Sun Cluster 2.2 Software Installation Guide

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303-4900
U.S.A.

Part No: 806–1008
April 1 1999

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 and Solaris 7 software environments. When configured properly, the hardware and software together provide highly available data services. Sun Cluster™ depends upon the mirroring, diskset capabilities, and other functionality provided by a volume manager. Sun Cluster supports three volume managers, Solstice DiskSuite™, Sun StorEdge Volume Manager™ (SSVM), and Cluster Volume Manager (CVM).

This book documents the guidelines for planning and setting up the configuration, and the procedures for installing and configuring the Sun Cluster software. This book is intended to be used with the hardware and software books listed in Section 0.3 “Related Documentation” on page 0-xiv.

Who Should Use This Book

This book is intended for Sun representatives and others 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 knowledge of the UNIX® system and expertise with the volume manager software (Solstice DiskSuite, SSVM, or CVM) used with Sun Cluster.

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 divided into general sections that each cover a major installation topic. This includes cluster planning, cluster software installation and upgrade, and data services installation. Depending on the chapter, you may see only overview or task information, or a combination of both.

Most of the overview information about Sun Cluster is contained in the beginning chapters, and the other chapters provide step-by-step instructions on installation tasks to perform.

Each of the data service chapters contains information about installing and configuring the data service.

The appendixes include configuration worksheets, and information specific to different volume managers.

Related Documentation

The documents listed in Table P–1 contain relevant information for the Sun Cluster system administrator and service provider.

TABLE P–1

Product Family	Title	Part Number
Sun Cluster	*Sun Cluster 2.2 System Administration Guide*	805-4238
	Sun Cluster 2.2 API Developer's Guide	805-4241
	Sun Cluster 2.2 Error Messages Manual	805-4242
	Sun Cluster 2.2 Release Notes	805-4243
Hardware	*Sun Enterprise Cluster System Site Preparation, Planning, and Installation Guide*	805-6512
	Sun Enterprise Cluster Hardware Service Manual	805-6511

TABLE P–1 *(continued)*

Product Family	Title	Part Number
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
SSVM	*Sun StorEdge Volume Manager 2.6 User's Guide*	805-5705
	Sun StorEdge Volume Manager 2.6 System Administrator's Guide	805-5706
CVM	*Sun Cluster 2.2 Cluster Volume Manager Guide*	805-4240
Solaris	*SPARC: Installing Solaris Software*	801-6109
	Solaris System Administration Guide	805-3115

Typographic Conventions

Table P–2 describes the typographic conventions used in this book.

TABLE P–2

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

Shell Prompts in Command Examples

Table P–3 shows the default system prompt and superuser prompt for the C shell, Bourne shell, and Korn shell.

TABLE P–3

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, contact your service provider and provide the following information:

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

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

- `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

Understanding the Sun Cluster Environment

This chapter provides an overview of the Sun Cluster product.

1.1 Sun Cluster Overview

The Sun Cluster system is a software environment that provides high availability (HA) support for data services and parallel database access on a cluster of servers (Sun Cluster servers). The Sun Cluster servers run the Solaris 2.6 or Solaris 7 operating environment, Sun Cluster framework software, disk volume management software, and HA data services or parallel database applications (OPS or XPS).

Sun Cluster framework software provides hardware and software failure detection, Sun Cluster system administration, system *failover* and automatic restart of data services in the event of a failure. Sun Cluster software includes a set of HA data services and an Application Programming Interface (API) that can be used to create other HA data services by integrating them with the Sun Cluster framework.

Shared disk architecture used with Sun Cluster parallel databases provide increased availability by allowing users to simultaneously access a single database through several cluster nodes. If a node fails, users can continue to access the data through another node without any significant delay.

The Sun Cluster system uses Solstice DiskSuite, Sun StorEdge Volume Manager (SSVM), or Cluster Volume Manager (CVM) software to administer *multihost disks*—disks that are accessible from multiple Sun Cluster servers. The volume management software provides disk mirroring, concatenation, striping, and hot sparing. SSVM and CVM also provide RAID5 capability.

The purpose of the Sun Cluster system is to avoid the loss of service by managing failures. This is accomplished by adding hardware redundancy and software monitoring and restart capabilities; these measures reduce single points of failure in the system. A single-point failure is the failure of a hardware or software component that causes the entire system to be inaccessible to client applications.

With redundant hardware, every hardware component has a backup that can take over for a failed component. The fault monitors regularly probe the Sun Cluster framework and the highly available data services, and quickly detect failures. In the case of HA data services, HA fault monitors respond to failures either by moving data services running on a failed node to another node, or, if the node has not failed, by attempting to restart the data services on the same node.

Sun Cluster configurations tolerate the following types of single-point failures:

- Server operating environment failure because of a crash or a panic
- Data service failure
- Server hardware failure
- Network interface failure
- Disk media failure

1.2 Hardware Configuration Components

HA and parallel database configurations are composed of similar hardware and software components. The hardware components include:

- Cluster nodes
- Private interconnects
- Public networks
- Local disks
- Multihost disks
- Terminal Concentrator or System Service Processor (SSP)
- Administrative Workstation

Details on all of these components are described in the following sections.

1.2.1 Cluster Nodes

Cluster nodes are the Sun Enterprise™ servers that run data services and parallel database applications. Sun Cluster supports 2-, 3-, and 4-node clusters.

1.2.2 Cluster Interconnect

The cluster interconnect provides a reliable internode communication channel used for vital locking and heartbeat information. The interconnect is used for maintaining cluster availability, synchronization, and integrity. The cluster interconnect is composed of two private links. These links are redundant; only one is required for cluster operation. If all nodes are up and a single private interconnect is lost, cluster operation will continue. However, when a node joins the cluster, both private interconnects must be operational for the join to complete successfully.

Note - By convention throughout this guide the network adapter interfaces `hme1` and `hme2` are shown as the cluster interconnect. Your interface names can vary depending on your hardware platform and your private network configuration. The requirement is that the two private interconnects do not share the same controller and thus cannot be disrupted by a single point of failure.

Clusters can use either the Scalable Coherent Interface (SCI) or Fast Ethernet as the private interconnect medium. However, support for mixed configurations (that is, both SCI and Ethernet private interconnects in the same cluster) is not supported.

1.2.2.1 The Switch Management Agent

The Switch Management Agent (SMA) is a cluster module that maintains communication channels over the private interconnect. It monitors the private interconnect and performs a failover of the logical adapter on the surviving private network if it detects a failure. In the case of more than one failure, SMA notifies the Cluster Membership Monitor which will take any action needed to change the cluster membership.

Clustered environments have different communication needs depending on the types of data services they support. Clusters providing only HA data services need only the heartbeat and minimal cluster configuration traffic over the private interconnect, and for these configurations Fast Ethernet is more than adequate. Clusters providing parallel database services send substantial amounts of traffic over the private interconnect. These applications benefit from the increased throughput of SCI.

SMA for SCI Clusters

The Scalable Coherent Interface (SCI) is a memory-based high-speed interconnect that enables sharing of memory among cluster nodes. The SCI private interconnect consists of Transmission Control Protocol/Internet Protocol (TCP/IP) network interfaces based on SCI.

Clusters of all sizes may be connected through a switch or hub. However, only two-node clusters may be connected point-to-point. The Switch Management Agent (SMA) software component manages sessions for the SCI links and switches.

There are three basic SCI topologies supported in Sun Cluster (Figure 1–1 and Figure 1–2):

- Three- or four-node cluster that requires two SCI switches
- Two-node cluster connected point-to-point
- Two-node switched cluster (degenerate case of the four-node cluster that allows for future expansion of cluster nodes with minimal interruption)

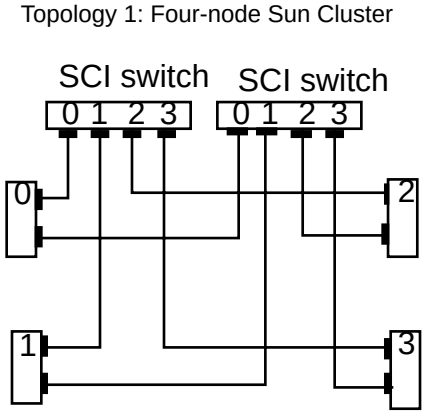
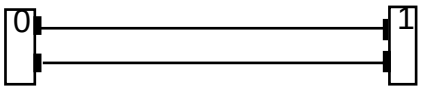


Figure 1–1 SCI Cluster Topology for Four Nodes

Topology 2: Two-node Sun Cluster with point-to-



Topology 3: Two-node Sun Cluster with

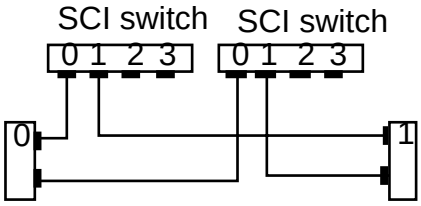


Figure 1–2 SCI Cluster Topologies for Two Nodes

SMA for Ethernet Clusters

Clusters of all sizes may be connected through a switch or hub. However, only two-node clusters may be connected point-to-point. The Switch Management Agent (SMA) software component manages communications over the Ethernet switches or hubs.

There are three basic Ethernet topologies supported in Sun Cluster (Figure 1–3 and Figure 1–4):

- Three- or four-node cluster that requires two Ethernet switches or hubs
- Two-node point-to-point cluster
- Two-node cluster with Ethernet switches or hubs (a degenerate case of the four-node cluster that allows for future expansion of cluster nodes with minimal interruption)

Topology 1: Four-node

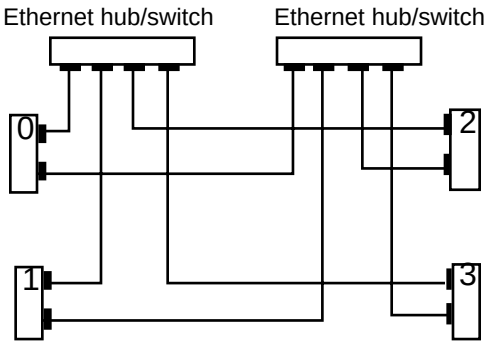
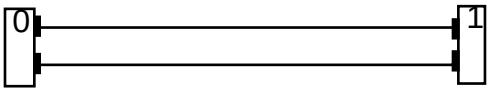


Figure 1–3 Ethernet Cluster Topology for Four Nodes

Topology 2: Two-node point-to-point, Ethernet



Topology 3: Two-node switched, Ethernet

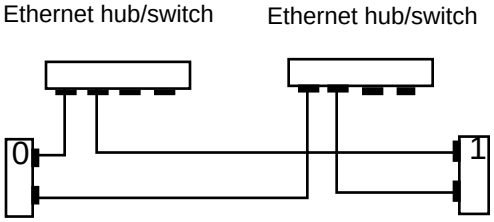


Figure 1–4 Ethernet Cluster Topologies for Two Nodes

1.2.3 `/etc/nsswitch.conf` File Entries

You must modify the `/etc/nsswitch.conf` file to ensure that "services," "group," and "hosts" lookups are always directed to the `/etc` files. This is done as part of the Sun Cluster installation described in Chapter 3.

The following shows an example `/etc/nsswitch.conf` file using NIS+ as the name service:

```
services: files nisplus
```

This entry must be before other services entries. Refer to the `nsswitch.conf(4)` man page for more information.

You must update `/etc/nsswitch.conf` manually by using your favorite editor. You can use the Cluster Console to update all nodes at one time. Refer to the chapter on Sun Cluster administration tools in the *Sun Cluster 2.2 System Administration Guide* for more information on the Cluster Console.

1.2.4 Public Networks

Cluster access to a Sun cluster is achieved by connecting the cluster nodes to one or more public networks. You can have any number of public networks attached to your cluster nodes, but the public network(s) must connect to every node in the cluster, regardless of the cluster topology. Figure 1–5 shows a four-node configuration with a single public network (192.9.200). Each physical host has an IP address on the public network.

One public network is designated as the *primary public networ*k and other public networks are called *secondary public networks*. Each network is also referred to as a *subnetwork* or *subnet*. The physical network adapter (`hme0`) is also shown in Figure 1–5. By convention throughout this guide, `hme0` is shown for the primary public network interface. This can vary depending on your hardware platform and your public network configuration.

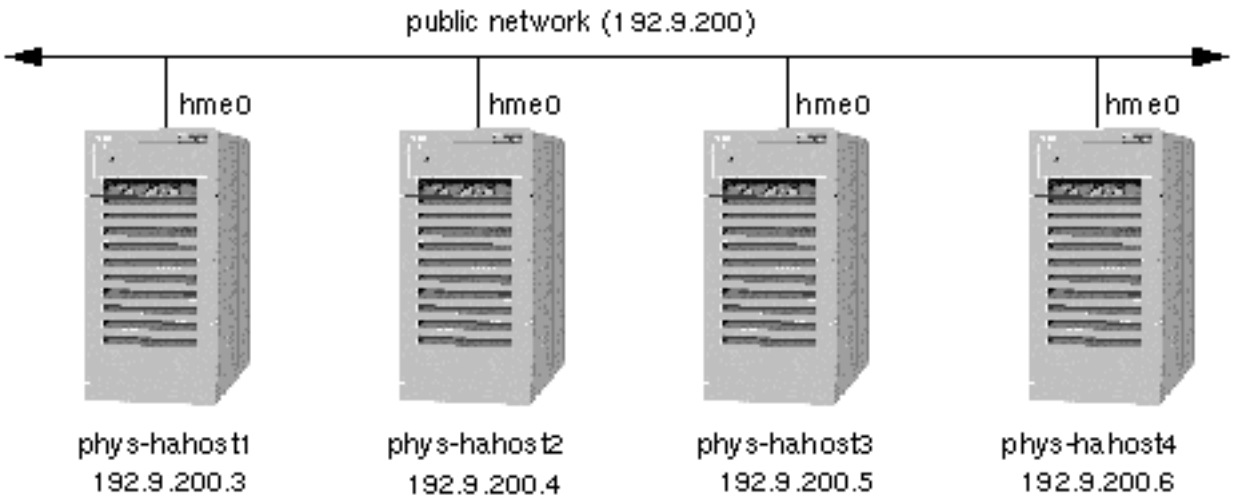


Figure 1–5 Four-Node Cluster With a Single Public Network Connection

Figure 1–6 shows the same configuration with the addition of a second public network (192.9.201). An additional physical host name and IP address must be assigned on each Sun Cluster server for each additional public network.

