `vxtrace` command is used to display detailed trace information about errors or I/O operations. This level of detail is generally not necessary but is included here for completeness. The following options can be used to control the display:

- **`vxtrace -o disk`**

Trace all physical disk I/O operations.

- **`vxtrace -o disk c3t98d0`**

Trace all I/O operations to the physical disk `c3t98d0`.

- **`vxtrace hist2`**

Trace all virtual device I/O operations associated with the volume `hist2`.

- **`vxtrace -o dev hist2`**

Trace virtual disk device I/O to the device associated with volume `hist2`.



Sun Educational Services

Performance Analysis

The SSVM performance analysis process includes the following steps:

1. Resetting the existing data.
2. Maintaining normal system load during evaluation.
3. Identifying SSVM objects with an unusually large number of I/O operations.
4. Identifying SSVM objects with excessive read or write times.
5. Performing detailed traces using the `vxtrace` command.

Performance Analysis

Once performance data has been gathered, it can be used to determine and optimize system configuration for efficient use of system resources.

It should be noted that a volume or disk with elevated read or write access times is not necessarily a problem. If the slow response is not causing any apparent problems for users or applications, then there might not be anything that needs fixing.

Performance Analysis

Preparation

Before obtaining statistics, clear (reset) all existing statistics with the `vxstat -r` command. Clearing statistics eliminates any differences between volume or disks due to volumes being created, and also removes statistics from booting which are not normally of interest.

After clearing the statistics, allow the system to run during typical system activity. When monitoring a system that is used for multiple purposes, try not to exercise any one application more than it would be exercised normally.

It can also be beneficial to take periodic *snapshots* of the volume statistics to help identify the source of irregular system load problems.

A single volume that has excessive I/O rates can cause performance degradation on other volumes associated with the same physical disk drives.

Performance Analysis

Volume Statistics

You can use the `vxstat` command as follows to help identify volumes with an unusually large number of operations or excessive read or write times:

```
# vxstat -g bench
```

		OPERATIONS		BLOCKS		AVG TIME(ms)	
TYP	NAME	READ	WRITE	READ	WRITE	READ	WRITE
vol	acct	473	11	57252	44	4.0	20.9
vol	brch	23	11	92	44	33.0	20.0
vol	ctrl	367	18000	1675	72000	9.3	21.7
vol	hist1	23	11	92	44	97.0	24.5
vol	hist2	23	11	92	44	54.8	22.7
vol	hist3	23	11	92	44	103.5	25.5
vol	log1	9	27217	9	409716	16.7	21.1
vol	log2	7	8830	7	159769	15.7	24.3
vol	rb1	123	13	492	52	30.7	83.1
vol	rb2	23	11	92	44	149.1	25.5
vol	sys	26933	86156	177688	344632	15.0	310.7
vol	t11r	23	11	92	44	39.6	24.5

Disk Statistics

The `vxstat` command can also summarize operations according to physical disk drives. For example:

```
# vxstat -g bench -d
```

		OPERATIONS		BLOCKS		AVG TIME(ms)	
TYP	NAME	READ	WRITE	READ	WRITE	READ	WRITE
dm	c3t98d0	14330	140370	120348	986785	15.4	185.6
dm	c3t100d0	13881	140370	117971	986785	15.4	187.7
dm	c3t113d0	0	0	0	0	0.0	0.0
dm	c3t115d0	0	0	0	0	0.0	0.0
dm	c3t117d0	0	0	0	0	0.0	0.0

Performance Analysis

Trace Information

After identifying a volume that has a I/O related problem, you can use the `vxtrace` command to determine which system process is responsible for the I/O requests. The volume of interest in this example is named `ctrl`.

```
# vxtrace -o dev ctrl
40122 START write vdev ctrl block 16 len 4 concurrency 1 pid 10689
40122 END write vdev ctrl op 40122 block 16 len 4 time 1
40123 START write vdev ctrl block 16 len 4 concurrency 1 pid 10689
40123 END write vdev ctrl op 40123 block 16 len 4 time 2
40124 START write vdev ctrl block 16 len 4 concurrency 1 pid 10689
40124 END write vdev ctrl op 40124 block 16 len 4 time 4
40125 START write vdev ctrl block 16 len 4 concurrency 1 pid 10689
40125 END write vdev ctrl op 40125 block 16 len 4 time 0
^C
#
#
# ps -ef |grep 10689
oracle 10689 1 0 20:05:21 ? 0:03 ora_ckpt_bench
```



Sun Educational Services

RAID-5 Write Performance

To maximize write performance in RAID-5 volumes you must understand the following write operations and how they affect performance:

- Read-modify-write
- Reconstruct-write
- Full-stripe write

RAID-5 Write Performance

The RAID-5 write process is controlled according to how much data will be written into a full stripe width. The optimum write performance is obtained when full stripes are written.

Read-Modify-Write Operations

When less than 50 percent of the data disks are undergoing writes in a single I/O, the read-modify-write sequence is used.

This is the default operation for RAID-5 volumes.

RAID-5 Write Performance

Read-Modify-Write Operations

As shown in Figure 8-6, the read-modify-write involves several steps including:

1. The stripes to be modified are read into a buffer.
2. The parity information is read into a buffer.
3. Exclusive OR (XOR) operations are performed.
4. The new data and parity are written in a single write.

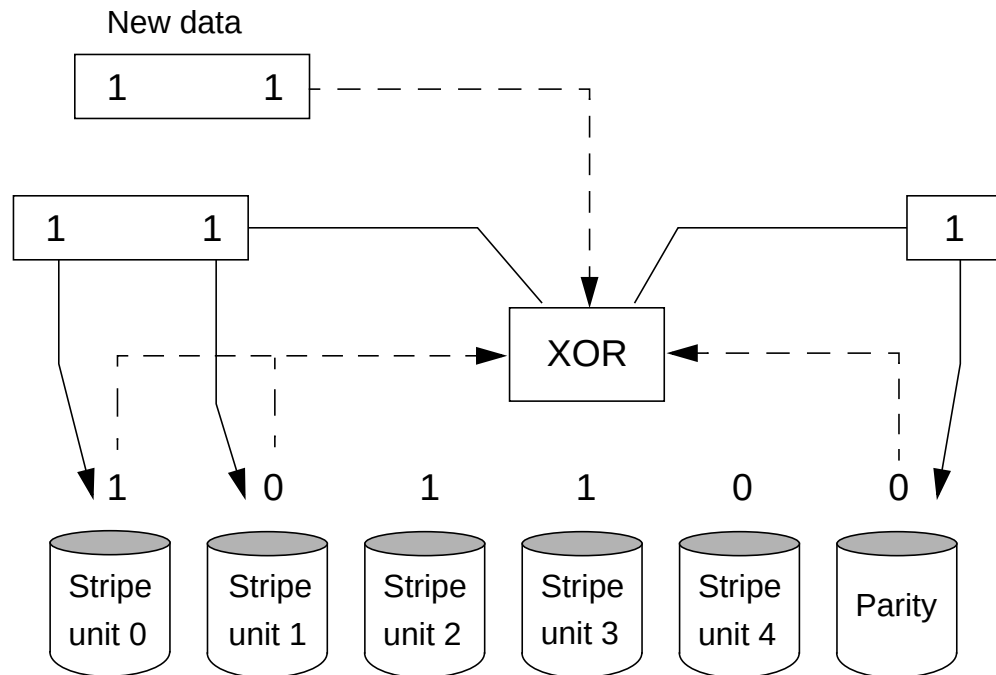


Figure 8-6 Read-Modify-Write Operation

At least three I/O operations are necessary in the example shown in Figure 8-6. Also, additional XOR calculations are necessary to account for the data in stripe units 2, 3, and 4 that was not read.

Generally, the read-modify-write method is the least efficient way of writing to RAID-5 structures.

RAID-5 Write Performance

Reconstruct-Write Operations

If more than 50 percent of the data stripe is going to be modified, the reconstruct-write method is used.

As shown in Figure 8-7, the reconstruct-write involves different steps including:

1. Only unaffected data is read into a buffer.
2. XOR is applied to the new data and the unaffected data.
3. New parity and data are written in a single write.

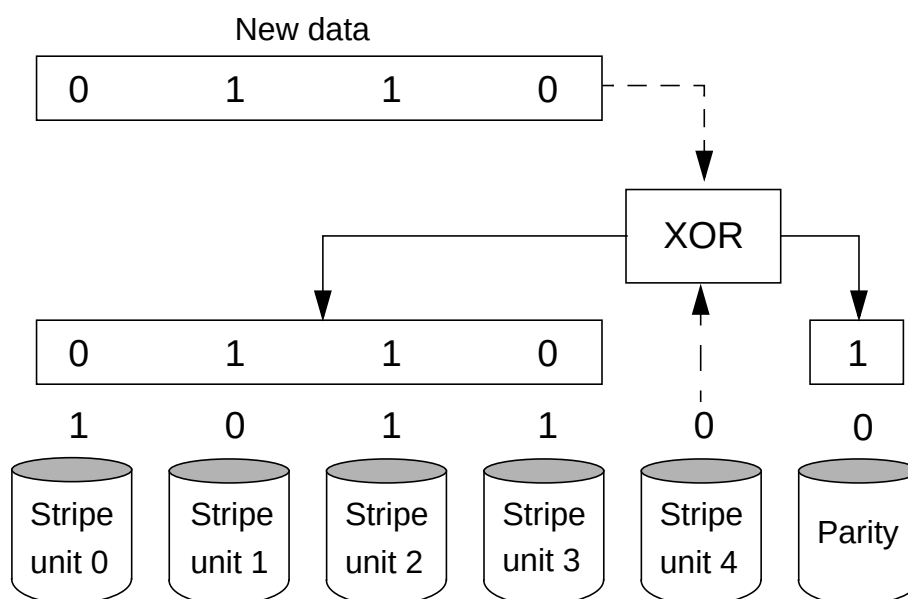


Figure 8-7 Reconstruct-Write Operation

Only two I/O operations are necessary in the example shown in Figure 8-7.

Generally, the reconstruct-write operation is more efficient than the read-modify-write sequence.

RAID-5 Write Performance

Full-Stripe Write Operations

When large writes that cover an entire data stripe are issued, the read-modify-write and reconstruct-write procedures are bypassed in favor of a full-stripe-write. A full-stripe-write is faster than the other RAID-5 write procedures because it does not require any read operations. As shown in Figure 8-8, a full stripe write procedure consists of the following steps:

1. XOR is applied to the new data to produce new parity.
2. The new data and parity are written in a single write.

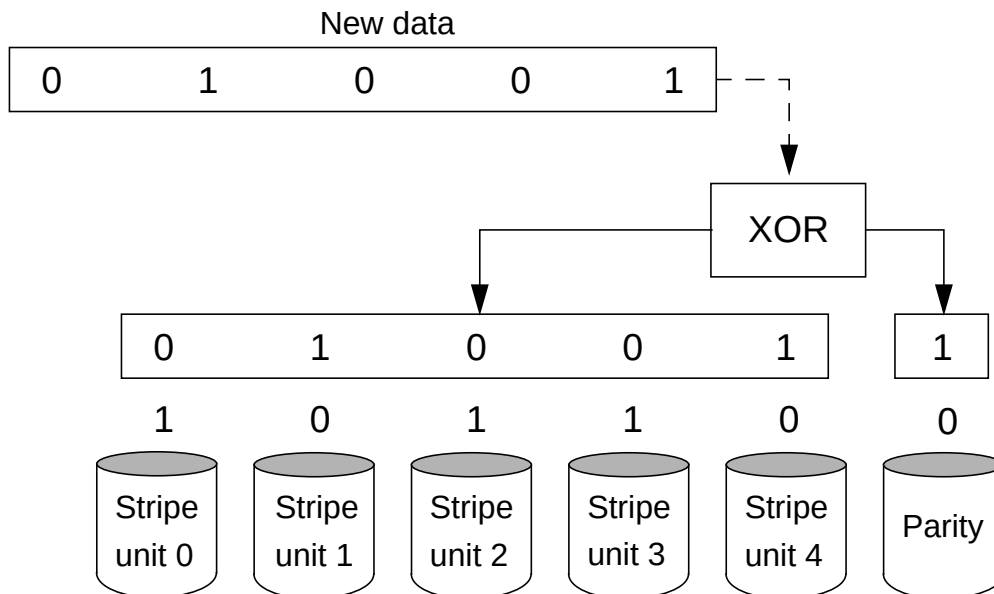


Figure 8-8 Full-Stripe Write Operation

Only a single write operation is necessary in the example shown in Figure 8-8.

Note – In some cases, it is beneficial to reduce the number of RAID-5 columns to force more full-stripe write operations. This can enhance overall write performance for some applications that use random-length writes.

Exercise: Demonstrating Performance Differences



Exercise objective – In this exercise, you will:

- Observe the performance differences between the three types of RAID-5 write operations.
- Watch the instructor perform the tasks on the SSVM server system or remotely log in to the server and have the instructor direct output to your monitor.

Preparation

These steps are completed by the instructor:

1. Prepare a list of five disk names that will be used to build a five-column, RAID-5 volume that is 30 Mbytes in size.

Note – Do not create it yet.

disk01: _____ disk02: _____
disk03: _____ disk04: _____
disk05: _____

2. Log in remotely or locally to the SSVM server.
3. Add `/usr/proc/bin` to the search path.

```
# PATH=$PATH:/usr/proc/bin  
# export PATH
```

4. Have the students remotely log in to the server if possible.
5. Have the students give you their terminal identifiers by typing the `tty` command. Record them here:

_____	_____	_____
_____	_____	_____
_____	_____	_____

Exercise: Demonstrating Performance Differences

Task – Performing the Demonstration

Your instructor will use the following steps:

1. Direct the output of your window to all of the student systems.

```
# script /dev/null |tee /dev/pts/5|tee /dev/pts/8
```

Note – Later, you will probably have to kill all of the `tee` processes which should take care of the `script` processes and end everything.

2. Create a five-column, no log, RAID-5 volume with 30 Mbytes.

```
# vxassist -g disk_group make r5demo 30m layout=raid5,nolog disk01 \
disk02 disk03 disk04 disk05
```

Note – The default stripe unit size is 32k. The stripe width counts up as 16k, 32k, 48, 64k, and 80k, which is full stripe width.

3. Make a link to the new volume. (This saves typing time later.)

```
# ln /dev/vx/rdsk/disk_group/r5demo raidvol
```

4. Create a 20-Mbyte test file (check for space first).

```
# mkfile 20m /testfile
```

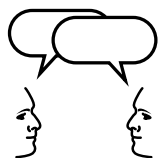
5. Repeat the following command sequence and each time increment the block size of the `dd` command (32, 48, 80, 81).

```
# vxstat -g disk_group -r r5demo
# /usr/proc/bin/ptime dd if=/testfile of=/raidvol bs=32k
# vxstat -g disk_group -f MRF -v r5demo
```

Note – As you move past the 50 percent stripe write into full stripe write, the I/O should move through the three write categories (M, R, and F).

Exercise: Demonstrating Performance Differences

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences
- Interpretations
- Conclusions
- Applications

Check Your Progress

Before continuing, check that you are able to accomplish or answer the following:

- ☐ Describe how data assignment planning can improve system performance
- ☐ List the volume configurations that can improve read and write performance
- ☐ List the SSVM commands that are used to gather performance information
- ☐ Describe the three types of RAID-5 write procedures
- ☐ List the three types of RAID-5 write procedures in order of performance efficiency

Think Beyond

Are there features in some of your applications that might help with increasing performance?

Instead of tuning for performance, are there user related strategies that might help reduce system loads?

Objectives

Upon completion of this module, you should be able to:

- Discuss the features and benefits of the RAID Manager software
- Define the terms:
 - ▼ *Logical unit*
 - ▼ *Drive group*
 - ▼ *RAID module*
- Discuss hot spare usage
- Describe the data reconstruction process
- Describe RAID Manager device naming conventions
- Define caching control options

Relevance



Discussion – The following questions are relevant to understanding the content of this module:

- Which RAID Manager component functions as the link between the user interface and the Solaris kernel?
- How does a RAID level differ from a RAID module?
- What is the maximum number of disk drives that can be configured into a drive group?
- How is the RAID level of a specific LUN determined?
- How do hot spares aid in the reconstruction process?
- How does cache memory increase overall storage performance?
- When would a storage administrator use the CLI instead of the GUI?
- What changes to the standard Solaris device naming conventions are required when using the RAID Manager software?

Additional Resources



Additional resources – The following reference can provide additional details on the topics discussed in this module:

- <http://docs.sun.com/ab2>

 Sun Educational Services

RAID Manager Components and Features

- Major components
 - User interfaces
 - RAID Manager (RM) engine
 - Redundant dual active controller (RDAC) driver
- RAID Manager features
 - Solstice DiskSuite™ compatible features
 - Volume Manager compatible features
 - Unsupported features

RAID Manager Components and Features

This section focuses on the RAID Manager software in terms of:

- Major components
 - ▼ User interfaces
 - ▼ RAID Manager (RM) engine
 - ▼ Redundant dual active controller (RDAC) driver
- RAID Manager features
 - ▼ Solstice DiskSuite™ compatible features
 - ▼ Volume manager compatible features
 - ▼ Unsupported features

RAID Manager Components and Features

RAID Manager Components

As illustrated in Figure 9-1, the RAID Manager software has three major components:

- User interface – Both a GUI and CLI are provided, and either can be used for array configuration, monitoring, and maintenance.
- RM engine – The RM engine runs on the host server and is the RAID Manager link between the user interface and the Solaris kernel.
- RDAC driver – The RDAC driver is a kernel-level driver that enables automatic failover to the second controller when one controller fails. The RDAC driver logically resides above the Solaris SCSI driver in the Solaris kernel.

This differs from DMP where there are two simultaneously active paths, one to each controller. With RDAC, each controller has access to the same SCSI drives. Thus, should the path to one controller fail, a second path is brought into play. The definition of the logical drives which maps a set of physical drives fails over to the alternative controller.

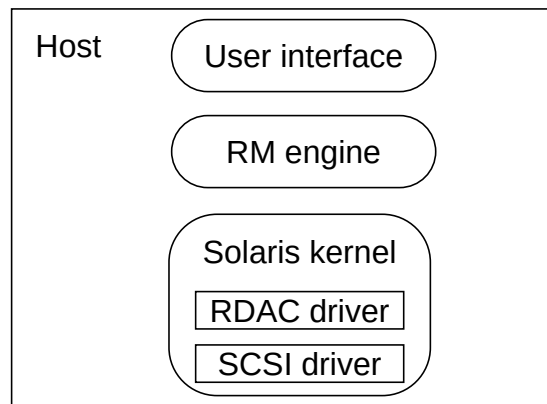


Figure 9-1 RAID Manager Components

RAID Manager Components and Features

RAID Manager Features

- Virtual disks – Virtual disks are a logical grouping of one or more physical disks presented as one device to the operating system, application, and user.
- Striping – This is the ability to group multiple disks together. The addressing is interleaved among the disks.
- Mirroring – Multiple copies of the data are kept. In Solstice DiskSuite (SDS) each copy is called a submirror; in Volume Manager (VM) each copy is contained in a plex.
- RAID 5 – Data is protected by parity. The parity is interspersed among the data.
- Hot spares – Disks / partitions can replace failed disks / partitions. Data is reconstructed on the hot spare. In SDS, hot spares are not permanent; they go back to the hot spare pool. In VM, they are permanent replacements.
- Disk grouping – RM6 has drive groups, which provide pools of drives from which to define logical drives. Drive groups cannot be exported between hosts.

Note – In Solstice DiskSuite, this is called a diskset. In Volume Manager, it is called a disk group. The difference between these and the RM6 drive groups is that these can be used to move groups of disks between hosts, and is the basis for high availability solutions.

- GUI – A graphical user interface is supported.
- Free space management – Because RM6 is based on physical partitions, it is easy to determine what is available. With VM, you have to total the free space in holes and at the end of disks.



Sun Educational Services

Definitions

- RAID module
- Drive group
- Logical unit (LUN)
- Drive group numbering
- Hot spare drive

Definitions

Because some terms are used repeatedly in the storage industry, you should learn the definitions of some common storage terms as they relate to the RAID Manager. Some of these terms are:

- RAID module
- Drive group
- LUN
- Drive group numbering
- Hot spare drive

Definitions

RAID Module

A *RAID module* is defined as a set of controllers, disk drives, and associated power supplies and cooling units. In other words, a StorEdge A3000 or A1000 is a RAID module.

RAID modules are selected when performing various administrative tasks such as configuring, obtaining status, or recovering.

RAID module numbers are assigned in the order in which the host system detects them and uses the hostname in the name of the RAID module; for example, mars_001, mars_002, and so on.

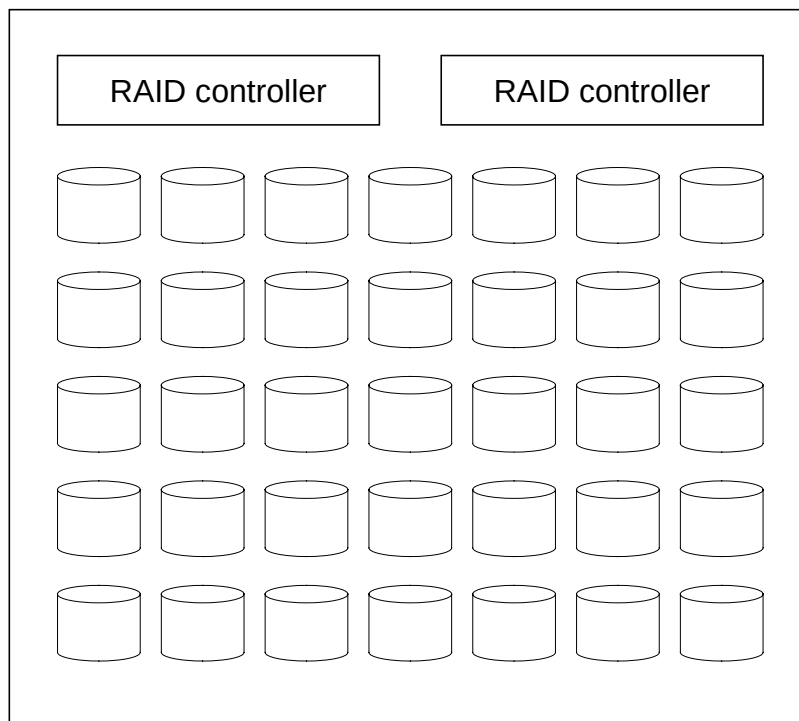


Figure 9-2 RAID Module

Definitions

Drive Group

A *drive group* is a set of physical drives in a particular RAID module. Drive groups define a pool of space in which all the logical volumes are of the same RAID level, the RAID level of the drive group. In addition, when the host system has access to the drives through two controllers, load balancing between the controllers is achieved through sharing drive groups between controllers. Drive groups are defined from the Configuration window. Three types of drive groups are available:

- Configured drive groups – These groups have been configured into one or more logical units with the same RAID level. There is a maximum of 20 disk drives per drive group.
- Unassigned drive group – This group consists of all the disk drives that are not currently configured into logical units or hot spares drive groups.
- Hot spare drive group – This group consists of all disk drives that can be assigned as hot spares. Like the configured drive groups, hot spare drive groups can consist of a maximum of 20 disk drives.

Definitions

Drive Group

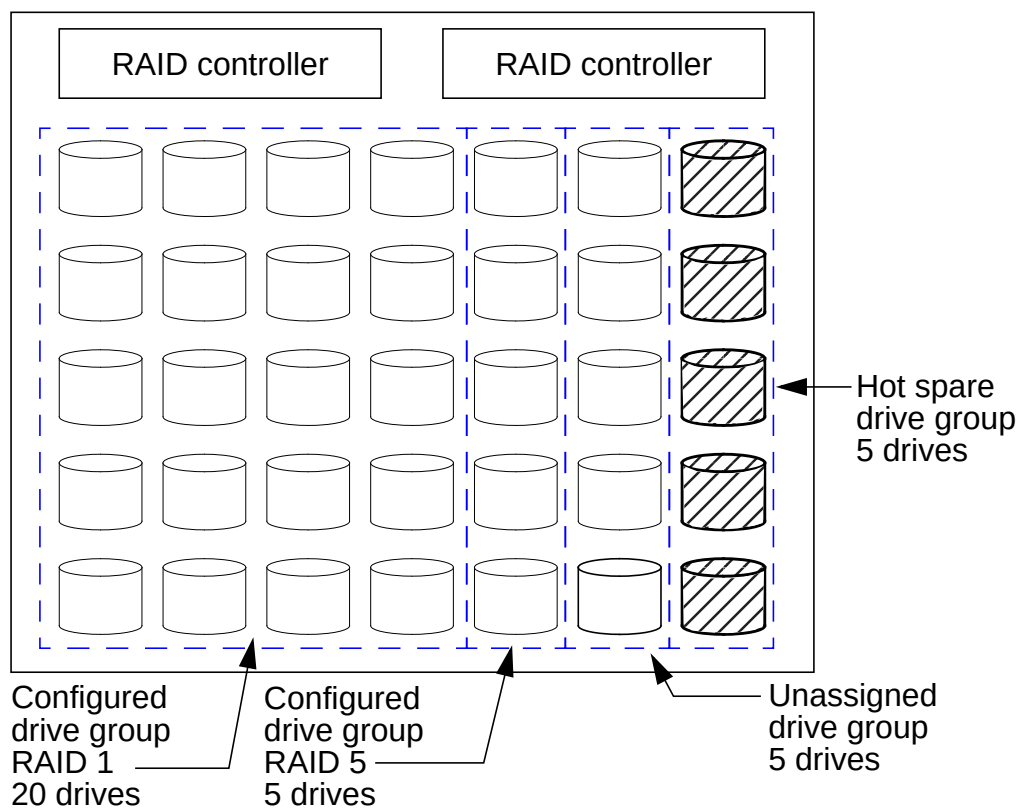


Figure 9-3 RAID Manager Drive Groups

Definitions

Logical Unit (LUN)

Defining a LUN

A *LUN* spans one or more drives and is configured into either RAID 0, 1, 3 or 5. The RAID level is determined by the drive group to which the LUN is associated. The RM6 LUN is similar to a volume in VM.

- A drive group can contain one or more LUNs.
- Every LUN within a drive group shares the same physical drives and RAID level.
- Each LUN is seen by the operating system as *one* virtual drive and may include up to 20 physical disk drives.
- Each LUN can be sliced into multiple partitions using the `format` command, because the Solaris OS sees the LUN as a drive.
- Under the Solaris 2.6 OS, a maximum of 16 LUNs are permitted on each RAID module (StorEdge A3000).

This drive group is configured with 20 disk drives that are divided into four LUNs (LUN 0, 1, 2, and 3).

LUN 0 is c1t5d0s2
LUN 1 is c1t5d1s2
LUN 2 is c1t5d2s2
LUN 3 is c1t5d3s2

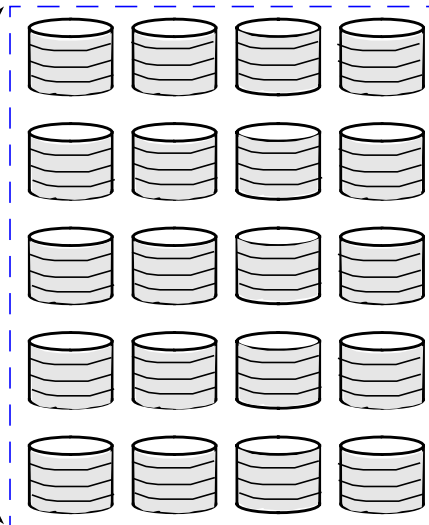
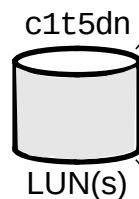


Figure 9-4 Logical Unit

Definitions

Logical Unit

Configuring LUNs

The LUN configuration information is stored redundantly on three disk drives in the configuration. The location of these drives is not known to the user and is managed by the controller. The LUN information is not user accessible (it is not stored in a text file). It is visible only through the RM6 Configuration Manager application.

The controller serial numbers are stored as part of the LUN configuration information. This information is checked during the Start-of-Day test.

