# Sun Cluster 2.2 System Administration Guide

901 San Antonio Road
Palo Alto, , CA  94303-4900
USA 650 960-1300 Fax 650 969-9131

Part Number  806-5343
July 2000, Revision 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.

CETTE PUBLICATION EST FOURNIE "EN L'ETAT" ET AUCUNE GARANTIE, EXPRESSE OU IMPLICITE, N'EST ACCORDEE, Y COMPRIS DES GARANTIES CONCERNANT LA VALEUR MARCHANDE, L'APTITUDE DE LA PUBLICATION A REPONDRE A UNE UTILISATION PARTICULIERE, OU LE FAIT QU'ELLE NE SOIT PAS CONTREFAISANTE DE PRODUIT DE TIERS. CE DENI DE GARANTIE NE S'APPLIQUERAIT PAS, DANS LA MESURE OU IL SERAIT TENU JURIDIQUEMENT NUL ET NON AVENU.

Adobe PostScript

Please
Recycle

# Contents

# Preface

Sun™ Cluster 2.2 is a software product that supports specific two- to four-node server hardware configurations. It is compatible with the Solaris™ 2.6, Solaris 7, and Solaris 8 operating environments. When configured properly, the hardware and software together provide highly available data services and parallel database access. Sun Cluster depends upon the mirroring and disk group capabilities and other functionality provided by a volume manager. Sun Cluster supports Solstice™ DiskSuite and VERITAS Volume Manager (VxVM). The VxVM cluster feature is supported for use with the Oracle Parallel Server data service.

This book documents the procedures for setting up hardware and installing, configuring, and administering the Sun Cluster software. This book is intended to be used with the hardware and software books listed under “Related Documentation” on page 16.

## Who Should Use This Book

This book is intended for Sun representatives and system administrators whose duties include installing and maintaining Sun Cluster 2.2 configurations. The instructions and discussions are complex and intended for a technically advanced audience.

The instructions in this book assume readers have expertise with one of the supported volume managers used with Sun Cluster.

System administrators with UNIX® system experience will find this book useful when learning to administer Sun Cluster 2.2 configurations.

**Note -** Junior or less-experienced system administrators should not attempt to install, configure, or administer Sun Cluster 2.2 configurations.

# How This Book Is Organized

This book is split into parts that each cover a major system administration topic. Each part contains chapters that provide both overview and task information.

Most of the overview information about a topic is described in the first few chapters of each part, and the subsequent chapters provide step-by-step instructions for completing system administration tasks.

# Related Documentation

The documents listed in Table P–1 contain information that might be helpful to the system administrator or service provider. You should also have available the hardware installation and service manuals for your cluster hardware.

**TABLE P–1** List of Related Documentation

| Product Family | Title | Part Number |
|---|---|---|
| Sun Cluster | *Sun Cluster 2.2 Software Installation Guide* | 806-5342 |
| | *Sun Cluster 2.2 API Developer's Guide* | 806-5344 |
| | *Sun Cluster 2.2 Error Messages Manual* | 805-4242 |
| | *Sun Cluster 2.2 Release Notes* | 806-5345 |
| | *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide* | 806-5346 |
| | *Sun Cluster 2.2 Hardware Service Manual* | 806-5347 |
| Solstice DiskSuite | *Solstice DiskSuite 4.2 Installation/Product Notes* | 805-5960 |
| | *Solstice DiskSuite 4.2 User's Guide* | 805-5961 |
| | *Solstice DiskSuite 4.2 Reference* | 805-5962 |

**TABLE P–1** List of Related Documentation *(continued)*

| Product Family | Title | Part Number |
|---|---|---|
| | *Solstice DiskSuite 4.2.1 Reference Guide* | 806-3204 |
| | *Solstice DiskSuite 4.2.1 User's Guide* | 806-3205 |
| | *Solstice DiskSuite 4.2.1 Installation and Product Notes* | 806-3206 |
| VERITAS Volume Manager | *Veritas Volume Manager for Solaris* | 100-001125A (VERITAS) |

## Typographic Conventions

The following table describes the typographic conventions used in this book.

**TABLE P–2** Typographic Conventions

| Typeface or Symbol | Meaning | Example |
|---|---|---|
| `Typewriter` | The names of commands, files, and directories; on-screen computer output. | Edit your `.login` file.<br>Use `ls -a` to list all files.<br>`machine_name% You have mail.` |
| **`boldface`** | What you type, contrasted with on-screen computer output. | `machine_name%` **`su`**<br>`Password:` |
| *italic* | Command-line placeholder:<br>replace with a real name or value.<br>Book titles, new words or terms, or words to be emphasized. | To delete a file, type `rm` *filename*. |

## Shell Prompts in Command Examples

The following table shows the default system prompt and superuser prompt for the C shell, Bourne shell, and Korn shell.

**TABLE P–3** Shell Prompts

| **Shell** | **Prompt** |
|---|---|
| C shell prompt | `machine_name%` |
| C shell superuser prompt | `machine_name#` |
| Bourne shell and Korn shell prompt | `$` |
| Bourne shell and Korn shell superuser prompt | `#` |

## Getting Help

If you have problems installing or using Sun Cluster software, contact your service provider and provide the following information:

- Your name and electronic mail address (if available)
- Your company name, address, and phone number
- The model and serial numbers of your systems
- The release number of the operating system (for example, Solaris 2.6)
- The release number of the Sun Cluster software (for example, Sun Cluster 2.2)

Use the following commands to gather information on your system for your service provider:

**TABLE P–4** Getting Help

| | |
|---|---|
| `prtconf -v` | Displays the size of the system memory and reports information about peripheral devices |
| `psrinfo -v` | Displays information about processors |
| `showrev --p` | Reports which patches are installed |
| `prtdiag -v` | Displays system diagnostic information |

Also have available the contents of the `/var/adm/messages` file.

CHAPTER **1**

# Preparing for Sun Cluster Administration

This chapter describes procedures used to prepare for administration of the Sun Cluster configuration. Some of the procedures documented in this chapter are dependent on your volume management software (Solstice DiskSuite or VERITAS Volume Manager). Those that are dependent on the volume manager include the volume manager name in the procedure title. This chapter contains the following sections:

- "Saving Disk Partition Information (Solstice DiskSuite)" on page 21
- "Saving and Restoring VTOC Information (Solstice DiskSuite)" on page 22
- "Saving Device Configuration Information" on page 23
- "Instance Names and Numbering" on page 24
- "Logging Into the Servers as `root`" on page 26

## Saving Disk Partition Information (Solstice DiskSuite)

Maintain disk partitioning information for all nodes and multihost disks in the Sun Cluster configuration. Keep this information up-to-date as new disks are added to the disksets and when any of the disks are repartitioned. You need this information to perform disk replacement.

The disk partitioning information for the local disks is not as critical because the local disks on all Sun Cluster nodes should have been partitioned identically. Most likely, you can obtain the local disk partition information from another Sun Cluster node if a local disk fails.

When a multihost disk is replaced, the replacement disk must have the same partitioning as the disk it is replacing. Depending on how a disk has failed, this information might not be available when replacement is performed. Therefore, it is especially important to retain a record of the disk partitioning information if you have several different partitioning schemes in your disksets.

---

**Note -** Though VxVM does not impose this restriction, it is still a good idea to save this information.

---

A simple way to save disk partitioning information is shown in the following sample script. This type of script should be run after the Sun Cluster software has been configured. In this example, the files containing the volume table of contents (VTOC) information are written to the local `/etc/opt/SUNWcluster/vtoc` directory by the `prtvtoc(1M)` command.

```
#! /bin/sh
DIR=/etc/opt/SUNWcluster/vtoc
mkdir -p $DIR
cd /dev/rdsk
for i in *s7
do prtvtoc $i >$DIR/$i || rm $DIR/$i
done
```

Each of the disks in a Solstice DiskSuite diskset is required to have a Slice 7. This slice contains the metadevice state database replicas.

If a local disk also has a valid Slice 7, the VTOC information also will be saved by the sample script. However, this should not occur for the boot disk, because typically a boot disk does not have a valid Slice 7.

---

**Note -** Make certain that the script is run while none of the disks is owned by another Sun Cluster node. The script will work if the logical hosts are in maintenance mode, if the logical hosts are owned by the local host, or if Sun Cluster is not running.

---

## Saving and Restoring VTOC Information (Solstice DiskSuite)

When you save the VTOC information for all multihost disks, this information can be used when a disk is replaced. The sample script shown in the following example uses the VTOC information saved by the script shown below to give the replacement disk the same partitioning as the failed disk. Use the actual names of the disk or disks to be added in place of `c1t0d0s7` and `c1t0d1s7` in the example. Specify multiple disks as a space-delimited list.

```
#! /bin/sh
DIR=/etc/opt/SUNWcluster/vtoc
cd /dev/rdsk
for i in c1t0d0s7 c1t0d1s7
do fmthard -s $DIR/$i $i
done
```

**Note -** The replacement drive must be of the same size and geometry (generally the same model from the same manufacturer) as the failed drive. Otherwise the original VTOC might not be appropriate for the replacement drive.

If you did not record this VTOC information, but you have mirrored slices on a disk-by-disk basis (for example, the VTOCs of both sides of the mirror are the same), it is possible to copy the VTOC information from the other submirror disk to the replacement disk. For this procedure to be successful, the replacement disk must be in maintenance mode, or must be owned by the same host as the failed disk, or Sun Cluster must be stopped. This procedure is shown in the following example.

```
#! /bin/sh
cd /dev/rdsk
OTHER_MIRROR_DISK=c2t0d0s7
REPLACEMENT_DISK=c1t0d0s7
prtvtoc $OTHER_MIRROR_DISK | fmthard -s - $REPLACEMENT_DISK
```

If you did not save the VTOC information and did not mirror on a disk-by-disk basis, you can examine the component sizes reported by the `metaset(1M)` command and reverse engineer the VTOC information. Because the computations used in this procedure are complex, the procedure should be performed only by a trained service representative.

## Saving Device Configuration Information

Record the `/etc/path_to_inst` and the `/etc/name_to_major` information on removable media (floppy disk or backup tape).

The `path_to_inst(4)` file contains the minor unit numbers for disks in each multihost disk expansion unit. This information will be necessary if the boot disk on any Sun Cluster node fails and has to be replaced.

- **Solstice DiskSuite –** For configurations that do not use the Disk ID (DID) driver, the `/etc/name_to_major` file contains the major device numbers for multihost disks. For example, Solstice DiskSuite relies upon the major numbers remaining the same across Solaris operating system installs. This applies only to clusters that upgraded from HA 1.3 to Sun Cluster 2.2. Refer to the Solstice DiskSuite appendix in the *Sun Cluster 2.2 Software Installation Guide* for more information.
- **VxVM –** To avoid "Stale File handle" errors at the client on NFS failovers, make sure that the `vxio` driver has identical pseudo-device major numbers on all cluster nodes. This number can be found in the `/etc/name_to_major` file after you complete the installation. Refer to the chapters on Sun Cluster HA for NFS and on configuring VxVM in the *Sun Cluster 2.2 Software Installation Guide* for more information.

## Instance Names and Numbering

Instance names are occasionally reported in driver error messages. An instance name refers to system devices such as `ssd20` or `hme5`.

You can determine the binding of an instance name to a physical name by looking at `/var/adm/messages` or `dmesg(1M)` output:

```
ssd20 at SUNW,pln0:
ssd20 is /io-unit@f,e0200000/sbi@0,0/SUNW,soc@3,0/SUNW,pln@a0000800,20183777 \

/ssd@4,0

le5 at lebuffer5: SBus3 slot 0 0x60000 SBus level 4 sparc ipl 7
le5 is /io-unit@f,e3200000/sbi@0,0/lebuffer@0,40000/le@0,60000
```

Once an instance name has been assigned to a device, it remains bound to that device.

Instance numbers are encoded in a device's minor number. To keep instance numbers persistent across reboots, the system records them in the `/etc/path_to_inst` file. This file is read only at boot time and is currently updated by the `add_drv(1M)` and `drvconfig(1M)` commands. For additional information refer to the `path_to_inst(4)` man page.

When you install the Solaris operating environment on a node, instance numbers can change if hardware was added or removed since the last Solaris installation. For this reason, use caution whenever you add or remove devices such as SBus or FC/OM cards on Sun Cluster nodes. It is important to maintain the same configuration of existing devices, so that the system is not confused in the event of a reinstall or reconfiguration reboot.

Instance number problems can arise in a configuration. For example, consider a Sun Cluster configuration that consists of three SPARCstorage™ Arrays with Fibre Channel/SBus (FC/S) cards installed in SBus slots 1, 2, and 4 on each of the nodes. The controller numbers are c1, c2, and c3. If the system administrator adds another SPARCstorage Array to the configuration using a FC/S card in SBus slot 3, the corresponding controller number will be c4. If Solaris is reinstalled on one of the nodes, the controller numbers c3 and c4 will refer to different SPARCstorage Arrays. The other Sun Cluster node will still refer to the SPARCstorage Arrays with the original instance numbers. Solstice DiskSuite will not communicate with the disks connected to the c3 and c4 controllers.

Other problems can arise with instance numbering associated with the Ethernet connections. For example, each of the Sun Cluster nodes has three Ethernet SBus cards installed in slots 1, 2, and 3, and the instance numbers are hme1, hme2, and hme3. If the middle card (hme2) is removed and Solaris is reinstalled, the third SBus card will be renamed from hme3 to hme2.

## Performing Reconfiguration Reboots

During some of the administrative procedures documented in this book, you are instructed to perform a reconfiguration reboot by using the OpenBoot™ PROM `boot -r` command, or by creating the `/reconfigure` file on the node and then rebooting.

---

**Note -** It is not necessary to perform a reconfiguration reboot to add disks to an existing multihost disk expansion unit.

---

Avoid performing Solaris reconfiguration reboots when any hardware (especially a multihost disk expansion unit or disk) is powered off or otherwise defective. In such situations, the reconfiguration reboot removes the inodes in `/devices` and symbolic links in `/dev/dsk` and `/dev/rdsk` associated with the disk devices. These disks become inaccessible to Solaris until a later reconfiguration reboot. A subsequent reconfiguration reboot might not restore the original controller minor unit numbering, and therefore might the volume manager software to reject the disks. When the original numbering is restored, the volume manager software can access the associated objects.

If all hardware is operational, you can perform a reconfiguration reboot safely to add a disk controller to a node. You must add such controllers symmetrically to both nodes (though a temporary unbalance is allowed while the nodes are upgraded). Similarly, if all hardware is operational, it is safe to perform a reconfiguration reboot to remove hardware.

**Note -** For the Sun StorEdge A3000, in the case of a single controller failure, you should replace the failed controller as soon as possible. Other administration tasks that would normally require a `boot –-r` (such as after adding a new SCSI device) should be deferred until the failed controller has been replaced and brought back online, and all logical unit numbers (LUN) have been balanced back to their previous state when the failover occurred. Refer to the Sun StorEdge A3000 documentation for more information.

## Logging Into the Servers as `root`

If you want to log in to Sun Cluster nodes as `root` through a terminal other than the console, you must edit the `/etc/default/login` file and comment out the following line:

```
CONSOLE=/dev/console
```

This enables `root` logins using `rlogin(1)`, `telnet(1)`, and other programs.

CHAPTER **2**

# Sun Cluster Administration Tools

This chapter provides information about the following topics.

Administering the Sun Cluster software is facilitated by three Graphical User Interfaces (GUIs):

**Cluster Control Panel –** Launches the Cluster Console and other system administration tools.

**Cluster Console –** Executes commands on multiple nodes in the cluster simultaneously to simplify cluster administration.

**Sun Cluster Manager –** Monitors the current status of all nodes in the cluster via a HotJava browser.

Refer to the online help for complete documentation on these GUIs. You can also use utilities to monitor Sun Cluster software.

## Monitoring Utilities

You can use the Sun Cluster `hastat(1M)`utility, in addition to the `/var/adm/messages` files, to monitor a Sun Cluster configuration. You can also use the Sun Cluster Manager graphical user interface, which shows the status of major cluster components and subcomponents. For more information about the Sun Cluster

Manager, refer to "Using Sun Cluster Manager " on page 42. Sun Cluster also provides an SNMP agent that can be used to monitor up to 32 clusters at the same time. See Appendix C.

If you are running Solstice DiskSuite, you can also use the `metastat(1M)`, `metadb(1M)`, `metatool(1M)`, `medstat(1M)`, and `mdlogd(1M)` utilities to monitor the status of your disksets. The SNMP-based Solstice DiskSuite log daemon, `mdlogd(1M)`, generates a generic SNMP trap when Solstice DiskSuite logs a message to the `syslog` file. You can configure `mdlogd(1M)` to send a trap only when certain messages are logged by specifying a regular expression in the `mdlogd.cf(4)` configuration file. The trap is sent to the administrative host specified in the configuration file. The administrative host must be running a network management application such as Solstice SunNet Manager™ . You can use `mdlogd(1M)` if you don't want to run `metastat(1M)` periodically or scan the `syslog` output looking for Solstice DiskSuite errors or warnings. See the `mdlogd(1M)` man page for more information.

If you are running VxVM, you can use the `vxprint`, `vxstat`, `vxtrace`, `vxnotify`, and `vxva` utilities. Refer to your volume management software documentation for information about these utilities.

---

**Note -** For information about troubleshooting and repairing defective components, refer to the appropriate hardware documentation.

---

## Monitoring the Configuration With `hastat(1M)`

The `hastat(1M)` program displays the current state of the configuration. The program displays status information for the hosts, logical hosts, private networks, public networks, data services, local disks, and disksets, along with the most recent error messages. The `hastat(1M)` program extracts Sun Cluster-related error messages from the `/var/adm/messages` file and outputs the last few messages from each host if `-m` is specified. Because the recent error messages list is a filtered extract of the log messages, the context of some messages might be lost. Check the `/var/adm/messages` file for a complete list of the messages. The following pages show an example of output from `hastat(1M)`:

```
# hastat -m 10

HIGH AVAILABILITY CONFIGURATION AND STATUS
-------------------------------------------

LIST OF NODES CONFIGURED IN <ha-host1> CLUSTER
      phys-host1 phys-host2

CURRENT MEMBERS OF THE CLUSTER

```

**(continued)**

(Continuation)

```
     phys-host1 is a cluster member
     phys-host2 is a cluster member

CONFIGURATION STATE OF THE CLUSTER

     Configuration State on phys-host1: Stable
     Configuration State on phys-host2: Stable

UPTIME OF NODES IN THE CLUSTER

     uptime of phys-host1:         12:47pm  up 12 day(s), 21:11,  1 user,
load average: 0.21, 0.15, 0.14
     uptime of phys-host2:         12:46pm  up 12 day(s),  3:15,  3 users,
load average: 0.40, 0.20, 0.16
```

```
LOGICAL HOSTS MASTERED BY THE CLUSTER MEMBERS

Logical Hosts Mastered on phys-host1:
        ha-host-1
Loghost Hosts for which phys-host1 is Backup Node:
        ha-host2

Logical Hosts Mastered on phys-host2:
        ha-host2
Loghost Hosts for which phys-host2 is Backup Node:
        ha-host1

LOGICAL HOSTS IN MAINTENANCE STATE

     None

STATUS OF PRIVATE NETS IN THE CLUSTER

     Status of Interconnects on phys-host1:
        interconnect0: selected
        interconnect1: up
     Status of private nets on phys-host1:
        To phys-host1 - UP
        To phys-host2 - UP

     Status of Interconnects on phys-host2:
        interconnect0: selected
        interconnect1: up
     Status of private nets on phys-host2:
        To phys-host1 - UP
        To phys-host2 - UP

STATUS OF PUBLIC NETS IN THE CLUSTER
```

(continued)

(Continuation)

```
Status of Public Network On phys-host1:

bkggrp  r_adp   status  fo_time live_adp
nafo0   le0     OK      NEVER   le0

Status of Public Network On phys-host2:

bkggrp  r_adp   status  fo_time live_adp
nafo0   le0     OK      NEVER   le0
```

```
STATUS OF SERVICES RUNNING ON LOGICAL HOSTS IN THE CLUSTER

       Status Of Registered Data Services
       q:                       Off
       p:                       Off
       nfs:                     On
       oracle:                  On
       dns:                     On
       nshttp:                  Off
       nsldap:                  On

      Status Of Data Services Running On phys-host1
      Data Service HA-NFS:
      On Logical Host ha-host1:      Ok

      Status Of Data Services Running On phys-host2
      Data Service HA-NFS:
      On Logical Host ha-host2:      Ok

       Data Service ‘‘oracle’’:
       Database Status on phys-host2:
       SC22FILE - running;

       No Status Method for Data Service ‘‘dns’’

       RECENT  ERROR MESSAGES FROM THE CLUSTER

       Recent Error Messages on phys-host1
       ...
       Recent Error Messages on phys-host2
       ...
```

## Checking Message Files

The Sun Cluster software writes messages to the /var/adm/messages file, in addition to reporting messages to the console. The following is an example of the messages reported when a disk error occurs.

```
...
Jun 1 16:15:26 host1 unix: WARNING: /io-unit@f,e1200000/sbi@0.0/SUNW,pln@a0000000,741022/
ssd@3,4(ssd49):
Jun 1 16:15:26 host1 unix: Error for command ‘write(I))’ Err
Jun 1 16:15:27 host1 unix: or Level: Fatal
Jun 1 16:15:27 host1 unix: Requested Block 144004, Error Block: 715559
Jun 1 16:15:27 host1 unix: Sense Key: Media Error
Jun 1 16:15:27 host1 unix: Vendor ‘CONNER’:
Jun 1 16:15:27 host1 unix: ASC=0x10(ID CRC or ECC error),ASCQ=0x0,FRU=0x15
...
```

**Note -** Because Solaris and Sun Cluster error messages are written to the /var/adm/messages file, the /var directory might become full. Refer to “Maintaining the /var File System” on page 94 for the procedure to correct this problem.

## Highly Available Data Service Utilities

In addition, Sun Cluster provides utilities for configuring and administering the highly available data services. The utilities are described in their associated man pages. The utilities include:

- cconsole(1) – Starts the cluster console GUI.
- ccp(1) -- Starts the cluster control panel GUI.
- ctelnet(1) – Starts a telnet session.
- crlogin(1) – Starts an rlogin session.
- chosts(1) – Expands a cluster name into a list of hosts belonging to the cluster.
- cports(1) – Expands a host name into a *host, node, port* triplet. Used by cconsole(1) to identify the serial port consoles of the named hosts via the terminal server returned in the triplets.
- scconf(1M) – Creates or modifies configuration information.

# Online Help System

Each Sun Cluster administration tool includes detailed online help. To access the online help, launch one of the administration tools from the administrative workstation and select Help from the menu bar.

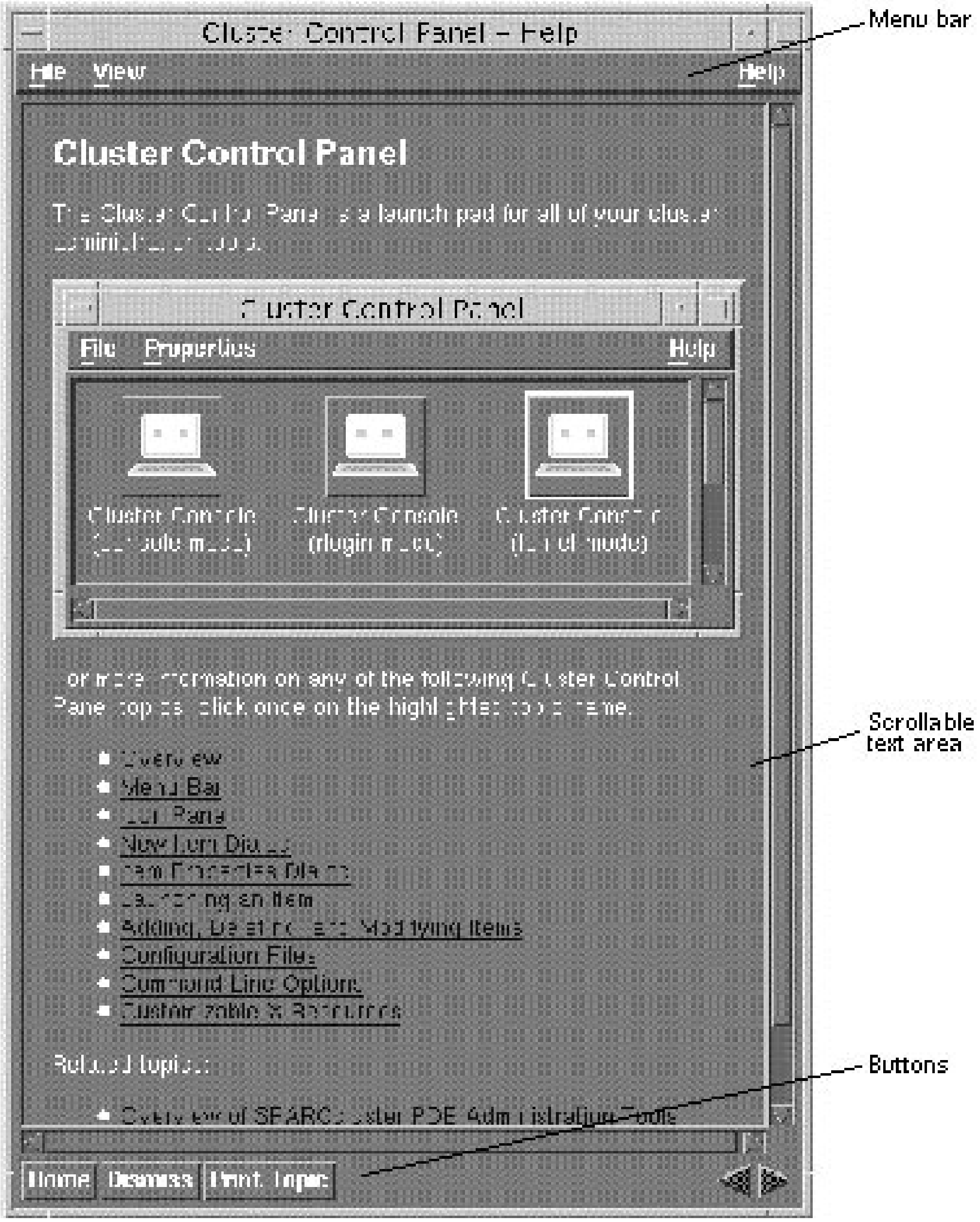
Alternatively, double-click on the Help icon in the Cluster Control Panel.

Help topics cover the administration tools in detail, as well as some administration tasks. Also see Chapter 4, for additional detailed instructions on performing specific tasks.

Figure 2–1 shows a sample Help window for the Cluster Control Panel. The text covers a specific topic. When you first launch the Help window from a tool, it displays the top-level home topic. Afterwards, the Help window displays the last topic you viewed. Hypertext links to other topics are displayed as underlined, colored text.

Clicking once on a hypertext link displays the text for that topic. The online help system also includes an automatic history list feature that remembers the information you previously accessed. Display this list by choosing Topic History from the View menu.

The Help window has a scrollable text area, a menu bar, and several buttons. Each of these items is described in the following sections.


Cluster Control Panel – Help
File
View
Help
Menu bar
Cluster Control Panel
File
Properties
Help
Cluster Console (console mode)
Cluster Console (rlogin mode)
Cluster Console (telnet mode)
Scrollable text area
Buttons
Home
Dismiss
Print Topic

Cluster Control Panel

The Cluster Control Panel is a launch pad for all of your cluster administration tools.

For more information on any of the following Cluster Control Panel topics, click once on the highlighted topic name.

- Overview
- Menu Bar
- Icon Pane
- New Item Dialog
- Item Properties Dialog
- Launching an Item
- Adding, Deleting, and Modifying Items
- Configuration Files
- Command Line Options
- Customizable X Resources

Related topics:

- Overview of SPARCcluster PDB Administration Tools

*Figure 2–1* Sample Help Window Home Page for the Cluster Control Panel

With Help, you can access each pull-down menu by:

- Clicking on the menu name

- Pressing the mnemonic, or underlined character in a menu or entry name (only when the pull-down menu is visible)
- Using an accelerator, or the sequence of keys to the right of menu items

You can customize the mnemonics and accelerators; refer to the online help for more information.

The tables in this section list each menu item, define the menu functions, and list the accelerators (keyboard combinations).

## Help Window Menu Bar Items

The Help window menu contains the File, View, and Help menu items. You display the menus for these items by selecting them.

### File Menu

The File menu contains these items:

**TABLE 2–1** File Menu Items

| **Item** | **Function** | **Accelerator** |
|---|---|---|
| Print Topic | Prints the topic currently displayed in the Help window scrolled text area. | Alt + R |
| Dismiss | Dismisses the Help window. | Alt + D |

### View Menu

The View menu contains these items:

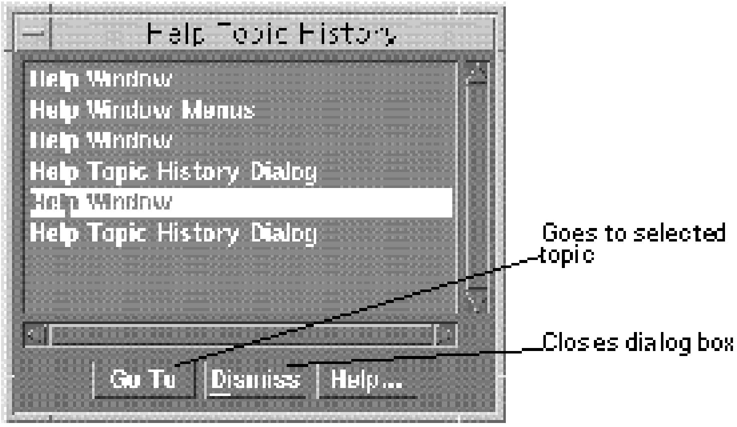
**TABLE 2–2** View Menu Items

| **Item** | **Function** | **Accelerator** |
|---|---|---|
| Previous Topic | Displays the previous help topic (if any). | Alt + P |
| Next Topic | Displays the next help topic (if any). | Alt + N |

**TABLE 2–2** View Menu Items *(continued)*

| Item | Function | Accelerator |
|---|---|---|
| Home Topic | Displays the home (top level) topic. | Alt + O |
| Topic History... | Displays the Help Topic History dialog box which allows you to navigate easily through the help topics you have already viewed. The uppermost topic in the scrolled list is the first topic you displayed. The bottommost topic is the last topic you have viewed in the current path. The highlighted topic is the current topic. | Alt + I |

To display the dialog box, select View → Topic History... (Figure 2–2).



Goes to selected topic

Closes dialog box

*Figure 2–2* Help Topic History of the Help Window

## Help Menu

The Help menu contains these items:

**TABLE 2–3** Help Menu Items

| Item | Function |
|---|---|
| Help on Help... | Describes the Help window and explains how to use it. |
| About... | Displays the About Box, which contains information about the application, such as the version number. |

## Help Window Buttons

The following tables lists the help window buttons and their functions.

**TABLE 2–4** Help Menu Items

| Button | Function |
|---|---|
| Home | Displays the home page for the application. |
| Dismiss | Dismisses the Help window. |
| Print Topic | Prints the current topic on your default printer. |
| ◀ | Steps back through the displayed Help topics to the previous topic. Clicking on the left arrow repeatedly steps the display back through the Help windows until the first topic you viewed is redisplayed. Topics are “remembered” by the automatic Help history list. |
| ▶ | Steps forward through the displayed Help topics, one at a time, to the last topic in the list. |

# Cluster Control Panel

The Cluster Control Panel (CCP) is a GUI that enables you to launch the Cluster Console, and other system administration tools. The CCP contains icons that represent these tools.

## ▼ How to Launch the Cluster Control Panel

After you have installed the Sun Cluster client software on the administrative workstation, use this procedure to run an application from the CCP.

1. **As superuser, add the Sun Cluster tools directory** `/opt/SUNWcluster/bin` **to the path on the administrative workstation.**

---

**Note -** For E10000 Platforms, you must first log into the System Service Processor (SSP) and connect by using the `netcon` command. Once connected, enter Shift~@ to unlock the console and gain write access. Then proceed to Step 2.

---

2. **In a shell window on your workstation, bring up the CCP.**

   Specify the name of the cluster to be monitored:

```
# ccp clustername
```

---

**Note -** If the Sun Cluster tools are not installed in the default location of `/opt/SUNWcluster`, the environment variable `$CLUSTER_HOME` must be set to the alternate location.

---

## Cluster Control Panel Items

The CCP (shown in the following figure) has a menu bar and an icon pane that displays all of the tools currently in the control panel. From the menu bar you can add, delete, or modify the tools.

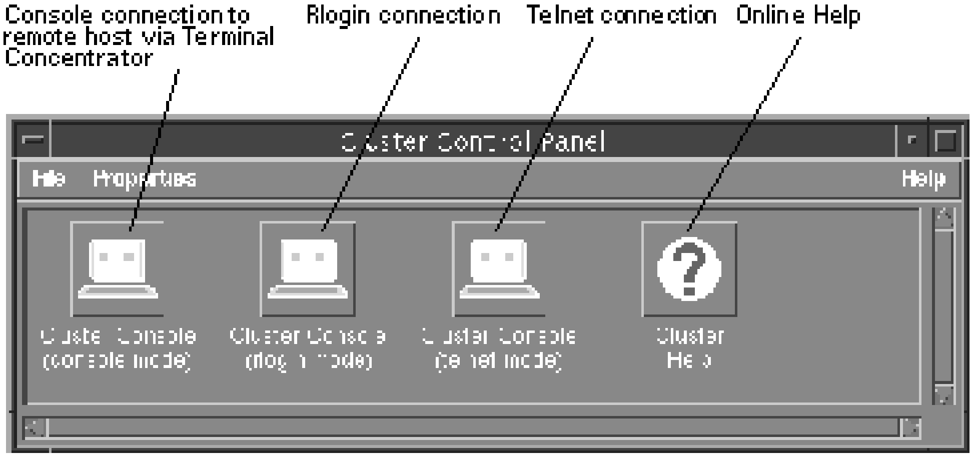

Console connection to remote host via Terminal Concentrator
Rlogin connection
Telnet connection
Online Help
Cluster Control Panel
File
Properties
Help
Cluster Console (console mode)
Cluster Console (rlogin mode)
Cluster Console (telnet mode)
Cluster Help

*Figure 2–3* Sample Cluster Control Panel

From the File or Properties menu you can:

- Add a new item
- Delete an item
- Modify an item

For detailed information about the CCP, refer to the online help.

For information about the programs represented by these tools and their usage, see “Cluster Console” on page 39. For information about using the HotJava browser to monitor cluster configurations, see “Using Sun Cluster Manager ” on page 42.

## Cluster Control Panel Configuration File Locations

The CCP stores properties and other information in configuration files within a configuration directory. By default, the configuration directory is located at `/opt/SUNWcluster/etc/ccp`.

---

**Note -** You must be `root` (superuser) to write to this default location. Only `root` can add, delete, or change CCP items in this configuration directory.

---

You can, however, create your own configuration directory and define its location using the environment variable `$CCP_CONFIG_DIR`. The `$CCP_CONFIG_DIR` variable specifies the configuration directory in which the configuration files containing item properties are stored. If the path name is not set, it defaults to the standard location, `/opt/SUNWcluster/etc/ccp`. To create your configuration directory, create a new directory and set the environment variable `$CCP_CONFIG_DIR` to the full path name of the new directory.

These files do not need to be edited manually, because they are created, modified, or deleted by `ccp` whenever you create, modify, or delete an item.

## Cluster Console

The Cluster Console (CC) GUI enables you to run commands on multiple nodes simultaneously, simplifying cluster administration. The Cluster Console displays one terminal window for each cluster node, plus a small Common window that you can use to control all windows simultaneously.

Different types of remote sessions enable you to connect to the console of the host, or remotely log in by using `rlogin` or `telnet`. Hosts can be specified on the command line and added or deleted from the Select Hosts dialog box after the program is running. The session type can be specified only on the command line. Once started, the session type cannot be changed.

You can issue commands to multiple hosts from the Common window, and you can issue commands to a single host from a terminal window. Terminal windows use VT100 terminal emulation.

Alternatively, you can turn off all hosts in the Hosts menu except the one you want to access, then issue commands from the Common window text field.

### ▼ How to Launch the Cluster Console

You can launch the Cluster Console from the CCP (see “Cluster Control Panel” on page 36) or from the command line in a shell window. If an optional parameter is specified, a terminal window is created for each host in the cluster or for each host specified.

1. **Type `cconsole` to initiate remote console access:**

```
% cconsole [clustername | hostname...]
```

1. **Type `ctelnet` to establish a `telnet(1)` connection from the console:**

```
% ctelnet [clustername | hostname...]
```

1. **Launch** `crlogin` **with your user name to establish an** `rlogin(1)` **connection from the console:**

```
% crlogin -l user name [clustername | hostname...]
```

All three of the preceding commands also take the standard X/Motif command-line arguments. Once the Cluster Console has started, the Console window is displayed.

For detailed information about the Cluster Console, refer to the online help.

## The Common Window Menu Bar

The Common window (shown in the following figure) is the primary window used to send input to all nodes. The Common window is always displayed when you launch the Cluster Console.

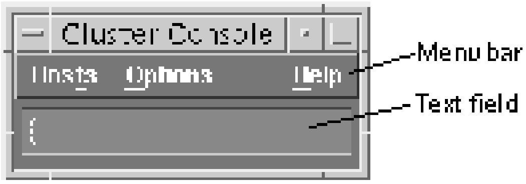


*Figure 2–4* Common Window Menu Bar of the Cluster Console

This window has a menu bar with three menus and a text field for command entry. From the Hosts menu you can use the Select dialog box to:

- Add a host
- Add all hosts in a cluster
- Remove a host

From the Options menu, you can group or ungroup the Common Window and the terminal windows.

## Configuration Files Used by the Cluster Console

Two configuration files are used by the Cluster Console: `clusters` and `serialports`. These can be `/etc` files or NIS/NIS+ databases. The advantage to using a NIS+ environment is that you can run the Cluster Console on multiple Administrative Workstations. Refer to your NIS/NIS+ system administration documentation for complete information about NIS/NIS+.

## The clusters File

The `clusters` file maps a cluster name to the list of host names that comprise the cluster. Each line in the file specifies a cluster, as in this example:

```
planets      mercury venus earth mars
wine         zinfandel merlot chardonnay riesling
```

The `clusters` file is used by all three session types of the Cluster Console (`cconsole`, `ctelnet`, and `crlogin`) to map cluster names to host names on the command line and in the Select Hosts dialog box. For additional information, see "Modifying the `clusters` File" on page 70.

## The serialports File

The `serialports` file maps a host name to the Terminal Concentrator and Terminal Concentrator serial port to which the host is connected. Each line in this database specifies a serial port of the host.

Sample `serialports` file database entries for the Sun Enterprise 10000 are:

```
mercury    systemserviceprocessorname    23
venus      systemserviceprocessorname    23
earth      systemserviceprocessorname    23
mars       systemserviceprocessorname    23
```

Sample `serialports` file database entries for all other nodes are:

```
mercury        planets-tc   5002
venus          planets-tc   5003
earth          planets-tc   5004
mars           planets-tc   5005
```

The `serialports` file is used only by the `cconsole` variation of this program to determine which Terminal Concentrator and port to connect to for hosts or clusters that are specified in the command line or the Select Hosts dialog box.

In the preceding example, node mercury is connected to planets-tc Port 2, while node venus is connected to planets-tc Port 3. Port 1 is reserved for the administration of the Terminal Concentrator.

For additional information, see “Modifying the serialports File” on page 71.

# Using Sun Cluster Manager

Sun Cluster Manager (SCM) is the cluster management tool for Sun Cluster 2.2. SCM provides a single interface to many of Sun Cluster’s command line monitoring features. SCM reports information about:

- SCM-monitored alarms
- syslog messages on each cluster node
- Cluster resources, including parallel data services, logical hosts, registered HA services, cluster nodes, and volume managers

SCM consists of two parts—the SCM server software and the SCM Graphical User Interface (GUI). The SCM server must run on each node in the cluster. The SCM GUI runs as either an application or an applet. If run as an applet, SCM can be displayed through a browser compliant with the Java Development Kit (JDK™), such as HotJava™ or Netscape. The browser can run on any machine, including the cluster nodes. For the most current information about which JDK versions are supported, see your service provider or the latest version of the *Sun Cluster 2.2 Release Notes*.

Use the information in the following sections to configure and run SCM.

## Running SCM As an Application

To run SCM as an application, perform the following procedure.

### ▼ How to Run SCM As an Application

1. **On the administrative workstation, install the SCM package (SUNWscmgr) from the Sun Cluster 2.2 product CD.**

2. **Install the latest version of the SCM patch on all cluster nodes and on the administrative workstation.**

   For the most current information about patches and patch numbers, see the *Sun Cluster 2.2 Release Notes*, your service provider, or the Sun patch web site, http://sunsolve.sun.com.

3. **Run the SCM application.**

   Start the application by running the following command from any cluster node, where *cluster_node* is a current member of the cluster.

```
# /opt/SUNWcluster/bin/scmgr cluster_node
```

Once the SCM application is running, you can access the online help for information about menu navigation, tasks, and reference. To display the Help window from the SCM application, select Help Contents from the Help menu. Alternatively, click on the Help icon in the tool bar above the folder icon.

Refer to the `scmgr(1M)` man page for more information about running SCM.

## Running SCM As an Applet

To run SCM as an applet, you must perform these tasks, which are described in more detail in the procedures that follow.

1. On all cluster nodes, install the latest version of the SCM patch from SunSolve. For the most current information about patches and patch numbers, see the *Sun Cluster 2.2 Release Notes*, your service provider, or the Sun patch web site, `http://sunsolve.sun.com`.
2. Install a browser on the administrative workstation. Supported browsers for Sun Cluster 2.2 4/00 Release are HotJava (version 1.1.4 or greater) and Netscape (version 4.5 or greater).
3. If you installed HotJava as your browser, install the JDK on the administrative workstation.
4. Install and configure a web server on all cluster nodes.
5. Start the SCM applet by typing the appropriate URL in the browser. Make sure the host you specify in the URL is a current member of the cluster.

---

**Note -** If you use the HotJava browser shipped with your Solaris 2.6 or 2.7 operating system, you might experience problems when using the menus. For example, after you make a menu selection, the selection might remain visible on the browser. Refer to the *Sun Cluster 2.2 Release Notes* for more information about SCM open issues. Solaris 8 software does not support a HotJava browser. To run SCM with Solaris 8 software, you must use another browser, such as Netscape. See “How to Run the SCM Applet in a Netscape Browser From a Cluster Node” on page 47.

---

**Note -** If you use the HotJava browser with SCM, you should have at least 40 Megabytes of free swap space. If you find that swap space gets low, restarting the HotJava browser can help.

To run SCM, you must have the correct versions of HotJava and the JDK packages (`SUNWjvrt` and `SUNWjvjit`) installed on the system containing the HotJava browser. Check your version numbers against those listed in the following table. Also see the latest version of the *Sun Cluster 2.2 Release Notes* for possible updates to this information:

**TABLE 2–5** SCM Requirements: JDK and HotJava

| Operating Environment | Java Developer Kit (JDK) version | HotJava Version | Netscape Version |
|---|---|---|---|
| Solaris 2.6 | 1.1.6 or later | 1.1.4 or later | 4.5 or later |
| Solaris 7 | 1.1.6 or later | 1.1.4 or later | 4.5 or later |
| Solaris 8 | 1.2 or later | not supported | 4.5 or later |

You can choose to:

- Run the HotJava browser on a cluster node. If you choose this option, you will have to restart HotJava on a different node if the node running HotJava crashes.
- Install a web server on each node in the cluster. If you choose this option, you will need to type an appropriate URL for another node into the HotJava browser if you see the `Lost Connection` dialog.

Use the following procedures as necessary to set up your preferred configuration.

## ▼ How to Set Up the JDK

1. **Determine the current Java version by typing the following at the console prompt on a cluster server.**

```
# java -version
```

2. **If necessary, download a later version of the JDK.**

If the system displays a version of Java lower than 1.1.6, follow the instructions to download the JDK version 1.1.6 (or later) software from the following URL:

`http://www.sun.com/solaris/java`

## How to Download HotJava

1. **From the machine running the HotJava browser, select About HotJava from the Help menu.**

   If the browser displays a version lower than 1.1.4, or if you do not have a HotJava browser, follow the instructions to download the HotJava software version 1.1.4 (or later) from the following URL:

   `http://java.sun.com/products/hotjava/index.html`

## How to Run the SCM Applet in a HotJava Browser From a Cluster Node

1. **Run your HotJava browser on a cluster node.**

   The HotJava browser is located in `/usr/dt/bin`.

2. **Remotely display the HotJava browser on an X windows workstation.**

3. **In the HotJava browser, set the applet security preferences:**

   a. **Choose Applet Security from Preferences on the Edit menu.**

   b. **Click Medium Security as the Default setting for Unsigned applets.**

4. **When you are ready to begin monitoring the cluster with SCM, type the appropriate URL. For example:**

```
file:/opt/SUNWcluster/scmgr/index.html
```

5. **Click OK on dialog boxes that ask for permission to access certain files, ports, and so forth.**

---

**Note -** HotJava takes some time to download and run the applet. No status information will appear during this time.

---

Refer to the online help for complete information about menu navigation, tasks, and reference.

## How to Run the SCM Applet in a HotJava Browser From the Administrative Workstation

1. **Run your HotJava browser on a node in the cluster.**

   The HotJava browser is located in `/usr/dt/bin`.

2. **Set up and start a web server on all cluser nodes.**

   For details, see "How to Set Up a Web Server to Run With SCM" on page 47.

3. **Set the applet security preferences in your HotJava browser:**

   a. **Choose Applet Security from Preferences on the Edit menu.**

   b. **Click Medium Security as the Default setting for Unsigned applets.**

4. **When you are ready to begin monitoring the cluster with SCM, type the appropriate URL.**

   ```
   http://cluster_node/link_to_scm/index.html
   ```

5. **Click OK on dialog boxes that ask for permission to access certain files, ports, and so forth from the remote display workstation to the cluster node on which the browser is started.**

   ---

   **Note -** HotJava takes some time to download and run the applet. No status information will appear during this time.

   ---

Refer to the online help for complete information about menu navigation, tasks, and reference.

## How to Run the SCM Applet in a Netscape Browser From a Cluster Node

1. **Install Netscape on the cluster nodes.**

2. **Install SCM and the required SCM patch on the cluster nodes.**

   To install SCM, use `scinstall(1M)`. The `scinstall(1M)` command installs the SCM package (`SUNWscmgr`) as part of the server package set. To get the SCM patch, see your service representative or the SunSolve web site:

   `http://sunsolve.sun.com/`

3. **Add the following lines to the `preferences.js` file, if necessary.**

   The file is located in the `$HOME/.netscape` directory. If the preferences are not included in the file already, add the following lines:

```
user_pref(‘‘security.lower_java_network_security_by_trusting_proxies’’, true);
user_pref(‘‘signed.applets.codebase_principal_support’’, true);
```

4. **On a cluster node, set your `DISPLAY` environment variable so that the Netscape browser is displayed remotely on your X Windows workstation, and then run the Netscape browser on that cluster node.**

5. **When you are ready to begin monitoring the cluster with SCM, enter the appropriate URL.**

```
file:/opt/SUNWcluster/scmgr/index.html
```

6. **Click Grant on Java Security dialog boxes that ask for permission to access certain files, ports, and so forth from the remote display workstation.**

   Refer to the online help for complete information about menu navigation, tasks, and reference.

## How to Set Up a Web Server to Run With SCM

If you choose, you can install a web server on the cluster nodes to run with SCM.

**Note -** If you are running the Sun Cluster HA for Netscape HTTP service and an HTTP server on SCM, you must configure the HTTP servers to listen on different ports, to prevent a port conflict between the HTTP servers.

1. **Install a web server on all nodes in the cluster.**

2. **Follow the web server's configuration procedure to make sure that SCM's `index.html` file is accessible to the clients.**

   The client applet for SCM is in the `index.html` file in the `/opt/SUNWcluster/scmgr` directory. For example, go to your HTTP server's `document_root` and create a link to the `/opt/SUNWcluster/scmgr` directory.

3. **Run the HotJava browser from your workstation.**

4. **Set the applet security preferences in your HotJava browser:**

   a. **Choose Applet Security from Preferences on the Edit menu.**

   b. **Click Medium Security as the Default setting for Unsigned applets.**

5. **When you are ready to begin monitoring the cluster with SCM, type the appropriate URL.**

   For example, if you had created a link from the web server's `document_root` directory to the `/opt/SUNWcluster/scmgr` directory, you would type the following URL:

   ```
   http://cluster_node/scmgr/index.html
   ```

6. **Click OK on dialog boxes that ask for permission to access certain files, ports, and so forth on the cluster node on which the browser is started.**

   **Note -** HotJava takes some time to download and run the applet. No status information will appear during this time.

Refer to the online help for complete information about menu navigation, tasks, and reference.

## Accessing SCM Online Help

SCM provides online help information about menu navigation, tasks, and reference. This help is available whether you are running SCM as an application or an applet.

To display the Help window from SCM, select Help Contents from the Help menu. Alternatively, click on the Help icon (question mark) in the tool bar above the folder.

If necessary, you can run online help in a separate browser by typing the following URL:

```
file:/opt/SUNWcluster/scmgr/help/locale/en/main.howtotopics.html
```

For example, if you had created a link from the web server's `document_root` directory to the `/opt/SUNWcluster/scmgr` directory, you would type the following URL:

```
http://clusternode/scmgr/help/locale/en/main.howtotopics.html
```

When you finish viewing the online help, close its HotJava browswer. Selecting online help again brings up a new browser and loads the help.

CHAPTER **3**

# Modifying the Sun Cluster Configuration

This chapter provides instructions on the following topics.

# Adding and Removing Cluster Nodes

When you add or remove cluster nodes, you must reconfigure the Sun Cluster software. When you installed the cluster originally, you specified the number of "active" nodes and the number of "potential" nodes in the cluster, using the `scinstall(1M)` command. Use the procedures in this section to add "potential" nodes and to remove "active" nodes.

To add nodes that were not already specified as potential nodes, you must halt and reconfigure the entire cluster.

## ▼ How to Add a Cluster Node

Use this procedure only for nodes that were already specified as "potential" during initial installation.

1. **Use the `scinstall(1M)` command to install Sun Cluster 2.2 on the node you are adding.**

   Use the installation procedures described in the *Sun Cluster 2.2 Software Installation Guide*, but note the following when responding to `scinstall(1M)` prompts:

   - When asked for the number of active nodes, include the node you are adding now in the total.
   - You will not be prompted for shared Cluster Configuration Database (CCD) information, since the new cluster will have greater than two nodes.
   - (VxVM with direct attached devices only) When prompted for the node lock port, provide the designated node lock device and port.
   - (VxVM only) Do not select a quorum device when prompted. Instead, select `complex` mode and then `N`. Later, you will run the `scconf -q` command to configure the quorum device.
   - (VxVM only) Select `ask` when prompted to choose a cluster partitioning behavior.