The names by which physical hosts are known on the public network are their *primary physical host names*. The names by which physical hosts are known on a secondary public network are their *secondary physical host names*. In Figure 1–6 the primary physical host names are labeled `phys-hahost[1-4]`. The secondary physical host names are labeled `phys-hahost[1-4]-201`, where the suffix `-201` identifies the network. Physical host naming conventions are described in more detail in Chapter 2.

The network adapter `hme3` is shown to be used by all nodes as the interface to the secondary public network. The adapter interface can be any suitable interface; `hme3` is shown here as an example.

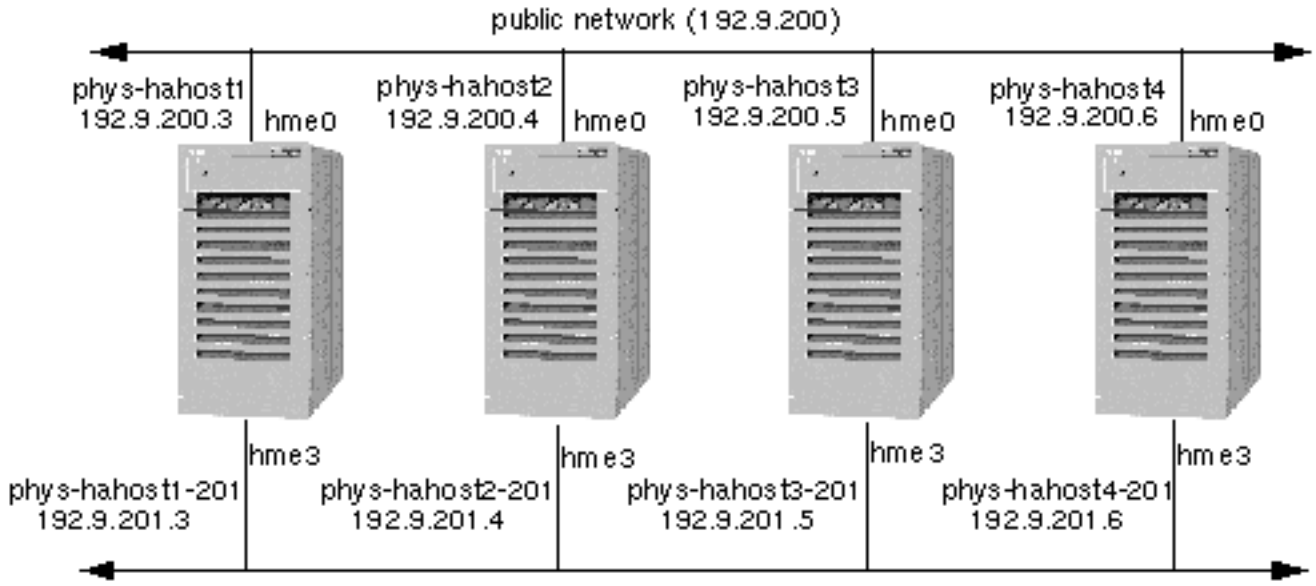


Figure 1–6 Four-Node Cluster With Two Public Networks

1.2.5 Local Disks

Each Sun Cluster server has one or more disks that are accessible only from that server. These are called *local disks*. They contain the Sun Cluster software environment and the Solaris operating environment.

Note - Sun Cluster supports the capability to boot from a disk inside a multihost SPARCstorage™ Array (SSA) and does not require a private boot disk. The Sun Cluster software supports SSAs that have both local (private) and shared disks.

Figure 1–7 shows a two-node configuration including the local disks.

Local disks can be mirrored, but mirroring is not required. Refer to Chapter 2, for a detailed discussion about mirroring the local disks.

1.2.6 Multihost Disks

In all Sun Cluster configurations, two or more nodes are physically connected to a set of shared, or *multihost*, disks. The shared disks are grouped across *disk expansion units*. Disk expansion units are the physical disk enclosures. Sun Cluster supports various disk expansion units: Sun StorEdge™ MultiPack, Sun StorEdge A3000, and Sun StorEdge A5000 units, for example. Figure 1–7 shows two hosts, both physically connected to a set of disk expansion units. It is not required that all cluster nodes are physically connected to all disk expansion units.

In HA configurations, the multihost disks contain the data for highly available data services. A server can access data on a multihost disk when it is the current master of that disk. In the event of failure of one of the Sun Cluster servers, the data services fail over to another server in the cluster. At failover, the data services that were running on the failed node are started on another node without user intervention and with only minor service interruption. The system administrator can switch over data services manually at any time from one Sun Cluster server to another. Refer to Section 1.5.10 “System Failover and Switchover” on page 1-42, for more details on failover and switchover.

In parallel database configurations, the multihost disks contain the data used by the relational database application. Multiple servers access the multihost disk simultaneously. User processes are prevented from corrupting shared data by the Oracle UNIX Dynamic Lock Manager (DLM). If one server connected to a multihost disk fails, the cluster software recognizes the failure and routes user queries through one of the remaining servers.

All multihost disks with the exception of the Sun StorEdge A3000 (with RAID5) must be mirrored. Figure 1–7 shows a multihost disk configuration.

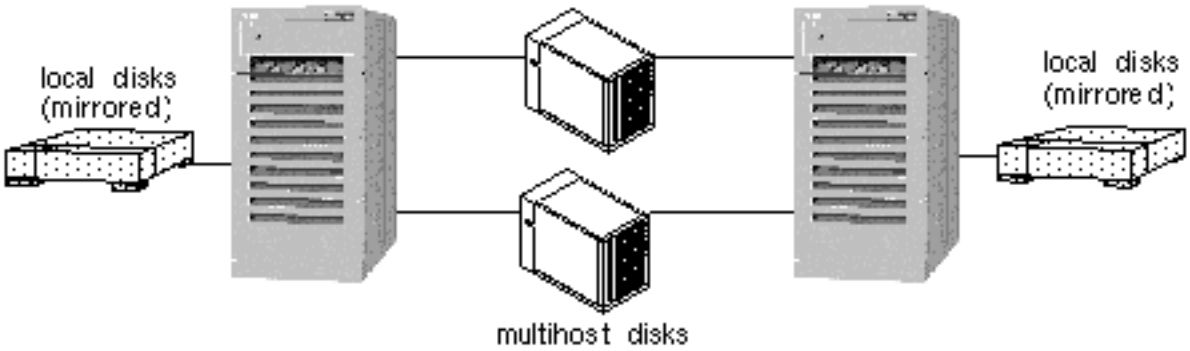


Figure 1–7 Local and Multihost Disks

1.2.7 Terminal Concentrator or System Service Processor and Administrative Workstation

The *Terminal Concentrator* is a device used to connect all cluster node console serial ports to a single workstation. The Terminal Concentrator turns the console serial ports on cluster nodes into telnet-accessible devices. You can telnet to an address on the Terminal Concentrator, and see a boot-PROM-prompt capable console window.

The *System Service Processor* (SSP) provides console access for Sun Enterprise 10000 servers. The SSP is a Solaris workstation on an Ethernet network that is especially configured to support the Sun Enterprise 10000. The SSP is used as the *administrative workstation* for Sun Cluster configurations using the Sun Enterprise 10000. Using the Sun Enterprise 10000 Network Console feature, any workstation in the network can open a host console session.

The Cluster Console connects a `telnet(1M)` session to the SSP, allowing you to log into the SSP and start a netcon session to control the domain. Refer to your Sun Enterprise 10000 documentation for more information on the SSP.

The Terminal Concentrator and System Service Processor are used to shut down nodes in certain failure scenarios as part of the failure fencing process. See Section 1.3.4.1 “Failure Fencing (SSVM and CVM)” on page 1-18, for more details.

The administrative workstation is used to provide console interfaces from all of the nodes in the cluster. This can be any workstation capable of running a Cluster Console session.

See the *Sun Cluster 2.2 System Administration Guide* and the Terminal Concentrator documentation for further information on these interfaces.

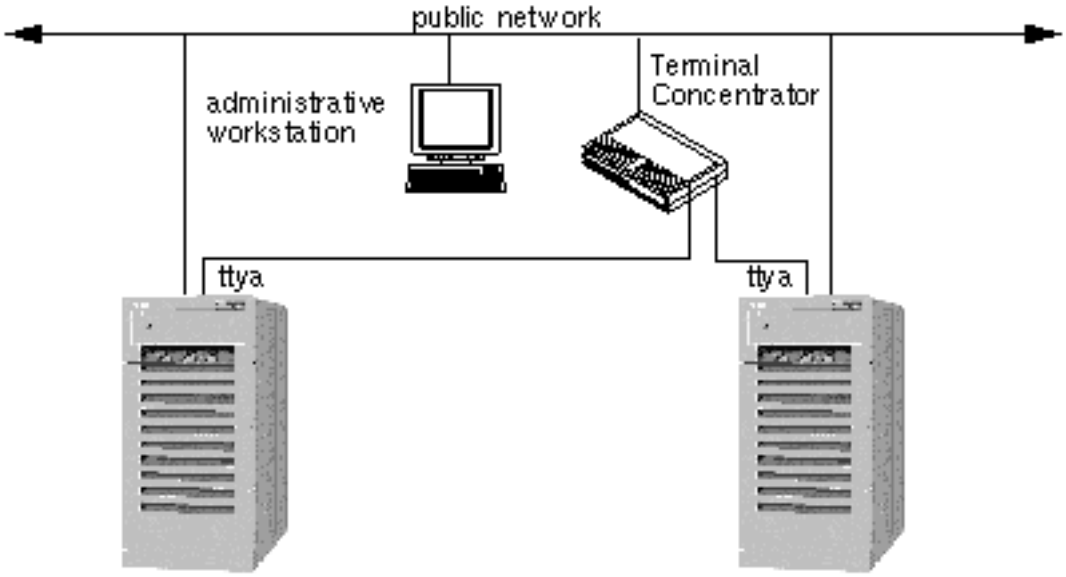


Figure 1–8 Terminal Concentrator and Administrative Workstation

1.3 Quorum, Quorum Devices, and Failure Fencing

Quorum is a term that is often used in the clustering world, and it is a concept that comes into play quite often in distributed systems. Fundamentally speaking, it is no different from the quorum that is required in Congress to pass a piece of legislation —obtaining majority consensus to agree on an issue. The notion of what number constitutes an acceptable quorum can vary from issue to issue; some may require a simple 50+ percent of the votes, while others may require a 2/3-majority. Exactly the same notion applies to a set of communicating processes in a distributed system. To ensure the system operates effectively and to make critical decisions about the behavior of the system, the set of processes need to agree on the desired quorum and then try to obtain consensus on some underlying issue by communicating messages until a quorum is obtained.

In Sun Cluster, two different types of quorums are used.

- The Cluster Membership Monitor (CMM) needs to obtain quorum about the set of cluster nodes that can participate in the cluster membership. The quorum is referred to as the *CMM quorum*, or *cluster quorum*.
- The Cluster Configuration Database (CCD) needs to obtain quorum to elect a valid and consistent copy of the CCD.

1.3.1 CMM Quorum

The Sun Cluster and Solstice HA clustering products determined CMM quorum by different methods. In previous Sun Cluster releases including Sun Cluster 2.0 and 2.1, the cluster framework determined CMM quorum. In Solstice HA, quorum was determined by the volume manager, Solstice DiskSuite. Sun Cluster 2.2 is an integrated release based on both Sun Cluster 2.x and Solstice HA 1.x. In Sun Cluster 2.2, determining CMM quorum depends on the volume manager, Solstice DiskSuite, SSVM, or CVM. If Solstice DiskSuite is the volume manager, CMM quorum is determined by a quorum of *metadevice state database replicas* managed by Solstice DiskSuite. If SSVM or CVM are used as the volume manager, CMM quorum is determined by the cluster framework.

For Sun Cluster 2.2, CMM quorum is determined by the following:.

- In clusters using SSVM and CVM as their volume manager, the cluster quorum is agreed upon based on the number of participating nodes and another independent device. In two-node clusters, a *quorum device* provides a third vote toward quorum. In greater-than-two-node clusters, an exclusive lock mechanism, a *nodelock*, is used to decide quorum if the cluster becomes split.
- In clusters using Solstice DiskSuite as the volume manager, cluster quorum is agreed upon based on metadevice state database replicas or *mediators*. In configurations with at least three disk strings, the metadevice state database replicas can always determine whether a node is part of the cluster quorum. In two-node configurations with only two disk strings, the concept of a mediator was developed. Mediators work in a similar manner to the quorum device in SSVM and CVM. Refer to the chapter on mediators in the *Sun Cluster 2.2 System Administration Guide* for details.

It is necessary to determine cluster quorum when nodes join or leave the cluster and in the event that the cluster interconnect (the redundant private links between nodes) fails. In Solstice HA 1.x, cluster interconnect failure was considered a double failure and the software guaranteed to preserve data integrity, but did not guarantee that the cluster could continue without user intervention. Manual intervention for dual failures was part of the system design. It was determined to be the safest method to ensure data integrity in contrast to an automatic response that might preserve availability but compromise data integrity.