Using LUN Partitioning

Following is the default partition table, created by format, of a 20-drive RAID 5 LUN:

Current partition table (original):

Total disk cylinders available: 38882 + 2 (reserved cylinders)

Part	Tag	Flag	Cylinders	Size	Blocks	
0	root	wm	0 - 63	128.00MB	(64/0/0)	262144
1	swap	wu	64 - 127	128.00MB	(64/0/0)	262144
2	backup	wu	0 - 38881	75.94GB	(38882/0/0)	159260672
3	unassigned	wm	0	0	(0/0/0)	0
4	unassigned	wm	0	0	(0/0/0)	0
5	unassigned	wm	0	0	(0/0/0)	0
6	usr	wm	128 - 38881	75.69GB	(38754/0/0)	158736384
7	unassigned	wm	0	0	(0/0/0)	

Definitions

Drive Group Numbering

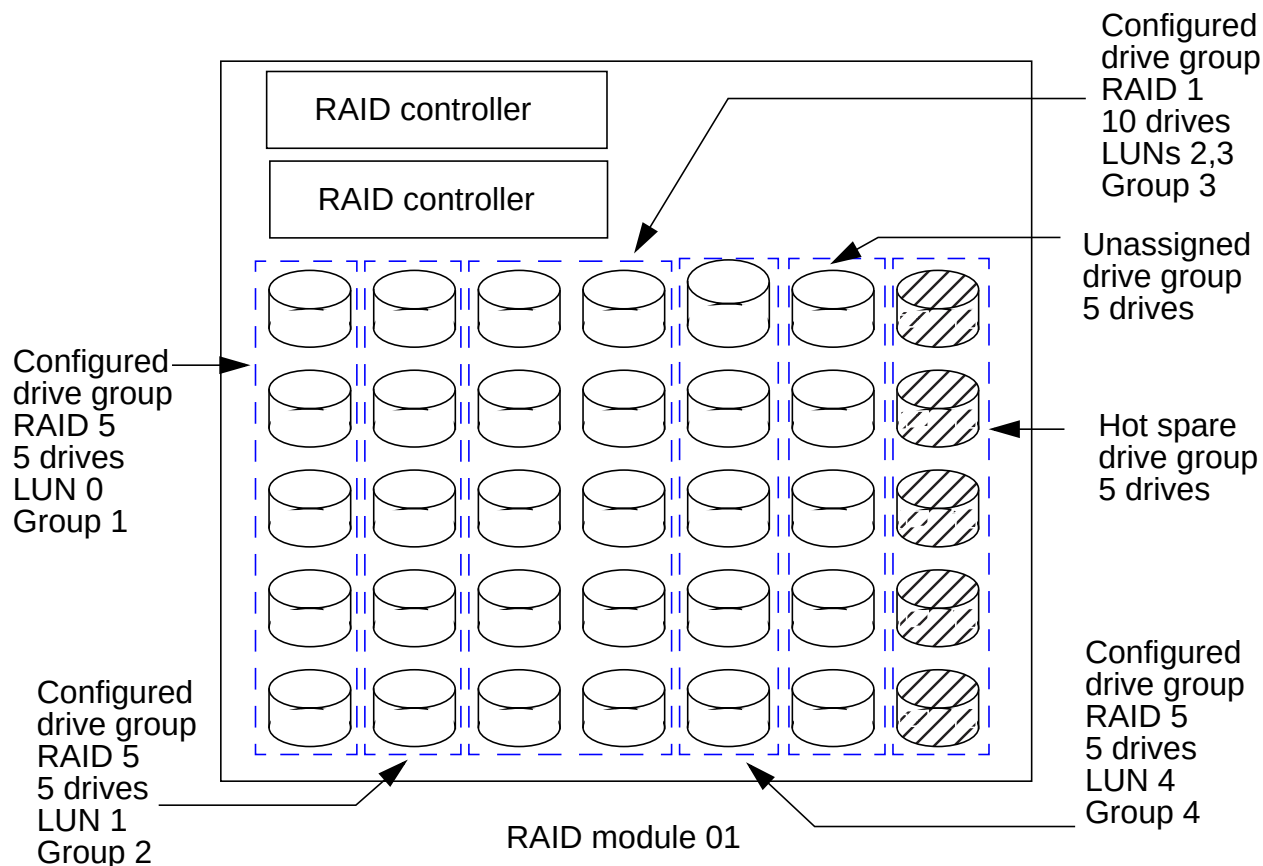
Each configured drive group is assigned a number. Drive group numbering is dynamic and the lowest numbered LUN, LUN 0, is always in group 1.

The drive group that contains LUN 0 will always be drive group 1. If LUN 0 is removed from group 1, and all the other LUNs are being used, and a new LUN is added to group 3 (because there was unused disk space in that group), then when LUN 0 is added to group 3 it will become group 1. Drive group 2 will contain the next lowest numbered LUN.

This reconfiguration of the group number has absolutely no effect on the way the drives are mounted or used, the LUNs and their controller associations remain unchanged.

Definitions

Drive Group Numbering



- Drive group numbering
 - ▼ Each configured drive group is assigned a number.
 - ▼ Drive group numbering starts with the lowest numbered LUN.
 - ▼ Due to the order of assignment, the LUN numbers may not be contiguous with the groups.
 - ▼ Drive groups can renumber automatically after deleting and creating LUNs. Note that LUN numbers are used by the administrator when creating file systems, and so on. Drive group numbers are internal to RM6 configuration. Drive group renumbering has no effect on the Solaris OS's view of the LUNs.

Definitions

Hot Spare Drive

A *hot spare* drive is a drive that contains no data and acts as a standby in case a drive fails in a RAID-1 (mirrored), RAID-3, or RAID-5 logical unit.

When a drive fails in a RAID-1, RAID-3, or RAID-5 logical unit, a hot spare drive automatically replaces the failed drive and the data is regenerated and written to the hot spare. When the failed drive is replaced and the recovery process is completed, the hot spare drive automatically returns to a standby status.

Hot spares are not dedicated to a specific drive group or LUN. They can be used to replace any failed drive in the RAID module with the same or smaller capacity.

Hot spare drives provide additional redundancy and allow deferred maintenance. Depending on how many hot spares you configure, a logical unit could remain fully operational and still have multiple failed drives, each one being covered by a hot spare.

Note – Hot spare drives cannot be used to replace failed drives in a RAID-0 logical unit

There is no redundancy in a RAID-0 (striped) logical unit, so therefore there is no way to reconstruct the data on the hot spare. If a drive fails in a RAID-0 logical unit, the data will remain unavailable until the failed disk is replaced and the data is restored from backup media.

Definitions

Hot Spare Drive

In summary:

- A hot spare drive is a drive that contains no data and acts as a standby in case a drive fails in a RAID-1 (mirrored), RAID-3, or RAID-5 logical unit.
- When a drive fails, a hot spare drive replaces the failed drive.
- When the failed drive is replaced and the recovery process is completed, the hot spare drive automatically returns to standby status.
- Hot spares can be used to replace any failed drive in a RAID module with the same or smaller capacity.
- Hot spare drives provide additional redundancy and enable deferred maintenance.



Sun Educational Services

RAID Reconstruction

- Degraded mode
- Reconstruction
- Hot spares
- RAID 1 (mirroring) LUN difference

RAID Reconstruction

This includes:

- Degraded mode
- Reconstruction
- Hot spares
- RAID-1 (mirroring) LUN difference

RAID Reconstruction

Degraded Mode

When a RAID-3 or RAID-5 LUN experiences a single disk failure, it can continue to make data available, although in a degraded mode. Remember from the previous RAID examples in Module 3, “Introduction to Managing Data,” that RAID 3 and RAID 5 utilize a parity scheme that rebuilds missing data in the event the RAID set suffers a single disk failure. The parity disk cannot rebuild missing data if the RAID set suffers failures on multiple disks.

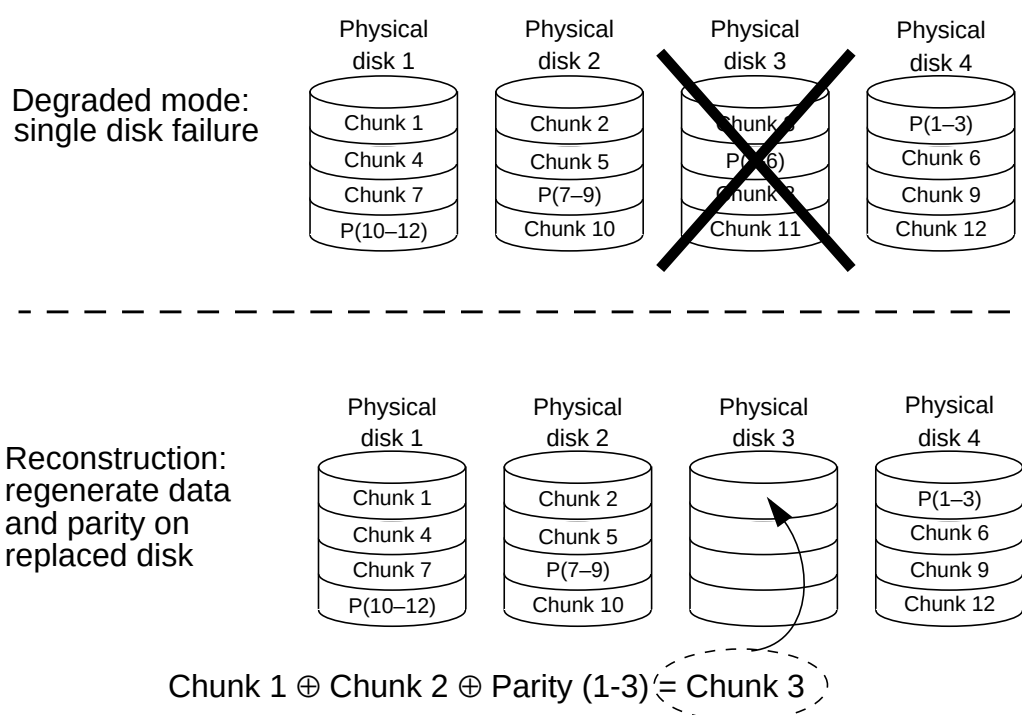


Figure 9-5 RAID Reconstruction

In the RAID-5 example in Figure 9-5, physical disk 3 has failed and there are no hot spares. If an application accesses data on an operational disk (for example, chunk 2), the I/O operation proceeds normally. If, however, the application accesses data on the failed disk (for example, chunk 3), the controller needs to regenerate that data using the remaining data and parity in that stripe. In this example, an XOR is performed on chunk1, chunk2, and the parity for chunks 1, 2, and 3 to regenerate the original chunk3 data.

RAID Reconstruction

Reconstruction

Reconstruction is the process used to restore a degraded RAID-1, -3, or 5 logical unit to its original state after you replace a single failed drive. The example on the preceding page shows chunk 3 being regenerated using the remaining data chunks and parity in the stripe.

Reconstruction occurs when you initiate recovery by physically replacing a failed drive in a RAID-1, -3, or -5 logical unit.

Hot Spares

If you have hot spares configured and you lose a single disk in a RAID-3 or RAID-5 LUN, the LUN continues to be available (although in a non-optimal state) while the data is reconstructed on the hot spare. The reconstruction process is as described previously. When this reconstruction finishes, the LUN status is optimal.

When the original failed disk is replaced, the data is copied from the hot spare to the replaced disk. When this copy is complete, the hot spare returns to standby status.

When a RAID set LUN has been reconstructed using a hot spare disk, the LUN status will be reported as optimal. After replacement of the original failed disk, the LUN is automatically rebuilt onto the replacement drive group component disk. Because the RAID set LUN is already reporting optimal status, having been reconstructed using a hot spare, the LUN remains accessible (with an optimal status) during the rebuild of the replacement disk from the hot spare disk.

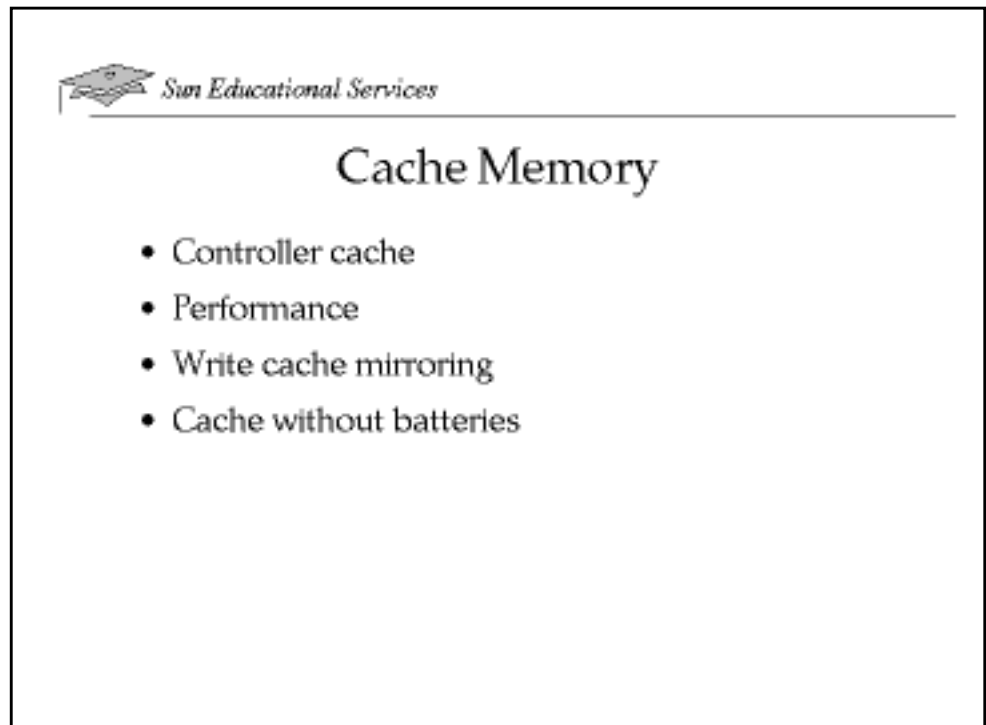
RAID Reconstruction

RAID 1 (Mirroring) LUN Difference

RAID 1 (mirroring) LUNs can also continue to make data available in the event of a single disk failure. You do not need to recalculate the data, however, because you have a good copy on the other mirror in the LUN. In degraded mode (loss of a single disk), all reads and writes are performed on the surviving mirror. When the failed disk is replaced, the data is copied from the surviving mirror to the replaced disk.

In summary:

- Reconstruction is the process used to restore a degraded RAID-1, -3, or -5 logical unit to its original state after you replace a single failed drive.
- During reconstruction, the controller:
 - ▼ Recalculates data on the replaced drive using data/parity from the other drives in the logical unit (RAID 3 or RAID 5).
 - ▼ Writes this data to the replace drive.
- Reconstruction occurs automatically when you initiate recovery by physically replacing a failed drive in a RAID-1, -3, or -5 LUN.



Cache Memory

Cache memory is an area on the controller used for intermediate storage of read and write data.

Controller Cache

By default, each controller has 64 Mbytes of cache. This can be upgraded to 128 Mbytes of cache per controller.

Performance

Cache memory can increase overall performance. Data for read operations may already be in cache, eliminating the need to access the drive itself. Write operations are considered complete once written to the cache. This also improves performance as the application does not need to wait for the data to be written to disk.

Cache Memory

Write Cache Mirroring

When enabled, cached data is written to the cache memory of both controllers so that when a controller fails, the second controller completes all outstanding write operations.

Cache Without Batteries

There are several conditions, such as low battery power, where the controller may temporarily turn off the cache settings until the condition is back to normal. If you set the *cache without batteries* option, the controller will override this safeguard and continue to use caching even without the battery backup.



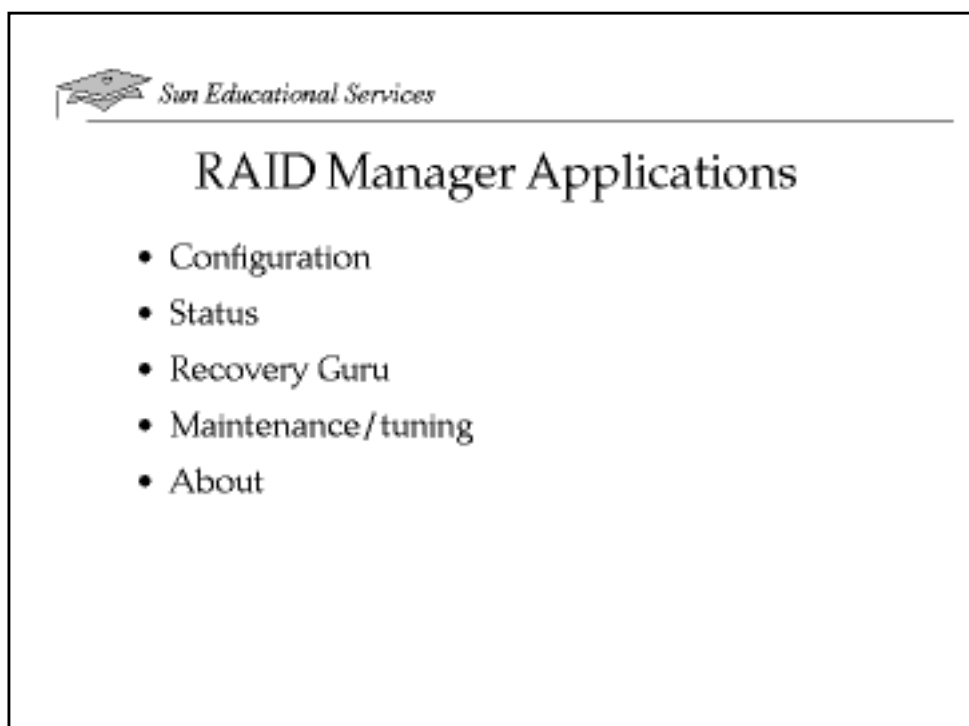
Caution – If you select *cache without batteries* and you do not have a uninterruptable power supply for protection, you could lose data if a power failure occurs.

Note – Caching should be enabled for performance reasons. Write cache mirroring should be enabled for data protection in the event of a controller failure. Caching can be controlled on a per LUN basis. Users should confirm the caching status for each newly created LUN.

Cache Memory

In summary:

- Cache memory is an area on the controller used for intermediate storage of read and write data.
- Cache memory can increase overall performance:
 - ▼ Data for read operations may already be in cache, eliminating the need to access the drive itself.
 - ▼ Write operations are considered complete once written to the cache.
- Write cache mirroring:
 - ▼ When enabled, writes cached data to the cache memory of both controllers so that when a controller fails, the second controller completes all outstanding write operations
- Cache without batteries:
 - ▼ Has the controller temporarily turn off caching when the batteries are low or completely discharged
 - ▼ Overrides this safeguard and continues to use caching



RAID Manager Applications

The RM6 GUI has four applications that can be initiated from icons. The applications are:

- Configuration
- Status
- Recovery Guru
- Maintenance/Tuning
- About

RAID Manager Applications

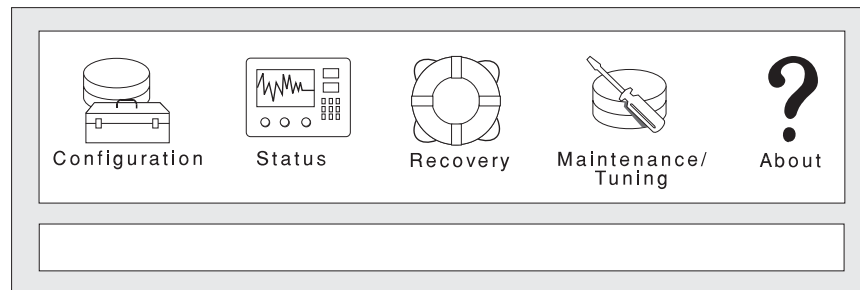


Figure 9-6 RAID Manager Top-level Display

Configuration

This application is primarily used to specify the configuration. Users can specify how physical drives in the array are to be allocated to logical units for data storage, and which RAID levels are to be used. Users can also specify which disks are to be configured as hot spares.

Status

This application permits an administrator to determine if an array has any abnormal or unusual status conditions associated with it. Three kinds of status information are available:

- Message log viewing — Permits browsing and detailed viewing of accumulated history information pertaining to array exception conditions.
- “On-demand” health checking — Examines selected arrays for any fault conditions that need to be remedied.
- Reconstruction status — Permits viewing of reconstruction progress for logical units that have had failed drives replaced.

RAID Manager Applications

Recovery Guru

This application assists an administrator in the process of carrying out recovery operations on degraded hardware. The Recovery Guru “knows” about certain failure modes and attempts to lead the user through the necessary recovery steps, ensuring that the user goes about replacing components in the right manner. Combinations of multiple failed components are taken into account, if necessary.

Maintenance/Tuning

This application provides control of certain array management tasks that arise from time to time in a storage array configuration. These tasks include downloading controller firmware, validating array parity, and tuning the controller cache.

About

When you click on this icon, an RM6 window is activated that returns the version number of the RM6 software.



Sun Educational Services

Command-Line Interface

- It is useful when writing shell scripts.
- CLI is not as intuitive as the GUI.
 - CLI command syntax is available through the GUI Task Monitor.
- CLI and GUI tools perform the same functions.
- All commands are located in `/usr/lib/osa/bin` and can be referenced through the symbolic link at `/usr/sbin/osa -> /usr/lib/osa/bin`

Command-Line Interface

A command-line interface is also available for writing shell scripts for commonly performed operations or for performing certain functions when the GUI interface may not be readily available.

The GUI is generally more intuitive and easy to use. It masks the underlying complexity and reduces the chance for operator error during administration.

Not all tasks can be performed identically by both interfaces.

All of the commands are found in the `/usr/lib/osa/bin` file and can be referenced through the symbolic link:
`/usr/sbin/osa -> /usr/lib/osa/bin`

Command-Line Interface

Table 9-1 RM6 Commands

Command	Description
drivutil	This drive/LUN utility is used to manage drives/LUNs. It enables you to obtain drive/LUN information, revive a LUN, fail/unfail a drive, and obtain LUN reconstruction progress.
fwutil	This controller firmware download utility downloads appware, bootware, Fibre Channel code, or a non-volatile storage, random access memory (NVSRAM) file to a specified controller.
healthck	This health check utility performs a health check on the indicated RAID module(s) and displays a report to standard output.
lad	This list array devices utility identifies which RAID controllers and logical units are connected to the system.
logutil	This log format utility formats the error log file and displays a formatted version to the standard output.
nvutil	This NVSRAM display/modification utility permits the viewing and changing of RAID controller NVRAM settings, allowing for some customization of controller behavior. It verifies and fixes any NVSRAM settings that are not compatible with the storage management software.
parityck	This parity check/repair utility checks, and if necessary, repairs the parity information stored on the array. (While correct parity is vital to the operation of the array, the possibility of damage to parity is extremely unlikely.)
raidutil	This RAID configuration utility is the command-line counterpart to the graphical configuration application. It permits RAID LUN and hot spare creation and deletion to be performed from a command line or script.
rdacutil	This redundant disk array controller, management utility permits certain redundant controller operations such as LUN load balancing and controller failover and restoration to be performed from a command line or script.
storutil	This host store utility is used to perform certain operations on a region of the controller called host store. You can use this utility to set an independent controller configuration, change RAID modules' names, and clear information in the host store region.



Sun Educational Services

Device Naming Conventions

- Standard device naming conventions
- Logical links made at `/dev/(r)dsk` and `/dev/osa/dev/(r)dsk` to the `/devices` file
- Solaris OS restrictions:
 - There is a maximum of 16 LUNs per RAID module.
 - Each LUN can be partitioned the same as any disk drive.
 - There is a maximum of 32 LUNs per host bus adapter.

Device Naming Conventions

Standard Device Names

The RAID Manager software uses device addresses to refer to logical units. These addresses are determined by the location of the subsystem hardware.

As shown in Figure 9-7, the address indicates the SCSI host controller, SCSI ID number of the controller, LUN, and the slice number.

The RAID Manager software uses this device name in various screen displays. This address usually indicates the path to a particular logical unit. If you transfer LUN ownership between controllers as part of a maintenance/tuning procedure (LUN balancing), the device name will be automatically updated in the RAID software. However, the Solaris OS will continued to use the original path until a reconfiguration boot (`boot -r`) is performed.

Device Naming Conventions

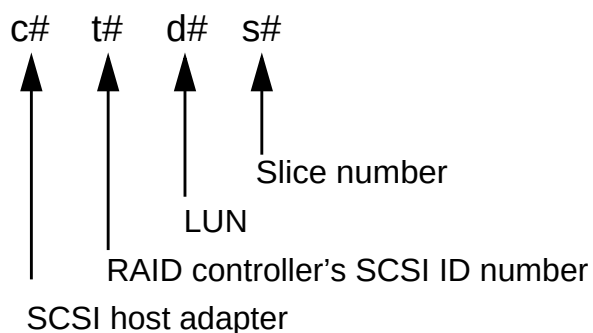


Figure 9-7 Naming Conventions

Figure 9-7 illustrates:

- Standard device naming conventions
- Logical links made at `/dev/(r)dsk` and `/dev/osa/dev/(r)dsk` to the `/devices` file
- Solaris OS restrictions:
 - ▼ Maximum 16 LUNs per RAID module
 - ▼ Each LUN can be partitioned the same as any disk drive
 - ▼ Maximum 32 LUNs per host bus adapter

Exercise: Reviewing RAID Manager Architecture



Exercise objective – In this exercise, you will:

- Answer questions related to the RAID Manager software architecture

Task

Answer the following questions:

1. If the host is connected to two controllers which are connected to the same set of drives, it is possible to distribute the load by:
 - a. Dividing LUNS within each drive group across the two controllers
 - b. Arranging for LUNS of similar types to exist on the same controller
 - c. Mirroring the cache
 - d. Associating each drive group to one of the controllers.
2. What four applications are present in the RM6 GUI? Which is used to create and delete LUNS?

Exercise: Reviewing RAID Manager Architecture

Task

3. Which GUI application enables you to recover a data redundant LUN when replacing a disk drive? What status would the LUN have before and immediately after the replacement?

4. How can the cache can be enabled?
- a. On a per drive group basis
 - b. On a controller basis
 - c. On a per LUN basis
 - d. On a per host basis
5. Can the cache be used if the cache battery is non-functional?

6. Cache write mirroring enhances:
- a. Performance
 - b. Security
 - c. Read performance only
 - d. There is no such thing as cache mirroring.

Exercise: Reviewing RAID Manager Architecture

Task

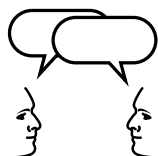
7. Can RM6 be used to perform an online backup?

8. Can RM6 be used to analyze performance of the drives?

9. Which of the following VM features have an equivalent under RM6?
 - a. Online volume growth
 - b. Striping
 - c. Concatenation
 - d. RAID 5
 - e. RAID 1
 - f. Hot spares
 - g. GUI
 - h. Dirty region log
10. In which directory are the RM6 commands found?

Exercise: Reviewing RAID Manager Architecture

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences
- Interpretations
- Conclusions
- Applications

Check Your Progress

Before continuing, check that you are able to accomplish or answer the following:

- ☐ Discuss the features and benefits of the RAID Manager software
- ☐ Define the terms
 - ▼ *Logical unit*
 - ▼ *Drive group*
 - ▼ *RAID module*
- ☐ Discuss hot spare usage
- ☐ Describe the data reconstruction process
- ☐ Describe RAID Manager device naming conventions
- ☐ Define caching control options

Think Beyond

You have a general understanding of what the RAID Manager software is and what it is used for, but how do you create drive groups, add and delete LUNs, create hot spares, and use the RM6 software features to recover from disk failures?

Sun StorEdge Volume Manager

Recovery Procedures

A 

This appendix is a summary of selected SSVM status and recovery information.