2. **(Scalable Coherent Interface [SCI] only) Update the `sm_config` template file to verify information about the new node.**

   This step is not necessary for Ethernet configurations.

   Nodes that were specified as "potential" during initial installation should have been included in the `sm_config` file with their host names commented out by the characters `_%`. Uncomment the name of the node you will be activating now. Make sure the configuration information in the file matches the physical layout of the node.

3. **(SCI only) Run `sm_config`.**

4. **(VxVM only) Set up the root disk group.**

   For details, see the VxVM appendix in the *Sun Cluster 2.2 Software Installation Guide*.

5. **(SDS only) Set up Solstice DiskSuite disksets.**

   For details, see the Solstice DiskSuite appendix in the *Sun Cluster 2.2 Software Installation Guide*.

6. **If you have a direct attached device connected to all nodes, set up the direct attached disk flag on the new node.**

   To set the direct attached flag correctly in the `cdb` files of all nodes, run the following command on all nodes in the cluster. In this example, the cluster name is `sc-cluster`:

   ```
   # scconf sc-cluster +D
   ```

7. **(VxVM only) Select a common quorum device.**

   If your volume manager is VxVM and if you have a direct attached device connected to all nodes, run the following command on all nodes and select a common quorum device.

   ```
   # scconf sc-cluster -q -D
   ```

   If you do not have a direct attached disk connected to all nodes, run the following command for each pair of nodes that shares a quorum device with the new node.

   ```
   # scconf -q
   ```

8. **(VxVM only) Set the node lock port on the new node.**

   If you just installed a direct attached disk, set the node lock port on all nodes.

   If the cluster contained a direct attached disk already, run the following command on only the new node. In this example, the cluster name is `sc-cluster` and the Terminal Concentrator is `cluster-tc`.

   ```
   # scconf sc-cluster -t cluster-tc -l port_number
   ```

9. **Stop the cluster.**

10. **Run the** scconf -A **command on all nodes to update the number of active nodes.**

    See the scconf(1M) man page for details. In this example, the cluster name is sc-cluster and the new total number of active nodes is three.

    ```
    # scconf sc-cluster -A 3
    ```

11. **(VxVM only) Remove the shared CCD if it exists, since it is necessary only with two-node clusters.**

    Run the following command on all nodes.

    ```
    # scconf sc-cluster -S none
    ```

12. **Use** ftp **(in binary mode) to copy the** cdb **file from an existing node to the new node.**

    The cdb files normally reside in /etc/opt/SUNWclus/conf/*clustername*.cdb.

13. **Reboot the new node.**

14. **Start the cluster.**

    Run the following command from any single node.

```
# scadmin startcluster phys-hahost sc-cluster
```

Then run the following command on all other nodes.

```
# scadmin startnode
```

## ▼ How to Remove a Cluster Node

The scconf(1M) command enables you to remove nodes by decrementing the number you specified as active nodes when you installed the cluster software with the scinstall(1M) command. For this procedure, you must run the scconf(1M) command on all nodes in the cluster.

1. **For an HA configuration, switch over all logical hosts currently mastered by the node to be removed.**

   For parallel database configurations, skip this step.

```
# haswitch phys-hahost3 hahost1
```

2. **Run the** `scconf -A` **command to exclude the node.**

   Run the `scconf(1M)` command on all cluster nodes. See the `scconf(1M)` man page for details.

   ---

   **Note -** In this command, the number you specify does not represent a node number. Instead, this number represents the total number of cluster nodes that will be active after the `scconf` operation. The `scconf` operation always removes from the cluster the node with the highest node number. There is no procedure to remove, for example, node number 2 from a three-node cluster.

   ---

In this example, the cluster name is `sc-cluster` and the total number of active nodes after the `scconf` operation is two.

```
# scconf sc-cluster -A 2
```

# Changing the Name of a Cluster Node

The names of the cluster nodes can be changed using the `scconf(1M)` command. Refer to the `scconf(1M)` man page for additional information.

## ▼ How to Change the Name of a Cluster Node

1. **Find the names of the current cluster nodes.**

   You can run the `scconf -p` command on any node that is an active cluster member.

```
# scconf clustername -p
Current Configuration for Cluster clustername:
       Hosts in cluster: phys-hahost1 phys_hahost2 phys-hahost3
       Private Network Interfaces for
           phys-hahost1: be0 be1
```

**(continued)**

(Continuation)

```
            phys-hahost2: be0 be1
            phys-hahost3: hme0 hme1
```

2. **Run the** `scconf -h` **command on all nodes in the cluster.**

   Run the `scconf(1M)` command on all nodes. See the `scconf(1M)` man page for details.

   ```
   # scconf -h clustername hostname0 [...hostname3]
   ```

   The new node names should be specified in the order shown by the `scconf -p` command. For example, to change the name of `phys-hahost3` to `phys-opshost1`, you would run the following command on all cluster nodes.

   ```
   # scconf -h sccluster phys-hahost1 phys-hahost2 phys-opshost1
   ```

# Changing the Private Network Interfaces

The private network interfaces of the nodes in the cluster can be changed using the `scconf(1M)` command. Refer to the `scconf(1M)` man page for additional information.

## ▼ How to Change the Private Network Interfaces

1. **Use the** `scconf(1M)` **command for all nodes in the cluster.**

   For example:

   ```
   # scconf planets -i mercury scid0 scid1
   # scconf planets -i venus   scid0 scid1
   # scconf planets -i pluto   scid0 scid1
   # scconf planets -i jupiter scid0 scid1
   ```

(Continuation)

After running these commands, all four nodes mercury, venus, pluto, and jupiter use interfaces scid0 and scid1.

**Caution -** While the cluster is up, you should not use the ifconfig(1M) command. This action causes unpredictable behavior on a running system.

# Printing the Cluster Configuration

The cluster configuration information can be printed using scconf(1M). Refer to the scconf(1M) man page for more information.

## How to Print the Cluster Configuration

1. **Use the scconf(1M) command on any node that is an active cluster member.**

   For example:

```
# scconf planets -p
```

A response similar to the following is displayed. (Depending on your type of private interconnect, your messages may state hme instead of scid.)

```
Current Configuration for Cluster planets:
  Hosts in cluster: mercury venus pluto jupiter

  Private Network Interfaces for
      mercury: scid0 scid1
      venus: scid0 scid1
      pluto: scid2 scid3
      jupiter: scid2 scid3
```

(continued)

# Adding and Removing Logical Hosts

Logical hosts are the objects that fail over if a node fails. Each logical host is composed of a disk group or groups, a relocatable IP address, and a logical host name. Logical hosts are only used in configurations with HA data services. There are no logical hosts in a parallel database configuration.

You add or remove logical hosts by updating the logical host configuration information stored by the cluster, and then reconfiguring the cluster. When you configure the cluster originally, you provide `scinstall(1M)` with information about your logical host configuration. Once the cluster is up, there are two ways to change this information:

- Rerun the `scinstall(1M)` command – The `scinstall(1M)` command provides a menu-based interface to the `scconf(1M)` command and is the recommended way to modify your logical host configuration. You must run `scinstall(1M)` as `root`.
- Run the `scconf(1M)` command – If you choose to use the `scconf(1M)` command, refer to the man page for options and other information. If you are going to set up a logical host with more than one disk group, you must use the `scconf(1M)` command to configure it.

## ▼ How to Add a Logical Host to the Cluster

As part of the process of adding a logical host, you will be asked to provide the following information:

- The names of the primary public network controllers for the cluster nodes.
- Whether the cluster serves any secondary public subnets.
- Whether you want to initialize Public Network Management (PNM) on the current cluster node. You would only re-initialize PNM if you have added a network

controller, or changed your controller configuration while adding the new logical host.

- The name of the new logical host.
- The name of the default master for the new logical host.
- The name of the disk group included in the logical host.
- Whether to enable automatic failback for the new logical host. Automatic failback means that in the event that this logical host fails over to a backup node, it will be remastered by its default master once the failed node is back in the cluster. See "Disabling Automatic Switchover" on page 90 for details.
- The disk group name for the new logical host.

Gather the answers to these questions before starting the procedure. Note that you must have already set up your disk group to be used by the new logical host. Refer to the appropriate appendix in the *Sun Cluster 2.2 Software Installation Guide* for your volume manager for details.

Use the following procedure to add a logical host to a cluster.

1. **Run the `scinstall(1M)` command and select the Change option from the main menu.**

```
# scinstall
Assuming a default cluster name of planets
Note: Cluster planets is currently running.
      Install/Uninstall actions are disabled during
      cluster operation.

        <<Press return to continue>>

        Checking on installed package state
........................

============ Main Menu =================

1) Change  - Modify cluster or data service configuration.
2) Verify  - Verify installed package sets.
3) List    - List installed package sets.

4) Quit    - Quit this program.
5) Help    - The help screen for this menu.

Please choose one of the menu items: [5]:  1
```

2. **From the Change menu, choose the Logical Hosts option.**

```
=========== Changes Menu ================

Choices from this menu:

1) Logical Hosts       - Change the logical hosts configuration.
2) NAFO                - Re-initialize the NAFO configuration.

3) Close  - Close this menu to return to the Main Menu.
4) Quit   - Exit this program.
5) Help   - Show the Help screen.

Please choose a displayed option: [5] 1
```

This will display the Logical Hosts Configuration menu.

**3. From the Logical Hosts Configuration menu, select the Add option.**

```
====== Logical Hosts Configuration ======

1) Add     - Add a logical host to the cluster.
2) Remove  - Remove a logical host from the cluster.
3) List    - List the logical hosts in the cluster.

4) Close   - Return to the previous menu.
5) Quit    - Exit.


Please choose an option:  1
```

You will be asked a series of questions regarding the new logical host.

**4. Respond to the prompts with the required information.**

Once the `scinstall(1M)` portion of this procedure is complete, you will be returned to the Logical Hosts Configuration menu.

```
What is the primary public network controller for ‘‘phys-hahost1’’?
What is the primary public network controller for ‘‘phys-hahost2’’?
Does the cluster serve any secondary public subnets (yes/no) [no]?
Re-initialize NAFO on ‘‘phys-hahost1’’ with one ctlr per group

(yes/no)?
What is the name of the new logical host?  hahost1
What is the name of the default master for ‘‘hahost1’’? phys-hahost1
Enable automatic failback for ‘‘hahost1’’ (yes/no) [no]?
```

**(continued)**

(Continuation)

```
Disk group name for logical host ‘‘hahost1’’ [hahost1]?
Is it okay to add logical host ‘‘hahost1’’ now (yes/no) [yes]?
/etc/opt/SUNWcluster/conf/ha.cdb
Checking node status...
```

5. **Create a new HA administrative file system and update the** `/etc/opt/SUNWcluster/conf/hanfs/vfstab.`***logicalhost*** **file.**

   When you add a new logical host, you must set up a file system on a disk group within the logical host to store administrative information. The steps for setting up the HA administrative file system differ depending on your volume manager. These steps are described in the volume manager appendixes of the *Sun Cluster 2.2 Software Installation Guide*.

---

**Note -** Do not use host name aliases for the logical hosts. NFS clients mounting Sun Cluster file systems using logical host name aliases might experience `statd` lock recovery problems.

---

## ▼ How to Remove a Logical Host From the Cluster

To remove a logical host from the cluster configuration, the cluster must be up and there must be no data services registered for the logical host.

1. **Stop all data service applications running on the logical host to be removed.**

```
# hareg -n dataservice
```

2. **Unregister the data service.**

```
# hareg -u dataservice
```

3. **Remove the logical host from the cluster.**

   Run the `scinstall(1M)` command as described in the *Sun Cluster 2.2 Software Installation Guide* and select the Change option from the main menu.

```
# scinstall
Assuming a default cluster name of planets
Note: Cluster planets is currently running.
      Install/Uninstall actions are disabled during
      cluster operation.

        <<Press return to continue>>

        Checking on installed package state
........................

============ Main Menu =================

1) Change  - Modify cluster or data service configuration.
2) Verify  - Verify installed package sets.
3) List    - List installed package sets.

4) Quit    - Quit this program.
5) Help    - The help screen for this menu.

Please choose one of the menu items: [5]:  1
```

4. **From the Change menu, select the Logical Hosts option.**

```
=========== Changes Menu ================

Choices from this menu:

1) Logical Hosts       - Change the logical hosts configuration.
2) NAFO                - Re-initialize the NAFO configuration.

3) Close  - Close this menu to return to the Main Menu.
4) Quit   - Exit this program.
5) Help   - Show the Help screen.

Please choose a displayed option: [5] 1
```

This will display the Logical Hosts Configuration menu.

5. **From the Logical Hosts Configuration menu, select the Remove option.**

```
====== Logical Hosts Configuration ======

1) Add     - Add a logical host to the cluster.
2) Remove  - Remove a logical host from the cluster.
3) List    - List the logical hosts in the cluster.

4) Close   - Return to the previous menu.
5) Quit    - Exit.

Please choose an option:  2
```

This displays a list of configured logical hosts.

6. **Enter the name of the logical host to be removed from the list of configured logical hosts.**

```
The list of logical hosts is:

        hahost1
        hahost2

Which one do you want to remove?  hahost1
```

The procedure is now complete and you are returned to the Logical Hosts Configuration menu.

7. **As `root`, delete the `/etc/opt/SUNWcluster/conf/hanfs/vfstab`.*logicalhost* file that was created when the logical host was added to the cluster configuration.**

## Changing the Logical Host IP Address

You can change a logical host IP address by removing and adding the logical host with the new IP address by using the procedures in “Adding and Removing Logical Hosts” on page 58, or by using the procedure in this section.

Refer to the `scconf(1M)` man page for more information.

### ▼ How to Change the Logical Host IP Address

The following steps must be done on a single node that is a cluster member.

1. **Remove the existing logical host entry from the configuration files by running the following command on all nodes.**

   ```
   # scconf clustername -L logicalhost -r
   ```

2. **Create a new logical host entry, which uses the same logical host name but with the new IP address, by running the following command on all cluster nodes.**

   ```
   # scconf clustername -L logicalhost -n nodelist -g diskgroup -i interfaces_and_IP
   ```

## Forcing a Cluster Reconfiguration

You can force a cluster reconfiguration by using the `haswitch(1M)` command, or by changing the cluster membership by using the `scconf(1M)` command.

### ▼ How to Force a Cluster Reconfiguration

1. **Force a cluster reconfiguration by running the** `haswitch(1M)` **command on any node that is a cluster member. For example:**

   ```
   # haswitch -r
   ```

   See the `haswitch(1M)` man page for details.

# Configuring Sun Cluster Data Services

This section provides procedures for configuring Sun Cluster data services. These data services are normally configured with the logical hosts as part of the cluster installation. However, you can also configure logical hosts and data services after the installation. For more detailed information about a particular data service, see the individual chapter covering the data service in the *Sun Cluster 2.2 Software Installation Guide*.

---

**Note -** All commands used in this section can be run from any node that is a cluster member, even a node that cannot master the specified logical hosts or run the specified data services. You can run the commands even if there is only one node in the cluster membership.

---

**Caution -** The commands used in this section update the CCD even without a quorum. Consequently, updates to the CCD can be lost if nodes are shut down and brought up in the wrong sequence. Therefore, the node that was the last to leave the cluster must be the first node brought back into the cluster by using the `scadmin startcluster` command. For more information about the CCD, see the *Sun Cluster 2.2 Software Installation Guide*.

---

## ▼ How to Configure a Sun Cluster Data Service

1. **Verify that the following tasks have been completed.**

   - The logical hosts that will run the data services are configured. Refer to "Adding and Removing Logical Hosts" on page 58, for details on configuring a logical host.
   - The necessary disk groups, logical volumes, and file systems are set up. Refer to the *Sun Cluster 2.2 Software Installation Guide* for details.
   - The HA administrative file system and `vfstab`.*logicalhost* file have been set up. This procedure varies depending on your volume manager. Refer to the appendix describing how to configure your volume manager in the *Sun Cluster 2.2 Software Installation Guide*.

2. **Register the data service.**

   Register the Sun Cluster data service(s) associated with the logical host(s).

```
# hareg -s -r dataservice [-h logicalhost]
```

This assumes that the data service has already been installed and its methods are available.

When the `-h` option is used in the `hareg -r` command, the data service is configured only on the logical hosts specified by the *logicalhost* argument. If the `-h` option is not specified, then the data service is configured on all currently-existing logical hosts. See the `hareg(1M)` man page for details.

---

**Note -** If the data service is to be associated with any logical hosts that are created after the registration of the data service, run `scconf -s` on all cluster nodes to extend the set of logical hosts associated with the data service.

---

3. **Configure the data service using the interactive command** `hadsconfig(1M)` **and your data service documentation.**

```
# hadsconfig
```

4. **Start the data service.**

```
# hareg -y dataservice
```

## Unconfiguring Sun Cluster Data Services

Use this procedure to unconfigure Sun Cluster data services. For more detailed information about a data service, see the individual chapter covering the data service in the *Sun Cluster 2.2 Software Installation Guide.*

### How to Unconfigure Sun Cluster Data Services

1. **Stop all data service applications to be unconfigured.**

   Perform your normal shutdown procedures for the data service application.

2. **If the data service is a database management system (DBMS), stop all fault monitors.**

3. **Stop the data service on all logical hosts.**

```
# hareg -n dataservice
```

4. **Unregister the data service.**

```
# hareg -u dataservice
```

---

**Note -** If the `hareg -u` command fails, it can leave the Cluster Configuration Database (CCD) in an inconsistent state. If this occurs, run `scconf` *clustername* `-R` *dataservice* on all cluster nodes to forcibly remove the data service from the CCD.

---

5. **(Optional) Remove the logical hosts from the cluster configuration.**

   You can remove a logical host from the cluster configuration only if all data services are disassociated from it.

   Use one of the following methods to remove a logical host.

   Run this `scconf(1M)` command on one node that is a cluster member:

```
# scconf clustername -L logicalhost -r
```

   Or run the `scinstall(1M)` command as described in “Adding and Removing Logical Hosts” on page 58. If you use `scinstall(1M)`, you do not need to perform the cluster reconfiguration shown in Step 6 on page 67.

6. **Perform a cluster reconfiguration by using the** `haswitch(1M)` **command.**

```
# haswitch -r
```

   At your discretion, you may choose to remove or rename the `vfstab.`*logicalhost* and `dfstab.`*logicalhost* files associated with the logical host you removed, and reclaim the space occupied by its volumes and file systems. These files are untouched by the `scconf(1M)` remove operation.

# Adding Sun Cluster Data Services

You can add a data service to an existing cluster by running the `scinstall(1M)` command. See Chapter 3 in the *Sun Cluster 2.2 Software Installation Guide* for details.

To add a data service to a two-node cluster with shared CCD, additional steps apply. Use the following procedure to add a data service to such a cluster.

## ▼ How to Add a Data Service to a Two-Node Cluster With Shared CCD

1. **Unshare the shared CCD.**

   You must reconfigure the cluster to unshare the CCD before you add any new data services. Run the following command on both nodes, as root, while both nodes are in the cluster:

   ```
   phys-hahost1# /opt/SUNWcluster/bin/scconf clustername -S none
   phys-hahost2# /opt/SUNWcluster/bin/scconf clustername -S none
   ```

   ---

   **Note -** You must unshare the CCD. If you attempt to add a data service while the CCD is in shared state, only the local `ccd.database` file will be updated, and not the shared CCD file. This attempt will cause registration of the new data service to fail.

   ---

2. **Add the new data services, using the following commands.**

   Run all commands as root. In these examples, the node names are `phys-hahost1` and `phys-hahost2`.

   a. **Stop the cluster on the first node.**

   ```
   phys-hahost1# scadmin stopnode
   ```

   b. **Use `scinstall(1M)` to add the new data service package to the first node.**

   See Chapter 3 of the *Sun Cluster 2.2 Software Installation Guide* for details. This step automatically updates the local CCD file.

```
phys-hahost1# scinstall
```

c. **Stop the cluster on the second node.**

**Note -** The existing data services will be unavailable to clients after you stop the cluster on the second node and until you restart the cluster on the first node.

```
phys-hahost2# scadmin stopnode
```

d. **Restart the cluster on the first node.**

```
phys-hahost2# scadmin startcluster phys-hahost1 clustername
```

e. **Use** `scinstall(1M)` **to add the new data service package to the second node. See Chapter 3 of the *Sun Cluster 2.2 Software Installation Guide* for details. This step automatically updates the local CCD file.**

```
phys-hahost2# scinstall
```

f. **Add the second node to the cluster.**

```
phys-hahost2# scadmin startnode
```

3. **Reinstate the shared CCD.**

Run the `scconf(1M)` command on both nodes, as root. See the `scconf(1M)` man page for details.

```
phys-hahost1# /opt/SUNWcluster/bin/scconf clustername -S ccdvol
phys-hahost2# /opt/SUNWcluster/bin/scconf clustername -S ccdvol
```

Then run the `confccdssa(1M)` command on one node only, as root. See the `confccdssa(1M)` man page for details.

```
phys-hahost1# /opt/SUNWcluster/bin/confccdssa
```

## Modifying the `clusters` File

The `/etc/clusters` file contains information about the known clusters in the local naming domain. This file, which maps a cluster name to the list of host names in the cluster, can be created as an NIS or NIS+ map, or locally in the `/etc` directory.

The `/etc/clusters` file requires updating only if:

- Any host names change
- More clusters are added to your Sun Cluster

For more information about NIS and NIS+ maps, refer to the *NIS/NIS+ System Administration Guide*. Refer to the *Sun Cluster 2.2 Software Installation Guide* for information about creating the `/etc/clusters` file. NIS/NIS+ files must be changed on the NIS/NIS+ server.

### ▼ How to Modify the `clusters` File

1. **Edit the `/etc/clusters` file to add the cluster name and physical host names of all nodes.**

   For example, to create a cluster named `hacluster` that consists of node 0 `phys-hahost1`, node 1 `phys-hahost2`, node 2 `phys-hahost3`, and node 3 `phys-hahost4`, add this entry to the file:

   ```
   # Sun Enterprise Cluster nodes
     hacluster phys-hahost1 phys-hahost2 phys-hahost3 phys-hahost4
   ```

   The `/etc/clusters` files must be identical on all nodes. Make the same modifications on all nodes.

### ▼ How to Create the `clusters` Table

1. **In an NIS+ environment, you must create a `clusters` table. The entries in this table are the same as the entries in the `/etc/clusters` file.**

   For example, to create a `clusters` table in a domain named *mydomain* in an NIS+ environment, use the following command:

```
# nistbladm -c key-value key=SI value= clusters.mydomain.
```

---

**Note -** The trailing period (`.`) at the end of the `nistbladm` command is required.

---

## Modifying the `serialports` File

The `serialports` file maps a host name to the Terminal Concentrator and Terminal Concentrator serial port to which the console of the host is connected. This file can be created as an NIS or NIS+ map or locally in the `/etc` directory.

The `serialports` file requires updating only if:

- The host name(s) changes
- The Terminal Concentrator name changes
- The port number of the host on the Terminal Concentrator changes
- Additional hosts are being added to this Terminal Concentrator
- More cluster nodes are added

Refer to the *Sun Cluster 2.2 Software Installation Guide* for information about creating the `/etc/serialports` file. For more information about NIS and NIS+ maps, refer to the *NIS/NIS+ System Administration Guide.*

### ▼ How to Modify the `serialports` File

1. **As root, create a `serialports` file in the `/etc` directory.**

2. **For a Sun Enterprise™ 10000 system, enter *hostname sspname* `23` in the `serialports` file. For all other hardware systems, enter *hostname terminal_concentrator serial_port* in the `serialports` file.**

For Sun Enterprise 10000:

```
# Sun Enterprise Cluster nodes
  phys-hahost1 sspname 23
  phys-hahost2 sspname 23
  phys-hahost3 sspname 23
  phys-hahost4 sspname 23
```

For all other hardware systems:

```
# Sun Enterprise Cluster nodes
  phys-hahost1 hacluster-tc    5002
  phys-hahost2 hacluster-tc    5003
  phys-hahost3 hacluster-tc    5004
  phys-hahost4 hacluster-tc    5005
```

## ▼ How to Create the `serialports` Table

1. **In an NIS+ environment, you need to create a** `serialports` **table. The entries in this table are the same as the entries in the** `/etc/serialports` **file.**

   To create a `serialports` table in a domain named *mydomain* in an NIS+ environment, use the following command:

```
# nistbladm -c key-value key=SI value=clusters.mydomain.
```

---

**Note -** The trailing period (`.`) at the end of the `nistbladm` command is required.

---

## Changing TC/SSP Information

When installing Sun Cluster software, information about the Terminal Concentrator (TC) or a System Service Processor (SSP) is required. This information is stored in the cluster configuration database (CCD).

This information is used to:

- Forcibly terminate hung nodes (as in failure fencing)
- Implement a cluster-wide locking mechanism which prevents partitioned nodes from joining the cluster

Both these mechanisms serve to protect data integrity in the case of four-node clusters with directly attached storage devices.

Use the `scconf(1M)` command to change the TC or SSP information associated with a particular node, as described in the following procedures.

For more information about installing and configuring the TC or SSP, see the terminal concentrator chapter in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide.*

### How to Change TC/SSP Information

To change TC or SSP information, run the `scconf(1M)` command on all cluster nodes. For each node, supply the appropriate new information. The following examples show `scconf(1M)` command syntax for each type of information change.

1. **Node architecture type and IP address – Supply the cluster name, the host name, the new architecture type, and the new IP address.**

```
# scconf clustername -H hostname -d E10000 -t new_ip_address
```

---

**Note -** Multiple hosts may be connected to the same TC; the `-H` option affects only the information associated with the host you specify in the command line.

---

2. **Password for a TC or SSP – Supply the cluster name, the IP address, and the new password.**

```
# scconf clustername -t ip_address -P
ip_address (129.34.123.51) Password:
```

3. **Port number for an SSP console – Supply the cluster name, host name, and new port number.**
   - If a Terminal Concentrator is being used, specify an unused TC port from `1` to *N*.
   - If an SSP is being used, a value of `-1` must be specified.

```
# scconf clustername -H hostname -p new_port_number
```

4. **TC name or IP address – Supply the cluster name, host name, and new TC name or IP address.**

```
# scconf clustername -H hostname -t new_tc_name|new_ip_address
```

For additional information about changing TC or SSP information, see the `scconf(1M)` man page and Chapter 8.

# Changing the Quorum Device

A quorum device is used only in VxVM configurations. It is not used in Solstice DiskSuite configurations.

The `scconf -q` command can be used to change the quorum device to either a disk or a controller. This option is useful if the quorum device needs servicing. Refer to the `scconf(1M)` man page for details.

**Note -** If the quorum device is a disk, the `scconf -q` command *must* be used whenever the disk address (in the form c*x*t*y*d*z*s2) changes, even if the serial number of the disk is preserved. This change in disk address can happen if the SBus slot of the drive controller changes.

**Caution -** Do not use the `scconf -q` option to modify the quorum device topology while the cluster is running. You cannot add or remove a quorum device between any two cluster nodes. Specifically, between a pair of nodes in the cluster, you cannot add a quorum device when there previously was no quorum device, and you cannot specify "no quorum device" if there currently is a quorum device. However, you can change a quorum device (for example, from a disk to another disk) in a running cluster by using the `scconf -q` option.

## ▼ How to Change the Quorum Device

1. **Before servicing the device, you can change the quorum device to a different device by running the `scconf -q` command on all cluster nodes.**

   For example, to change the quorum device in the cluster `haclust` for nodes phys-hahost1 and `phys-hahost2`, run the `scconf(1M)` command as shown below.

```
# scconf haclust -q phys-hahost1 phys-hahost2
Select quorum device for nodes 0 (phys-hahost1) and 1 (phys-
hahost2).
Type the number corresponding to the desired selection.
For example: 1<CR>

 1) DISK:c2t2d0s2:01943825
 2) DISK:c2t3d0s2:09064321
 3) DISK:c2t4d0s2:02171369
 4) DISK:c2t5d0s2:02149886
 5) DISK:c2t8d0s2:09062992
 6) DISK:c2t9d0s2:02166472
 7) DISK:c3t2d0s2:02183692
 8) DISK:c3t3d0s2:02183488
 9) DISK:c3t4d0s2:02160277
10) DISK:c3t5d0s2:02166396
11) DISK:c3t8d0s2:02164352
12) DISK:c3t9d0s2:02164312
Quorum device: 12
```

The `-q` option probes the list of devices attached to each node and lists the devices that the two nodes share. The quorum device can then be selected from this list.

To enable probing of devices attached to remote hosts, the local `/.rhosts` file is modified to enable `rsh(1)` permissions. The permissions are removed after the command completes.

**Note -** This behavior occurs only if this command is run from all the nodes at the same time. If you do not want remote root access capability, use the -m option.

2. **You may choose either an SSA controller or a disk from this list as the quorum device.**

   If you choose an SSA controller, the list of disks in that controller is displayed.

3. **If you chose an SSA controller in Step 2 on page 76, you are given the option to select a disk from this SSA as the quorum device.**

   If no disk is chosen in this step, the SSA controller chosen in the previous step is retained as the quorum device.

   The -q option also checks for the case where a node might have a reservation on the quorum device, due to some other node not being part of the membership. In this case, the -q option releases the reservation on the old quorum device and reserves the new quorum device.

**Note -** All the specified nodes must be booted for the scconf -q command to run successfully. If any of the nodes is not booted, the command probes and presents the list of all devices on the local node. Be sure to select a shared device as the quorum device.

If you already know the name of the device to use as the quorum device, use the -m option to specify the new device.

```
# scconf clustername -q -m quorum-device hostname1 hostname2
```

The quorum device can either be an SSA controller's World Wide Name (WWN), or a disk identifier of the form *WWN*.*disk-serial-id* for disks in SSAs, or a disk identifier of the form *disk-address*:*disk-serial-id* for non-SSA disks. The disk-address is of the form c*x*t*y*d*z*s2. You can use the finddevices(1M) command to obtain the serial numbers of SSA or non-SSA disks.

If you have a cluster with more than two nodes where all nodes share a common quorum device, you can use the -q -D options to specify a new common quorum device.

```
# scconf clustername -q -D
```

Since all the hosts in the cluster share a common device, specifying a list of hosts is unnecessary.

This is an interactive option that probes the list of devices attached to each host and then presents the list of shared devices. Select the quorum device from this list.

---

**Note -** All the active hosts defined in the cluster must be booted for the `scconf -q -D` command to be successful. If any of the hosts are not booted, the command probes and presents the list of all devices on the local host. Be sure to select a shared device as the quorum device.

---

The `-q -D` combination also checks for the case where a node of the cluster may have a reservation on the quorum device, due to some other node not being part of the cluster. In this case, the reservation on the old quorum device is released and the new quorum device is reserved.

If this command is run from all the nodes at the same time via the `cconsole` and `crlogin` GUI interfaces, then the local `/.rhosts` file is modified to enable `rsh(1)` permissions. This enables the probing of devices attached to remote hosts. The permissions are removed after the command completes.

The `-m` option can be added if remote root access capability is not desired. The `-m` option configures the quorum device and is given as the last argument to the command for the specified nodes.

```
# scconf clustername -q -D -m quorum-device
```

The quorum device is a disk identifier of the form `cxtydzs2`:*disk-serial-ID*. Use the `finddevices(1M)` command to obtain the serial numbers of disks.

## Configuring Timeouts for Cluster Transition Steps

Sun Cluster has configurable timeouts for the cluster transition steps where logical hosts of the HA framework are taken over and given up as cluster membership changes. Adapt these timeouts as needed to effectively handle configurations consisting of large numbers of data services on each node. It is impractical to have constant timeout values for a wide variety of configurations, unless the timeouts are set to a very large default value.

There are essentially two considerations when tuning timeouts:

- Number of logical hosts per cluster node

- Number of data services on a logical host

It is difficult to estimate what the correct value should be for a particular installation. These values should be arrived at by trial and error. You can use as guidelines the cluster console messages related to the beginning and end of each cluster transition step. They should give you a fairly good idea of how long a step takes to execute.

The timeouts need to account for the worst case scenario. When you configure cluster timeouts, take into consideration the maximum number of logical hosts that a cluster node can potentially master at any time.

For example, in an N+1 configuration, the standby node can potentially master all the logical hosts of the other cluster nodes. In this case, the reconfiguration timeouts must be large enough to accommodate the time needed to master all of the logical hosts configured in the cluster.

## ▼ How to Adjust Cluster Timeouts

1. **Adjust the cluster reconfiguration timeouts by using the `scconf -T` command.**

   For example, to change the configurable transition step timeout values to 500 seconds, you would run the following command on all cluster nodes.

```
# scconf clustername -T 500
```

The default values for these steps are 720 seconds. Use the `ssconf -p` command to see the current timeout values.

Within the reconfiguration steps, the time taken to master a single logical host can vary depending on how many data services are configured on each logical host. If there is insufficient time to master a logical host—if the `loghost_timeout` parameter is too small—messages similar to the following appear on the console:

```
ID[SUNWcluster.ccd.ccdd.5001]: error freeze cmd =
command /opt/SUNWcluster/bin/loghost_sync timed out.
```

The cluster framework makes a “best effort” to bring the system to a consistent state by attempting to give up the logical host. If this is not successful, the node may abort from the cluster to prevent inconsistencies.

2. **Use the `scconf -l` option to adjust the `loghost_timeout` parameter.**

   The default is 180 seconds.

**Note -** The reconfiguration step timeouts can never be less than the `loghost_timeout` value. Otherwise, an error results and the cluster configuration file is not modified. This requirement is verified by the `scconf -T` or `scconf -l` options. A warning is printed if either of these timeouts is set to 100 seconds or less.

CHAPTER **4**

# General Sun Cluster Administration

This chapter provides instructions for the following topics.

## Starting the Cluster and Cluster Nodes

The `scadmin startcluster` command is used to make a node the first member of the cluster. This node becomes node 0 of the cluster. The other Sun Cluster nodes are started by a single command, `scadmin startnode`. This command starts the programs required for multinode synchronization, and coordinates integration of the other nodes with the first node (if the Sun Cluster software is already running on the first node). You can remove nodes from the cluster by using the `scadmin` command with the `stopnode` option on the node that you are removing from the cluster.

Make the local node the first member node in the cluster. This node must be a configured node of the cluster in order to run the `scadmin startcluster` command successfully. This command *must* complete successfully before any other nodes can join the cluster. If the local node aborts for any reason while the subsequent nodes are joining the cluster, the result might be a corrupted CCD. If this scenario occurs, restore the CCD using the procedure “How to Restore the CCD” on page 101.

To make the local node a configured cluster node, see “Adding and Removing Cluster Nodes” on page 52.

## How to Start the Cluster

It is important that no other nodes are running the cluster software at this time. If this node detects that another cluster node is active, the local node aborts.

1. **Start the first node of the cluster by using the `scadmin(1M)` command.**

```
# scadmin startcluster localnode clustername
```

The `startcluster` option does not run if *localnode* does not match the name of the node on which the command runs. See the `scadmin(1M)` man page for details.

For example:

```
phys-hahost1# scadmin startcluster phys-hahost1 haclust
Node specified is phys-hahost1
Cluster specified is haclust

=========================== WARNING============================
=                     Creating a new cluster                  =
===============================================================

You are attempting to start up the cluster node "phys-hahost1" as
the only node in a new cluster.  It is important that no other
cluster nodes be active at this time.  If this node hears from
other cluster nodes, this node will abort. Other nodes may only
join after this command has completed successfully.  Data
corruption may occur if more than one cluster is active.

Do you want to continue [y,n,?] y
```

If you receive a `reconfig.4013` error message, then either there is already a node in a cluster, or another node is still in the process of going down. Run the `get_node_status(1M)` command on the node that might be up to determine that node’s status.

2. **Add all other nodes to the cluster.**

   Run the following command on each node in the cluster, sequentially.

```
# scadmin startnode
```

If you receive the following `reconfig.4015` error message, there might be no existing cluster. Restart the cluster by using the `scadmin startcluster` *localnode* command.

```
SUNWcluster.clustd.reconf.4015
‘‘Aborting--no existing or intact cluster to join.’’
```

Alternately, there may be a partition or node failure. (For example, a third node is attempting to join a two-node cluster when one of the two nodes fails.) If this happens, wait until the failures have completed. Fix the problems, if any, and then attempt to rejoin the cluster.

If any required software packages are missing, the command fails and the console displays a message similar to the following:

```
Assuming a default cluster name of haclust
Error: Required SC package ‘SUNWccm’ not installed!
Aborting cluster startup.
```

For information about installing the Sun Cluster software packages, refer to the *Sun Cluster 2.2 Software Installation Guide.*

## Stopping the Cluster and Cluster Nodes

Putting a node in any mode other than multiuser, or halting or rebooting the node, requires stopping the Sun Cluster membership monitor. Then your site’s preferred method can be used for further node maintenance.

Stopping the cluster requires stopping the membership monitor on all cluster nodes by running the `scadmin stopnode` command on all nodes simultaneously.

- You can stop the membership monitor only when no logical hosts are owned by the local Sun Cluster node.
- To stop the membership monitor on one node, switch over the logical host(s) to another node using the `haswitch(1M)` command and stop the membership monitor by typing the following command:

```
phys-hahost1# haswitch destination_host logicalhost
phys-hahost1# scadmin stopnode
```

If a logical host is owned by the node when the `scadmin stopnode` command is run, ownership will be transferred to another node that can master the logical host before the membership monitor is stopped. If the other possible master of the logical host is down, the `scadmin stopnode` command will shut down the data services in addition to stopping the membership monitor.

After the `scadmin stopnode` command runs, Sun Cluster will remain stopped, even across system reboots, until the `scadmin startnode` command is run.

The `scadmin stopnode` command removes the node from the cluster. In the absence of other simultaneous failures, you may shut down as many nodes as you choose without losing quorum among the remaining nodes. (If quorum is lost, the entire cluster shuts down.)

If you shut down a node for disk maintenance, you also must prepare the boot disk or data disk using the procedures described in Chapter 10 for boot disks, or those described in your volume manager documentation for data disks.

You might have to shut down one or more Sun Cluster nodes to perform hardware maintenance procedures such as adding or removing SBus cards. The following sections describe the procedure for shutting down a single node or the entire cluster.

---

**Note -** In a cluster with more than two nodes and with direct-attached storage, a problem can occur if the last node in the cluster panics or exits the cluster unusually (without performing the `stopnode` transition). In such a case, all nodes have been removed from the cluster and the cluster no longer exists, but because the last node left the cluster in an unusual manner, it still holds the nodelock. A subsequent invocation of the `scadmin startcluster` command will fail to acquire the nodelock. To work around this problem, manually clear the nodelock before restarting the cluster, using the procedure “How to Clear a Nodelock Freeze After a Cluster Panic” on page 86.

---

## How to Stop Sun Cluster on a Cluster Node

1. **If it is not necessary to have the data remain available, place the logical hosts (disk groups) into maintenance mode.**

   ```
   phys-hahost2# haswitch -m logicalhost
   ```

   Refer to the `haswitch(1M)` man page for details.

   ---

   **Note -** It is possible to halt a Sun Cluster node by using the `halt(1M)` command, allowing a failover to restore the logical host services on the backup node. However, the `halt(1M)` operation might cause the node to panic. The `haswitch(1M)` command offers a more reliable method of switching ownership of the logical hosts.

   ---

2. **Stop Sun Cluster on one node without stopping services running on the other nodes in the cluster.**

   ```
   phys-hahost1# scadmin stopnode
   ```

   ---

   **Note -** When you stop a node, the following error message might be displayed: `in.rdiscd[517]: setsockopt (IP_DROP_MEMBERSHIP): Cannot assign requested address` The error is caused by a timing issue between the `in.rdiscd` daemon and the IP module. It is harmless and can be ignored safely.

   ---

3. **Halt the node.**

   ```
   phys-hahost1# halt
   ```

   The node is now ready for maintenance work.

## How to Stop Sun Cluster on All Nodes

You might want to shut down all nodes in a Sun Cluster configuration if a hazardous environmental condition exists, such as a cooling failure or a severe lightning storm.

1. **Stop the membership monitor on all nodes by using the** `scadmin(1M)` **command.**

Run this command on the console of each node in the cluster. Allow each node to exit the cluster and the remaining nodes to reconfigure completely before you run the command on the next node

```
phys-hahost1# scadmin stopnode
...
```

.

2. **Halt all nodes using `halt(1M)`.**

```
phys-hahost1# halt
...
```

## How to Halt a Sun Cluster Node

1. **Shut down any Sun Cluster node by using the `halt(1M)` command or the `uadmin(1M)` command.**

   If the membership monitor is running when a node is shut down, the node will most likely take a “Failfast timeout” and display the following message:

```
panic[cpu9]/thread=0x50f939e0: Failfast timeout - unit
```

   You can avoid this by stopping the membership monitor before shutting down the node. Refer to the procedure, “How to Stop Sun Cluster on All Nodes” on page 85, for additional information.

## How to Clear a Nodelock Freeze After a Cluster Panic

In a cluster with more than two nodes and with direct-attached storage, a problem can occur if the last node in the cluster panics or exits the cluster unusually (without performing the `stopnode` transition). In such a case, all nodes have been removed from the cluster and the cluster no longer exists. However, because the last node left the cluster in an unusual manner, it still holds the nodelock. A subsequent invocation of the `scadmin startcluster` command will fail to acquire the nodelock.

To work around this problem, manually clear the nodelock before restarting the cluster. Use the following procedure to manually clear the nodelock and restart the cluster, after the cluster has aborted completely.

1. **As root, display the cluster configuration.**

```
# scconf clustername -p
```

Look for this line in the output:

```
clustername Locking TC/SSP, port  : A.B.C.D, E
```

- If *E* is a positive number, the nodelock is on Terminal Concentrator *A.B.C.D* and Port *E*. Proceed to Step 2 on page 87.
- If *E* is -1, the lock is on an SSP. Proceed to Step 3 on page 88.

2. **For a nodelock on a Terminal Concentrator (TC), perform the following steps.**

   a. **Start a telnet connection to Terminal Concentrator *tc-name*.**

```
$ telnet tc-name
Trying 192.9.75.51...
Connected to tc-name.
Escape character is '^]'.
```

Enter Return to continue.

   b. **Specify `cli` (command line interface).**

```
Enter Annex port name or number: cli
```

   c. **Log in as root.**

   d. **Run the `admin` command.**

```
annex# admin
```

   e. **Reset Port *E*.**

```
admin : reset E
```

**f. Close the telnet connection.**

```
annex# hangup
```

**g. Proceed to Step 4 on page 89.**

**3. For a nodelock on a System Service Processor (SSP), perform the following steps.**

**a. Connect to the SSP.**

```
$ telnet ssp-name
```

**b. Log in as user `ssp`.**

**c. Display information about the *clustername*`.lock` file by using the following command. (This file is a symbolic link to `/proc/`*csh.pid*.)**

```
$ ls -l /var/tmp/clustername.lock
```

**d. Search for the process *csh.pid*.**

```
$ ps -ef | grep csh.pid
```

**e. If the *csh.pid* process exists in the `ps -ef` output, kill the process by using the following command.**

```
$ kill -9 csh.pid
```

**f. Delete the *clustername*`.lock` file.**

```
$ rm -f /var/tmp/clustername.lock
```

**g. Log out of the SSP.**

4. **Restart the cluster.**

```
$ scadmin startcluster
```

## Stopping the Membership Monitor While Running RDBMS Instances

Database server instances can run on a node only after you have invoked the `startnode` option and the node has successfully joined the cluster. All database instances should be shut down before the `stopnode` option is invoked.

---

**Note -** If you are running Oracle7 Parallel Server, Oracle8 Parallel Server, or Informix XPS, refer to your product documentation for shutdown procedures.

---

If the `stopnode` command is executed while the Oracle7 or Oracle8 instance is still running on the node, `stopnode` will hang and the following message is displayed on the console:

```
ID[vxclust]: stop: waiting for applications to end
```

The Oracle7 or Oracle8 instance must be shut down for the `stopnode` command to terminate successfully.

If the `stopnode` command is executed while the Informix-Online XPS instance is still running on the node, the database hangs and becomes unusable.

# Switching Over Logical Hosts

The `haswitch(1M)` command is used to switch over the specified logical hosts (and associated disk groups, data services, and logical IP addresses) to the node specified by the destination host. For example, the following command switches over logical hosts `hahost1` and `hahost2` to both be mastered by `phys-hahost1`.

```
# haswitch phys-hahost1 hahost1 hahost2
```

If the logical host has more than one data service configured on it, you cannot selectively switch over just one data service, or a subset of the data services. Your only option is to switch over *all* the data services on the logical host.

**Caution -** If a failover or switchover occurs while a logical host's file system is busy, the logical host fails over only partially; some of the disk group remains on the original target physical host. Do not attempt a switchover if a logical host's file system is busy. Also, do not access any host's file system locally, because file locking does not work correctly when both NFS locks and local locks are present.

**Note -** Both the destination host and the current master of the logical host *must* be in the cluster membership. Otherwise, the command fails.

# Disabling Automatic Switchover

In clusters providing HA data services, automatic switchover can be set up for the situation where a node fails, the logical hosts it mastered are switched over to another node, and later the failed node returns to the cluster. Logical hosts will automatically be remastered by their default master, unless you configure them to remain mastered by the host to which they were switched.

If you do not want a logical host to be automatically switched back to its original master, use the `-m` option of the `scconf(1M)` command. Refer to the `scconf(1M)` man page for details.

**Note -** To disable automatic switchover for a logical host, you need only run the `scconf(1M)` command on a single node that is an active member of the cluster.

```
# scconf clustername -L logicalhost -n node1,node2 -g dg1 -i qe0,qe0,logaddr1 -m
```

# Putting Logical Hosts in Maintenance Mode

Maintenance mode is useful for some administration tasks on file systems and disk groups. To put the disk groups of a logical host into maintenance mode, use the -m option to the haswitch(1M) command.

---

**Note -** Unlike other types of ownership of a logical host, maintenance mode persists across node reboots.

---

For example, this command puts logical host hahost1 in maintenance mode.

```
phys-hahost2# haswitch -m hahost1
```

This command stops the data services associated with hahost1 on the Sun Cluster node that currently owns the disk group, and also halts the fault monitoring programs associated with hahost1 on all Sun Cluster nodes. The command also executes a umount(1M) of any Sun Cluster file systems on the logical host. The associated disk group ownership is released.

This command runs on any host, regardless of current ownership of the logical host and disk group.

You can remove a logical host from maintenance mode by performing a switchover specifying the physical host that is to own the disk group. For example, you could use the following command to remove hahost1 from maintenance mode:

```
phys-hahost1# haswitch phys-hahost1 hahost1
```

# Recovering From Cluster Partitions

Multiple failures (including network partitions) might result in subsets of cluster members attempting to remain in the cluster. Usually, these subsets have lost partial or total communication with each other. In such cases, the software attempts to ensure that there is only one resultant valid cluster. To achieve this, the software might cause some or all nodes to abort. The following discussion explains the criterion used to make these decisions.

The quorum criterion is defined as a subset with at least half the members of the *original* set of cluster nodes (not only the configured nodes). If a subset does not meet the quorum criterion, the nodes in the subset abort themselves and a `reconfig.4014` error message is displayed. Failure to meet the quorum criterion could be due to a network partition or to a simultaneous failure of more than half of the nodes.

---

**Note -** Valid clusters only contain nodes that can communicate with each other over private networks.

---

Consider a four-node cluster that partitions itself into two subsets: one subset consists of one node, while the other subset consists of three nodes. Each subset attempts to meet the quorum criterion. The first subset has only one node (out of the original four) and does not meet the quorum criterion. Hence, the node in the first subset shuts down. The second subset has three nodes (out of the original four), meets the quorum criterion, and therefore stays up.

Alternatively, consider a two-node cluster with a quorum device. If there is a partition in such a configuration, then one node and the quorum device meet the quorum criterion and the cluster stays up.

## Split-Brain Partitions (VxVM Only)

A split-brain partition occurs if a subset has exactly half the cluster members. (The split-brain partition does not include the scenario of a two-node cluster with a quorum device.) During initial installation of Sun Cluster, you were prompted to choose your preferred type of recovery from a split-brain scenario. Your choices were `ask` and `select`. If you chose `ask`, then if a split-brain partition occurs, the system asks you for a decision about which nodes should stay up. If you chose `select`, the system automatically selects for you which cluster members should stay up.

If you chose an automatic selection policy to deal with split-brain situations, your options were `Lowest Nodeid` or `Highest Nodeid`. If you chose `Lowest Nodeid`, then the subset containing the node with the lowest ID value becomes the new cluster. If you chose `Highest Nodeid`, then the subset containing the node with the highest ID value becomes the new cluster. For more details, see Chapter 3 in the *Sun Cluster 2.2 Software Installation Guide.*

In either case, you must manually abort the nodes in all other subsets.

If you did not choose an automatic selection policy or if the system prompts you for input at the time of the partition, then the system displays the following error message.

```
SUNWcluster.clustd.reconf.3010
‘‘*** ISSUE ABORTPARTITION OR CONTINUEPARTITION *** Proposed cluster: xxx  Unreachable nodes: yyy’’
```

Additionally, a message similar to the following is displayed on the console every ten seconds:

```
*** ISSUE ABORTPARTITION OR CONTINUEPARTITION ***
If the unreachable nodes have formed a cluster, issue ABORTPARTITION.
(scadmin abortpartition <localnode> <clustername>)
You may allow the proposed cluster to form by issuing CONTINUEPARTITION.
(scadmin continuepartition <localnode> <clustername>)
Proposed cluster partition:  0  Unreachable nodes: 1
```

If you did not enable automatic selection, use the following procedure to choose the new cluster.

**Note -** To restart the cluster after a split-brain failure, you must wait for the stopped node to come up entirely (it might undergo automatic reconfiguration or reboot) before you bring it back into the cluster using the `scadmin startnode` command.

## ▼ How to Choose a New Cluster

1. **Determine which subset should form the new cluster. Run the following command on one node in the subset that should abort.**

```
# scadmin abortpartition
```

When the `abortpartition` command is issued on one node, the Cluster Membership Monitor (CMM) propagates that command to all the nodes in that partition. Therefore, if *all* nodes in that partition receive the command, they all abort. However, if some of the nodes in the partition cannot be contacted by the CMM, then they have to be manually aborted. Run the `scadmin abortpartition` command on any remaining nodes that do not abort.

2. **Run the following command on one node in the subset that should stay up.**

```
# scadmin continuepartition
```

**Note -** A further reconfiguration occurs if there has been another failure within the new cluster. At all times, only one cluster is active.

## Maintaining the `/var` File System

Because Solaris and Sun Cluster software error messages are written to the `/var/adm/messages` file, the `/var` file system can become full. If the `/var` file system becomes full while the node is running, the node will continue to run, but you probably will not be able to log into the node with the full `/var` file system. If the node goes down, Sun Cluster will not start and a login will not be possible. If this happens, you must reboot in single-user mode (`boot -s`).

If the node reports a full `/var` file system and continues to run Sun Cluster services, follow the steps outlined in the following procedure.

### ▼ How to Repair a Full `/var` File System

In this example, `phys-hahost1` has a full `/var` file system.

1. **Perform a switchover.**

   Move all logical hosts off the node experiencing the problem.