The Sun Cluster 2.x software attempted to preserve data integrity and to also maintain cluster availability without user intervention. To preserve cluster availability, Sun Cluster 2.x implemented several new processes. These included using quorum devices, and the Terminal Concentrator or System Service Processor. Note that since Solstice HA 1,x used Solstice DiskSuite to determine cluster quorum, in Sun Cluster 2.2, the volume manager is the primary factor in determining cluster quorum and what occurs when the cluster interconnect fails. The results of a cluster interconnect failure are described in Section 1.3.3 “Quorum Devices (SSVM and CVM)” on page 1-14.

1.3.2 CCD Quorum

The Cluster Configuration Database (CCD) needs to obtain quorum to elect a valid and consistent copy of the CCD. Refer to Section 1.5.6 “Cluster Configuration Database” on page 1-36, for an overview of the CCD.

Sun Cluster does not have a storage topology that guarantees direct access from all cluster nodes to underlying storage devices for all configurations. This precludes the possibility of using a single logical volume to store the CCD database, which would guarantee that updates would be propagated correctly across restarts of the cluster framework.The CCD communicates with its peers through the cluster interconnect, and this logical link is unavailable on nodes that are not cluster members. We will illustrate the CCD quorum requirement with a simple example.

Assume a three-node cluster consisting of nodes A, B, and C. Node A exits the cluster leaving B and C as the surviving cluster members. The CCD is updated and the updates are propagated to nodes B and C. Now, nodes B and C leave the cluster. Subsequently, node A is restarted. However, A does not have the most recent copy of the CCD database because it has no means of knowing the updates that happened on nodes B and C after it left the cluster membership the last time around. In fact, irrespective of which node is started first, it is not possible to determine in an unambiguous manner which node has the most recent copy of the CCD database. Only when all three nodes are restarted is there sufficient information to determine the most recent copy of the CCD.If a valid CCD could not be elected, all query or update operations on the CCD would fail with an invalid CCD error. In practice, starting all cluster nodes before determining a valid copy of the CCD is too restrictive a condition.

This condition can be relaxed by imposing a restriction on the update operation. If N is the number of currently configured nodes in the cluster, at least floor(n/2)+1[1] nodes must be up for updates to be propagated. In this case, it is sufficient for ceiling(n/2)[2] identical copies to be present to elect a valid database on a cluster restart. The valid CCD is then propagated to all cluster nodes that do not already have it.

Note that even if the CCD is invalid, a node is allowed to join the cluster. However, the CCD can neither be updated or queried in this state. This implies that all components of the cluster framework that rely on the CCD remain in a dysfunctional state. In particular, logical hosts cannot be mastered and data services cannot be activated in this state. The CCD is enabled only after sufficient number of nodes join the cluster for quorum to be reached. Alternatively, an administrator can restore the CCD database with the maximum CCD generation number.

CCD quorum problems can be avoided if at least one or more nodes stay up during a reconfiguration. In this case, the valid copy on any of these nodes will be propagated to the newly joining nodes. Another alternative is to ensure that the cluster is started up on the node that has the most recent copy of the CCD database. Nevertheless, it is quite possible that after a system crash while a database update

1. floor(n) = n, if (n modulo 1 = 0), = n - (n modulo 1), if (n modulo 1 != 0)
2. ceiling(n) = n, if (n modulo 1 = 0), = n +1 - (n modulo 1), if (n modulo 1 != 0)

was in progress, the recovery algorithm finds inconsistent CCD copies. In such cases, it is the responsibility of the administrator to restore the database using the `ccdadm(1M)` restore option. The CCD also provides a checkpoint facility to backup the current contents of the database. It is good practice to make a backup copy of the CCD database after any change to system configuration. The backup copy can then be used to subsequently restore the database. The CCD is quite small compared to conventional relational databases and the backup and restore operations take no more than a few seconds to complete.

1.3.2.1 CCD Quorum in Two-Node Clusters

In the case of two-node clusters, the previously discussed quorum majority rule would require both nodes to be cluster members for updates to succeed, which is too restrictive. On the other hand, if updates are allowed in this configuration while only one node is up, the database will have to be manually made consistent before restarting the cluster. This can be accomplished by either restarting the node that has the most recent copy first, or restoring the database with the `ccdadm(1M)` restore operation after both nodes have joined. In the latter case, even though both nodes will be able to join the cluster membership, the CCD will be in an invalid state until the restore operation is complete.

This problem is solved by configuring persistent storage for the database on a shared disk device. The shared copy is used only when a single node is active. When the second node joins, the shared CCD copy is copied into the local copy on each node.

Whenever one of the nodes leave, the shared copy is reactivated by copying the local CCD into the shared copy. This enables updates only when a single node is in the cluster membership and also ensures reliable propagation of updates across cluster restarts.

The downside of using a shared storage device for the shared copy of the CCD is that two disks need to be allocated exclusively for this purpose, because the volume manager precludes these disks from being used for any other purpose. The usage of the two disks can be avoided if application downtime caused by the procedural limitations described above are understood and can be tolerated in a production environment.

Similar to the Sun Cluster 2.2 integration issues with the CMM quorum, a shared CCD is not supported in all Sun Cluster configurations. If Solstice DiskSuite is the volume manager, the shared CCD is not supported. Because the shared CCD is only used when one node is active, the failure addressed by the shared CCD is not common.

1.3.3 Quorum Devices (SSVM and CVM)

In certain cases—for example, in a two-node cluster when both cluster interconnects fail and both cluster nodes are still members—Sun Cluster needs assistance from a

hardware device to solve the problem of cluster quorum. This device is called the *quorum device*.

Quorum devices must be used in clusters running either Sun StorEdge Volume Manager (SSVM) or Cluster Volume Manager (CVM) as the volume manager, regardless of the number of cluster nodes. Solstice DiskSuite assures cluster quorum through the use of its own metadevice state database replicas, and as such, does not need a quorum device. Quorum devices are neither required nor supported in Solstice DiskSuite configurations. When you install a cluster using Solstice DiskSuite, the `scinstall(1M)` program will not ask for, or accept a quorum device.

The quorum device is merely a disk or a controller which is specified during the cluster installation procedure by using the `scinstall(1M)` command. The quorum device is a logical concept; there is nothing special about the specific piece of hardware chosen as the quorum device. However, the quorum device must be in its own disk group to be imported and exported independently. SSVM does not allow a portion of a disk to be in a separate disk group, so an entire disk and its plex (mirror) are required for the quorum device. Since you cannot be sure which node will have the quorum device imported at any time, it cannot usefully store data besides the data needed for the quorum.

A quorum device ensures that at any point in time only one node can update the multihost disks that are shared between nodes. The quorum device comes into use if the cluster interconnect is lost between nodes. Each node (or set of nodes in a greater than two-node cluster) should not attempt to update shared data unless it can establish that it is part of the majority quorum. The nodes take a vote, or quorum, to decide which nodes remain in the cluster. Each node determines how many other nodes it can communicate with. If it can communicate with more than half of the cluster, then it is in the majority quorum and is allowed to remain a cluster member. If it is not in the majority quorum, the node aborts from the cluster.

The quorum device acts as the “third vote” to prevent a tie. For example, in a two-node cluster, if the cluster interconnect is lost, each node will “race” to reserve the quorum device. Figure 1–9 shows a two-node cluster with a quorum device located in one of the multihost disk enclosures.

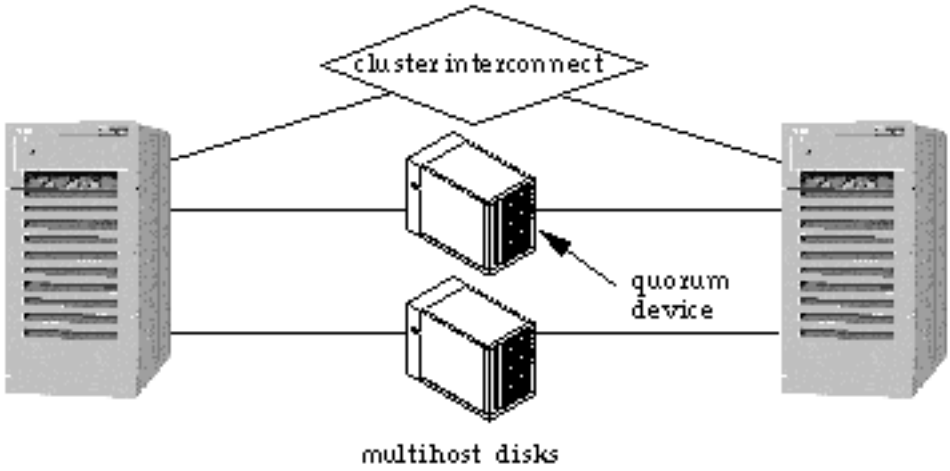


Figure 1–9 Two-node Cluster with Quorum Device

The node that reserves the quorum device then has two votes toward quorum versus the remaining node that has only one vote. The node with the quorum will then start its own cluster (mastering the multihost disks) and the other node will abort.

Before each cluster reconfiguration, the set of nodes and the quorum device vote to approve the new system configuration. Reconfiguration proceeds only if a majority quorum is reached. After a reconfiguration, a node remains in the cluster only if it is part of the majority partition.

Note - In greater than two-node clusters, each set of nodes that share access to multihost disks must be configured to use a quorum device.

The concept of quorum device changes somewhat in greater than two-node clusters. If there is an even split for nodes that do not share a quorum device—referred to as a "split-brain" partition—you must be able to decide which set of nodes will become a new cluster and which set will abort. This situation is not handed by the quorum device. Instead, as part of the installation process, when you configure the quorum device(s), you are asked questions that determine what will happen when such a partition occurs. One of two events occurs in this partition situation depending on whether you requested to have the cluster software automatically select the new cluster membership or whether you specified manual intervention.

- Automatic selection – If, during installation, you choose `select`, then the software automatically selects which subset is aborted based on either the `Lowest Nodeid` or the `Highest Nodeid`. If you chose the `Lowest Nodeid`, then the subset containing the node with the lowest node ID value automatically becomes the new cluster. If you chose the `Highest Nodeid`, then the subset containing the node with the highest node ID value automatically becomes the new cluster. You must manually abort all other subsets.

- Manual intervention – At the time of the partition, the system prompts you to choose the new cluster.

For example, consider a four-node cluster (that might or might not share a storage device common to all nodes) where a network failure results in node 0 and 1 communicating with each other and nodes 2 and 3 communicating with each other. In this situation, the automatic or manual decision of quorum would be used. The cluster monitor software is quite intelligent. It tries to determine on its own which nodes should be cluster members and which should not. It resorts to the quorum device to break a tie or the manual and automatic selection of cluster domains only in extreme situations.

Note - The failure of a quorum device is similar to the failure of a node in a two-node cluster.

The quorum device on its own cannot account for all scenarios where a decision must be made on cluster membership. For example, consider a fully operational three-node cluster, where all of the nodes share access to the multihost disks, such as the Sun StorEdge A5000. If one node aborts or loses both cluster interconnects, and the other two nodes are still able to communicate to each other, the two remaining nodes do not have to reserve the quorum device to break a tie. Instead, the majority voting that comes into play (two votes out of three) determines that the two nodes that can communicate with each other can form the cluster. However, the two nodes that form the cluster must still prevent the crashed or hung node from coming back online and corrupting the shared data. They do this by using a technique called *failure fencing*, as described in Section 1.3.4 “Failure Fencing ” on page 1-17.

1.3.4 Failure Fencing

In any clustering system, once a node is no longer in the cluster, it must be prevented from continuing to write to the multihost disks. Otherwise, data corruption could ensue. The surviving nodes of the cluster need to be able to start reading from and writing to the multihost disk. If the node that is no longer in the cluster is continuing to write to the multihost disk, its writes would confuse and ultimately corrupt the updates that the surviving nodes are performing.

Preventing a node that is no longer in the cluster from writing to the disk is called *failure fencing*. Failure fencing is very important for ensuring data integrity by preventing an isolated node from coming up in its own partition as a separate cluster when the actual cluster exists in a different partition.

Caution - It is very important to prevent the faulty node from performing I/O as the two cluster nodes now have very different views. The faulty node’s cluster view includes both cluster members (because it has not been reconfigured), while the surviving node’s cluster view consists of a one-node cluster (itself).

In a two-node cluster, if one node hangs or fails, the other node detects the missing heartbeats from the faulty node and reconfigures itself to become the sole cluster member. Part of this reconfiguration involves fencing the shared devices to prevent it from performing I/O on the multihost disks. In all Sun Cluster configurations with only two-nodes, this is accomplished through the use of SCSI-2 reservations on the multihost disks. The surviving node reserves the disks and prevents the failed node from performing I/O on the reserved disks. The semantics of the SCSI-2 reservation is that it is atomic in nature and if two nodes simultaneously attempt to reserve the device, one is guaranteed to succeed and the other guaranteed to fail.

1.3.4.1 Failure Fencing (SSVM and CVM)

Failure fencing is done differently depending on the cluster topology. The simplest case is a two-node cluster.

Failure Fencing Two-Node Clusters

In a two-node cluster, the quorum device determines which node remains in the cluster and the failed node is prevented from starting its own cluster because it cannot reserve the quorum device. SCSI-2 reservation is used to fence a failed node and prevent it from updating the multihost disks.

Failure Fencing Greater Than Two-Node Clusters

The difficulty with the SCSI-2 reservation model used in two-node clusters is that the SCSI reservations are host-specific. If a host has issued reservations on shared devices, it effectively shuts out every other node that can access the device, faulty or not. Consequently, this model breaks down when more than two nodes are connected to the multihost disks in a shared disk environment such as OPS.