Summary

This appendix contains information that will help you:

- Check the status of volumes and plexes
- Configure debugging parameters
- Move data off a failing disk
- Replace a failed RAID 5 disk and recover the volume
- Replace a failed disk in a mirror and recover the volume
- Replace a failed SSVM disk
- Recover from boot problems

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The online manual pages for `luxadm` (1M), `vxdiskadm`(1M), `vxctl`(1M), `vxconfigd`(1M), `vxiod`(1M), and `vxmend`(1M).
- *Sun StorEdge Volume Manager 2.5 System Administrator's Guide*
- *Sun StorEdge Volume Manager 2.5 User's Guide*

Detecting Failed Physical Disks

You can identify a failed physical disk three ways:

- Have SSVM notify the administrator via email.
- Use the GUI to look at the status of physical disks.
- Use the `vxprint` command to display information and status.

When a physical disk fails, the SSVM disk associated with that disk also enters an error state. All volumes using the SSVM disk are affected.

If hot relocation is in use, the disk failure is detected (and recovered from) and the administrator is notified via email. If hot relocation is not enabled or you miss the email message, you can check disk status in two ways:

- Use the `vxprint` and `vxdisk list` commands and check the output for failed disks and volumes
- Use the GUI to look at the status of physical disks, SSVM disks, and volumes

Once the failed disk is identified, data may need to be recovered (or regenerated) for volumes using the failed disk. After the data recovery issues (discussed in the following pages) are addressed, the disk drive can be replaced.

Plex States

You must be able to determine plex state conditions in order to locate specific problems, errors, and failure conditions.

Plex states reflect whether or not plexes are complete and consistent copies of the volume contents. The SSVM utilities maintain the plex state.

A system administrator can modify the state of a plex to keep the volume from being modified. For example, if a disk with a particular plex on it begins to show aberrant behavior, the plex can be temporarily disabled.

Understanding plex states can help you:

- Identify whether the volume contents have been initialized to a known state
- Determine if a plex contains a valid copy of the volume contents
- Track whether a plex was in active use at the time of a system failure
- Monitor operations on plexes

Plex States

Table A-1 Plex State Definitions

Plex State	Definition
EMPTY	Volume creation sets all plexes associated with a volume to the EMPTY state to indicate the plex is not yet initialized.
CLEAN	A plex is in the CLEAN state when it is known to contain a consistent copy of the volume contents and an operation has disabled the volume. No action is required to guarantee data consistency.
ACTIVE	Two situations can cause a plex to be in the ACTIVE state: (1) when the volume is started and the plex is fully participating in I/O and (2) when the volume was stopped due to a system crash and the plex was active. When the volume is started, recovery procedures will update the plex contents.
STALE	The plex may not have the complete and current volume contents. If an I/O error occurs on a plex, the kernel stops using and updating the plex, placing it in the STALE state.
OFFLINE	The plex is detached (but still associated with the volume). If changes are made to the volume, this plex will go into the STALE state.
TEMP	This state enables some plex operations that cannot occur atomically. It is set by utilities during the course of some operations.
TEMPRM	This is similar to TEMP. Some subdisk operations require a temporary plex. When the operation is complete, the temporary plex is removed.
TEMPRMSD	This state is used when new plexes are being attached.
IOFAIL	This state is associated with persistent state logging. Most likely an I/O failure has occurred. The plex is disqualified from the recovery process.
LOG	The plex contains a dirty region log or a RAID-5 log. It is informational.

Volume States

Some of the generic volume states (Table A-2) are similar to plex states.

Table A-2 Generic Volume State Definitions

Generic Volume State	Definition
CLEAN	The volume is not started (kernel state is DISABLED) and its plexes are synchronized.
ACTIVE	The volume has been started (kernel state is currently ENABLED) or was in that state when the machine was rebooted. If the volume is currently ENABLED, the state of its plexes at any moment is not certain (the contents are being modified). If the volume is currently DISABLED, the plexes are not guaranteed to be consistent, but will be made consistent when the volume is started.
EMPTY	The volume contents are not initialized. The kernel state is always DISABLED when the volume is EMPTY.
SYNC	The volume is either in read-writeback recovery mode (kernel state is ENABLED) or was in the mode when the machine rebooted (kernel state is DISABLED). With read-writeback recovery, plex consistency is recovered by reading data from blocks of one plex and writing the data to all other writable plexes in the volume. If the volume is ENABLED, this means the plexes are being resynchronized using this read-writeback procedure. If the volume is DISABLED, the plexes were being resynchronized using the read-writeback procedure when the machine rebooted. In this case, the plexes still need to be resynchronized.
NEEDSYNC	The volume will require a resynchronization the next time it is started.

RAID-5 Volume States

RAID-5 volumes (Table A-3) have their own set of volume states.

Table A-3 RAID-5 Volume State Definitions

RAID-5 Volume State	Definition
CLEAN	The volume is not started (kernel state is DISABLED) and its parity is good. The RAID-5 plex stripes are consistent.
ACTIVE	The volume has been started (kernel state was ENABLED) or was in use when the machine rebooted. If the volume is currently ENABLED, the state of the plex is uncertain (the volume is in use). If the volume is currently DISABLED, the parity cannot be guaranteed to be synchronized.
EMPTY	The volume contents are not initialized. The kernel state is always DISABLED when the volume is EMPTY.
SYNC	The volume is either undergoing a parity resynchronization (kernel state is currently ENABLED) or was having its parity resynchronized when the machine rebooted (kernel state is DISABLED).
NEEDSYNC	The volume will require a parity resynchronization the next time it is started.
REPLAY	The volume is in a transient state as part of a log replay. A log replay occurs when it becomes necessary to use logged parity and data. (The RAID-5 volume is running in degraded mode.)

Moving Data From a Failing Disk

If a physical disk starts to behave strangely, you can move its data to another physical disk in the same disk group before the failure becomes a hard failure.

Before you proceed with disk evacuation, verify that the evacuation process is not going to create either of the following conflicts:

- Both volume mirrors on the same physical disk drive
- More than one stripe column of a striped or RAID-5 volume on the same disk drive

Preparing for an Evacuation

Before starting the evacuation:

- Find out what volume the failing plex is associated with and the name of the disks that are associated with it.
- Find out the disk group associated with the failing disk drive.
- Determine if there are any other volumes associated with the failing disk drive.
- Find a new disk with enough free space to perform the evacuation.
- Check for any volume conflicts associated with the new disk.

Performing an Evacuation

The evacuation process can be performed from VMSA as follows:

1. Select the disk that contains the objects and data to be moved.
2. Choose Disks ► Evacuate from the Selected menu.
3. Enter the destination disk in the Evacuate Disk dialogue box.

Note – You can also use `vxdiskadm` option 7 or the `vxevac` command directly to perform a disk evacuation.

Recovering a Volume

If a system crash or an I/O error corrupts one or more plexes of a volume and no plex is CLEAN or ACTIVE, mark one of the plexes CLEAN and instruct the system to use that plex as the source for repairing the others.



Caution – If neither plex is in a clean state, you cannot determine which one is the best to select. Both might be damaged. This is a last-resort procedure. You will probably lose data.

To place a plex in the CLEAN state, use the `vxmend` command. (For more information, see the online manual page for `vxmend` (1M).)

Sometimes a system crash or I/O error results in a volume with no CLEAN or ACTIVE plex. Use this procedure to place a plex in the CLEAN state and thereby instruct the system to use it as the source to repair other plexes.

1. Identify the failed/failing plex.
 2. Note the plex name.
-
3. Use the `vxmend` command to place the plex in the CLEAN state.

```
# vxmend fix clean plex-name
```

For example:

```
# vxmend fix clean vol01-02
```

Recovering a RAID-5 Volume (A5000)

This procedure assumes the failed SSVM disk contains only the RAID-5 subdisk that needs to be evacuated and that hot sparing and hot relocation have not (or cannot) take place. (A manual fix is required.) A more detailed example of moving a fully populated SSVM disk follows.

When the RAID-5 subdisk is evacuated to a new SSVM disk, the data for that subdisk is regenerated automatically using the data and parity from the other components of the RAID-5 volume. You would

1. Identify the SSVM disk that has failed.
2. Identify the physical disk containing the failed SSVM disk.
3. Unmount all file systems on disks in the tray with the faulty drive.
4. Move the failed RAID-5 subdisk from the failed SSVM disk to a good SSVM disk. (Evacuate the subdisk.)
5. Remove the failed SSVM disk.
6. Locate the relocated RAID-5 subdisk on the *new* SSVM disk. The data is regenerated using parity calculations.
7. Use `vxdiskadm` and `luxadm` commands to replace the failed disk.

Recovering a RAID-5 Volume (SPARCstorage Array)

This procedure assumes the failed SSVM disk contains only the RAID-5 subdisk that needs to be evacuated and that hot sparing and hot relocation have not (or cannot) take place. (A manual fix is required.)

Note – A more detailed example of moving a fully populated SSVM disk follows.

When the RAID-5 subdisk is evacuated to a new SSVM disk, the data for that subdisk is regenerated automatically using the data and parity from the other components of the RAID-5 volume. You would

1. Identify the SSVM disk that has failed.
2. Identify the physical disk containing the failed SSVM disk.
3. Unmount all file systems on disks in the tray with the faulty drive.
4. Stop all database processes that are accessing disks in the tray with the faulty drive.
5. Stop all other processes that are accessing disks in the tray with the faulty drive.
6. Flush or purge any outstanding writes from non-volatile random access memory (NVRAM), if necessary.
7. Move the failed RAID-5 subdisk from the failed SSVM disk to a good SSVM disk. (Evacuate the subdisk.)
8. Remove the failed SSVM disk.
9. Locate the relocated RAID-5 subdisk on the *new* SSVM disk. The data is regenerated using parity calculations.
10. Spin down the drives in the tray containing the failed disk.
11. Replace the failed disk.
12. Spin up the drives in the tray containing the replaced disk.

Recovering a Mirror (A5000)

This procedure assumes the failed SSVM disk contains only the subdisk from the mirrored volume that needs to be evacuated and that hot sparing and hot relocation have not (or cannot) take place. (A manual fix is required.)

Note – A more detailed example of moving a fully populated SSVM disk follows.

SSVM does not regenerate data in this case unless you reconfigure another plex and attach it to the mirror—at which point the “new” submirror is fully synchronized. You would:

1. Identify the SSVM disk that has failed.
2. Identify the physical disk containing the failed SSVM disk.
3. Unmount all file systems on disks in the tray with the faulty drive.
4. Remove the plex containing the failed subdisk from the volume.
5. Remove the failed SSVM disk.
6. Using `vxdiskadm` and `luxadm`, replace the failed disk.
7. Create a new plex to replace the failed plex.
8. Attach the new plex to the volume. The plex is fully synchronized.

Note – Sometimes `vxdiskadm` will not start and you must use `vxmend` to clear the `putil` and `tutil` fields (`vxmend clear putil all`).

Recovering a Mirror (SPARCstorage Array)

This procedure assumes the failed SSVM disk contains only the RAID-5 subdisk that needs to be evacuated and that hot sparing and hot relocation cannot take place. (A manual fix is required.)

Note – A more detailed example of moving a fully populated SSVM disk follows.

SSVM does not regenerate data in this case unless you reconfigure another plex and attach it to the mirror—at which point the “new” submirror is fully synchronized. You should:

1. Identify the SSVM disk that has failed.
2. Identify the physical disk containing the failed SSVM disk.
3. Stop all database processes that are accessing disks in the tray with the faulty drive.
4. Stop all other processes that are accessing disks in the tray with the faulty drive.
5. Unmount all file systems on disks in the tray with the faulty drive.
6. Flush or purge any outstanding writes from NVRAM, if necessary.
7. Remove the plex containing the failed subdisk from the volume.
8. Remove the failed SSVM disk.
9. Spin down the drives in the tray containing the failed disk.
10. Replace the failed disk.
11. Spin up the drives in the tray containing the replaced disk.
12. Create a new plex to replace the failed plex.
13. Attach the new plex to the volume. The plex is fully synchronized.

Note – Sometimes `vxdiskadm` will not start and you must use `vxmend` to clear the `putil` and `tutil` fields (`vxmend clear putil all`).

Replacing a Failed SSVM Disk (A5000)

This is perhaps the most likely scenario to be faced in the field. When a SSVM disk fails, each volume which makes use of the SSVM disk will be in an error state. If hot sparing and hot relocation cannot be accomplished (not activated or spare disks are not available), the volumes can be repaired by hand.

Once the SSVM disk is identified, each subdisk on the SSVM disk must be examined and relocated, if possible. You should:

1. Identify the SSVM disk that has failed.
2. Identify the physical disk containing the failed SSVM disk.
3. Unmount all file systems on disk with the faulty drive.
4. Check each subdisk on the failed SSVM disk:
 - a. If the subdisk is part of a mirror, remove the plex containing the failed subdisk.
 - b. If the subdisk is part of a RAID-5 volume, evacuate the subdisk to a known good SSVM disk.
 - c. If the subdisk is part of a simple or striped volume, remove the volume.
 - d. If the subdisk is a logging subdisk, remove the log from the volume.
5. Remove the failed SSVM disk.
6. Replace the failed disk using `vxdiskadm` and `luxadm` utilities.
7. For each mirror affected by the failed SSVM disk:
 - a. Create a new plex to replace the failed plex.
 - b. Attach the new plex to the volume. The plex is fully synchronized.

Replacing a Failed SSVM Disk (SPARCstorage Array)

This is another scenario that is likely to be faced in the field. When a SSVM disk fails, each volume which makes use of the SSVM disk will be in an error state. If hot sparing and hot relocation cannot be accomplished (not activated or spare disks are not available), the volumes can be repaired by hand.

Once the SSVM disk is identified, each subdisk on the SSVM disk must be examined and relocated, if possible. You would:

1. Identify the SSVM disk that has failed.
2. Identify the physical disk containing the failed SSVM disk.
3. Unmount all file systems on disks in the tray with the faulty drive.
4. Stop all database processes that are accessing disks in the tray with the faulty drive.
5. Stop all other processes that are accessing disks in the tray with the faulty drive.
6. Flush or purge any outstanding writes from NVRAM, if necessary.
7. Check each subdisk on the failed SSVM disk:
 - a. If the subdisk is part of a mirror, remove the plex containing the failed subdisk.
 - b. If the subdisk is part of a RAID-5 volume, evacuate the subdisk to a known good SSVM disk.
 - c. If the subdisk is part of a simple or striped volume, remove the volume.
 - d. If the subdisk is a logging subdisk, remove the log from the volume.

Replacing a Failed SSVM Disk (SPARCstorage Array)

8. Remove the failed SSVM disk.
9. Spin down the drives in the tray containing the failed disk.
10. Replace the failed disk.
11. Spin up the drives in the tray containing the replaced disk.
12. For each mirror affected by the failed SSVM disk:
 - a. Create a new plex to replace the failed plex.
 - b. Attach the new plex to the volume. The plex is fully synchronized.

Booting After a Failure – Booting From a Mirror

You can boot from a mirror if:

- The boot disk is mirrored and under SSVM control (root is encapsulated).
- The boot disk fails.

To boot from a mirror of the boot disk:

1. Check for aliased SSVM disks using the `devalias` command at the OpenBoot™ prompt.

Note – Disks that are suitable mirrors of the boot disk will be listed with the name `vx-medianame`, where *medianame* represents the disk media name for the disk containing the boot disk mirror.

2. Boot using the alias name.

ok **boot** *alias-name*

If the root disk is mirrored, an alternate boot disk can be used to boot the system if the primary boot disk fails.

To boot the system from a mirror of the boot disk:

- Check for aliased SSVM disks using the `devalias` command at the OpenBoot prompt. Disks that are suitable mirrors of the boot disk will be listed with the name `vx-medianame`, where *medianame* represents the disk media name for the disk containing the boot disk mirror.
- Boot using the alias name.

If a selected disk contains a root mirror that is stale, `vxconfigd` will display an error message stating that the mirror is unusable and it will list any nonstale alternate disks from which to boot.

Once the system boots, the boot disk and its mirrors can be repaired.

Replacing a Failed Boot Disk

If the boot disk is under SSVM control and the boot disk fails, it needs to be replaced.

The first step is to boot the system from an alternate boot disk (such as a mirror of the boot disk) or boot device.

If the failed disk is not detached from its device, manually detach it using the `vxdiskadm` command. Use the Remove a Disk for Replacement function in addition to the `luxadm remove_device` utility.

The replacement disk must be at least as large as the failed disk to ensure it can hold all the information required.

Using the `luxadm` utility to complete the disk replacement, boot the system. Run the `vxdiskadm` command and select the Replace a Failed or Removed Disk function. This replaces the failed disk with the new device that was just added. Then:

1. Boot the system from an alternate boot device (or a mirror of the boot disk, if the boot disk was mirrored using SSVM).
2. Detach the failing disk.

Use the Remove a Disk for Replacement function of the `vxdiskadm` command.

3. Shut down the system and replace the failed disk.
4. Boot the system.
5. Partition the new boot disk and do a dump/restore from the surviving mirror.

Note – Do not use the `vxdiskadm` Replace a Failed or Removed Disk function. This will create a partitionless boot disk that looks just like the mirror. You can never again boot directly from a partition on the primary boot disk. You must use the dump/restore technique to avoid this problem.

Moving a Storage Array to Another Host

Use this procedure to move a storage array managed by SSVM to another host.

Note – This procedure assumes that the rootdg disk group is also on the storage array. If it is not, only a simple disk group import is needed.

In this example, SSVM is running on Host_A with one storage array connected. Host_A fails, and you want to move the storage array to another host, Host_B. Use these steps:

1. Install the SSVM software on Host_B, if necessary.
2. Remove the install-db file in
/etc/vx/reconfig.d/state.d
3. Disconnect the storage array from Host_A and connect it to Host_B.
4. Perform a reconfiguration reboot to build the device tree.
5. Start SSVM on Host_B.

```
# vxiod set 10
# vxconfigd -m disable
```

6. On Host_B, remove the old host name and add the new host name to the configuration.

```
# vxdctl init Host_A
```

Note – This temporarily changes the host name in the volboot file of Host_B to the hostname of Host_A.

```
# vxdctl enable
# vxdctl hostid Host_B
```

Note – This changes the hostname in both the volboot file and the array disk private regions back to the new host, Host_B.

```
# vxdctl enable
```


Sun StorEdge Volume Manager

Boot Disk Encapsulation

B 

This appendix summarizes the prerequisites and the process for encapsulating a system boot disk.

Summary

This appendix contains information that will help you

- Identify the optimum boot disk configuration
- List the boot disk encapsulation prerequisites
- Perform the steps necessary to encapsulate a system boot disk
- Verify copies have been made of all important pre-encapsulation configuration files
- Boot from the boot disk mirror

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- The online manual pages for `luxadm(1M)`, `vxdiskadm(1M)`, `vxctl(1M)`, `vxconfigd(1M)`, `vxiod(1M)`, and `vxmend(1M)`.
- *Sun StorEdge Volume Manager 2.5 System Administrator's Guide*
- *Sun StorEdge Volume Manager 2.5 User's Guide*

Boot Disk Encapsulation Overview

When you install the Sun StorEdge Volume Manager software on a system you can place your system boot disk under SSVM control in two different ways:

- Using the `vxinstall` program during the initial software installation
- Using the VMSA interface after the initial installation

Preferred Boot Disk Configuration

Although there are many possible boot disk variations, this appendix focuses on the preferred boot disk configuration shown in Figure B-1.

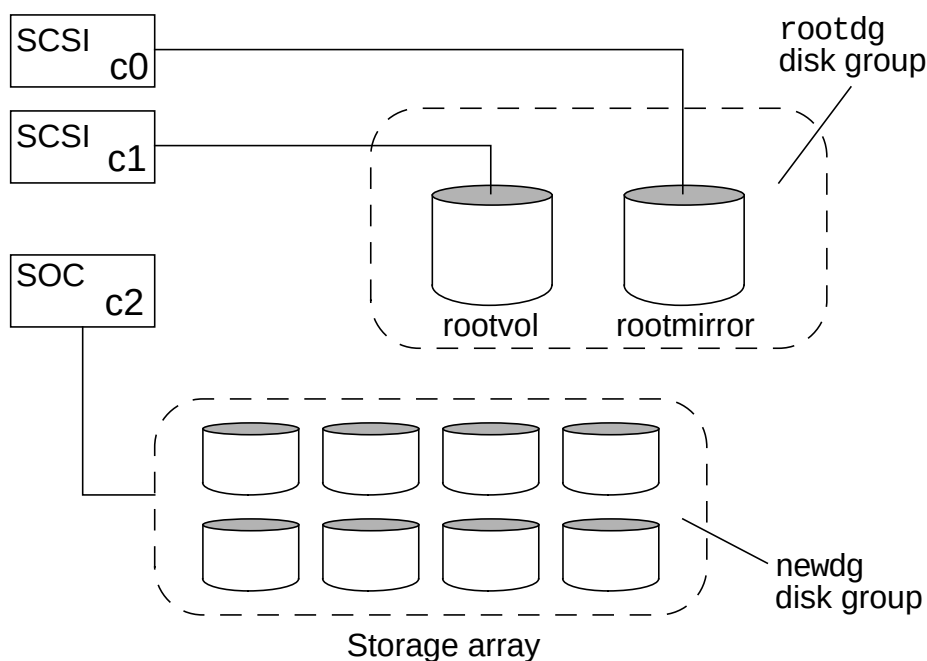


Figure B-1 Preferred Boot Disk Configuration

The preferred configuration has the following features:

- The boot disk and mirror are on separate interfaces.
- The boot disk and mirror are not in a storage array.
- Only the boot disk and mirror are in the `rootdg` disk group.

Boot Disk Encapsulation Overview

Prerequisites for Boot Disk Encapsulation

In order for the boot disk encapsulation process to succeed, the following prerequisites must be met:

- The disk must have at least two unused slices
- The boot disk must not have any slices in use other than the following:
 - ▼ root
 - ▼ swap
 - ▼ var
 - ▼ opt
 - ▼ usr

An additional prerequisite that is desirable but not mandatory is there should be at least 1024 sectors at the beginning or end of the disk. This is needed for the private region. SSVM will take the space from the end of the swap partition if necessary.

Primary and Mirror Configuration Differences

When you encapsulate your system boot disk, the location of all data remains unchanged even though the partition map is modified.

When you mirror the encapsulated boot disk, the location of the data on the mirror is probably very different from the original boot disk.

During encapsulation, a copy of the system boot disk partition map is made so that the disk can be returned to a state that allows booting directly from a slice.

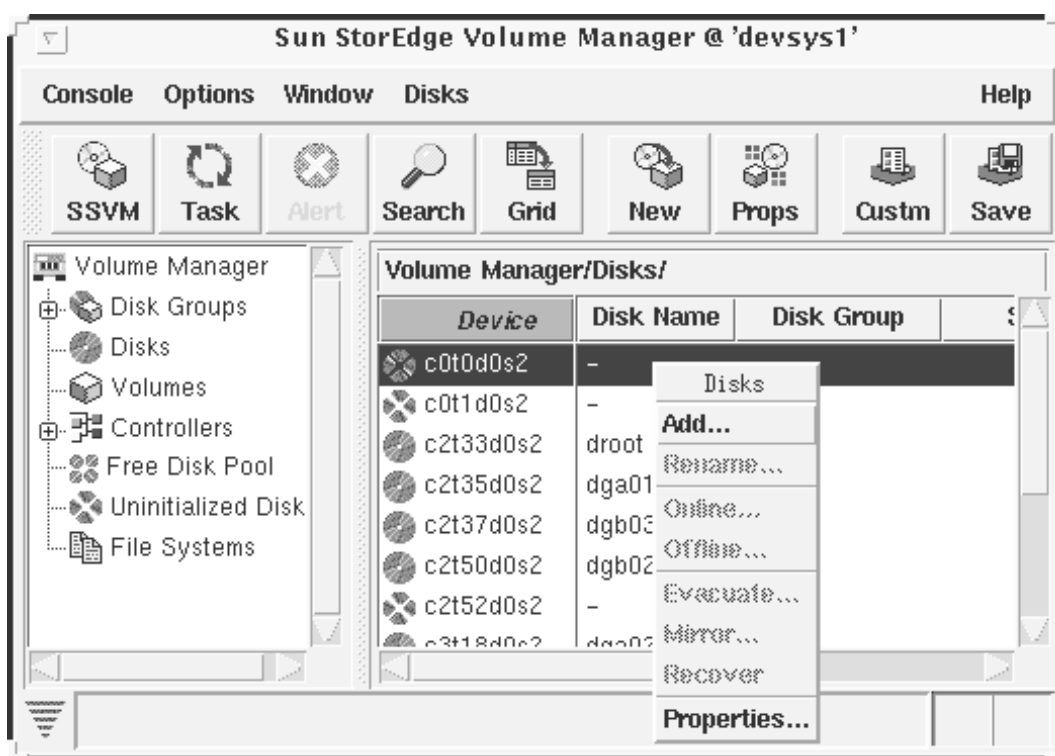
The mirror of the boot disk cannot be returned to a sliced configuration. You must boot from its associated SSVM device.

Encapsulating the Boot Disk Using VMSA

The boot disk encapsulation process is easy to perform using the VMSA graphical interface. The process is exactly the same as adding a new disk to a disk group except that the SSVM software is “aware” that you are adding a disk that has mounted file systems. The software is also “aware” that this is the system boot disk.

The following steps are part of a typical encapsulation process:

1. Highlight the system boot disk in the Grid area and select Add from the pop-up menu.



Encapsulating the Boot Disk Using VMSA

2. Enter the appropriate information in the Add Disk form.

Add Disk(s) @ 'devsys1'

Scan Disks

Disk Device(s): c0t0d0s2 Browse...

Add Disk(s) to:

- Existing Disk Group
- New Disk Group
- Free Disk Pool

Select/Create Disk Group

Disk Group Name: rootdg Browse...

☐ Create Shared

Options:

Disk Name(s): root01

Comment: Primary Boot Disk

Ok Apply Cancel Help

3. Select the desired reboot operation.

Add Disk

These disks may be in use:
c0t0d0s2 has a file system.

You may encapsulate the disks. This will preserve the data on the disks as new volumes. This requires a system reboot to take effect.

Begin Encapsulation, do not reboot

Encapsulate and reboot now

Disk Initialization is not available for disks with mounted partitions or swap areas.

Initialize

Cancel

Encapsulating the Boot Disk Using VMSA

Note – Until the system is rebooted, no changes are made. All the necessary information has been stored in the /etc/vx directory. It might be more convenient for you to complete the initial portion of the encapsulation process and wait until later to do the system reboot.