   ```
   phys-hahost2# haswitch phys-hahost2 hahost1 hahost2
   ```

2. **Remove the node from the cluster membership.**

   If you have an active login to `phys-hahost1`, enter the following:

   ```
   phys-hahost1# scadmin stopnode
   ```

   If you do not have an active login to `phys-hahost1`, halt the node.

3. **Reboot the node in single-user mode.**

   ```
   (0) ok boot -s
   INIT: SINGLE USER MODE
   ```

(Continuation)

```

Type Ctrl-d to proceed with normal startup,
(or give root password for system maintenance): root_password
Entering System Maintenance Mode

Sun Microsystems Inc. SunOS 5.6 Generic August 1997
```

4. **Perform the steps you would normally take to clear the full file system.**

5. **After the file system is cleared, enter multiuser mode.**

```
# exit
```

6. **Use the** `scadmin startnode` **command to cause the node to rejoin the configuration.**

```
# scadmin startnode
```

## Administering the Time in Sun Cluster Configurations

We recommend that you use Network Time Protocol (NTP) to maintain time synchronization between cluster nodes if NTP comes with your Solaris operating environment.

**Caution -** An administrator cannot adjust the time of the nodes in a Sun Cluster configuration. Never attempt to perform a time change using the `date(1)`, `rdate(1M)`, or `xntpdate(1M)` commands.

In the Sun Cluster environment, the cluster nodes can run as NTP clients. You must have an NTP server set up and configured outside the cluster to use NTP; the cluster nodes cannot be configured to be NTP servers. Refer to the `xntpd(1M)` man page for information about NTP clients and servers.

If you are running cluster nodes as NTP clients, make sure that there are no crontab(1) entries that call ntpdate(1M). It is safer to run xntpd(1M) on the clients because that keeps the clocks in sync without making large jumps forward or backward.

# Replacing a Failed Node

Complete the following steps when one node has a hardware failure and needs to be replaced with a new node.

**Note -** This procedure assumes the root disk of the failed node is still operational and can be used. If your failed root disk is not mirrored, contact your local Sun Enterprise Service representative or your local authorized service provider for assistance.

## ▼ How to Replace a Failed Node

If the failed node is not operational, start at Step 5 on page 97.

1. **If you have a parallel database configuration, stop the database.**

   **Note -** Refer to the appropriate documentation for your data services. All HA applications are automatically shut down with the scadmin stopnode command.

2. **Use the Cluster Console to open a terminal window.**

3. **As root, enter the following command in the terminal window.**

   This command removes the node from the cluster, stops the Sun Cluster software, and disables the volume manager on that node.

```
# scadmin stopnode
```

4. **Halt the operating system on the node.**

   Refer to your Solaris system administration documentation if necessary.

5. **Power off the node.**

   Refer to your hardware service manual for more information.

---

**Caution -** Do not disconnect any cables from the failed node at this time.

---

6. **Remove the boot disk from the failed node.**

   Refer to your hardware service manual for more information.

7. **Place the boot disk in the identical slot in the new node.**

   The root disk should be accessible at the same address as before. Refer to your hardware service manual for more information.

   ---

   **Note -** Be sure that the new node has the same IP address as the failed system. You may need to modify the boot servers or `arp` servers to remap the IP address to the new Ethernet address. For more information, refer to the *NIS+ and DNS Setup and Configuration Guide*.

   ---

8. **Power on the new node.**

   Refer to your hardware service manual for more information.

9. **If the node automatically boots, shut down the operating system and take the system to the OpenBoot PROM monitor.**

   For more information, refer to the `shutdown(1M)` man page.

10. **Make sure that every** `scsi-initiator-id` **is set correctly.**

    See Chapter 4 in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide* for the detailed procedure to set the `scsi-initiator-id`.

11. **Power off the new node.**

    Refer to your hardware service manual for more information.

12. **On the surviving node that shares the multihost disks with the failed node, detach all of the disks in one disk expansion unit attached to the failed node.**

    Refer to your hardware service manual for more information.

13. **Power off the disk expansion unit.**

    Refer to your hardware service manual for more information.

**Note -** As you replace the failed node, messages similar to the following might appear on the system console. Disregard these messages, because they might not indicate a problem.

```
Nov  3 17:44:00 updb10a unix: WARNING: /sbus@1f,0/SUNW,fas@0,8800000/sd@2,0 (sd17):
Nov  3 17:44:00 updb10a unix: SCSI transport failed: reason "incomplete": retrying \ command
Nov  3 17:44:03 updb10a unix: WARNING: /sbus@1f,0/SUNW,fas@0,8800000/sd@2,0 (sd17):
Nov  3 17:44:03 updb10a unix:   disk not responding to selection
```

14. **Detach the SCSI cable from the failed node and attach it to the corresponding slot on the new node.**

    Refer to your hardware service manual for more information.

15. **Power on the disk expansion unit.**

    Refer to your hardware service manual for more information.

16. **Reattach all of the disks you detached in Step 12 on page 97.**

    Refer to your hardware service manual for more information.

17. **Wait for volume recovery to complete on all the volumes in the disk expansion unit before detaching the corresponding mirror disk expansion unit.**

    Use your volume manager software to determine when volume recovery has occurred.

18. **Repeat Step 12 on page 97 through Step 17 on page 98 for all of the remaining disk expansion units.**

19. **Power on the replaced (new) node.**

    Refer to your hardware service manual for more information.

20. **Reboot the node and wait for the system to come up.**

```
<#0> boot
```

21. **Determine the Ethernet address on the replaced (new) node.**

```
# /usr/sbin/arp nodename
```

22. **Determine the node ID of the replaced node.**

    By the process of elimination, you can determine which node is not in the cluster. The node IDs should be numbered consecutively starting with node 0.

```
# get_node_status
sc: included in running cluster
node id: 0
membership: 0
interconnect0: unknown
interconnect1: unknown
vm_type: vxvm
vm_on_node: master
vm: up
db: down
```

23. **Inform the cluster system of the new Ethernet address (of the replaced node) by entering the following command on all the cluster nodes.**

```
# scconf clustername -N node-id ethernet-address-of-host
```

    Continuing with the example in Step 22 on page 99, the node ID is `1`:

```
# scconf clustername -N 1 ethernet-address-of-host
```

24. **Start up the replaced node.**

```
# scadmin startnode
```

25. **In parallel database configuration, restart the database.**

---

**Note -** Refer to the appropriate documentation for your data services. All HA applications are automatically started with the `scadmin startcluster` and `scadmin startnode` commands.

---

## Replacing a Failed Terminal Concentrator

The Terminal Concentrator need not be operational for the cluster to stay up. If the Terminal Concentrator fails, the cluster itself does not fail.

You can replace a failed Terminal Concentrator without affecting the cluster. If the new Terminal Concentrator has retained the same name, IP address, and password as the original, then no cluster commands are required. Simply plug in the new Terminal Concentrator and it will work as expected.

If the replacement Terminal Concentrator has a new name, IP address, or password, use the `scconf(1M)` command as described in "Changing TC/SSP Information" on page 73, to change this information in the cluster database. This can be done with the cluster running without affecting cluster operations.

## Administering the Cluster Configuration Database

The `ccdadm(1M)` command is used to perform administrative procedures on the Cluster Configuration Database (CCD). Refer to the `ccdadm(1M)` man page for additional information.

---

**Note -** As `root`, you can run the `ccdadm(1M)` command from any active node. This command updates all the nodes in your cluster.

---

It is good practice to checkpoint the CCD using the `-c` option (checkpoint) to `ccdadm(1M)` each time cluster configuration is updated. The CCD is extensively used by the Sun Cluster framework to store configuration data related to logical hosts and HA data services. The CCD is also used to store the network adapter configuration data used by PNM. We strongly recommended that after any changes to the HA or PNM configuration of the cluster, you capture the current valid snapshot of the CCD by using the `-c` option as an insurance against problems that can occur under fault scenarios in the future. This requirement is no different from requiring database administrators or system administrators to frequently backup their data to avoid catastrophes in the future due to unforeseen circumstances.

## ▼ How to Verify CCD Global Consistency

1. **Run the `-v` option whenever there may be a problem with the Dynamic CCD.**

   This option compares the consistency record of each CCD copy on all the cluster nodes, enabling you to verify that the database is consistent across all the nodes. CCD queries are disabled while the verification is in progress.

```
# ccdadm clustername -v
```

## ▼ How to Back Up the CCD

1. **Run the `-c` option once a week or whenever you back up the CCD.**

   This option makes a backup copy of the Dynamic CCD. The backup copy subsequently can be used to restore the Dynamic CCD by using the `-r` option. See "How to Restore the CCD" on page 101 for more information.

   **Note -** When backing up the CCD, put all logical hosts in maintenance mode before running the `ccdadm -c` command. The logical hosts must be in maintenance mode when restoring the CCD database. Therefore, having a backup file similar to the restore state will prevent unnecessary errors or problems.

```
# ccdadm clustername -c checkpoint-filename
```

In this command, *checkpoint-filename* is the name of your backup copy.

## ▼ How to Restore the CCD

Run `ccdadm(1M)` with the `-r` option whenever the CCD has been corrupted. This option discards the current copy of the Dynamic CCD and restores it with the contents of the restore file you supply. Use this command to initialize or restore the Dynamic CCD after the `ccdd(1M)` reconfiguration algorithm failed to elect a valid CCD copy upon cluster restart. The CCD is then marked valid.

1. **If necessary, disable the quorum.**

   See "How to Enable or Disable the CCD Quorum" on page 102 for more information.

```
# ccdadm clustername -q off
```

2. **Put the logical hosts in maintenance mode.**

```
# haswitch -m logicalhosts
```

3. **Restore the CCD.**

   In this command, *restore-filename* is the name of the file you are restoring.

```
# ccdadm clustername -r restore-filename
```

4. **If necessary, turn the CCD quorum back on.**

```
# ccdadm clustername -q on
```

5. **Bring the logical hosts back online.**

   For example:

```
# haswitch phys-host1 logicalhost1
# haswitch phys-host2 logicalhost2
```

## ▼ How to Enable or Disable the CCD Quorum

1. **Typically, the cluster software requires a quorum before updating the CCD. The `-q` option enables you to disable this restriction and to update the CCD with any number of nodes.**

   Run this option to enable or disable a quorum when updating or restoring the Dynamic CCD. The `quorum_flag` is a toggle: `on` (to enable) or `off` (to disable) a quorum. By default, the quorum is enabled.

   For example, if you have three physical nodes, you need at least two nodes to perform updates. Because of a hardware failure, you can bring up only one node. The cluster software does not enable you to update the CCD. If, however, you run the `ccdadm -q` command, you can toggle off the software control, and update the CCD.

```
# ccdadm clustername -q on|off
```

## How to Purify the CCD

1. **The `-p` option enables you to purify (verify the contents and check the syntax of) the CCD database file. Run this option whenever there is a syntax error in the CCD database file.**

```
# ccdadm -p CCD-filename
```

The `-p` option reports any format errors in the candidate file and writes a corrected version of the file into the file *filename*.`pure`. You can then restore this “pure” file as the new CCD database. See “How to Restore the CCD” on page 101 for more information.

## Troubleshooting the CCD

The system logs errors in the CCD to the `/var/opt/SUNWcluster/ccd/ccd.log` file. Critical error messages are also passed to the Cluster Console. Additionally, in the rare case of a crash, the software creates a core file under `/var/opt/SUNWcluster/ccd`.

The following is an example of the `ccd.log` file.

```
lpc204# cat ccd.log
Apr 16 14:54:05 lpc204 ID[SUNWcluster.ccd.ccdd.1005]: (info) starting ‘START’ transition with time-
out 10000
Apr 16 14:54:05 lpc204 ID[SUNWcluster.ccd.ccdd.1005]: (info) completed ‘START’ transition with status 0
Apr 16 14:54:06 lpc204 ID[SUNWcluster.ccd.ccdd.1005]: (info) starting ‘STEP1’ transition with time-
out 20000
Apr 16 14:54:06 lpc204 ID[SUNWcluster.ccd.ccdd.1000]: (info) Nodeid = 0 Up = 0 Gennum = 0 Date = Feb 14 10h30m00 1997 Re
Apr 16 14:54:06 lpc204 ID[SUNWcluster.ccd.ccdd.1002]: (info) start reconfiguration elected CCD from Nodeid = 0
Apr 16 14:54:06 lpc204 ID[SUNWcluster.ccd.ccdd.1004]: (info) the init CCD database is consistent
Apr 16 14:54:06 lpc204 ID[SUNWcluster.ccd.ccdd.1001]: (info) Node is up as a one-
node cluster after scadmin startcluster; skipping ccd quorum test
Apr 16 14:54:06 lpc204 ID[SUNWcluster.ccd.ccdd.1005]: (info) completed ‘STEP1’ transition with status 0
```

The following table lists the most common error messages with suggestions for resolving the problem. Refer to the *Sun Cluster 2.2 Error Messages Manual* for the complete list of error messages.

**TABLE 4–1** Common Error Messages for the Cluster Configuration Database

| Number Range | Explanation | Action |
|---|---|---|
| 4200 | Cannot open file | Restore the CCD by running the `ccdadm -r` command. |
| 4302 | File not found | Restore the CCD by running the `ccdadm -r` command. |
| 4307 | Inconsistent Init CCD | Remove, then reinstall, the Sun Cluster software. |
| 4402 | Error registering RPC server | Check your public network (networking problem). |
| 4403 | RPC client create failed | Check your public network (networking problem). |
| 5000 | System execution error | The synchronization script has an error. Check the permissions on the script. |
| 5300 | Invalid CCD, needs to be restored | Restore the CCD by running the `ccdadm -r` command. |
| 5304 | Error running freeze command | There are incorrect arguments in the executed synchronization script. Check that the format of the script is correct. |
| 5306 | Cluster pointer is null | This message indicates that the cluster does not exist (`ccdadm` *cluster*). Check that you typed the cluster name correctly. |

# Reserving Shared Disks (VxVM)

The list of disks maintained by the volume manager is used as the set of devices for failure fencing. If there are no disk groups present in a system, there are no devices for failure fencing (there is effectively no data to be protected). However, when new shared disk groups are imported while one or more nodes are not in the cluster, the cluster must be informed that an extra set of devices need failure fencing.

## How to Reserve Shared Devices (VxVM)

1. **When new shared disk groups are imported while one or more nodes are not in the cluster, the cluster must be informed that an extra set of devices need failure fencing. This is accomplished by running the** `scadmin resdisk` **command from a node that can access the new disk group(s).**

```
# scadmin resdisks
```

This command reserves all the devices connected to a node if no other node (that has connectivity to the same set of devices) is in the cluster membership. That is, reservations are affected only if one and only one node, out of all possible nodes that have direct physical connectivity to the devices, is in the cluster membership. If this condition is false, the `scadmin resdisks` command has no effect. The command also fails if a cluster reconfiguration is in progress. Reservations on shared devices are automatically released when this one node is shut down, or when other nodes, with direct connectivity to the shared devices, join the cluster membership.

---

**Note -** It is unnecessary to run the `scadmin resdisks` command if shared disk groups are imported while all nodes are in the cluster. Reservations and failure fencing are not relevant if full cluster membership is present.

---

However, if a shared disk group is deported, the reservations on the shared devices in the deported disk group are not released. These reservations are not released until either the node that does the reservations is shut down, or the other node, with which it shares devices, joins the cluster.

To enable the set of disks belonging to the deported disk group to be used immediately, enter the following two commands in succession on all cluster nodes, after deporting the shared disk group:

```
# scadmin reldisks
# scadmin resdisks
```

The first command releases reservations on all shared devices. The second command effectively redoes the reservations based on the currently imported set of disk groups, and automatically excludes the set of disks associated with deported disk groups.

CHAPTER 5

# Recovering From Power Loss

This chapter describes different power loss scenarios and the steps you take to return the system to normal operation. The topics in this chapter are listed below.

Maintaining Sun Cluster configurations includes handling failures such as power loss. A power loss can shut down an entire Sun Cluster configuration, or one or more components within a configuration. Sun Cluster nodes behave differently depending on which components lose power. The following sections describe typical scenarios and expected behavior.

## Recovering From Total Power Loss

In Sun Cluster configurations with a single power source, a power failure takes down all Sun Cluster nodes along with their multihost disk expansion units. When all nodes lose power, the entire configuration fails.

In a total-failure scenario, there are two ways in which the cluster hardware might come back up.

- A Sun Cluster node reboots before the Terminal Concentrator. Any errors reported when the node is rebooting are stored in the `/var/adm/messages` file or the error log pointed to in the `/etc/syslog.conf` file.
- A Sun Cluster node reboots before the multihost disk expansion unit. The associated disks will not be accessible. One or more nodes must be rebooted after the multihost disk expansion unit comes up. Once the nodes are up, run the

hastat(1M) command and use your volume management software to search for errors that occurred due to the power outage.

# Recovering From Partial Power Loss

If the Sun Cluster nodes and the multihost disk expansion units have separate power sources, a failure can take down one or more components. Several scenarios can occur. The most likely cases are:

- The power to one Sun Cluster node fails, taking down only the node.
- The power to one multihost disk expansion unit fails, taking down only the expansion unit.
- The power to one Sun Cluster node fails, taking down at least one multihost disk expansion unit.
- The power to one Sun Cluster node fails, taking down the node, at least one multihost disk expansion unit, and the Terminal Concentrator.

## Failure of One Node

If separate power sources are used on the nodes and the multihost disk expansion units, and you lose power to only one of the nodes, the other node detects the failure and initiates a takeover.

When power is restored to the node that failed, it reboots. You must rejoin the cluster by using the `scadmin startnode` command. Then perform a manual switchover by using the `haswitch(1M)` command to restore the default logical host ownership.

## Failure of a Multihost Disk Expansion Unit

If you lose power to one of the multihost disk expansion units, your volume management software detects errors on the affected disks and takes action to put them into an error state. Disk mirroring masks this failure from the Sun Cluster fault monitoring. No switchover or takeover occurs.

When power is returned to the multihost disk expansion unit, perform the procedure documented in Chapter 11, or Chapter 12.

## Failure of One Server and One Multihost Disk Expansion Unit

If power is lost to one of the Sun Cluster nodes and one multihost disk expansion unit, a secondary node immediately initiates a takeover.

After the power is restored, you must reboot the node, rejoin the node to the configuration by using the `scadmin startnode` command, and then begin monitoring activity. If manual switchover is configured, use the `haswitch(1M)` command to manually return ownership of the diskset to the node that had lost power. Refer to “Switching Over Logical Hosts” on page 89, for more information.

After the diskset ownership has been returned to the default master, any multihost disks that reported errors must be returned to service. Use the instructions provided in the chapters on your disk expansion unit to return the multihost disks to service.

**Note -** The node might reboot before the multihost disk expansion unit. Therefore, the associated disks will not be accessible. Reboot the node after the multihost disk expansion unit comes up.

# Powering On the System

Applying power to system cabinets, nodes, and boot disks varies, depending on the type of cabinet being used, and the manner in which the nodes receive AC power.

For disk arrays that do not receive AC power from an independent power source, AC power is applied when the system cabinet is powered on.

For specific power-on procedures for Sun StorEdge MultiPacks, refer to the *Sun StorEdge MultiPack Service Manual*.

The Terminal Concentrator that receives AC power from the system cabinet is turned on when power is applied to the cabinet. Otherwise, the Terminal Concentrator must be powered on independently.

CHAPTER 6

# Administering Network Interfaces

This chapter provides a description of the Public Network Management (PNM) feature of Sun Cluster, and instructions for adding or replacing network interface components. The topics in this chapter are listed below.

- "Public Network Management Overview" on page 111
- "Setting Up and Administering Public Network Management" on page 114
- "Troubleshooting PNM Errors" on page 121
- "Adding and Removing Network Interfaces" on page 122
- "Administering the Switch Management Agent" on page 127

## Public Network Management Overview

The PNM feature of Sun Cluster uses fault monitoring and failover to prevent loss of node availability due to single network adapter or cable failure. PNM fault monitoring runs in local-node or cluster-wide mode to check the status of nodes, network adapters, cables, and network traffic. PNM failover uses sets of network adapters called *backup groups* to provide redundant connections between a cluster node and the public network. The fault monitoring and failover capabilities work together to ensure availability of services.

If your configuration includes HA data services, you must enable PNM; HA data services are dependent on PNM fault monitoring. When an HA data service experiences availability problems, it queries PNM through the cluster framework to see whether the problem is related to the public network connections. If it is, the data services wait until PNM has resolved the problem. If the problem is not with the public network, the data services invoke their own failover mechanism.

The PNM package, SUNWpnm, is installed during initial Sun Cluster software installation. The commands associated with PNM include:

- pnmset(1M)– Used to set up PNM before or after you configure the cluster, and to check the correctness of an existing PNM configuration.
- pnmstat(1M)– Used to check the status of the network and adapters.
- pnmconf(1M)– Used to display the PNM network interface configuration and status.
- pnmrtop(1M)– Displays the backup group name or pseudo adapter name (such as nafo1) associated with the real adapter name (such as hme2) that you supply to the command.
- pnmptor(1M)– Displays the real adapter name (such as hme2) associated with the pseudo adapter name or backup group name (such as nafo1) that you supply to the command.
- pnmd(1M)– The PNM daemon.

See the associated man pages for details.

## PNM Fault Monitoring and Failover

PNM monitors the state of the public network and the network adapters associated with each node in the cluster, and reports dubious or errored states. When PNM detects lack of response from a *primary* adapter (the adapter currently carrying network traffic to and from the node) it fails over the network service to another working adapter in the adapter backup group for that node. PNM then performs some checks to determine whether the fault is with the adapter or the network.

If the adapter is faulty, PNM sends error messages to syslog(3), which are in turn detected by the Cluster Manager and displayed to the user through a GUI. After a failed adapter is fixed, it is automatically tested and reinstated in the backup group at the next cluster reconfiguration. If the entire adapter backup group is down, then the Sun Cluster framework invokes a failover of the node to retain availability. If an error occurs outside of PNM's control, such as the failure of a whole subnet, then a normal failover and cluster reconfiguration will occur.

PNM monitoring runs in two modes, cluster-aware and cluster-unaware. PNM runs in cluster-aware mode when the cluster is operational. It uses the Cluster Configuration Database (CCD) to monitor status of the network (for more information on the CCD, see the overview chapter in the *Sun Cluster 2.2 Software Installation Guide*). PNM uses the CCD to distinguish between public network failure and local adapter failure. See Appendix B for more information on logical host failover initiated by public network failure.

PNM runs in cluster-unaware mode when the cluster is not operational. In this mode, PNM is unable to use the CCD and therefore cannot distinguish between adapter and network failure. In cluster-unaware mode, PNM simply detects a problem with the local network connection.

You can check the status of the public network and adapters with the PNM monitoring command, pnmstat(1M). See the man page for details.

## Backup Groups

Backup groups are sets of network adapters that provide redundant connections between a single cluster node and the public network. You configure backup groups during initial installation by using the scinstall(1M) command, or after initial installation by using the pnmset(1M) command. PNM allows you to configure as many redundant adapters as you want on a single host.

To configure backup groups initially, you run pnmset(1M) as root before the cluster is started. The command runs as an interactive script to configure and verify backup groups. It also selects one adapter to be used as the primary, or active, adapter. The pnmset(1M) command names backup groups nafo*n*, where *n* is an integer you assign. The command stores backup group information in the /etc/pnmconfig file.

To change an existing PNM configuration on a cluster node, you must remove the node from the cluster and then run the pnmset(1M) command. PNM monitors and incorporates changes in backup group membership dynamically.

**Note -** The /etc/pnmconfig file is not removed even if the SUNWpnm package is removed, for example, during a software upgrade. That is, the backup group membership information is preserved during software upgrades and you are not required to run the pnmset(1M) utility again, unless you want to modify backup group membership.

## Updates to nsswitch.conf

When configuring PNM with a backup network adapter, the /etc/nsswitch.conf file should have one of the following entries for the netmasks entry.

**TABLE 6–1** Name Service Entry Choices for the `/etc/nsswitch.conf` File

| Name Service Used | `netmasks` Entry |
|---|---|
| None | `netmasks: files` |
| `nis` | `netmasks: files [NOTFOUND=return] nis` |
| `nisplus` | `netmasks: files [NOTFOUND=return] nisplus` |

The above settings will ensure that the `netmasks` setting will not be looked up in an NIS/NIS+ lookup table. This is important if the adapter that has failed is the primary public network and thus would not be available to provide the requested information. If the netmasks entry is not set in the prescribed manner, failover to the backup adapter will not succeed.

**Caution -** The preceding changes have the effect of using the local files (`/etc/netmasks` and `/etc/groups`) for lookup tables. The NIS/NIS+ services will only be used when the local files are unavailable. Therefore, these files must be kept up-to-date with their NIS/NIS+ versions. Failure to update them makes the expected values in these files inaccessible on the cluster nodes.

## Setting Up and Administering Public Network Management

This section provides procedures for setting up PNM and configuring backup groups.

### ▼ How to Set Up PNM

These are the high-level steps to set up PNM:

- Setting up the node hardware to allow multiple network adapters per node, per subnet.
- Installing the Sun Cluster and PNM packages if they have not been installed already.

- Starting the cluster.
- Verifying the default network interfaces.
- Establishing PNM backup groups using the `pnmset(1M)` command.
- Verifying the PNM configuration.

These are the detailed steps to set up PNM.

1. **Set up the node hardware so that you have multiple network adapters on a single node using the same subnet.**

   Refer to your Sun Cluster hardware documentation to set up your network adapters.

2. **If the Sun Cluster node software packages have not been installed already, install them by using the `scinstall(1M)` command.**

   The `scinstall(1M)` command runs interactively to install the package set you select. The PNM package, `SUNWpnm`, is part of the node package set. See the *Sun Cluster 2.2 Software Installation Guide* for the detailed cluster installation procedure.

3. **Register the default network interface on each node, if you did not do so already.**

   You must register one default network interface per node in the interface database associated with each node, and verify that the interface is plumbed and functioning correctly.

   a. **Create an interface database on each node and register the primary public network interfaces.**

   Create a file in the `/etc` directory on each node to use as the interface database. Name the file `hostname`.*interface*, where *interface* is your interface type, such as qfe, hme, etc. Then add one line containing the host name for that node. For example, on node `phys-hahost1` with a default interface `qfe1`, create a file `/etc/hostname.qfe1` containing the following line:

```
phys-hahost1
```

   b. **In the `/etc/hosts` file on each node, associate the primary public network interface name with an IP address.**

   In this example, the primary physical host name is `phys-hahost1`:

```
129.146.75.200 phys-hahost1-qfe1
```

   If your system uses a naming mechanism other than `/etc/hosts`, refer to the appropriate section in the *TCP/IP and Data Communications Administration Guid*e to perform the equivalent function.

4. **Establish PNM backup groups by using the `pnmset(1M)` command.**

   Run the `pnmset(1M)` interactive script to set up backup groups.

---

**Caution -** If you have configured logical hosts and data services already, you must stop the HA data services before changing the backup group membership with `pnmset(1M)`. If you do not stop the data services before running the `pnmset(1M)` command, serious problems and data service failures can result.

---

a. **Run the `pnmset(1M)` command.**

```
phys-hahost1# /opt/SUNWpnm/bin/pnmset
```

b. **Enter the total number of backup groups you want to configure.**

   Normally this number corresponds with the number of public subnets.

```
In the following dialog, you will be prompted to configure public network management.

do you want to continue ... [y/n]: y

How many NAFO backup groups on the host [1]: 2
```

c. **Assign backup group numbers.**

   At the prompt, supply an integer between 0 and the maximum of 255. The `pnmset(1M)` command appends this number to the string `nafo` to form the backup group name.

```
Enter backup group number [0]: 0
```

d. **Assign adapters to backup groups.**

```
Please enter all network adapters under nafo0:
qe0 qe1
...
```

Continue by assigning backup group numbers and adapters for all other backup groups in the configuration.

e. **Allow the `pnmset(1M)` command to test your adapter configuration.**

The `pnmset(1M)` command tests the correctness of your adapter configuration. In this example, the backup group contains one active adapter and two redundant adapters.

```
The following test will evaluate the correctness of the customer NAFO configuration...
name duplication test passed


Check nafo0... < 20 seconds
qe0 is active
remote address = 192.168.142.1
nafo0 test passed


Check nafo1... < 20 seconds
qe3 is active
remote address = 192.168.143.1
test qe4 wait...
test qe2 wait...
nafo1 test passed
phys-hahost1#
```

Once the configuration is verified, the PNM daemon `pnmd(1M)` automatically notes the configuration changes and starts monitoring the interfaces.

---

**Note -** Only one adapter within a backup group should be plumbed and have an entry in the `/etc/hostname.`*adapter* file. Do not assign IP addresses to the backup adapters; they should not be plumbed.

---

**Note -** PNM uses broadcast `ping(1M)` to monitor networks, which in turn uses broadcast ICMP (Internet Control Message Protocol) packets to communicate with other remote hosts. Some routers do not forward broadcast ICMP packets; consequently, PNM's fault detection behavior is affected. See the *Sun Cluster 2.2 Release Notes* for a workaround to this problem.

---

5. **Start the cluster by using the `scadmin(1M)` command.**

Run the following command on one node:

```
# scadmin startcluster physical-hostname sc-cluster
```

Then add all other nodes to the cluster by running the following command from all other nodes:

```
# scadmin startnode
```

6. **Verify the PNM configuration by using the** `pnmstat(1M)` **command.**

```
phys-hahost1# /opt/SUNWpnm/bin/pnmstat -l
bkggrp  r_adp   status  fo_time live_adp
nafo0   hme0    OK      NEVER   hme0
phys-hahost1#
```

You have now completed the initial setup of PNM.

## ▼ How to Reconfigure PNM

Use this procedure to reconfigure an existing PNM configuration by adding or removing network adapters. Follow these steps to administer one node at a time, so that Sun Cluster services remain available during the procedure.

1. **Stop the Sun Cluster software on the node to be reconfigured.**

```
phys-hahost1# scadmin stopnode
```

2. **Add or remove the network adapters.**

   Use the procedures described in “Adding and Removing Network Interfaces” on page 122.

3. **Run the** `pnmset(1M)` **command to reconfigure backup groups.**

   Use the `pnmset(1M)` command to reconfigure backup groups as described in Step 4 on page 116 of the procedure “How to Set Up PNM” on page 114.

```
phys-hahost1# pnmset
```

4. **Restart the Sun Cluster software on the node.**

   Restart the node by running the following command from the administrative workstation:

```
phys-hahost1# scadmin startnode
```

5. **Repeat Step 1 through Step 4 for each node you want to reconfigure.**

## ▼ How to Check the Status of Backup Groups

You can use the pnmptor(1M) and pnmrtop(1M) commands to check the status of local backup groups only, and the pnmstat(1M) command to check the status of local or remote backup groups.

1. **Run the** pnmptor(1M) **command to find the backup group to which an adapter belongs.**

   The pnmptor(1M) command maps a pseudo adapter name that you supply to a real adapter name. In this example, the system output shows that pseudo adapter name nafo0 is associated with the active adapter hme2:

```
phys-hahost1# pnmptor nafo0
hme2
```

1. **Run the** pnmrtop(1M) **command to find the active adapter associated with a given backup group.**

   In this example, the system output shows that adapter hme1 belongs to backup group nafo0:

```
phys-hahost1# pnmrtop hme1
nafo0
```

1. **Run the** pnmstat(1M) **command to determine the status of a backup group.**

   Use the -c option to determine the status of a backup group on the local host:

```
phys-hahost1# pnmstat -c nafo0
OK
NEVER
hme2
```

Use the following syntax to determine the status of a backup group on a remote host:

```
phys-hahost1# pnmstat -sh remotehost -c nafo1
OK
NEVER
qe1
```

**Note -** It is important to use the `-s` and `-h` options together. The `-s` option forces `pnmstat(1M)` to communicate over the private interconnect. If the `-s` option is omitted, `pnmstat(1M)` queries over the public interconnect. Both *remotehost* and the host on which you run `pnmstat(1M)` must be cluster members.

Whether checking the local or remote host, the `pnmstat(1M)` command reports the status, history, and current active adapter. See the man page for more details.

## PNM Configurable Parameters

The following table describes the PNM parameters that are user-configurable. Configure these parameters after you have installed PNM, but before you bring up the cluster, by manually editing the configuration file `/opt/SUNWcluster/conf/TEMPLATE.cdb` on all nodes in the cluster. You can edit the file on one node and copy the file to all other nodes, or use the Cluster Console to modify the file on all nodes simultaneously. You can display the current PNM configuration with `pnmd -t`. See the `pnmd(1M)` man page for details.

**TABLE 6–2** PNM Configurable Parameters

| | |
|---|---|
| `pnmd.inactive_time` | The time, in seconds, between fault probes. The default interval is 5 seconds. |
| `pnmd.ping_timeout` | The time, in seconds, after which a fault probe will time out. The default timeout value is 4 seconds. |
| `pnmd.repeat_test` | The number of times that PNM will retry a failed probe before deciding there is a problem. The default repeat quantity is 3. |
| `pnmd.slow_network` | The latency, in seconds, between the listening phase and actively probing phase of a fault probe. The default latency period is 2 seconds. If your network is slow, causing PNM to initiate spurious takeovers, consider increasing this latency period. |

# Troubleshooting PNM Errors

The following errors are those most commonly returned by PNM.

```
PNM rpc svc failed
```

This error indicates that the PNM daemon has not been started. Restart the PNM daemon with the following command. The *node-id* is the value returned by the `/opt/SUNWcluster/bin/get_node_status` command.

```
# /opt/SUNWpnm/bin/pnmd -s -c cluster-name -l node-id
```

```
PNM not started
```

This message indicates that no backup groups have been configured. Use the `pnmset(1M)` command to create backup groups.

```
No nafoXX
```

This message indicates that you have specified an illegal backup group name. Use the pnmrtop(1M) command to determine the backup group names associated with a given adapter. Rerun the command and supply it with a valid backup group name.

```
PNM configure error
```

This message indicates that either the PNM daemon was unable to configure an adapter, or that there is a formatting error in the configuration file, /etc/pnmconfig. Check the syslog messages and take the actions specified by Sun Cluster Manager. For more information on Sun Cluster Manager, see Chapter 2.

```
Program error
```

This message indicates that the PNM daemon was unable to execute a system call. Check the syslog messages and take the actions specified by Sun Cluster Manager. For more information on Sun Cluster Manager, see Chapter 2.

## Adding and Removing Network Interfaces

The procedures in this section can be used to add or remove public network interface cards within a cluster configuration.

To add or remove a network interface to or from the control of a logical host, you must modify each logical host configured to use that interface. You change a logical host's configuration by completely removing the logical host from the cluster, then adding it again with the required changes. You can reconfigure a logical host with either the scconf(1M) or scinstall(1M) command. The examples in this section use the scconf(1M) command. Refer to "Adding and Removing Logical Hosts" on page 58, for the logical host configuration steps using the scinstall(1M) command.

### Adding a Network Interface

Adding a network interface requires unconfiguring and reconfiguring all logical hosts associated with the interface. Note that all data services will be inaccessible for a short period of time during the procedure.

## How to Add a Network Interface

On each node that will receive a new network interface card, perform the following steps.

1. **Stop the cluster software.**

   ```
   phys-hahost# scadmin stopnode
   ```

2. **Add the new interface card, using the instructions included with the card.**

3. **Configure the new network interface on each node.**

   This step is necessary only if the new interface will be part of a logical host. Skip this step if your configuration does not include logical hosts.

   ```
   phys-hahost# pnmset
   ```

   For Ethernet, create a new `/etc/hostname.`*if* file for each new interface on each node, and run the `ifconfig(1M)` command as you normally would in a non-cluster environment.

   ---

   **Note -** When you configure a set of network interfaces to be used by different logical hosts within a cluster, you must connect all interfaces in the set to the same subnet.

   ---

4. **Start the cluster software.**

   If all nodes have been stopped, run the `scadmin startcluster` command on node 0 and then the `scadmin startnode` command on all other nodes. If at least one node has not had the cluster software stopped, run the `scadmin startnode` command on the remaining nodes.

   ```
   phys-hahost# scadmin startnode
   ```

   If the new interfaces are being added to already existing backup groups, the procedure is complete.

   If you modified the backup group configuration, you must bring the cluster back into normal operation and reconfigure each logical host that will be using the new set of network controllers. You will unconfigure and reconfigure each logical host, so run the `scconf -p` command to print out the current configuration before starting these steps. You can run the `scconf -p` command on any node that is an active cluster member; it does not need to be run on all cluster nodes.

To unconfigure and reconfigure the logical host, you can use either the `scconf(1M)` command as shown in these examples, or the `scinstall(1M)` command as described in “Adding and Removing Cluster Nodes” on page 52.

5. **Notify users that data services on the affected logical hosts will be unavailable for a short period.**

6. **Save copies of the** `/etc/opt/SUNWcluster/conf/ccd.database` **files on each node, in case you need to restore the original configuration.**

7. **Turn off the data services.**

```
phys-hahost# hareg -n dataservice
```

8. **Unregister the data services.**

```
phys-hahost# hareg -u dataservice
```

9. **Remove the logical host from the cluster.**

   Run this command on any node that is an active cluster member. You do not need to run this command on all cluster nodes.

```
phys-hahost# scconf clustername -L logicalhost -r
```

10. **Reconfigure the logical host to include the new interface.**

    Run this command on any node that is an active cluster member. You do not need to run this command on all cluster nodes.

```
phys-hahost# scconf clustername -L logicalhost -n nodelist -g dglist -i logaddrinfo
```

    The *logaddrinfo* field is where you define the new interface name. Refer to the listing taken from the `scconf -p` command output to reconstruct each logical host.

11. **Register the data services.**

```
phys-hahost# hareg [-s] -r dataservice
```

12. **Turn on the data services.**

```
phys-hahost# hareg -y dataservice
```

13. **Check access to the data services.**

14. **Notify users that the data services are once again available.**

This completes the process of adding a network interface.

## Removing a Network Interface

Use the following procedure to remove a public network interface from a cluster.

- If you have an OPS configuration, no action is required within the cluster to remove a network interface. However, if you want to remove the network adapters from the cluster nodes, perform the following procedure.
- If you have an HA configuration, you will need to use this procedure to unconfigure and reconfigure any logical host that uses the network interface to be removed. This requires that all data services will be inaccessible for a short period of time during the procedure.

## ▼ How to Remove a Network Interface

While all nodes are participating in the cluster, perform the following steps on one node only.

1. **Identify which logical hosts must be reconfigured to exclude the network interface.**

   All of these logical hosts will need to be unconfigured then reconfigured. Run the `scconf -p` command to print out a list of logical hosts in the current configuration; save this list for later use. You do not need to run the `scconf -p` command on all cluster nodes. You can run it on any node that is an active cluster member.

2. **Run the `pnmset(1M)` command to display the current PNM configuration.**

3. **Remove the controller from a backup group, if necessary.**

   If the controller to be removed is part of a backup group, remove the controller from all logical hosts, then run the `pnmset(1M)` command to remove the controller from the backup group.

4. **Notify users that any data services on the affected logical hosts will be unavailable for a short period.**

5. **Turn off the data services.**

```
phys-hahost# hareg -n dataservice
```

6. **Unregister the data services.**

```
phys-hahost# hareg -u dataservice
```

7. **Remove the logical host from the cluster.**

---

**Note -** To unconfigure and reconfigure the logical host (Step 7 on page 126 and Step 8 on page 126), you can either run the `scconf(1M)` command as shown, or run the `scinstall(1M)` command as described in "Adding and Removing Cluster Nodes" on page 52.

---

You can run this command on any node that is an active cluster member. You do not need to run it on all cluster nodes.

```
phys-hahost# scconf clustername -L logicalhost -r
```

8. **Reconfigure the logical host to include the new interface.**

You can run this command on any node that is an active cluster member. You do not need to run it on all cluster nodes.

```
phys-hahost# scconf clustername -L logicalhost -n nodelist -g dglist -i logaddrinfo
```

The *logaddrinfo* field is where you define the new interface name. Refer to the listing taken from the `scconf -p` command output to reconstruct each logical host.

9. **If the controller being removed was part of a backup group, rerun the** `pnmset(1M)` **command.**

Rerun the `pnmset(1M)` command and exclude the controller being removed.

10. **(Optional) If you are removing the network adapter from the nodes, perform the following steps on each affected node:**

a. **Stop the cluster software.**

```
phys-hahost# scadmin stopnode
```

b. **Halt the node and remove the interface card.**

c. **Boot the node.**

d. **Perform the Solaris system administration tasks you would normally perform to remove a network interface (remove *hostname.if* file, update /etc/hosts, etc).**

e. **Restart the cluster software. If all nodes were brought down, start the first node using the** scadmin startcluster **command. If at least one node is still running the cluster software, restart the other nodes.**

```
phys-hahost# scadmin startnode
```

11. **Register the data services.**

```
phys-hahost# hareg -r dataservice
```

12. **Turn on the data services.**

```
phys-hahost# hareg -y dataservice
```

13. **Check access to the data services.**

14. **Notify users that the data services are once again available.**

# Administering the Switch Management Agent

The Switch Management Agent (SMA) is a cluster module that maintains communication channels over the cluster private interconnect. It monitors the private interconnect and invokes a failover to a backup network if it detects a failure.

Note the following limitations before beginning the procedure:

- On SC2000/SS1000 nodes, do not install more than one SCI card on a single system board. More than one SCI card can cause spurious link resets on the SCI interconnect.
- On E10000 nodes, an SCI card should not be the only card on an SBus.
- On Sun StorEdge A3000 configurations, do not install SCI adapters and other A3000 hosts adapters on the same SBus.

See also Appendix B in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide*.

## ▼ How to Add Switches and SCI Cards

Use this procedure to add switches and SCI cards to cluster nodes. See the `sm_config(1M)` man page for details.

1. **Edit the** `sm_config` **template file to include the configuration changes.**

   Normally, the template file is located in `/opt/SUNWsma/bin/Examples`.

2. **Configure the SCI SBus cards by running the** `sm_config(1M)` **command from one of the nodes.**

   Rerun the command a second time to ensure that SCI node IDs and IP addresses are assigned correctly to the cluster nodes. Incorrect assignments can cause miscommunication between the nodes.

3. **Reboot the new nodes.**

## SCI Software Troubleshooting

If a problem occurs with the SCI software, verify that the following are true:

- The `sm_config(1M)` template file matches the hardware configuration (SCI link and switch) and cluster topology.
- The `sm_config(1M)` command can be run successfully from one of the cluster nodes.
- Any reconfigured nodes were rebooted after the `sm_config(1M)` command was executed.

Also note the following problems and solutions:

- Some applications, such as Oracle Parallel Server (OPS), require an unusually high shared memory minimum to be specified in the `/etc/system` file. If the field `shmsys:shminfo_shmmin` in the `/etc/system` file is set to a value greater than 200 bytes, the `sm_config(1M)` command will not be able to acquire shared memory because it requests fewer bytes than the minimum the system can allocate. As a result, the system call made by the sm_config(1M) command fails, and the command is aborted.

To work around this problem, edit the /etc/system file and set the value of shmsys:shminfo_shmmin to less than 200. Then reboot the machine for the new values to take effect.

- If you encounter semsys warnings and core dumps, check to see whether the semaphore values contained in the semsys:seminfo_* fields in the /etc/system file match the actual physical limits of the machine.

For more information about SCI components, see Appendix B in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide.*

## ▼ How to Verify Connectivity Between Nodes

There are two ways to verify the connectivity between nodes: by running get_ci_status(1M) or by running ping(1).

1. **Run the** get_ci_status(1M) **command on all cluster nodes.**

   Example output for get_ci_status(1M) is shown below.

```
# /opt/SUNWsma/bin/get_ci_status
sma: sci #0: sbus_slot# 1; adapter_id 8 (0x08); ip_address 1; switch_id# 0; port_id# 0; Adapter Status - UP; Link Status
 UP
sma: sci #1: sbus_slot# 2; adapter_id 12 (0x0c); ip_address 17; switch_id# 1; port_id# 0; Adapter Status - UP; Link Stat
 UP
sma: Switch_id# 0
sma: port_id# 1: host_name = interconn2; adapter_id = 72; active | operational
sma: port_id# 2: host_name = interconn3; adapter_id = 136; active | operational
sma: port_id# 3: host_name = interconn4; adapter_id = 200; active | operational
sma: Switch_id# 1
sma: port_id# 1: host_name = interconn2; adapter_id = 76; active | operational
sma: port_id# 2: host_name = interconn3; adapter_id = 140; active | operational
sma: port_id# 3: host_name = interconn4; adapter_id = 204; active | operational
#
```

   The first four lines indicate the status of the local node (in this case, interconn1). It is communicating with both switch_id# 0 and switch_id# 1 (Link Status - UP).

```
sma: sci #0: sbus_slot# 1; adapter_id 8 (0x08); ip_address 1; switch_id# 0; port_id# 0; Adapter Status - UP; Link Statu
 UP
sma: sci #1: sbus_slot# 2; adapter_id 12 (0x0c); ip_address 17; switch_id# 1; port_id# 0; Adapter Status - UP; Link Sta
 UP
```

The rest of the output indicates the global status of the other nodes in the cluster. All the ports on the two switches are communicating with their nodes. If there is a problem with the hardware, `inactive` is displayed (instead of `active`). If there is a problem with the software, `inoperational` is displayed (instead of `operational`).

```
sma: Switch_id# 0
sma: port_id# 1: host_name = interconn2; adapter_id = 72; active | operational
sma: port_id# 2: host_name = interconn3; adapter_id = 136; active | operational
sma: port_id# 3: host_name = interconn4; adapter_id = 200; active | operational
sma: Switch_id# 1
sma: port_id# 1: host_name = interconn2; adapter_id = 76; active | operational
sma: port_id# 2: host_name = interconn3; adapter_id = 140; active | operational
sma: port_id# 3: host_name = interconn4; adapter_id = 204; active | operational
#
```

1. **Run the `ping(1)` command on all the IP addresses of remote nodes.**

   Example output for `ping(1)` is shown below.

```
# ping IP-address
```

The IP addresses are found in the `/etc/sma.ip` file. Be sure to run the `ping(1)` command for each node in the cluster.

The `ping(1)` command returns an “alive” message indicating that the two ends are communicating without a problem. Otherwise, an error message is displayed.

For example,

```
# ping 204.152.65.2
204.152.65.2 is alive
```

## How to Verify the SCI Interface Configuration

1. **Run the** `ifconfig -a` **command to verify that all SCI interfaces are up and that the cluster nodes have the correct IP addresses.**

   The last 8 bits of the IP address should match the `IP` field value in the `/etc/sma.config` file.

```
# ifconfig -a
lo0: flags=849<UP,LOOPBACK,RUNNING,MULTICAST> mtu 8232
 inet 127.0.0.1 netmask ff000000
hme0: flags=863<UP,BROADCAST,NOTRAILERS,RUNNING,MULTICAST> mtu 1500
 inet 129.146.238.55 netmask ffffff00 broadcast 129.146.238.255
 ether 8:0:20:7b:fa:0
scid0: flags=80cl<UP,RUNNING,NOARP,PRIVATE> mtu 16321
 inet 204.152.65.1 netmask fffffff0
scid1: flags=80cl<UP,RUNNING,NOARP,PRIVATE> mtu 16321
 inet 204.152.65.17 netmask fffffff0
```

CHAPTER **7**

# Administering Server Components

This chapter describes the software procedure for adding or removing Sun Cluster node components. The topics in this chapter are listed below.

- "Replacing System Boards" on page 133
- "Adding Board-Level Modules" on page 134
- "Replacing SBus Cards" on page 135

## Replacing System Boards

The Solstice DiskSuite component of Sun Cluster is sensitive to device numbering and can become confused if system boards are moved around. Refer to Chapter 1, for more information on instance names and numbering.

When the node is booted initially, the multihost disk expansion unit entries in the `/dev` directory are tied to the connection slot.

For example, when the node is booted, system board 0 and SBus slot 1 will be part of the identity of the multihost disk expansion unit. If the board or SBus card is shuffled to a new location, Solstice DiskSuite will be confused because Solaris will assign new controller numbers to the SBus controllers when they are in a new location.

**Note -** The SBus cards can be moved as long as the type of SBus card in a slot remains the same.

Shuffling the fiber cables that lead to the multihost disk expansion units also can create problems. When SBus cards are switched, you must also reconnect the multihost disk expansion units back to the same SBus slot they were connected to before the changes.

## Adding Board-Level Modules

Adding or replacing board-level modules such as SIMMs and CPUs involves both software and hardware procedures.

### ▼ How to Add Board-Level Modules

1. **Stop Sun Cluster on the node that is to receive the board-level module.**

   In this example, `phys-hahost2` will be the first to receive the board-level module.

```
phys-hahost2# scadmin stopnode
```

2. **Halt the node.**

```
phys-hahost2# halt
```

3. **Power off the node.**

4. **Install the board-level module using the instructions in the appropriate hardware manual.**

5. **Power on the node.**

6. **Perform a reconfiguration reboot.**

```
ok boot -r
```

7. **Start the cluster software on the node.**

```
# scadmin startnode
```

8. **Repeat Step 1 on page 134 through Step 7 on page 134 on the other Sun Cluster nodes that need a similar hardware upgrade.**

9. **Switch each logical host back to its default master, if necessary.**

If manual mode is not set, an automatic switchback will occur.

```
phys-hahost2# haswitch phys-hahost1 hahost1
```

# Replacing SBus Cards

Replacement of SBus cards in Sun Cluster nodes can be done by switching over the data services to the node that is functioning and performing the hardware procedure to replace the board. The logical hosts can be switched back to the default masters following the procedure.

## ▼ How to Replace SBus Cards

1. **Switch ownership of the logical hosts from the Sun Cluster node that needs an SBus card replaced.**

   For instance, if the board is being replaced on the physical host `phys-hahost2`, enter the following:

   ```
   phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
   ```

2. **Stop Sun Cluster on the affected node.**

   Run the `scadmin(1M)` command with the `stopnode` option on the host that has the failed SBus card

```
phys-hahost2# scadmin stopnode
```

.

3. **Halt and power off the affected node.**

4. **Perform the hardware replacement procedure.**

   Refer to the instructions in the appropriate hardware service manual to replace the SBus card.

5. **Power on the node and start the cluster software on the node.**

```
# scadmin startnode
```

The node automatically will rejoin the Sun Cluster configuration.

6. **Switch the logical hosts back to the default masters, if necessary.**

   If manual mode is not set, an automatic switchback will occur.

   ```
   phys-hahost1# haswitch phys-hahost2 hahost2
   ```

CHAPTER **8**

# Administering the Terminal Concentrator

This chapter provides instructions for using the Terminal Concentrator when performing administration of Sun Cluster configurations. See also Chapter 5 in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide*.

The topics in this chapter are listed below.

- "Connecting to the Sun Cluster Console" on page 137
- "Resetting a Terminal Concentrator Connection" on page 140
- "Entering the OpenBoot PROM on a Sun Cluster Server" on page 141
- "Troubleshooting the Terminal Concentrator" on page 142
- "Changing TC/SSP Information" on page 146

## Connecting to the Sun Cluster Console

You can perform administrative tasks from a window connected to any Sun Cluster node. The procedures for initial setup of a Terminal Concentrator and how to set up security are in the hardware planning and installation manual for your Sun Cluster node and the Terminal Concentrator documentation.