For example, if one node hangs in a three-node cluster, the other two nodes reconfigure. However, neither of the surviving nodes can issue SCSI reservations to protect the underlying shared devices from the faulty node, as this action also shuts out the other surviving node. But without the reservations, the faulty node might “wake up” and issue I/O to the shared devices, despite the fact that its view of the cluster is no longer current.

Consider a four-node cluster with storage devices directly accessible from all the cluster nodes. If one node hangs, and the other three nodes reconfigure, none of them can issue the reservations to protect the underlying devices from the faulty node, as the reservations will also prevent some of the valid cluster members from issuing any I/O to the devices. But without the reservations, we have the real danger of the faulty node “waking up” and issuing I/O to shared devices despite the fact that its view of the cluster is no longer current.

Now consider the problem of split-brain situations. In the case of a four-node cluster, a variety of interconnect failures are possible. We will define a *partition* as a set of cluster nodes where each node can communicate with every other member within that partition, but not with any other cluster node that is outside the partition. There can be situations where, due to interconnect failures, two partitions are formed with two nodes in one partition and two nodes in the other partition, or with three nodes in one partition and one node in the other partition. Or there can even be cases where a four-node cluster can degenerate into four different partitions with one node in each partition. In all such cases, Sun Cluster attempts to arrive at a consistent distributed consensus on which partition should stay up and which partition should abort. Consider the following two cases.

Case 1. Two partitions with two nodes in each partition. As in the case of the one-one split in the case of a two-node cluster, the CMMs in either partition do not have quorum to conclude decisively which partition should stay up and which partition should abort. To meet the goals of data integrity and high availability, both partitions should not stay up and both partitions should not go down. As in the case of a two-node cluster, is it possible to adjudicate by means of an external device (the quorum disk) A designated node in each partition can race for the reservation on the designated quorum device, and whichever partition wins would be declared the winner. However, the node that successfully obtains the reservation on the quorum device shuts out the other node in its own partition from accessing the device due to the nature of the SCSI-2 reservation. because the quorum device contains useful data, this is not a desirable thing to do.

Case 2. Two partitions with three nodes in one partition and one node in the other partition. Even though the majority partition in this case has adequate quorum, the crux of the problem here is that the single isolated node has no idea what happened to the other three nodes. Perhaps they formed a valid cluster and this node should abort. But perhaps they did not; perhaps all three nodes did really fail for some reason. In this case, the single isolated node must stay up to maintain availability. With total loss of communication and without an external device to mediate, it is impossible to decide. Racing for the reservation of a configured external quorum device leads to a situation worse than in case 1. If one of the nodes in the majority partition reserves the quorum device, it shuts out the other two nodes in its own partition from accessing the device. But what is worse is that if the single isolated node wins the race for the reservation, it may lead to the loss of three potentially healthy nodes from the cluster. Once again, the disk reservation solution does not work well.

The inability to use the disk reservation technique also renders the system vulnerable to the formation of multiple independent clusters, each in its own isolated partition, in the presence of interconnect failures and operator errors. Consider case 2 above: Assume that the CMMs or some external entity somehow decides that the three nodes in the majority partition should stay up and the single isolated node should abort. Assume at some later point in time, the administrator now attempts to start up the aborted node without repairing the interconnect. The node would still be unable to communicate with any of the surviving members, and thinking it is the only node

in the cluster, attempt to reserve the quorum device. It would succeed because there are no quorum reservations in effect, due to the reasons elucidated above, and form its own independent cluster with itself as the sole cluster member.

Therefore, the simple quorum reservation scheme is unusable for three and four node clusters with storage devices directly accessible from all nodes. We need new techniques to solve the following three problems:

1. How to resolve all split-brain situations in three and four node clusters?
2. How do we failure fence faulty nodes from shared devices?
3. How do we prevent isolated partitions from forming multiple independent clusters?

To solve the different types of split-brain situations in three and four node clusters, a combination of heuristics and manual intervention is used, with the caveat that operator error during the manual intervention phase can destroy the integrity of the cluster. In Section 1.3.3 “Quorum Devices (SSVM and CVM)” on page 1-14, we discussed the policies that can be specified to determine what occurs in the event of a cluster partition for greater than two node clusters. If you choose the interventionist policy, the CMMs on all partitions will suspend all cluster operations in each partition waiting for manual operator input as to which partition is to continue to form a valid cluster and which partition should abort. It is the operator’s responsibility to let a desired partition to continue and to abort all other partitions. Allowing more than one partition to form a valid cluster can result in irretrievable data corruption.

If you choose a pre-deterministic policy, a preferred node (either the highest or lowest node id in the cluster) is requested, and when a split-brain situation occurs, the partition containing the preferred node will automatically become the new cluster if it is able to do so. All other partitions must be manually aborted by operator intervention. The selected quorum device is used solely to break the tie in the case of a split-brain for two-node clusters. Note that this situation can happen even in a configured four-node cluster, where only two cluster members are active when a split-brain situation occurs. The quorum device still plays a role, but in a much more limited capacity.

Once a partition has been selected to stay up, the next question is how to effectively protect the data from other partitions that should have aborted. Even though we require the operator to abort all other partitions, the command to abort the partition may not succeed immediately, and without an effective failure fencing mechanism, there is always the danger of hung nodes suddenly “waking up” and issuing pending I/O to shared devices before processing the abort request. In this case, the faulty nodes are reset before a valid cluster is formed in some partition.

To prevent a failed node from “waking up” and issuing I/O to the multihost disks, the faulty node is forcefully terminated by one of the surviving nodes, and dropped down to the OpenBoot PROM via the Terminal Concentrator or System Service Processor (Sun Enterprise 10000 systems), and the hung image of the operating system is terminated. This terminal operation prevents you from accidentally resuming a system by typing `go` at the Boot PROM. The surviving cluster members

wait for a positive acknowledgment from the termination operation before proceeding with the cluster reconfiguration process.

If there is no response to the termination command, then the hardware power sequencer (if present) is tripped to power cycle the faulty node. If tripping is not successful, then the system displays the following message requesting information to continue cluster reconfiguration:

```
/opt/SUNWcluster/bin/scadmin continuepartition localnode clustername
\007*** ISSUE ABORTPARTITION OR CONTINUEPARTITION ***
You must ensure that the unreachable node no longer has access to the
shared data. You may allow the proposed cluster to form after you have
ensured that the unreachable node has aborted or is down.
Proposed cluster partition:
```

Note - You should ensure that the faulty node has been successfully terminated before issuing the `scadmin continuepartition` command on the surviving nodes.

Partitioned, isolated, and terminated nodes do eventually boot up, and if due to some oversight, if the administrator tries to join the node into the cluster without repairing the interconnects, we must prevent this node from forming a valid cluster partition of its own if the node is unable to communicate with the existing cluster.

Assume a case where two partitions are formed with three nodes in one partition and one node in the other partition. A designated node in the majority partition terminates the isolated node and the three nodes form a valid cluster in their own partition. The isolated node, on booting up, tries to form a cluster of its own due to an administrator running the `startcluster(1M)` command and replying in the affirmative when asked for confirmation. Because the isolated node believes it is the only node in the cluster, it tries to reserve the quorum device and actually succeeds in doing so, because there are three nodes in the valid partition and none of them can reserve the quorum device without locking the others out.

To resolve this problem, a new concept is needed, a *nodelock*. For this concept, a designated node in the cluster membership opens a `telnet(1)` session to an unused port in the Terminal Concentrator as part of its cluster reconfiguration and keeps this session alive as long as it is a cluster member. If this node were to leave the membership, the nodelock is passed on to one of the remaining cluster members. In the above example, if the isolated node were to try to form its own cluster, it would try to acquire this lock and fail, because one of the nodes in the existing membership (in the other partition) would be holding this lock. If one of the valid cluster members is unable to acquire this lock for some reason, it would not be considered a fatal, but would be logged as an error requiring immediate attention. The locking facility should be considered as a safety feature rather than a mechanism critical to the operation of the cluster, and its failure should not be considered

catastrophic. In order to speedily detect faults in this area, monitoring processes in the Sun Cluster framework monitor whether the Terminal Concentrator is accessible from the cluster nodes.

1.3.4.2 *Failure Fencing (Solstice DiskSuite)*

In Sun Cluster configurations using Solstice DiskSuite as the volume manager, it is Solstice DiskSuite itself that determines cluster quorum and provides failure fencing. There is no distinction between different cluster topologies for failure fencing. That is, two-node and greater than two-node clusters are treated identically. This is possible for two reasons:

- Unlike the situation with Cluster Volume Manager, there is no concept of shared disks in the HA environment. At most, only one node can master a diskset at any given time. This precludes the situation where more than one node would need to access the diskset after a node fails.
- The split-brain situation, where the cluster interconnect fails is viewed as a double failure (both private links have failed). In the event of a split-brain situation, the guarantee is that data integrity will be maintained. There is no guarantee that the cluster will be able to continue without user intervention. For example, in a three-node cluster where all nodes are attached to the diskset, and split brain develops, it is possible that one node will crash, two nodes will crash, or all three nodes will crash. This algorithm has been shown to have higher actual availability than any algorithm employing a quorum device.

Disk fencing is accomplished in the following manner.

1. After a node is removed from the cluster, a remaining node does a `SCSI` reserve of the disk. After this, other nodes—including the one no longer in the cluster—are prevented by the disk itself to read or write to the disk. The disk will return a `Reservation_Conflict` error to the read or write command. In Solstice DiskSuite configurations, the `SCSI` reserve is accomplished by issuing the Sun multihost ioctl `MHIOCTKOWN`.
2. Nodes that are in the cluster continuously enable the `MHIOCENFAILFAST` ioctl for the disks that they are accessing. This ioctl is a directive to the disk driver, giving the node the capability to panic itself if it cannot access the disk due to the disk being reserved by some other node. The `MHIOCENFAILFAST` ioctl causes the driver to check the error return from every read and write that this node issues to the disk for the `Reservation_Conflict` error code, and it also periodically, in background, issues a test operation to the disk to check for `Reservation_Conflict`. Both the foreground and the background control flow paths will panic should `Reservation_Conflict` be returned.
3. The `MHIOCENFAILFAST` ioctl is not specific to dual-hosted disks. If the node that has enabled the `MHIOCENFAILFAST` for a disk loses access to that disk due to another node reserving the disk (by SCSI-2 exclusive reserve), the node panics.

This solution to disk fencing relies on the SCSI-2 concept of disk reservation, which requires that a disk be reserved by exactly one single node.

For Solstice DiskSuite configurations, the installation program `scinstall(1M)` does not prompt for a quorum device, a node preference, or to select a failure fencing policy as is done in SSVM and CVM configurations. When Solstice DiskSuite is specified as the volume manager, you cannot configure direct-attach devices, that is, devices that directly attach to more than two nodes. Disks can only be connected to pairs of nodes.

Note - Although the `scconf(1M)` command allows you to specify the `+D` flag to enable configuring direct-attach devices, you should not do so in Solstice DiskSuite configurations.

1.3.5 Preventing Partitioned Clusters (SSVM and CVM)

Two-Node Clusters

If lost interconnects occur in a two-node cluster, both nodes attempt to start the cluster reconfiguration process with only the local node in the cluster membership (because each has lost the heartbeat from the other node). The first node that succeeds in reserving the configured quorum device remains as the sole surviving member of the cluster. The node that failed to reserve the quorum device aborts.

If you try to start up the aborted node without repairing the faulty interconnect, the aborted node (which is still unable to contact the surviving node) attempts to reserve the quorum device, because it sees itself as the only node in the cluster. This attempt will fail because the reservation on the quorum device is held by the other node. This action effectively prevents a partitioned node from forming its own cluster.

Three- or Four-Node Clusters

If a node drops out of a four-node cluster as a result of a reset issued via the terminal concentrator (TC), the surviving cluster nodes are unable to reserve the quorum device, since the reservation by any other node prevents the two healthy nodes from accessing the device. However, if you erroneously ran the `scadmin startcluster` command on the partitioned node, the partitioned node would form its own cluster, since it is unable to communicate with any other node. There are no quorum reservations in effect to prevent it from forming its own cluster.

Instead of the quorum scheme, Sun Cluster resorts to a cluster-wide lock (nodelock) mechanism. An unused port in the TC of the cluster, or the SSP, is used. (Multiple TCs are used for campus-wide clusters.) During installation, you choose the TC or SSP for this node-locking mechanism. This information is stored in the CCD. One of

the cluster members always holds this lock for the lifetime of a cluster activation; that is, from the time the first node successfully forms a new cluster until the last node leaves the cluster. If the node holding the lock fails, the lock is automatically moved to another node.

The only function of the nodelock is to prevent operator error from starting a new cluster in a split-brain scenario.

Note - The first node joining the cluster aborts if it is unable to obtain this lock. However, node failures or aborts do not occur if the second and subsequent nodes of the cluster are unable to obtain this lock.

Node locking functions in this way:

- If the first node to form a new cluster is unable to acquire this lock, it aborts with the following message:

```
[SUNWcluster.reconf.nodelock.4002] $clustname Failed to obtain
NodeLock status = ??
```

- If the first node to form a new cluster acquires this lock, the following message is displayed:

```
[SUNWcluster.reconf.nodelock.1000] $clustname Obtained Nodelock
```

- If one of the current nodes in a cluster is unable to acquire this lock during the course of a reconfiguration, an error message is logged on the system console:

```
[SUNWcluster.reconf.nodelock.3004] $clustname WARNING: Failed to Force obtain
NodeLock status = ??
```

Caution - This message warns you that the lock could not be acquired. You need to diagnose and fix this error as soon as possible to prevent possible future problems.