4. Ignore the following misleading message. No modifications have yet been made to the system boot disk.

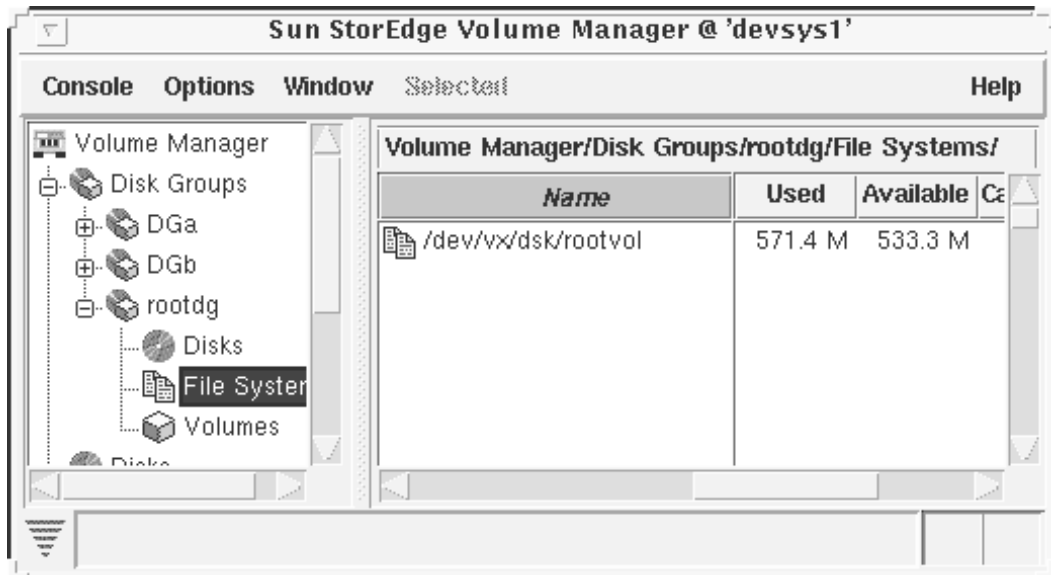


5. When you are ready, reboot your system and verify the following messages are displayed:

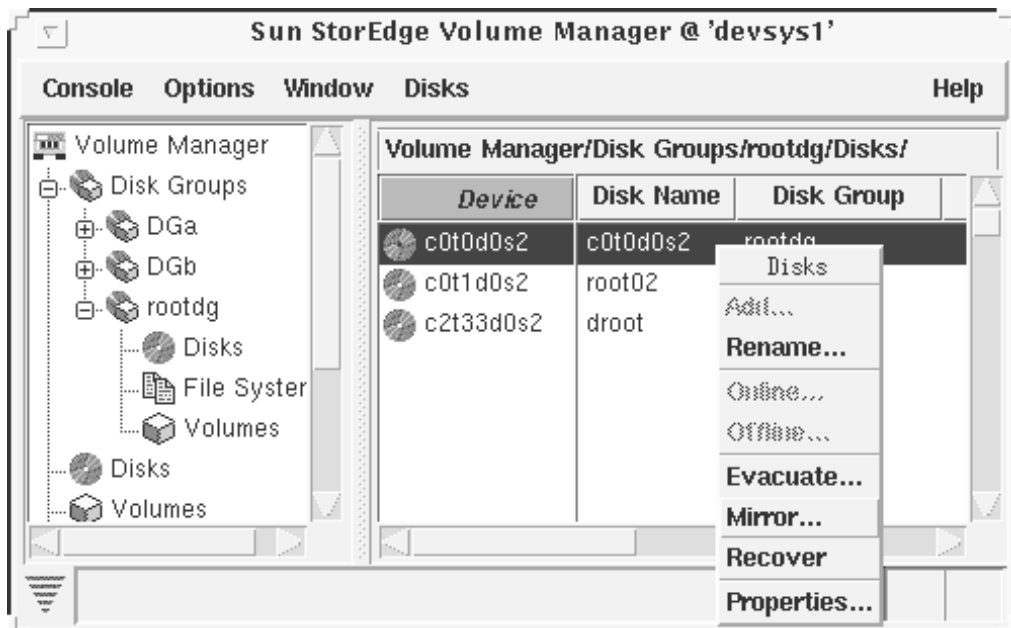
```
VxVM starting in boot mode...
configuring network interfaces: hme0.
Hostname: devsys1
VxVM starting special volumes (swapvol)...
VxVM general startup...
vxvm: NOTE: Setting partition /dev/dsk/c0t0d0s1 as the dump device.
  Dump content: kernel pages
  Dump device: /dev/dsk/c0t0d0s1 (dedicated)
Savecore directory: /var/crash/devsys1
  Savecore enabled: yes
```

Encapsulating the Boot Disk Using VMSA

6. Verify the rootvol volume has been created successfully.



7. Highlight the system boot disk again and select Mirror from the pop-up menu.



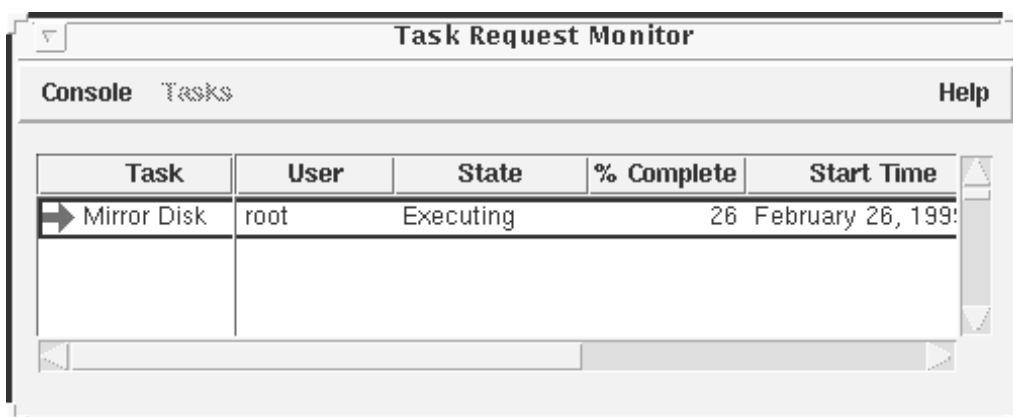
Encapsulating the Boot Disk Using VMSA

8. Enter the mirror target disk name in the Mirror Disk form or use the Browse button.



The dialog box is titled "Mirror Disk @ 'devsys1'". It contains two input fields: "Disk Name:" with the text "rootdg/c0t0d0s2" and "Target Disk(s):" with the text "root02". Each input field has a "Browse..." button to its right. At the bottom of the dialog are four buttons: "Ok", "Apply", "Cancel", and "Help".

9. Open the Task Request Monitor and observe the progress of the mirroring operation.



The window is titled "Task Request Monitor". It has a tabbed interface with "Console" and "Tasks" tabs, and a "Help" button in the top right corner. The "Tasks" tab is active, showing a table with the following data:

Task	User	State	% Complete	Start Time
➡ Mirror Disk	root	Executing	26	February 26, 199:

Note – The mirroring operation can take quite a while, depending on how large the volumes are. Each of the five possible volumes will be mirrored in order.

Encapsulating the Boot Disk Using VMSA

10. Verify the status of the boot disk volumes with the `vxprint` command.

```
# vxprint -g rootdg
```

TY	NAME	ASSOC	KSTATE	LENGTH	PLOFFS	STATE
dg	rootdg	rootdg	-	-	-	-
dm	c0t0d0s2	c0t0d0s2	-	8378640	-	-
v	rootvol	root	ENABLED	2458080	-	ACTIVE
pl	rootvol-02	rootvol	ENABLED	2458080	-	ACTIVE
sd	root02-01	rootvol-02	ENABLED	2458080	0	-
pl	rootvol-01	rootvol	ENABLED	2458080	-	ACTIVE
sd	c0t0d0s2-B0	rootvol-01	ENABLED	1	0	-
sd	c0t0d0s2-02	rootvol-01	ENABLED	2458079	1	-
v	swapvol	swap	ENABLED	369360	-	ACTIVE
pl	swapvol-02	swapvol	ENABLED	369360	-	ACTIVE
sd	root02-02	swapvol-02	ENABLED	369360	0	-
pl	swapvol-01	swapvol	ENABLED	369360	-	ACTIVE
sd	c0t0d0s2-01	swapvol-01	ENABLED	369360	0	-

Encapsulation Files

A number of files are used during the boot disk encapsulation process.

Files in the /etc/vx Directory

The following files are created when the boot disk is first encapsulated but before the system is rebooted:

- /etc/vx/disks-cap-part

This file contains only the path `c0t0d0` and points to which device is to be reconfigured during the system reboot.

- /etc/vx/reconfig.d/disk.d/c0t0d0/newpart

This file contains the new partitioning and the SSVM commands that will be used during the system reboot.

```
# volume manager partitioning for drive c0t0d0
0 0x2 0x200      0 2458080
1 0x3 0x201 2458080 369360
2 0x5 0x200      0 8380800
3 0xe 0x201      0 8380800
4 0xf 0x201 8378640 2160
5 0x0 0x000      0 0
6 0x0 0x000      0 0
7 0x0 0x000      0 0
# vxmake vol rootvol plex=rootvol-%00 usetype=root logtype=none
# vxmake plex rootvol-%00 sd=c0t0d0s2-B0,c0t0d0s2-%00
# vxmake sd c0t0d0s2-%00 disk=c0t0d0s2 offset=0 len=2458079
# vxmake sd c0t0d0s2-B0 disk=c0t0d0s2 offset=8378639 len=1 putil0=Block0
comment="Remap of block 0
# vxvol start rootvol
# rename c0t0d0s0 rootvol
# vxmake vol swapvol plex=swapvol-%01 usetype=swap
# vxmake plex swapvol-%01 sd=c0t0d0s2-%01
# vxmake sd c0t0d0s2-%01 disk=c0t0d0s2 offset=2458079 len=369360
# vxvol start swapvol
# rename c0t0d0s1 swapvol
```

Encapsulation Files

Files in the /etc/vx Directory

- /etc/vx/reconfig.d/disk.d/c0t0d0/vtoc

This file contains the original vtoc listing of the boot disk. The vtoc listing will be similar to the following one:

#THE PARTITIONING OF /dev/rdisk/c0t0d0s2 IS AS FOLLOWS:

#SLICE	TAG	FLAGS	START	SIZE
0	0x2	0x200	0	2458080
1	0x3	0x201	2458080	369360
2	0x5	0x200	0	8380800
3	0x0	0x000	0	0
4	0x0	0x000	0	0
5	0x0	0x000	0	0
6	0x0	0x000	0	0
7	0x0	0x000	0	0

- /etc/vx/reconfig.d/state.d/init-cap-part

This is a temporary state file that will be removed after the encapsulation process has completed.

- /etc/vx/reconfig.d/state.d/reconfig

This is a temporary state file that will be removed after the encapsulation process has completed and will be replaced with blank file named root-done.

- /etc/vx/reconfig.d/saveconf.d/etc/system

This is a copy of the original /etc/system file.

Encapsulation Files

The /etc/vfstab File

A backup copy of the /etc/vfstab file is made before the new boot disk path names are configured. This /etc/vfstab file is typical for a boot disk with a single partition root file system.

#device #to mount #	device to fsck	mount point	FS type	fsck pass	mount at	mount bootoptions
fd	-	/dev/fd	-	no	-	-
/proc	-	/proc	proc	-	no	-
/dev/vx/dsk/swapvol	-	-	swap	-	no	-
/dev/vx/dsk/rootvol	/dev/vx/rdisk/rootvol	/	ufs	1	no	-
swap	-	/tmp	tmpfs	-	yes	-
#						
#NOTE: volume rootvol (/) encapsulated partition c0t0d0s0						
#NOTE: volume swapvol (swap) encapsulated partition c0t0d0s1						

Boot PROM Changes

When the system boot disk is encapsulated, you can no longer boot directly from a boot disk partition. The SSVM software creates two new boot aliases for you so that you can boot from the primary system boot disk, or if a failure occurs, from the surviving mirror. You can examine the new boot aliases as follows:

```
# eeprom | grep devalias
devalias vx-rootdisk /sbus@1f,0/SUNW,fas@e,88000000/sd@1,0:a
devalias vx-rootmir /sbus@1f,0/SUNW,fas@e,88000000/sd@0,0:a
```

If your primary boot disk fails, you can boot from the surviving mirror as follows:

```
ok boot vx-rootmir
```

Un-Encapsulating the Boot Disk

About the only time you might want to un-encapsulate the system boot disk is if you are removing the SSVM software. Before using the `vxunroot` command to un-encapsulate the boot disk, make sure the following actions have been taken:

- All boot disk volumes have been unmirrored.
- All non-root file systems, volumes, plexes, and subdisks have been removed.

If you forget to prepare the boot disk, the `vxunroot` command performs a very thorough check before starting.

The `vxunroot` command performs these basic functions:

- Checks for any unacceptable structures on the boot disk
- Returns the boot disk partition map to its original state
- Returns the `/etc/system` file to its original state
- Returns the `/etc/vfstab` file to its original state
- Returns the OpenBoot PROM device aliases to their original state

Sun StorEdge Volume Manager and RAID Manager

C 

This appendix provides an overview of using RAID Manager software in conjunction with Sun StorEdge Volume Manager.

Summary

This appendix contains information that will help you:

- List the advantages of utilizing SSVM with the RAID Manager software used to configure the A3000 and A1000 storage systems
- Describe the supported configurations for SSVM + RAID Manager

Additional Resources



Additional resources – The following references can provide additional details on the topics discussed in this module:

- Sun Microsystems, Inc. Sun Microsystems Computer Company. *Sun StorEdge RAID Manager 6.1.1 User's Guide*.
- Sun Microsystems, Inc., Sun Microsystems Computer Company. *Sun StorEdge RAID Manager 6.1.1 Installation and Support Guide for Solaris*.
- Storage Product Business Unit. *SSVM/RM6/Solaris How to Make Sense of It All*.

SSVM and RAID Manager

The RAID Manager software is used to configure A3000 and the A1000 storage components. Sun StorEdge Volume Manager may or may not be used “on top of” the RM software. The RM software is required for these two storage units, the SSVM is optional. There are a few terms which are specific to the RM and warrant explanation:

- RAID module – A set of drives, a set of controller(s), applicable power supplies, and fans. For example, a unit with five drive trays, each containing seven drives, and two controllers would be considered a RAID module.
- Drive group – A physical set of drives in the RAID module. These groups are identified during configuration. There are three types:
 - ▼ An unassigned drive group (not been configured into LUNs or hot spares)
 - ▼ A hot spare drive group (identified as hot spares)
 - ▼ A configured drive group (configured into one or more LUNs with the same RAID level).
- Logical unit – The basic structure you create on the RAID modules to store and retrieve data. A LUN spans one or more drives and is configured into RAID 0, 1, 3, or 5. More than one LUN may reside within a drive group and all LUNs in the same drive group share the same physical drives and RAID level.

Determining What Is Seen by the System

How are logical units viewed by the system? To determine what controllers and LUNs are attached to a system, use the following command:

```
# /etc/raid/bin/lad  
c1t5d0s0 1T62549100 LUNS: 0 1 2 6
```

SSVM and RAID Manager

Determining What Is Seen by the System

The Solaris OS “sees” the A1000 and A3000 disk arrays as logical disks. Each LUN that the `format` utility “sees” can be made up of one or more disks. The previous output of the `lad` utility lists four LUNs. If you ran `format`, listings for `c1t5d0`, `c1t5d1`, `c1t5d2`, and `c1t5d6` would be displayed.

Under SSVM, you can use the `vxdisk list` command to view the association between the physical device, for example, `c1t5d0`, to the SSVM disk, `disk01`.

Installing Sun StorEdge Volume Manager

Follow all of the documentation regarding installation, including any release notes, for the storage, RM software, and SSVM.

The Sun StorEdge Volume Manager should be installed only after the following steps have been completed and validated:

- The Sun StorEdge A3000 and A1000 disk arrays are properly attached to the host computer.
- The Sun StorEdge A3000 and A1000 RAID Manager software is properly installed.
- The LUNs are properly configured using RAID Manager.
- The host system was rebooted using `-r` to rescan for new devices.
- Upon reboot, the RAID Manager software has recognized the configured A3000/A1000 devices and has created the appropriate Solaris OS device nodes.
- The SSVM software has been installed.

Note – SSVM volumes configured using devices from A3000/A1000 cannot be a part of the default disk group, `rootdg`. Configure these devices to non-`rootdg` disk groups.

SSVM and RAID Manager

Using Sun StorEdge Volume Manager With RAID Manager

When SSVM is used in conjunction with RAID Manager, an increase in availability, performance, and manageability can be obtained.

- The SSVM performance statistics capability can be used to monitor the activity on the volumes and re-allocate storage if deemed necessary. Each LUN within the A3000/ A1000 looks to both the Solaris OS and SSVM as a single physical device. Due to this structure, it is difficult to determine performance on a spindle basis, however, learning of volume activity and potential imbalances between the LUNs can be handled efficiently through the Volume Manager.
- The size of a single file system or database tablespace is limited by the maximum size of a single LUN in a controller-based RAID product such as the A3000 or A1000. Used in conjunction with SSVM, different LUNs can be combined to a larger “multi-LUN” volume.
- Data movement between LUNs is made easier with SSVM. If SSVM was not in use, any LUN reconfiguration requires interruption of data access. If using SSVM, data may be copied through SSVM as a mirror. The advantage is all writes are delivered to all mirrors, keeping all copies up-to-date at all times. Once the mirror resync is completed, one of the mirrors (plexes) can be removed to free up the associated LUN(s). The users will not suffer any data loss.
- Backup can also be enhanced by using the previous method to sync a copy of the data to be backed up or by using SSVM’s snapshot utility to accomplish the same task – which results in a minimal amount of data access interruption.
- SSVM can partition a LUN into many subdisks that can be as small as a sector. Without SSVM, partitioning of the LUN is subject to the same restrictions as the Solaris OS – eight partitions.

SSVM and RAID Manager

Determining Supported Configurations

Mirroring

Some of the supported configurations are:

- SSVM, used to mirror non-mirrored LUNs provides:
 - ▼ Data redundancy
 - ▼ Data migration (online movement of data between LUNs)
- SSVM, used to mirror multiple A3000s and/or A1000s provides:
 - ▼ Centralized storage management
 - ▼ Data migration ability
- SSVM three-way mirroring provides:
 - ▼ Three-way mirroring on the A3000/ A1000 disk arrays, which is not possible without SSVM
 - ▼ Snapshot for backups
 - ▼ Data migration
 - ▼ Use of one plex as a consistent and stable backup source
- SSVM mirroring across A3000/ A1000 LUNs configured as stripes provides:
 - ▼ Better mirror performance by off-loading the stripe breakup off of the storage subsystem (A3000 or A1000 disk array)
 - ▼ Improved mirror performance because the host can pass larger data transfers through the hosts' drivers, as opposed to a series of smaller stripe units based on data transfers
- SSVM mirroring across A3000/ A1000 LUNs configured as RAID 5 through RAID Manager provides:
 - ▼ Data redundancy
 - ▼ Centralized storage management

SSVM and RAID Manager

Determining Supported Configurations

Striping

Some of the supported configurations are:

- SSVM striping across LUNs configured as RAID 5 provides:
 - ▼ Improved performance through striping
 - ▼ Centralized storage management
- SSVM striping across multiple A3000/ A1000 subsystems provides:
 - ▼ Improved performance through striping
 - ▼ Centralized storage management
- SSVM striping across A3000/ A1000 LUNs configured as mirrors provides:
 - ▼ Improved performance through striping
 - ▼ Faster mirror resynchronization recovery time
 - ▼ Lower exposure to data loss with loss of a plex in a mirror (better redundancy)

Determining Unsupported Configurations

A configuration that is not supported is:

- An SSVM RAID-5 and A3000/ A1000 RAID Manager RAID-5 configuration provides:
 - ▼ Poor performance without gaining any data reliability or availability

SSVM and RAID Manager

Using SSVM Hot Relocation and RAID Manager Hot Sparing

The hot sparing ability through the RAID Manager software allows the storage system to automatically react to I/O failures *internal* to the array box to restore access to a LUN. SSVM reacts to failures through the host side and if appropriate disk space is available, will recover redundant volumes.

If a disk failure occurs on the A3000/A1000, the array hot sparing—if enabled—will provide the disk failure redundancy. If the data redundancy is provided through SSVM (in RAID-5 or mirrored volumes only), then SSVM hot relocation can provide disk or partial disk failure redundancy protection.

The most complete solution is to implement hot sparing through the RAID Manager and then hot relocation through SSVM.

The Veritas VxFS File System

D 

This appendix summarizes the major features and characteristics of the Veritas file system (VxFS).

Summary

This appendix contains information that will help you:

- Identify the features of the VxFS
- Define how the VxFS intent logging feature alleviates the need for frequent file system checking
- Use the VxFS `fsadm` utility to defragment disks on demand
- Expand active file systems

Additional Resources



Additional resources – The following reference can provide additional details on the topics discussed in this module:

- *VERITAS File System (VxFS) System Administrator's Guide*, Release 2.2 (March 1996).

Introduction to VxFS

VxFS is an extent-based, intent-logging file system intended for use with Solaris 7 OS. It provides enhancements that increase Solaris 7 OS usability and better equip the UNIX system to handle commercial environments where high performance and availability are important and large volumes of data must be managed. A few of the VxFS features are:

- Fast file system recovery
- Online system administration
- Online backup
- Enhanced file system performance
- Extent-based allocation

These are discussed in the following sections. This appendix concludes with a discussion of the VxFS disk layout.

Fast File System Recovery

The UNIX file system relies on full structural verification by the `fsck` command to recover from a system failure. This means checking the entire structure of a file system, verifying that it is intact, and correcting any inconsistencies that are found. This can be very time-consuming.

The VxFS file system provides recovery only seconds after a system failure by using a tracking feature called intent logging. *Intent logging* is a logging scheme that records pending changes to the file system structure. During system recovery from a failure, the intent log for each file system is scanned and operations that were pending are completed. The file system can then be mounted without a full structural check of the entire system. When the disk has a hardware failure, the intent log may not be enough to recover and in such cases, a full `fsck` check must be performed, but often, when failure is due to software rather than hardware, a system can be recovered in seconds.

Online System Administration

A VxFS file system can be defragmented and resized while it remains online and accessible to users.

Defragmentation

The UFS file system uses the concept of cylinder groups to limit fragmentation. These are self-contained sections of a file system that are composed of inodes, data blocks, and bitmaps that indicate free inodes and data blocks. Allocation strategies in UFS attempt to place inodes and related data blocks near each other. This reduces fragmentation, but does not eliminate it. Over time, the original ordering of free resources can be lost and as files are added and removed, gaps between used areas of disk can still occur.

The VxFS file system provides a utility called `fsadm` to defragment a disk without requiring that the disk be unmounted first. It can be run on demand and should be scheduled as a regular `CRON` job. It removes unused space from directories, makes small files contiguous, and consolidates free blocks for use.

Resizing

In UFS file systems, when a file system becomes too small or too large for its assigned portion of disk, there are three things that can be done:

- Users can be moved to new or different file systems.
- Subdirectories of a file system can be moved to other file systems.
- An entire file system can be backed up and then restored to a re-sized file system.

VxFS in conjunction with SSVM enables a file system to be expanded while being accessed.

Online Backup

The VxFS file system provides a method for performing online backups of data using the “snapshot” feature. An image of a mounted file system is created by mounting another file system, which then becomes an exact read-only copy of the first file system. The original file system is said to be snapped, and the copy is called the snapshot.

Enhanced File System Performance

Standard UFS file systems use block-based allocation schemes and provide good random access to files and reasonable latency on small files. For larger files, however, this block-based architecture limits throughput. The VxFS file system improves file system performance by using a different allocation scheme and by providing increased user control over allocation and I/O and caching policies. The following VxFS features provide this improved performance:

- Extent-based allocation
- Enhanced mount options
- Data-synchronous I/O
- Direct I/O
- Caching advisories
- Enhanced directory features
- Explicit file alignment, extent size, and preallocation controls

Extent-based allocation is described in the following section.

Extent-based Allocation

Disk space is allocated by the system in 512-byte sectors, which are grouped together to form a logical block. Logical blocks can be 1024, 4096, or 8192 bytes. The default is 1024.

An *extent* is defined as one or more adjacent blocks of data within the file system. It is presented as an address-length pair, which identifies the starting block address and the length of the extent (in blocks). When storage is added to a file on a VxFS system, it is grouped in extents, as opposed to being allocated a block at a time (as is done with UFS file systems).

By allocating disk space in extents, disk I/O to and from a file can be done in units of multiple blocks. This type of I/O can occur if storage is allocated in units of consecutive blocks. For sequential I/O, multiple-block operations are considerably faster than block-at-a-time operations. Almost all disk drives accept I/O operations of multiple blocks.

Extent allocation makes the interpretation of addressed blocks from the inode structure only slightly different from that of block-based inodes. The UFS inode references data in block sizes, whereas the VxFS inode references data in extents, which may be multiple blocks. Otherwise, the UFS inode contains the addresses of 12 direct blocks, 1 indirect block, and 1 double-indirect block. The VxFS inode contains addresses of 10 direct extents and 2 indirect-address extents. The first indirect-address extent is used for single indirection; the second is used for double indirection.

Disk Layout

This section describes the structural elements of the file system that exist in fixed locations on the disk.

Figure D-1 illustrates the basic VxFS Version 2 disk layout.

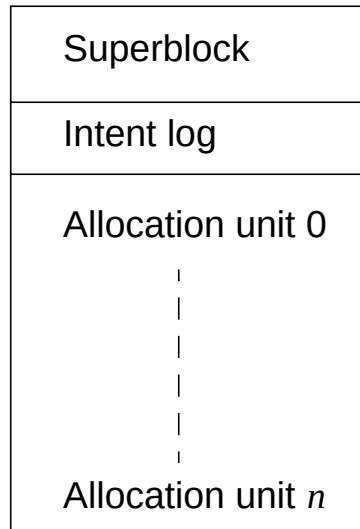


Figure D-1 VxVFS Disk Layout

The disk is composed of:

- The superblock
- The object-location table
- The intent log
- A replica of the object location table
- One or more allocation units

Disk Layout

Superblock

The superblock contains important information about the file system, such as:

- File system type
- Creation and modification dates
- Label information
- Information about the size and layout of the file system
- Count of available resources
- File system, disk-layout version number
- Pointers to the object-location table and its replica

The superblock is always in a fixed location, offset from the start of the file system by 8192 bytes. This fixed location enables utilities to easily locate the superblock when necessary. The superblock is 1024 bytes long.

Copies of the superblock are kept in allocation-unit headers. These copies can be used for recovery purposes if the superblock is corrupted or destroyed.

Object-Location Table

The object-location table can be considered an extension of the superblock. It contains information used at mount time to locate file system structures that are not in fixed locations. It is typically located immediately after the superblock and is 8 Kbytes long.

The object-location table is replicated and its replica is located immediately after the intent log. This separation of original and replica minimizes the potential for losing both copies of the information in the event of localized disk damage.

Disk Layout

Intent Log

The intent log is a circular activity log with a default size of 512 blocks. If the file system is less than 4 Mbytes, the log size will be reduced to avoid wasting space. The intent log contains records of the intention of the system to update a file system structure. An update to the file system structure (a transaction) is divided into separate subfunctions for each data structure that needs to be updated. A composite log record of the transaction which contains the subfunctions that constitute the transaction is created.

The intent log contains records for all pending changes to the file system structure, and insures that the log records are written to disk in advance of the changes to the file system. Once the intent log has been written, the transaction's other updates to the file system can be written in any order. In the event of a system failure, the pending changes to the file system are either nullified or completed by the `fsck` utility. The intent log generally only records changes to the file system structure. File-data changes are not normally logged.

During system recovery, the existence of this log makes it possible for recovery to occur much more quickly than if the entire disk structure had to be checked and validated by the `fsck` command, as is the case with standard UFS file systems.

Disk Layout

Allocation Unit

An allocation unit is a group of consecutive blocks in a file system that contain a resource summary, a free-resource map, data blocks, and a copy of the superblock. An allocation unit is similar in concept to the UFS cylinder group. Each component of an allocation unit begins on a block boundary. All of the Version 2 allocation-unit components deal with the allocation of disk space. Those components of the Version 1 allocation unit that deal with inode allocation have been relocated elsewhere for Version 2. In particular, the inode list now resides in an inode-list file and the inode allocation information now resides in an inode-allocation unit. The allocation unit is illustrated in Figure D-2.

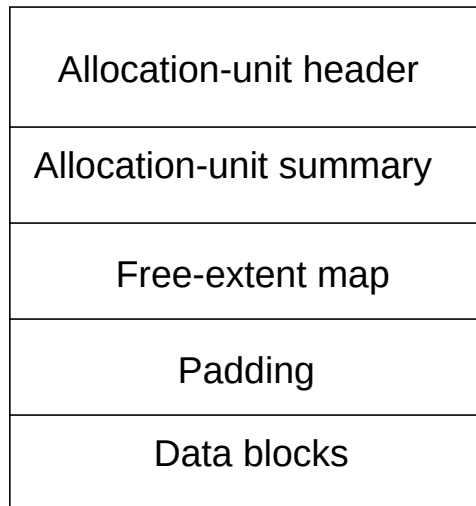


Figure D-2 Version 2 Allocation

One or more allocation units exist per file system. Allocation units are located after the object-location table replica. The number and size of allocation units can be specified when the file system is made. All of the allocation units, except possibly the last one, are of equal size. If space is limited, the last allocation unit can have a partial set of data blocks to allow use of all remaining blocks.

Disk Layout

Allocation Unit

Allocation-Unit Header

The allocation-unit header contains a copy of the file system's superblock that is used to verify that the allocation unit matches the superblock of the file system. The superblock copies contained in allocation-unit headers can also be used for recovery purposes if the superblock is corrupted or destroyed. The allocation unit-header occupies the first block of each allocation unit.