The following procedure describes how to create connections from the administrative workstation in a Sun Cluster configuration.

Because a `shelltool(1)` can be of variable size and the connection is made through a serial-port console interface, the console port is incapable of determining the window size of the `shelltool(1)` from which the connection was made. You must set the window size manually on the nodes for any applications that require information about row and column quantities.

## How to Connect to the Sun Cluster Console

1. **Open a** `shelltool(1)` **window on the desktop of a workstation.**

2. **Run the tput(1) command and note the size of the** `shelltool(1)` **window.**

   These numbers will be used in Step 6 on page 139.

   ```
   # tput lines
   35
   # tput cols
   80
   ```

3. **Enter the following command to open a** `telnet(1)` **connection to one of the Sun Cluster nodes, through the Terminal Concentrator.**

   ```
   # telnet terminal-concentrator-name 5002
   Trying 192.9.200.1 ...
   Connected to 192.9.200.1.
   Escape character is '^]'.
   ```

---

**Note -** Port numbers are configuration dependent. Typically, ports 2 and 3 (5002 and 5003 in the examples) are used for the first Solaris cluster at a site.

---

4. **Open another** `shelltool(1)` **window and enter the following command to open a** `telnet(1)` **connection to the other node.**

   ```
   # telnet terminal-concentrator-name 5003
   Trying 192.9.200.1 ...
   Connected to 192.9.200.1.
   Escape character is '^]'.
   ```

**Note -** If you set up security as described in the hardware planning and installation guide for your Sun Cluster node, you will be prompted for the port password. After establishing the connection, you will be prompted for the login name and password.

5. **Log in to the node.**

```
Console login: root
Password: root-password
```

6. **Use the `stty(1)` command to reset the terminal rows and cols values to those found in Step 2 on page 138.**

```
# stty rows 35
# stty cols 80
```

7. **Set the `TERM` environment variable to the appropriate value based on the type of window used in Step 1 on page 145.**

   For example, if you are using an xterm window, type:

```
# TERM=xterm; export TERM (sh or ksh)
or
# setenv TERM xterm (csh)
```

# Resetting a Terminal Concentrator Connection

This section provides instructions for resetting a Terminal Concentrator connection.

If another user has a connection to the Sun Cluster node console port on the Terminal Concentrator, you can reset the port to disconnect that user. This procedure will be useful if you need to immediately perform an administrative task.

If you cannot connect to the Terminal Concentrator, the following message appears:

```
# telnet terminal-concentrator-name 5002
Trying 192.9.200.1 ...
telnet: Unable to connect to remote host: Connection refused
#
```

If you use the port selector, you might see a port busy message.

## ▼ How to Reset a Terminal Concentrator Connection

1. **Press an extra return after making the connection and select the command line interface (`cli`) to connect to the Terminal Concentrator.**

   The `annex:` prompt appears.

```
# telnet terminal-concentrator-name
...
Enter Annex port name or number: cli
...
annex:
```

2. **Enter the `su` command and password.**

   By default, the password is the IP address of the Terminal Concentrator.

```
annex: su
Password:
```

3. **Determine which port you want to reset.**

   The port in this example is Port 2. Use the Terminal Concentrator’s built-in `who` command to show connections.

```
annex# who
Port   What   User   Location     When   Idle   Address
2   PSVR   ---   ---     ---   1:27   192.9.75.12
v1   CLI   ---   ---     ---      192.9.76.10
```

4. **Reset the port.**

   Use the Terminal Concentrator’s built-in `reset` command to reset the port. This example breaks the connection on Port 2.

```
annex# admin reset 2
```

5. **Disconnect from the Terminal Concentrator.**

```
annex# hangup
```

6. **Reconnect to the port.**

```
# telnet terminal-concentrator-name 5002
```

## Entering the OpenBoot PROM on a Sun Cluster Server

This section contains information for entering the OpenBoot PROM from the Terminal Concentrator.

## ▼ How to Enter the OpenBoot PROM

1. **Connect to the port.**

```
# telnet terminal-concentrator-name 5002
Trying 192.9.200.1 ...
Connected to 129.9.200.1 .
Escape character is '^]'.
```

2. **Stop the cluster software, if necessary, by using the** `scadmin stopnode` **command, and then halt the system.**

   Halt the system gracefully by using the `halt(1M)` command.

```
# halt
```

   If halting the system with the `halt(1M)` command is not possible, then enter the `telnet(1)` command mode. The default `telnet(1)` escape character is `Control-]`.

3. **Send a break to the node.**

```
telnet> send brk
```

4. **Execute the OpenBoot PROM commands.**

# Troubleshooting the Terminal Concentrator

This section describes troubleshooting techniques associated with the Terminal Concentrator. See also Chapter 5 in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide.*

## Port Configuration Access Errors

A `connect: Connection refused` message while trying to access a particular Terminal Concentrator port using `telnet(1)` can have two possible causes:

- The port is being used by someone else.
- The port is misconfigured and not accepting network connections.

## How to Correct a Port Configuration Access Error

1. **Telnet to the Terminal Concentrator without specifying the port, and then interactively specify the port.**

```
# telnet terminal-concentrator-name
Trying ip_address ..
Connected to 192.9.200.1
Escape character is "^]".
[you may have to enter a RETURN to see the following prompts]

Rotaries Defined:
      cli                              -

Enter Annex port name or number: 2
```

If you see the following message, the port is in use.

```
Port(s) busy, do you wish to wait? (y/n) [y]:
```

If you see the following message, the port is misconfigured.

```
Port 2
Error: Permission denied.
```

If the port is in use, reset the Terminal Concentrator connections using the instructions in “Resetting a Terminal Concentrator Connection” on page 140.

If the port is misconfigured, do the following:

a. **Select the command-line interpreter (`cli`) and become the Terminal Concentrator superuser.**

```
Enter Annex port name or number: cli

Annex Command Line Interpreter   *   Copyright 1991 Xylogics, Inc.

annex: su
Password:
```

**b. As the Terminal Concentrator superuser, reset the port mode.**

```
annex# admin
Annex administration MICRO-XL-UX R7.0.1, 8 ports
admin: port 2
admin: set port mode slave
 You may need to reset the appropriate port, Annex subsystem or
 reboot the Annex for changes to take effect.
admin: reset 2
admin:
```

The port is now configured correctly.

For more information about the Terminal Concentrator administrative commands, see the *Sun Terminal Concentrator General Reference Guide.*

## Random Interruptions to Terminal Concentrator Connections

Terminal concentrator connections made through a router can experience intermittent interruptions. These connections might come alive for random periods, then go dead again. When the connection is dead, any new Terminal Concentrator connection attempts will time out. The Terminal Concentrator will show no signs of rebooting. Subsequently, a needed route might be re-established, only to disappear again. The problem is due to Terminal Concentrator routing table overflow and loss of network connection.

This is not a problem for connections made from a host that resides on the same network as the Terminal Concentrator.

The solution is to establish a default route within the Terminal Concentrator and disable the `routed` feature. You must disable the `routed` feature to prevent the default route from being lost. The following procedure shows how to do this. See the Terminal Concentrator documentation for additional information.

The `config.annex` file is created in the Terminal Concentrator's EEPROM file system and defines the default route to be used. You also can use the `config.annex` file to define rotaries that allow a symbolic name to be used instead of a port number. Disable the `routed` feature using the Terminal Concentrator's `set` command.

## ▼ How to Establish a Default Route

1. **Open a `shelltool(1)` connection to the Terminal Concentrator.**

```
# telnet terminal-concentrator-name
Trying 192.9.200.2 ...
Connected to xx-tc.
Escape character is '^]'.


Rotaries Defined:
    cli                              -

Enter Annex port name or number: cli


Annex Command Line Interpreter   *   Copyright 1991 Xylogics, Inc.
```

2. **Enter the `su` command and administrative password.**

   By default, the password is the IP address of the Terminal Concentrator.

```
annex: su
Password: administrative-password
```

3. **Edit the `config.annex` file.**

```
annex# edit config.annex
```

4. **Enter the highlighted information appearing in the following example, substituting the appropriate IP address for your default router:**

```
Ctrl-W:save and exit Ctrl-X: exit Ctrl-F:page down Ctrl-B:page up
%gateway
net default gateway 192.9.200.2 metric 1 active ^W
```

5. **Disable the local `routed`.**

```
annex# admin set annex routed n
   You may need to reset the appropriate port, Annex subsystem or
   reboot the Annex for changes to take effect.
annex#
```

6. **Reboot the Terminal Concentrator.**

```
annex# boot
```

It takes a few minutes for the Terminal Concentrator to boot. During this time, the Sun Cluster node consoles are inaccessible.

## Changing TC/SSP Information

In Sun Cluster 2.2, information about the Terminal Concentrator (TC) or a System Service Processor (SSP) (Sun Enterprise 10000 only) is required during installation. The TC or SSP information is stored in the cluster configuration file.

This information is used to:

- Forcibly terminate hung nodes
- Implement a cluster-wide locking mechanism which prevents partitioned nodes from joining the cluster

Both these mechanisms serve to protect data integrity in the case of four-node clusters with directly attached storage devices.

---

**Note -** If you are using Solstice DiskSuite, `tcmon` and `quorum` will be disabled, and TC information is not required.

---

The `scconf(1m)` command enables you to change this information in the cluster configuration file if, for example, modifications are made to this part of the cluster hardware configuration.

For additional information on changing TC or SSP information, see Table 8–1 and the `scconf(1M)` man page. See also Chapter 5 in the *Sun Cluster 2.2 Hardware Site Preparation, Planning, and Installation Guide*.

**Note -** These commands must be run on all cluster nodes.

**TABLE 8–1** Modifying the Host Information for All Hosts Associated With a TC or SSP

| To Accomplish This... | Run This Command |
|---|---|
| Replace the IP address/name of a TC | `scconf(1m) -t -i` *new-ip-address old-IP-address* \| *TC-name* |
| Supply a new password | `scconf(1m) -t -P` *old-IP-address* \| *TC-name* |
| Change the port number used for the cluster-wide locking mechanism (TC only) | `scconf(1m) -t -l` *new-port old-IP-address* \| *TC-name* |

## How to Change Host Information

1. **Run the `scconf -H` command to change the information associated with a particular host. For example, to change a given host's architecture type and specify a new IP address for its SSP (or TC), run this command on all cluster nodes, where `-d` specifies the new architecture (Sun Enterprise 10000) associated with the host, and `-t` specifies a new IP address or host name (`foo-ssp`) for the SSP (or TC) connected to the host:**

```
# scconf clustername -H foo -d E10000 -t foo-ssp
```

## How to Specify a Port Number for an SSP or TC

1. **Run the `scconf -p` command on all cluster nodes to specify a port number for this host's console for this SSP (or TC).**

```
# scconf clustername -H hostname -p port-number
```

For example:

```
# scconf clustername -H foo -p 10
```

Multiple hosts may be connected to the same TC, and the `-H` option only affects the information associated with a particular host.

## ▼ How to Change the Configuration of a TC

1. **Run the** `scconf -t` **command on all cluster nodes to change the configuration of a particular TC in the system. For example, to change a TC IP address, use the following command, in which** `-i` **specifies a new IP address (**`129.34.123.52`**) for the specified Terminal Concentrator (or SSP), and** `-l` **specifies a new port (**`8`**) used for locking purposes in failure fencing:**

   ```
   # scconf clustername -t foo-tc -i 129.34.123.52 -l -8
   ```

   If a Terminal Concentrator is being used, an unused TC port from `2` to *n* is specified, where *n* is the number of ports in the TC. If a SSP is being used, a value of `-1` must be specified.

## ▼ How to Specify a New Password for an SSP or TC

1. **Run the** `scconf -P` **command on all cluster nodes to specify a new password for this SSP (or TC).**

```
# scconf clustername -t foo-ssp -P
foo-ssp(129.34.123.51) Password:*****
```

**Note -** If you change the user password on the SSP or TC, you also need to notify Sun Cluster software of the change by running this procedure from each cluster node. Otherwise, failure fencing might not work properly when a faulty node needs to be brought down forcibly by a “send break” from the SSP or TC.

CHAPTER **9**

# Using Dual-String Mediators

This chapter describes the Solstice DiskSuite feature that allows Sun Cluster to run highly available data services using only two disk strings. The topics in this chapter are listed below. Refer to the Solstice DiskSuite documentation for more information about the Solstice DiskSuite features and concepts.

- "Overview of Mediators" on page 151
- "Configuring Mediators" on page 152
- "Failures Addressed by Mediators" on page 153
- "Administering Mediators" on page 156

## Overview of Mediators

The requirement for Sun Cluster is that a dual-string configuration must survive the failure of a single node or a single string of drives without user intervention.

In a dual-string configuration, metadevice state database replicas are always placed such that exactly half of the replicas are on one string and half are on a second string. A quorum (half + 1 or more) of the replicas is required to guarantee that the most current data is being presented. In the dual-string configuration, if one string becomes unavailable, a quorum of the replicas will not be available.

A mediator is a host (node) that stores mediator data. Mediator data provides information about the location of other mediators and contains a commit count that is identical to the commit count stored in the database replicas. This commit count is used to confirm that the mediator data is in sync with the data in the database replicas. Mediator data is individually verified before use.

Solstice DiskSuite requires a replica quorum (half + 1) to determine when “safe” operating conditions exist. This guarantees data correctness. With a dual-string configuration, it is possible that only one string is accessible. In this situation it is impossible to get a replica quorum. If mediators are used and a mediator quorum is present, the mediator data can help you determine whether the data on the accessible string is up-to-date and safe to use.

The introduction of mediators enables the Sun Cluster software to ensure that the most current data is presented in the case of a single string failure in a dual-string configuration.

## Golden Mediators

To avoid unnecessary user intervention in some dual-string failure scenarios, the concept of a golden mediator has been implemented. If exactly half of the database replicas are accessible and an event occurs that causes the mediator hosts to be updated, two mediator updates are attempted. The first update attempts to change the commit count and to set the mediator to not golden. The second update occurs if and only if during the first phase, all mediator hosts were successfully contacted and the number of replicas that were accessible (and which had their commit count advanced) were exactly half of the total number of replicas. If all the conditions are met, the second update sets the mediator status to golden. The golden status enables a takeover to proceed, without user intervention, to the host with the golden status. If the status is not golden, the data will be set to read-only, and user intervention is required for a takeover or failover to succeed. For the user to initiate a takeover or failover, exactly half of the replicas must be accessible.

The golden state is stored in volatile memory (RAM) only. Once a takeover occurs, the mediator data is once again updated. If any mediator hosts cannot be updated, the golden state is revoked. Since the state is in RAM only, a reboot of a mediator host causes the golden state to be revoked. The default state for mediators is not golden.

# Configuring Mediators

Figure 9–1 shows a Sun Cluster system configured with two strings and mediators on two Sun Cluster nodes.

Regardless of the number of nodes, there are still only two mediator hosts in the cluster. The mediator hosts are the same for all disksets using mediators in a given cluster, even when a mediator host is not a member of the server set capable of mastering the diskset.

To simplify the presentation, the configurations shown here use only one diskset and a symmetric configuration. The number of disksets is not significant in these sample scenarios. In the stable state, the diskset is mastered by `phys-hahost1`.

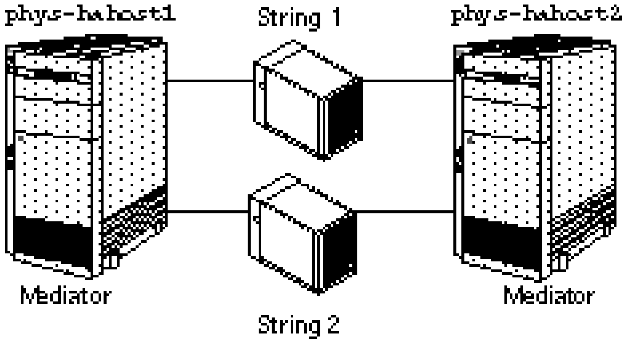


*Figure 9–1* Sun Cluster System in Steady State With Mediators

Normally, if half + 1 of the database replicas are accessible, then mediators are not used. When exactly half of the replicas are accessible, the mediator's commit count can be used to determine whether the accessible half is the most up to date. To guarantee that the correct mediator commit count is being used, both of the mediators must be accessible, or the mediator must be golden. Half + 1 of the mediators constitutes a *mediator quorum*. The mediator quorum is independent of the replica quorum.

# Failures Addressed by Mediators

With mediators, it is possible to recover from single failures, as well as some double failures. Since Sun Cluster only guarantees automatic recovery from single failures, only the single-failure recovery situation is covered here in detail. The double failure scenarios are included, but only general recovery processes are described.

Figure 9–1 shows a dual-string configuration in the stable state. Note that mediators are established on both Sun Cluster nodes, so both nodes must be up for a mediator quorum to exist and for mediators to be used. If one Sun Cluster node fails, a replica quorum will exist. If a takeover of the diskset is necessary, the takeover will occur without the use of mediators.

The following sections show various failure scenarios and describe how mediators help recover from these failures.

## Single Server Failure

Figure 9–2 shows the situation where one Sun Cluster node fails. In this case, the mediator software is not used since there is a replica quorum available. Sun Cluster node phys-hahost2 will take over the diskset previously mastered by phys-hahost1.

The process for recovery in this scenario is identical to the process followed when one Sun Cluster node fails and there are more than two disk strings. No administrator action is required except perhaps to switch over the diskset after phys-hahost1 rejoins the cluster. See the haswitch(1M) man page for more information about the switchover procedure.

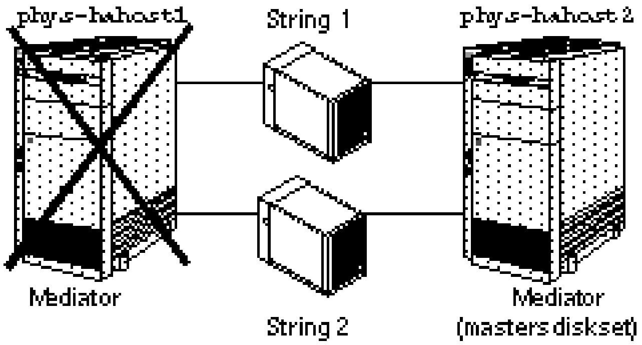


*Figure 9–2* Single Sun Cluster Server Failure With Mediators

## Single String Failure

Figure 9–3 illustrates the case where, starting from the steady state shown in Figure 9–1, a single string fails. When String 1 fails, the mediator hosts on both phys-hahost1 and phys-hahost2 will be updated to reflect the event, and the system will continue to run as follows:

- No takeover occurs.
- Sun Cluster node phys-hahost1 continues to own the diskset.
- Because String 1 failed, it must be resynchronized with String 2. For more information about the resynchronization process, refer to the *Solstice DiskSuite User's Guide* and the metareplace(1M) man page.

The commit count is incremented and the mediators remain golden.

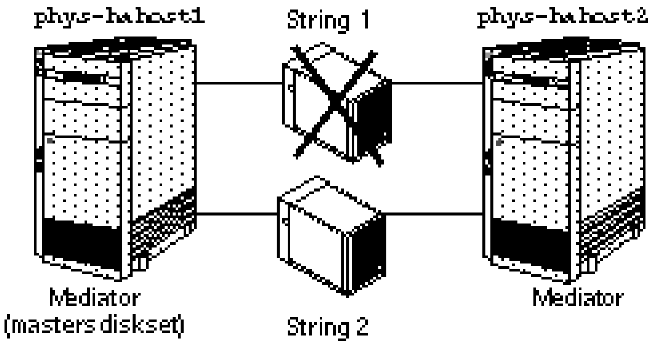


*Figure 9–3* Single String Failure With Mediators

The administration required in this scenario is the same that is required when a single string fails in the three or more string configuration. Refer to the relevant chapter on administration of your disk expansion unit for details on these procedures.

## Host and String Failure

Figure 9–4 shows a double failure where both String 1 and `phys-hahost2` fail. If the failure sequence is such that the string fails first, and later the host fails, the mediator on `phys-hahost1` could be golden. In this case, we have the following conditions:

- The mediator on `phys-hahost1` is golden.
- Half of the mediators are available.
- Half of the replicas are accessible.
- The mediator commit count on `phys-hahost1` matches the commit count found in the replicas on String 2

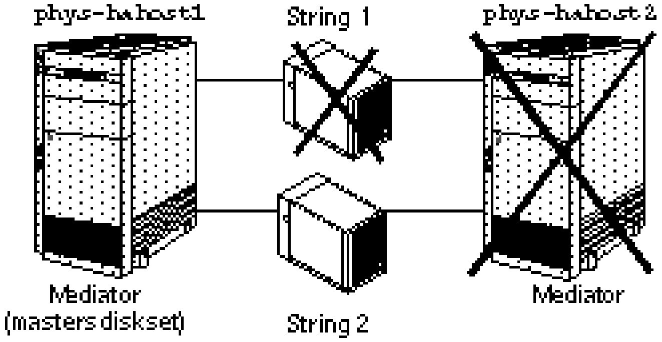


*Figure 9–4* Multiple Failure – One Server and One String

This type of failure is recovered automatically by Sun Cluster. If `phys-hahost2` mastered the diskset, `phys-hahost1` will take over mastery of the diskset. Otherwise, mastery of the diskset will be retained by `phys-hahost1`. After String 1 is fixed, the data on String 1 must be resynchronized with the data on String 2. For more information about the resynchronization process, refer to the *Solstice DiskSuite User's Guide* and the `metareplace(1M)` man page.

**Caution -** Although you can recover from this scenario, you must be sure to restore the failed components immediately since a third failure will cause the cluster to be unavailable.

If the mediator on `phys-hahost1` is not golden, this case is not automatically recovered by Sun Cluster and requires administrative intervention. In this case, Sun Cluster generates an error message and the logical host is put into maintenance mode (read-only). If this or any other multiple failure occurs, contact your service provider to assist you.

# Administering Mediators

Administer mediator hosts with the `medstat(1M)` and `metaset(1M)` commands. Use these commands to add or delete mediator hosts, and to check and fix mediator data. See the `medstat(1M)`, `metaset(1M)`, and `mediator(7)` man pages for details.

## ▼ How to Add Mediator Hosts

Use this procedure after you have installed and configured Solstice DiskSuite.

1. **Start the cluster software on all nodes.**

   On the first node:

   ```
   # scadmin startcluster
   ```

   On all remaining nodes:

   ```
   # scadmin startnode
   ```

2. **Determine the name of the private link for each node.**

   Use `grep(1)` to identify the private link included in the *clustername*.`cdb` file.

   ```
   hahost1# grep ‘‘^cluster.node.0.hostname’’ \
   /etc/opt/SUNWcluster/conf/clustername.cdb
   cluster.node.0.hostname : hahost0
   phys-hahost1# grep ‘‘cluster.node.0.hahost0’’ \
   /etc/opt/SUNWcluster/conf/clustername.cdb | grep 204
   204.152.65.33

   hahost1# grep ‘‘^cluster.node.1.hostname’’ \
   /etc/opt/SUNWcluster/conf/clustername.cdb
   cluster.node.1.hostname : hahost1
   hahost1# grep ‘‘cluster.node.1.hahost1’’ \
   /etc/opt/SUNWcluster/conf/clustername.cdb | grep 204
   204.152.65.34
   ```

   In this example, 204.152.65.33 is the private link for `hahost0` and 204.152.65.34 is the private link for `hahost1`.

3. **Configure mediators using the** `metaset(1M)` **command.**

   Add each host with connectivity to the diskset as a mediator for that diskset. Run each command on the host currently mastering the diskset. You can use the `hastat(1M)` command to determine the current master of the diskset. The information returned by `hastat(1M)` for the logical host identifies the diskset master.

```
hahost1# metaset -s disksetA -a -m hahost0,204.152.65.33
hahost1# metaset -s disksetA -a -m hahost1,204.152.65.34
hahost1# metaset -s disksetB -a -m hahost0,204.152.65.33
hahost1# metaset -s disksetB -a -m hahost1,204.152.65.34
hahost1# metaset -s disksetC -a -m hahost0,204.152.65.33
hahost1# metaset -s disksetC -a -m hahost1,204.152.65.34
```

The `metaset(1M)` command treats the private link as an alias.

## How to Check the Status of Mediator Data

1. **Run the `medstat(1M)` command.**

```
phys-hahost1# medstat -s diskset
```

See the `medstat(1M)` man page to interpret the output. If the output indicates that the mediator data for any one of the mediator hosts for a given diskset is bad, refer to the following procedure to fix the problem.

## How to Fix Bad Mediator Data

---

**Note -** The `medstat(1M)` command checks the status of mediators. Use this procedure if `medstat(1M)` reports that a mediator host is bad.

---

1. **Remove the bad mediator host(s) from all affected diskset(s).**

   Log into the Sun Cluster node that owns the affected diskset and enter:

```
phys-hahost1# metaset -s diskset -d -m bad_mediator_host
```

2. **Restore the mediator host and its aliases:**

```
phys-hahost1# metaset -s diskset -a -m bad_mediator_host, physical_host_alias,...
```

**Note -** The private links must be assigned as mediator host aliases. Specify the physical host IP address first, and then the HA private link on the metaset(1M) command line. See the mediator(7) man page for details on this use of the metaset(1M) command.

## Handling Failures Without Automatic Recovery

Certain double-failure scenarios exist that do not allow for automatic recovery by Sun Cluster. They include the following:

- Both a node and a string have failed in a dual string configuration, but the mediator on the surviving node was not golden. This scenario is further described in “Host and String Failure” on page 155.
- Mediator data is bad, stale, or non-existent on one or both of the nodes and one of the strings in a dual string configuration fails. The next attempt to take ownership of the affected logical host(s) will fail.
- A string fails in a dual string configuration, but the number of good replicas on the surviving string does not represent at least half of the total replica count for the failed diskset. The next attempt by DiskSuite to update these replicas will result in a system panic.
- A failure with no automatic recovery has occurred, and an attempt is made to bring the affected logical host(s) out of maintenance mode before manual recovery procedures have been completed.

It is very important to monitor the state of the disksets, replicas, and mediators regularly. The medstat(1M) command is useful for this purpose. Bad mediator data, replicas, and disks should always be repaired immediately to avoid the risk of potentially damaging multiple failure scenarios.

When a failure of this type does occur, one of the following sets of error messages will be logged:

```
ERROR: metaset -s <diskset> -f -t exited with code 66
ERROR: Stale database for diskset <diskset>
NOTICE: Diskset <diskset> released

ERROR: metaset -s <diskset> -f -t exited with code 2
ERROR: Tagged data encountered for diskset <diskset>
NOTICE: Diskset <diskset> released

ERROR: metaset -s <diskset> -f -t exited with code 3
ERROR: Only 50% replicas and 50% mediator hosts available for diskset <diskset>
NOTICE: Diskset <diskset> released
```

Eventually, the following set of messages also will be issued:

```
ERROR: Could not take ownership of logical host(s) <lhost>, so switching into maintenance mode
ERROR: Once in maintenance mode, a logical host stays in maintenance mode until the admin interv
ERROR: The admin must investigate/
repair the problem and if appropriate use haswitch command to move the logical host(s) out of ma
```

Note that for a dual failure of this nature, high availability goals are sacrificed in favor of attempting to preserve data integrity. Your data might be unavailable for some time. In addition, it is not possible to guarantee complete data recovery or integrity.

Your service provider should be contacted immediately. Only an authorized service representative should attempt manual recovery from this type of dual failure. A carefully planned and well coordinated effort is essential to data recovery. Do nothing until your service representative arrives at the site.

Your service provider will inspect the log messages, evaluate the problem, and, possibly, repair any damaged hardware. Your service provider might then be able to regain access to the data by using some of the special `metaset(1M)` options described on the `mediator(7)` man page. However, such options should be used with extreme care to avoid recovery of the wrong data.

**Caution -** Attempts to alternate access between the two strings should be avoided at all costs; such attempts will make the situation worse.

Before restoring client access to the data, exercise any available validation procedures on the entire dataset or on any data affected by recent transactions against the dataset.

Before you run the `haswitch(1M)` command to return any logical host from maintenance mode, make sure that you release ownership of the associated diskset.

## Error Log Messages Associated With Mediators

The following `syslog` or console messages indicate that there is a problem with mediators or mediator data. Use the procedure "How to Fix Bad Mediator Data" on page 158 to address the problem.

```
Attention required -
 medstat shows bad mediator data on host %s for diskset %s

Attention required -
 medstat finds a fatal error in probing mediator data on host %s for diskset %s!

Attention required - medstat failed for diskset %s
```

**(Continuation)**

CHAPTER **10**

# Administering Sun Cluster Local Disks

This chapter provides instructions for administering Sun Cluster local disks. Some of the procedures documented in this chapter are dependent on your volume management software (Solstice DiskSuite or VxVM). Those that are dependent on the volume manager include the volume manager name in the procedure title.

This chapter includes the following topics:

- "Restoring a Local Boot Disk From Backup" on page 163
- "Replacing a Local Non-Boot Disk" on page 167

Sun Cluster administration involves monitoring the status of the configuration. (See Chapter 2, for information about monitoring methods.) The monitoring process might reveal problems with the local disks. The following sections provide instructions for correcting these problems.

For multihost disk administration procedures, see the administration chapter for your particular disk expansion unit. Also refer to your volume manager software documentation when you are replacing or repairing hardware in the Sun Cluster configuration.

## Restoring a Local Boot Disk From Backup

Some situations require you to replace a cluster node's boot disk, such as when a software problem leaves the boot disk in an unknown state, an operating system upgrade fails, or a hardware problem occurs. Use the following procedures to restore the boot disk to a known state, or to replace the disk.

**Note -** These procedures assume the existence of a backup copy of the boot disk.

## ▼ How to Restore a Local Boot Disk From Backup (Solstice DiskSuite)

When the physical hosts are in the same cluster, this procedure is performed on the local host while another host provides data services for all hosts. In this example, we use two physical hosts `phys-hahost1` and `phys-hahost2`, and two logical hosts `hahost1` and `hahost2`.

These are the high-level steps to restore a boot disk from backup in a Solstice DiskSuite configuration.

- Removing the host containing the boot disk from the disksets
- Restoring the boot disk from backup
- Renewing or creating replicas on the restored disk
- Adding the host back to the disksets
- Starting Sun Cluster on that host
- Switching over the logical host to its default master (if manual mode is set for switchback)

These are the detailed steps to restore a boot disk from backup in a Solstice DiskSuite configuration. In this example, `phys-hahost1` contains the disk to be restored. The boot disk is not mirrored.

1. **Halt the host requiring the restore.**

2. **On the other hosts in the cluster, use the** `metaset(1M)` **command to remove the host being restored from the disksets.**

   In this example, the `metaset(1M)` command is run from the other host in the cluster, `phys-hahost2`.

```
phys-hahost2# metaset -s hahost1 -f -d -h phys-hahost1
phys-hahost2# metaset -s hahost2 -f -d -h phys-hahost1
```

3. **Restore the boot disk on the host being restored from the backup media.**

Follow the procedure to restore files and file systems, as found in your Solaris system administration documentation to restore the boot disk file system.

4. **Reboot the host being restored.**

5. **Remove old DiskSuite replicas and reboot.**

   If you are replacing a failed disk, old replicas will not be present. If you are restoring a disk, run the `metadb(1M)` command to check whether old replicas are present. If so, delete the old replicas.

   ---

   **Note -** The default location for replicas is Slice 7. However, you are not required to place replicas on Slice 7.

   ```
   phys-hahost1# metadb -d -f c0t3d0s7
   phys-hahost1# reboot
   ```

   ---

6. **Create new DiskSuite replicas on the restored disk with the** `metadb(1M)` **command.**

   ```
   phys-hahost1# metadb -afc 3 c0t3d0s7
   ```

7. **Add the restored host to the diskset or disksets, from the sibling host.**

   ```
   phys-hahost2# metaset -s hahost1 -a -h phys-hahost1
   phys-hahost2# metaset -s hahost2 -a -h phys-hahost1
   ```

8. **Start Sun Cluster on the restored host.**

   ```
   phys-hahost1# scadmin startnode
   ```

9. **Switch back the logical hosts to the default master, if necessary.**

   If manual mode is not set, an automatic switchback will occur.

```
phys-hahost1# haswitch phys-hahost1 hahost1
```

## ▼ How to Restore a Local Boot Disk From Backup (VxVM)

When the physical hosts are in the same cluster, this procedure is performed on the local host while another host provides data services for all hosts. In this example, we use two physical hosts phys-hahost1 and phys-hahost2, and two logical hosts hahost1 and hahost2. In this example, the boot disk is not mirrored.

These are the high-level steps to restore a boot disk from backup in a VxVM configuration.

- Halting the host requiring the restore
- Restoring the boot disk from backup
- Starting Sun Cluster on that host
- Switching over the logical host to its default master (if manual mode is set for switchback)

These are the detailed steps to restore a boot disk from backup in a VxVM configuration. In this example, phys-hahost1 contains the disk to be restored.

1. **Halt the host requiring the restore.**

2. **Restore the boot disk on the host being restored from the backup media.**

   Follow the procedure to restore files and file systems as found in your Solaris system administration documentation to restore the boot disk file system.

3. **Reboot the host being restored.**

   The reboot causes the host to discover all the devices.

   ---

   **Note -** If the disks are reserved, it may be necessary to run vxdctl enable at a later time, when reservations are released.

   ---

4. **Start Sun Cluster on the local host.**

   ```
   phys-hahost1# scadmin startnode
   ```

5. **Switch back the logical hosts to the default master, if necessary.**

   If manual mode is not set, an automatic switchback will occur.

```
phys-hahost1# haswitch phys-hahost1 hahost1
```

# Replacing a Local Non-Boot Disk

This section describes the replacement of a failed local disk that does not contain the Solaris operating environment.

In general, if a local non-boot disk fails, you recover using a backup copy to restore the data to a new disk.

The procedures for restoring a local boot disk are described in “How to Restore a Local Boot Disk From Backup (Solstice DiskSuite)” on page 164, and in “How to Restore a Local Boot Disk From Backup (VxVM)” on page 166.

These are the high-level steps to replace a failed local non-boot disk.

- (Optional) Stopping Sun Cluster on the node with the bad disk, and shutting down that node
- Replacing the disk
- Formatting and partitioning the new disk
- Restoring data from a backup copy
- Starting Sun Cluster on that host
- Switching over the logical host to its default master (if manual mode is set for switchback)

## ▼ How to Replace a Local Non-Boot Disk

These are the detailed steps to replace a failed local non-boot disk. In this example, `phys-hahost2` contains the disk that failed.

1. **(Optional) Shut down the Sun Cluster services on the node with the failed disk and halt the node.**

   You may not need to perform this step if the node boots from a SPARCstorage Array disk. However, if the disk to be replaced is on the same SCSI bus as the functioning boot disk, you must shut down Sun Cluster and halt the node.

```
# scadmin stopnode
...
```

**(Continuation)**

```
# halt
```

2. **Perform the disk replacement.**

   Use the procedure described in the service manual for your Sun Cluster node.

3. **Start the node in single-user mode.**

4. **Run the** `format(1M)` **or** `fmthard(1M)` **command to repartition the new disk.**

   Make sure that you partition the new disk exactly as the disk that was replaced. (Saving the disk format information is outlined in Chapter 1.)

5. **Run the** `newfs(1M)` **command on the new slices to create file systems.**

6. **Run the** `mount(1M)` **command to mount the appropriate file systems.**

   Specify the device and mount points for each file system.

7. **Restore data from a backup copy.**

   Use the instructions in you Solaris system administration documentation to perform this step.

8. **Reboot the node.**

9. **Start Sun Cluster on the local host.**

   ```
   phys-hahost1# scadmin startnode
   ```

10. **Switch back the logical hosts to the default master, if necessary.**

    If manual mode is not set, an automatic switchback will occur.

    ```
    phys-hahost2# haswitch phys-hahost2 hahost2
    ```

CHAPTER **11**

# Administering SPARCstorage Arrays

This chapter provides instructions for administering SPARCstorage Array Model 100 Series, SPARCstorage Array Model 200 Series with differential SCSI, and SPARCstorage Array Model 200 Series with RSM™ disk trays. See also Chapter 3 in the *Sun Cluster 2.2 Hardware Service Manual*.

This chapter includes the following topics.

- "Recovering From Power Loss" on page 169
- "Repairing a Lost SPARCstorage Array Connection" on page 174
- "Adding a SPARCstorage Array" on page 175
- "Administering SPARCstorage Array Trays" on page 177
- "Replacing a SPARCstorage Array Controller and Changing the World Wide Name" on page 184
- "Administering SPARCstorage Array Disks" on page 193
- "Administering SPARCstorage Array NVRAM" on page 216

Use the service manual for your SPARCstorage Array and your volume manager documentation when replacing or repairing SPARCstorage Array hardware in the Sun Cluster configuration.

## Recovering From Power Loss

When power is lost to one SPARCstorage Array, I/O operations generate errors that are detected by the volume management software. Errors are not reported until I/O transactions are made to the disk. Hot spare activity can be initiated if affected devices are set up for hot sparing.

You should monitor the configuration for these events. See Chapter 2, for more information on monitoring the configuration.

## ▼ How to Recover From Power Loss (Solstice DiskSuite)

These are the high-level steps to recover from power loss to a SPARCstorage Array in a Solstice DiskSuite configuration:

- Identifying the errored replicas
- Returning the errored replicas to service
- Identifying the errored devices
- Returning the errored devices to service
- Resyncing the disks

These are the detailed steps to recover from power loss to a SPARCstorage Array in a Solstice DiskSuite configuration.

1. **When power is restored, run the** `metadb(1M)` **command to identify the errored replicas.**

   ```
   # metadb -s diskset
   ```

2. **Return replicas to service.**

   After the loss of power, all metadevice state database replicas on the affected SPARCstorage Array chassis enter an errored state. Because metadevice state database replica recovery is not automatic, it is safest to perform the recovery immediately after the SPARCstorage Array returns to service. Otherwise, a new failure can cause a majority of replicas to be out of service and cause a kernel panic. This is the expected behavior of Solstice DiskSuite when too few replicas are available.

   While these errored replicas will be reclaimed at the next takeover (`haswitch(1M)` or `reboot(1M)`), it is best to return them to service manually by first deleting them, then adding them back.

   ---

   **Note -** Make sure that you add back the same number of replicas that were deleted on each slice. You can delete multiple replicas with a single `metadb(1M)` command. If you need multiple copies of replicas on one slice, you must add them in one invocation of `metadb(1M)` using the `-c` flag.

   ---

3. **Run the** `metastat(1M)` **command to identify the errored metadevices.**

```
# metastat -s diskset
```

4. **Return errored metadevices to service by using the `metareplace(1M)` command, which will cause a resync of the disks.**

```
# metareplace -s diskset -e mirror component
```

The `-e` option transitions the component (slice) to the `Available` state and performs a resync.

Components that have been replaced by a hot spare should be replaced last by using the `metareplace(1M)` command. If the hot spare is replaced first, it could replace another errored submirror as soon as it becomes available.

You can perform a resync on only one component of a submirror (metadevice) at a time. If all components of a submirror were affected by the power outage, each component must be replaced separately. It takes approximately 10 minutes to resync a 1.05GB disk.

If more than one diskset was affected by the power outage, you can resync each diskset's affected submirrors concurrently. Log into each host separately to recover that host's diskset by running the `metareplace(1M)` command on each.

---

**Note -** Depending on the number of submirrors and the number of components in these submirrors, the resync actions can require a considerable amount of time. A single submirror made up of 30 1.05GB drives might take about five hours to complete. A more manageable configuration made up of five component submirrors might take only 50 minutes to complete.

---

## ▼ How to Recover From Power Loss (VxVM)

Power failures can detach disk drives and cause plexes to become detached, and thus, unavailable. In a mirror, however, the volume remains active, because the remaining plex(es) in the volume are still available. It is possible to reattach the disk drives and recover from this condition without halting nodes in the cluster.

These are the high-level steps to recover from power loss to a SPARCstorage Array in an VxVM configuration:

- Determining the errored plex(es) by using the `vxprint` and `vxdisk` commands
- Fixing the problem that caused the power loss
- Running the `drvconfig` and `disks` commands to create the `/devices` and `/dev` entries

- Scanning the current disk configuration
- Reattaching disks that had transient errors
- Verifying there are no more errors
- (Optional) For shared disk groups, running the vxdg command for each disk that was powered off
- Starting volume recovery

These are the detailed steps to recover from power loss to a SPARCstorage Array in an VxVM configuration.

1. **Run the vxprint command to view the errored plexes.**

   Optionally, specify a disk group with the -g *diskgroup* option.

2. **Run the vxdisk command to identify the errored disks.**

```
# vxdisk list
DEVICE       TYPE      DISK         GROUP        STATUS
...
-            -         c1t5d0       toi          failed was:c1t5d0s2
...
```

3. **Fix the condition that resulted in the problem so that power is restored to all failed disks.**

   Be sure that the disks are spun up before proceeding.

4. **Enter the following commands on all nodes in the cluster.**

   In some cases, the drive(s) must be rediscovered by the node(s).

```
# drvconfig
# disks
```

5. **Enter the following commands on all nodes in the cluster.**

   VxVM must scan the current disk configuration again.

```
# vxdctl enable
# vxdisk -a online
```

6. **Enter the following command on all nodes in the cluster.**

   ---

   **Note -** If you are using VxVM cluster feature (used with Oracle Parallel Server), enter the command first on the master node, then on the slave nodes.

   ---

   This will reattach disks that had transitory failures.

```
# vxreattach
```

7. **Verify the output of the `vxdisk` command to see if there are any more errors.**

```
# vxdisk list
```

   If there are still errors, rerun the `vxreattach` command as described in Step 6 on page 173.

8. **VxVM cluster feature (OPS) only: If you have shared disk groups, and if media was replaced from the master node, repeat the following command for each disk that has been disconnected.**

   The physical disk and the volume manager access name for that disk must be reconnected.

```
# vxdg -g disk-group-name -k adddisk medianame=accessname
```

   The values for `medianame` and `accessname` appear at the end of the `vxdisk list` command output.

   For example:

```
# vxdg -g toi -k adddisk c1t5d0=c1t5d0s2
# vxdg -g toi -k adddisk c1t5d1=c1t5d1s2
# vxdg -g toi -k adddisk c1t5d2=c1t5d2s2
# vxdg -g toi -k adddisk c1t5d3=c1t5d3s2
# vxdg -g toi -k adddisk c1t5d4=c1t5d4s2
```

   You can also use the `vxdiskadm` command, or the graphical user interface, to reconnect the disks.

9. **From the node, or from the master node for shared disk groups, start volume recovery.**

```
# vxrecover -bv [-g diskgroup]
```

10. **(Optional) Run the** `vxprint -g` **command to view the changes.**

# Repairing a Lost SPARCstorage Array Connection

When a connection from a SPARCstorage Array to one of the hosts fails, the failure is probably due to a fiber optic cable, an SBus FC/S card, or an FC/OM module.

The host on which the failure occurred will begin generating errors when the failure is discovered. Later accesses to the SPARCstorage Array will generate additional errors. The host behaves as though power had been lost to the SPARCstorage Array.

I/O operations from the other nodes in the cluster are unaffected by this type of failure.

To diagnose the failure, inspect the SPARCstorage Array's display. The display will show whether the A or B connection has been lost. Use the procedures for testing the FC/S card and FC/OM modules in the service manual for your Sun Cluster node to determine which component failed. For hardware debugging, free up one Sun Cluster node and the SPARCstorage Array that appears to be down.

## ▼ How to Repair a Lost SPARCstorage Array Connection

1. **Prepare the Sun Cluster system for component replacement.**

   Depending on the cause of the connection loss, prepare the Sun Cluster system with one of the following procedures.

   - If the failed component is an FC/S card or the FC/OM module for an FC/S card, see Chapter 7, to prepare the Sun Cluster node for power down.
   - If the problem is a bad fiber optic cable, the volume management software will have detected the problem and prepared the system for cable replacement.
   - If the SPARCstorage Array FC/OM module has failed, use either the procedure "How to Take a SPARCstorage Array Tray Out of Service (Solstice DiskSuite)"

on page 177 or “How to Take a SPARCstorage Array Tray Out of Service (VxVM)” on page 179, on each SPARCstorage Array tray to prepare the entire SPARCstorage Array.

2. **Replace the failed component.**

   If the fiber optic cable, SBus FC/S card, or an FC/OM module fails, refer to the service manual for your Sun Cluster node for detailed instructions on replacing them.

3. **Recover from volume management software errors.**

   Use the procedures described in “Recovering From Power Loss” on page 169.

# Adding a SPARCstorage Array

You can add SPARCstorage Arrays to a Sun Cluster configuration at any time.

You must review the disk group configuration in your cluster before adding a SPARCstorage Array. To determine the impact of the SPARCstorage Array on your disk group configuration, refer to the configuration planning information in the *Sun Cluster 2.2 Software Installation Guide*.

## ▼ How to Add a SPARCstorage Array

1. **Shut down the cluster node that is to receive the new SPARCstorage Array.**

   Use the procedure “How to Stop Sun Cluster on a Cluster Node” on page 85 to shut down the node.

2. **Install the Fibre Channel SBus card (FC/S) in the node.**

   Use the instructions in the hardware service manual for your Sun Cluster node to install the FC/S card.

   **Note -** Install the FC/S card in the first available empty SBus slot, following all other cards in the node. This will ensure the controller numbering will be preserved if the Solaris operating environment is reinstalled. Refer to “Instance Names and Numbering” on page 24, for more information.

3. **Connect the cables to the SPARCstorage Array and FC/S card.**

Use the instructions in the hardware service manual for your Sun Cluster node.

4. **Perform a reconfiguration reboot of the node.**

```
ok boot -r
```

5. **Run the** haswitch(1M) **command to switch ownership of all logical hosts that can be mastered to the rebooted node.**

```
phys-hahost1# haswitch phys-hahost2 hahost1 hahost2
```

6. **Repeat Step 1 on page 175 through Step 4 on page 176 on the other nodes that are connected to this SPARCstorage Array.**

7. **Switch ownership of the logical hosts back to the appropriate default master, if necessary.**

```
phys-hahost1# haswitch phys-hahost2 hahost2
```

8. **Add the disks in the SPARCstorage Arrays to the selected disk group(s).**

   Use the instructions in your volume manager documentation to add the disks to the selected disk group(s). Also, refer the *Sun Cluster 2.2 Software Installation Guide* for information on Solstice DiskSuite and VxVM.

9. **(Solstice DiskSuite configurations only) After adding the disks to the diskset by using the** metaset(1M) **command, run the** scadmin(1M) **command to reserve and enable failfast on the specified disks.**

```
phys-hahost1# scadmin reserve cNtXdYsZ
```

# Administering SPARCstorage Array Trays

This section describes procedures for administering SPARCstorage Array trays. Use the procedures described in your node hardware manual to identify the tray associated with the failed component.

To guard against data loss and a failure that might require you to replace the entire SPARCstorage Array chassis, set up all mirrors so that a single chassis contains only one submirror.

---

**Note -** There are several different SPARCstorage Array models supported by Sun Cluster. The procedures in this section are only applicable to the SPARCstorage Array 100 series.

---

## How to Take a SPARCstorage Array Tray Out of Service (Solstice DiskSuite)

Before removing a SPARCstorage Array tray, you must halt all I/O and spin down all drives in the tray. The drives automatically spin up if I/O requests are made, so it is necessary to stop all I/O before the drives are spun down.

These are the high-level steps to take a SPARCstorage Array tray out of service in a Solstice DiskSuite configuration:

- Switching logical hosts to one cluster node
- Stopping I/O to the affected tray
- Identifying any replicas, hot spares, and submirrors on the affected tray
- Flushing NVRAM data, if appropriate
- Spinning down and removing the tray

If the entire SPARCstorage Array is being serviced, you must perform these steps on each tray.

These are the detailed steps to take a SPARCstorage Array tray out of service in a Solstice DiskSuite configuration.

1. **Switch ownership of the affected logical hosts to other nodes by using the** `haswitch(1M)` **command.**

```
phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
```

The SPARCstorage Array tray to be removed might contain disks included in more than one logical host. If this is the case, switch ownership of all logical hosts with disks using this tray to another node in the cluster. The `luxadm(1M)` command will be used later to spin down the disks. In this example, the `haswitch(1M)` command switched the logical hosts to `phys-hahost1`, enabling `phys-hahost2` to perform the administrative functions.

2. **Use the** `metastat(1M)` **command on all affected logical hosts to identify all submirrors containing slices on the tray to be removed.**

```
phys-hahost1# metastat -s disksetname
```

3. **Stop I/O to the submirrors whose components (slices) are on the affected tray.**

   Use the `metaoffline(1M)` command for this step. This takes the submirror offline. You can use the `metadetach(1M)` command to stop the I/O, but the resync cost is greater.

   When the submirrors on a tray are taken offline, the corresponding mirrors provide only one-way mirroring (that is, there will be no data redundancy). (A three-way mirror does not have this problem.) When the mirror is brought back online, an automatic resync occurs.

   With all affected submirrors offline, I/O to the tray is stopped.

4. **Use the** `metadb(1M)` **command to identify any replicas on the tray.**

   Save the `metadb(1M)` output to use when you replace the tray.

5. **Use the** `metahs(1M)` **command to identify any available hot spare devices and associated submirrors.**

   Save the `metahs(1M)` output to use when you replace the tray.

6. **If NVRAM is enabled, flush the NVRAM data on the appropriate controller, tray, or disk(s).**

```
phys-hahost1# luxadm sync_cache pathname
```

   A confirmation appears, indicating that NVRAM data has been flushed. See “Flushing and Purging NVRAM” on page 219, for details on flushing NVRAM data.

7. **Spin down the tray using the** `luxadm stop` **command.**

   When the tray lock light is out, remove the tray and perform the required service.

```
phys-hahost1# luxadm stop c1
```

## How to Take a SPARCstorage Array Tray Out of Service (VxVM)

Before removing a SPARCstorage Array tray, you must halt all I/O and spin down all drives in the tray. The drives automatically spin up if I/O requests are made, so it is necessary to stop all I/O before the drives are spun down.

These are the high-level steps to take a SPARCstorage Array tray out of service in an VxVM configuration:

- Switching logical hosts to one cluster node
- Identifying VxVM objects on the affected tray
- Stopping I/O to the affected tray
- Flushing NVRAM data, if appropriate
- Spinning down and removing the tray

If the entire SPARCstorage Array is being serviced, you must perform these steps on each tray.

These are the detailed steps to take a SPARCstorage Array tray out of service in an VxVM configuration.

1. **Switch ownership of the affected logical hosts to other nodes by using the `haswitch(1M)` command.**

   ```
   phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
   ```

   The SPARCstorage Array tray to be removed might contain disks included in more than one logical host. If this is the case, switch ownership of all logical hosts with disks using this tray to another node in the cluster. The `luxadm(1M)` command will be used later to spin down the disks. In this example, the `haswitch(1M)` command switched the logical hosts to `phys-hahost1`, enabling `phys-hahost1` to perform the administrative functions.