- If a partitioned node tries to form its own cluster (by using the `scadmin startcluster` command), it is unable to acquire the cluster lock if the cluster is active in some other partition. Failure to acquire this lock causes this node to abort.

1.4 Configurations Supported by Sun Cluster

A cluster is composed of a set of physical hosts, or *nodes*. Throughout the Sun Cluster documentation, cluster nodes also are referred to as Sun Cluster servers.

The Sun Cluster hardware configuration supports *symmetric*, *asymmetric*, *clustered pairs*, *ring*, *N+1 (star)*, or *N to N* (*scalable)* topologies. Each of these is described in detail later in this chapter.

A symmetric configuration has only two nodes. Both servers are configured identically and, generally, both provide data services during normal operation. See Figure 1–12.

A two-node configuration where one server operates as the hot-standby server for the other is referred to as an asymmetric configuration. This configuration is treated as an N+1 configuration where N=1.

Clustered pairs are two pairs of Sun Cluster nodes operating under a single cluster administrative framework. See Figure 1–13.

The ring configuration allows for one primary and one backup server to be specified for each set of data services. All disk storage is dual-hosted and physically attached to exactly two cluster nodes. The nodes and storage are connected alternately, in a ring. This is ideal for configuring multiple online highly available data services. See Figure 1–14.

An N+1 or star configuration is composed of two or more nodes. One node in the N+1 configuration (the +1 node) might be configured to be inactive until there is a failure of another node. In this configuration, the +1 node operates as a “hot-standby.” The remaining nodes are “active” in normal operation. The examples in this chapter assume that the hot-standby node is not running data services in normal operation. However there is no requirement that the +1 node not run data services in normal operation. See Figure 1–15.

An N to N, or scalable configuration has all servers directly connected to a set of shared disks. This is the most flexible configuration because data services can fail over to any of the other servers. See Figure 1–16.

1.4.1 High Availability and Parallel Database Configurations

Sun Cluster supports HA data service and parallel database configurations. HA and parallel databases can also be combined within a single cluster, with some restrictions.

Data services in an HA configuration are made highly available by having multiple hosts connected to the same physical disk enclosure. The status of each host is monitored over private interconnects. If one of the hosts fails, another host connected to the same shared storage device can take over the data service work previously done by the failed host. Figure 1–10 shows an example of a highly available data service configuration.

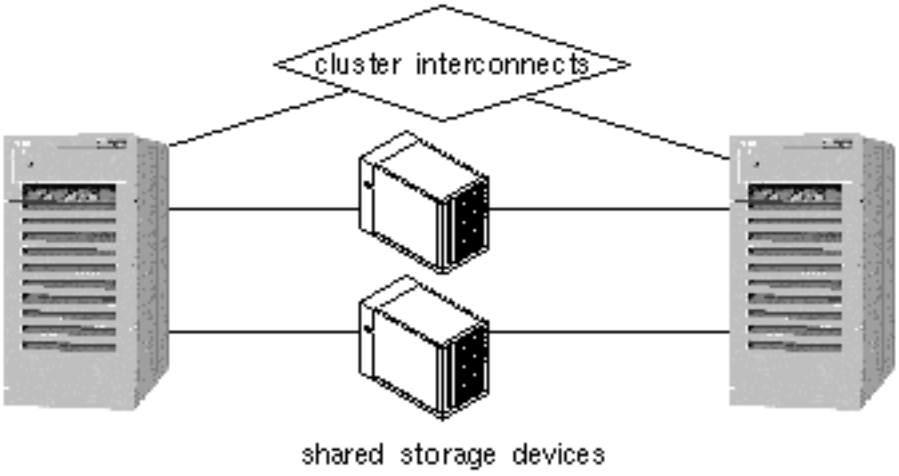


Figure 1–10 Highly Available Data Services Configuration

Oracle Parallel Server (OPS) enables a relational database to be highly available by enabling multiple hosts to access the same data on shared storage devices. Traffic to the shared disk enclosures is controlled by a DLM that prevents two processes from accessing the same data at the same time. High availability is attained by redirecting database access traffic from a failed host to one of the remaining nodes. Figure 1–11 shows an example of a highly available OPS configuration. The private interconnect can be either the Scalable Coherent Interface (SCI) or Fast Ethernet.

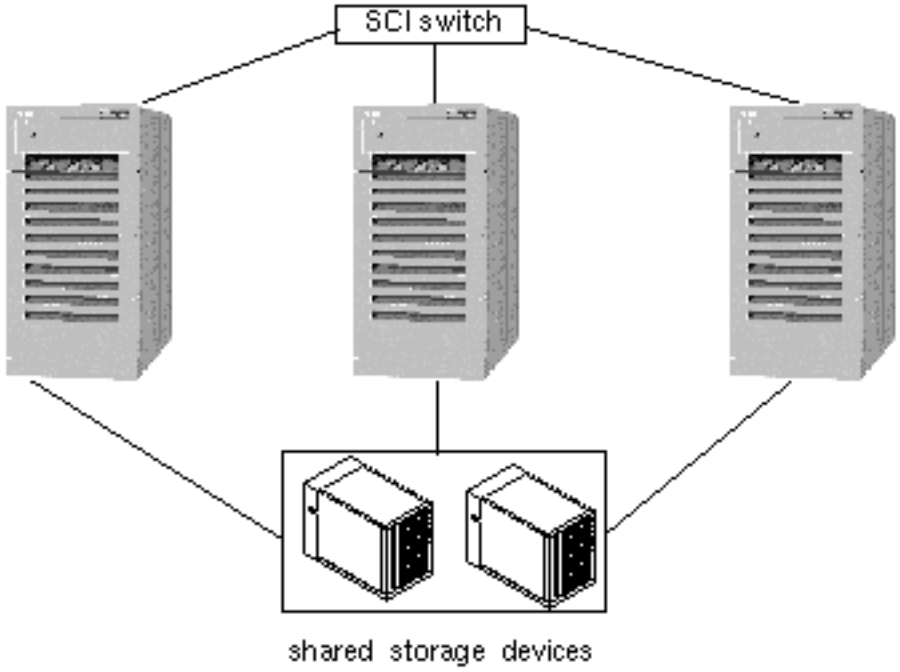


Figure 1–11 OPS Database Configuration

The Informix-Online XPS parallel database permits parallel access by partitioning the relational database across shared storage devices. Multiple host processes can access the same database simultaneously provided they do not access data stored in the same partition. Access to a particular partition is through a single host, so if that host fails, no access is possible to that partition of data. For this reason, Informix-Online XPS is a parallel database, but cannot be configured to be highly available in a Sun Cluster.

1.4.2 Symmetric and Asymmetric Configurations

Symmetric and asymmetric HA configuration, by definition, consists of exactly two nodes. Highly available data services run on one or both nodes. Figure 1–12 shows a two-node configuration. This example configuration consists of two active nodes (`phys-hahost1` and `phys-hahost2`) that are referred to as *siblings.*

Both nodes are physically connected to a set of multihost disks.

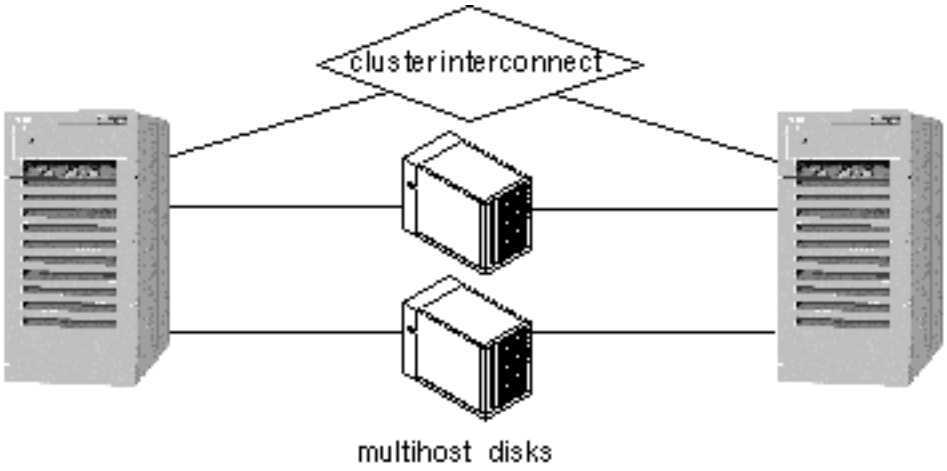


Figure 1–12 Two-node Configuration

1.4.3 Clustered Pairs Configuration

The clustered pairs configuration is a variation on the symmetric configuration. In this configuration, there are two pairs of servers, with each pair operating independently. However, all of the servers are connected by the private interconnects and are under control of the Sun Cluster software.

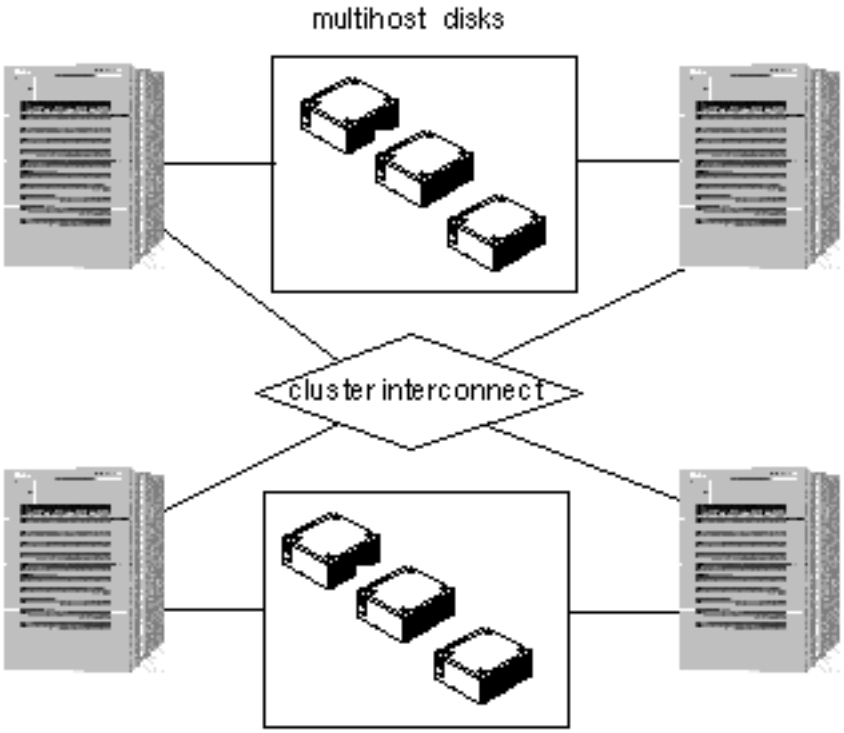


Figure 1–13 Clustered Pairs Configuration

Clustered pairs can be configured so that you can run a parallel database application on one pair and HA data services on the other. Failover is only possible across a pair.

1.4.4 Ring Configuration

The ring configuration allows for one primary and one backup server to be specified for each set of data services. The backup for a given data service is a node adjacent to the primary. All disk storage is dual-hosted and physically attached to exactly two cluster nodes. The nodes and storage are connected alternately, in a ring. All nodes can be actively running applications. Each node is both a primary for one set of data services and a backup for another set. Figure 1–14 shows a four-node ring configuration.

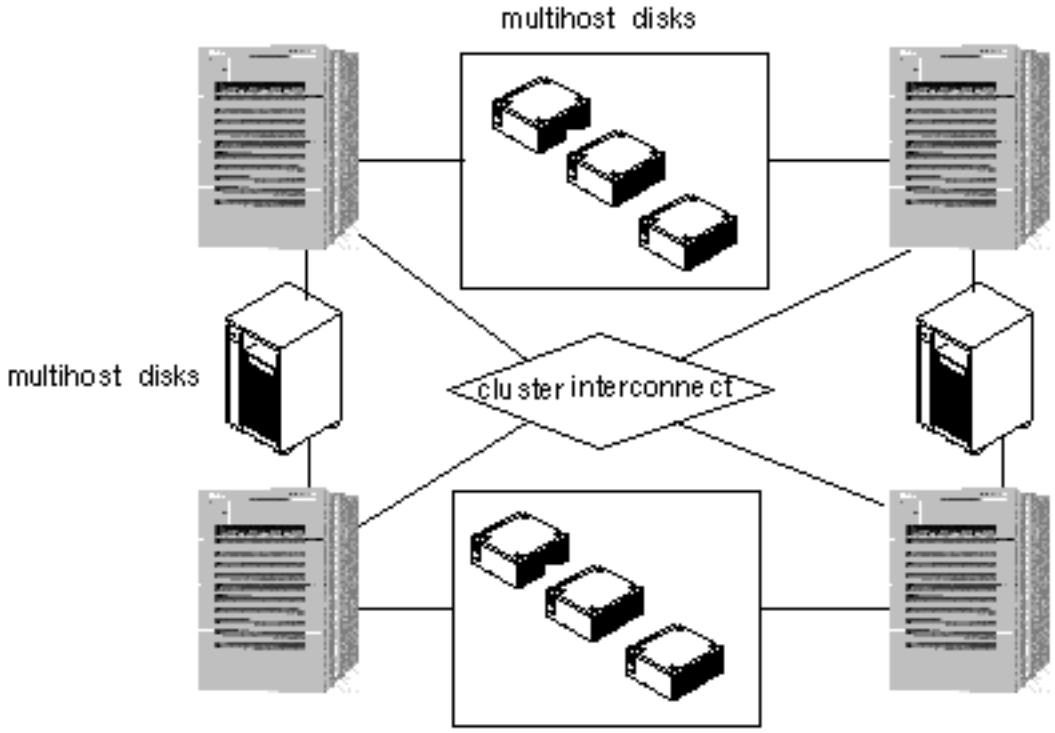


Figure 1–14 Ring Configuration

Note - A restriction to the ring configuration is that you cannot run multiple RDBMS data services on the same node.