Allocation-Unit Summary

The allocation-unit summary summarizes the resources (data blocks) used in the allocation unit. This includes information on the number of free extents of each size in the allocation unit and a flag indicating the status of the summary.

Free-Extent Map

The free-extent map is a series of independent 512-byte bitmaps that are each referred to as a free extent map section. Each section is broken down into multiple regions. The first region of 2048 bits represents a section of 2048 one-block extents. The second region of 1024 bits represents a section of 1024 two-block extents. This regioning continues for all powers of 2 up to the single bit that represents one 2048-block extent.

The one-block bitmaps always represent the true allocation of blocks from the allocation unit. The remaining bitmaps remap these same blocks, in a "binary-buddy" scheme, in increasingly larger sized groups. As smaller extents are needed, the larger groups of blocks mapped by the buddy maps are broken apart to create the smaller extents.

Disk Layout

Allocation Unit

Padding

It may be desirable to align data blocks to a physical boundary. To facilitate this, the system administrator can specify that a gap be left between the end of the free-extent map and the first data block.

Data Blocks

The balance of the allocation unit is occupied by data blocks. Data blocks contain the actual data stored in files and directories.

Objectives

Upon completion of this appendix, you should be able to:

- Create a drive group
- Add a LUN to an existing drive group
- Create a hot spare pool
- Delete a LUN
- Recover from a failure such as a failed RAID set

Relevance



Discussion – The following questions are related to understanding the content of this appendix:

- How does adding a LUN to an existing drive group differ from adding a LUN while creating a new drive group?
- Why should hot spare spools be distributed across I/O busses?
- How does the Recovery Guru aid you in RAID set recovery?
- What is the key concern when deleting LUNs?

Additional Resources



Additional resources – The following reference can provide additional details on the topics discussed in this module:

- <http://docs.sun.com/ab2/@DSCBrowse?storage=1&currentsubject=Hardware>



Sun Educational Services

Starting RM6

The applications that can be selected using RM6 software are:

- Configuration
- Status
- Recovery
- Maintenance/tuning
- About (RM6)

Starting RM6

The RAID Manager (RM6) is started from the command line using the following command:

```
# rm6 &
```

Note – If your PATH variable has not been updated since the RAID Manager software installation, the fully qualified pathname is /usr/sbin/osa/rm6.

The applications that can be selected using RM6 software are:

- Configuration
- Status
- Recovery
- Maintenance / tuning
- About (RM6)

Starting RM6

Figure E-1 illustrates the layout of the selectable icons on the RAID Manager GUI.

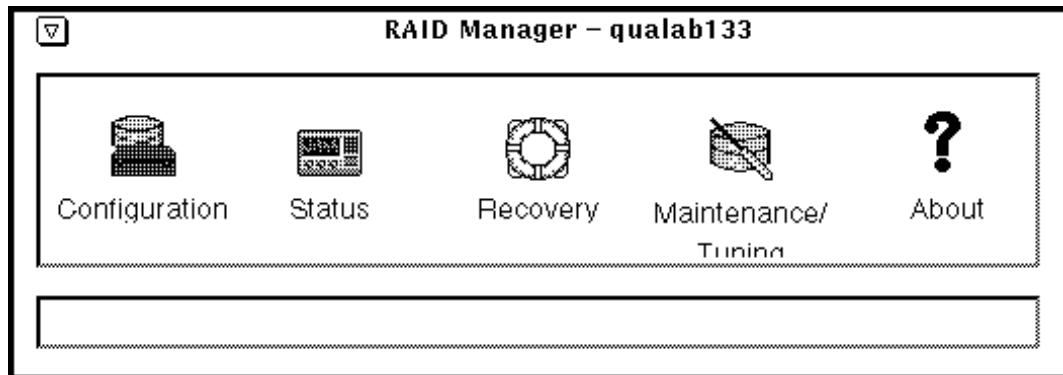


Figure E-1 RAID Manager Top-level Screen

The applications can be described as follows:

- Configuration
 - ▼ List/locate drives
 - ▼ Create LUN
 - ▼ Create hot spare
 - ▼ Delete drive groups, LUNs, or hot spares
- Status
 - ▼ Message log
 - ▼ Health check
 - ▼ LUN reconstruction

Starting RM6

- Recovery
 - ▼ Recovery Guru
 - ▼ Manual parity check and repair
 - ▼ Manual recovery
- Maintenance / Tuning
 - ▼ LUN reconstruction rate
 - ▼ LUN balancing
 - ▼ Controller mode
 - ▼ Caching parameters
 - ▼ Firmware upgrade options
 - ▼ Automatic parity
- About
 - ▼ Software version information

Note – The configuration issues and use of the Recovery Guru are addressed in this appendix. The remaining (maintenance) topics are addressed in the SM-250: *Sun StorEdge Configuration and Troubleshooting* course.



Sun Educational Services

Creating a Drive Group

To create a new drive group, you must:

- Open the Configuration window.
- Select unassigned drives.
- Configure the drive group options.
- Confirm the option selection.
- Wait for the drive group RAID configuration to complete.

Creating a Drive Group

To create a new drive group, you must:

- Open the configuration window.
- Select unassigned drives.
- Configure the drive group options.
- Confirm the option selection.
- Wait for the drive group RAID configuration to complete.

Creating a Drive Group

You select the configuration application from the RAID Manager top-level window.

1. Click on the Configuration icon.

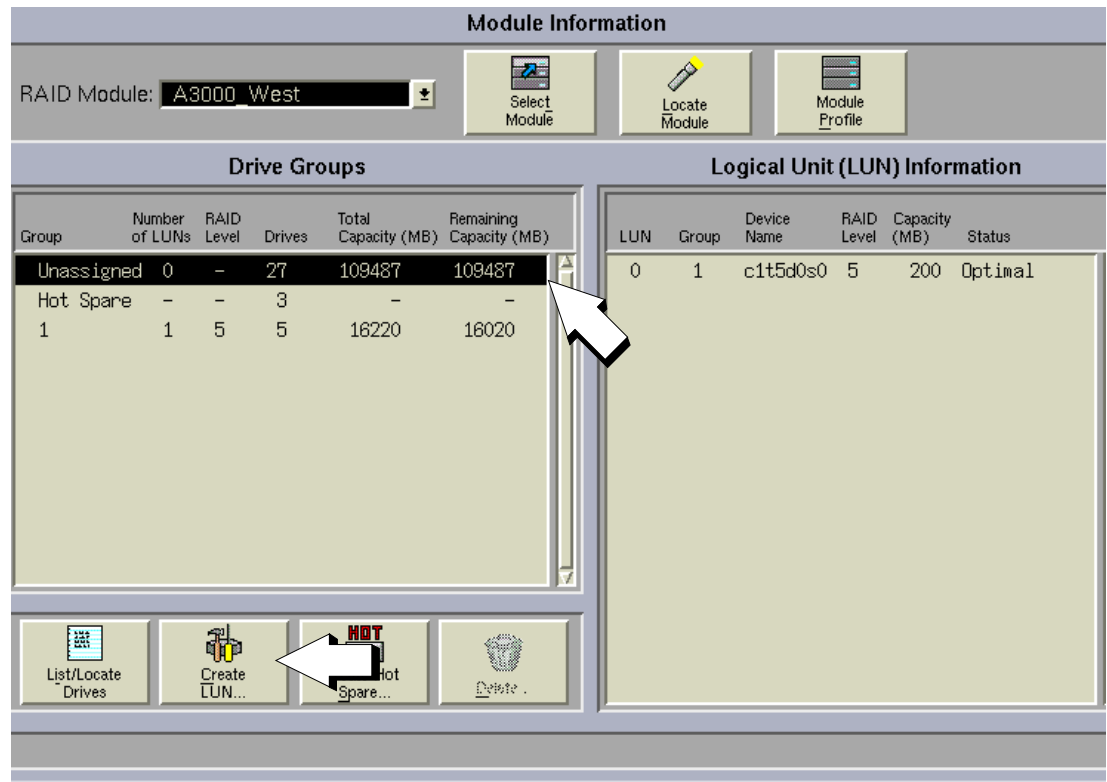
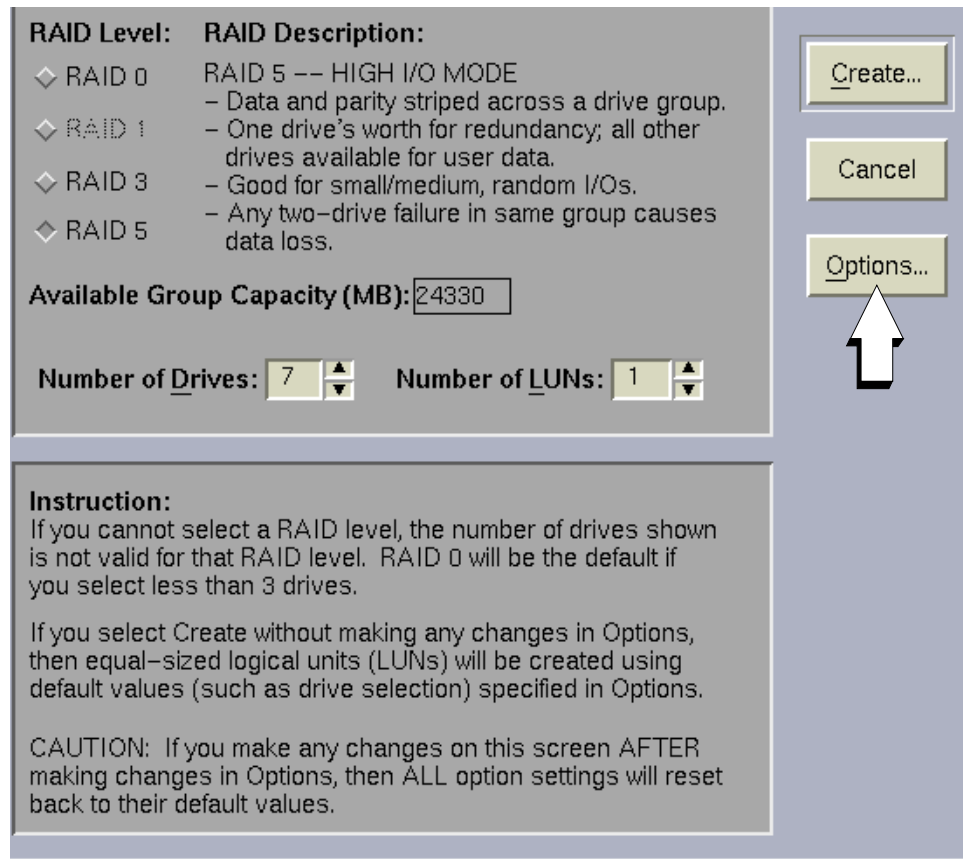


Figure E-2 Create a New Drive Group

2. Click on the unassigned disk drives.
3. Click on the Create LUN gadget.

Creating a Drive Group

4. Select the desired RAID level for this drive group.



RAID Level:	RAID Description:
◇ RAID 0	RAID 5 -- HIGH I/O MODE
◇ RAID 1	- Data and parity striped across a drive group.
◇ RAID 3	- One drive's worth for redundancy; all other drives available for user data.
◇ RAID 5	- Good for small/medium, random I/Os.
	- Any two-drive failure in same group causes data loss.

Available Group Capacity (MB):

Number of Drives: Number of LUNs:

Instruction:
If you cannot select a RAID level, the number of drives shown is not valid for that RAID level. RAID 0 will be the default if you select less than 3 drives.

If you select Create without making any changes in Options, then equal-sized logical units (LUNs) will be created using default values (such as drive selection) specified in Options.

CAUTION: If you make any changes on this screen AFTER making changes in Options, then ALL option settings will reset back to their default values.

Figure E-3 Assigning Drive Group RAID Level

5. Select the desired number of disk drives to include in this drive group.

Note – A drive group can contain a maximum of 20 disk drives.

6. Select the number of LUNs to create.
7. Click on the Options box to further define the new drive group's parameters.

Creating a Drive Group

There are five option windows. Clicking on OK from any of these windows returns you to the Create Drive Group window, and any options that are not modified will use the default values.

8. Click on LUN Capacity (if it is not already selected).

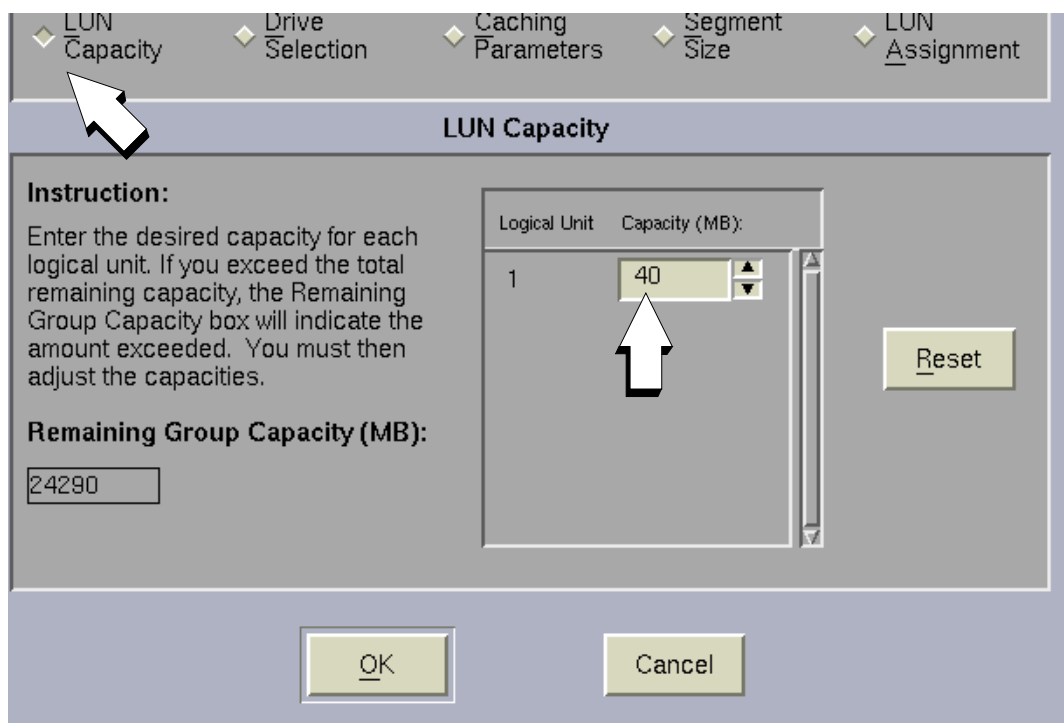


Figure E-4 LUN Capacity Option Specification

9. Type in the desired capacity of this LUN.

Note – As each LUN's storage capacity is allocated, the remaining drive group capacity decreases by a like amount.

From this window, the next logical choice is the Drive Selection window.

Creating a Drive Group

10. Click on Drive Selection.

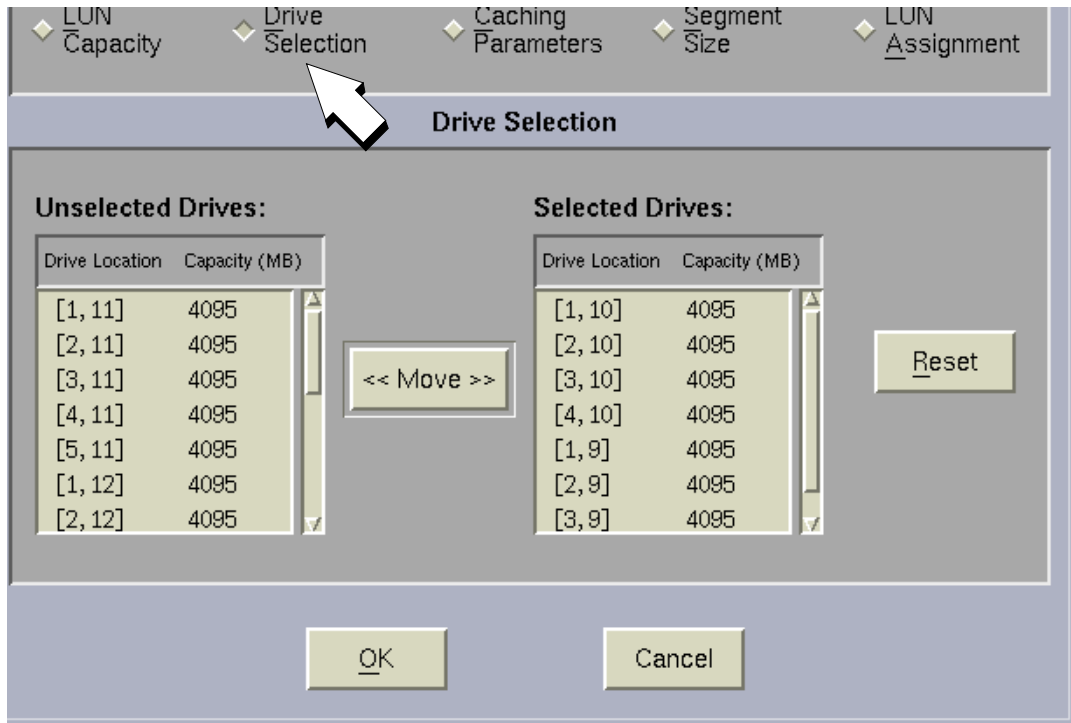


Figure E-5 Drive Selection Window

The RM6 software attempts to distribute the disks evenly among the available (unassigned) disks. This window provides you with the capability to adjust this selection.

You can move disks between the selected and the unselected columns by clicking on the disk and then clicking on the Move gadget.

The next option is the caching parameters.

Creating a Drive Group

11. Click on Caching Parameters.

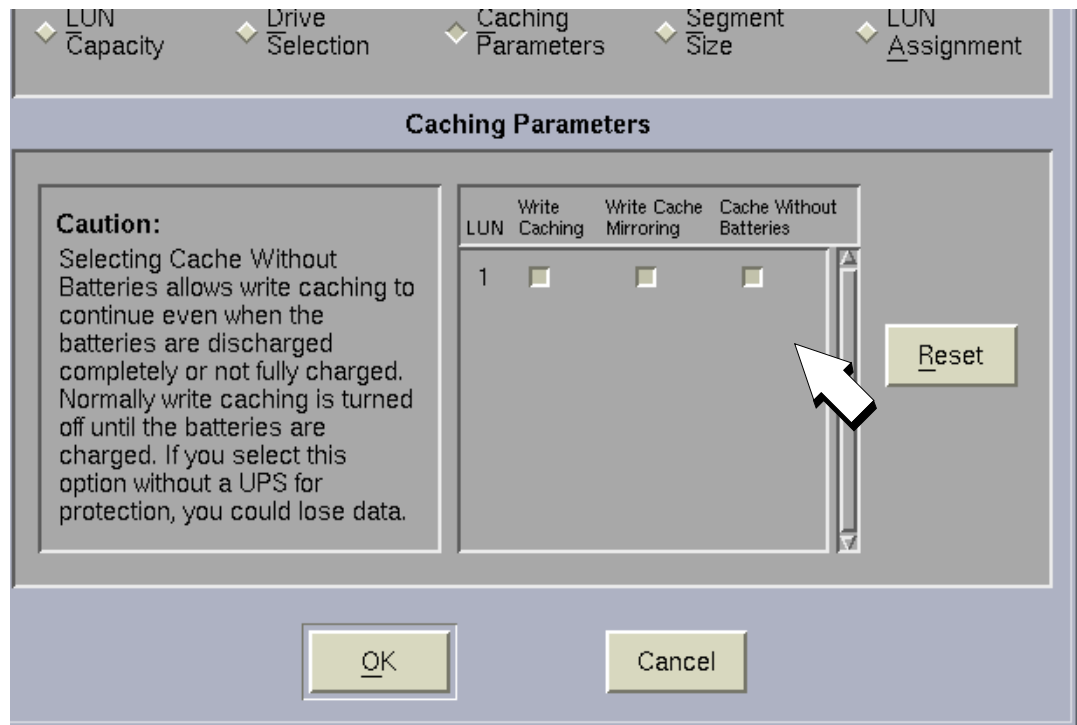


Figure E-6 Caching Parameters Window

Use this option to view or modify three caching parameters for LUNs on a selected RAID module:

- **Write Caching** – Enables write operations from the host to be stored in the controller's cache memory. The use of write caching increases overall performance because a write operation from the host machine is considered completed once it is written to the cache.

Creating a Drive Group

- Write Cache Mirroring – Enables cached data to be mirrored across two redundant controllers with the same size cache. The data written to the cache memory of one controller is also written to the cache memory of the other controller. Therefore, if one controller fails, the other can complete all outstanding write operations.
- Cache Without Batteries – Enables write caching to continue even if the batteries are discharged completely, not fully charged, or not present. If you select this option without a UPS for additional protection, you could lose data if a power failure occurs.

The next option is the segment size.

Creating a Drive Group

12. Click on Segment Size.

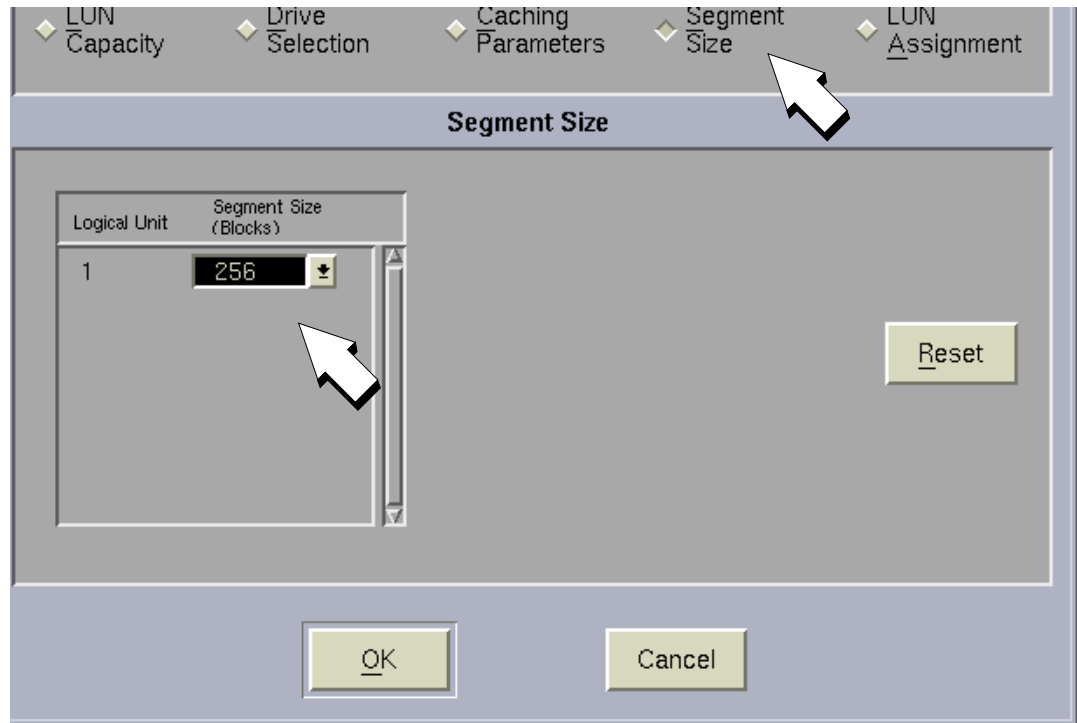


Figure E-7 Segment Size Window

A segment is the amount of data the controller writes on a single drive in a LUN before writing data on the next drive. The segment size is composed of blocks; one block equals 512 bytes. Normally, you should use the default segment size shown because the values provided are based on the RAID level specified for the drive group/LUNs.

The last option parameter is the LUN assignment.

Creating a Drive Group

13. Click on LUN Assignment.

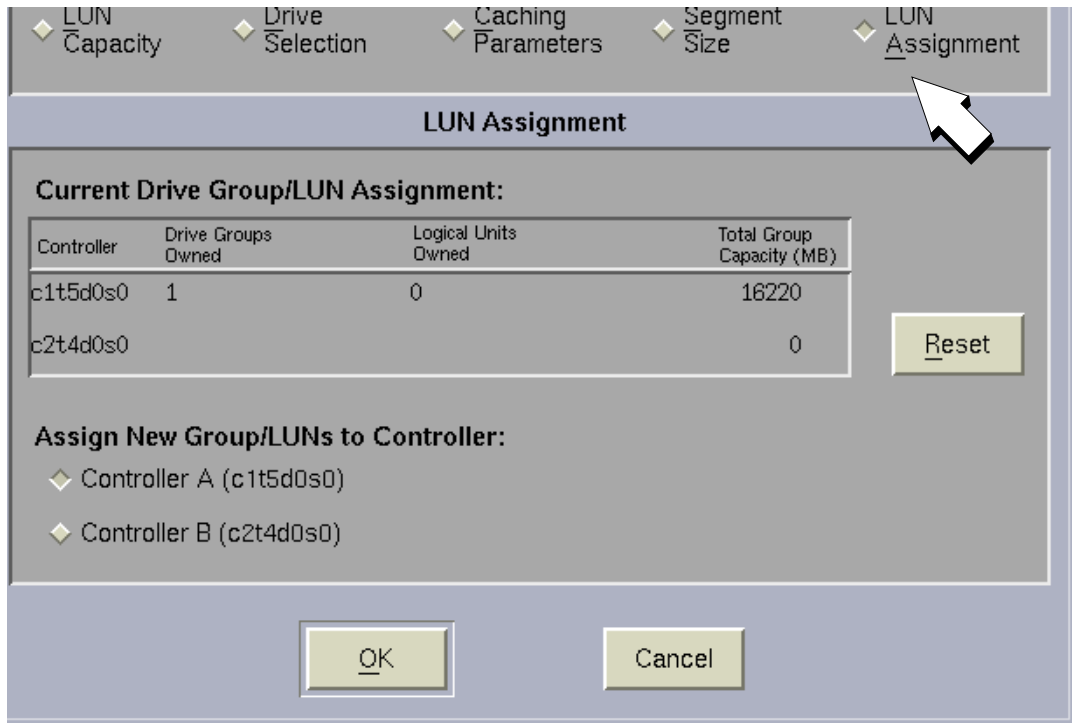


Figure E-8 LUN Capacity Window

This window enables you to change which controller owns the new drive group/LUN(s) you create.

Note – This option is dimmed if there are not two active controllers in the RAID module, if you are creating additional LUNs on an existing drive group, or if the module has an independent controller configuration.

Creating a Drive Group

The display shows you which controller owns the current drive groups/LUNs.

Normally, you should use the default controller selected under the Assign New Group/LUNs To Controller area. The only reason to change the default is to be sure that a particular controller owns a specific drive group/LUNs.

The capacity shown is the total capacity available on the drive group. It is not the total capacity of the LUNs configured on the drive group unless the LUNs have used all of the capacity.

Unless you use this option, the LUNs are balanced across active controller pairs on a drive group basis. The odd-numbered drive groups are assigned to one active controller and the even-numbered drive groups are assigned to the other active controller.

14. When you are satisfied with the LUN option parameter settings, click on the OK gadget on any of the option screens to return to the Create LUN window.