2. **Identify all the volumes and corresponding plexes on the disks in the tray which is being taken out of service.**

   a. **From the physical device address `c`*N*`t`*N*`d`*N*, obtain the controller number and the target number.**

      For example, if the device address is c3t2d0, the controller number is 3 and the target is 2.

   b. **Identify VxVM devices on the affected tray from a `vxdisk list` output.**

      If the target is 0 or 1, identify all devices with physical addresses beginning with c*N*t0 and c*N*t1. If the target is 2 or 3, identify all devices with physical addresses beginning with c*N*t2 and c*N*t3. If the target is 4 or 5, identify all devices with physical addresses beginning with c*N*t4 and c*N*t5. Here is an example of how `vxdisk` can be used to obtain the information.

```
# vxdisk -g diskgroup -q list | egrep c3t2\|c3t3 | nawk '{print $3}'
```

c. **Identify all plexes on the above devices by using the appropriate version (`csh`, `ksh`, or Bourne shell) of the following command.**

```
PLLIST=`vxprint -ptq -g diskgroup -e '(aslist.sd_dm_name in   (''c3t2d0'',''c3t3d0'',''c3t3d1'')) && (pl_kstate=ENABLED)'
```

For `csh`, the syntax is `set PLLIST ....` For `ksh`, the syntax is `export PLLIST= ....` The Bourne shell requires the command `export PLLIST` after the variable is set.

3. **After you have set the variable, stop I/O to the volumes whose components (subdisks) are on the tray.**

Make sure all volumes associated with that tray are detached (mirrored or RAID5 configurations) or stopped (simple plexes). Issue the following command to detach a mirrored plex.

```
# vxplex det ${PLLIST}
```

An alternate command for detaching each plex in a tray is:

```
# vxplex -g diskgroup -v volume det plex
```

To stop I/O to simple plexes, unmount any file systems or stop database access.

---

**Note -** Mirrored volumes will still be active because the other half of the mirror is still available.

---

4. **If NVRAM is enabled, flush the NVRAM data on the appropriate controller, tray, or disk(s). Otherwise, skip to Step 5 on page 181.**

```
# luxadm sync_cache pathname
```

A confirmation appears, indicating that NVRAM data has been flushed. See “Flushing and Purging NVRAM” on page 219, for details on flushing NVRAM data.

5. **To remove the tray, use the `luxadm stop` command to spin it down.**

   When the tray lock light is out, remove the tray and perform the required service.

   ```
   # luxadm stop c1
   ```

## How to Return a SPARCstorage Array Tray to Service (Solstice DiskSuite)

These are the high-level steps to return a SPARCstorage Array tray back to service in a Solstice DiskSuite configuration.

- Spinning up the drives
- Restoring all replicas, submirrors, and hot spares
- Switching each logical host back to its default master

If the entire SPARCstorage Array has been serviced, you must perform these steps on each tray.

These are the detailed steps to return a SPARCstorage Array tray back to service in a Solstice DiskSuite configuration.

1. **If the SPARCstorage Array was removed, spin up the drives in the SPARCstorage Array tray. Otherwise, skip to Step 3 on page 182.**

   When you have completed work on a SPARCstorage Array tray, replace the tray in the chassis. The disks will spin up automatically. However, if the disks fail to spin up, run the `luxadm(1M) start` command to manually spin up the entire tray. There is a short delay (several seconds) between invocation of the command and spin-up of drives in the SPARCstorage Array. In this example, `c1` is the controller ID:

   ```
   phys-hahost1# luxadm start c1
   ```

2. **Add all metadevice state database replicas that were deleted from disks on this tray.**

   Use the information saved from Step 4 on page 178 in the procedure “How to Take a SPARCstorage Array Tray Out of Service (Solstice DiskSuite)” on page 177 to restore the metadevice state database replicas.

   ```
   phys-hahost1# metadb -s hahost1 -a deleted-replicas
   ```

   To add multiple replicas to the same slice, use the `-c` option.

3. **After the disks spin up, place online all the submirrors that were taken offline.**

   Use the `metaonline(1M)` command appropriate for the disks in this tray.

   ```
   phys-hahost1# metaonline -s hahost1 d15 d35
   phys-hahost1# metaonline -s hahost1 d24 d54
   ...
   ```

   When the `metaonline(1M)` command is run, an optimized resync operation automatically brings the submirrors up-to-date. The optimized resync copies only those regions of the disk that were modified while the submirror was offline. This is typically a very small fraction of the submirror capacity.

   Run `metaonline(1M)` as many times as necessary to bring back online all of the submirrors.

   ---

   **Note -** If you used the `metadetach(1M)` command to detach the submirror rather than `metaoffline(1M)`, you must synchronize the entire submirror using the `metattach(1M)` command. This typically takes about 10 minutes per Gigabyte of data.

   ---

4. **Add back all hot spares that were deleted when the SPARCstorage Array was taken out of service.**

   Use the `metahs(1M)` command as appropriate for your hot spare configuration. Use the information saved from Step 5 on page 178 in the procedure “How to Take a SPARCstorage Array Tray Out of Service (Solstice DiskSuite)” on page 177 to replace your hot spares.

   ```
   phys-hahost1# metahs -s hahost1 -a hotsparepool cNtXdYsZ
   ```

5. **Switch each logical host back to its default master, if necessary.**

   ```
   phys-hahost1# haswitch phys-hahost2 hahost2
   ```

## How to Return a SPARCstorage Array Tray to Service (VxVM)

These are the high-level steps to return a SPARCstorage Array tray back to service in a VxVM configuration:

- Spinning up the drives
- Restoring VxVM objects
- Switching each logical host back to its default master

If the entire SPARCstorage Array has been serviced, you must perform these steps on each tray.

These are the detailed steps to return a SPARCstorage Array tray back to service in an VxVM configuration.

1. **If the SPARCstorage Array was removed, spin up the drives in the SPARCstorage Array tray. Otherwise, skip to Step 2 on page 183.**

   When you have completed work on a SPARCstorage Array tray, replace the tray in the chassis. The disks will spin up automatically. However, if the disks fail to spin up, run the `luxadm(1M) start` command to manually spin up the entire tray. There is a short delay (several seconds) between invocation of the command and spin-up of drives in the SPARCstorage Array. In this example, `c1` is the controller ID.

   ```
   phys-hahost1# luxadm start c1
   ```

2. **After the disks spin up, monitor the volume management recovery.**

   Previously affected volumes that are on the tray should begin to come back online, and the rebuilding of data should start automatically within a few minutes. If necessary, use the `vxreattach` and `vxrecover` commands to reattach disks and recover from error. Refer to the respective man pages for more information.

   ---

   **Note -** DRL subdisks that were detached must be manually reattached.

   ---

3. **Switch each logical host back to its default master, if necessary.**

   ```
   phys-hahost1# haswitch phys-hahost2 hahost2
   ```

# Replacing a SPARCstorage Array Controller and Changing the World Wide Name

The SPARCstorage Array controller has a unique identifier known as the World Wide Name (WWN) that identifies the controller to Solaris software. Therefore, when SPARCstorage Array failures make it necessary to replace the controller or the entire chassis containing the controller, special procedures apply.

The WWN is like the host ID stored in the host IDPROM of a SPARC machine. The last four digits of the SPARCstorage Array WWN are displayed on the LCD panel of the chassis. The WWN is part of the `/devices` path associated with the SPARCstorage Array and its component drives.

If you must replace the SPARCstorage Array controller or the entire chassis, the Sun Cluster nodes will discover the new WWN when they are rebooted. To avoid confusion of the upper layers of Sun Cluster software by the new WWN, change the WWN of the new controller to be the WWN of the old controller. (This is similar to swapping the IDPROM when replacing a system board in a SPARC machine.)

Consider the following situations when deciding which WWN replacement procedure to use:

- The procedure described in “How to Change a SPARCstorage Array World Wide Name Using a Maintenance System” on page 185 makes use of a separate maintenance system that enables the controller to be changed without stopping cluster nodes.
- If the SPARCstorage Array has not entirely failed or is being swapped for some other reason, prepare for the swap by performing the steps described in “Administering SPARCstorage Array Trays” on page 177, for each tray in the SPARCstorage Array. Then use the procedure described in “How to Change a SPARCstorage Array World Wide Name” on page 189.
- If the SPARCstorage Array controller has failed entirely, your volume management software has already prepared for the swap. In this case, you can use the procedure described in “How to Change a SPARCstorage Array World Wide Name Using a Maintenance System” on page 185.

## ▼ How to Change a SPARCstorage Array World Wide Name Using a Maintenance System

This procedure describes how to change a SPARCstorage Array controller and replace its WWN with the WWN of the failed controller. This procedure enables you to replace a SPARCstorage Array controller without taking down any nodes in the cluster.

This procedure makes use of a “maintenance system,” which can be any Sun Microsystems architecture capable of supporting a SPARCstorage Array. The presence of a maintenance system enables you to complete this procedure without taking down any nodes in the cluster.

This system should be loaded with the same version of the Solaris operating environment as the cluster nodes, and should have all applicable patches. It also should have available a CD-ROM drive, a Fibre Channel SBus Card (FC/S), and a Fibre Channel Optical Module (FC/OM). The system should have the proper FCODE and hardware revisions. Alternately, you can boot the maintenance system over the net.

---

**Note -** If a “maintenance system” is not available, use one of the cluster nodes for this purpose by following the steps in this procedure.

---

These are the high-level steps to change a SPARCstorage Array World Wide Name (WWN) using a maintenance system:

- (Optional) If the controller is the quorum device, using the `scconf(1M)` command to select a new quorum device
- Obtaining the WWN of the previous array
- Detaching the optical cables and replacing the controller or array
- Attaching the optical cable from the maintenance system to the new controller
- Booting the maintenance system with “mini-unix” from a Solaris CD-ROM
- Downloading the original WWN
- Resetting the SSA
- Shutting down the maintenance system
- Attaching the SSA controller to the cluster nodes
- Checking the new controller’s firmware level from the cluster node
- (Optional) If necessary, upgrading the new controller’s firmware from the cluster node
- Bringing the SSA tray online and performing volume management recovery

These are the detailed steps to change a SPARCstorage Array World Wide Name by using a maintenance system.

1. **If the failed SPARCstorage Array controller is the quorum controller, select a new quorum controller by using the `scconf(1M)` command.**

   Refer to the `scconf(1M)` man page for more information.

2. **Determine the WWN of the broken SPARCstorage Array.**

   If the SPARCstorage Array is powered down, use the following instructions to obtain the WWN.

   The WWN is composed of 12 hexadecimal digits. The digits are shown as part of the device path component. They are the last 12 digits following the characters `pln@a0`, excluding the comma. Use the `ls(1)` command on a cluster node connected to the SSA to identify the current WWN.

   ```
   # ls -l /dev/rdsk/cNt0d0s0
   ...SUNW,pln@a0000000,7412bf ...
   ```

   In this example, the WWN for the SPARCstorage Array being replaced is `0000007412bf`. The variable *N* in the device name represents the controller number for the broken SPARCstorage Array. The string "`t0d0s0`" is just an example. Use a device name that you know exists on the SPARCstorage Array, or use `/dev/rdsk/c`*N*`*` to match all devices.

   If the SPARCstorage Array is up and running, you can obtain the WWN by using the `luxadm(1M)` command.

   When you run `luxadm(1M)` with the `display` option and specify a controller, all the information about the SPARCstorage Array is displayed. The serial number reported by `luxadm(1M)` is the WWN.

   ```
   # /usr/sbin/luxadm display cN
   ```

3. **Detach the optical cable from the faulty SPARCstorage Array Controller.**

4. **Replace the faulty controller.**

   Use the instructions in your SPARCstorage Array service manual to perform this step.

   If the SPARCstorage Array has not failed entirely or is being swapped for a reason other than controller failure, prepare for the swap by performing the steps described in "Administering SPARCstorage Array Trays" on page 177, for each tray in the SPARCstorage Array.

   If the SPARCstorage Array controller has failed entirely, your volume manager has already prepared for the swap.

5. **Attach the optical cable from the maintenance system to the new controller.**

6. **Enter the OpenBoot PROM on the maintenance system and boot it with “mini-unix.”**

   Do this from the distribution CD (or its network equivalent) to put the maintenance system into single-user mode and to obtain an in-memory version of the device structure that contains the new SPARCstorage Array WWN.

```
<#0> ok boot cdrom -s

or

<#0> ok boot netqe1 -s
```

   Use “mini-unix” to avoid making any permanent device information changes.

7. **Run the** `luxadm download` **command to set the WWN.**

```
# /usr/sbin/luxadm -s -w WWN download cN
```

   *WWN* is the 12-digit WWN of the replaced controller and *N* is the controller number from `c`*N*`t`*X*`d`*X* in the device name. You should have obtained the WWN in Step 2 on page 186.

   **Note -** The leading zeros must be entered as part of the WWN to make a total of 12 digits.

**Caution -** Do not interrupt the download process. Wait for the shell prompt after completion of the `luxadm(1M)`command.

8. **After the prompt is redisplayed, reset the SSA.**

   The new address should appear in the window on the SPARCstorage Array.

9. **Shut down the maintenance system.**

10. **Reattach the SPARCstorage Array controller to the cluster nodes.**

11. **Verify the SPARCstorage Array firmware level from the cluster node.**

Use the luxadm(1M) command to determine the current version of the firmware. Specify the controller number (*N* in the example) to the luxadm(1M) command.

```
# /usr/sbin/luxadm display cN
```

---

**Note -** If the Solaris system detects an old version of firmware on your system, it displays a message on the console and in /var/adm/messages similar to the following: NOTICE: pln0: Old SSA firmware has been detected (Ver:3.11) : Expected (Ver:3.12) - Please upgrade

---

12. **(Optional) To upgrade the controller's firmware, follow these steps.**

    a. **Download the proper firmware. Refer to the** README **file in the firmware patch for details.**

```
# /usr/sbin/ssaadm download -f path/ssafirmware cN
```

where *path* is the path to the directory where the firmware is stored and *N* is the controller number. For example:

```
# /usr/sbin/ssaadm download -f /usr/lib/firmware/ssa/ssafirmware cN
```

b. **Reset the SPARCstorage Array by pressing the SYS OK button on the unit.**

There will be a short delay while the unit reboots.

c. **Verify the firmware level again (using Step 11 on page 187). If the firmware level or WWN is still incorrect, repeat Step 12 on page 188 using a different controller.**

13. **Begin volume manager recovery.**

Refer to "Administering SPARCstorage Array Trays" on page 177. Wait until the SPARCstorage Array is online for all nodes, and all nodes can see all the disks.

## ▼ How to Change a SPARCstorage Array World Wide Name

---

**Caution -** This procedure will not work if the root disk is encapsulated by VxVM, or if the boot disk of one of the nodes is on this SPARCstorage Array. For those situations, use the procedure "How to Change a SPARCstorage Array World Wide Name Using a Maintenance System" on page 185.

---

**Note -** If a quorum controller fails, you must select a new quorum controller before shutting down a node.

---

These are the high-level steps to change a SPARCstorage Array World Wide Name:

- (Optional) If the controller is the quorum device, using the `scconf(1M)` command to select a new quorum device
- Switching ownership of logical hosts away from the node on which the repair procedure will be performed or controller being replaced
- Obtaining the WWN of the previous array
- Replacing the controller or array
- Stopping Sun Cluster software and halting the node that does not own the disks
- With "mini-unix," rebooting the node that does not own the disks
- Determining the controller number for the new array
- Setting the new WWN and resetting the array
- Rebooting the other cluster nodes, if necessary
- Performing volume management recovery

These are the detailed steps to change a SPARCstorage Array World Wide Name.

1. **If the failed SPARCstorage Array controller is the quorum controller, select a new quorum controller by using the `scconf(1M)` command.**

   Refer to the `scconf(1M)` man page for more information.

2. **On the cluster node that is connected to the SSA being repaired, stop the Sun Cluster software and halt the system.**

   Use the `scadmin(1M)` command to switch ownership of all logical hosts to the other nodes in the cluster, and to stop Sun Cluster. Then run the `halt(1M)` command to stop the machine.

   In this example, `phys-hahost2` is the node from which the repair procedure is performed.

```
phys-hahost2# scadmin stopnode
...
phys-hahost2# halt
```

3. **Determine the WWN of the broken SPARCstorage Array.**

   If the SPARCstorage Array is powered down, use the following instructions to obtain the WWN.

   The WWN is composed of 12 hexadecimal digits. The digits are shown as part of the device path component containing the characters `pln@a0`. They are the last 12 digits following the characters `pln@a0`, excluding the comma. Use the `ls(1)` command on a cluster node connected to the SSA to identify the current WWN.

   ```
   phys-hahost1# ls -l /dev/rdsk/cNt0d0s0
   ...SUNW,pln@a0000000,7412bf ...
   ```

   In this example, the WWN for the SPARCstorage Array being replaced is 0000007412bf. The variable *N* in the device name represents the controller number for the broken SPARCstorage Array. The string `t0d0s0` is just an example. Use a device name that you know exists on the SPARCstorage Array, or use `/dev/rdsk/cN*` to match all devices.

   If the SPARCstorage Array is up and running, you can obtain the WWN by using the `luxadm(1M)` command.

   When you run `luxadm(1M)` with the `display` option and specify a controller, all the information about the SPARCstorage Array is displayed. The serial number reported by `luxadm(1M)` is the WWN.

   ```
   phys-hahost1# /usr/sbin/luxadm display cN
   ```

4. **Replace the controller or SPARCstorage Array.**

   Use the instructions in your SPARCstorage Array service manual to perform this step.

   If the SPARCstorage Array has not failed completely or is being swapped for a reason other than controller failure, prepare for the swap by performing the steps described in "Administering SPARCstorage Array Trays" on page 177, for each tray in the SPARCstorage Array.

   If the SPARCstorage Array controller has failed completely, your volume manager has already prepared for the swap.

5. **Enter the OpenBoot PROM on the halted node and boot it with “mini-unix.”**

   Do this from the distribution CD (or net equivalent) to put the host into single-user mode and to obtain an in-memory version of the device structure that contains the new SPARCstorage Array WWN.

   ```
   <#0> ok boot cdrom -s
   or

   <#0> ok boot netqe1 -s
   ```

   Use “mini-unix” to avoid making any permanent device information changes to the cluster node.

6. **Determine the controller number for the new SPARCstorage Array.**

   Use the `ls(1)` command and the four digits displayed on the LCD display of the new SPARCstorage Array to identify the controller number.

   In this example, the four digits shown on the LCD display are `143b`. Note that the device name `c*t0d0s0` uses pattern matching for the controller number but specifies a slice that is known to exist. This reduces the number of lines generated in the output.

   ```
   # ls -l /dev/rdsk/c*t0d0s0 | grep -i 143b
   lrwxrwxrwx   1 root     root          98 Mar 14 13:38 /dev/rdsk/c3t0d0s0 -> ../../devices/iommu@f
   sbus@f,e0001000/SUNW,soc@3,0/SUNW,pln@a0000000,74143b/ssd@0,0:a,raw
   ```

   In this example, 3 (from `/dev/rdsk/c3...`) is the controller number of the new SPARCstorage Array under “mini-unix”.

---

**Note -** The hex digits in the LCD display are in mixed case—letters `A`, `C`, `E`, and `F` are in upper case, and letters `b` and `d` are in lower case. The example uses `grep -i` to ignore case in the comparison.

---

7. **Run the `luxadm download` command to set the WWN.**

   Use the controller number determined in Step 6 on page 191. For example, the following command would change the WWN from the current value to the value determined in Step 3 on page 190, `0000007412bf`. The SPARCstorage Array controller is Controller 3.

   ```
   phys-hahost2# /usr/sbin/luxadm download -w 0000007412bf c3
   ```

**Note -** The leading zeros must be entered as part of the WWN to make a total of 12 digits.

**Caution -** Do not interrupt the download process. Wait for the shell prompt after completion of the `luxadm(1M)` command.

8. **Reset the SPARCstorage Array by pressing the SYS OK button on the unit.**

   There will be a short delay while the unit reboots and begins communicating with the Sun Cluster nodes.

9. **Abort "mini-unix" and boot the host normally.**

   Send a break to the console, and boot the machine.

10. **Verify the SPARCstorage Array firmware level from the cluster node.**

    Use the `luxadm(1M)` command to determine the current version of the firmware. Specify the controller number (*N* in the example) to the `luxadm(1M)` command.

```
phys-hahost2# /usr/sbin/luxadm display cN
```

**Note -** If the Solaris system detects an old version of firmware on your system, it displays a message on the console and in `/var/adm/messages` similar to the following: `NOTICE: pln0: Old SSA firmware has been detected (Ver:3.11) : Expected (Ver:3.12) - Please upgrade`

11. **(Optional) To upgrade the controller's firmware, follow these steps.**

    a. **Download the proper firmware. Refer to the `README` file in the firmware patch for details.**

```
# /usr/sbin/ssaadm download -f path/ssafirmware cN
```

where *path* is the path to the directory where the firmware is stored and *N* is the controller number. For example:

```
# /usr/sbin/ssaadm download -f /usr/lib/firmware/ssa/ssafirmware cN
```

b. **Reset the SPARCstorage Array using the SYS OK button on the unit.**

   There will be a short delay while the unit reboots.

c. **Re-verify the firmware level (see Step 10 on page 192). If either the firmware level or WWN are still incorrect, then repeat Step 11 on page 192 using a different controller.**

12. **Start the node.**

```
phys-hahost2# scadmin startnode
```

13. **Switch back the logical hosts to the default master, if necessary.**

14. **Complete the replacement by restoring the volume manager components onto the repaired SPARCstorage Array.**

    This procedure is described in “Administering SPARCstorage Array Trays” on page 177.

15. **Reboot the other nodes in the cluster, if necessary.**

    You might need to reboot the other cluster nodes, if they are unable to recognize all disks in the SPARCstorage Array following the replacement. If this is the case, use the `scadmin stopnode` command to stop Sun Cluster activity, then reboot. After the reboot, if necessary, switch the logical hosts back to their default masters. See the `scadmin(1M)` man page for more information.

# Administering SPARCstorage Array Disks

As part of standard Sun Cluster administration, you should monitor the status of the configuration. See Chapter 2, for information about monitoring methods. During the monitoring process you might discover problems with multihost disks. The following sections provide instructions for correcting these problems.

Sun Cluster supports these SSA disk types:

- 100 series
- 200 series with differential SCSI tray
- 200 series with RSM (214 RSM)

Depending on which type you have and the electrical and mechanical characteristics of this disk enclosure, adding a disk might require you to prepare all disks connected to a particular controller, all disks in a particular array tray, or only the disk being added. For example, in the SPARCstorage Array 200 series with the differential SCSI tray, you must prepare the array controller and the disk enclosure. In the SPARCstorage Array 200 series with RSM (214 RSM), you need to prepare only the new disk. In the SPARCstorage Array 110, you must prepare a single tray.

If you have a SPARCstorage Array 100 series array, follow the steps as documented. If you have a SPARCstorage Array 200 series array with differential SCSI tray, you must bring down all disks attached to the array controller that will connect to the new disk. This means you repeat all of the tray-specific steps for all disk enclosures attached to the array controller that will connect to the new disk. If you have a SPARCstorage Array 214 RSM, you need not perform any of the tray-specific steps, since individual disk drives can be installed without affecting other disks.

Refer to the hardware service manual for your multihost disk expansion unit for a description of your disk enclosure.

## Adding a SPARCstorage Array Disk

Depending upon the disk enclosure, adding SPARCstorage Array (SSA) multihost disks might involve taking off line all volume manager objects in the affected disk tray or disk enclosure. Additionally, the disk tray or disk enclosure might contain disks from more than one disk group, requiring that a single node own all of the affected disk groups.

## ▼ How to Add a SPARCstorage Array Disk (Solstice DiskSuite)

These are the high-level steps to add a multihost disk in a Solstice DiskSuite configuration:

- Switching logical hosts to one cluster node
- Identifying the controller for this new disk, and locating an empty slot in the tray or enclosure
- For Model 100 series SPARCstorage Arrays, preparing the disk enclosure for removal of a disk tray
- For Model 200 series SPARCstorage Arrays with wide differential SCSI disk trays, powering down the controller and all attached disks
- Deleting all hot spares from the affected drives
- Deleting all metadevice state databases from the affected drives
- Taking offline all metadevices containing affected drives
- Spinning down all affected drives

- Adding the new disk
- Returning the affected drives to service
  - Spinning up all drives
  - Bringing back online all affected metadevices
  - Adding back all deleted hot spares
  - Recreating all deleted metadevices
- Performing the administrative actions to prepare the disk for use by Sun Cluster
  - Creating the `/devices` special files and `/dev/dsk` and `/dev/rdsk` links
  - Running the `scdidadm -r` command
  - Adding the disk to the diskset
  - Formatting and partitioning the disk, if necessary
  - Performing the volume manager-related administrative tasks

These are the detailed steps to add a new multihost disk to a Solstice DiskSuite configuration.

1. **Switch ownership of the logical host that will include the new disk to other nodes in the cluster.**

   Switch over any logical hosts with disks in the tray you are removing.

   ```
   phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
   ```

2. **Determine the controller number of the tray to which the disk will be added.**

   SPARCstorage Arrays are assigned World Wide Names (WWN). The WWN on the front of the SPARCstorage Array also appears as part of the `/devices` entry, which is linked by pointer to the `/dev` entry containing the controller number. For example:

   ```
   phys-hahost1# ls -l /dev/rdsk | grep -i WWN | tail -1
   ```

   If the WWN on the front of the SPARCstorage Array is `36cc`, the following output will display, and the controller number would be `c2`:

   ```
   phys-hahost1# ls -l /dev/rdsk | grep -i 36cc | tail -1
   lrwxrwxrwx  1 root   root       94 Jun 25 22:39 c2t5d2s7 -> ../../devices/io-unit@f,e1200000/sbi@
   SUNW,soc@3,0/SUNW,pln@a0000800,201836cc/ssd@5,2:h,raw
   ```

3. **Use the `luxadm(1M)` command with the `display` option to view the empty slots.**

```
phys-hahost1# luxadm display c2

                     SPARCstorage Array Configuration
...
                          DEVICE STATUS
      TRAY 1                 TRAY 2                 TRAY 3
slot
1     Drive: 0,0             Drive: 2,0             Drive: 4,0
2     Drive: 0,1             Drive: 2,1             Drive: 4,1
3     NO SELECT              NO SELECT              NO SELECT
4     NO SELECT              NO SELECT              NO SELECT
5     NO SELECT              NO SELECT              NO SELECT
6     Drive: 1,0             Drive: 3,0             Drive: 5,0
7     Drive: 1,1             NO SELECT              NO SELECT
8     NO SELECT              NO SELECT              NO SELECT
9     NO SELECT              NO SELECT              NO SELECT
10    NO SELECT              NO SELECT              NO SELECT
...
```

The empty slots are shown with a `NO SELECT` status. The output shown here is from a SPARCstorage Array 110; your display will be slightly different if you are using a different series SPARCstorage Array.

Determine the tray to which you will add the new disk. If you can add the disk without affecting other drives, such as in the SPARCstorage Array 214 RSM, skip to Step 11 on page 198.

In the remainder of the procedure, Tray 2 is used as an example. The slot selected for the new disk is Tray 2 Slot 7. The new disk will be known as `c2t3d1`.

4. **Locate all hot spares affected by the installation.**

   To determine the status and location of all hot spares, run the `metahs(1M)` command with the `-i` option on each of the logical hosts.

```
phys-hahost1# metahs -s hahost1 -i
...
phys-hahost1# metahs -s hahost2 -i
...
```

---

**Note -** Save a list of the hot spares. The list is used later in this maintenance procedure. Be sure to note the hot spare devices and their hot spare pools.

---

5. **Use the `metahs(1M)` command with the `-d` option to delete all affected hot spares.**

Refer to the man page for details on the `metahs(1M)` command.

```
phys-hahost1# metahs -s hahost1 -d hot-spare-pool components
phys-hahost1# metahs -s hahost2 -d hot-spare-pool components
```

6. **Locate all metadevice state database replicas that are on affected disks.**

   Run the `metadb(1M)` command on each of the logical hosts to locate all metadevice state databases. Direct the output into temporary files.

```
phys-hahost1# metadb -s hahost1 > /usr/tmp/mddb1
phys-hahost1# metadb -s hahost2 > /usr/tmp/mddb2
```

   The output of `metadb(1M)` shows the location of metadevice state database replicas in this disk enclosure. Save this information for the step in which you restore the replicas.

7. **Delete the metadevice state database replicas that are on affected disks.**

   Keep a record of the number and locale of the replicas that you delete. The replicas must be restored in a later step.

```
phys-hahost1# metadb -s hahost1 -d replicas
phys-hahost1# metadb -s hahost2 -d replicas
```

8. **Run the `metastat(1M)` command to determine all the metadevice components on affected disks.**

   Direct the output from `metastat(1M)` to a temporary file so that you can use the information later when deleting and re-adding the metadevices.

```
phys-hahost1# metastat -s hahost1 > /usr/tmp/replicalog1
phys-hahost1# metastat -s hahost2 > /usr/tmp/replicalog2
```

9. **Take offline all submirrors containing affected disks.**

   Use the temporary files to create a script to take offline all affected submirrors in the disk expansion unit. If only a few submirrors exist, run the `metaoffline(1M)` command to take each offline. The following is a sample script.

```
#!/bin/sh
# metaoffline -s <diskset> <mirror> <submirror>

metaoffline -s hahost1 d15 d35
metaoffline -s hahost2 d15 d35
...
```

10. **Spin down the affected disks.**

    Spin down the SPARCstorage Array disks in the tray using the `luxadm(1M)` command.

```
phys-hahost1# luxadm stop -t 2 c2
```

11. **Add the new disk.**

    Use the instructions in your multihost disk expansion unit service manual to perform the hardware procedure of adding the disk. After the addition:

    - If your disk enclosure is a SPARCstorage Array 214 RSM, skip to Step 16 on page 199. (This type of disk can be added without affecting other drives.)
    - For all other SPARCstorage Array types, proceed with Step 12 on page 198.

12. **Make sure all disks in the tray spin up.**

    The disks in the SPARCstorage Array tray should spin up automatically, but if the tray fails to spin up within two minutes, force the action using the following command:

```
phys-hahost1# luxadm start -t 2 c2
```

13. **Bring the submirrors back online.**

    Modify the script that you created in Step 9 on page 198 to bring the submirrors back online.

```
#!/bin/sh
# metaonline -s <diskset> <mirror> <submirror>

metaonline -s hahost1 d15 d35
metaonline -s hahost2 d15 d35
...
```

14. **Restore the hot spares that were deleted in Step 5 on page 196.**

```
phys-hahost1# metahs -s hahost1 -a hot-spare-pool components
phys-hahost1# metahs -s hahost2 -a hot-spare-pool components
```

15. **Restore the original count of metadevice state database replicas to the devices in the tray.**

    The replicas were removed in Step 7 on page 197.

```
phys-hahost1# metadb -s hahost1 -a replicas
phys-hahost1# metadb -s hahost2 -a replicas
```

16. **Run the `drvconfig(1M)` and `disks(1M)` commands to create the new entries in `/devices`, `/dev/dsk`, and `/dev/rdsk` for all new disks.**

```
phys-hahost1# drvconfig
phys-hahost1# disks
```

17. **Switch ownership of the logical host to which this disk will be added to the other node that is connected to the SPARCstorage Array.**

    This assumes a topology in which each disk is connected to two nodes.

```
phys-hahost1# haswitch phys-hahost2 hahost2
```

18. **Run the** `drvconfig(1M)` **and** `disks(1M)` **commands on the cluster node that now owns the diskset to which this disk will be added.**

```
phys-hahost2# drvconfig
phys-hahost2# disks
```

19. **Run the** `scdidadm(1M)` **command to initialize the new disk for use by the DID pseudo driver.**

    You must run `scdidadm(1M)` on Node 0 in the cluster. Refer to the *Sun Cluster 2.2 Software Installation Guide* for details on the DID pseudo driver.

```
phys-hahost2# scdidadm -r
```

20. **Add the disk to a diskset.**

    The command syntax is as follows, where *diskset* is the name of the diskset containing the failed disk, and *drive* is the DID name of the disk in the form `d`*N* (for new installations of Sun Cluster), or `c`*N*`t`*Y*`d`*Z* (for installations that upgraded from HA 1.3):

```
# metaset -s diskset -a drive
```

**Caution -** The `metaset(1M)` command might repartition this disk automatically. See the Solstice DiskSuite documentation for more information.

21. **Use the** `scadmin(1M)` **command to reserve and enable failfast on the specified disk that has just been added to the diskset.**

```
phys-hahost2# scadmin reserve cNtXdYsZ
```

22. **Perform the usual administration actions on the new disk.**

    You can now perform the usual administration steps that bring a new drive into service. These include partitioning the disk, adding it to the configuration as a hot spare, or configuring it as a metadevice. See the Solstice DiskSuite documentation for more information on these tasks.

23. **If necessary, switch logical hosts back to their default masters.**

# How to Add a SPARCstorage Array Disk (VxVM)

These are the high-level steps to add a multihost disk to an VxVM configuration:

- Switching logical hosts to one cluster node
- Identifying the controller for this new disk and locating an empty slot in the tray or enclosure
- For Model 100 series SPARCstorage Arrays, preparing the disk enclosure for removal of a disk tray
- For Model 200 series SPARCstorage Arrays with wide differential SCSI disk trays, powering down the controller and all attached disks
- Identifying VxVM objects on the affected tray
- Stopping I/O to volumes with subdisks on the affected tray
- Adding the new disk
- Returning the affected drives to service
    - Spinning up all drives
    - Bringing back online all affected VxVM objects
- Performing the administrative actions to prepare the disk for use by Sun Cluster
    - Creating the `/devices` special files and `/dev/dsk` and `/dev/rdsk` links
    - Scanning for the new disk
    - Adding the disk to VM control
    - Formatting and partitioning the disk, if necessary
    - Performing the volume manager-related administrative tasks

These are the detailed steps to add a new multihost disk to an VxVM configuration.

1. **Switch ownership of the logical host that will include the new disk to another node in the cluster.**

   Switch over any logical hosts with disks in the tray you are removing.

   ```
   phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
   ```

   ---

   **Note -** In a mirrored configuration, you may not need to switch logical hosts as long as the node is not shut down.

   ---

2. **Determine the controller number of the tray to which the disk will be added.**

   SPARCstorage Arrays are assigned World Wide Names (WWN). The WWN on the front of the SPARCstorage Array also appears as part of the `/devices` entry, which is linked by pointer to the `/dev` entry containing the controller number. For example:

```
phys-hahost1# ls -l /dev/rdsk | grep -i WWN | tail -1
```

If the WWN on the front of the SPARCstorage Array is 36cc, the following output will display and the controller number would be c2:

```
phys-hahost1# ls -l /dev/rdsk | grep -i 36cc | tail -1
lrwxrwxrwx  1 root   root       94 Jun 25 22:39 c2t5d2s7 -> ../../devices/io-unit@f,e1200000/sbi
SUNW,soc@3,0/SUNW,pln@a0000800,201836cc/ssd@5,2:h,raw
phys-hahost1#
```

3. **Use the `luxadm(1M)` command with the `display` option to view the empty slots.**

   If you can add the disk without affecting other drives, skip to Step 11 on page 204.

```
phys-hahost1# luxadm display c2

                     SPARCstorage Array Configuration
...
                          DEVICE STATUS
      TRAY 1                 TRAY 2                 TRAY 3
slot
1     Drive: 0,0             Drive: 2,0             Drive: 4,0
2     Drive: 0,1             Drive: 2,1             Drive: 4,1
3     NO SELECT              NO SELECT              NO SELECT
4     NO SELECT              NO SELECT              NO SELECT
5     NO SELECT              NO SELECT              NO SELECT
6     Drive: 1,0             Drive: 3,0             Drive: 5,0
7     Drive: 1,1             NO SELECT              NO SELECT
8     NO SELECT              NO SELECT              NO SELECT
9     NO SELECT              NO SELECT              NO SELECT
10    NO SELECT              NO SELECT              NO SELECT
...
```

   The empty slots are shown with a NO SELECT status. The output shown here is from a SPARCstorage Array 110; your display will be slightly different if you are using a different series SPARCstorage Array.

   Determine the tray to which you will add the new disk.

   In the remainder of the procedure, Tray 2 is used as an example. The slot selected for the new disk is Tray 2 Slot 7. The new disk will be known as c2t3d1.

4. **Identify all the volumes and corresponding plexes on the disks in the tray which will contain the new disk.**

a. **From the physical device address cNtNdN, obtain the controller number and the target number.**

In this example, the controller number is 2 and the target is 3.

b. **Identify devices from a** `vxdisk list` **output.**

Here is an example of how `vxdisk` can be used to obtain the information.

```
# vxdisk -g diskgroup -q list | nawk ’/^c2/ {print $3}’
```

Record the volume media name for the disks from the output of the command.

c. **Identify all plexes on the above devices by using the appropriate version (**`csh`**,** `ksh`**, or Bourne shell) of the following command.**

```
PLLIST=`vxprint -ptq -g diskgroup -e ’(aslist.sd_dm_name in (‘‘c2t3d0’’)) && (pl_kstate=ENABLED)’ | nawk ’{print $2}’`
```

For `csh`, the syntax is `set PLLIST ....` For `ksh`, the syntax is `export PLLIST= ....` The Bourne shell requires the command `export PLLIST` after the variable is set.

5. **After you have set the variable, stop I/O to the volumes whose components (subdisks) are on the tray.**

Make sure all volumes associated with that tray are detached (mirrored or RAID5 configurations) or stopped (simple plexes). Issue the following command to detach a mirrored plex.

```
# vxplex -g diskgroup det ${PLLIST}
```

An alternate command for detaching each plex in a tray is:

```
# vxplex -g diskgroup -v volume det plex
```

To stop I/O to simple plexes, unmount any file systems or stop database access.

---

**Note -** Mirrored volumes will still be active because the other half of the mirror is still available.

---

6. **Add the new disk.**

Use the instructions in your multihost disk expansion unit service manual to perform the hardware procedure of adding the disk.

7. **Make sure all disks in the tray spin up.**

   The disks in the SPARCstorage Array tray should spin up automatically, but if the tray fails to spin up within two minutes, force the action with the following command:

   ```
   phys-hahost1# luxadm start -t 2 c2
   ```

8. **Run the** `drvconfig(1M)` **and** `disks(1M)` **commands to create the new entries in** `/devices`**,** `/dev/dsk`**, and** `/dev/rdsk` **for all new disks.**

   ```
   phys-hahost1# drvconfig
   phys-hahost1# disks
   ```

9. **Force the VxVM** `vxconfigd` **driver to scan for new disks.**

```
phys-hahost1# vxdctl enable
```

10. **Bring the new disk under VM control by using the** `vxdiskadd` **command.**

11. **Perform the usual administration actions on the new disk.**

    You can now perform the usual administration steps that bring a new drive into service. These include partitioning the disk, adding it to the configuration as a hot spare, or configuring it as a plex.

    This completes the procedure of adding a multihost disk to an existing SPARCstorage Array.

## Replacing a SPARCstorage Array Disk

This section describes replacing a SPARCstorage Array (SSA) multihost disk without interrupting Sun Cluster services (online replacement) when the volume manager is reporting problems such as:

- Components in the “Needs Maintenance” state
- Hot spare replacement
- Intermittent disk errors

## ▼ How to Replace a SPARCstorage Array Disk (Solstice DiskSuite)

These are the high-level steps to replace a multihost disk in a Solstice DiskSuite configuration. Some of the steps in this procedure apply only to configurations using SPARCstorage Array 100 series or SPARCstorage Array 200 series with the differential SCSI tray.

- Switching logical hosts to one cluster node
- Determining which disk needs replacement
- Determining which tray holds the disk to be replaced
- (SSA 100 and SSA 200 only) Detaching submirrors on the affected tray or disk enclosure
- (SSA 100 and SSA 200 only) Running `metaclear(1M)` on the detached submirrors
- (SSA 100 and SSA 200 only) Deleting available hot spares in the affected disk tray
- Removing the bad disk from the diskset
- (SSA 100 and SSA 200 only) Deleting any affected metadevice state database replicas on disks in the affected tray
- (SSA 100 and SSA 200 only) Producing a list of metadevices in the affected tray
- (SSA 100 and SSA 200 only) Using `metaoffline(1M)` on submirrors in the affected tray or submirrors using hot spares in the tray
- (SSA 100 and SSA 200 only) Flushing NVRAM, if enabled
- Spinning down the disk(s) and removing the tray or disk enclosure
- Replacing the disk drive
- Running the `scdidadm -R` command
- Adding the new disk to the diskset
- Reserving and enabling failfast on the new disk
- Partitioning the new disk
- (SSA 100 and SSA 200 only) Using the `metainit(1M)` command to initialize any devices that were cleared previously with the `metaclear(1M)` command
- (SSA 100 and SSA 200 only) Bringing offline mirrors back on line using `metaonline(1M)` and resynchronizing
- (SSA 100 and SSA 200 only) Attaching submirrors unattached previously
- (SSA 100 and SSA 200 only) Replacing any hot spares in use in the submirrors that have just been attached
- (SSA 100 and SSA 200 only) Returning the deleted hot spare devices to their original hot spare pools

- Running the metastat(1M) command to verify the problem has been fixed

These are the detailed steps to replace a failed multihost disk in a Solstice DiskSuite configuration.

1. **Switch ownership of the affected logical hosts to other nodes by using the** haswitch(1M) **command.**

```
phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
```

The SPARCstorage Array tray containing the failed disk might contain disks included in more than one logical host. If this is the case, switch ownership of all logical hosts with disks using this tray to another node in the cluster.

2. **Identify the disk to be replaced by examining** metastat(1M) **and** /var/adm/messages **output.**

When metastat(1M) reports that a device is in maintenance state or some of the components have been replaced by hot spares, you must locate and replace the device. A sample metastat(1M) output follows. In this example, device c3t3d4s0 is in maintenance state.

```
phys-hahost1# metastat -s hahost1
...
 d50:Submirror of hahost1/d40
      State: Needs Maintenance
      Stripe 0:
          Device       Start Block      Dbase      State          Hot Spare
          c3t3d4s0     0                No         Okay           c3t5d4s0
...
```

Check /var/adm/messages to see what kind of problem has been detected.

```
...
Jun 1 16:15:26 host1 unix: WARNING: /io-unit@f,e1200000/sbi@0.0/SUNW,pln@a0000000,741022/
ssd@3,4(ssd49):
Jun 1 16:15:26 host1 unix: Error for command ‘write(I))’ Err
Jun 1 16:15:27 host1 unix: or Level: Fatal
Jun 1 16:15:27 host1 unix: Requested Block 144004, Error Block: 715559
Jun 1 16:15:27 host1 unix: Sense Key: Media Error
Jun 1 16:15:27 host1 unix: Vendor ‘CONNER’:
Jun 1 16:15:27 host1 unix: ASC=0x10(ID CRC or ECC error),ASCQ=0x0,FRU=0x15
...
```

3. **Determine the location of the problem disk by running the `luxadm(1M)` command.**

   The `luxadm(1M)` command lists the trays and the drives associated with them. The output differs for each SPARCstorage Array series. This example shows output from a SPARCstorage Array 100 series array. The damaged drive is highlighted below.

```
phys-hahost1# luxadm display c3
         SPARCstorage Array Configuration
Controller path:

/devices/iommu@f,e0000000/sbus@f,e0001000/SUNW,soc@0,0/
SUNW,pln@a0000000,779a16:ctlr
         DEVICE STATUS
         TRAY1          TRAY2          TRAY3
Slot
1        Drive:0,0      Drive:2,0      Drive:4,0
2        Drive:0,1      Drive:2,1      Drive:4,1
3        Drive:0,2      Drive:2,2      Drive:4,2
4        Drive:0,3      Drive:2,3      Drive:4,3
5        Drive:0,4      Drive:2,4      Drive:4,4
6        Drive:1,0      Drive:3,0      Drive:5,0
7        Drive:1,1      Drive:3,1      Drive:5,1
8        Drive:1,2      Drive:3,2      Drive:5,2
9        Drive:1,3      Drive:3,3      Drive:5,3
10       Drive:1,4      Drive:3,4      Drive:5,4

         CONTROLLER STATUS
Vendor:    SUN
Product ID:  SSA110
Product Rev: 1.0
Firmware Rev: 3.9
Serial Num: 000000741022
Accumulate performance Statistics: Enabled
```

4. **Detach all submirrors with components on the disk being replaced.**

   If you are detaching a submirror that has a failed component, you must force the detach using the `metadetach -f` command. The following example command detaches submirror d50 from metamirror d40.

```
phys-hahost1# metadetach -s hahost1 -f d40 d50
```

5. **Use the `metaclear(1M)` command to clear the submirrors detached in Step 4 on page 207.**

```
phys-hahost1# metaclear -s hahost1 -f d50
```

6. **Before deleting replicas and hot spares, make a record of the location (slice), number of replicas, and hot spare information (names of the devices and list of devices that contain hot spare pools) so that the actions can be reversed following the disk replacement.**

7. **Delete all hot spares that have `Available` status and are in the same tray as the problem disk.**

   This includes all hot spares, regardless of their logical host assignment. In the following example, the `metahs(1M)` command reports hot spares on `hahost1`, but shows that none are present on `hahost2`"

```
phys-hahost1# metahs -s hahost1 -i
hahost1:hsp000 2 hot spares
        c1t4d0s0                Available       2026080 blocks
        c3t2d5s0                Available       2026080 blocks
phys-hahost1# metahs -s hahost1 -d hsp000 c3t2d4s0
hahost1:hsp000:
        Hotspare is deleted
phys-hahost1# metahs -s hahost2 -i
phys-hahost1#
hahost1:hsp000 1 hot spare
   c3t2d5s0                Available       2026080 blocks
```

8. **Use the `metaset(1M)` command to remove the failed disk from the diskset.**

   The syntax for the command is shown below. In this example, *diskset* is the name of the diskset containing the failed disk, and *drive* is the DID name of the disk in the form `d`*N* (for new installations of Sun Cluster), or `c`*N*`t`*Y*`d`*Z* (for installations that upgraded from HA 1.3).

```
# metaset -s diskset -d drive
```

   This operation can take up to fifteen minutes or more, depending on the size of your configuration and the number of disks.

9. **Delete any metadevice state database replicas that are on disks in the tray to be serviced.**

   The `metadb(1M)` command with the `-s` option reports replicas in a specified diskset.

```
phys-hahost1# metadb -s hahost1
phys-hahost1# metadb -s hahost2
phys-hahost1# metadb -s hahost1 -d replicas-in-tray
phys-hahost1# metadb -s hahost2 -d replicas-in-tray
```

(Continuation)

10. **Locate the submirrors using components that reside in the affected tray.**

    One method is to use the metastat(1M) command to create temporary files that contain the names of all metadevices. For example:

```
phys-hahost1# metastat -s hahost1 > /usr/tmp/hahost1.stat
phys-hahost1# metastat -s hahost2 > /usr/tmp/hahost2.stat
```

Search the temporary files for the components in question (c3t3dn and c3t2dn in this example). The information in the temporary files will look like this:

```
...
hahost1/d35: Submirror of hahost1/d15
   State: Okay
   Hot Spare pool: hahost1/hsp100
   Size: 2026080 blocks
   Stripe 0:
      Device      Start Block     Dbase     State      Hot Spare
      c3t3d3s0    0               No        Okay
hahost1/d54: Submirror of hahost1/d24
   State: Okay
   Hot Spare pool: hahost1/hsp106
   Size: 21168 blocks
   Stripe 0:
      Device      Start Block     Dbase     State      Hot Spare
      c3t3d3s6    0               No        Okay
...
```

11. **Take offline all other submirrors that have components in the affected tray.**

    Using the output from the temporary files in Step 10 on page 209, run the metaoffline(1M) command on all submirrors in the affected tray

```
phys-hahost1# metaoffline -s hahost1 d15 d35
phys-hahost1# metaoffline -s hahost1 d24 d54
...
```

.

Run `metaoffline(1M)` as many times as necessary to take all the submirrors off line. This forces Solstice DiskSuite to stop using the submirror components.

12. **If enabled, flush the NVRAM on the controller, tray, individual disk or disks.**

    ```
    phys-hahost1# luxadm sync_cache pathname
    ```

    A confirmation appears, indicating that NVRAM has been flushed. See "Flushing and Purging NVRAM" on page 219, for details on flushing NVRAM data.

13. **Spin down all disks in the affected SPARCstorage Array tray(s).**

    Use the `luxadm stop` command to spin down the disks. Refer to the `luxadm(1M)` man page for details.

    ```
    phys-hahost1# luxadm stop -t 2 c3
    ```

**Caution -** Do not run any Solstice DiskSuite commands while a SPARCstorage Array tray is spun down because the commands might have the side effect of spinning up some or all of the drives in the tray.

14. **Replace the disk.**

    Refer to the hardware service manuals for your SPARCstorage Array for details on this procedure.

15. **Update the DID driver's database with the new device ID.**

    Use the `-l` flag to `scdidadm(1M)` to identify the DID name for the lower-level device name of the drive to be replaced. Then update the DID drive database using the `-R` flag to `scdidadm(1M)`. Refer to the *Sun Cluster 2.2 Software Installation Guide* for details on the DID pseudo driver.

    ```
    phys-hahost1# scdidadm -o name -l /dev/rdsk/c3t3d4
    6 phys-hahost1:/dev/rdsk/c3t3d4 /dev/did/rdsk/d6
    phys-hahost1# scdidadm -R d6
    ```

    **(continued)**

(Continuation)

16. **Make sure all disks in the affected multihost disk expansion unit spin up.**

    The disks in the multihost disk expansion unit should spin up automatically. If the tray fails to spin up within two minutes, force the action by using the following command:

```
phys-hahost1# luxadm start -t 2 c3
```

17. **Add the new disk back into the diskset by using the** `metaset(1M)` **command.**

    This step automatically adds back the number of replicas that were deleted from the failed disk. The command syntax is as follows, where *diskset* is the name of the diskset containing the failed disk, and *drive* is the DID name of the disk in the form `d`*N* (for new installations of Sun Cluster), or `c`*N*`t`*Y*`d`*Z* (for installations that upgraded from HA 1.3):

```
# metaset -s diskset -a drive
```

18. **(Optional) If you deleted replicas that belonged to other disksets from disks that were in the same tray as the errored disk, use the** `metadb(1M)` **command to add back the replicas.**

```
phys-hahost1# metadb -s hahost2 -a deleted-replicas
```

    To add multiple replicas to the same slice, use the `-c` option.

19. **Use the** `scadmin(1M)` **command to reserve and enable failfast on the specified disk that has just been added to the diskset.**

```
phys-hahost2# scadmin reserve c3t3d4
```

20. **Use the** `format(1M)` **or** `fmthard(1M)` **command to repartition the new disk.**

    Make sure that you partition the new disk exactly as the disk that was replaced. (Saving the disk format information was recommended in Chapter 1.)