1.4.5 N+1 Configuration (Star)

An N+1 or star configuration includes some number of *active servers* and one *hot-standby server*. The active servers and hot-standby server do not have to be configured identically. Figure 1–15 shows an N+1 configuration. The active servers provide on-going data services while the hot-standby server waits for one or more of the active servers to fail. The hot-standby server is the only server in the configuration that has physical disk connections to all disk expansion units.

In the event of a failure of one active server, the data services from the failed server fail over to the hot-standby server. The hot-standby server then continues to provide data services until the data services are switched over manually to the original active server.

The hot-standby server need not be idle while waiting for another Sun Cluster server to fail. However, the hot-standby server should always have enough excess CPU capacity to handle the load should one of the active servers fail.

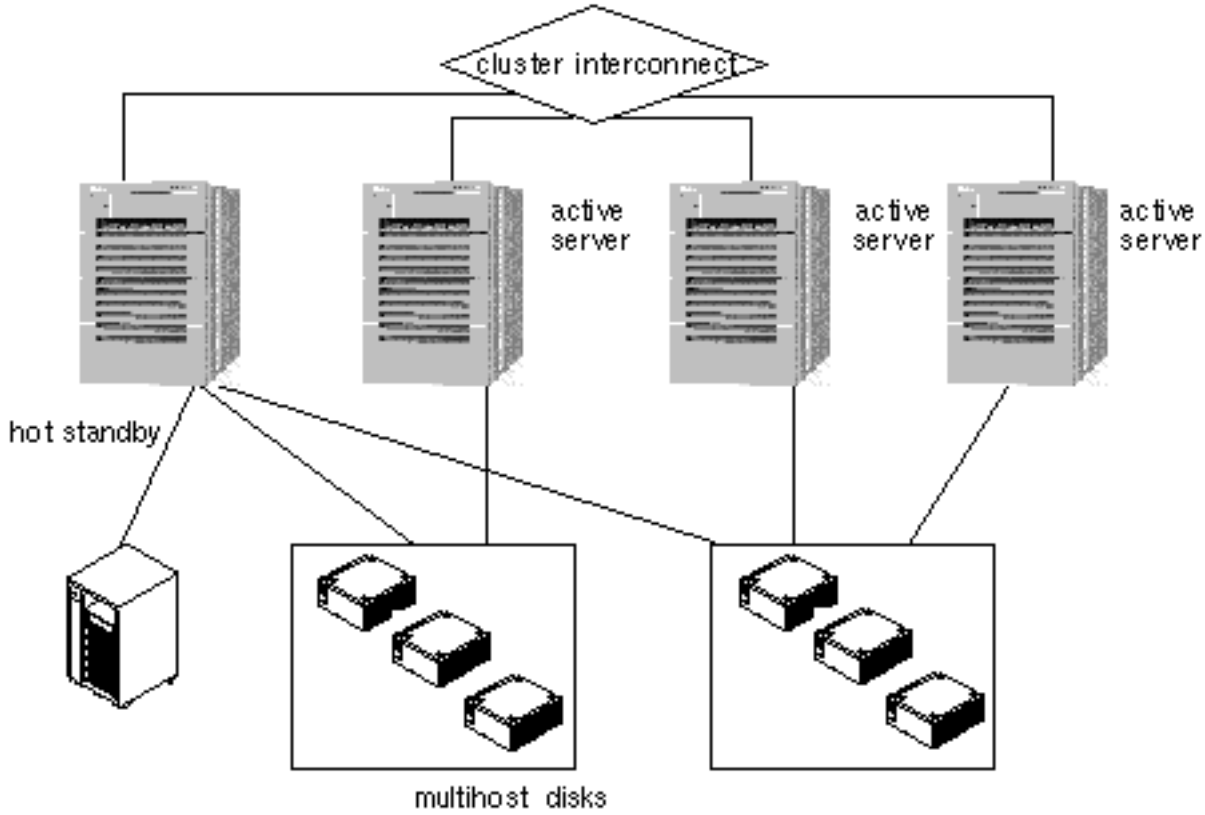


Figure 1–15 N+1 Configuration

1.4.6 N to N Configuration (Scalable)

An N to N or scalable configuration has all servers physically connected to all multihost disks. Data services can fail over from one node to a backup, to another backup, a feature known as cascading failover.

This is the highest redundancy configuration because data services can fail over to up to three other nodes.

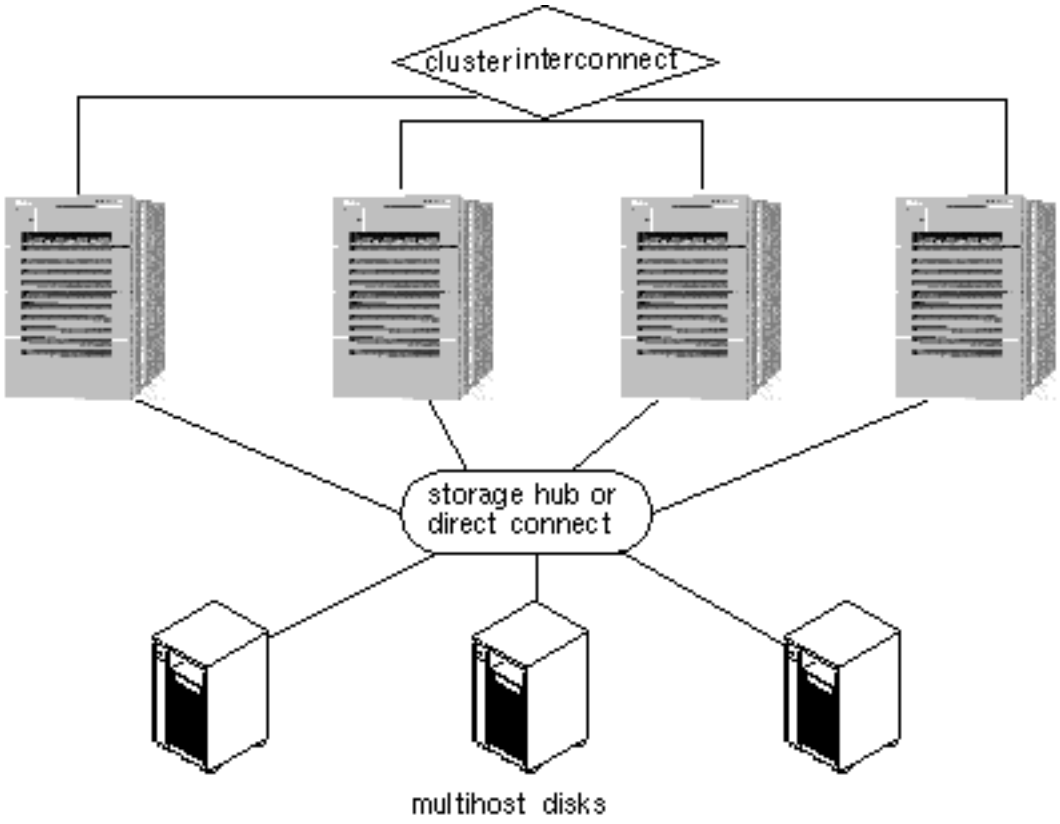


Figure 1–16 N to N Configuration

1.4.7 Campus Clustering

Sun Cluster features campus clustering, a cluster configuration that provides geographic site separation and enables recovery from certain types of failures, which is localized to a part of the campus.

The servers and storage devices may be physically located in the same server room, or geographically distributed across multiple sites. Geographical distribution improves protection of data from catastrophic failures, such as a fire, and thus improves overall data service availability.

For additional information on campus clustering, contact your local Sun sales representative.

1.5 Software Configuration Components

Sun Cluster includes the following software components:

- Cluster framework
- Data services

Associated with these software components are the following logical components:

- Logical hosts
- Disk groups

These components are described in the following sections.

1.5.1 Cluster Framework

Figure 1–17 shows the approximate layering of the various components that constitute the framework required to support HA data services in Sun Cluster. This diagram does not illustrate the relationship between the various components of Sun Cluster. The innermost core consists of the *Cluster Membership Monitor* (CMM), which keeps track of the current cluster membership. Whenever nodes leave or rejoin the cluster, the CMM on the cluster nodes go through a distributed membership protocol to agree on the new cluster membership. Once the new membership is established, the CMM orchestrates the reconfiguration of the other cluster components through the Sun Cluster framework.

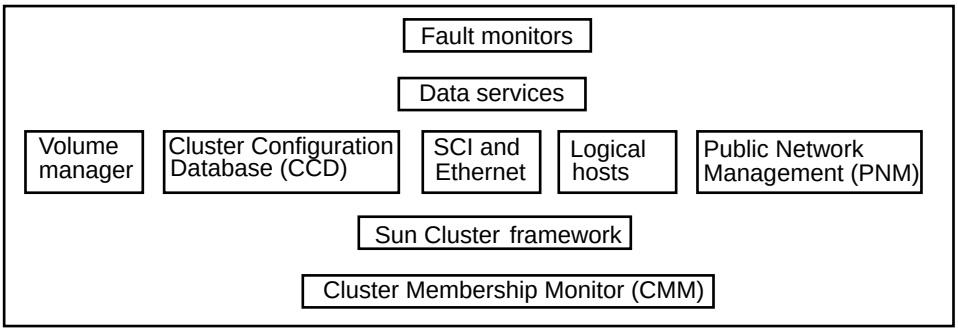


Figure 1–17 Sun Cluster Software Components

In an HA configuration, the membership monitor, *fault monitor*, and associated programs allow one Sun Cluster server to take over processing of all data services from the other Sun Cluster server when hardware or software fails. This is accomplished by causing a Sun Cluster server without the failure to take over mastery of the logical host associated with the failed Sun Cluster server. Some types

of failures do not cause failover. Disk drive failure does not typically result in a failover—mirroring handles this. Similarly, software failures detected by the fault monitors might cause a data service to be restarted on the same physical node rather than failing over to another node.

1.5.2 Fault Monitor Layer

The fault monitor layer consists of a fault daemon and the programs used to probe various parts of the data service. If the fault monitor layer detects a service failure, it can attempt to restart the service on the same node, or initiate a failover of the logical host, depending on how the data service is configured.

Under certain circumstances a data service fault monitor will not initiate a failover even though there has been an interruption of a service. These exceptions include:

- File systems under control of a logical host are being checked with `fsck(1M)`.
- The NFS™ file system is locked by using `lockfs(1M)`.
- The name service (NIS, NIS+, DNS) is not working. The name service exists outside the Sun Cluster configuration so you must ensure its reliability.

1.5.3 Data Services Layer

Sun Cluster includes a set of data services that have been made highly available by Sun. Sun Cluster provides a fault monitor at the data services layer. The level of fault detection provided by this fault monitor varies depending on the particular data service. The level is dependent on a number of factors; Refer to the *Sun Cluster 2.2 System Administration Guide* for details on how the fault monitor works with the Sun Cluster data services.

As the fault monitors probe the servers, they use the `local7` message facility. Messages generated by this facility can be viewed in the messages files or on the console, depending on how you have messages configured on the servers. See the `syslog.conf(4)` man page for details on setting up your messages configuration.

1.5.3.1 Data Services Supported by Sun Cluster

Sun Cluster provides HA support for various applications such as relational databases, parallel databases, internet services, and resource management data services. For the current list of data services and supported revision levels, see the *Sun Cluster 2.2 Release Notes* document or contact your Enterprise Service provider. The following data services are supported with this release of Sun Cluster:

- Sun Cluster HA for DNS
- Sun Cluster HA for Informix

- Sun Cluster HA for Lotus
- Sun Cluster HA for Netscape
  - Sun Cluster HA for Netscape HTTP
  - Sun Cluster HA for Netscape LDAP
  - Sun Cluster HA for Netscape Mail
  - Sun Cluster HA for Netscape News
- Sun Cluster HA for NFS
- Sun Cluster HA for Oracle
- Sun Cluster HA for SAP
- Sun Cluster HA for Sybase
- Sun Cluster HA for Tivoli
- Oracle Parallel Server
- Informix-Online XPS

1.5.3.2 Data Services API

Sun Cluster software includes an Application Programming Interface (API) permitting existing crash-tolerant data services to be made highly available under the Sun Cluster HA framework. Data services register methods (programs) that are called back by the HA framework at certain key points of cluster reconfigurations. Utilities are provided to permit data service methods to query the state of the Sun Cluster configuration and to initiate takeovers. Additional utilities make it convenient for a data service method to run a program while holding a file lock, run a program under a timeout, or automatically restart a program if it dies.

For more information on the data services API, refer to the *Sun Cluster 2.2 API Developer's Guide.*

1.5.4 Switch Management Agent

The Switch Management Agent (SMA) software component manages sessions for the SCI links and switches. Similarly, it manages communications over the Ethernet links and switches. In addition, SMA isolates applications from individual link failures and provides the notion of a logical link for all applications.

1.5.5 Cluster SNMP Agent

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 Sun Cluster SNMP agent is capable of monitoring several clusters (a maximum of 32) at the same time. In a typical Sun Cluster, you can manage the cluster from the administration workstation or System Service Processor (Sun Enterprise 10000). By installing the Sun Cluster SNMP agent on the administration workstation or System Service Processor, network traffic is regulated and the CPU power of the nodes is not wasted in transmitting SNMP packets.