Creating a Drive Group

15. Click on the Create button.

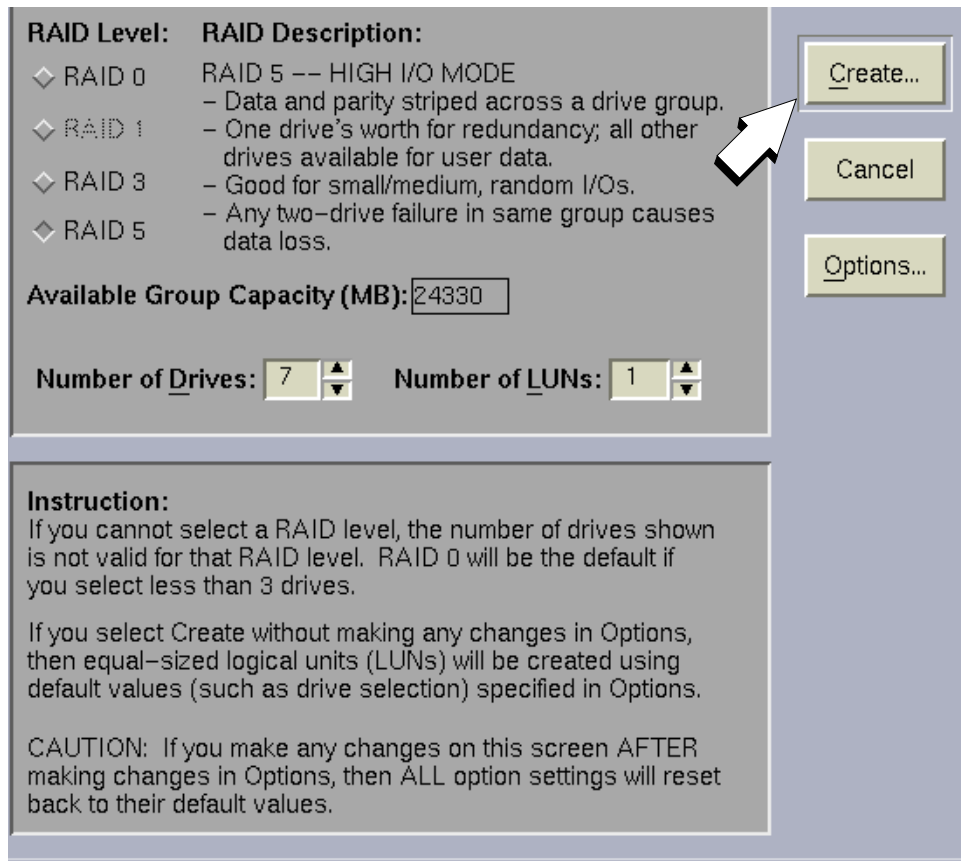


Figure E-9 Create LUN Window

16. Click on the OK button in the Confirmation window.

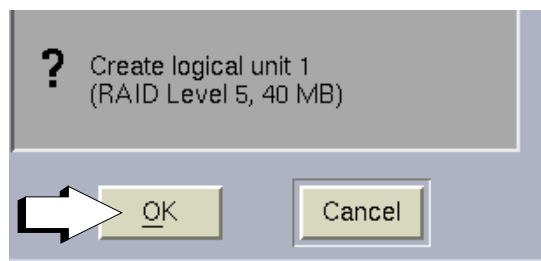


Figure E-10 Create LUN Confirmation Window

Creating a Drive Group

You are returned to the (Configuration) Module Information window, where you can observe the formatting status during the LUN initialization.

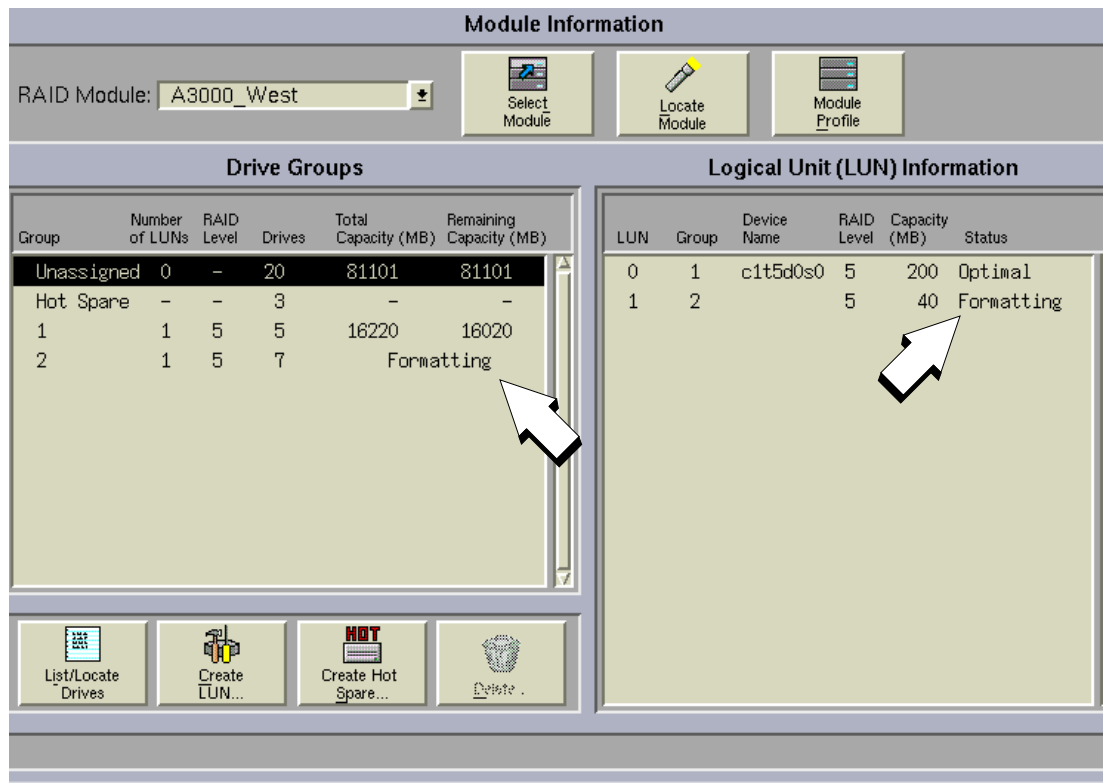
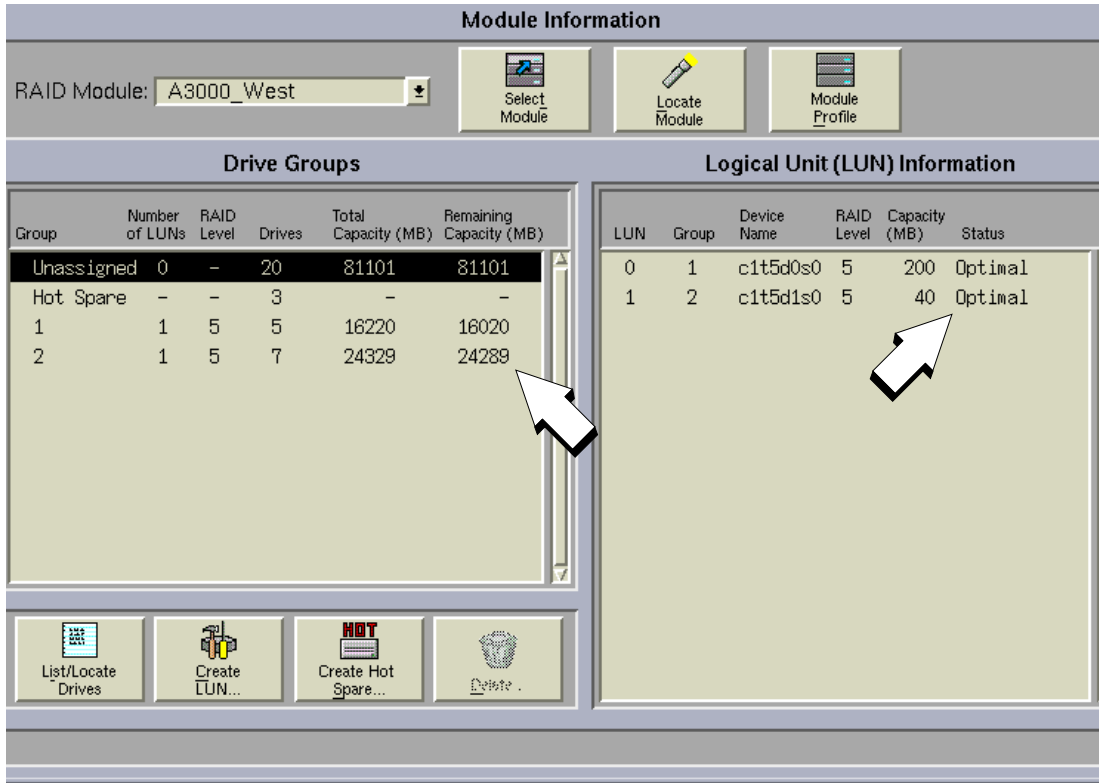


Figure E-11 LUN Formatting Display

Creating a Drive Group

When the LUN is initialized, the Module Information window contains information relative to the newly created drive group.



The screenshot shows the 'Module Information' window for RAID Module 'A3000_West'. It displays two main sections: 'Drive Groups' and 'Logical Unit (LUN) Information'.

Drive Groups Table:

Group	Number of LUNs	RAID Level	Drives	Total Capacity (MB)	Remaining Capacity (MB)
Unassigned	0	-	20	81101	81101
Hot Spare	-	-	3	-	-
1	1	5	5	16220	16020
2	1	5	7	24329	24289

Logical Unit (LUN) Information Table:

LUN	Group	Device Name	RAID Level	Capacity (MB)	Status
0	1	c1t5d0s0	5	200	Optimal
1	2	c1t5d1s0	5	40	Optimal

At the bottom of the window, there are four buttons: 'List/Locate Drives', 'Create LUN...', 'Create Hot Spare...', and 'Delete'.

Figure E-12 New Drive Group Display

Figure E-12 shows that drive group 2 has been created. Drive Group 2 has one LUN assigned (at this time), is initialized to RAID 5, contains seven disk drives, and has a total capacity of 24329 Mbytes. Of the total RAID-5 capacity of 24239 Mbytes, 40 Mbytes have been allocated to LUN 1, leaving 24289 Mbytes available for future allocation to additional LUNs in this drive group.



Sun Educational Services

Adding LUNs to an Existing Drive Group

To add a LUN to an existing drive group, you must:

- Open the Configuration window.
- Select the desired drive group.
- Specify the number of LUNs to add.
- Configure the optional LUN parameters.
- Confirm the option selection.
- Wait for the drive group RAID configuration to complete.

Adding LUNs to an Existing Drive Group

To add a LUN to an existing drive group, you must:

- Open the Configuration window.
- Select the desired drive group.
- Specify the number of LUNs to add.
- Configure the optional LUN parameters.
- Confirm the option selection.
- Wait for the drive group RAID configuration to complete.

Adding LUNs to an Existing Drive Group

You select the configuration application from the RAID Manager top-level window.

1. Click on the Configuration icon.

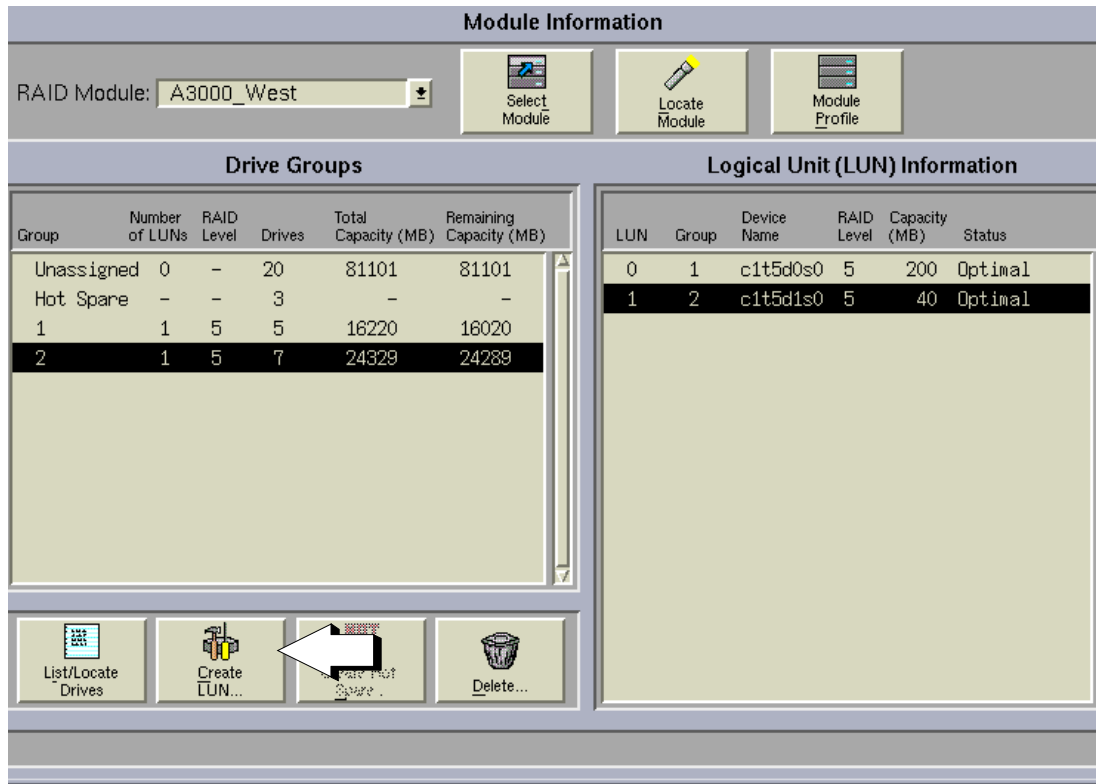


Figure E-13 Drive Group Selection Display

2. Click on the desired drive group.

Note – The selected drive group and all its assigned LUNs are highlighted.

3. Click on the Create LUN gadget.

Adding LUNs to an Existing Drive Group

You do not get the opportunity to select the desired RAID level for this drive group, because the RAID level is determined when the drive group is created. All subsequent LUNs created within this drive group will have the RAID level of the drive group.

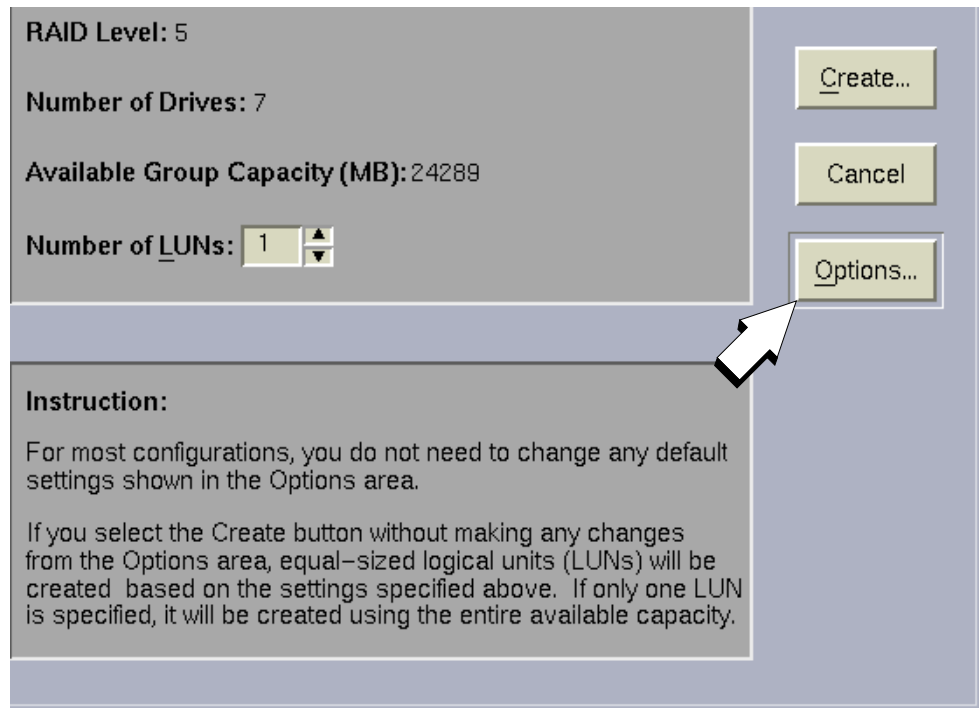


Figure E-14 Set LUN Count Display

4. Select the number of LUNs to create.
5. Click on the Options box to further define the LUN's parameters.

Adding LUNs to an Existing Drive Group

There are five option windows choices listed, however, the Drive Selection option and the LUN Assignment option are greyed out making them not selectable during the process of adding LUNs to existing drive groups. These two options are configured only when the drive group is created and cannot be subsequently modified when LUNs are added.

6. Click on LUN Capacity (if it is not already selected).

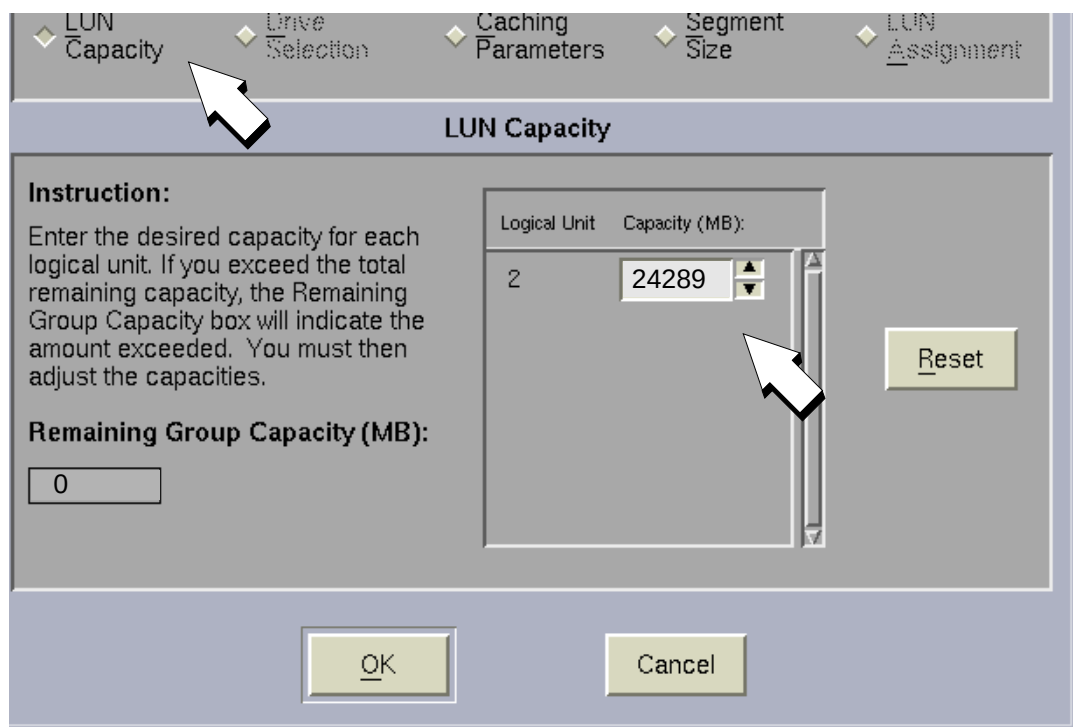


Figure E-15 LUN Capacity Display

7. Type in the desired capacity of this LUN.

Note – You can set this LUN to the remaining size available, so no additional LUNs can be added to this drive group. Remember that each drive group is limited to 16 LUN assignments.

8. From this window, you can accept the defaults for the remaining option parameters by clicking on OK.

Adding LUNs to an Existing Drive Group

You are returned to the Create LUN display.

9. Click on the Create button.

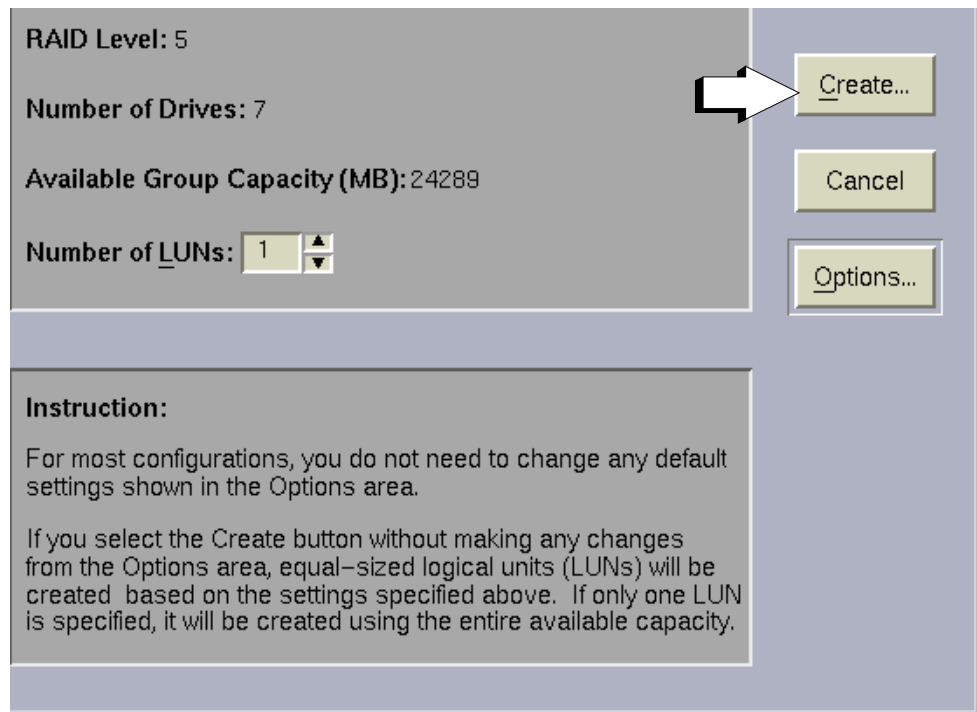


Figure E-16 Create LUN Display

10. Click on the OK button in the Confirmation window.

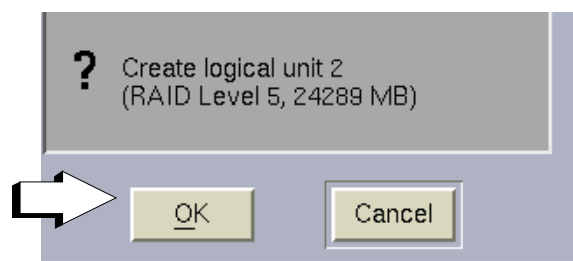


Figure E-17 Create LUN Confirmation Window

Adding LUNs to an Existing Drive Group

You are returned to the (Configuration) Module Information window, where you observe the formatting status during the LUN initialization.

Module Information

RAID Module: A3000_West

Drive Groups

Group	Number of LUNs	RAID Level	Drives	Total Capacity (MB)	Remaining Capacity (MB)
Unassigned	0	-	20	81101	81101
Hot Spare	-	-	3	-	-
1	1	5	5	16220	16020
2	2	5	7	Formatting	

Logical Unit (LUN) Information

LUN	Group	Device Name	RAID Level	Capacity (MB)	Status
0	1	c1t5d0s0	5	200	Optimal
1	2	c1t5d1s0	5	40	Optimal
2	2		5	24289	Formatting

Buttons: List/Locate Drives, Create LUN..., Create Hot Spare..., Delete.

Figure E-18 LUN Formatting Display

Adding LUNs to an Existing Drive Group

When the LUN is initialized, the Module Information window contains information that shows the new LUN in drive group 2.

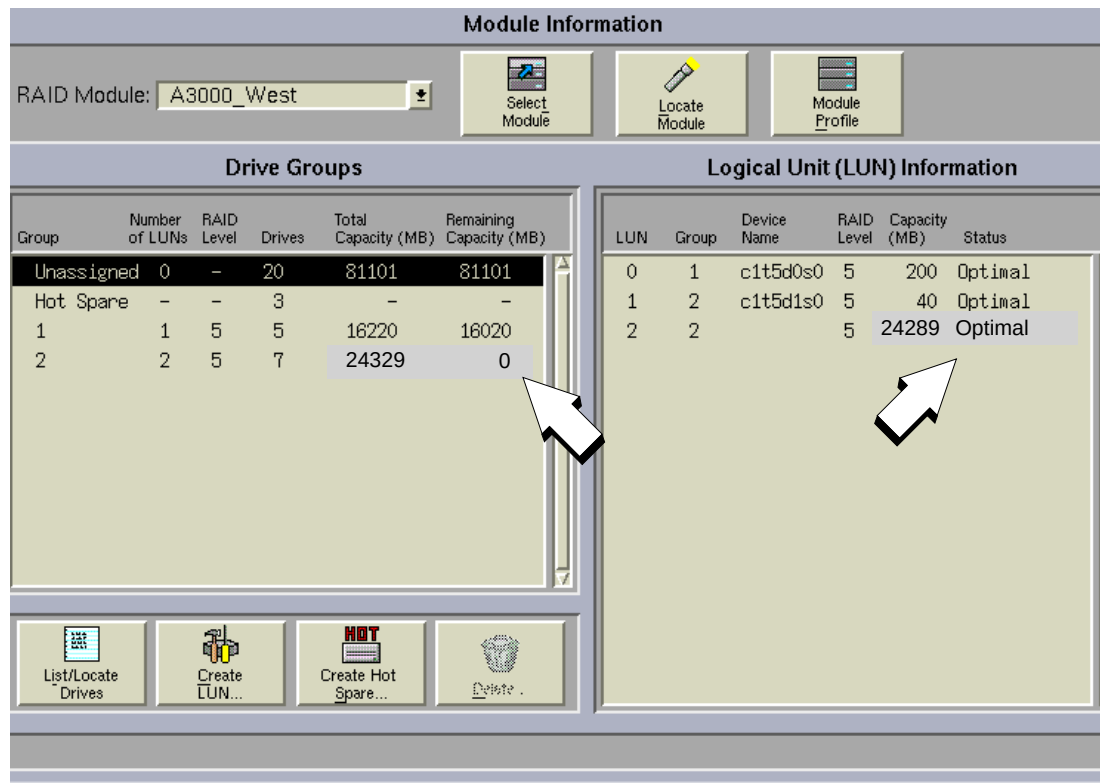


Figure E-19 Multiple LUN Display

Figure E-19 shows that an additional LUN has been added to drive group 2. Drive group 2 now shows that its total capacity remains unchanged at 24329 Mbytes with no remaining available capacity. This is due to the creation of the second LUN, which has absorbed the 24289 Mbytes that were available after the first LUN was created with the drive group. The LUN Information side of the display now registers LUN 1 and LUN 2 in drive group 2. Both LUNs are assigned the logical device `c1t5d1s0`, which is the RAID-5 volume that was created as drive group 2. The optimal status indicates the LUNs are currently available for storing data.



Sun Educational Services

Creating a Hot Spares Pool

To create a hot spare pool, you must:

- Open the Configuration window.
- Select unassigned drives.
- Indicate that you want to create hot spares.
- Ensure the hot spare drives have sufficient storage capacity.
- Confirm the option selection.

Creating a Hot Spares Pool

To create a hot spare pool, you must:

- Open the Configuration window.
- Select unassigned drives.
- Indicate that you want to create hot spares.
- Ensure the hot spare drives have sufficient storage capacity.
- Confirm the option selection.

Creating a Hot Spares Pool

Use this option to create hot spare drives from unassigned drives. These drives contain no data and act as standbys in case any drive fails in a RAID-1, -3, or -5 LUN in the RAID module. The hot spare drive adds another level of redundancy to your RAID module. Each RAID module can support as many hot spare drives as there are SCSI channels (usually two or five, depending on the model of your RAID module).



Caution – Hot spares cannot cover for drives with a larger capacity (that is, a 2-Gbyte hot spare drive cannot stand in for a 4-Gbyte failed drive). If your unassigned drive group contains drives with different capacities, then the Configuration application selects the first available drive when you select Create Hot Spare, which may not be the largest capacity.

If a drive fails, the hot spare drive automatically takes over for the failed drive until you replace it. Once you replace the failed drive, the hot spare drive automatically returns to a Standby status after reconstruction is completed on the new replacement drive.

Note – When you assign a drive as a hot spare, it is used for any configured RAID-1, -3, or -5 LUN that may fail in the RAID module. You cannot specify a hot spare for a particular drive group /LUN.

You can determine the status of the hot spare drives by highlighting the hot spare drive group in the main Configuration window and selecting List/Locate Drives.

Creating a Hot Spares Pool

You select the configuration application from the RAID Manager top-level window.