21. **Use the** `metainit(1M)` **command to reinitialize disks that were cleared in Step 5 on page 207.**

```
phys-hahost1# metainit -s hahost1 d50
```

22. **Bring online all submirrors that were taken off line in Step 11 on page 209.**

```
phys-hahost1# metaonline -s hahost1 d15 d35
phys-hahost1# metaonline -s hahost1 d24 d54
...
```

Run the metaonline(1M) command as many times as necessary to bring online all the submirrors.

When the submirrors are brought back online, Solstice DiskSuite automatically performs resyncs on all the submirrors, bringing all data up-to-date.

---

**Note -** Running the metastat(1M) command at this time would show that all metadevices with components residing in the affected tray are resyncing.

---

23. **Attach submirrors that were detached in Step 4 on page 207.**

Use the metattach(1M) command to perform this step. See the metattach(1M) man page for details.

```
phys-hahost1# metattach -s hahost1 d40 d50
```

24. **Replace any hot spares in use in the submirrors attached in Step 23 on page 212.**

If a submirror had a hot spare replacement in use before you detached the submirror, this hot spare replacement will be in effect after the submirror is reattached. This step returns the hot spare to Available status.

```
phys-hahost1# metareplace -s hahost1 -e d40 c3t3d4s0
```

25. **Restore all hot spares that were deleted in Step 7 on page 208.**

Use the metahs(1M) command to add back the hot spares. See the metahs(1M) man page for details.

```
phys-hahost1# metahs -s hahost1 -a hsp000 c3t2d5s0
```

26. **If necessary, switch logical hosts back to their default masters.**

```
phys-hahost1# haswitch phys-hahost2 hahost2
```

**27. Verify that the replacement corrected the problem.**

```
phys-hahost1# metastat -s hahost1
```

## ▼ How to Replace a SPARCstorage Array Disk (VxVM)

In a VxVM configuration, it is possible to replace a SPARCstorage Array disk without halting the system, as long as the configuration is mirrored.

---

**Note -** If you need to replace a disk in a bootable SPARCstorage Array, do not remove the SSA trays containing the boot disk of the hosts. Instead, shut down the host whose boot disk is present on that tray. Let the cluster software reconfigure the surviving nodes for failover to take effect before servicing the faulty disk. Refer to the *SPARCstorage Array User's Guide* for more information.

---

These are the high-level steps to replace a multihost disk in an VxVM environment using SPARCstorage Array 100 series disks.

- Identifying all volumes and corresponding plexes on the disks in the tray which contains the faulty disk
- Determining the controller and target number of the errored disk
- Identifying devices on the tray by using the `vxdisk list` command
- Identifying all plexes on the affected tray
- Detaching all plexes on the affected tray
- Removing the disk from its disk group
- Spinning down the disks in the tray
- Replacing the disk drive
- Spinning up the drives in the tray
- Initializing the replacement disk drive
- Scanning the current disk configuration
- Adding the replacement disk drive to the disk group
- Resynchronizing the volumes

These are the detailed steps to replace a multihost disk in an VxVM environment using SPARCstorage Array 100 series disks.

1. **If the replaced disk is a quorum device, use the `scconf -q` command to change the quorum device to a different disk.**

2. **Identify all the volumes and corresponding plexes on the disks in the tray which contains the faulty disk.**

   a. **From the physical device address `c`*N*`t`*N*`d`*N*, obtain the controller number and the target number.**

      For example, if the device address is c3t2d0, the controller number is 3 and the target is 2.

   b. **Identify devices from a `vxdisk list` output.**

      If the target is 0 or 1, identify all devices with physical addresses beginning with c*N*t0 and c*N*t1, where `N` is the controller number. If the target is 2 or 3, identify all devices with physical addresses beginning with c*N*t2 and c*N*t3. If the target is 4 or 5, identify all devices with physical addresses beginning with c*N*t4 and c*N*t5. Here is an example of how `vxdisk` can be used to obtain the information.

```
# vxdisk -g diskgroup -q list | egrep c3t2\|c3t3 | nawk ’{print $3}’
```

   c. **Record the volume media name for the faulty disk from the output of the command.**

      You will need this name in Step 10 on page 215.

   d. **Identify all plexes on the above devices by using the appropriate version (`csh`, `ksh`, or Bourne shell) of the following command.**

```
PLLIST=`vxprint -ptq -g diskgroup -e ’(aslist.sd_dm_name in   (‘‘c3t2d0’’,’’c3t3d0’’,’’c3t3d1’’)) && (pl_kstate=ENABLED)’
```

      For `csh`, the syntax is `set PLLIST ....` For `ksh`, the syntax is `export PLLIST= ....` The Bourne shell requires the command `export PLLIST` after the variable is set.

3. **After you have set the variable, stop I/O to the volumes whose components (subdisks) are on the tray.**

   Make sure all volumes associated with that tray are detached (mirrored or RAID5 configurations) or stopped (simple plexes). Issue the following command to detach a mirrored plex.

```
# vxplex det ${PLLIST}
```

An alternate command for detaching each plex in a tray is:

```
# vxplex -g diskgroup -v volume det plex
```

To stop I/O to simple plexes, unmount any file systems or stop database access.

---

**Note -** Mirrored volumes will still be active because the other half of the mirror is still available.

---

4. **Remove the disk from the disk group.**

```
# vxdg -g diskgroup rmdisk diskname
```

5. **Spin down the disks in the tray.**

```
# luxadm stop -t tray controller
```

6. **Replace the faulty disk.**

7. **Spin up the drives.**

```
# luxadm start -t tray controller
```

8. **Initialize the replacement disk.**

```
# vxdisksetup -i devicename
```

9. **Scan the current disk configuration again.**

   Enter the following commands on all nodes in the cluster.

```
# vxdctl enable
# vxdisk -a online
```

10. **Add the new disk to the disk group.**

    The *device-media-name* is the volume media name recorded in Step 2 on page 214c.

```
# vxdg -g diskgroup -k adddisk device-media-name=device-name
```

**11. Resynchronize the volumes.**

```
# vxrecover -g diskgroup -b -o
```

# Administering SPARCstorage Array NVRAM

NVRAM supports the fast write capability for SPARCstorage Arrays. Without NVRAM, synchronous write requests from a program must be committed to disk, and an acknowledgment must be received by the program, before another request can be submitted. The NVRAM caches write requests in non-volatile memory and periodically flushes the data to disk. Once the data is in the NVRAM, an acknowledgment is returned to the program just as if the data had been written to disk. This enhances performance of write-intensive applications using SPARCstorage Arrays.

The procedures described here use the command-line interface. However, in Solstice DiskSuite configurations, you also can use the `metatool` graphical user interface to administer NVRAM for a disk, tray, or controller. For more information on Solstice DiskSuite, see the Solstice DiskSuite documentation.

**Caution -** Use this functionality with care. It provides a powerful way to manage the SPARCstorage Array. Back up your data before using these procedures.

## Enabling and Disabling NVRAM

Fast writes can be configured:

- At the controller level, affecting all drives in the SPARCstorage Array
- At the drive level, setting fast write for an individual drive
- At the tray level, through the Solstice DiskSuite GUI

When fast write is enabled, it can be saved—across power cycles—as part of the SPARCstorage Array's configuration.

If the NVRAM battery is low, missing, or has failed, then fast write is disabled on the controller.

Before enabling fast write, you must stop all I/O to the controller or disk. In particular, ensure that diskset ownership has been released because an implicit I/O stream exists while ownership of a diskset is maintained. The following procedure explains how to stop all I/O.

Use the `luxadm(1M)` command to enable and disable NVRAM. Refer to the `luxadm(1M)` man page for complete information on this command.

---

**Note -** For VxVM cluster feature (as used with Oracle Parallel Server), you should disable NVRAM.

---

## ▼ How to Enable and Disable NVRAM

These are the high-level steps to enable or disable NVRAM:

- Ensuring that you have a current backup of all data
- Ensuring that you have `root` privilege
- Identifying the controller or disk on which to enable or disable NVRAM
- Stopping all I/O to the device
- Enabling or disabling NVRAM
- Bringing the device back up and resynchronizing the data

These are the detailed steps to enable or disable NVRAM.

1. **Identify the controller, tray, or individual disk whose NVRAM is to be enabled or disabled.**

   You can use the `luxadm(1M)` command to display information for a specified controller, tray, or individual disk. For example, the following display identifies all of the disks on Controller c2.

```
phys-hahost1# luxadm display c2
                     SPARCstorage Array Configuration

Controller path:
/devices/iommu@f,e0000000/sbus@f,e0001000/SUNW,soc@0,0/
SUNW,pln@a0000000,779a16:ctlr
                          DEVICE STATUS
      TRAY 1                 TRAY 2                 TRAY 3
slot
1     Drive: 0,0             Drive: 2,0             Drive: 4,0
2     Drive: 0,1             Drive: 2,1             Drive: 4,1
3     NO SELECT              NO SELECT              NO SELECT
4     NO SELECT              NO SELECT              NO SELECT
```

**(continued)**

**(Continuation)**

```
5     NO SELECT              NO SELECT              NO SELECT
6     Drive: 1,0             Drive: 3,0             Drive: 5,0
7     Drive: 1,1             NO SELECT              NO SELECT
8     NO SELECT              NO SELECT              NO SELECT
9     NO SELECT              NO SELECT              NO SELECT
10    NO SELECT              NO SELECT              NO SELECT
                         CONTROLLER STATUS
...
```

2. **Stop all I/O to the affected device.**

   Solstice DiskSuite:

   - For a controller or tray, refer to the applicable steps in the procedure "How to Take a SPARCstorage Array Tray Out of Service (Solstice DiskSuite)" on page 177.
   - For a disk, refer to the applicable steps in the procedure "How to Replace a SPARCstorage Array Disk (Solstice DiskSuite)" on page 205.

   VxVM:

   - For a controller or tray, refer to the applicable steps in the procedure "How to Take a SPARCstorage Array Tray Out of Service (VxVM)" on page 179.
   - For a disk, refer to the applicable steps in the procedure "How to Replace a SPARCstorage Array Disk (VxVM)" on page 213.

3. **Enable or disable fast write on the controller or individual disk.**

   Use one of the three options to the `luxadm(1M)` command, depending on whether you are enabling fast write for all writes, enabling fast write only for synchronous writes, or disabling fast write.

   - `-e` enables fast write for all writes
   - `-c` enables fast write for only synchronous writes
   - `-d` disables fast write

   The following example saves the NVRAM configuration across power cycles and enables fast write for all writes. See the `luxadm(1M)` man page for details on these options.

```
phys-hahost# luxadm fast_write -s -e pathname
```

   A confirmation appears, indicating that fast write has been enabled.

4. **Perform the steps needed to bring the component into normal operation under Sun Cluster.**

   Solstice DiskSuite:

   - For a controller or tray, refer to the applicable steps in the procedure "How to Return a SPARCstorage Array Tray to Service (Solstice DiskSuite)" on page 181.
   - For a disk, refer to the applicable steps in the procedure "How to Replace a SPARCstorage Array Disk (Solstice DiskSuite)" on page 205.

   VxVM:

   - For a controller or tray, refer to the applicable steps in the procedure "How to Return a SPARCstorage Array Tray to Service (VxVM)" on page 182.
   - For a disk, refer to the applicable steps in the procedure "How to Replace a SPARCstorage Array Disk (VxVM)" on page 213.

## Flushing and Purging NVRAM

The `luxadm sync_cache` command flushes any outstanding writes from NVRAM to the disk drive. If you get an error while flushing data, you must purge the data using the `luxadm purge` command. Purging data "throws away" any outstanding writes in NVRAM.

---

**Caution -** Purging fast write data should be performed with caution, and only when a drive has failed, as it could result in the loss of data.

---

If the NVRAM battery is low, missing, or has failed, then NVRAM is non-functional and data is lost.

## ▼ How to Flush and Purge NVRAM

These are the high-level steps to flush or purge all outstanding writes for the selected controller (and all disks) or individual writes from the NVRAM to disk:

- Ensuring you have a current backup of all data
- Ensuring you have `root` privilege
- Identifying the controller or disk on which to flush or purge writes
- Flushing or purging all outstanding writes
- Stopping all I/O to the device
- Bringing the device back into service under Sun Cluster

These are the detailed steps to flush or purge NVRAM data.

1. **Identify the controller or the individual disk to flush or purge.**

You can use the luxadm(1M) command to display information for a specified controller, tray, or individual disk. For example, the following display identifies all of the disks on Controller c2.

```
phys-hahost1# luxadm display c2
                     SPARCstorage Array Configuration

Controller path:
/devices/iommu@f,e0000000/sbus@f,e0001000/SUNW,soc@0,0/
SUNW,pln@a0000000,779a16:ctlr
                          DEVICE STATUS
      TRAY 1                 TRAY 2                 TRAY 3
slot
1     Drive: 0,0             Drive: 2,0             Drive: 4,0
2     Drive: 0,1             Drive: 2,1             Drive: 4,1
3     NO SELECT              NO SELECT              NO SELECT
4     NO SELECT              NO SELECT              NO SELECT
5     NO SELECT              NO SELECT              NO SELECT
6     Drive: 1,0             Drive: 3,0             Drive: 5,0
7     Drive: 1,1             NO SELECT              NO SELECT
8     NO SELECT              NO SELECT              NO SELECT
9     NO SELECT              NO SELECT              NO SELECT
10    NO SELECT              NO SELECT              NO SELECT
                         CONTROLLER STATUS
Vendor:        SUN
Product ID:    SSA110
Product Rev:   1.0
Firmware Rev:  3.9
Serial Num:    000000779A16
Accumulate Performance Statistics: Enabled
phys-hahost1#
```

2. **Stop all I/O to the affected device.**

   Solstice DiskSuite:

   - For a controller or tray, refer to the applicable steps in the procedure “How to Take a SPARCstorage Array Tray Out of Service (Solstice DiskSuite)” on page 177.
   - For a disk, refer to the applicable steps in the procedure “How to Replace a SPARCstorage Array Disk (Solstice DiskSuite)” on page 205.

   VxVM:

   - For a controller or tray, refer to the applicable steps in the procedure “How to Take a SPARCstorage Array Tray Out of Service (VxVM)” on page 179.
   - For a disk, refer to the applicable steps in the procedure “How to Replace a SPARCstorage Array Disk (VxVM)” on page 213.

3. **Flush or purge the NVRAM on a controller, tray, or individual disk.**

If you can access drives in the SPARCstorage Array, flush the NVRAM. Only purge the NVRAM if you can no longer access the SPARCstorage Array or disk.

```
phys-hahost1# luxadm sync_cache pathname
or
phys-hahost1# luxadm purge pathname
```

A confirmation appears, indicating that NVRAM has been flushed or purged.

4. **Perform the steps needed to bring the component into normal operation under Sun Cluster.**

   Solstice DiskSuite:

   - For a controller or tray, refer to the applicable steps in the procedure "How to Return a SPARCstorage Array Tray to Service (Solstice DiskSuite)" on page 181.
   - For a disk, refer to the applicable steps in the procedure "How to Replace a SPARCstorage Array Disk (Solstice DiskSuite)" on page 205.

   VxVM:

   - For a controller or tray, refer to the applicable steps in the procedure "How to Return a SPARCstorage Array Tray to Service (VxVM)" on page 182.
   - For a disk, refer to the applicable steps in the procedure "How to Replace a SPARCstorage Array Disk (VxVM)" on page 213.

CHAPTER **12**

# Administering Sun StorEdge MultiPacks and Sun StorEdge D1000s

This chapter provides instructions for administering Sun StorEdge™ MultiPack and Sun StorEdge D1000 disks. Some of the procedures documented in this chapter are dependent on your volume management software (Solstice DiskSuite or VxVM). These procedures include the volume manager name in their titles.

Use the service manual for your Sun StorEdge MultiPack and Sun StorEdge D1000 disks, and the volume management software documentation, when you are replacing or repairing disk hardware in the Sun Cluster configuration.

## Recovering From Power Loss

When power is lost to one Sun StorEdge MultiPack or Sun StorEdge D1000, I/O operations generate errors that are detected by your volume management software. Errors are not reported until I/O transactions are made to the disk.

You should monitor the configuration for these events using the commands described in Chapter 2.

## ▼ How to Recover From Power Loss (Solstice DiskSuite)

These are the high-level steps to recover from power loss to a disk enclosure in a Solstice DiskSuite environment:

- Identifying the errored replicas
- Returning the errored replicas to service
- Identifying the errored devices
- Returning the errored devices to service
- Resyncing the disks

These are the detailed steps to recover from power loss to a disk enclosure in a Solstice DiskSuite environment.

1. **When power is restored, use the** `metadb(1M)` **command to identify the errored replicas:**

```
# metadb -s diskset
```

2. **Return replicas to service.**

   After the loss of power, all metadevice state database replicas on the affected disk enclosure chassis enter an errored state. Because metadevice state database replica recovery is not automatic, it is safest to perform the recovery immediately after the disk enclosure returns to service. Otherwise, a new failure can cause a majority of replicas to be out of service and cause a kernel panic. This is the expected behavior of Solstice DiskSuite when too few replicas are available.

   While these errored replicas will be reclaimed at the next takeover (`haswitch(1M)` or `reboot(1M)`), you might want to return them to service manually by first deleting and then adding them back.

   **Note -** Make sure that you add back the same number of replicas that were deleted on each slice. You can delete multiple replicas with a single `metadb(1M)` command. If you need multiple copies of replicas on one slice, you must add them in one invocation of the `metadb(1M)` command using the `-c` flag.

3. **Use the** `metastat(1M)` **command to identify the errored metadevices.**

```
# metastat -s diskset
```

4. **Return errored metadevices to service using the** `metareplace(1M)` **command, and resync the disks.**

   ```
   # metareplace -s diskset -e mirror component
   ```

   The `-e` option transitions the component (slice) to the available state and performs a resync.

   Components that have been replaced by a hot spare should be the last devices replaced using the `metareplace(1M)` command. If the hot spare is replaced first, it could replace another errored submirror as soon as it becomes available.

   You can perform a resync on only one component of a submirror (metadevice) at a time. If all components of a submirror were affected by the power outage, each component must be replaced separately. It takes approximately 10 minutes to resync a 1.05GB disk.

   If both disksets in a symmetric configuration were affected by the power outage, you can resync each diskset's affected submirrors concurrently. Log into each host separately to recover that host's diskset by running `metareplace(1M)` on each.

   **Note -** Depending on the number of submirrors and the number of components in these submirrors, the resync actions can require a considerable amount of time. A single submirror made up of 30 1.05GB drives might take about five hours to complete, whereas a configuration made up of five component submirrors might take only 50 minutes to complete.

## ▼ How to Recover From Power Loss (VxVM)

Power failures can detach disk drives and cause plexes to become detached, and thus, unavailable. The volume remains active, however, because the remaining plexes in a mirrored volume are still available. It is possible to reattach the disk drives and recover from this condition without halting nodes in the cluster.

These are the high-level steps to recover from power loss to a disk enclosure in an VxVM configuration:

- Determining the errored plex(es) by using the `vxprint` and `vxdisk` commands
- Fixing the problem that caused the power loss
- Using the `drvconfig` and `disks` commands to create the `/devices` and `/dev` entries
- Scanning the current disk configuration
- Reattaching disks that had transient errors
- Verifying there are no more errors

- (Optional) For shared disk groups, running the `vxdg` command for each disk that was powered off
- Starting volume recovery

These are the detailed steps to recover from power loss to a disk enclosure in an VxVM configuration.

1. **Use the `vxprint` command to view the errored plexes.**

   Optionally, specify a diskgroup with the `-g` *diskgroup* option.

2. **Use the `vxdisk` command to identify the errored disks.**

```
# vxdisk list
DEVICE       TYPE      DISK         GROUP        STATUS
..
-            -         c1t5d0       toi          failed was:c1t5d0s2
...
```

3. **Fix the condition that resulted in the problem so that power is restored to all failed disks.**

   Be sure that the disks are spun up before proceeding.

4. **Enter the following commands on all nodes in the cluster.**

   In some cases, the drive(s) must be rediscovered by the node(s).

```
# drvconfig
# disks
```

5. **Enter the following commands on all nodes in the cluster.**

   The volume manager must scan the current disk configuration again.

```
# vxdctl enable
# vxdisk -a online
```

6. **Enter the following command on all nodes in the cluster.**

   ---

   **Note -** For VxVM cluster feature (as used with Oracle Parallel Server), enter the command on the master node first, then on the remaining nodes.

   ---

   This will reattach and initiate recovery on disks that had transitory failure.

```
# vxreattach -r
```

7. **Verify the output of the `vxdisk` command to see if there are any more errors.**

```
# vxdisk list
```

8. **If media was replaced, enter the following command from the master node for each disk that has been disconnected.**

   The physical disk and the volume manager access name for that disk must be reconnected.

   ```
   # vxdg -g diskgroup -k adddisk medianame=accessname
   ```

   The values for `medianame` and `accessname` appear at the end of the `vxdisk list` command output.

   For example:

   ```
   # vxdg -g toi -k adddisk c1t5d0=c1t5d0s2
   # vxdg -g toi -k adddisk c1t5d1=c1t5d1s2
   # vxdg -g toi -k adddisk c1t5d2=c1t5d2s2
   # vxdg -g toi -k adddisk c1t5d3=c1t5d3s2
   # vxdg -g toi -k adddisk c1t5d4=c1t5d4s2
   ```

   You can also use the `vxdiskadm` command, or the graphical user interface, to reattach the disks.

9. **From the node, start volume recovery.**

   ```
   # vxrecover -bv [-g diskgroup]
   ```

If you have shared disk groups, use the `-svc` options to the `vxrecover` command.

10. **(Optional) Use the `vxprint -g` command to view the changes.**

# Administering Sun StorEdge MultiPacks and Sun StorEdge D1000s

This section describes procedures for administering Sun StorEdge MultiPack and Sun StorEdge D1000 components. Use the procedures described in your server hardware manual to identify the failed component.

## Repairing a Lost Sun StorEdge MultiPack or Sun StorEdge D1000 Connection

When a connection from a disk enclosure to one of the cluster nodes fails, the failure is probably due to a bad SCSI-2 cable or an SBus card.

In any event, the node on which the failure occurred will begin generating errors when the failure is discovered. Later accesses to the disk enclosure will generate additional errors. The node will exhibit the same behavior as though power had been lost to the disk enclosure.

I/O operations from the other nodes in the cluster are unaffected by this type of failure.

To diagnose the failure, use the procedures for testing the card module in the service manual for your Sun Cluster node to determine which component failed. You should free up one node and the disk enclosure that appears to be down, for hardware debugging.

## ▼ How to Repair a Lost Sun StorEdge MultiPack or Sun StorEdge D1000 Connection

1. **Prepare the Sun Cluster system for component replacement.**

   Depending on the cause of the connection loss, prepare the Sun Cluster node with one of the following procedures.

- If the failed component is an SBus card, see Chapter 7, to prepare the Sun Cluster node for power down.
- If the problem is a bad SCSI-2 cable, the volume management software will have detected the problem and prepared the system for cable replacement.

2. **Replace the failed component.**

   If the SCSI-2 cable or SBus card fails, refer to the service manual for your Sun Cluster node for detailed instructions on replacing them.

3. **Recover from volume management software errors.**

   Use the procedures described in “Recovering From Power Loss” on page 223.

## Adding a Sun StorEdge MultiPack or Sun StorEdge D1000

You can add Sun StorEdge MultiPacks or Sun StorEdge D1000s to a Sun Cluster configuration at any time.

You must review the disk group configuration in your Sun Cluster configuration before adding a disk enclosure. The discussions in Chapter 2 in the *Sun Cluster 2.2 Software Installation Guide*, and in Appendix A, in this book, will help determine the impact of the disk enclosure on the configuration of disk groups.

### ▼ How to Add a Sun StorEdge MultiPack or Sun StorEdge D1000

1. **Shut down one of the cluster nodes.**

   Use the procedure in “Stopping the Cluster and Cluster Nodes” on page 83, to shut down the node.

2. **Install an additional SBus card in the node, if necessary.**

   Use the instructions in the hardware service manual for your Sun Cluster node to install the SBus card.

   ---

   **Note -** Install the SBus card in the first available empty SBus slot, following all other cards in the node. This ensures that the controller numbering will be preserved if the Solaris operating environment is reinstalled. Refer to “Instance Names and Numbering” on page 24, for more information.

   ---

3. **Connect the SCSI-2 cables to the disk enclosure.**

Use the instructions in the hardware service manual for your Sun Cluster node.

4. **Set the SCSI initiator ID, as appropriate.**

   Use the instructions in the hardware service manual for your Sun Cluster node.

5. **Perform a reconfiguration reboot of the node.**

```
ok boot -r
```

6. **Use the `haswitch(1M)` command to switch ownership of all logical hosts that can be mastered to the rebooted node.**

```
phys-hahost1# haswitch phys-hahost2 hahost1 hahost2
```

7. **Repeat Step 1 on page 229 through Step 5 on page 230 on other nodes connected to this disk enclosure.**

8. **Switch ownership of the logical hosts back to the appropriate default master if necessary.**

   For example:

```
phys-hahost1# haswitch phys-hahost2 hahost2
```

9. **Add the disks in the disk enclosures to the selected disk group.**

   Use the instructions in your volume manager documentation to add the disks to the selected disk group(s). Also, refer to appendixes in the *Sun Cluster 2.2 Software Installation Guide* for information on Solstice DiskSuite or VxVM.

# Administering Sun StorEdge MultiPack and Sun StorEdge D1000 Disks

As part of standard Sun Cluster administration, you should monitor the status of the configuration. See Chapter 2, for information about monitoring methods. During the monitoring process you might discover problems with multihost disks. The following procedures describe how to correct these problems.

Sun Cluster supports different disk types. Refer to the hardware service manual for your multihost disk expansion unit for a description of your disk enclosure.

## Adding Sun StorEdge MultiPack or Sun StorEdge D1000 Disks

In a symmetric configuration, the disk enclosure might contain disks from multiple disk groups and will require that a single node own all of the affected disk groups.

## ▼ How to Add a Sun StorEdge MultiPack or a Sun StorEdge D1000 Disk

These are the high-level steps to add a Sun StorEdge MultiPack or Sun StorEdge D1000 disk:

- Identifying the controller for this new disk and locating an empty slot in the disk enclosure
- Adding the new disk
- Performing the administrative actions to prepare the disk for use by Sun Cluster
    - Creating the `/devices` special files and `/dev/dsk` and `/dev/rdsk` links
    - Adding the disk to the disk group
    - Formatting and partitioning the disk, if necessary
    - Performing the volume management-related administrative tasks

These are the detailed steps to add a new Sun StorEdge MultiPack or Sun StorEdge D1000 disk.

1. **Determine the controller number of the disk enclosure to which the disk will be added.**

   Use the `mount(1M)` or `format(1M)` command to determine the controller number.

2. **Locate an appropriate empty disk slot in the disk enclosure for the disk being added.**

   Identify the empty slots either by observing the disk drive LEDs on the front of the disk enclosure, or by removing the left side cover of the unit. The target address IDs corresponding to the slots appear on the middle partition of the drive bay.

   In the following steps, Tray 2 is used as an example. The slot selected for the new disk is Tray 2 Slot 7. The new disk will be known as `c2t3d1`.

3. **Add the new disk.**

   Use the instructions in your disk enclosure unit service manual to perform the hardware procedure of adding the disk.

4. **Run the** `drvconfig(1M)` **and** `disks(1M)` **commands to create the new entries in** `/devices`**,** `/dev/dsk`**, and** `/dev/rdsk` **for all new disks.**

   ```
   phys-hahost1# drvconfig
   phys-hahost1# disks
   ```

5. **Switch ownership of the logical hosts to the other cluster node to which this disk is connected.**

   ```
   phys-hahost1# haswitch phys-hahost2 hahost1 hahost2
   ```

6. **Run the** `drvconfig(1M)` **and** `disks(1M)` **commands on the node that now owns the disk group to which the disk will be added.**

   ```
   phys-hahost2# drvconfig
   phys-hahost2# disks
   ```

7. **Add the disk to a disk group using your volume management software.**

   For Solstice DiskSuite, the command syntax is as follows, where *diskset* is the name of the diskset containing the failed disk, and *drive* is the DID name of the disk in the form `d`*N* (for new installations of Sun Cluster), or `c`*N*`t`*Y*`d`*Z* (for installations that upgraded from HA 1.3):

   ```
   # metaset -s diskset -a drive
   ```

   For VxVM, you can use the command line or graphical user interface to add the disk to the disk group.

**Caution -** If you are using Solstice DiskSuite, the `metaset(1M)` command might repartition this disk automatically. See the Solstice DiskSuite documentation for more information.

8. **(Solstice DiskSuite configurations only) After adding the disks to the diskset by using the `metaset(1M)` command, use the `scadmin(1M)` command to reserve and enable failfast on the specified disks.**

```
phys-hahost1# scadmin reserve drivename
```

9. **Perform the usual administration actions on the new disk.**

   You can now perform the usual administration steps that are performed when a new drive is brought into service. See your volume management software documentation for more information on these tasks.

10. **If necessary, switch logical hosts back to their default masters.**

## Replacing Sun StorEdge MultiPack or Sun StorEdge D1000 Disks

This section describes replacing a multihost disk without interrupting Sun Cluster services (online replacement) when the volume manager is reporting problems such as:

- Components in the `Needs Maintenance` state
- Hot spare replacement
- Intermittent disk errors

Consult your volume management software documentation for offline replacement procedures.

### ▼ How to Replace a Sun StorEdge MultiPack or Sun StorEdge D1000 Disk (Solstice DiskSuite)

Use the following procedure if you have determined that a disk has components in the `Needs Maintenance` state, a hot spare has replaced a component, or a disk is generating intermittent errors.

These are the high-level steps to replace a Sun StorEdge MultiPack or Sun StorEdge D1000 disk in a Solstice DiskSuite configuration:

- Determining which disk needs replacement
- Determining which disk expansion unit holds the disk to be replaced
- Removing the bad disk from the diskset
- Spinning down the disk and opening the disk enclosure
- Replacing the disk drive
- Running the `scdidadm -R` command
- Adding the new disk to the diskset
- Reserving and enabling failfast on the disk
- Partitioning the new disk
- Running the `metastat(1M)` command to verify the problem has been fixed

These are the detailed steps to replace a failed Sun StorEdge MultiPack or Sun StorEdge D1000 disk in a Solstice DiskSuite configuration.

1. **Run the procedure on the host that masters the diskset in which the bad disk resides. This might require you to switch over the diskset using the** `haswitch(1M)` **command.**

2. **Identify the disk to be replaced.**

   Use the `metastat(1M)` command and `/var/adm/messages` output.

   When `metastat(1M)` reports that a device is in maintenance state or some of the components have been replaced by hot spares, you must locate and replace the device. A sample `metastat(1M)` output follows. In this example, device `c3t3d4s0` is in maintenance state:

```
phys-hahost1# metastat -s hahost1
...
 d50:Submirror of hahost1/d40
      State: Needs Maintenance
      Stripe 0:
          Device       Start Block      Dbase      State          Hot Spare
          c3t3d4s0     0                No         Okay           c3t5d4s0
...
```

   Check `/var/adm/messages` to see what kind of problem has been detected.

```
...
Jun 1 16:15:26 host1 unix: WARNING: /io-unit@f,e1200000/sbi@0.0/SUNW,pln@a0000000,741022/
ssd@3,4(ssd49):
Jun 1 16:15:26 host1 unix: Error for command 'write(I))' Err
Jun 1 16:15:27 host1 unix: or Level: Fatal
Jun 1 16:15:27 host1 unix: Requested Block 144004, Error Block: 715559
Jun 1 16:15:27 host1 unix: Sense Key: Media Error
Jun 1 16:15:27 host1 unix: Vendor 'CONNER':
Jun 1 16:15:27 host1 unix: ASC=0x10(ID CRC or ECC error),ASCQ=0x0,FRU=0x15
...
```

3. **Determine the location of the problem disk.**

   Use the `mount(1M)` or `format(1M)` command to determine the controller number.

4. **If the problem disk contains replicas, make a record of the slice and number, then delete the replicas.**

   Use the `metadb(1M)` command to delete the replicas.

5. **Detach all submirrors with components on the disk being replaced.**

   If you are detaching a submirror that has a failed component, you must force the detach using the `metadetach -f` option. The following example detaches submirror d50 from metamirror d40.

   ```
   phys-hahost1# metadetach -s hahost1 -f d40 d50
   ```

6. **Use the `metaclear(1M)` command to clear the submirrors detached in Step 5 on page 235.**

   ```
   phys-hahost1# metaclear -s hahost1 -f d50
   ```

7. **If the problem disk contains hot spares, make a record of the names of devices and list of devices that contain hot spare pools, then delete the hot spares.**

   Use the `metahs(1M)` command to delete hot spares.

---

**Caution -** You need to record the information before deleting the objects so that the actions can be reversed following the disk replacement.

---

8. **Use the `metaset(1M)` command to remove the failed disk from the diskset.**

   The command syntax is as follows, where *diskset* is the name of the diskset containing the failed disk, and *drive* is the DID name of the disk in the form d*N* (for new installations of Sun Cluster), or c*N*t*Y*d*Z* (for installations that upgraded from HA 1.3):

   ```
   phys-hahost1# metaset -s diskset -d drive
   ```

   This can take up to fifteen minutes or more, depending on the size of your configuration and the number of disks.

9. **Replace the bad disk.**

   Refer to the hardware service manuals for your disk enclosure for details on this procedure.

10. **Make sure the new disk spins up.**

    The disk should spin up automatically.

11. **Update the DID driver's database with the new device ID.**

    ---

    **Note -** If you upgraded from HA 1.3, your installation does not use the DID driver, so skip this step.

    ---

    Use the `-l` flag to `scdidadm(1M)` to identify the DID name for the lower level device name of the drive to be replaced. Then update the DID drive database using the `-R` flag to `scdidadm(1M)`. Refer to the *Sun Cluster 2.2 Software Installation Guide* for details on the DID pseudo driver.

    ```
    phys-hahost1# scdidadm -o name -l /dev/rdsk/c3t3d4
    6 phys-hahost1:/dev/rdsk/c3t3d4 /dev/did/rdsk/d6
    phys-hahost1# scdidadm -R d6
    ```

12. **Add the new disk back into the diskset using the `metaset(1M)` command.**

    This step adds automatically adds back the proper number of replicas that were deleted from the failed disk. The syntax of the command is show below. In this example, *diskset* is the name of the diskset containing the failed disk and *drive* is the DID name of the disk in the form d*N* (for new installations of Sun Cluster), or c*N*t*Y*d*Z* (for installations that upgraded from HA 1.3).

    ```
    phys-hahost1# metaset -s diskset -a drive
    ```

This operation can take up to fifteen minutes or more, depending on the size of your configuration and the number of disks.

13. **Use the `scadmin(1M)` command to reserve and enable failfast on the specified disk that has just been added back to the diskset.**

```
phys-hahost1# scadmin reserve c3t3d4
```

14. **Use the `format(1M)` or `fmthard(1M)` command to repartition the new disk.**

    Make sure that you partition the new disk exactly as the disk that was replaced. (Saving the disk format information was recommended in Chapter 1.)

15. **Use the `metainit(1M)` command to reinitialize disks that were cleared in Step 6 on page 235.**

```
phys-hahost1# metainit -s hahost1 d50
```

16. **Attach submirrors that were detached in Step 5 on page 235.**

    Use the `metattach(1M)` command to perform this step. See the `metattach(1M)` man page for details.

```
phys-hahost1# metattach -s hahost1 d40 d50
```

17. **Restore all hot spares that were deleted in Step 7 on page 235.**

    Use `metahs(1M)` to add back the hot spares. See the `metahs(1M)` man page for details.

```
phys-hahost1# metahs -s hahost1 -a hsp000 c3t2d5s0
```

18. **Verify that the replacement corrected the problem.**

```
phys-hahost1# metastat -s hahost1
```

## How to Replace a Sun StorEdge MultiPack or Sun StorEdge D1000 Disk (VxVM)

These are the high-level steps to replace a Sun StorEdge MultiPack or Sun StorEdge D1000 disk in a VxVM configuration:

- Removing the failed disk in the disk enclosure by using the `vxdiskadm` command
- Replacing the failed disk
- Replacing the disk removed earlier by using the `vxdiskadm` command

---

**Note -** For systems not running shared disk groups, master node refers to the node that has imported the disk group.

---

1. **If you are running shared disk groups, determine the master and slave node by entering the following command on all nodes in the cluster:**

```
# vxdctl -c mode
```

---

**Note -** Complete the following steps from the master node.

---

2. **Determine if the disk in question had failures and is in the** `NODEVICE` **state.**

   If this is not the case, skip to Step 8 on page 240.

3. **Run the** `vxdiskadm` **utility and enter** `4` **(Remove a disk for replacement).**

   This option removes a physical disk while retaining the disk name. The utility then queries you for the particular device that you want to replace.

4. **Enter the disk name or** `list`**.**

   The following example illustrates the removal of disk `c2t8d0`.

```
Enter disk name [<disk>,list,q,?] list

Disk group: rootdg

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

...

Disk group: demo

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

dm c1t2d0       c2t2d0s2     sliced   1519     4152640  -
dm c1t3d0       c2t3d0s2     sliced   1519     4152640  -
dm c1t4d0       c2t4d0s2     sliced   1519     4152640  -
dm c1t5d0       c2t5d0s2     sliced   1519     4152640  -
dm c1t8d0       c2t8d0s2     sliced   1519     4152640  -
dm c1t9d0       c2t9d0s2     sliced   1519     4152640  -
dm c2t2d0       c1t2d0s2     sliced   1519     4152640  -
dm c2t3d0       c1t3d0s2     sliced   1519     4152640  -
dm c2t4d0       c1t4d0s2     sliced   1519     4152640  -
```

**(continued)**

(Continuation)

```
dm c2t5d0       c1t5d0s2     sliced   1519     4152640  -
dm c2t8d0       c1t8d0s2     sliced   1519     4152640  -
dm c2t9d0       c1t9d0s2     sliced   1519     4152640  -

Enter disk name [<disk>,list,q,?] c2t8d0

  The requested operation is to remove disk c2t8d0 from disk group
  demo.  The disk name will be kept, along with any volumes using
  the disk, allowing replacement of the disk.

  Select "Replace a failed or removed disk" from the main menu
  when you wish to replace the disk.
```

5. **Enter** `y` **or press Return to continue.**

```
Continue with operation? [y,n,q,?] (default: y) y
  Removal of disk c2t8d0 completed successfully.
```

6. **Enter** `q` **to quit the utility.**

```
Remove another disk? [y,n,q,?] (default: n) q
```

7. **Enter** `vxdisk list` **and** `vxprint` **to view the changes.**

   The example disk `c2t8d0` is removed.

```
# vxdisk list
.
c2t3d0s2     sliced    c1t3d0       demo         online shared
c2t4d0s2     sliced    c1t4d0       demo         online shared
c2t5d0s2     sliced    c1t5d0       demo         online shared
c2t8d0s2     sliced    c1t8d0       demo         online shared
c2t9d0s2     sliced    c1t9d0       demo         online shared
-            -         c2t8d0       demo         removed
# vxprint
.
dm c2t3d0       c1t3d0s2     -        4152640  -        -        -       -
dm c2t4d0       c1t4d0s2     -        4152640  -        -        -       -
dm c2t5d0       c1t5d0s2     -        4152640  -        -        -       -
dm c2t8d0       -            -        -        -        REMOVED  -       -
dm c2t9d0       c1t9d0s2     -        4152640  -        -        -       -
```

(continued)

**(Continuation)**

```
pl demo05-02    -           DISABLED 51200    -        REMOVED  -       -
sd c2t8d0-1     demo05-02   DISABLED 51200    0        REMOVED  -       -
.
.
.
```

8. **Replace the physical drive without powering off any component.**

   For further information, refer to the documentation accompanying the disk enclosure unit.

   ---

   **Note -** As you replace the drive, you may see messages on the system console similar to those in the following example. Do not become alarmed as these messages may not indicate a problem. Instead, proceed with the replacement as described in the next steps.

```
Nov  3 17:44:00 updb10a unix: WARNING: /sbus@1f,0/SUNW,fas@0,8800000/sd@2,0 (sd17):
Nov  3 17:44:00 updb10a unix:   SCSI transport failed: reason "incomplete": \

retrying command
Nov  3 17:44:03 updb10a unix: WARNING: /sbus@1f,0/SUNW,fas@0,8800000/sd@2,0 (sd17):
Nov  3 17:44:03 updb10a unix:   disk not responding to selection
```

---

9. **Run the `vxdiskadm` utility and enter `5` (Replace a failed or removed disk).**

10. **Enter the disk name.**

    You can enter `list` to see a list of disks in the `REMOVED` state.

**Note -** The disk may appear in the NODEVICE state if it had failures.

```
Select a removed or failed disk [<disk>,list,q,?] list

Disk group: rootdg

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

...

Disk group: demo

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

dm c2t8d0       -            -        -        -        REMOVED


Select a removed or failed disk [<disk>,list,q,?] c2t8d0
```

The vxdiskadm utility detects the new device and asks you whether the new device should replace the removed device.

**Note -** If there are other unused disks attached to the system, vxdiskadm also presents these disks as viable choices.

11. **Enter the device name, or if the utility lists the device as the default, press Return.**

```
  The following devices are available as replacements:

        c1t8d0s2

  You can choose one of these disks to replace c2t8d0.
  Choose "none" to initialize another disk to replace c2t8d0.

Choose a device, or select "none"
[<device>,none,q,?] (default: c1t8d0s2) <Return>

 The requested operation is to use the initialized device c1t8d0s2
 to replace the removed or failed disk c2t8d0 in disk group demo.
```

12. **Enter** `y` **or press Return to verify that you want this device (in the example,** `c1t8d0s2`**) to be the replacement disk.**

```
Continue with operation? [y,n,q,?] (default: y) <Return>

  Replacement of disk c2t8d0 in group demo with disk device
  c1t8d0s2 completed successfully.
```

13. **Enter** `n` **or press Return to quit this utility.**

```
Replace another disk? [y,n,q,?] (default: n)  <Return>
```

14. **Enter** `vxdisk list` **and** `vxprint` **to see the changes.**

    The example, disk `c2t8d0`, is no longer in the `REMOVED` state.

```
# vxdisk list
...
c2t2d0s2     sliced    c1t2d0       demo         online shared
c2t3d0s2     sliced    c1t3d0       demo         online shared
c2t4d0s2     sliced    c1t4d0       demo         online shared
c2t5d0s2     sliced    c1t5d0       demo         online shared
c2t8d0s2     sliced    c1t8d0       demo         online shared
c2t9d0s2     sliced    c1t9d0       demo         online shared

# vxprint
...
dm c2t4d0       c1t4d0s2     -        4152640  -        -        -       -
dm c2t5d0       c1t5d0s2     -        4152640  -        -        -       -
dm c2t8d0       c1t8d0s2     -        4152640  -        -        -       -
dm c2t9d0       c1t9d0s2     -        4152640  -        -        -       -
...
```

## Replacing Sun StorEdge MultiPack or Sun StorEdge D1000 Enclosures

This section describes how to replace an entire Sun StorEdge MultiPack or Sun StorEdge D1000 enclosure running VxVM.

## ▼ How to Replace a Sun StorEdge MultiPack or Sun StorEdge D1000 Enclosure (VxVM)

These are the high-level steps for replacing an entire failed Sun StorEdge MultiPack or Sun StorEdge D1000 in a VxVM configuration:

- Removing all the disks in the defective disk enclosure by using the `vxdiskadm` command
- Replacing the failed disk enclosure
- Replacing all the disks removed earlier into the new disk enclosure by using the `vxdiskadm` command

---

**Note -** For systems not running shared disk groups, master node refers to the node that has imported the disk group.

---

1. **If you are running shared disk groups, determine the master and slave node by entering the following command on all nodes in the cluster:**

```
# vxdctl -c mode
```

---

**Note -** Complete the following steps from the master node.

---

2. **Remove all the disks on the failed disk enclosure by running the** `vxdiskadm` **utility and entering** `4` **(Remove a disk for replacement).**

---

**Note -** This option enables you to remove only one disk at a time. Repeat this procedure for each disk.

---

3. **Enter the** `list` **command.**

In the following example, assume that the disk enclosure on controller `c2` needs replacement. Based on the `list` output, the VxVM names for these disks are `c2t2d0`, `c2t3d0`, `c2t4d0`, `c2t5d0`, `c2t8d0`, and `c2t9d0.`

```
Remove a disk for replacement
Menu: VolumeManager/Disk/RemoveForReplace

  Use this menu operation to remove a physical disk from a disk
  group, while retaining the disk name.  This changes the state
  for the disk name to a "removed" disk.  If there are any
  initialized disks that are not part of a disk group, you will be
```

**(continued)**

**(Continuation)**

```
  given the option of using one of these disks as a replacement.

Enter disk name [<disk>,list,q,?] list
Disk group: rootdg

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

...

Disk group: demo

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

dm c1t2d0       c2t2d0s2     sliced   1519     4152640  -
dm c1t3d0       c2t3d0s2     sliced   1519     4152640  -
dm c1t4d0       c2t4d0s2     sliced   1519     4152640  -
dm c1t5d0       c2t5d0s2     sliced   1519     4152640  -
dm c1t8d0       c2t8d0s2     sliced   1519     4152640  -
dm c1t9d0       c2t9d0s2     sliced   1519     4152640  -
dm c2t2d0       c1t2d0s2     sliced   1519     4152640  -
dm c2t3d0       c1t3d0s2     sliced   1519     4152640  -
dm c2t4d0       c1t4d0s2     sliced   1519     4152640  -
dm c2t5d0       c1t5d0s2     sliced   1519     4152640  -
dm c2t8d0       c1t8d0s2     sliced   1519     4152640  -
dm c2t9d0       c1t9d0s2     sliced   1519     4152640  -
```

4. **Enter the disk name (in this example,** `c2t2d0`**).**

```
Enter disk name [<disk>,list,q,?] c2t2d0


  The following volumes will lose mirrors as a result of this
  operation:

        demo-1

  No data on these volumes will be lost.

  The requested operation is to remove disk c2t2d0 from disk group
  demo.  The disk name will be kept, along with any volumes using
  the disk, allowing replacement of the disk.

  Select "Replace a failed or removed disk" from the main menu
  when you wish to replace the disk.
```

5. **Enter** `y` **or press Return to verify that you want to replace the disk.**

```
Continue with operation? [y,n,q,?] (default: y) <Return>

  Removal of disk c2t2d0 completed successfully.
```

6. **Enter** `y` **to continue.**

```
Remove another disk? [y,n,q,?] (default: n) y

Remove a disk for replacement
Menu: VolumeManager/Disk/RemoveForReplace

  Use this menu operation to remove a physical disk from a disk
  group, while retaining the disk name.  This changes the state
  for the disk name to a "removed" disk.  If there are any
  initialized disks that are not part of a disk group, you will be
  given the option of using one of these disks as a replacement.
```

7. **Enter the next example disk name,** `c2t3d0`**.**

```
Enter disk name [<disk>,list,q,?] c2t3d0

  The following volumes will lose mirrors as a result of this
  operation:

        demo-2

  No data on these volumes will be lost.

The following devices are available as replacements:

        c1t2d0

  You can choose one of these disks now, to replace c2t3d0.
  Select "none" if you do not wish to select a replacement disk.
```

8. **Enter** `none`**, if necessary.**

**Note -** This query arises whenever the utility recognizes a good disk in the system. If there are no good disks, you will not see this query.

```
Choose a device, or select "none"
[<device>,none,q,?] (default: c1t2d0) none
```

9. **Enter** `y` **or press Return to verify that you want to remove the disk.**

```
The requested operation is to remove disk c2t3d0 from disk group
  demo.  The disk name will be kept, along with any volumes using
  the disk, allowing replacement of the disk.

  Select "Replace a failed or removed disk" from the main menu
  when you wish to replace the disk.

Continue with operation? [y,n,q,?] (default: y) <Return>

  Removal of disk c2t3d0 completed successfully.
```

10. **Repeat Step 6 on page 245 through Step 9 on page 246 for each disk you identified in Step 3 on page 243.**

11. **Power off and replace the disk enclosure.**

    For more information, refer to the disk enclosure documentation.

**Note -** As you replace the disk enclosure, you may see messages on the system console similar to those in the following example. Do not become alarmed, as these messages may not indicate a problem. Instead, proceed with the replacement as described in the next section.

```
Nov  3 17:44:00 updb10a unix: WARNING: /sbus@1f,0/SUNW,fas@0,8800000/sd@2,0 (sd17):
Nov  3 17:44:00 updb10a unix:   SCSI transport failed: reason "incomplete": \

retrying command
Nov  3 17:44:03 updb10a unix: WARNING: /sbus@1f,0/SUNW,fas@0,8800000/sd@2,0 (sd17):
Nov  3 17:44:03 updb10a unix:   disk not responding to selection
```

12. **Power on the disk enclosure.**

    For more information, refer to your disk enclosure service manual.

13. **Attach all the disks removed earlier by running the** `vxdiskadm` **utility and entering** `5` **(Replace a failed or removed disk).**

**Note -** This option enables you to replace only one disk at a time. Repeat this procedure for each disk.

14. **Enter the** `list` **command to see a list of disk names now in the** `REMOVED` **state.**

```
Replace a failed or removed disk
Menu: VolumeManager/Disk/ReplaceDisk

 Use this menu operation to specify a replacement disk for a disk
 that you removed with the "Remove a disk for replacement" menu
 operation, or that failed during use.  You will be prompted for
 a disk name to replace and a disk device to use as a replacement.
 You can choose an uninitialized disk, in which case the disk will
 be initialized, or you can choose a disk that you have already
 initialized using the Add or initialize a disk menu operation.

Select a removed or failed disk [<disk>,list,q,?] list

Disk group: rootdg

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

...
```

**(continued)**

**(Continuation)**

```


Disk group: demo

DM NAME         DEVICE       TYPE     PRIVLEN  PUBLEN   STATE

dm c2t2d0       -            -        -        -        REMOVED
dm c2t3d0       -            -        -        -        REMOVED
dm c2t4d0       -            -        -        -        REMOVED
dm c2t5d0       -            -        -        -        REMOVED
dm c2t8d0       -            -        -        -        REMOVED
dm c2t9d0       -            -        -        -        REMOVED
```

15. **Enter the disk name (in this example, `c2t2d0`).**

```
Select a removed or failed disk [<disk>,list,q,?] c2t2d0

  The following devices are available as replacements:

        c1t2d0s2 c1t3d0s2 c1t4d0s2 c1t5d0s2 c1t8d0s2 c1t9d0s2
```

The `vxdiskadm` utility detects the new devices and asks you whether the new devices should replace the removed devices.

16. **Enter the “replacement” or “new” device name, or if the utility lists the device as the default, press Return.**

```
  You can choose one of these disks to replace c2t2d0.
  Choose "none" to initialize another disk to replace c2t2d0.

Choose a device, or select "none"
[<device>,none,q,?] (default: c1t2d0s2) <Return>
```

17. **Enter `y` or press Return to verify that you want this device (in the example, `c1t2d0s2`) to be the replacement disk.**

```
 The requested operation is to use the initialized device c1t2d0s2
  to replace the removed or failed disk c2t2d0 in disk group demo.

Continue with operation? [y,n,q,?] (default: y) <Return>

  Replacement of disk c2t2d0 in group demo with disk device
  c1t2d0s2 completed successfully.
```

**18. Enter `y` to continue.**

```
Replace another disk? [y,n,q,?] (default: n) y
```

Repeat Step 15 on page 248 through Step 18 on page 249 for each of the `REMOVED/NODEVICE` disk names.