1.5.6 Cluster Configuration Database

The Cluster Configuration Database (CCD) is a highly available, replicated database that is used to store internal configuration data for Sun Cluster configuration needs. The CCD is for Sun Cluster internal use—it is not a public interface and you should not attempt to update it directly.

The CCD relies on the Cluster Membership Monitor (CMM) service to determine the current cluster membership and determine its consistency domain, that is, the set of nodes that must have a consistent copy of the database and that propagate updates. The CCD database is divided into an *Initial* (Init) and a *Dynamic* database.

The purpose of the Init CCD database is storage of *non-modifiable* boot configuration parameters whose values are set during the CCD package installation (`scinstall`). The Dynamic CCD contains the remaining database entries. Unlike the Init CCD, entries in the Dynamic CCD can be updated at any time with the restrictions that the CCD database is recovered (that is, the cluster is up) and the CCD has quorum. (See Section 1.5.6.1 “CCD Operation” on page 1-37, for the definition of quorum.)

The Init CCD (`/etc/opt/SUNWcluster/conf/ccd.database.init`) is also used to store data for components that are started before the CCD is up. This means that queries to the Init CCD can occur before the CCD database has recovered and global consistency is checked.

The Dynamic CCD (`/etc/opt/SUNWcluster/conf/ccd.database`) contains the remaining database entries. The CCD guarantees the consistent replication of the Dynamic CCD across all of the nodes of its consistency domain.

The CCD database is replicated on all the nodes to guarantee its availability in case of a node failure. CCD daemons establish communications among themselves to synchronize and serialize database operations within the CCD consistency domain. Database updates and query operations can be issued from any node—the CCD does not have a single point of control.

In addition, the CCD offers:

- Cluster-wide repository (the same view from every node)

- Distributed framework for updates
  - Local consistency (consistency record)
  - Global consistency (automatic replication)
- Database recovery and resynchronization

1.5.6.1 CCD Operation

The CCD guarantees a consistent replication of the database across all the nodes of the elected *consistency domain*. Only nodes that are found to have a valid copy of the CCD are allowed to be in the cluster. Consistency checks are performed at two levels, local and global. Locally, each replicated database copy has a self-contained consistency record that stores the checksum and length of the database. This consistency record validates the local database copy in case of an update or database recovery. The consistency record timestamps the last update of the database.

The CCD also performs a global consistency check to verify that every node has an identical copy of the database. The CCD daemons exchange and verify their consistency record. During a cluster restart, a quorum voting scheme is used for recovering the database. The recovery process determines how many nodes have a valid copy of the CCD (the local consistency is checked through the consistency record), and how many copies are identical (have the same checksum and length).

A *quorum majority* (when more than half the nodes are up) must be found within the default consistency domain to guarantee that the CCD copy is current.

Note - A quorum majority is required to perform updates to the CCD.

The equation $Q = [Na/2]+1$ specifies the number of nodes required to perform updates to the CCD. *Na* is the number of nodes physically present in the cluster. These nodes might be physically present, but not running the cluster software.

In the case of a two-node cluster with Cluster Volume Manager or Sun StorEdge Volume Manager, quorum may be maintained with only one node up by the use of a shared CCD volume. In a shared-CCD configuration, one copy of the CCD is kept on the local disk of each node and another copy is kept on in a special disk group that can be shared between the nodes. In normal operation, only the copies on the local disks are used, but if one node fails, the shared CCD is used to maintain CCD quorum with only one node in the cluster. When the failed node rejoins the cluster, it is updated with the current copy of the shared CCD. Refer to Chapter 3, for details on setting up a shared CCD volume in a two-node cluster.

If one node stays up, its valid CCD can be propagated to the newly joining nodes. The CCD recovery algorithm guarantees that the CCD database is up only if a valid copy is found and is correctly replicated on all the nodes. If the recovery fails, you must intervene and decide which one of the CCD copies is the valid one. The elected copy can then be used to restore the database via the `ccdadm -r` command. See the

Sun Cluster 2.2 System Administration Guide for the procedures used to administer the CCD.

Note - The CCD provides a backup facility, `ccdadm(1M)`, to checkpoint the current content of the database. The backup copy can subsequently be used to restore the database. Refer to the `ccdadm(1M)` man page for details.

1.5.7 Volume Managers

Sun Cluster supports three volume managers: Solstice DiskSuite, Sun StorEdge Volume Manager (SSVM), and Cluster Volume Manager (CVM). These volume managers provide mirroring, concatenating, and striping for use by Sun Cluster. SSVM and CVM also enable you to set up and administer RAID5 under Sun Cluster. Volume managers organize disks into *disk groups* that can then be administered as a unit.

The Sun StorEdge A3000 disk expansion unit also has the capability of mirroring, concatenation, and striping all within the Sun StorEdge A3000 hardware. You must use SSVM or CVM to manage disksets on the Sun StorEdge A3000. You also must use SSVM or CVM if you want to concatenate or stripe over several Sun StorEdge A3000s or mirror between Sun StorEdge A3000s.

For information on your particular volume manager refer to your volume manager documentation.

1.5.7.1 Disk Groups

Disk groups are sets of mirrored or RAID5 configurations composed of shared disks. All data service and parallel database data is stored in disk groups on the shared disks. Mirrors within disk groups are generally organized such that each half of a mirror is physically located within a separate disk expansion unit and connected to a separate controller or host adapter. This eliminates a single disk or disk expansion unit as a single point of failure.

Disk groups may either be used for raw data storage, or for file systems, or both.

1.5.8 Logical Hosts

In HA configurations, Sun Cluster supports the concept of a *logical host*. A logical host is a set of resources that can move as a unit between Sun Cluster servers. In Sun Cluster, the resources include a collection of network host names and their associated IP addresses plus one or more groups of disks (a disk group). In non-HA cluster environments, such as OPS configurations, an IP address is permanently mapped to

a particular host system. Client applications access their data by specifying the IP address of the host running the server application.

In Sun Cluster, an IP address is assigned to a logical host and is temporarily associated with whatever host system the application server is currently running on. These IP addresses are *relocatable*—that is, they can move from one node to another. In the Sun Cluster environment, clients specify the logical hosts's relocatable IP addresses to connect to an application rather than the IP address of the physical host system.

In Figure 1–18, logical host `hahost1` is defined by the network host name `hahost1`, the relocatable IP address 192.9.200.1, and the disk group `diskgroup1`. Note that the logical host name and the disk group name do not have to be the same.

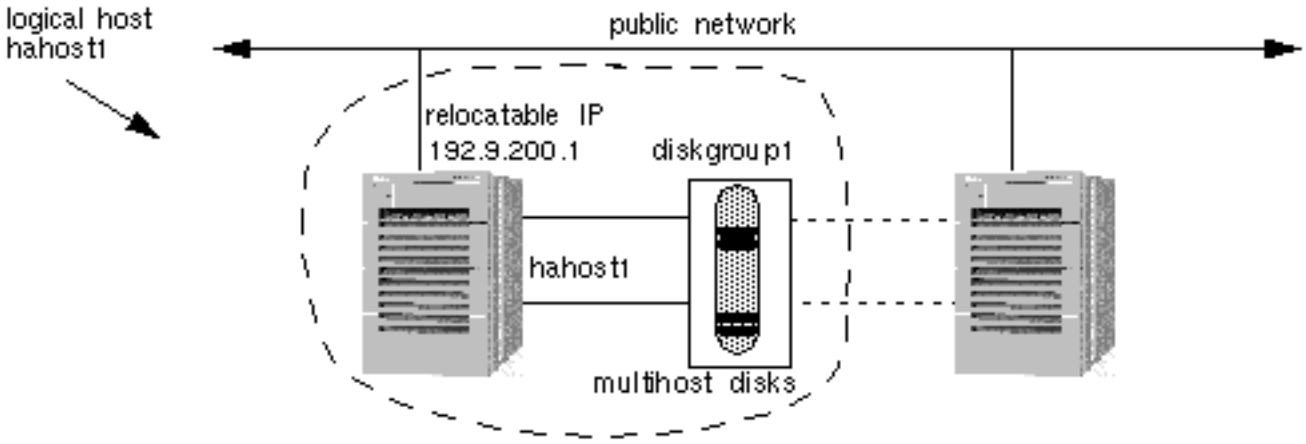


Figure 1–18 Logical Hosts

Logical hosts have one logical host name and one relocatable IP address on each public network. The name by which a logical host is known on the primary public network is its *primary logical host nam*e. The names by which logical hosts are known on secondary public networks are *secondary logical host names*. Figure 1–19 shows the host names and relocatable IP addresses for the two logical hosts with primary logical host names `hahost1` and `hahost2`. In this figure, secondary logical host names use a suffix that consists of the last component of the network number (201). For example, `hahost1-201` is the secondary logical host name for logical host `hahost1`.

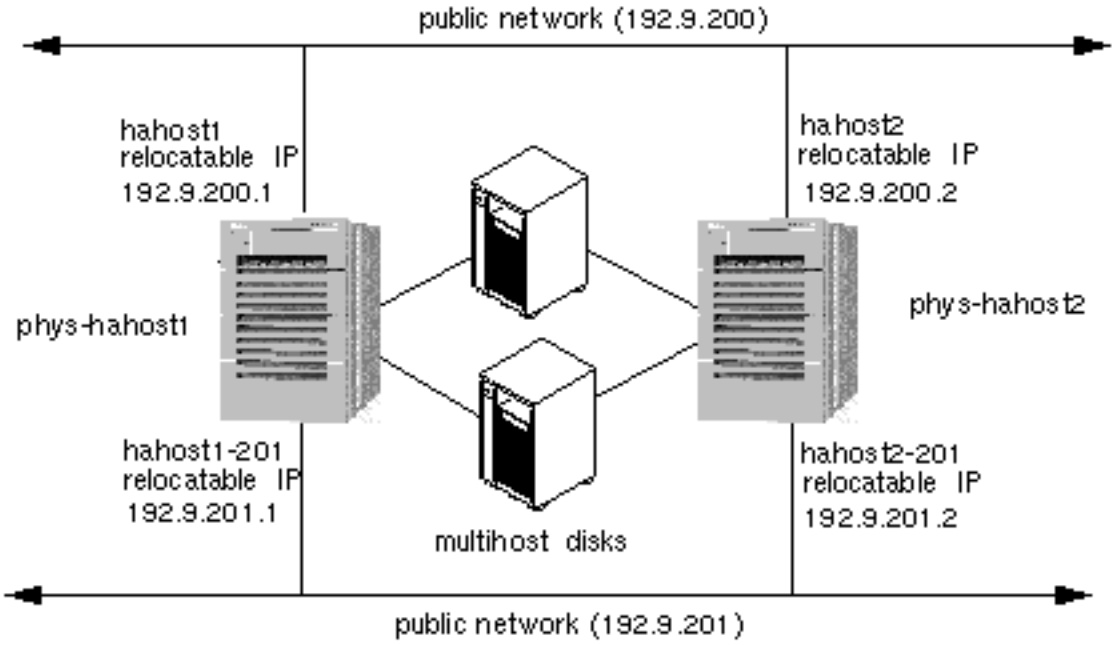


Figure 1–19 Logical Hosts on Multiple Public Networks

Logical hosts are mastered by physical hosts. Only the physical host that currently masters a logical host can access the logical host's disk groups. A physical host can master multiple logical hosts, but each logical host can be mastered by only one physical host at a time. Any physical host that is capable of mastering a particular logical host is referred to as a *potential master* of that logical host.

A data service makes its services accessible to clients on the network by advertising a well-known logical host name associated with the physical host. The logical host names are part of the IP name space at a site, but do not have a specific physical dedicated to them. The clients use these logical host names to access the services provided by the data service.

Figure 1–20 shows a configuration with multiple data services located on a single logical host's disk group. In this example, assume logical host `hahost2` is currently mastered by `phys-hahost2`. In this configuration, if `phys-hahost2` fails, both of the Sun Cluster HA for Netscape data services (`dg2-http` and `dg2-news`) will fail over to `phys-hahost1`.

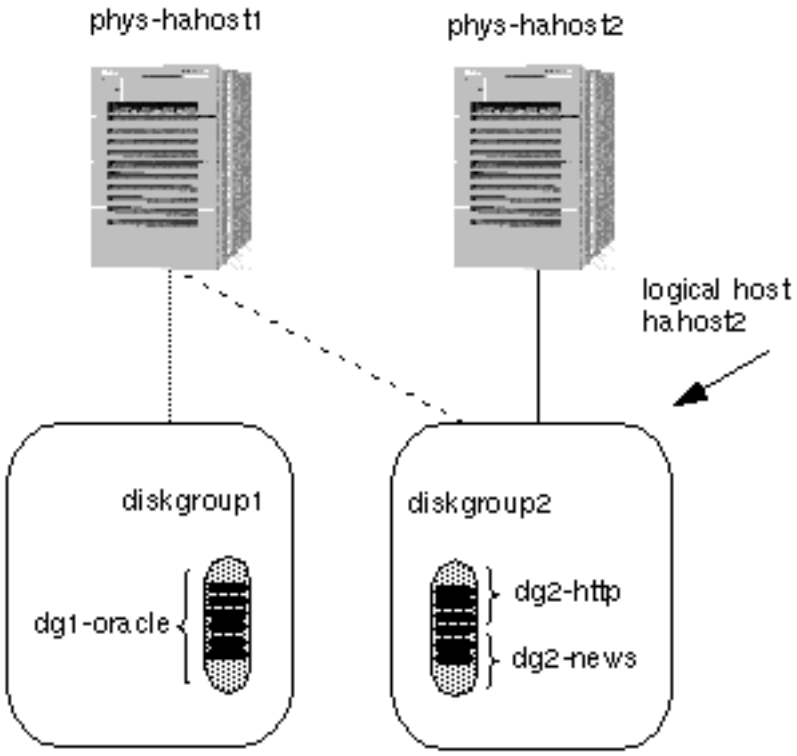


Figure 1–20 Logical Hosts, Disksets, and Data Service Files

Read the discussion in Chapter 2, for a list of issues to consider when deciding how to configure your data services on the logical hosts.