1. Click on the Configuration icon.

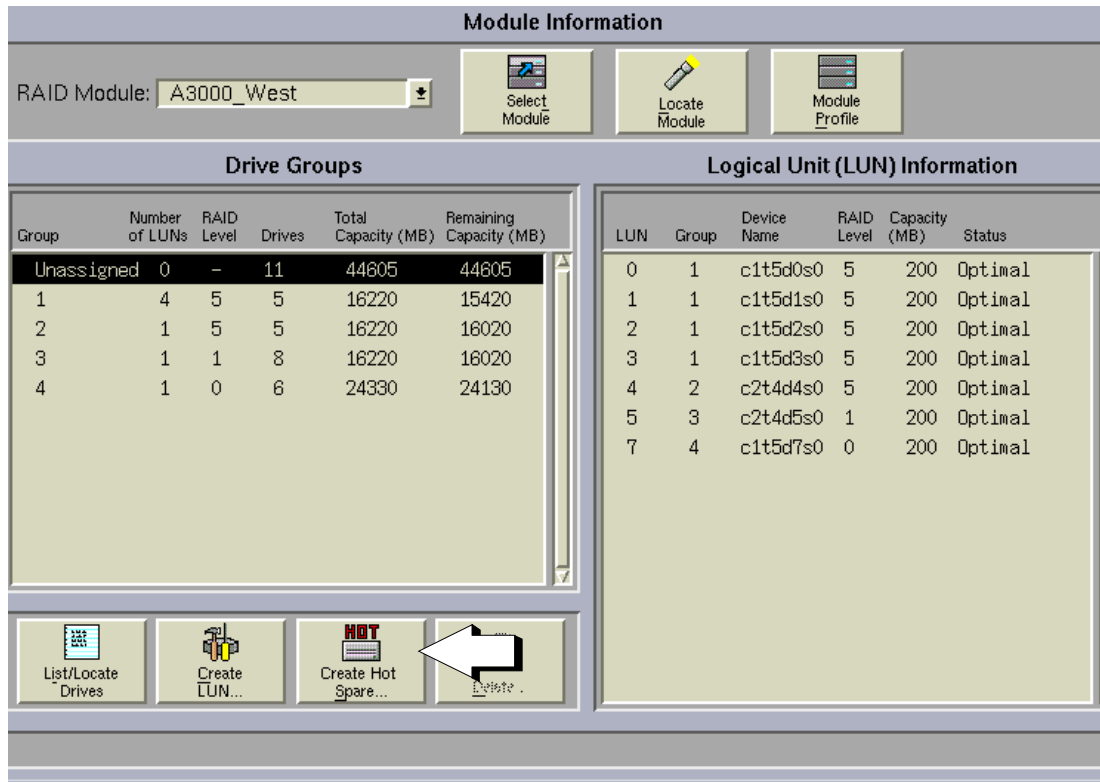


Figure E-20 Configuration Window

2. Click on the unassigned disk drives.
3. Click on the Create Hot Spare button.

Creating a Hot Spares Pool

4. Select the desired number of disk drives to include in this hot spare pool.

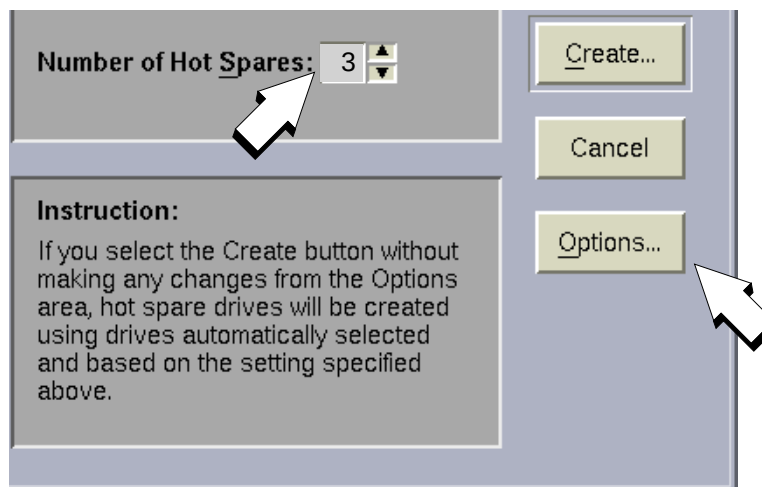


Figure E-21 Create Hot Spare Window

Note – You should have one hot spare per SCSI channel for each RAID module.

5. Click on the Options box.

Creating a Hot Spares Pool

From this window, determine which drives were allocated by the RM6 software to be used as hot spares.

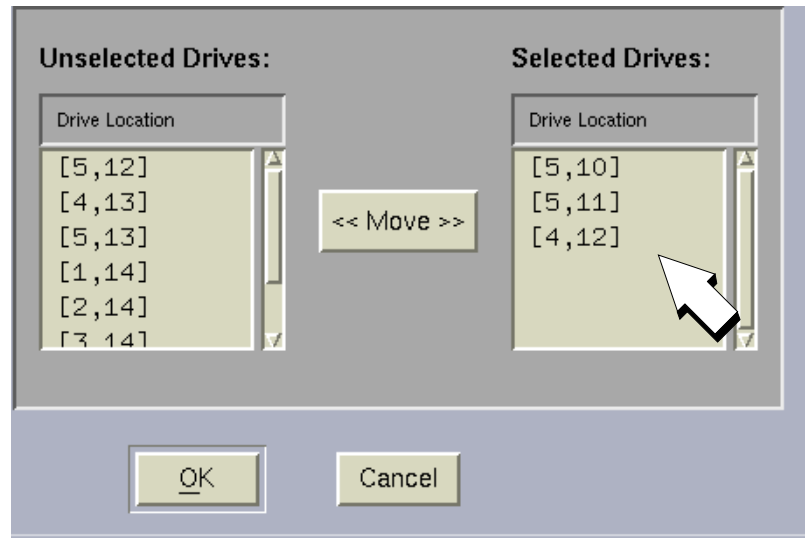


Figure E-22 Hot Spare Drive Selection Display



Caution – You must use drives that have a storage capacity that is equal to or greater than the drives that the hot spare would replace.

Note – To determine the storage capacity of the drives, select the unassigned disks on the Module Information window and click on the List/Locate Drives button. Drive capacity will be displayed.

6. To choose other drives, highlight the drive and click on the Move button.
7. When you are satisfied with the drive selection, click on the OK button.

Creating a Hot Spares Pool

8. Click on the Create button.

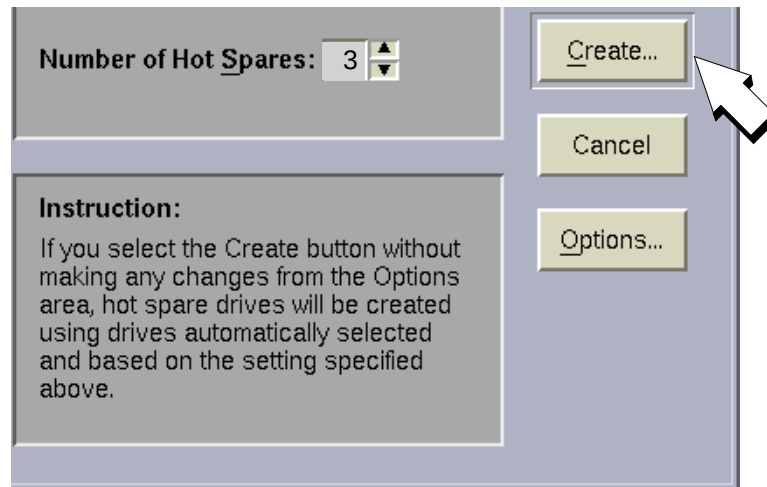


Figure E-23 Create Hot Spares Display

9. Click on the OK button to confirm hot spare assignment.

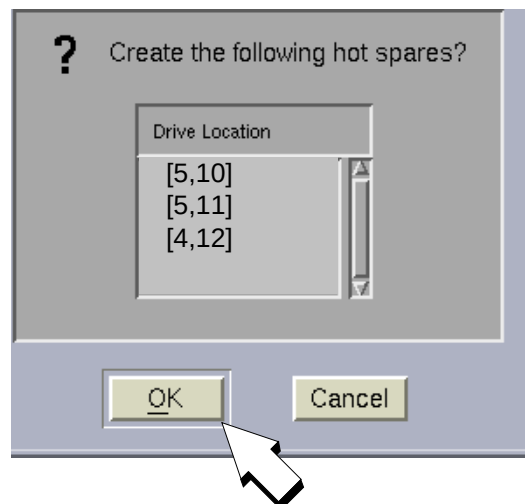
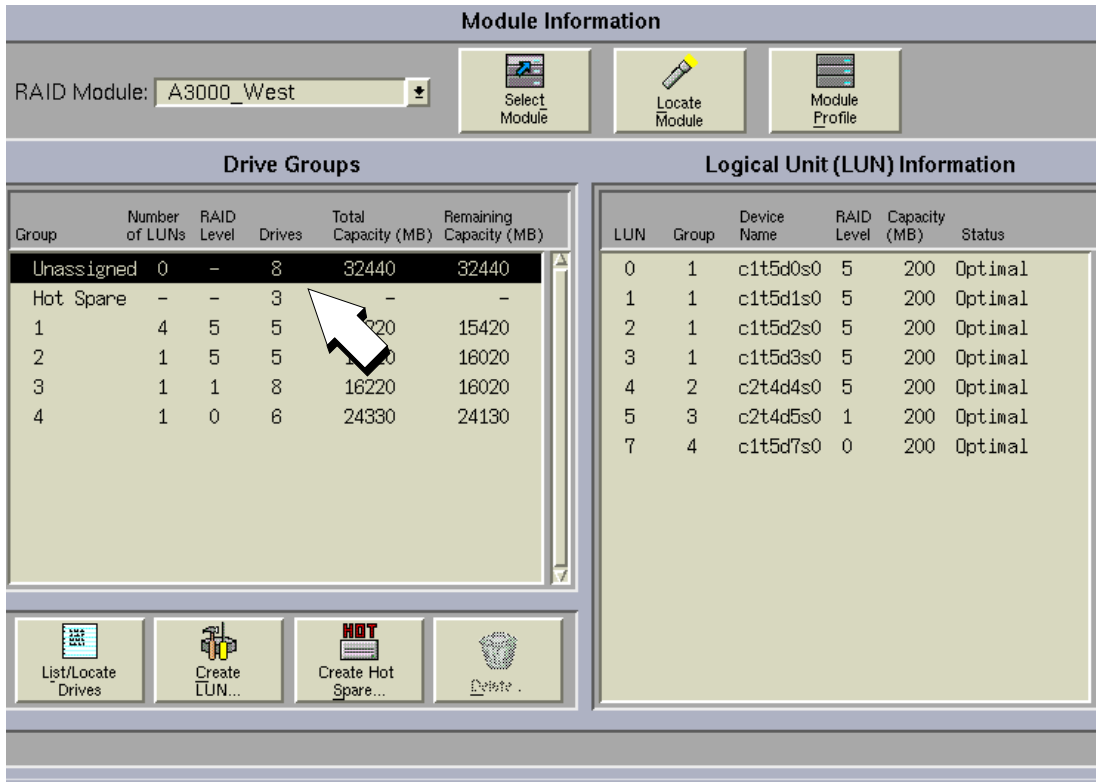


Figure E-24 Hot Spare Confirmation Display

Creating a Hot Spares Pool

You are returned to the (Configuration) Module Information window.



The screenshot shows the 'Module Information' window for RAID Module 'A3000_West'. It displays two main sections: 'Drive Groups' and 'Logical Unit (LUN) Information'.

Drive Groups Table:

Group	Number of LUNs	RAID Level	Drives	Total Capacity (MB)	Remaining Capacity (MB)
Unassigned	0	-	8	32440	32440
Hot Spare	-	-	3	-	-
1	4	5	5	15420	15420
2	1	5	5	16020	16020
3	1	1	8	16220	16020
4	1	0	6	24330	24130

Logical Unit (LUN) Information Table:

LUN	Group	Device Name	RAID Level	Capacity (MB)	Status
0	1	c1t5d0s0	5	200	Optimal
1	1	c1t5d1s0	5	200	Optimal
2	1	c1t5d2s0	5	200	Optimal
3	1	c1t5d3s0	5	200	Optimal
4	2	c2t4d4s0	5	200	Optimal
5	3	c2t4d5s0	1	200	Optimal
7	4	c1t5d7s0	0	200	Optimal

At the bottom of the window, there are four buttons: 'List/Locate Drives', 'Create LUN...', 'Create Hot Spare...', and 'Delete'.

Figure E-25 Hot Spares Listed Display

Figure E-25 shows a reduction in the number of unassigned disk drives and the creation of the hot spare disk pool. This display does not reflect any statistics for the hot spares other than the number of disks assigned.



Sun Educational Services

Deleting a LUN

To delete a LUN, you must:

- Open the Configuration window.
- Select the desired drive group.
- Specify delete LUN.
- Specify which LUN or LUNs to delete.
- Confirm the remaining storage capacity increases following the deletion.

Deleting a LUN

You can use this option to delete all the LUNs in a drive group, individual LUNs within a drive group, or hot spare drives (if supported).

To delete a LUN, you must:

- Open the Configuration window.
- Select the desired drive group.
- Specify delete LUN.
- Specify which LUN or LUNs to delete.
- Confirm the remaining storage capacity increases following the deletion.

Deleting a LUN



Caution – Deleting all LUNs in a drive group causes the loss of all data on each LUN in that drive group. Deleting one LUN in the drive group (for example, to change segment size or capacity), causes data loss on only that one LUN.



Caution – Because deleting LUNs causes data loss, back up data on all the LUNs in any drive group you are deleting. This operation also deletes any file systems mounted on the LUNs.



Caution – You must first stop I/Os to the affected RAID module and ensure no other users are on the system.

You delete all LUNs or the only LUN in a drive group if you want to:

- Change the RAID level or number of drives of that drive group
 - ▼ You delete the LUNs and then use Create LUN to re-create them.
- Free up capacity

You delete individual LUNs in a drive group if you want to:

- Change the segment size or capacity of an individual LUN
 - ▼ You delete the individual LUN and then use Create LUN to re-create them.
- Free up capacity

You delete a standby hot spare drive if you want to:

- Return it to an unassigned status and make it available for LUN creation

Deleting a LUN

Delete is dimmed for either of the following reasons:

- You selected an unassigned drive group. You cannot delete an unassigned drive group.
- You selected a hot spare drive group and all of the hot spares are currently being used. You cannot delete a hot spare drive that is being used because doing so would delete the data contained on it, and would cause the LUN to have a Degraded or Dead status.

After clicking on Delete, a list of LUNs displays for the drive group you selected. You can select any or all of these LUNs to delete.

Once you have deleted LUNs or hot spare drives, the Drive Groups area of the main Configuration window displays one of the following:

- The drives return to the unassigned drive group if you did any of the following:
 - ▼ Deleted all of the LUNs in a drive group
 - ▼ Deleted the only LUN in the drive group
 - ▼ Deleted a hot spare drive
- There will be additional remaining capacity on the drive group if you deleted some, but not all, of the LUNs in a drive group.

Deleting a LUN

You select the Configuration application from the RAID Manager top-level window.

1. Click on the Configuration icon.

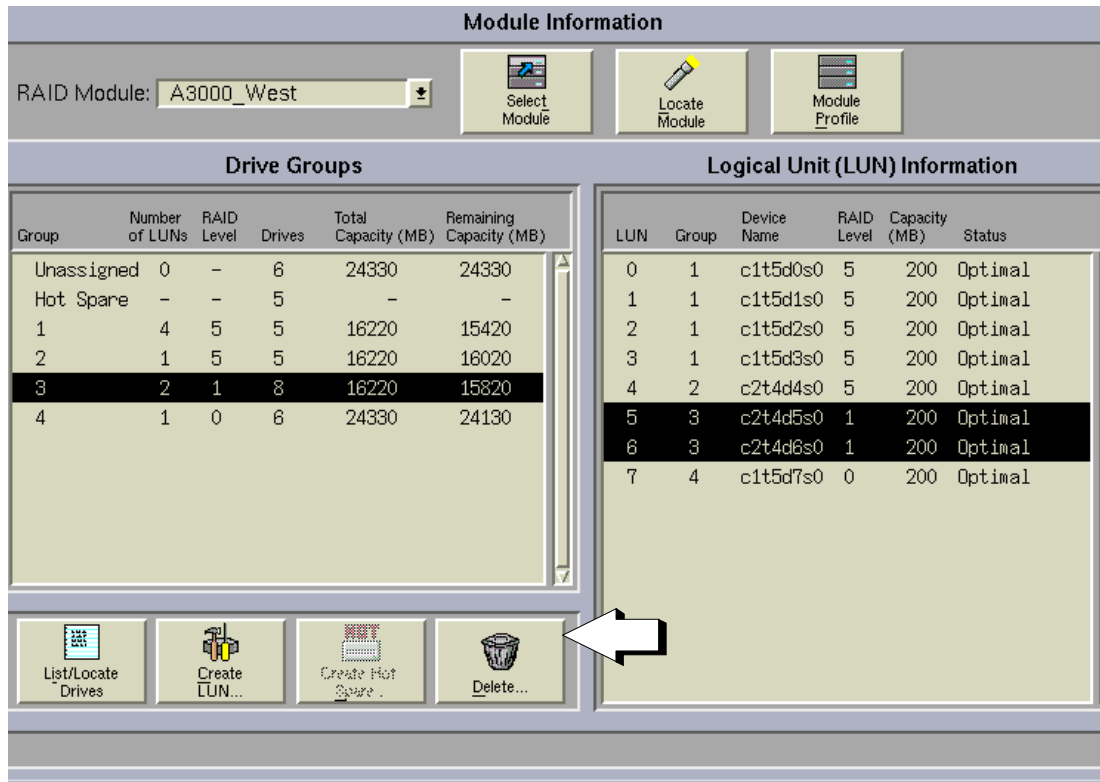


Figure E-26 Configuration Drive Group Display

2. Click on the desired drive group.
3. Click on the Delete button.

Deleting a LUN

The LUN listing is displayed for the selected drive group.

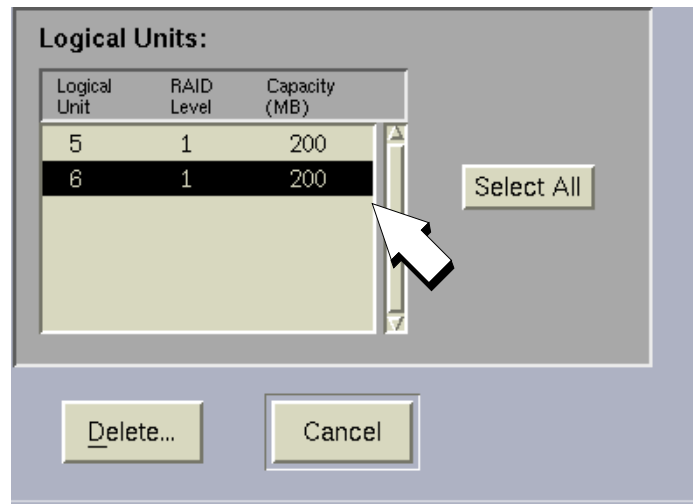


Figure E-27 Drive Group LUN Listing Display

4. You can select all the LUNs within the drive group (click on the Select All button), or any combination of available LUNs (click on the LUN listing within the display).

Deleting a LUN

5. A confirmation is displayed, with an appropriate warning, giving you a final chance to cancel the delete operation. To delete, click on the OK button.

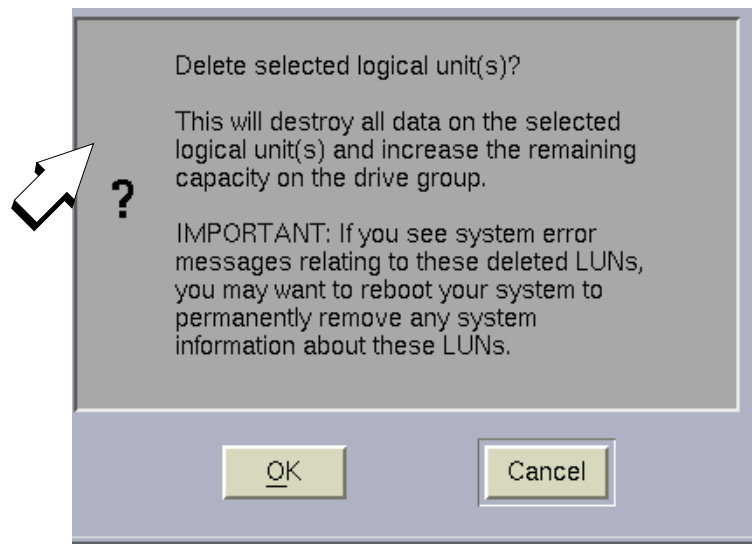


Figure E-28 Deletion Confirmation Display

This display will vary slightly depending on your LUN configuration, as follows:

- If you have selected some, but not all, of the LUNs in the specified drive group, the banner will state:

Delete selected logical unit(s)?

- If you have selected a drive group that contains a single LUN, the banner will state:

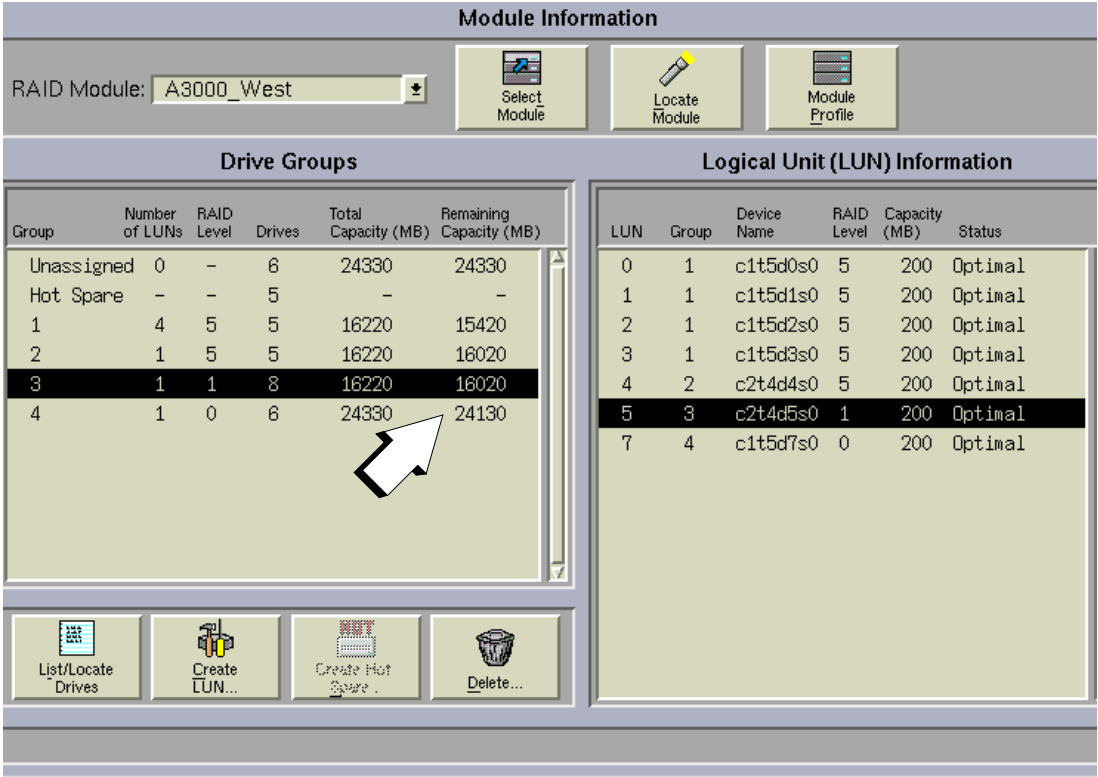
Delete logical unit?

- If you have selected all LUNs within a specified drive group that contains multiple LUNs, the banner will state:

Delete all logical units?

Deleting a LUN

6. After the selected LUNs are deleted, check the remaining storage capacity of the drive group to ensure that it increased appropriately.



The screenshot displays the RAID Manager interface. At the top, the 'Module Information' section shows the 'RAID Module' as 'A3000_West'. Below this are three buttons: 'Select Module', 'Locate Module', and 'Module Profile'. The main area is divided into two panels. The left panel, 'Drive Groups', contains a table with columns: Group, Number of LUNs, RAID Level, Drives, Total Capacity (MB), and Remaining Capacity (MB). The right panel, 'Logical Unit (LUN) Information', contains a table with columns: LUN, Group, Device Name, RAID Level, Capacity (MB), and Status. A mouse cursor is pointing at the 'Remaining Capacity (MB)' of Group 4 in the Drive Groups table.

Group	Number of LUNs	RAID Level	Drives	Total Capacity (MB)	Remaining Capacity (MB)
Unassigned	0	-	6	24330	24330
Hot Spare	-	-	5	-	-
1	4	5	5	16220	15420
2	1	5	5	16220	16020
3	1	1	8	16220	16020
4	1	0	6	24330	24130

LUN	Group	Device Name	RAID Level	Capacity (MB)	Status
0	1	c1t5d0s0	5	200	Optimal
1	1	c1t5d1s0	5	200	Optimal
2	1	c1t5d2s0	5	200	Optimal
3	1	c1t5d3s0	5	200	Optimal
4	2	c2t4d4s0	5	200	Optimal
5	3	c2t4d5s0	1	200	Optimal
7	4	c1t5d7s0	0	200	Optimal

At the bottom of the interface are four buttons: 'List/Locate Drives', 'Create LUN...', 'Create Hot Spare...', and 'Delete...'.

Figure E-29 Drive Group Capacity Display



Recovering Failures

To recover a failure, you must:

- Open the Recovery window.
- Perform a health check to uncover the failure.
- Specify fix the failure.
- Follow the Recovery Guru repair procedures.
- Confirm the recovery with another health check.

Recovering Failures

To recover a failure, you must:

- Open the Recovery window.
- Perform a health check to uncover the failure.
- Specify fix the failure.
- Follow the Recovery Guru repair procedures.
- Confirm the recovery with another health check.

Recovering Failures

Table E-1 Drive Status

Drive Status	Indication	Action
Optimal	The drive is functioning normally.	No action required.
Failed	The drive has failed and is no longer functioning.	Use Recovery Guru to replace the drive as soon as possible.
In Use or Spare	The hot spare drive is currently in use and is taking over for the drive specified in the brackets. The In Use [x,y] status is shown only in List/Locate Drives when you select the hot spare group. The Spare [x,y] status is the same as In Use but is shown in all other screens where drives are displayed.	No action required for the hot spare drive; however, if the drive is being used, it means that the affected logical unit has at least one failed drive. Use Recovery Guru to correct the problem drive as soon as possible.
Offline	The controller has placed the drive Offline because data reconstruction failed and a read error occurred for one or more drives in the LUN. The affected logical unit is Dead, and all its drives are probably either Failed or Offline.	Use Recovery Guru to correct the problem.
Standby or Spare-Stdby	The hot spare drive is currently not in use. The Standby status is shown only in List/ Locate Drives when you select the hot spare group. The Spare-Stdby status is the same as Standby but is shown in all other screens where drives are displayed.	No action required.

Recovering Failures

Table E-1 Drive Status (Continued)

Drive Status	Indication	Action
Replaced	The drive has been replaced, is being formatted, or is reconstructing.	No action required.
Mismatch	The controller has sensed that the drive has some parameters different than expected, such as sector size, SCSI channel, or ID.	Verify that the drive is the correct kind.
Unresponsive	The controller is unable to communicate with a drive that is part of a drive group containing LUNs. You can determine which drive is Unresponsive using Module Profile►Drives in all applications, List/Locate Drives in the Configuration Application, Recovery Guru, Options►Manual Recovery►Drives in the Recovery Application, or Health Check in the Status Application.	Determine which drive is Unresponsive, then manually fail it using Manual Recovery►Drives.

Recovering Failures

Table E-2 LUN Status

Logical Unit Status	Indication	Action
Optimal	The LUN is operating normally.	No action required.
Formatting	The LUN is not available because it is being formatted.	No action required.
Reconstructing	The controller is currently reconstructing a drive on the LUN.	No action required.
Degraded	A single drive in a drive group has failed on a RAID-1, -3, or -5 LUN and the LUN is now functioning in a degraded mode	You can still access your data; however, use Recovery Guru to replace the failed drive as soon as possible.
Dead	The LUN is no longer functioning. Furthermore, all the LUNs in the drive group are Dead also. This is the most serious status a LUN can have and you will lose data unless the LUN status changed from Degraded because you replaced the wrong drive accidentally.	Use Recovery Guru and follow the step-by-step instructions provided.
Inaccessible	The LUN is not available because it is part of a drive group /LUN owned by the alternate controller in an independent controller RAID module. It cannot be accessed using this software from the current host.	If you need to perform an operation on this drive group /LUN, you need to use the software on the host machine connected to the controller that owns that drive group.
Locked	The LUN is not available because an operation has obtained exclusive access to it (such as LUN creation).	No action required.