CHAPTER 13

# Administering Sun StorEdge A3x00s and Sun StorEdge A1000s

This chapter provides instructions for administering the Sun StorEdge A3x00 and Sun StorEdge A1000 expansion units and disks included in your Sun Cluster configuration. Both expansion units support RAID5 hardware. The Sun StorEdge A3x00 has two RAID5 controllers, and the Sun StorEdge A1000 has only one. Most administration procedures are the same for both expansion units.

Use the service manual for your Sun StorEdge A3x00 or Sun StorEdge A1000 disks, and your volume management software documentation, when you are replacing or repairing disk hardware in the Sun Cluster configuration.

## Power Sources

The Sun StorEdge A3x00 or Sun StorEdge A1000 disk expansion unit includes redundant power sequencers. Each power sequencer supplies power to half of the expansion unit's components, so power loss to one of these power sources does not affect system availability. The controller module has redundant power supplies, so a single power failure in the controller module also will not affect system availability.

There are no special procedures required by Sun Cluster to recover in the event of a failure of the entire disk expansion unit. Follow the procedures in the disk expansion service manual to bring them back into service.

# Adding Sun StorEdge A3x00 or Sun StorEdge A1000 Enclosures

This section describes procedures used to add a Sun StorEdge A3x00 or a Sun StorEdge A1000 expansion unit. These procedures can be done with the Sun Cluster up and running.

**Note -** To upgrade the firmware in expansion units, refer to the hardware service manual for your system.

## ▼ How to Add a Disk Enclosure

1. **Switch over all logical hosts to one of the Sun Cluster nodes in the cluster that will be receiving the new disk enclosure.**

```
phys-hahost1# haswitch phys-hahost2 hahost1 hahost2
```

2. **Stop Sun Cluster on the node that no longer masters any logical hosts.**

```
phys-hahost1# scadmin stopnode
```

3. **Add the disk interface card (UDWIS) to the Sun Cluster node.**

   Use the instructions that come with the UDWIS interface card.

4. **Cable the new UDWIS card to the disk enclosure.**

5. **Perform a reconfiguration reboot on the node with the new UDWIS card.**

```
phys-hahost1# boot -r
```

6. **Start Sun Cluster on the node.**

```
phys-hahost1# scadmin startnode
```

7. **Switch over the logical hosts to another node in the cluster.**

```
phys-hahost1# haswitch phys-hahost1 hahost1 hahost2
```

8. **Perform Step 2 on page 252 through Step 6 on page 252 on the node that no longer masters the logical hosts.**

9. **Configure the disk subsystems into the cluster.**

   Use the expansion unit and VxVM documentation to set up the disk configuration.

10. **Run the** `haswitch(1M)` **command on both nodes to incorporate the new configuration into Sun Cluster.**

```
# haswitch -r
```

## Administering Sun StorEdge A3x00 or Sun StorEdge A1000 Disks

The procedures for administering Sun StorEdge A3x00 or Sun StorEdge A1000 disks in a Sun Cluster configuration are identical to those used with nodes that are not clustered. Refer to your expansion unit documentation for procedures to add, replace, or repair disks or disk components in your disk expansion unit.

### Adding a Sun StorEdge A3x00 or Sun StorEdge A1000 Disk

When adding drives to a Sun StorEdge A3x00 or a Sun StorEdge A1000, add the drives while the system is up and running. Do not reboot the system. Doing so may cause a loss of configuration information on the new drives, and a loss of data and logical unit (LUN) configuration on the existing drives.

If you do see a problem, such as not seeing pre-existing LUNs after adding drives and rebooting the system, then remove the newly-added drives, restart the system, then add them one at a time.

This problem does not occur when the drives are added to an “unused” disk group on a running system, after which it is safe to reboot the system. Because hot-plugging is fully supported on the Sun StorEdge A3x00 and the Sun StorEdge A1000, this is the accepted procedure.

## Replacing a Sun StorEdge A3x00 or Sun StorEdge A1000 Disk

The Sun Cluster sees Sun StorEdge A3x00 or Sun StorEdge A1000 disks as logical units (LUNs) and not as physical disks. Because of this:

- As long as the LUN is available, no action is necessary when replacing a failed physical disk.
- If the LUN is not available (or not in an optimal state) and was used as a quorum device, you need to use the `scconf -q` command to change the quorum device to a different LUN (disk) before proceeding with the disk replacement procedure.

For additional information regarding logical units and physical disks, refer to your expansion unit documentation. See also Chapter 3 and Chapter 9 in the *Sun Cluster 2.2 Hardware Service Manual.*

Note that some administrative procedures on the Sun StorEdge A3x00 or Sun StorEdge A1000s require replacement of UDWIS cards. For this procedure, see Chapter 12 in the *Sun Cluster 2.2 Hardware Service Manual.*

CHAPTER 14

# Administering Sun StorEdge A5000s

This chapter provides instructions for administering Sun StorEdge A5000 disks.

- "Recovering From Power Loss" on page 255
- "Administering Sun StorEdge A5000s " on page 259
- "Administering Sun StorEdge A5000 Disks" on page 261

Use the service manual for your Sun StorEdge A5000 disks and your volume management software documentation when you are replacing or repairing disk hardware in the Sun Cluster configuration.

## Recovering From Power Loss

When power is lost to one Sun StorEdge A5000, I/O operations generate errors that are detected by your volume management software. Errors are not reported until I/O transactions are made to the disk.

You should monitor the configuration for these events using the commands described in Chapter 2.

### ▼ How to Recover From Power Loss (Solstice DiskSuite)

These are the high-level steps to recover from power loss to a disk enclosure in a Solstice DiskSuite environment:

- Identifying the errored replicas

- Returning the errored replicas to service
- Identifying the errored devices
- Returning the errored devices to service
- Resyncing the disks

These are the detailed steps to recover from power loss to a disk enclosure in a Solstice DiskSuite environment.

1. **When power is restored, use the** `metadb(1M)` **command to identify the errored replicas:**

```
# metadb -s diskset
```

2. **Return replicas to service.**

   After the loss of power, all metadevice state database replicas on the affected disk enclosure chassis enter an errored state. Because metadevice state database replica recovery is not automatic, it is safest to perform the recovery immediately after the disk enclosure returns to service. Otherwise, a new failure can cause a majority of replicas to be out of service and cause a kernel panic. This is the expected behavior of Solstice DiskSuite when too few replicas are available.

   While these errored replicas will be reclaimed at the next takeover (`haswitch(1M)` or `reboot(1M)`), you might want to return them to service manually by first deleting and then adding them back.

   **Note -** Make sure that you add back the same number of replicas that were deleted on each slice. You can delete multiple replicas with a single `metadb(1M)` command. If you need multiple copies of replicas on one slice, you must add them in one invocation of the `metadb(1M)` command using the `-c` flag.

3. **Use the** `metastat(1M)` **command to identify the errored metadevices.**

```
# metastat -s diskset
```

4. **Return errored metadevices to service using the** `metareplace(1M)` **command, and resync the disks.**

```
# metareplace -s diskset -e mirror component
```

   The `-e` option transitions the component (slice) to the available state and performs a resync.

Components that have been replaced by a hot spare should be the last devices replaced using the metareplace(1M) command. If the hot spare is replaced first, it could replace another errored submirror as soon as it becomes available.

You can perform a resync on only one component of a submirror (metadevice) at a time. If all components of a submirror were affected by the power outage, each component must be replaced separately. It takes approximately 10 minutes to resync a 1.05GB disk.

If both disksets in a symmetric configuration were affected by the power outage, you can resync each diskset's affected submirrors concurrently. Log into each host separately to recover that host's diskset by running metareplace(1M) on each.

---

**Note -** Depending on the number of submirrors and the number of components in these submirrors, the resync actions can require a considerable amount of time. A single submirror made up of 30 1.05GB drives might take about five hours to complete. A more manageable configuration made up of five component submirrors might take only 50 minutes to complete.

---

## ▼ How to Recover From Power Loss (VxVM)

Power failures can detach disk drives and cause plexes to become detached, and thus, unavailable. The volume remains active, however, because the remaining plexes in a mirrored volume are still available. It is possible to reattach the disk drives and recover from this condition without halting nodes in the cluster.

These are the high-level steps to recover from power loss to a disk enclosure in an VxVM configuration:

- Determining the errored plex(es) by using the vxprint and vxdisk commands
- Fixing the problem that caused the power loss
- Using the drvconfig and disks commands to create the /devices and /dev entries
- Scanning the current disk configuration
- Reattaching disks that had transient errors
- Verifying there are no more errors
- (Optional) For shared disk groups, running the vxdg command for each disk that was powered off
- Starting volume recovery

These are the detailed steps to recover from power loss to a disk enclosure in an VxVM configuration.

1. **Use the vxprint command to view the errored plexes.**

Optionally, specify a disk group with the `-g` *diskgroup* option.

2. **Use the `vxdisk` command to identify the errored disks.**

```
# vxdisk list
DEVICE       TYPE      DISK         GROUP        STATUS
..
-            -         c1t5d0       toi          failed was:c1t5d0s2
...
```

3. **Fix the condition that resulted in the problem so that power is restored to all failed disks.**

   Be sure that the disks are spun up before proceeding.

4. **Enter the following commands on all nodes in the cluster.**

   In some cases, the drive(s) must be rediscovered by the node(s).

```
# drvconfig
# disks
```

5. **Enter the following commands on all nodes in the cluster.**

   The volume manager must scan the current disk configuration again.

```
# vxdctl enable
# vxdisk -a online
```

6. **Enter the following command first on the master node, then on the remaining nodes in the cluster.**

   This will reattach disks that had transitory failures.

```
# vxreattach
```

7. **Verify the output of the `vxdisk` command to see if there are any more errors.**

```
# vxdisk list
```

8. **If media was replaced, from the master node enter the following command for each disk that has been disconnected.**

   The physical disk and the volume manager access name for that disk must be reconnected.

```
# vxdg -g diskgroup -k adddisk medianame=accessname
```

   The values for `medianame` and `accessname` appear at the end of the `vxdisk list` command output.

   For example:

```
# vxdg -g toi -k adddisk c1t5d0=c1t5d0s2
# vxdg -g toi -k adddisk c1t5d1=c1t5d1s2
# vxdg -g toi -k adddisk c1t5d2=c1t5d2s2
# vxdg -g toi -k adddisk c1t5d3=c1t5d3s2
# vxdg -g toi -k adddisk c1t5d4=c1t5d4s2
```

   You can also use the `vxdiskadm` command, or the graphical user interface, to reattach the disks.

9. **From the node, start volume recovery.**

   If you have shared disk groups, use the `-svc` options to the `vxrecover` command.

```
# vxrecover -bv [-g diskgroup]
```

10. **(Optional) Use the** `vxprint -g` **command to view the changes.**

# Administering Sun StorEdge A5000s

This section describes procedures for administering Sun StorEdge A5000 components. Use the procedures described in your server hardware manual to identify the failed component.

# Repairing a Lost Sun StorEdge A5000 Connection

When a connection from a disk enclosure to one of the cluster nodes fails, the failure is probably due to a bad SCSI-2 cable or an SBus card.

In any event, the node on which the failure occurred will begin generating errors when the failure is discovered. Later accesses to the disk enclosure will generate additional errors. The node will exhibit the same behavior as though power had been lost to the disk enclosure. I/O operations from the other nodes in the cluster are unaffected by this type of failure.

To diagnose the failure, use the procedures for testing the card module in the service manual for your Sun Cluster node to determine which component failed. You should free up one node and the disk enclosure that appears to be down, for hardware debugging.

## ▼ How to Repair a Lost Sun StorEdge A5000 Connection

1. **Prepare the Sun Cluster system for component replacement.**

   Depending on the cause of the connection loss, prepare the Sun Cluster node with one of the following procedures.

   - If the failed component is an SBus FC-100 host adapter, see Chapter 7, to prepare the Sun Cluster node for power down.
   - If the problem is a bad FC-100 fiber optic cable, the volume management software will have detected the problem and prepared the system for cable replacement.

2. **Replace the failed component.**

   If the FC-100 fiber optic cable or SBus FC-100 host adapter fails, refer to the *Sun StorEdge A5000 Installation and Service Manual* for detailed instructions on replacing them.

3. **Recover from volume management software errors.**

   Use the procedures described in “Recovering From Power Loss” on page 255.

This completes the procedure for repairing a lost connection.

# Administering Sun StorEdge A5000 Disks

This section describes how to add and replace Sun StorEdge A5000 disks in a Sun Cluster configuration.

## Adding or Replacing Sun StorEdge A5000 Disks

When adding or replacing Sun StorEdge A5000 disks, be sure to refer to the documentation that came with your system for more information.

**Note -** When replacing a failed A5000 disk under VxVM control, you cannot simply pull out the disk and replace it with a new one. This is because each disk has a unique World Wide Name (WWN). See “Replacing a SPARCstorage Array Controller and Changing the World Wide Name” on page 184, for more information about the WWN.

### ▼ How to Add a Sun StorEdge A5000 Disk (Solstice DiskSuite)

1. **Use the `luxadm` command to insert the new disk.**

   Physically install the new disk or disks when prompted. Repeat for each node that is physically connected to the array.

```
# luxadm insert enclosure.slot
```

2. **Insert the new disk drive and enter Return.**

3. **Use format to create a disk label and repartition, if needed.**

4. **Use `scdidadm(1M)` to discover the new disk and create a DID instance for it.**

   This should be run from node1 only. See the `scdidadm(1M)` man page for details.

```
# scdidadm -r -H node2,node3...
```

This completes the disk addition procedure.

## ▼ How to Add a Sun StorEdge A5000 Disk (VxVM)

1. **Use the** `luxadm` **command to prepare the loop for a new device.**

   Physically install the new disk or disks when prompted.

```
# luxadm insert
```

2. **Notify VxVM of the new disk.**

```
# vxdctl enable
```

3. **Use the** `vxdiskadm` **command to bring the new disk(s) into VxVM control.**

   Enter **1** (Add or initialize one or more disks).

This completes the disk addition procedure.

## ▼ How to Replace a Sun StorEdge A5000 Disk (Solstice DiskSuite)

1. **Identify all metadevices or applications using the failing disk.**

   If the metadevices are mirrored or RAID5, the disk can be replaced without stopping the metadevices. Otherwise all I/O to the disk must be stopped using the appropriate commands. For example, use the `umount(1M)` command to unmount a file system on a stripe or concatenation.

2. **Preserve the disk label, if necessary.**

   For example:

```
# prvtoc /dev/rdsk/c1t3d0s2 > /tmp/c1t3d0.vtoc
```

3. **(optional) Use** `metareplace` **to replace the disk slices if the disk has not been hot-spared.**

   For example:

```
# metareplace d1 c1t3d0s2 c1t2d0s2
d1: device c1t3d0s2 is replaced with c1t2d0s2
```

4. **Use `luxadm -F` to remove the disk.**

   The `-F` option is required because Solstice DiskSuite does not offline disks. Repeat the command for all hosts, if the disk is multihosted. For example:

```
# luxadm remove -F /dev/rdsk/c1t3d0s2
WARNING!!! Please ensure that no filesystems are mounted on these device(s).  All data on these devices should have been

1: Box Name ‘‘macs1’’ rear slot 1

Please enter ‘q’ to Quit or <Return> to Continue: stopping:  Drive in ‘‘macs1’’ rear  slot 1....Done

offlining: Drive in ‘‘macs1’’ rear  slot 1....Done

Hit <Return> after removing the device(s).
```

---

**Note -** The FPM icon for the disk drive to be removed should be blinking. The amber LED under the disk drive should also be blinking.

---

5. **Remove the disk drive and enter Return.**

   The output should look similar to the following:

```
Hit <Return> after removing the device(s).

Drive in Box Name ‘‘macs1’’ rear slot 1

Removing Logical Nodes:

Removing c1t3d0s0 Removing c1t3d0s1 Removing c1t3d0s2 Removing c1t3d0s3 Removing c1t3d0s4 Removing c1t3d0s5 Removing c1t
#
```

6. **Repeat Step 4 on page 263 for all nodes, if the disk array is in a multi-host configuration.**

7. **Use the `luxadm insert` command to insert the new disk.**

Repeat for all nodes. The output should be similar to the following:

```
# luxadm insert macs1,r1
The list of devices which will be inserted is:

1: Box Name ‘‘macs1’’ rear slot 1

Please enter ‘q’ to Quit or <Return> to Continue: Hit <Return> after inserting the device(s).
```

8. **Insert the disk drive and enter Return.**

   The output should be similar to the following:

```
Hit <Return> after inserting the device(s).  Drive in Box Name ‘‘macs1’’ rear slot 1  Logical Nodes under /dev/dsk and
rdsk : c1t3d0s0 c1t3d0s1 c1t3d0s2 c1t3d0s3 c1t3d0s4 c1t3d0s5 c1t3d0s6 c1t3d0s7 c2t3d0s0 c2t3d0s1 c2t3d0s2 c2t3d0s3 c2t3
```

**Note -** The FPM icon for the disk drive you replaced should be lit. In addition, the green LED under the disk drive should be blinking.

9. **Use** `scdidadm(1M)` **to update the DID pseudo device information.**

   On all nodes connected to the disk, execute the following command to update new Disk ID information.

```
# scdidadm -R DID_instance
```

where *DID_instance* is the instance number of the disk that was replaced. Refer to the `scdidadm(1M)` man page for more information.

10. **Reboot all nodes connected to the new disk.**

    To avoid down time, use the `haswitch(1M)` command to switch ownership of all logical hosts that can be mastered by the node to be rebooted. For example,

```
# haswitch phys-hahost2 hahost1 hahost2
```

11. **Label the disk, if necessary.**

    For example:

```
# cat /tmp/c1t3d0.vtoc | fmthard -s - /dev/rdsk/c1t3d0s2
fmthard:  New volume table of contents now in place.
```

12. **Replace the `metadb`, if necessary.**

    For example:

```
# metadb -d c1t3d0s0; metadb -a c1t3d0s0
```

13. **Enable the new disk slices with `metareplace -e`.**

    For example:

```
# metareplace -e d0 c1t3d0s0
d0: device c1t3d0s0 is enabled
```

This completes the disk replacement procedure.

## ▼ How to Replace a Sun StorEdge A5000 Disk (VxVM)

1. **Identify all volumes or applications using the failing disk.**

   If the volumes are mirrored or RAID5, the disk can be replaced without stopping the volume. Otherwise all I/O to the disk must be stopped using the appropriate commands. For example, use the `umount(1M)` command to unmount a file system on a stripe or concatenation.

2. **Use the `vxdiskadm` command to replace and offline a disk device.**

   For VxVM, perform these commands on the machine mastering the logical host owning the disk group.

   Enter **4** (Remove a disk for replacement), then enter **11** (Disable [offline] a disk device).

   You can use the graphical user interface instead, if you prefer it.

3. **Use the `luxadm` command to remove the device and device nodes.**

   This command is interactive and will prompt you to physically remove the disk. Perform the command on every node connected to the array. For example:

```
# luxadm remove_device -F /dev/rdsk/c2t20d0s2
```

4. **Physically replace the disk, then use the `luxadm` command to insert the new disk.**

   This creates the new device and device nodes. Perform the command on every node connected to the array. For example:

```
# luxadm insert_device ratbert,r4
```

5. **Notify the volume manager of the new disk.**

```
# vxdctl enable
```

6. **Use the `vxdiskadm` command to bring the new disk under VxVM control.**

   Enter **5** (Replace a failed or removed disk).

7. **(Optional) The volume can now be restored if necessary.**

This completes the disk replacement procedure.

APPENDIX **A**

# Administering Volume Managers

This appendix provides instructions for administering Solstice DiskSuite disksets and metadevices, and for administering VERITAS Volume Manager objects. The procedures documented in this appendix are dependent on your volume management software.

## Using Solstice DiskSuite in the Sun Cluster Environment

This section describes using DiskSuite to administer:

- Disksets
- Disks in disksets
- Multi-host metadevices
- Local metadevices

Refer to the Solstice DiskSuite documentation for a complete discussion of administering DiskSuite objects.

# Metadevice and Diskset Administration

Metadevices and disksets are created and administered using either Solstice DiskSuite command-line utilities or the DiskSuite Tool (`metatool(1M)`) graphical user interface.

Read the information in this chapter before using the Solstice DiskSuite documentation to administer disksets and metadevices in a Sun Cluster configuration.

Disksets are groups of disks. The primary administration task that you perform on disksets involves adding and removing disks.

Before using a disk that you have placed in a diskset, you must set up a metadevice using the disk's slices. A metadevice can be a concatenation, stripe, mirror, or UFS logging device (also called a trans device). You can also create hot spare pools that contain slices to serve as replacements when a metadevice is errored.

---

**Note -** Metadevice names begin with `d` and are followed by a number. By default in a Sun Cluster configuration, there are 128 unique metadevices in the range 0 to 127. Each UFS logging device that you create will use at least seven metadevice names. Therefore, in a large Sun Cluster configuration, you might need more than the 128 default metadevice names. For instructions on changing the default quantity, refer to the Solstice DiskSuite documentation. Hot spare pool names begin with `hsp` and are followed by a number. You can have up to 1,000 hot spare pools ranging from `hsp000` to `hsp999`.

---

## About Disksets

This section provides overview information on disksets and their relationship to logical hosts, and procedures on how to add and remove disks from the diskset associated with the logical host.

Sun Cluster logical hosts are mastered by physical hosts. Only the physical host that currently masters a logical host can access the logical host's diskset. When a physical host masters a logical host's diskset, it is said to have ownership of the diskset. In general, Sun Cluster takes care of diskset ownership. However, if the logical host is in maintenance state, as reported by the `hastat(1M)` command, you can use the DiskSuite `metaset -t` command to manually take diskset ownership. Before returning the logical host to service, release diskset ownership with the `metaset -r` command.

---

**Note -** If the logical hosts are up and running, you should never perform diskset administration using either the `-t` (take ownership) or `-r` (release ownership) options of the `metaset(1M)` command. These options are used internally by the Sun Cluster software and must be coordinated between the cluster nodes.

---

## Adding a Disk to a Diskset

If the disk being added to a diskset will be used as a submirror, you must have two disks available on two different multihost disk expansion units to allow for mirroring. However, if the disk will be used as a hot spare, you can add a single disk.

## ▼ How to Add a Disk to a Diskset (Solstice DiskSuite)

1. **Ensure that no data is on the disk.**

   This is important because the partition table will be rewritten and space for a metadevice state database replica will be allocated on the disk.

2. **Insert the disk device into the multihost disk expansion unit.**

   Use the instructions in the hardware documentation for your disk expansion unit for information on disk addition and removal procedures.

3. **Add the disk to a diskset.**

   The syntax for the command is shown below. In this example, *diskset* is the name of the diskset to which the disk is to be added, and *drive* is the DID name of the disk in the form `d`*N* (for new installations of Sun Cluster), or `c`*N*`t`*Y*`d`*Z* (for installations that upgraded from HA 1.3).

   ```
   # metaset -s diskset -a drive
   ```

4. **After adding the disks to the diskset by using the `metaset(1M)` command, use the `scadmin(1M)` command to reserve and enable failfast on the specified disks.**

```
phys-hahost1# scadmin reserve drivename
```

## Removing a Disk From a Diskset

You can remove a disk from a diskset at any time, as long as none of the slices on the disk are currently in use in metadevices or hot spare pools.

## ▼ How to Remove a Disk From a Diskset (Solstice DiskSuite)

1. **Use the** `metastat(1M)` **command to ensure that none of the slices are in use as metadevices or as hot spares.**

2. **Use the** `metaset(1M)` **command to remove the target disk from the diskset.**

   The syntax for the command is shown below. In this example, *diskset* is the name of the diskset containing the (failed) disk to be removed, and *drive* is the DID name of the disk in the form `d`*N* (for new installations of Sun Cluster), or `c`*N*`t`*Y*`d`*Z* (for installations that upgraded from HA 1.3).

   ```
   # metaset -s diskset -d drive
   ```

   This operation can take up to fifteen minutes or more, depending on the size of your configuration and the number of disks.

# Administering Multihost Metadevices

The following sections contain information about the differences between administering metadevices in the multihost Sun Cluster environment and in a single-host environment.

Unless noted in the following sections, you can use the instructions in the Solstice DiskSuite documentation.

**Note -** The instructions in the Solstice DiskSuite books are relevant only for single-host configurations.

The following sections describe the Solstice DiskSuite command-line programs to use when performing a task. Optionally, you can use the `metatool(1M)` graphical user interface for all the tasks unless directed otherwise. Use the `-s` option when running `metatool(1M)`, because it allows you to specify the diskset name.

## Managing Metadevices

For ongoing management of metadevices, you must constantly monitor the metadevices for errors in operation, as discussed in “Monitoring Utilities” on page 27.

When `hastat(1M)` reports a problem with a diskset, use the `metastat(1M)` command to locate the errored metadevice.

You must use the `-s` option when running either `metastat(1M)` or `metatool(1M)`, so that you can specify the diskset name.

**Note -** You should save the metadevice configuration information when you make changes to the configuration. Use `metastat -p` to create output similar to what is in the `md.tab` file and then save the output. Refer to “Saving Disk Partition Information (Solstice DiskSuite)” on page 21, for details on saving partitioning data.

## Adding a Mirror to a Diskset

Mirrored metadevices can be used as part of a logging UFS file system for Sun Cluster highly available applications.

Idle slices on disks within a diskset can be configured into metadevices by using the `metainit(1M)` command.

## Removing a Mirror From a Diskset

Sun Cluster highly available database applications can use raw mirrored metadevices for database storage. While these are not mentioned in the `dfstab.`*logicalhost* file or in the `vfstab` file for each logical host, they appear in the related Sun Cluster database configuration files. The mirror must be removed from these files, and the Sun Cluster database system must stop using the mirror. Then the mirror can be deleted by using the `metaclear(1M)` command.

## Taking Submirrors Offline

If you are using SPARCstorage Arrays, note that before replacing or adding a disk drive in a SPARCstorage Array tray, all metadevices on that tray must be taken offline.

In symmetric configurations, taking the submirrors offline for maintenance is complex because disks from each of the two disksets might be in the same tray in the SPARCstorage Array. You must take the metadevices from each diskset offline before removing the tray.

Use the `metaoffline(1M)` command to take offline all submirrors on every disk in the tray.

## Creating New Metadevices

After a disk is added to a diskset, create new metadevices using `metainit(1M)` or `metatool(1M)`. If the new devices will be hot spares, use the `metahs(1M)` command to place the hot spares in a hot spare pool.

## Replacing Errored Components

When replacing an errored metadevice component, use the `metareplace(1M)` command.

A replacement slice (or disk) must be available. This could be an existing device that is not in use, or a new device that you have added to the diskset.

You also can return to service drives that have sustained transient errors (for example, as a result of a chassis power failure) by using the `metareplace -e` command.

## Deleting Metadevices

Before deleting a metadevice, verify that none of the components in the metadevice is in use by Sun Cluster HA for NFS. Then use the `metaclear(1M)` command to delete the metadevice.

## Growing Metadevices

To grow a metadevice, you must have a least two slices (disks) in different multihost disk expansion units available. Each of the two new slices should be added to a different submirror with the `metainit(1M)` command. You then use the `growfs(1M)` command to grow the file system.

**Caution -** When the `growfs(1M)` command is running, clients might experience interruptions of service.

If a takeover occurs while the file system is growing, the file system will not be grown. You must reissue the `growfs(1M)` command after the takeover completes.

**Note -** The file system that contains /*logicalhost*/`statmon` cannot be grown. Because the `statd(1M)` program modifies this directory, it would be blocked for extended periods while the file system is growing. This would have unpredictable effects on the network file locking protocol. This is a problem only for configurations using Sun Cluster HA for NFS.

## Managing Hot Spare Pools

You can add or delete hot spare devices to or from hot spare pools at any time, as long as they are not in use. In addition, you can create new hot spare pools and associate them with submirrors using the `metahs(1M)` command.

## Managing UFS Logs

All UFS logs on multihost disks are mirrored. When a submirror fails, it is reported as an errored component. Repair the failure using either `metareplace(1M)` or `metatool(1M)`.

If the entire mirror that contains the UFS log fails, you must unmount the file system, back up any accessible data, repair the error, repair the file system (using `fsck(1M)`), and remount the file system.

## Adding UFS Logging to a Logical Host

All UFS file systems within a logical host must be logging UFS file systems to ensure that the failover or `haswitch(1M)` timeout criteria can be met. This facilitates fast switchovers and takeovers.

The logging UFS file system is set up by creating a trans device with a mirrored logging device and a mirrored UFS master file system. Both the logging device and UFS master device must be mirrored.

Typically, Slice 6 of each drive in a diskset can be used as a UFS log. The slices can be used for UFS log submirrors. If the slices are smaller than the log size you want, several can be concatenated. Typically, one Mbyte per 100 Mbytes is adequate for UFS logs, up to a maximum of 64 Mbytes. Ideally, log slices should be drive-disjoint from the UFS master device.

---

**Note -** If you must repartition the disk to gain space for UFS logs, then preserve the existing Slice 7, which starts on Cylinder 0 and contains at least two Mbytes. This space is required and reserved for metadevice state database replicas. The `Tag` and `Flag` fields (as reported by the `format(1M)` command) must be preserved for Slice 7. The `metaset(1M)` command sets the `Tag` and `Flag` fields correctly when the initial configuration is built.

---

After the trans device has been configured, create the UFS file system using `newfs(1M)` on the trans device.

After the `newfs` process is completed, add the UFS file system to the `vfstab` file for the logical host, by editing the `/etc/opt/SUNWcluster/conf/hanfs/vfstab.`*logicalhost* file to update the administrative and multihost UFS file system information.

Make sure that the `vfstab`.*logicalhost* files of all cluster nodes contain the same information. Use the `cconsole(1)` facility to make simultaneous edits to `vfstab`.*logicalhost* files on all nodes in the cluster.

Here's a sample `vfstab`.*logicalhost* file showing the administrative file system and four other UFS file systems:

```
#device                 device                   mount       FS   fsck  mount mount
#to mount                to fsck                    point       type pass  all   options#
/dev/md/hahost1/dsk/d11  /dev/md/hahost1/rdsk/d11 /hahost1    ufs  1     no   -
/dev/md/hahost1/dsk/d1   /dev/md/hahost1/rdsk/d1  /hahost1/1  ufs  1     no   -
/dev/md/hahost1/dsk/d2   /dev/md/hahost1/rdsk/d2  /hahost1/2  ufs  1     no   -
/dev/md/hahost1/dsk/d3   /dev/md/hahostt1/rdsk/d3 /hahost1/3  ufs  1     no   -
/dev/md/hahost1/dsk/d4   /dev/md/hahost1/rdsk/d4  /hahost1/4  ufs  1     no   -
```

If the file system will be shared by Sun Cluster HA for NFS, follow the procedure for sharing NFS file systems as described in Chapter 11 in the *Sun Cluster 2.2 Software Installation Guide*.

The new file system will be mounted automatically at the next membership monitor reconfiguration. To force membership reconfiguration, use the following command:

```
# haswitch -r
```

## Administering Local Metadevices

Local disks can be mirrored. If a single mirror fails, use the instructions in the Solstice DiskSuite documentation to replace the failed mirror and resynchronize the replacement disk with the good disk.

## Destructive Metadevice Actions

The metadevice actions that are not supported in Sun Cluster configurations include:

- Creation of a one-way mirror in a diskset
- Creation of a configuration with too few metadevice state database replicas on the local disks
- Modification of metadevice state database replicas on multihost disks, unless there are explicit instructions to do so in this or another Sun Cluster book

# Using VxVM in the Sun Cluster Environment

VERITAS Volume Manager (VxVM) and the VxVM cluster feature are variations of the same volume manager. The VxVM cluster feature is only used in Oracle Parallel Server (OPS) configurations. This section describes using disks under the control of the volume manager to administer:

- Volume manager disks
- Disk groups
- Subdisks
- Plexes
- Volumes

Refer to the appropriate section for a complete discussion of administering these objects.

## Objects Administration Overview (VxVM)

Objects under the control of a volume manager are created and administered using either command-line utilities or the Visual Administrator graphical user interface.

Read the information in this chapter before using the VxVM documentation to administer objects under the control of a volume manager in a Sun Cluster configuration. The procedures presented here are one method for performing the following tasks. Use the method that works best for your particular configuration.

These objects generally have the following relationship:

- Disks are placed under volume manger control and are grouped into disk groups.
- One or more subdisks (each representing a specific portion of a disk) are combined to form plexes, or mirrors.
- A volume is composed of one or more plexes.

The default disk group is `rootdg` (the root disk group). You can create additional disk groups as necessary. The primary administration tasks that you perform on disk groups involve adding and removing disks.

Before using a disk that you have placed in a disk group, you must set up disks and subdisks (under volume manager control) to build plexes, or mirrors, using the physical disk’s slices. A plex can be a concatenation or stripe.

With VxVM, applications access volumes (created on volume manager disks) rather than slices.

The following sections describe the VxVM command-line programs to use when performing a task. Optionally, you can use the graphical user interface for all the tasks unless directed otherwise.

---

**Note -** On nodes running Sun Cluster HA data services, never manually run the `vxdg import` or `deport` options on a disk group that is under the control of Sun Cluster, unless the logical host for that disk group is in maintenance mode. Before manually importing or deporting a disk group, you must either stop Sun Cluster on all nodes that can master that disk group (by running `scadmin stopnode` on all such nodes), or use the `haswitch -m` command to switch any corresponding logical host into maintenance mode. When you are ready to return control of the disk group to Sun Cluster, the safest course is to deport the disk group before running `scadmin startnode` or before using `haswitch(1M)` to place the logical host back under the control of Sun Cluster.

---

## Administering Disks

Before a disk can be used by VxVM, it must be identified, or initialized, as a disk that is under control of a volume manager. A fully initialized disk can be added to a disk group, used to replace a previously failed disk, or used to create a new disk group.

## ▼ How to Initialize and Configure a Disk (VxVM)

1. **Ensure that no data is on the disk.**

   This is important because existing data is destroyed if the disk is initialized.

2. **Insert the disk device and install it in the disk enclosure by following the instructions in the accompanying hardware documentation.**

3. **Initialize the disk and add it to a disk group.**

   This is commonly done by using either the `vxdiskadm` menus or the graphical user interface. Alternately, you can use the command line utilities `vxdisksetup` and `vxdg addisk` to initialize the disk and place it in a disk group.

### Taking a Disk Offline

Occasionally, you may need to take a physical disk offline. If the disk is corrupted, you need to disable it and remove it. You also must disable a disk before moving the physical disk device to another location to be connected to another system.

To take a physical disk offline, first remove the disk from its disk group. Then place the disk offline by using the `vxdisk(1M)` command.

### Removing a Disk

You can remove a disk to move it to another system, or you may remove the disk because the disk is failing or has failed. Alternatively, if the volumes are no longer needed, they can be removed.

To remove a disk from the disk group, use the `vxdg(1M)` command. To remove the disk from volume manager control by removing the private and pubic partitions, use the `vxdiskunsetup(1M)` command. Refer to the `vxdg(1M)` and `vxdiskunsetup(1M)` man pages for complete information on these commands.

## Administering Disk Groups

For VxVM, it is most convenient to create and populate disk groups from the active node that is the default master of the particular disk group. In an N+1 configuration, each of these default master nodes shares multihost disk connectivity with only one other node in the cluster, the hot-standby node. By using these nodes to populate the disk groups, you avoid the risk of generating improperly configured groups.

### Creating a Disk Group (VxVM)

You can use either the `vxdiskadm` menus or the graphical user interface to create a new disk group. Alternately, you can use the command-line utility `vxdg init`.

Once the disk groups have been created and populated, each one should be deported by using the `vxdg deport` command. Then, each group should be imported onto the hot-standby node by using the `-t` option. The `-t` option is important, as it prevents the import from persisting across the next boot. All VxVM plexes and volumes should be created, and volumes started, before continuing.

### Moving a Disk to a Different Disk Group (VxVM)

Use the following procedure to move a disk to a different disk group.

## How to Move a Disk to a Different Disk Group (VxVM)

To move a disk between disk groups, remove the disk from one disk group and add it to the other.

This example moves the physical disk `c1t0d1` from disk group `acct` to disk group `log_node1` by using command-line utilities.

1. **Use the `vxprint(1M)` command to determine if the disk is in use.**

```
# vxprint -g acct
TY NAME         ASSOC        KSTATE   LENGTH   PLOFFS   STATE    TUTIL0  PUTIL0
dg acct         acct         -        -        -        -        -       -

dm c1t0d0       c1t0d0s2     -        2050272  -        -        -       -
dm c1t0d1       c1t0d1s2     -        2050272  -        -        -       -
dm c2t0d0       c2t0d0s2     -        2050272  -        -        -       -
dm c2t0d1       c2t0d1s2     -        2050272  -        -        -       -

v  newvol       gen          ENABLED  204800   -        ACTIVE   -       -
pl newvol-01    newvol       ENABLED  205632   -        ACTIVE   -       -
sd c1t0d1-01    newvol-01    ENABLED  205632   0        -        -       -
pl newvol-02    newvol       ENABLED  205632   -        ACTIVE   -       -
sd c2t0d1-01    newvol-02    ENABLED  205632   0        -        -       -

v  vol01        gen          ENABLED  1024000  -        ACTIVE   -       -
pl vol01-01     vol01        ENABLED  1024128  -        ACTIVE   -       -
sd c1t0d0-01    vol01-01     ENABLED  1024128  0        -        -       -
pl vol01-02     vol01        ENABLED  1024128  -        ACTIVE   -       -
sd c2t0d0-01    vol01-02     ENABLED  1024128  0        -        -       -
```

2. **Use the `vxedit(1M)` command to remove the volume to free up the `c1t0d1` disk.**

   You must run the `vxedit` command from the node mastering the shared disk group.

```
# vxedit -g acct -fr rm newvol
```

   The `-f` option forces an operation. The `-r` option makes the operation recursive.

3. **Remove the `c1t0d1` disk from the `acct` disk group.**

   You must run the `vxdg` command from the node mastering the shared disk group.

```
# vxdg -g acct rmdisk c1t0d1
```

4. **Add the** `c1t0d1` **disk to the** `log_node1` **disk group.**

```
# vxdg -g log_node1 adddisk c1t0d1
```

**Caution -** This procedure does not save the configuration or data on the disk.

This is the `acct` disk group after `c1t0d1` is removed.

```
# vxprint -g acct
TY NAME         ASSOC        KSTATE   LENGTH   PLOFFS   STATE    TUTIL0  PUTIL0
dg acct         acct         -        -        -        -        -       -

dm c1t0d0       c1t0d0s2     -        2050272  -        -        -       -
dm c2t0d0       c2t0d0s2     -        2050272  -        -        -       -
dm c2t0d1       c2t0d1s2     -        2050272  -        -        -       -

v  vol01        gen          ENABLED  1024000  -        ACTIVE   -       -
pl vol01-01     vol01        ENABLED  1024128  -        ACTIVE   -       -
sd c1t0d0-01    vol01-01     ENABLED  1024128  0        -        -       -
pl vol01-02     vol01        ENABLED  1024128  -        ACTIVE   -       -
sd c2t0d0-01    vol01-02     ENABLED  1024128  0        -        -       -
```

This is the `log_node1` disk group after `c1t0d1` is added.

```
# vxprint -g log_node1
TY NAME         ASSOC        KSTATE   LENGTH   PLOFFS   STATE    TUTIL0  PUTIL0
dg log_node1    log_node1    -        -        -        -        -       -

dm c1t0d1       c1t0d1s2     -        2050272  -        -        -       -
dm c1t3d0       c1t3d0s2     -        2050272  -        -        -       -
dm c2t3d0       c2t3d0s2     -        2050272  -        -        -       -
#
```

To change permissions or ownership of volumes, you must use the `vxedit` command.

**Caution -** Do not use `chmod` or `chgrp`. The permissions and ownership set by `chmod` or `chgrp` are automatically reset to `root` during a reboot.

Here is an example of the permissions and ownership of the volumes vol01 and vol02 in the /dev/vx/rdsk directory before a change.

```
# ls -l
crw-------     1   root  root  nnn,nnnnn   date   time  vol01
crw-------     1   root  root  nnn,nnnnn   date   time  vol02
...
```

This an example for changing the permissions and ownership for vol01.

```
# vxedit -g group_name set mode=755 user=oracle vol01
```

After the edit, note how the permissions and ownership have changed.

```
# ls -l
crwxr-xr-x     1  oracle  root  nnn,nnnnn   date   time  vol01
crw-------     1  root  root  nnn,nnnnn    date   time  vol02
...
```

## Administering VxVM Objects

Volumes, or virtual disks, can contain file systems or applications such as databases. A volume can consist of up to 32 plexes, each of which contains one or more subdisks. In order for a volume to be usable, it must have at least one associated plex with at least one associated subdisk. Note that all subdisks within a volume must belong to the same disk group.

### Creating Volumes and Adding Mirrors to Volumes

Use the graphical user interface or the command-line utility vxassist(1M) to create volumes in each disk group, and to create an associated mirror for each volume.

The actual size of a VxVM device is slightly less than the full disk drive size. VxVM reserve a small amount of space for private use, called the private region.

**Note -** The use of the same volume name is allowed if the volumes belong to different disk groups.

## Adding Dirty Region Logging

Dirty Region Logging (DRL) is an optional property of a volume, used to provide a speedy recovery of mirrored volumes after a system failure. DRL keeps track of the regions that have changed due to I/O writes to a mirrored volume and uses this information to recover only the portions of the volume that need to be recovered.

## Creating a Log File for an Existing Volume

Log subdisks are used to store the dirty region log of a volume that has DRL enabled. A volume with DRL has at least one log subdisk; multiple log subdisks can be used to mirror the dirty region log. Each log subdisk is associated with one of the volume's plexes. Only one log subdisk can exist per plex. If the plex contains only a log subdisk and no data subdisks, that plex can be referred to as a log plex. The log subdisk can also be associated with a regular plex containing data subdisks, in which case the log subdisk risks becoming unavailable in the event that the plex must be detached due to the failure of one of its data subdisks.

Use the graphical user interface or the command-line utility `vxassist(1M)` to create a log for an existing volume.

## Using Hot-Relocation

Hot-relocation is the ability of a system to automatically react to I/O failures on redundant (mirrored or RAID5) volume manager objects, and to restore redundancy and access to those objects. Hot-relocation is supported only on configurations using VxVM. VxVM detects I/O failures on volume manager objects and relocates the affected subdisks to disks designated as spare disks or free space within the disk group. VxVM then reconstructs the objects that existed before the failure and makes them redundant and accessible again.

When a partial disk failure occurs (that is, a failure affecting only some subdisks on a disk), redundant data on the failed portion of the disk is relocated, and the existing volumes consisting of the unaffected portions of the disk remain accessible.

---

**Note -** Hot-relocation is performed only for redundant (mirrored or RAID5) subdisks on a failed disk. Non-redundant subdisks on a failed disk are not relocated, but you are notified of their failure.

---

A spare disk must be initialized and placed in a disk group as a spare before it can be used for replacement purposes. If no disks have been designated as spares when a failure occurs, VxVM automatically uses any available free space in the disk group in which the failure occurs. If there is not enough spare disk space, a combination of spare space and free space is used. You can designate one or more disks as hot-relocation spares within each disk group. Disks can be designated as spares with the `vxedit(1M)` command.

## Using VxFS File Systems

You can configure and specify either UFS or VxFS file systems associated with a logical host's disk groups on volumes of type `fsgen`. When a cluster node masters a logical host, the logical host's file systems associated with the disk groups are mounted on the mastering node's specified mount points.

During a logical host reconfiguration sequence, it is necessary to check file systems with the `fsck(1M)` command. Though this process is performed in non-interactive parallel mode on UFS file systems, it can affect the overall time of the reconfiguration sequence. The logging feature of UFS, SDS, and VxFS file systems greatly reduce the time that `fsck(1M)` takes prior to mounting file systems.

When the switchover of a data service is required along with volume recovery, the recovery takes longer than allowed in the reconfiguration steps. This causes step time-outs and the node aborts.

Consequently, when setting up mirrored volumes, always add a DRL log to decrease volume recovery time in the event of a system crash. When mirrored volumes are used in the cluster environment, DRL must be assigned for volumes greater than 500 Mbytes.

Use VxFS if large file systems (greater than 500 Mbytes) are used for HA data services. Under most circumstances, VxFS is not bundled with Sun Cluster and must be purchased separately from VERITAS.

---

**Note -** Although it is possible to configure logical hosts with very small mirrored file systems, you should use Dirty Region Logging (DRL) or VxFS file systems because of the possibility of time-outs as the size of the file system increases.

---

## Growing a File System

To grow a striped or RAID5 volume containing a file system, you must have the free space on the same number of disks that are currently in the stripe or RAID5 volume. For example, if you have four 1GB disks striped together (giving you a 4GB file system), and you wish to add 1GB of space (to yield a 5GB filesystem), you must have four new disks, each with at least .25GB of free space. In other words, you can not add one disk to a 4-disk stripe.

The VxVM graphical user interface will choose the disks on which to grow your file system. To select the specific disks on which to grow the file system, use the command line interface instead.

UFS file systems cannot be shrunk. The only way to "shrink" a file system is to recreate the volume, run `newfs` on the volume, and then restore the data from backup.

## Administering Local Mirrors

Local disks can be mirrored. If a single mirror fails, use the instructions in your volume manager documentation to replace the failed mirror and resynchronize the replacement disk with the good disk.

# Backing Up Multihost Data Using Solstice Backup

This section contains suggestions for using Solstice Backup™ to back up Sun Cluster file systems.

Solstice Backup is designed to run each copy of the server software on a single server. Solstice Backup expects files to be recovered using the same physical server from which they were backed up.

Solstice Backup has considerable data about the physical machines (host names and host IDs) corresponding to the server and clients. Solstice Backup's information about the underlying physical machines on which the logical hosts are configured affects how it stores client indexes.

Do not put the Solstice Backup `/nsr` database on the multihost disks. Conflicts can arise if two different Solstice Backup servers attempt to access the same `/nsr` database.

Because of the way Solstice Backup stores client indexes, do not back up a particular client using different Solstice Backup servers on different days. Make sure that a particular logical host is always mastered by the same physical server whenever backups are performed. This will enable future recover operations to succeed.

---

**Note -** By default, Sun Cluster systems will not generate the full file system list for your backup configuration. If the save set list consists of the keyword All, then the `/etc/vfstab` file will be examined to determine which file systems should be saved. Because Sun Cluster `vfstab` files are kept in `/etc/opt/SUNWcluster/conf/hanfs` by default, Solstice Backup will not find them unless you explicitly list the Sun Cluster files systems to be saved. When you are testing your backup procedures, verify that all of the Sun Cluster file systems that need to be backed up appear in the Solstice Backup file system list.

---

Four methods of configuring Solstice Backup are presented here. You might prefer one depending on your particular Sun Cluster configuration. Switchover times could influence your decision. Once you decide on a method, continue using that method so that future recover operations will succeed.

Here is a description of the configuration methods:

- Use a non-cluster node, non-high availability server configured as a Solstice Backup server.

Configure an additional server apart from the Sun Cluster servers to act as the Solstice Backup server. Configure the logical hosts as clients of the server. For best results, always ensure that the logical hosts are configured on their respective default masters before doing the daily backup. This might require a switchover. Having the logical hosts mastered by alternate servers on different days (possibly as the result of a takeover) could cause Solstice Backup to become confused upon attempting a recover operation, due to the way Solstice Backup stores client indexes.

- Use one Sun Cluster server configured to perform local backups.

Configure one of the Sun Cluster servers to perform local backups. Always switch the logical hosts to the Solstice Backup server before performing the daily backup. That is, if `phys-hahost1` and `phys-hahost2` are the Sun Cluster servers, and `phys-hahost1` is the Solstice Backup server, always switch the logical hosts to phys-hahost1 before performing backups. When backups are complete, switch back the logical host normally mastered by `phys-hahost2`.

- Use the Sun Cluster servers configured as Solstice Backup servers.

Configure each Sun Cluster server to perform local backups of the logical host it masters by default. Always ensure that the logical hosts are configured on their respective default masters before performing the daily backup. This might require a switchover. Having the logical hosts mastered by alternate servers on different days (possibly as the result of a takeover) could cause Solstice Backup to become confused upon attempting a recover operation, due to the way Solstice Backup stores client indexes.

- Use one Sun Cluster server configured as the Solstice Backup server.

Configure one Sun Cluster server to back up its logical host locally and to back up its sibling’s logical host over the network. Always ensure that the logical hosts are configured on their respective default masters before doing the daily backup. This might require a switchover. Having the logical hosts mastered by alternate servers on different days (possibly as the result of a takeover) could cause Solstice Backup to become confused upon attempting a recover operation, due to the way Solstice Backup stores client indexes.

In all four of the above backup options, you can have another server configured to temporarily perform backups in the event the designated Solstice Backup server is down. Note that you will not be able to use the temporary Solstice Backup server to recover files backed up by the normal Solstice Backup server, and that you cannot recover files backed up by the temporary server from the normal backup server.

APPENDIX **B**

# Sun Cluster Fault Detection

This appendix describes fault detection for Sun Cluster, and includes the following topics:

- "Fault Detection Overview" on page 286
- "Public Network Monitoring (PNM)" on page 288
- "Sun Cluster Fault Probes" on page 289
- "Data Service-Specific Fault Probes" on page 289

This section presents an overview of Sun Cluster fault detection. This fault detection encompasses three general approaches:

- A heartbeat mechanism
- Fault monitoring of networks
- Fault monitoring of specific data services

Fault monitoring performs sanity checks to ensure that the faulty node is the one being blamed for a problem, and not the healthy node.

Some of the information presented is specific to this release of Sun Cluster, and is expected to change as the product evolves. The time estimates given to detect various faults are rough approximations and are intended only to give the reader a general understanding of how Sun Cluster behaves. This document is not intended to be a program logic manual for the internals of Sun Cluster nor does it describe a programming interface.

# Fault Detection Overview

As noted in the basic Sun Cluster architecture discussion, when one server goes down the other server takes over. This raises an important issue: how does one server recognize that another server is down?

Sun Cluster uses three methods of fault detection.

- Heartbeat and SMA link monitoring – These monitors run over the private links. For Ethernet, there are two monitors: an SMA link monitor and a cluster membership monitor. For SCI, there are three monitors: an SMA link monitor, a cluster membership monitor, and a low-level SCI heartbeat monitor.
- Network fault monitoring – All servers' public network connections are monitored: if a server cannot communicate over the public network because of a hardware or software problem, then another server in the server set will take over.
- Data service-specific fault probes – Each Sun Cluster data service performs fault detection that is specific for that data service. This last method addresses the issue of whether the data service is performing useful work, not just the low-level question of whether the machine and operating system appear to be running.

For the second and third methods, one server is probing the other server for a response. After detecting an apparent problem, the probing server carries out a number of sanity checks of itself before forcibly taking over from the other server. These sanity checks try to ensure that a problem on the probing server is not the real cause of the lack of response from the other server. These sanity checks are provided by `hactl(1M)`, a library subroutine that is part of the Sun Cluster base framework; hence, data service-specific fault detection code need only call `hactl(1M)` to perform sanity checks on the probing server. See the `hactl(1M)` man page for details.

## The Heartbeat Mechanism: Cluster Membership Monitor

Sun Cluster uses a heartbeat mechanism. The heartbeat processing is performed by a real-time high-priority process which is pinned in memory, that is, it is not subject to paging. This process is called the *cluster membership monitor*. In a `ps(1)` listing, its name appears as `clustd`.

Each server sends out an "I am alive" message, or heartbeat, over both private links approximately once every two seconds. In addition, each server is listening for the heartbeat messages from other servers, on both private links. Receiving the heartbeat on either private link is sufficient evidence that another server is running. A server will decide that another server is down if it does not hear a heartbeat message from that server for a sufficiently long period of time, approximately 12 seconds.

In the overall fault detection strategy, the cluster membership monitor heartbeat mechanism is the first line of defense. The absence of the heartbeat will immediately detect hardware crashes and operating system panics. It might also detect some gross operating system problems, for example, leaking away all communication buffers. The heartbeat mechanism is also Sun Cluster's fastest fault detection method. Because the cluster membership monitor runs at real-time priority and because it is pinned in memory, a relatively short timeout for the absence of heartbeats is justified. Conversely, for the other fault detection methods, Sun Cluster must avoid labelling a server as being down when it is merely very slow. For those methods, relatively long timeouts of several minutes are used, and, in some cases, two or more such timeouts are required before Sun Cluster will perform a takeover.

The fact that the cluster membership monitor runs at real-time priority and is pinned in memory leads to the paradox that the membership monitor might be alive even though its server is performing no useful work at the data service level. This motivates the data service-specific fault monitoring, as described in "Data Service-Specific Fault Probes" on page 289.

## Sanity Checking of Probing Node

The network fault probing and data service-specific fault probing require each node to probe another node for a response. Before doing a takeover, the probing node performs a number of basic sanity checks of itself. These checks attempt to ensure that the problem does not really lie with the probing node. They also try to ensure that taking over from the server that seems to be having a problem really will improve the situation. Without the sanity checks, the problem of *false takeovers* would likely arise. That is, a sick node would wrongly blame another node for lack of response and would take over from the healthier server.

The probing node performs the following sanity checks on itself before doing a takeover from another node:

- The probing node checks its own ability to use the public network, as described in "Public Network Monitoring (PNM)" on page 288.

- The probing node also checks whether its own HA data services are responding. All the HA data services that the probing node is already running are checked. If any are not responsive, takeover is inhibited, on the assumption that the probing node will not do any better trying to run another node's services if it can't run its own. Furthermore, the failure of the probing node's own HA data services to respond might be an indication of some underlying problem with the probing node that could be causing the probe of the other node to fail. Sun Cluster HA for NFS provides an important example of this phenomenon: to lock a file on another node, the probing node's own `lockd` and `statd` daemons must be working. By checking the response of its `lockd` and `statd` daemons, the probing node rules out the scenario where its own daemons' failure to respond makes the other node look unresponsive.

# Public Network Monitoring (PNM)

The PNM component has two primary functions:

- To monitor the status of configured adapters on a node and report general adapter or network failures
- To fail over transparently to other backup adapters on a node when the primary adapter fails

PNM is implemented as a daemon (`pnmd`) which periodically gathers network statistics on the set of public network interfaces in a node. If the results indicate any abnormalities, `pnmd` attempts to distinguish between the following three cases:

- The network is quiescent.
- The network is down.
- The network interface is down.

PNM then does a multicast `ping`. PNM places the results of its findings in the CCD and compares the local results with the results of the other nodes (which are also placed in the CCD). This comparison is used to determine whether the network is down or whether the network interface is faulty. If PNM detects that the network interface is faulty and backup adapters are configured, it performs the network adapter failover.

---

**Note -** The multicast `ping` initiated by PNM might not be understood by any non-Sun hardware components present in the configuration. Therefore, you should directly connect a Sun network appliance to the network being monitored.

---

The results of PNM monitoring are used by various entities. The network adapter failover component of PNM uses the monitoring results to decide whether an adapter failover would be useful. For example, if the network is experiencing a failure, no adapter failover is performed. Fault monitors associated with SC HA data services and the API call `hactl` use the PNM facility to diagnose the cause of data service failures. The information returned by PNM is used to decide whether to migrate the data service, and to determine the location of the data service after migration.

The `syslog` messages written by the PNM facility on detection of adapter failures are read by the SC Manager, which translates the messages into graphic icons and displays them through the graphical user interface.

You also can run the PNM utilities on the command line to determine the status of network components. For more information, see the man pages `pnmset(1M)`, `pnmstat(1M)`, `pnmptor(1M)pnmrtop(1M)`, , and `pnmd(1M)`.

# Sun Cluster Fault Probes

PNM monitors the health of the public network and will switch to backup connections when necessary. However, in the event of the total loss of public network access, PNM will not provide data service or logical host failover. In such a case, PNM will report the loss but it is up to an external fault probe to handle switching between backup nodes.

If you are using VxVM as your volume manager, the Sun Cluster framework is responsible for monitoring each Network Adapter Failover (NAFO) backup group defined per logical host, and initiating a switchover to a backup node when either of the following conditions are met:

- There is total loss of the public network (all NAFO backup groups are unavailable) and the backup node has at least one NAFO group available.
- There is partial loss of the public network—at least one NAFO backup group is still active when more than one (multiple subnets) are defined for a logical host—and the backup node has a greater number of valid, active NAFO backup groups.

If neither of these conditions are met, Sun Cluster will not attempt a switchover.

If your volume manager is Solstice DiskSuite, loss of public network causes the disconnected node to abort and causes the logical hosts mastered by that node to migrate to the backup node.

The Sun Cluster framework monitors the public networks only while the configuration includes a logical host and while a data service is in the “on” state and registered on that logical host. Only those NAFO backup groups that are in use by a logical host are monitored.

# Data Service-Specific Fault Probes

The motivation for performing data service-specific fault probing is that although the server node and operating system are running, the software or hardware might be in such a confused state that no useful work at the data service level is occurring. In the overall architecture, the total failure of the node or operating system is detected by the cluster membership monitor’s heartbeat mechanism. However, a node might be working well enough for the heartbeat mechanism to continue to execute even though the data service is not doing useful work.

Conversely, the data service-specific fault probes do not need to detect the state where one node has crashed or has stopped sending cluster heartbeat messages. The assumption is made that the cluster membership monitor detects such states, and the data service fault probes themselves contain no logic for handling these states.

A data service fault probe behaves like a client of the data service. A fault probe running on a machine monitors both the data service exported by that machine and, more importantly, the data service exported by another server. A sick server is not reliable enough to detect its own sickness, so each server is monitoring another node in addition to itself.

In addition to behaving like a client, a data service-specific fault probe will also, in some cases, use statistics from the data service as an indication that useful work is or is not occurring. A probe might also check for the existence of certain processes that are crucial to a particular data service.

Typically, the fault probes react to the absence of service by forcing one server to take over from another. In some cases, the fault probes will first attempt to restart the data service on the original machine before doing the takeover. If many restarts occur within a short time, the indication is that the machine has serious problems. In this case, a takeover by another server is executed immediately, without attempting another local restart.

## Sun Cluster HA for NFS Fault Probes

The probing server runs two types of periodic probes against another server's NFS service.

1. The probing server sends a `NULL RPC` to all daemon processes on the target node that are required to provide NFS service; these daemons are `rpcbind`, `mountd`, `nfsd`, `lockd`, and `statd`.
2. The probing server does an end-to-end test: it tries to mount an NFS file system from the other node, and then to read and write a file in that file system. It does this end-to-end test for every file system that the other node is currently sharing. Because the mount is expensive, it is executed less often than the other probes.

If any of these probes fail, the probing node will consider doing a takeover from the serving node. However, certain conditions might inhibit the takeover from occurring immediately:

- Grace period for local restart – Before doing the takeover, the probing node waits for a short time period that is intended to:
  - Give the victim node a chance to notice its own sickness and fix the problem by doing a local restart of its own daemons
  - Give the victim node a chance to be less busy (if it is merely overloaded)

After waiting, the prober retries the probe, going on with takeover consideration only if it fails again. In general, two entire timeouts of the basic probe are required for a takeover, to allow for a slow server.

- Multiple public networks – If the other node is on multiple public networks, the probing node will try the probe on at least two of them.
- Locks – Some backup utilities exploit the `lockfs(1M)` facility, which locks out various types of updates on a file system, so that backup can take a snapshot of an unchanging file system. Unfortunately, in the NFS context, `lockfs(1M)` makes a file system appear unavailable; NFS clients will see the condition `NFS server not responding`. Before doing a takeover, the probing node queries the other node to find out whether the file system is in `lockfs` state, and, if so, takeover is inhibited. The takeover is inhibited because the `lockfs` is part of a normal, intended administrative procedure for doing backup. Note that not all backup utilities use `lockfs`; some permit NFS service to continue uninterrupted.
- Daemons – Unresponsiveness of `lockd` and `statd` daemons does not cause a takeover. The `lockd` and `statd` daemons, together, provide network locking for NFS files. If these daemons are unresponsive, the condition is merely logged to `syslog`, and a takeover does not occur. `lockd` and `statd`, in the course of their normal work, must perform RPCs to client machines, so that a dead or partitioned client can cause `lockd` and `statd` to hang for long periods of time. Thus, a bad client can make `lockd` and `statd` on the server look sick. And if a takeover by the probing server were to occur, the probing server would probably be stalled by the bad client in the same way. With the current model, a bad client will not cause a false takeover.

After passing these Sun Cluster HA for NFS-specific tests, the process of considering whether or not to do a takeover continues with calls to `hactl(1M)` (see “Sanity Checking of Probing Node” on page 287).

The probing server also checks its own NFS service. The logic is similar to the probes of the other server, but instead of doing takeovers, error messages are logged to syslog and an attempt is made to restart any daemons whose process no longer exists. In other words, the restart of a daemon process is performed only when the daemon process has exited or crashed. The restart of a daemon process is not attempted if the daemon process still exists but is not responding, because that would require killing the daemon without knowing which data structures it is updating. The restart is also not done if a local restart has been attempted too recently (within the last hour). Instead, the other server is told to consider doing a takeover (provided the other server passes its own sanity checks). Finally, the `rpcbind` daemon is never restarted, because there is no way to inform processes that had registered with `rpcbind` that they need to re-register.

## HA-DBMS Fault Probes

The fault probes for Sun Cluster HA for Oracle, Sun Cluster HA for Sybase and Sun Cluster HA for Informix perform similarly to monitor the database server. The HA-DBMS fault probes are configured by running one of the utilities, `haoracle(1M)`, `hasybase(1M)`, or `hainformix(1M)`. (See the online man pages for a detailed description of the options for these utilities.)

Once the utilities are configured and activated, two processes are started on the local node and two processes are started on the remote node simulating a client access. The remote fault probe is initiated by the `ha_dbms_serv` daemon and is started when `hareg -y` *dataservicename* is initiated.

The HA-DBMS module uses two methods to monitor whether the DBMS service is available. First, HA-DBMS extracts statistics from the DBMS itself:

- In Oracle, the `V$SYSSTAT` table is queried.
- In Sybase, the global variables `@@io_busy`, `@@pack_received`, `@@pack_sent`, `@@total_read`, `@@total_write`, and `@@connections` are queried.
- In Informix, the `SYSPROFILE` table is queried.

If the extracted statistics indicate that work is being performed for clients, then no other probing of the DBMS is required. Second, if the DBMS statistics show that no work is occurring, then HA-DBMS submits a small test transaction to the DBMS. If all clients happen to be idle, the DBMS statistics would show no work occurring; that is, the test transaction distinguishes the situation of the database being hung from the legitimately idle situation. Because the test transaction is executed only when the statistics show no activity, it imposes no overhead on an active database. The test transaction consists of:

- Creating a table by the name of either `HA_DBMS_REM` or `HA_DBMS_LOC`
- Inserting values into the created table
- Updating the inserted value
- Dropping the created table

HA-DBMS carefully filters the error codes returned by the DBMS, using a table that describes which codes should or should not cause a takeover. For example, in the case of Sun Cluster HA for Oracle, the scenario of `table space full` does not cause a takeover, because an administrator must intervene to fix this condition. (If a takeover were to occur, the new master server would encounter the same `table space full` condition.)

On the other hand, an error return code such as `could not allocate Unix semaphore` causes Sun Cluster HA for Oracle to attempt to restart ORACLE locally on this server machine. If a local restart has occurred too recently, then the other machine takes over instead (after first passing its own sanity checks).

## Sun Cluster HA for Netscape Fault Probes

The fault monitors for all of the Sun Cluster HA for Netscape data services share a common methodology for fault monitoring of the data service instance. All use the concept of remote and local fault monitoring.

The fault monitor process running on the node which currently masters the logical host that the data service is running on is called the local fault monitor. The fault monitor process running on a node which is a possible master of the logical host is called a remote fault monitor.

Sun Cluster HA for Netscape fault monitors periodically perform a simple data service operation with the server. If the operation fails or times out, that particular probe is declared to have failed.

When a probe fails, the local fault probe attempts to restart the data service locally. This is usually sufficient to restore the data service. The remote probe keeps a record of the probe failure but does not take any action. Upon two successive failures of the probe (indicating that a restart of the data service did not correct the problem), the remote probe invokes the `hactl(1M)` command in “takeover” mode to initiate a failover of the logical host. Some Netscape data services use a sliding window algorithm of probe successes and failures, in which a pre-configured number of failures within the window causes the probe to take action.

You can use the `hadsconfig(1M)` command to tune probe interval and timeout values for Sun Cluster HA for Netscape fault monitors. Reducing the probe interval value for fault probing results in faster detection of problems, but it also might result in spurious failovers due to transient problems. Similarly, reducing the probe timeout value results in faster detection of problems related to the data service instances, but also might result in spurious takeovers if the data service is merely busy due to heavy load. For most situations, the default values for these parameters are sufficient. The parameters are described in the `hadsconfig(1M)` man page and in the configuration sections of each data service chapter in the *Sun Cluster 2.2 Software Installation Guide*.

## Sun Cluster HA for DNS Fault Probes

The Sun Cluster HA for DNS fault probe performs an `nslookup` operation to check the health of the Sun Cluster HA for DNS server. It looks up the domain name of the Sun Cluster HA for DNS logical host from the Sun Cluster HA for DNS server. Depending upon the configuration of your `/etc/resolv.conf` file, `nslookup` might contact other servers if the primary Sun Cluster HA for DNS server is down. Thus, the `nslookup` operation might succeed, even when the primary Sun Cluster HA for DNS server is down. To guard against this, the fault probe verifies whether replies come from the primary Sun Cluster HA for DNS server or other servers.

## Sun Cluster HA for Netscape HTTP Fault Probes

The Sun Cluster HA for Netscape HTTP fault probe checks the health of the `http` server by trying to connect to it on the logical host address on the configured port. Note that the fault monitor uses the port number specified to `hadsconfig(1M)` during configuration of the `nshttp` service instance.

## Sun Cluster HA for Netscape News Fault Probes

The Sun Cluster HA for Netscape News fault probe checks the health of the news server by connecting to it on the logical host IP addresses and the `nntp` port number. It then attempts to execute the NNTP `date` command on the news server, and expects a response from the server within the specified probe timeout period.

## Sun Cluster HA for Netscape Mail or Message Server Fault Probes

The Sun Cluster HA for Netscape Mail or Message Server fault probe checks the health of the mail or message server by probing it on all three service ports served by the server, namely the SMTP, IMAP, and POP3 ports:

- SMTP (port 25)—Executes an SMTP "hello" message on the server and then executes a `quit` command.
- IMAP (port 143)—Executes an IMAP4 `CAPABILITY` command followed by an IMAP4 `LOGOUT` command.
- POP3 (port 110)—Executes a `quit` command.

For all of these tests, the fault probe expects a response string from the server within the probe timeout interval. Note that a probe failure on any of the above three service ports is considered a failure of the server. To avoid spurious failovers, the `nsmail` fault probe uses a sliding window algorithm for tracking probe failures and successes. If the number for probe failures in the sliding window is greater than a pre-configured number, a takeover is initiated by the remote probe.

## Sun Cluster HA for Netscape LDAP Fault Probes

The Sun Cluster HA for Netscape LDAP local probe can perform a variable number of local restarts before initiating a failover. The local restart mechanism uses a sliding window algorithm; only when the number of retries is exhausted within that window does a failover occur.

The Sun Cluster HA for Netscape LDAP remote probe uses a simple telnet connection to the LDAP port to check the status of the server. The LDAP port number is the one specified during initial set-up with `hadsconfig(1M)`.

The local probe:

- Probes the server by running a monitoring script. The script performs a search for the LDAP common name "monitor." The common name is defined by the Directory Server and is used only for monitoring. The probe uses the `ldapsearch` utility to perform this operation.
- Tries to restart the server locally, upon detecting a problem with the server.
- Initiates the `hactl(1M)` command in the `giveup` mode upon deciding that the local node cannot reliably run the directory server instance, while the remote

probe initiates the hactl(1M) command in the takeover mode. If there are multiple possible masters of the logical host, all of the remote probes invoke the takeover operation in unison. However, after the takeover, the underlying framework ensures that a unique master node is chosen for the Directory Server.

## Sun Cluster HA for Lotus Fault Probes

The Sun Cluster HA for Lotus fault probe has two parts—a local probe that runs on the node on which the Lotus Domino server processes are currently running, and a remote probe that runs on all other nodes that are possible masters of the Lotus Domino server's logical host.

Both probes use a simple telnet connection to the Lotus Domino port to check the status of the Domino server. If a probe fails to connect, it initiates a failover or takeover by invoking the hactl(1M) command.

The local fault probe can perform three local restarts before initiating a failover. The local restart mechanism uses a sliding time window algorithm; only when the number of retries is exhausted within that window does a failover occur.

## Sun Cluster HA for Tivoli Fault Probes

Sun Cluster HA for Tivoli uses only a local fault probe. It runs on the node on which the Tivoli object dispatcher, the oserv daemon, is currently running.

The fault probe uses the Tivoli command wping to check the status of the monitored oserv daemon. The wping of an oserv daemon can fail for the following reasons:

- The monitored oserv daemon is not running.
- The oserv daemon on the server dies while monitoring a client oserv daemon.
- Proper Tivoli roles (authorization) have not been set for the administrative user. See the *Sun Cluster 2.2 Software Installation Guide* for details about Tivoli.

If the local probe fails to ping the oserv daemon, it initiates a failover by invoking the hactl(1M) command. The fault probe will perform one local restart before initiating a failover.

## Sun Cluster HA for SAP Fault Probes

The Sun Cluster HA for SAP fault probe monitors the availability of the Central Instance, specifically the message server, the enqueue server, and the dispatcher. The probe monitors only the local node by checking for the existence of the critical SAP processes. It also uses the SAP utility lgtst to verify that the SAP message server is reachable.

Upon detecting a problem, such as when a process dies prematurely or lgtst reports an error, the fault probe will first try to restart SAP on the local node for a configurable number of times (configurable through hadsconfig(1M)). If the number of restarts that the user has configured has been exhausted, then the fault probe initiates a switchover by calling hactl(1M), if this instance has been configured to allow failover (also configurable through hadsconfig(1M)). The Central Instance is shut down before the switchover occurs, and then is restarted on the remote node after the switchover is complete.

## Displaying LOG_DB_WARNING Messages for the SAP Probe

The Sun Cluster HA for SAP parameter LOG_DB_WARNING determines whether warning messages should be displayed if the Sun Cluster HA for SAP probe cannot connect to the database. When LOG_DB_WARNING is set to y and the probe cannot connect to the database, a message is logged at the warning level in the local0 facility. By default, the syslogd(1M) daemon does not display these messages to /dev/console or to /var/adm/messages. To see these warnings, you must modify the /etc/syslog.conf file to display messages of local0.warning priority. For example:

```
...
*.err;kern.notice;auth.notice;local0.warning /dev/console
*.err;kern.debug;daemon.notice;mail.crit;local0.warning /var/adm/messages
...
```

After modifying the file, you must restart syslogd(1M). See the syslog.conf(1M) and syslogd(1M) man pages for more information.

APPENDIX **C**

# Using Sun Cluster SNMP Management Solutions

This appendix describes how to use SNMP to monitor the behavior of a Sun Cluster configuration, and includes the following topics.

You can use the following SNMP management solutions to monitor Sun Cluster configurations:

- Sun Cluster SNMP Agent
- Domain Manager
- Enterprise Manager
- Sun Net Manager
- SNMP compliant HP OpenView

# Cluster SNMP Agent and Cluster Management Information Base

Sun Cluster includes a Simple Network Management Protocol (SNMP) agent, along with a Management Information Base (MIB), for the cluster. The name of the agent file is `snmpd` (SNMP daemon) and the name of the MIB is `sun.mib`.

The cluster SNMP agent is a proxy agent that is capable of monitoring several clusters (a maximum of 32) at the same time. You can manage a typical Sun Cluster from the administration workstation or System Service Processor (SSP). By installing the cluster SNMP agent on the administrative workstation or SSP, network traffic is regulated and the CPU power of the nodes is not wasted in transmitting SNMP packets.

The `snmpd` daemon:

- Is an `RFC 1157`-compliant SNMP agent.
- Is dedicated to support the Sun Cluster (SC) MIB extensions under the enterprise group of Sun Microsystems, Inc.
- Provides the cluster `sun.mib` in ASCII format.
- Supports SNMP protocol operations including `GET-REQUEST`, `GETNEXT-REQUEST` and `TRAP`.
- Provides the Super Monitor agent `smond` for data collection.

The Super Monitor daemon, `smond`, collects hardware configuration information and critical cluster events by connecting to the `in.mond` daemon from each of the member nodes of the cluster(s). The `smond` daemon then reports the same information to the SNMP daemon (`snmpd`).

**Note -** You need to configure only one `smond` daemon to collect cluster information for several clusters.

The `SUNWcsnmp` package contains the following:

- `/opt/SUNWcluster/bin/snmpd` and `/opt/SUNWcluster/bin/smond` binaries
- ASCII `/opt/SUNWcluster/etc/sun.mib` file
- `/opt/SUNWcluster/bin/init.snmpd` script (`snmpd` control)
- `/var/opt/SUNWcluster/snmpd.conf` file (SNMP configuration)
- `/opt/SUNWcluster/etc/snmp.traps` file (SNMP traps)
- `/opt/SUNWcluster/etc/sun-snmp.schema` file (SunNet Manager schema)
- `/opt/SUNWcluster/bin/smond_conf` script (`smond` configuration)

- /opt/SUNWcluster/bin/smond_ctl script (smond control)
- Applicable man pages

For additional information on the snmpd and smond daemons, see the associated man pages.

# Cluster Management Information Base

The Management Information Base (MIB) is a collection of objects that can be accessed through a network management protocol. The definition of the objects should be in a generic and consistent manner so that various management platforms can read and parse the definition.

Run the snmpd daemon on the management server, which is the cluster administration workstation, or on any client. This agent provides information (gathered from smond) for all the SNMP attributes defined in the cluster MIB. This MIB file is typically compiled into an "SNMP-aware" network manager like the SunNet Manager Console. See "Changing the snmpd.conf File" on page 309.

The sun.mib file provides information about clusters in the following tables:

- clustersTable
- clusterNodesTable
- switchesTable
- portsTable
- lhostTable
- dsTable
- dsinstTable

**Note -** In the preceding bullets, time refers to the local time on the SNMP server (in which the table is maintained). Thus, the time indicates when any attribute change is reported on the server.

## The clustersTable Attributes

The clusters table consists of entries for all of the monitored clusters. Each entry in the table contains specific attributes that provide cluster information. See Table C–1 for the clustersTable attributes.

**TABLE C–1** `clustersTable` Attributes

| Attribute Names | Description |
|---|---|
| `clusterName` | The name of the cluster |
| `clusterDescr` | A description of the cluster |
| `clusterVersion` | The release version of the cluster |
| `numNodes` | The number of nodes in the cluster |
| `nodeNames` | The names of all the nodes in the cluster, separated by commas |
| `quorumDevices` | The names of all the quorum devices in the cluster, separated by commas |
| `clusterLastUpdate` | The last time any of the attributes of this entry were modified |

## The `clusterNodesTable` Attributes

The cluster nodes table consists of the known nodes of all of the monitored clusters. Each entry contains specific information about the node. See Table C–2 for the `clusterNodesTable` attributes.

**Note -** When using a cross-reference, note that the `belongsToCluster` attribute acts as the key reference between this table and the `clustersTable`.

**TABLE C–2** `clusterNodesTable` Attributes

| Attribute Names | Description |
|---|---|
| `nodeName` | The host name of the node. |
| `belongsToCluster` | The name of the cluster (to which this node belongs). |

**TABLE C–2** `clusterNodesTable` Attributes *(continued)*

| Attribute Names | Description |
|---|---|
| scState | State of the Sun Cluster software component on this node (Stopped, Aborted, In Transition, Included, Excluded, or Unknown). An enterprise specific trap signals a change in state. |
| vmState | State of the volume manager software component on this node. An enterprise specific trap signals a change in state. |
| dbState | State of the database software component on this node (Down, Up, or Unknown). An enterprise specific trap signals a change in state. |
| vmType | The type of volume manager currently being used on this node. |
| vmOnNode | Mode of the VxVM software component on this node (Master, Slave, or Unknown). An enterprise specific trap signals a change in state. This attribute is not valid for clusters with other volume managers. |
| `nodeLastUpdate` | The last time any of the attributes of this entry were modified. |

## The `switchesTable` Attributes

The switches table consists of entries for all of the switches. Each entry in the table contains specific information about a switch in a cluster. See Table C–3 for the `switchesTable` attributes.

**TABLE C–3** `switchesTable` Attributes

| Attribute Names | Description |
|---|---|
| `switchName` | The name of the switch |
| `numPorts` | The number of ports on the switch |
| `connectedNodes` | The names of all the nodes that are presently connected to the ports of the switch |
| `switchLastUpdate` | The last time any of the switch attributes of this entry were modified |

## The portsTable Attributes

The ports table consists of entries for all of the switch ports. Each entry in the table contains specific information about a port within a switch. See Table C–4 for the portsTable attributes.

**Note -** When using a cross-reference, note that the belongsToSwitch attribute acts as the key reference between this table and the switchesTable.

**TABLE C–4** portsTable Attributes

| Attribute Names | Description |
|---|---|
| portId | The port ID or number |
| belongsToSwitch | The name of the switch (to which this port belongs) |
| connectedNode | The name of the node (to which this port is presently connected) |
| nodeAdapterId | The adapter ID (of the SCI card) on the node to which this port is connected |
| portStatus | The status of the port (Active, Inactive, and so forth) |
| portLastUpdate | The last time any of the port attributes of this entry were modified |

## The lhostTable Attributes

The logical hosts table consists of entries for each logical host configured in the cluster. See Table C–5 for the lhostTable attributes.

**TABLE C–5** lhostTable Attributes

| Attribute Names | Description |
|---|---|
| lhostName | The name of the logical host |
| lhostMasters | The list of node names that constitute the logical host |

**TABLE C–5** `lhostTable` Attributes *(continued)*

| Attribute Names | Description |
|---|---|
| `lhostCurrMaster` | The name of the node that is currently the master for the logical host |
| `lhostDS` | The list of data services configured to run on this logical host |
| `lhostDG` | The disk groups configured on this logical host |
| `lhostLogicalIP` | The logical IP address associated with this logical host |
| `lhostStatus` | The current status of the logical host (UP or DOWN) |
| `lhostLastUpdate` | The last time any of the attributes of this entry were modified |

## The `dsTable` Attributes

The data services table consists of entries for all data services that are configured for all logical hosts in the monitored clusters. Each entry in the table consists of specific information about a data service configured on a logical host. See Table C–6 for the `dsTable` attributes.

**Note -** When using a cross-reference, note that the `dsonLhost` attribute acts as a key reference between this table and the `lhostTable`.

**TABLE C–6** `dsTable` Attributes

| Attribute Names | Description |
|---|---|
| `dsName` | The name of the data service. |
| `dsOnLhost` | The name of the logical host on which the data service is configured. |
| `dsReg` | The value is 1 or 0 depending on whether the data service is registered and configured to run (1) or not run (0). |

**TABLE C–6** dsTable Attributes *(continued)*

| Attribute Names | Description |
|---|---|
| dsStatus | The current status of the data service (ON / OFF / INST DOWN). |
| dsDep | The list of other data services on which this data service depends. |
| dsPkg | The package name for the data service. |
| dsLastUpdate | The last time any of the attributes of this entry were last modified. |

## The dsinstTable Attributes

The data service instance table consists of entries for all data service instances. See Table C–7 for the dsinstTable attributes.

**Note -** When using a cross-reference, note that the dsinstOfDS attribute can be used as a key reference between this table and the dsTable. Similarly, the dsinstOnLhost attribute can be used as a key reference between this table and the lhostTable.

**TABLE C–7** dsinstTable Attributes

| Attribute Names | Description |
|---|---|
| dsinstName | The name of the data service instance |
| dsinstOfDS | The name of the data service of which this is a data service instance |
| dsinstOnLhost | The name of the logical host on which this data service instance is running |
| dsinstStatus | The status of the data service instance |
| dsinstLastUpdate | The last time any of the attributes of this entry were modified |

# Cluster SNMP Daemon and Super Monitor Daemon Operation

The SNMP daemon operates under the following provisions:

- The `smond` daemon connects to `in.mond` on all of the requested cluster nodes.
- The `smond` daemon passes the collected `config` and `syslog` information to the `snmpd` daemon.
- The `snmpd` daemon fills in the cluster MIB tables (which are available to clients through `SNMP GET` operations).
- The `snmpd` daemon sends out enterprise-specific traps for critical cluster events when notified by `smond` `syslog` data.

# SNMP Traps

SNMP traps are asynchronous notifications generated by the SNMP agent that indicate an unintended change in the state of monitored objects.

The software generates Sun Cluster-specific traps for critical cluster events. These events are listed in the following tables.

Table C–8 lists the Sun Cluster traps reflecting the state of the cluster software on a node.

**TABLE C–8** Sun Cluster Traps Reflecting the Software on a Node

| Trap Number | Trap Name |
|---|---|
| 0 | `sc:stopped` |
| 1 | `sc:aborted` |
| 2 | `sc:in_transition` |
| 3 | `sc:included` |

**TABLE C–8** Sun Cluster Traps Reflecting the Software on a Node *(continued)*

| Trap Number | Trap Name |
|---|---|
| 4 | `sc:excluded` |
| 5 | `sc:unknown` |

Table C–9 lists the Sun Cluster traps reflecting the state of the volume manager on a node.

**TABLE C–9** Sun Cluster Traps Reflecting the Volume Manager on a Node

| Trap Number | Trap Name |
|---|---|
| 10 | `vm:down` |
| 11 | `vm:up` |
| 12 | `vm:unknown` |

Table C–10 lists the Sun Cluster traps reflecting the state of the database on a node.

**TABLE C–10** Sun Cluster Traps Reflecting the Database on a Node

| Trap Number | Trap Name |
|---|---|
| 20 | `db:down` |
| 21 | `db:up` |
| 22 | `db:unknown` |

Table C–11 lists the Sun Cluster traps reflecting the nature of VxVM cluster feature (master or slave) on a node.

**TABLE C–11** Sun Cluster Traps Reflecting VxVM on a Node

| Trap Number | Trap Name |
|---|---|
| 30 | `vm_on_node:master` |
| 31 | `vm_on_node:slave` |
| 32 | `vm_on_node:unknown` |

Table C–12 lists the Sun Cluster traps reflecting the states of a logical host.

**TABLE C–12** Sun Cluster Traps Reflecting the States of a Logical Host

| Trap Number | Trap Name |
|---|---|
| 40 | `lhost:givingup` |
| 41 | `lhost:given` |
| 42 | lhost:takingover |
| 43 | lhost:taken |
| 46 | `lhost:unknown` |

Table C–13 lists the Sun Cluster traps reflecting the states of a data service instance.

**TABLE C–13** Sun Cluster Traps Reflecting the States of a Data Service Instance

| Trap Number | Trap Name |
|---|---|
| 50 | `ds:started` |
| 51 | `ds:stopped` |
| 52 | ds:in-transition |
| 53 | ds:failed-locally |

**TABLE C–13** Sun Cluster Traps Reflecting the States of a Data Service Instance *(continued)*

| Trap Number | Trap Name |
|---|---|
| 54 | ds:failed-remotely |
| 57 | `ds:unknown` |

Table C–14 lists the Sun Cluster traps reflecting the states of the HA-NFS data service.

**TABLE C–14** Sun Cluster Traps Reflecting the States of the HA-NFS Data Service Instance

| Trap Number | Trap Name |
|---|---|
| 60 | `hanfs:start` |
| 61 | `hanfs:stop` |
| 70 | `hanfs:unknown` |

Table C–15 lists the Sun Cluster traps reflecting SNMP errors.

**TABLE C–15** Sun Cluster Traps Reflecting SNMP Errors

| Trap Number | Trap Name |
|---|---|
| 100 | `SOCKET_ERROR:node_out_of_system_resources` |
| 101 | `CONNECT_ERROR:node_out_of_system_resources` |
| 102 | `BADMOND_ERROR:node_running_bad/old_mond_version` |
| 103 | `NOMOND_ERROR:mond_not_installed_on_node` |
| 104 | `NOMONDYET_ERROR:mond_on_node_not_responding:node_may_be_rebooting` |
| 105 | `TIMEOUT_ERROR:timed_out_upon_trying_to_connect_to_nodes_mond` |
| 106 | `UNREACHABLE_ERROR:node’s_mond_unreachable:network_problems??` |

**TABLE C–15** Sun Cluster Traps Reflecting SNMP Errors *(continued)*

| Trap Number | Trap Name |
|---|---|
| 107 | READFAILED_ERROR:node_out_of_system_resources |
| 108 | NORESPONSE_ERROR:node_out_of_system_resources |
| 109 | BADRESPONSE_ERROR:unexpected_welcome_message_from_node's_mond |
| 110 | SHUTDOWN_ERROR:node's_mond_shutdown |
| 200 | Fatal:super_monitor_daemon(smond)_exited! |

For trap numbers 100-110, check the faulty node and fix the problem. For trap number 200, see "SNMP Troubleshooting" on page 313.

# Changing the snmpd.conf File

The snmpd.conf file is used for configuration information. Each entry in the file consists of a keyword followed by a parameter string. The default values in the file should suit your needs.

## ▼ How to Change the snmpd.conf FileHow to Change the snmpd.conf File

1. **Edit the** snmpd.conf **file.**

   For details on the descriptions of the keywords, refer to the snmpd(7) man page.

2. **After making any changes to the** snmpd.conf **file, stop the** smond **and** snmpd **programs, then restart the scripts by entering:**

```
# /opt/SUNWcluster/bin/smond_ctl stop
# /opt/SUNWcluster/bin/init.snmpd stop
# /opt/SUNWcluster/bin/init.snmpd start
```

**(continued)**

**(Continuation)**

```
# /opt/SUNWcluster/bin/smond_ctl start
```

An example `snmpd.conf` file follows.

```
sysdescr        Sun SNMP Agent, SPARCstation 10, Company
                             Property Number 123456
syscontact  Coby Phelps
sysLocation  Room 123
#
system-group-read-community     public
system-group-write-community    private
#
read-community  all_public
write-community all_private
#
trap            localhost
trap-community  SNMP-trap
#
#kernel-file    /vmunix
#
managers        lvs golden
```

# Configuring the Cluster SNMP Agent Port

By default, the cluster SNMP agent listens on User Datagram Protocol (UDP) Port 161 for requests from the SNMP manager, for example, SunNet Manager Console. You can change this port by using the `-p` option to the `snmpd` and `smond` daemons.

Both the `snmpd` and `smond` daemons must be configured on the *same* port in order to function properly.

**Caution -** If you are installing the cluster SNMP agent on an SSP or an Administrative workstation running Solaris 2.6 or compatible versions, always configure the `snmpd` and the `smond` programs on a port other than the default UDP port 161.

For example, with the SSP, the cluster SNMP agent interferes with the SSP SNMP agent which also uses UDP port 161. This interference could result in the loss of RAS features of the Sun Enterprise 10000 server.

## How to Configure the Cluster SNMP Agent PortHow to Configure the Cluster SNMP Agent Port

To configure the cluster SNMP agent on a port other than the default UDP Port 161, perform the following steps.

1. **Edit the** `/opt/SUNWcluster/bin/init.snmpd` **file and change the value of the** `CSNMP_PORT` **variable from 161 to the desired value.**

2. **Edit the** `/opt/SUNWcluster/bin/smond_ctl` **file and change the value of the** `CSNMP_PORT` **variable from 161 to the same value you chose in Step 1 on page 311.**

3. **Stop and then restart both the** `snmpd` **and** `smond` **daemons for the changes to take effect.**

```
# /opt/SUNWcluster/bin/smond_ctl stop
# /opt/SUNWcluster/bin/init.snmpd stop
# /opt/SUNWcluster/bin/smond_ctl start
# /opt/SUNWcluster/bin/init.snmpd start
```

**Note -** Configuration files specific to the SNMP manager may need to be edited for SNMP manager to become aware of the new port number. Refer to your SNMP manager documentation for more information. Alternatively, you can configure the master SNMP agent on the Administrative workstation to start the cluster SNMP proxy agent as a subagent on a port other than 161. See the *Solstice Enterprise Agents User's Guide* or the `snmpdx(1M)` man page for information on how to configure the master SNMP agent.

# Using the SNMP Agent With SunNet Manager

The cluster SNMP agent has been qualified with the SunNet Manager. Perform the following procedures prior to using SunNet Manager to monitor clusters.

---

**Note -** These procedures assume you are using the UDP port 161 for SNMP. If you changed the port number as described in “Configuring the Cluster SNMP Agent Port” on page 310, you need to run the SunNet Manager SNMP proxy agent, `na.snmp` to use the alternate port.

---

## ▼ How to Use the SNMP Agent With SunNet Manager to Monitor ClustersHow to Use the SNMP Agent With SunNet Manager to Monitor Clusters

1. **Copy the cluster MIB** `/opt/SUNWcluster/etc/sun.mib` **to** `/opt/SUNWconn/snm/agents/cluster.mib` **on the SunNet Manager console.**

2. **On the SunNet Manager console run** `mib2schema` **on the copied** `cluster.mib` **file:**

```
# /opt/SUNWconn/snm/bin/mib2schema cluster.mib
```

3. **On the Sun Cluster Administrative workstation, edit the** `snmpd.conf` **file and set the parameter string in the** `trap` **keyword to the name of the SunNet Manager console.**

   For more information on editing the snmpd.conf file, refer to “Changing the `snmpd.conf` File” on page 309.

4. **Run the** `smond_conf` **command on the Sun Cluster Administrative workstation for each cluster you want to monitor. For example:**

```
# /opt/SUNWcluster/bin/smond_conf -h [clustername ...]
```

5. **Set the proxy for** `cluster-snmp` **to be the name of the SunNet Manager console.**

**Note -** In order to monitor clusters, you must also monitor the Administrative workstation using SunNet Manager.

### ▼ How to Reconfigure smond to Monitor a Different ClusterHow to Reconfigure smond to Monitor a Different Cluster

You can reconfigure the smond daemon to monitor a different cluster.

1. **Stop the** snmpd **daemon by using:**

```
# /opt/SUNWcluster/bin/init.snmpd stop
```

2. **Reconfigure the** smond **daemon by using:**

```
# /opt/SUNWcluster/bin/smond_conf -h [clustername ...]
```

3. **Start the** snmpd **daemon by using:**

```
# /opt/SUNWcluster/bin/init.snmpd start
```

4. **Start the** smond **daemon by using:**

```
# /opt/SUNWcluster/bin/smond_ctl start
```

## SNMP Troubleshooting

If the Cluster MIB tables are not filled in your application, or if you receive trap number 200, be sure that the snmpd and smond daemons are running by entering:

```
# ps -ef | grep snmpd
# ps -ef | grep smond
```

You do not see any output if the daemons are not running.

If the daemons are not running, enter:

```
# /opt/SUNWcluster/bin/init.snmpd start
# /opt/SUNWcluster/bin/smond_ctl start
```

APPENDIX **D**

# Glossary

**Active server** A node in the Sun Cluster configuration that is providing highly available data services.

**Administrative workstation** A workstation that is either outside the cluster or one of the cluster nodes that is used to run cluster administrative software.

**Backup group** Used by network adapter failover (NAFO). A set of network adapters on the same subnet. Adapters within a set provide backup for each other.

**CCD quorum** The set of Cluster Configuration Databases needed to elect a valid and consistent copy of the Cluster Configuration Database.

**Cluster** Two to four nodes configured together to run either parallel database software or highly available data services.

**Cluster Configuration Database (CCD)** A highly-available, replicated database that can be used to store data for HA data services and other Sun Cluster configuration needs.

**Cluster interconnect** The private network interface between cluster nodes.

**Cluster Membership Monitor (CMM)** The software that maintains a consistent cluster membership roster to avoid database corruption and subsequent transmission of corrupted or inconsistent data to clients. When nodes join or leave the cluster, thus changing the membership, CMM processes on the nodes coordinate global reconfiguration of various system services.

**Cluster node** A physical machine that is part of a Sun cluster. Also referred to as a cluster host or cluster server.

**Cluster quorum** The set of cluster nodes that can participate in the cluster membership.

**Cluster reconfiguration** An ordered multistep process that is invoked whenever there is a significant change in cluster state, such as takeover, switchover, or a physical host reboot. During cluster reconfiguration, the Sun Cluster software coordinates all of the physical hosts that are up and communicating. Those hosts agree on which logical host(s) should be mastered by which physical hosts.

**Cluster pair topology** Two pairs of Sun Cluster nodes operating under a single cluster administrative framework.

**Cluster SNMP agent** The cluster Simple Network Management Protocol (SNMP) agent is used to monitor several clusters (a maximum of 32) at the same time.

**CMM quorum** See cluster quorum.

**Concatenation** A metadevice created by sequentially mapping blocks on several physical slices (partitions) to a logical device. Two or more physical components can be concatenated. The slices are accessed sequentially rather than interlaced (as with stripes).

**Data service** A network service that implements read-write access to disk-based data from clients on a network. NFS is an example of a data service. The data service may be composed of multiple processes that work together.

**Default master** The node that is configured to master a disk group when the logical hosts are configured.

**Direct attached device** A disk storage unit that is physically connected to all nodes in the cluster.

**Distributed Lock Manager (DLM)** Locking software used in a shared disk Oracle7 or Oracle8 Parallel Server (OPS) environment. The DLM enables Oracle processes running on different nodes to synchronize database access. The DLM is designed for high availability; if a process or node crashes, the remaining nodes do not have to be shut down and restarted. A quick reconfiguration of the DLM is performed to recover from such a failure.

**Disk expansion unit** The physical storage enclosure that holds the multihost disks. For example, SPARCstorage Arrays, Sun StorEdge MultiPacks, Sun StorEdge A3000s and Sun StorEdge A5000s.

**Disk group**
A well defined group of multhost disks that move as a unit between two servers in an HA configuration. This can be either a Solstice DiskSuite diskset or a VERITAS Volume Manager disk group.

**Diskset**
See disk group.

**DiskSuite state database**
A replicated database that is used to store the configuration of metadevices and the state of these metadevices.

**Fault detection**
Sun Cluster programs that detect two types of failures. The first type includes low-level failures such as system panics and hardware faults (that is, failures that cause the entire server to be inoperable). These failures can be detected quickly. The second type of failures are related to data service. These types of failures take longer to detect.

**Fault monitor**
A fault daemon and the programs used to probe various parts of data services.

**Fibre channel connections**
Fibre connections connect the nodes with the SPARCstorage Arrays.

**Golden mediator**
In Solstice DiskSuite configurations, the in-core state of a mediator host set if specific conditions were met when the mediator data was last updated. The state permits take operations to proceed even when a quorum of mediator hosts is not available.

**HA Administrative file system**
A special file system created on each logical host when Sun Cluster is first installed. It is used by Sun Cluster and by layered data services to store copies of their administrative data.

**Heartbeat**
A periodic message sent between the several membership monitors to each other. Lack of a heartbeat after a specified interval and number of retries may trigger a takeover.

**Highly available data service**
A data service that appears to remain continuously available, despite single-point failures of server hardware or software components.

**Host**
A physical machine that can be part of a Sun cluster. In Sun Cluster documentation, host is synonymous with node.

**Hot standby server**
In an N+1 configuration, the node that is connected to all multihost disks in the cluster. The hot standby is also the administrative node. If one or more active nodes fail, the data services move from the failed node to the hot standby. However, there is no requirement that the +1 node cannot run data services in normal operation.

**Local disks**

Disks attached to a HA server but not included in a diskset. The local disks contain the Solaris distribution and the Sun Cluster and volume management software packages. Local disks must not contain data exported by the Sun Cluster data service.

**Logical host**

A set of resources that moves as a unit between HA servers. In the current product, the resources include a collection of network host names and their associated IP addresses plus a group of disks (a diskset). Each logical host is mastered by one physical host at a time.

**Logical host name**

The name assigned to one of the logical network interfaces. A logical host name is used by clients on the network to refer to the location of data and data services. The logical host name is the name for a path to the logical host. Because a host may be on multiple networks, there may be multiple logical host names for a single logical host.

**Logical network interface**

In the Internet architecture, a host may have one or more IP addresses. HA configures additional logical network interfaces to establish a mapping between several logical network interfaces and a single physical network interface. This allows a single physical network interface to respond to multiple logical network interfaces. This also enables the IP address to move from one HA server to the other in the event of a takeover or `haswitch(1M)`, without requiring additional hardware interfaces.

**Master**

The server with exclusive read and write access to a diskset. The current master host for the diskset runs the data service and has the logical IP addresses mapped to its Ethernet address.

**Mediator**

In a dual-string configuration, provides a “third vote” in determining whether access to the metadevice state database replicas can be granted or must be denied. Used only when exactly half of the metadevice state database replicas are accessible.

**Mediator host**

A host that is acting in the capacity of a “third vote” by running the `rpc.metamed(1M)` daemon and that has been added to a diskset.

**Mediator quorum**

The condition achieved when half + 1 of the mediator hosts are accessible.

**Membership monitor**

A process running on all HA servers that monitors the servers. The membership monitor sends and receives heartbeats to its sibling hosts. The monitor can initiate a takeover if the heartbeat stops. It also keeps track of which servers are active.

**Metadevice** A group of components accessed as a single logical device by concatenating, striping, mirroring, or logging the physical devices. Metadevices are sometimes called pseudo devices.

**Metadevice state database** Information kept in nonvolatile storage (on disk) for preserving the state and configuration of metadevices.

**Metadevice state database replica** A copy of the state database. Keeping multiple copies of the state database protects against the loss of state and configuration information. This information is critical to all metadevice operations.

**Mirroring** Replicating all writes made to a single logical device (the mirror) to multiple devices (the submirrors), while distributing read operations. This provides data redundancy in the event of a failure.

**Multihomed host** A host that is on more than one public network.

**Multihost disk** A disk configured for potential accessibility from multiple servers. Sun Cluster software enables data on a multihost disk to be exported to network clients via a highly available data service.

**Multihost disk expansion unit** See *Disk expansion unit*.

**N to N topology** All nodes are directly connected to a set of shared disks.

**N+1 topology** Some number (N) active servers and one (+1) hot-standby server. The active servers provide on-going data services and the hot-standby server takes over data service processing if one or more of the active servers fail.

**Node** A physical machine that can be part of a Sun cluster. In Sun Cluster documentation, it is synonymous with host or node.

**Nodelock** The mechanism used in greater than two-node clusters using VERITAS Volume Manager to failure fence failed nodes.

**Parallel database** A single database image that can be accessed concurrently through multiple hosts by multiple users.

**Partial failover** Failing over a subset of logical hosts mastered by a single physical host.

**Potential master** Any physical host that is capable of mastering a particular logical host.

**Primary logical host name** The name by which a logical host is known on the primary public network.

**Primary physical host name** The name by which a physical host is known on the primary public network.

**Primary public network** A name used to identify the first public network.

**Private links** The private network between nodes used to send and receive heartbeats between members of a server set.

**Quorum device** In VxVM configurations, the system votes by majority quorum to prevent network partitioning. Since it is impossible for two nodes to vote by majority quorum, a quorum device is included in the voting. This device could be either a controller or a disk.

**Replica** See metadevice state database replica.

**Replica quorum** A Solstice DiskSuite concept; the condition achieved when HALF + 1 of the metadevice state database replicas are accessible.

**Ring topology** One primary and one backup server is specified for each set of data services.

**Scalable Coherent Interface** A high speed interconnect used as a private network interface.

**Scalable topology** See *N to N topology*.

**Secondary logical host name** The name by which a logical host is known on a secondary public network.

**Secondary physical host name** The name by which a physical host is known on a secondary public network.

**Secondary public network** A name used to identify the second or subsequent public networks.

**Server** A physical machine that can be part of a Sun cluster. In Sun Cluster documentation, it is synonymous with host or node.

**Sibling host** One of the physical servers in a symmetric HA configuration.

**Solstice DiskSuite** A software product that provides data reliability through disk striping, concatenation, mirroring, UFS logging, dynamic growth of metadevices and file systems, and metadevice state database replicas.

**Stripe** Similar to concatenation, except the addressing of the component blocks is non-overlapped and interlaced on the slices (partitions), rather than placed sequentially. Striping is used to gain performance. By striping data across disks on separate controllers, multiple controllers can access data simultaneously.

**Submirror** A metadevice that is part of a mirror. See also mirroring.

**Sun Cluster** Software and hardware that enables several machines to act as read-write data servers while acting as backups for each other.

**Switch Management Agent (SMA)** The software component that manages sessions for the SCI and Ethernet links and switches.

**Switchover** The coordinated moving of a logical host from one operational HA server to the other. A switchover is initiated by an administrator using the `haswitch(1M)` command.

**Symmetric configuration** A two-node configuration where one server operates as the hot-standby server for the other.

**Takeover** The automatic moving of a logical host from one HA server to another after a failure has been detected. The HA server that has the failure is forced to give up mastery of the logical host.

**Terminal Concentrator** A device used to enable an administrative workstation to securely communicate with all nodes in the Sun Cluster.

**Trans device** In Solstice DiskSuite configurations, a pseudo device responsible for managing the contents of a UFS log.

**UFS** An acronym for the UNIX® file system.

**UFS logging** Recording UFS updates to a log (the logging device) before the updates are applied to the UFS (the master device).

**UFS logging device** In Solstice DiskSuite configurations, the component of a transdevice that contains the UFS log.

**UFS master device** In Solstice DiskSuite configurations, the component of a transdevice that contains the UFS file system.