Recovering Failures

Table E-3 Controller Status

Controller Status	Indication	Action
Optimal	The controller is operating normally.	No action required.
Offline	The controller is not receiving I/O data. Either it has been manually placed offline or the driver for redundant software support has placed it offline (if you have RDAC protection).	If you did not manually place the controller offline, it may need to be replaced. Use Recovery Guru and follow the step-by-step instructions provided.
Dead	There is a problem on the data path (interface cable/terminator, network card, controller, or the host adapter).	Use Recovery Guru to diagnose and correct the problem.

Recovering Failures

To determine if any of the hot spares are in use, indicating a failed RAID component:

1. Click on the Configuration icon (in the RM6 top-level window).

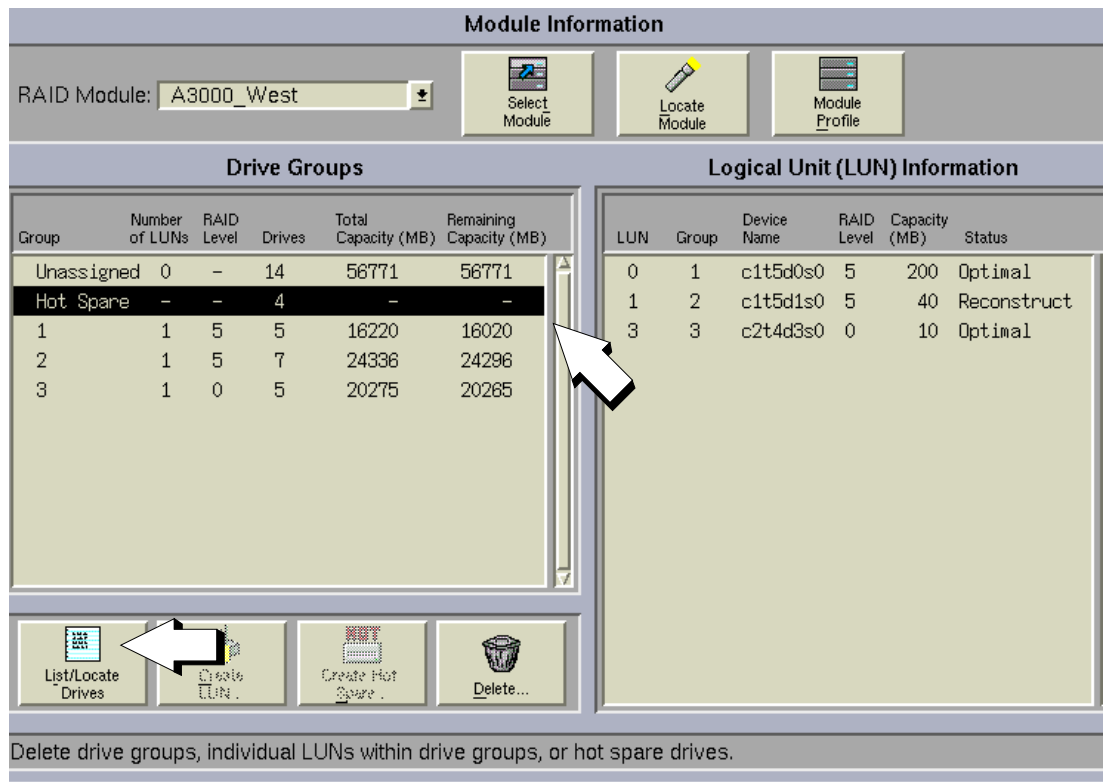


Figure E-30 Hot Spare List/Locate Display

2. Click on the List/Locate Drives button in the lower left corner.

This activity displays a Drive Status window.

Recovering Failures

This Drive Status window lists the status of the hot spare pool. Notice that the hot spare [2,11] is currently in use as a spare for the failed disk at [1,9].

Note – You can also use the List/Locate Drives button to flash the light-emitting diode (LED) of selected (or failed) drives to locate their chassis location.

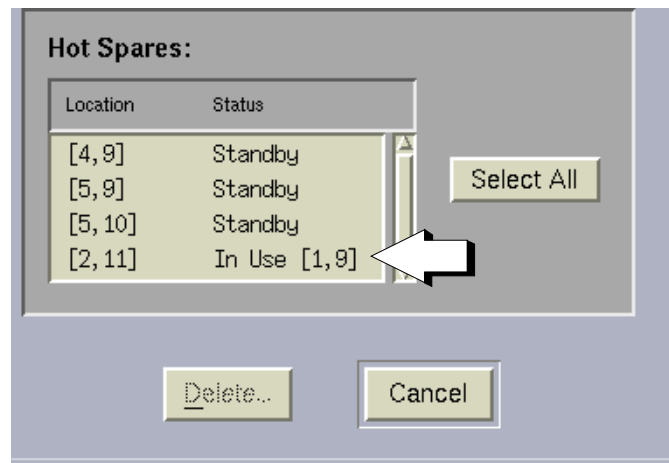


Figure E-31 Hot Spare In Use Display

In Figure E-30, the LUN 1 status is Reconstruct. This status is displayed because the failed RAID device was still reconstructing (rebuilding on the hot spare) when the status was checked. After the reconstruction has completed, the LUN 1 status will return to Optimal, and checking the hot spare status is then one method of determining if any LUNs are currently using a hot spare.

An alternative method is to use the Recovery Guru.

Recovering Failures

You select the recovery application from the RAID Manager top-level window.

1. Click on the Recovery icon.

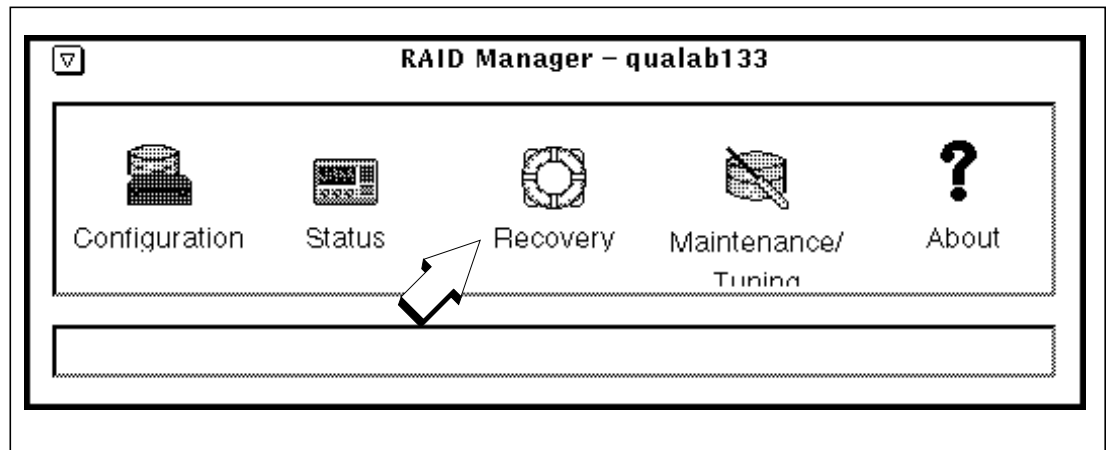


Figure E-32 Recovery Application Selection

This activity displays the Recovery Guru icon in the Module Information window.

Recovering Failures

Table E-4 Drive Failure Types

Failure Type	Probable Cause
Drive failure	One drive in a drive group has failed. A RAID module could show this failure on more than one line, as long as the failed drives belong to different drive groups. Caution – On a RAID-0 LUN, a single drive failure causes the loss of all data.
Multiple drive failure	More than one drive in the same drive group has failed on a RAID module.
Multiple offline / failed drives	One or more drives has been placed offline because data reconstruction failed and a read error occurred for one or more failed drives in the LUN.
Hot spare failure	A hot spare drive has failed while being used by a LUN on the RAID module. Note – This means that the drive the hot spare was covering for is also still failed, and the LUN has probably become Degraded.
Multiple unresponsive drives	The controller is unable to communicate with multiple drives in the selected RAID module. If you see this result, the drives' status in Module Profile►Drives is most likely Unresponsive. If the drives receive any I/O, the controller will fail them.
Unresponsive drive	The controller is unable to communicate with a drive in the selected RAID module. If you see this result, the drive status in Module Profile►Drives is most likely Unresponsive. If the drive receives any I/O, the controller will fail it.

Recovering Failures

Table E-5 Drive Trays Failure Types

Failure Type	Probable Cause
Drive tray – fan failure	A fan in one of the disk drive trays has failed. Replace the fan as soon as possible to keep the drives from overheating.
Drive tray – fan failures	Both fans in one of the disk drive trays have failed. Replace the fans as soon as possible to keep the drives from overheating.
Drive tray – power supply failure	A power supply in one of the disk drive trays has failed. Replace the power supply as soon as possible because a failure to a second power supply may cause the drive tray to shut down.
Drive tray – power supply failures	Both power supplies in one of the disk drive trays has failed. Replace the power supplies as soon as possible; the drive tray most likely has been shut down.
Drive tray – temperature exceeded	The maximum temperature allowed within a disk drive tray has been exceeded. Caution – This is a critical condition that may cause the drive tray to be automatically turned off if you do not resolve this condition within a short time.

Recovering Failures

Table E-6 Other Failure Types

Failure Type	Probable Cause
Channel failure	All of the drives on the same drive channel have Failed and /or are Unresponsive. Depending on how the logical units have been configured across these drives, the status of the logical units may be Dead, Degraded, or Optimal (if hot spare drives are in use).
Data path failure	A controller is not receiving I/O, which indicates some component along the data path has failed. For network versions, this means that the controller is not responding to the RAID Manager software. This failure could be the result of a problem with the interface cable/ terminator, controller, or the host adapter. The correct procedure for recovering from a data path failure varies depending on where the failure occurred. For example, the correct procedure for recovering from a controller failure depends on how many and what type of controllers the affected module has. Important – If you do not have RDAC protection, this failure type may not be displayed for every condition. Therefore, verify that the interface cable/ terminator or network card is not removed or damaged before proceeding with any controller-related recovery procedure.
Environmental card failure	An environmental card in one of the disk drive trays has failed. Caution – You may see a series of disk drive failures or a channel failure reported as well. You must service the environmental card first using the Recovery Guru. This recovery procedure will instruct you on how to fix the corresponding drive or channel failures; therefore, you should not use the Recovery Guru for the associated drive or channel failure entries.
Module component failure	Either single or multiple fans or power supplies have failed. Important – When recovering from a module component failure, wait for the controller to poll the module (default is 10 minutes) before re-selecting the Recovery Guru. Otherwise, this condition may continue to be reported as a failure.

Recovering Failures

2. You can leave the Recovery Guru default of All RAID Modules as the RAID Module selection, or you can choose a specific RAID module to test.
3. Click on the Recovery Guru button.

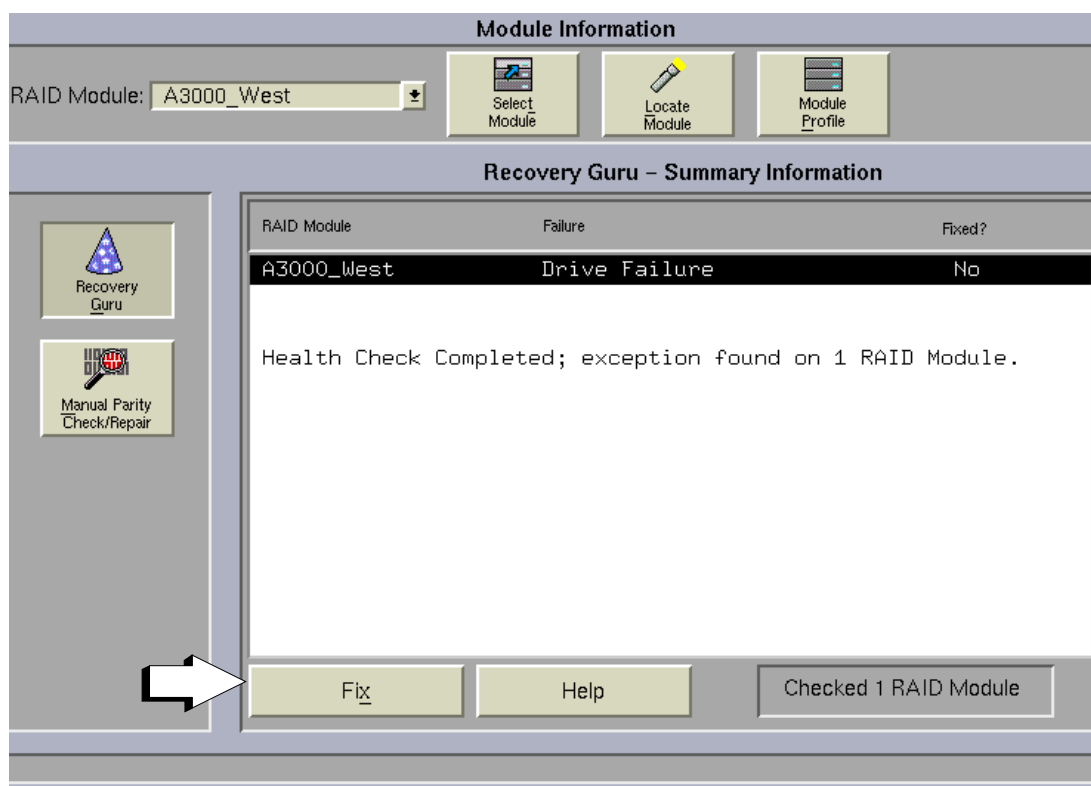


Figure E-33 Health Check Display

After clicking on the Recovery Guru button, RM6 runs a health check on the selected RAID modules, and returns any failures found. If no failures are found, the Fix button is not accessible. If any failures are found, clicking on the Fix button causes the Recovery Guru to begin displaying a series of screens that give step-by-step instructions on how to repair the failure.

Recovering Failures

This display provides information that identifies a failed RAID-5 component; it also tells you that a hot spare has been used to reconstruct the failed LUN.

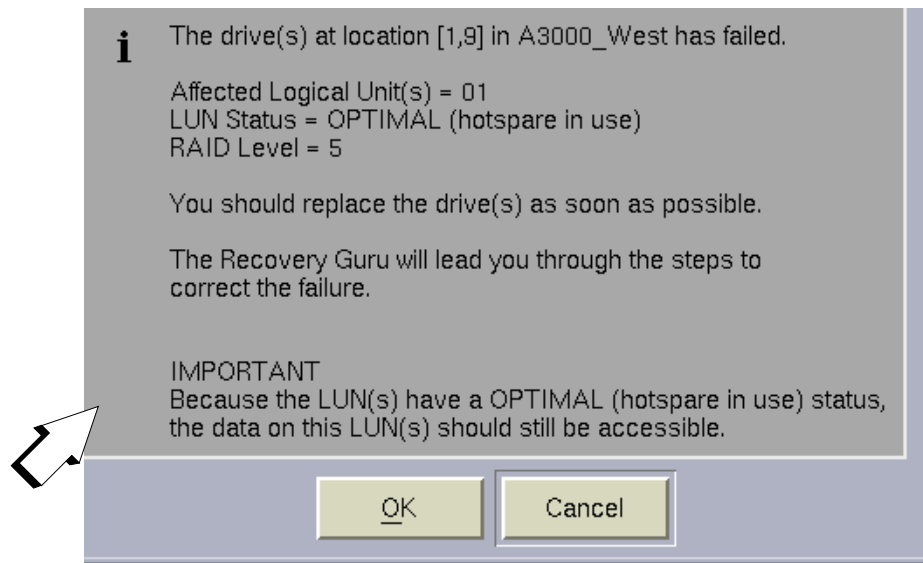


Figure E-34 Failure Identification Display

4. Click on the OK button to begin the recovery process.

Recovering Failures

5. Follow the displayed steps, and then click on the OK button.

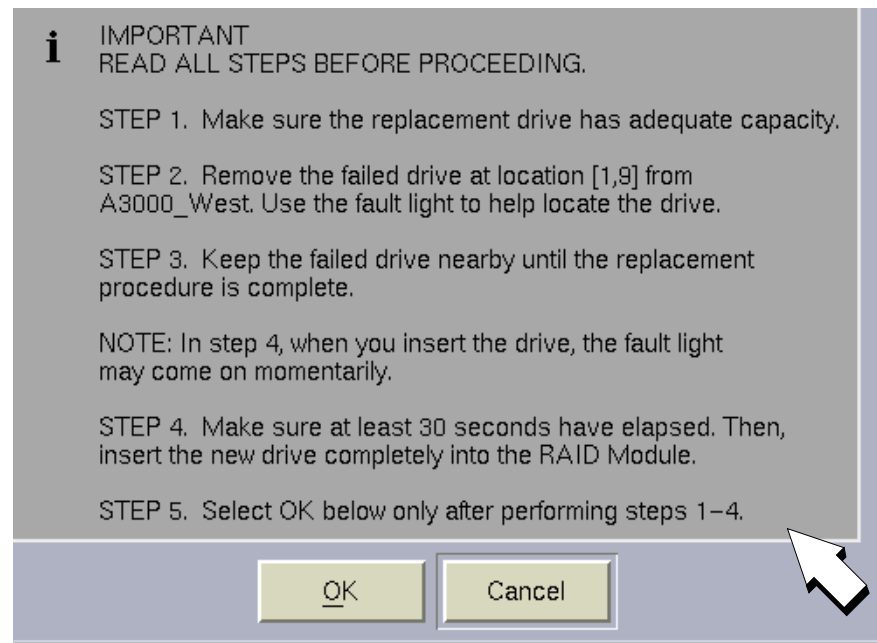


Figure E-35 Procedure to Repair Failure

Recovering Failures

The Recovery Guru checks to see that the drive is installed and spun up.

Note – The Recovery Guru does not check device capacity. It is assumed that you have followed instructions and replaced the failed component with an acceptable replacement device.

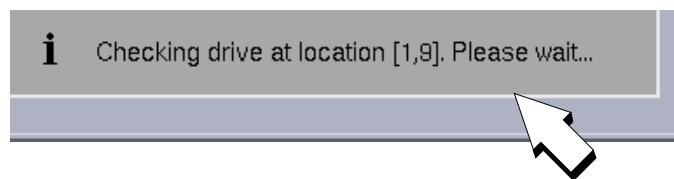


Figure E-36 Status Check Information Display

After checking the replacement device, data is copied from the hot spare to the replacement device.

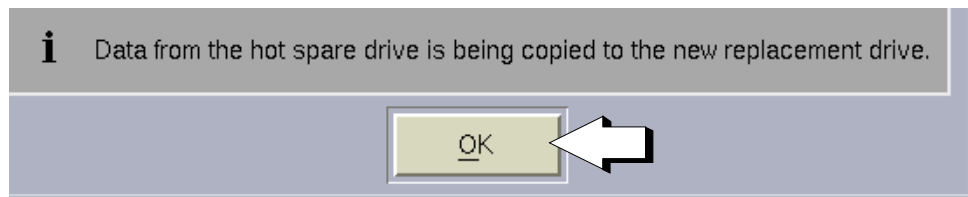


Figure E-37 Confirmation Display

Because the LUN was functioning on the hot spare, the LUN status was Optimal. The LUN status will remain Optimal throughout the data restoration activity.

Recovering Failures

6. Upon completion of the LUN restoration process, the hot spare status display shows that the hot spare that was in use has been returned to a standby status (and is ready for any subsequent failures).

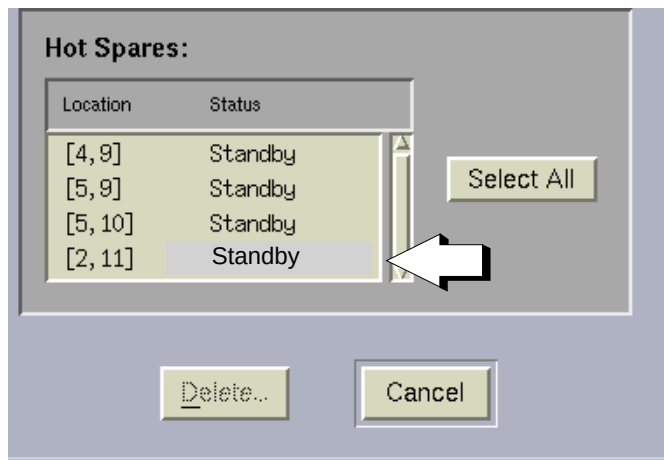


Figure E-38 Hot Spare Status

Exercise: Using RAID Manager Procedures

Exercise objective – In this exercise, you will:

- Create a drive group
- Add a LUN to an existing drive group
- Create a hot spare pool
- Delete a LUN
- Recover from a failure

Exercise: Using RAID Manager Procedures

Task – Creating a Drive Group

Complete the following steps:

1. Select the configuration application from the RAID Manager top-level window.
2. Select the unassigned disk drives.
3. Click on the Create LUN gadget.
4. Set the RAID level for this drive group equal to RAID 5.
5. Set the number of disk drives to include in this drive group to 4.
6. Set the number of LUNs to create to 1.
7. Click on the Options box to further define the new drive group's parameters.
8. Set the LUN capacity for this LUN to 50 Mbytes.
9. Click on Drive Selection and check to see if the selected disk drives are evenly distributed between the available buses.
10. Ensure the Write Caching Parameter is enabled.
11. Set the Segment Size to 256 blocks.
12. Ensure the LUN Assignment is set appropriately.
13. When you are satisfied with the LUN option parameter settings, click on the OK gadget on any of the option screens to display the Create LUN window.
14. In the Create LUN window, click on the Create button.
15. Click on the OK button in the Confirmation window.
16. Confirm the drive group is created, and the LUN is formatted. (Completion is indicated by a LUN status of Optimal.)

Exercise: Using RAID Manager Procedures

Task – Adding LUNs To An Existing Drive Group

Complete the following steps:

1. Select the configuration application from the RAID Manager top-level window.
2. Click on the desired drive group (the drive group that was created in the first section of this lab).
3. Click on the Create LUN gadget.
4. Set the number of LUNs to create to 2.
5. Click on the Options box to further define the LUNs parameters.
6. Set the LUN Capacity to 40 Mbytes each.

From this window, you can accept the defaults for the remaining option parameters by clicking on OK.

7. Click on the Create button.
8. Click on the OK button in the Confirmation window.
9. Observe successful completion of LUN formatting by waiting for LUN status of Optimal.
10. Ensure the new drive group now has three LUNs attached. The first LUN should be 50 Mbytes in size, and the other two LUNs should each be 40 Mbytes in size.

Note – You are requested to keep these LUN sizes small, during the lab exercise, due to the time required to format the LUNs during the creation process. Except for the formatting time requirement, the process works the same for a 40-Mbyte LUN as it does for a 4-Gbyte LUN.

Exercise: Using RAID Manager Procedures

Task – Creating a Hot Spares Pool

Use the following steps:

1. Select the Configuration application from the RAID Manager top-level window.
2. Click on the unassigned disk drives.
3. Click on the Create Hot Spare button.
4. Set the desired number of disk drives to include in this hot spare pool to 2.
5. Click on the Options box.
6. Ensure the hot spares are distributed so that one exists on each bus.
7. When you are satisfied with the drive selection, click on the OK button.
8. Click on the Create button.
9. Click on the OK button to confirm hot spare assignment.
10. Confirm the hot spare pool was created with the proper number of disk drives.

Exercise: Using RAID Manager Procedures

Task – Deleting a LUN

Complete the following steps:

1. Select the Configuration application from the RAID Manager top-level window.
2. Click on your drive group (that was created in the first segment of this exercise).
3. Click on the Delete button.
4. Select the first LUN listed within the drive group. (This is the 50-Mbyte LUN that was created with the drive group.)
5. A Confirmation window is displayed, with an appropriate warning, giving you a final chance to cancel the delete operation. To delete, click on the OK button.
6. Check the remaining storage capacity of the drive group to ensure that it increased by 50 Mbytes.

Exercise: Using RAID Manager Procedures

Task – Recovering Failures

Complete the following steps:

1. Select the recovery application from the RAID Manager top-level window.
2. Click on the Recovery icon.
3. Leave the Recovery Guru default of All RAID Modules as the RAID module selection, or choose a specific RAID module to test.
4. Click on the Recovery Guru button.
5. Click on the OK button to begin the recovery process.
6. Follow the displayed steps, and then click on the OK button.

Upon completion of the LUN restoration process, the hot spare status display shows that the hot spare that was in use has been returned to a standby status (and is ready for any subsequent failures).

Exercise: Using RAID Manager Procedures

Exercise Summary



Discussion – Take a few minutes to discuss what experiences, issues, or discoveries you had during the lab exercises.

- Experiences
- Interpretations
- Conclusions
- Applications

Check Your Progress

Before continuing, check that you are able to accomplish or answer the following:

- ☐ Create a drive group
- ☐ Add a LUN to an existing drive group
- ☐ Create a hot spare pool
- ☐ Delete a LUN
- ☐ Recover from a failure such as a failed RAID set

Think Beyond

Having completed this introduction to the RAID Manager 6 software architecture and procedures, you should be able to perform LUN manipulation on the disk arrays prior to configuring the arrays with the other supported volume managers.

Copyright 2000 Sun Microsystems Inc., 901 San Antonio Road, Palo Alto, California 94303, Etats-Unis. Tous droits réservés.

Ce produit ou document est protégé par un copyright et distribué avec des licences qui en restreignent l'utilisation, la copie, la distribution, et la décompilation. Aucune partie de ce produit ou document ne peut être reproduite sous aucune forme, par quelque moyen que ce soit, sans l'autorisation préalable et écrite de Sun et de ses bailleurs de licence, s'il y en a.

Le logiciel détenu par des tiers, et qui comprend la technologie relative aux polices de caractères, est protégé par un copyright et licencié par des fournisseurs de Sun.

Des parties de ce produit pourront être dérivées du système Berkeley 4.3 BSD licenciés par l'Université de Californie. UNIX est une marque déposée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X/Open Company Ltd.

Sun, Sun Microsystems, le logo Sun, Solaris, Sun StorEdge Volume Manager, Ultra, Answerbook, Java, NFS, Solstice DiskSuite, et OpenBoot sont des marques de fabrique ou des marques déposées de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays.

Toutes les marques SPARC sont utilisées sous licence sont des marques de fabrique ou des marques déposées de SPARC International, Inc. aux Etats-Unis et dans d'autres pays.

Les produits portant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

UNIX est une marque déposée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X/Open Company, Ltd.

L'interface d'utilisation graphique OPEN LOOK et Sun™ a été développée par Sun Microsystems, Inc. pour ses utilisateurs et licenciés. Sun reconnaît les efforts de pionniers de Xerox pour la recherche et le développement du concept des interfaces d'utilisation visuelle ou graphique pour l'industrie de l'informatique. Sun détient une licence non exclusive de Xerox sur l'interface d'utilisation graphique Xerox, cette licence couvrant également les licenciés de Sun qui mettent en place l'interface d'utilisation graphique OPEN LOOK et qui en outre se conforment aux licences écrites de Sun.

L'accord du gouvernement américain est requis avant l'exportation du produit.

Le système X Window est un produit de X Consortium, Inc.

LA DOCUMENTATION EST FOURNIE "EN L'ETAT" ET TOUTES AUTRES CONDITIONS, DECLARATIONS ET GARANTIES EXPRESSES OU TACITES SONT FORMELLEMENT EXCLUES, DANS LA MESURE AUTORISEE PAR LA LOI APPLICABLE, Y COMPRIS NOTAMMENT TOUTE GARANTIE IMPLICITE RELATIVE A LA QUALITE MARCHANDE, A L'APTITUDE A UNE UTILISATION PARTICULIERE OU A L'ABSENCE DE CONTREFAÇON.



Please
Recycle



Adobe PostScript