1.5.9 Public Network Management (PNM)

Some types of failures cause all logical hosts residing on that node, to be transferred to another node. The failure of a network adapter card, connector or cable between the node and the public network need not result in a node failover. Public Network Management (PNM) software in the Sun Cluster framework allows network adapters to be grouped into sets such that if one fails, another in its group takes over the servicing of network requests. A user will experience only a small delay while the error detection and failover mechanisms are in process.

In a configuration using PNM, there are multiple network interfaces on the same subnet. These interfaces make up a *backup group*. At any point, a network adapter can only be in one backup group and only one adapter within a backup group is active. When the current active adapter fails, the PNM software automatically switches the network services to use another adapter in the backup group. All adapters used for public networks should be in a backup group.

Note - Backup groups are also used to monitor the public nets even when same-node failover adapters are not present.

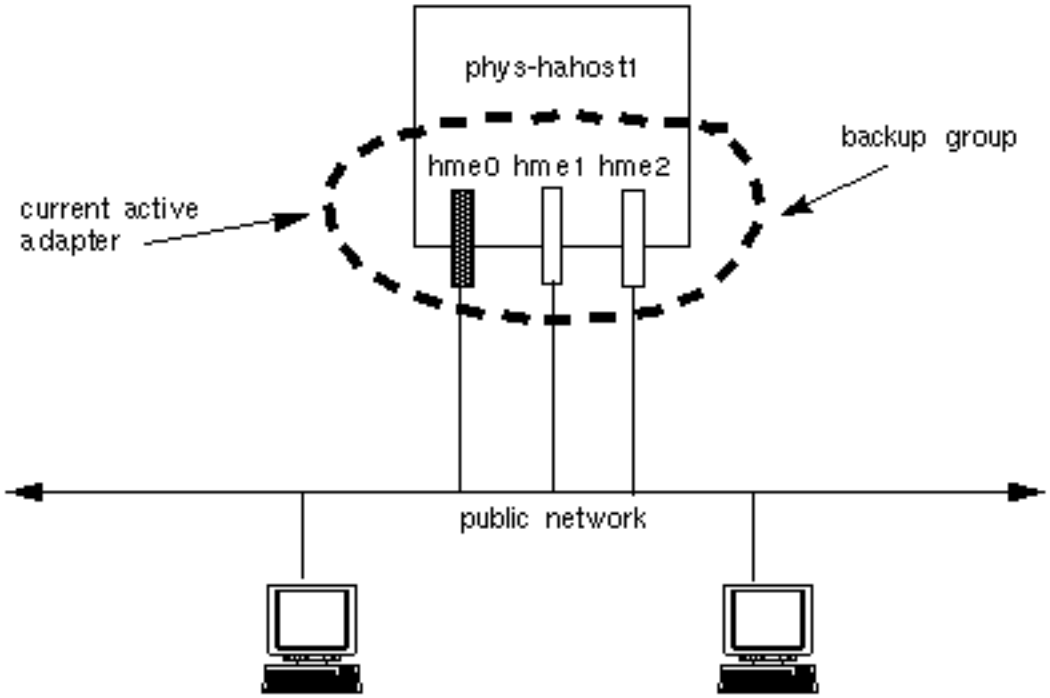


Figure 1–21 Network Adapter Failover Configuration

Refer to the *Sun Cluster 2.2 System Administration Guide* for information on setting up and administering PNM.

1.5.10 System Failover and Switchover

If a node fails in the Sun Cluster HA configuration, the data services running on the failed node are moved automatically to a working node in the failed node's server set. The failover software moves the IP addresses of the logical host(s) from the failed host to the working node. All data services that were running on logical hosts mastered by the failed host are moved.

The system administrator can manually switch over a logical host. The difference between failover and switchover is that the former is handled automatically by the Sun Cluster software when a node fails and the latter is done manually by the system administrator. A switchover might be performed to do periodic maintenance or to upgrade software on the cluster nodes.

Figure 1–22 shows a two-node configuration in normal operation. Note that each physical host masters a logical host (solid lines). The figure shows two clients accessing separate data services located on the two logical hosts.

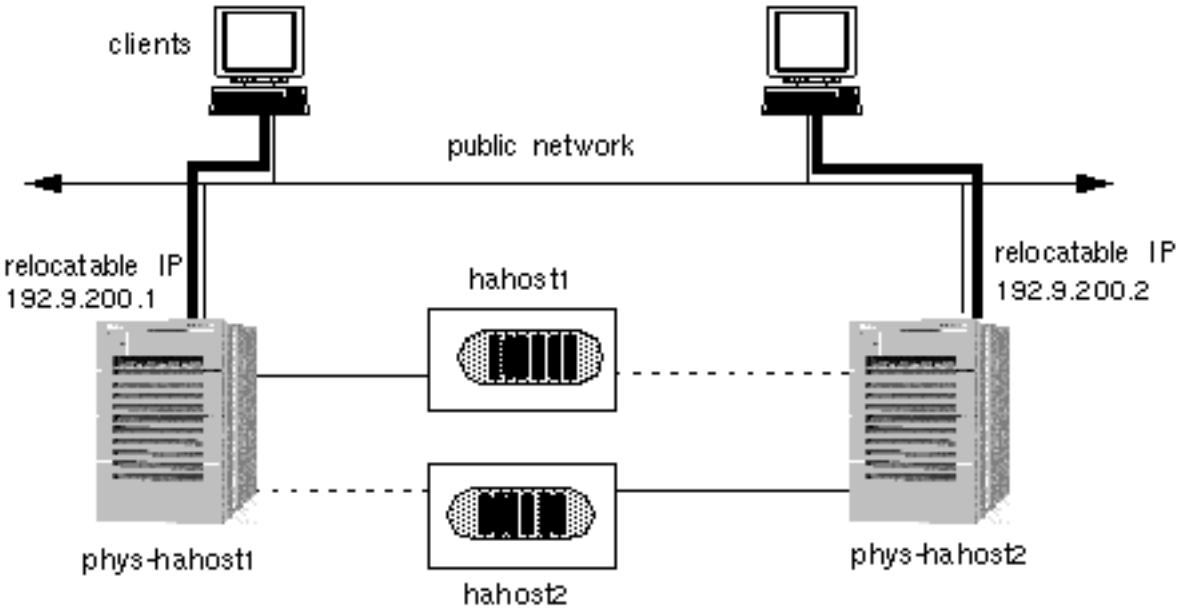


Figure 1–22 Symmetric Configuration Before Failover or Switchover

If `phys-hahost1` fails, the logical host `hahost1` will be relocated to `phys-hahost2`. The relocatable IP address for `hahost1` will move to `phys-hahost2` and data service requests will be directed to `phys-hahost2`. The clients accessing data on `hahost1` will experience a short delay while a *cluster reconfiguration* occurs. The new configuration that results is shown in Figure 1–23.

Note that the client system that previously accessed logical host `hahost1` on `phys-hahost1` continues to access the same logical host but now on `phys-hahost2`. In the failover case, this is automatically accomplished by the cluster reconfiguration. As a result of the failover, `phys-hahost2` now masters both logical hosts `hahost1` and `hahost2`. The associated disksets are now accessible only through `phys-hahost2`.

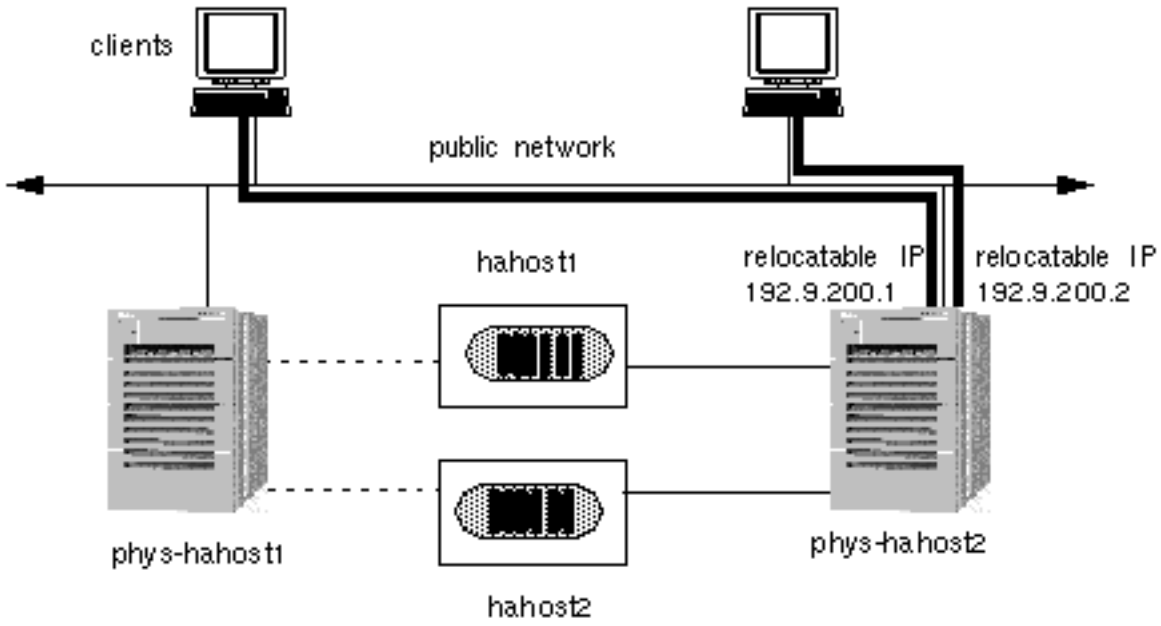


Figure 1–23 Symmetric Configuration After Failover or Switchover

1.5.10.1 Partial Failover

The fact that one physical host can master multiple logical hosts permits *partial failover* of data services. Figure 1–24 shows a star configuration that includes three physical hosts and five logical hosts. In this figure, the lines connecting the physical hosts and the logical hosts indicate which physical host currently masters which logical host (and disk groups).

The four logical hosts mastered by `phys-hahost1` (solid lines) can fail over individually to the hot-standby server. Note that the hot-standby server in Figure 1–24 has physical connections to all multihost disks, but currently does not master any logical hosts.

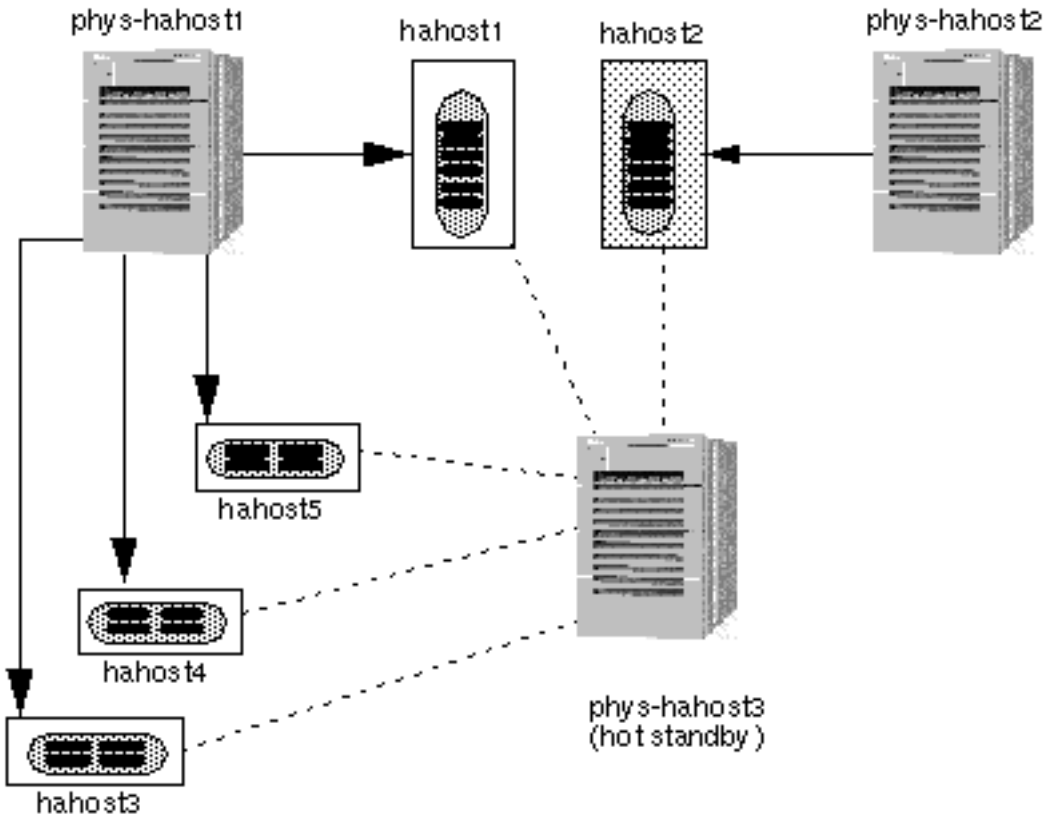


Figure 1–24 Before Partial Failover with Multiple Logical Hosts

Figure 1–25 shows the results of a partial failover where `hahost5` has failed over to the hot-standby server.

During partial failover, `phys-hahost1` relinquishes mastery of logical host `hahost5`. Then `phys-hahost3`, the hot-standby server, takes over mastery of this logical host.

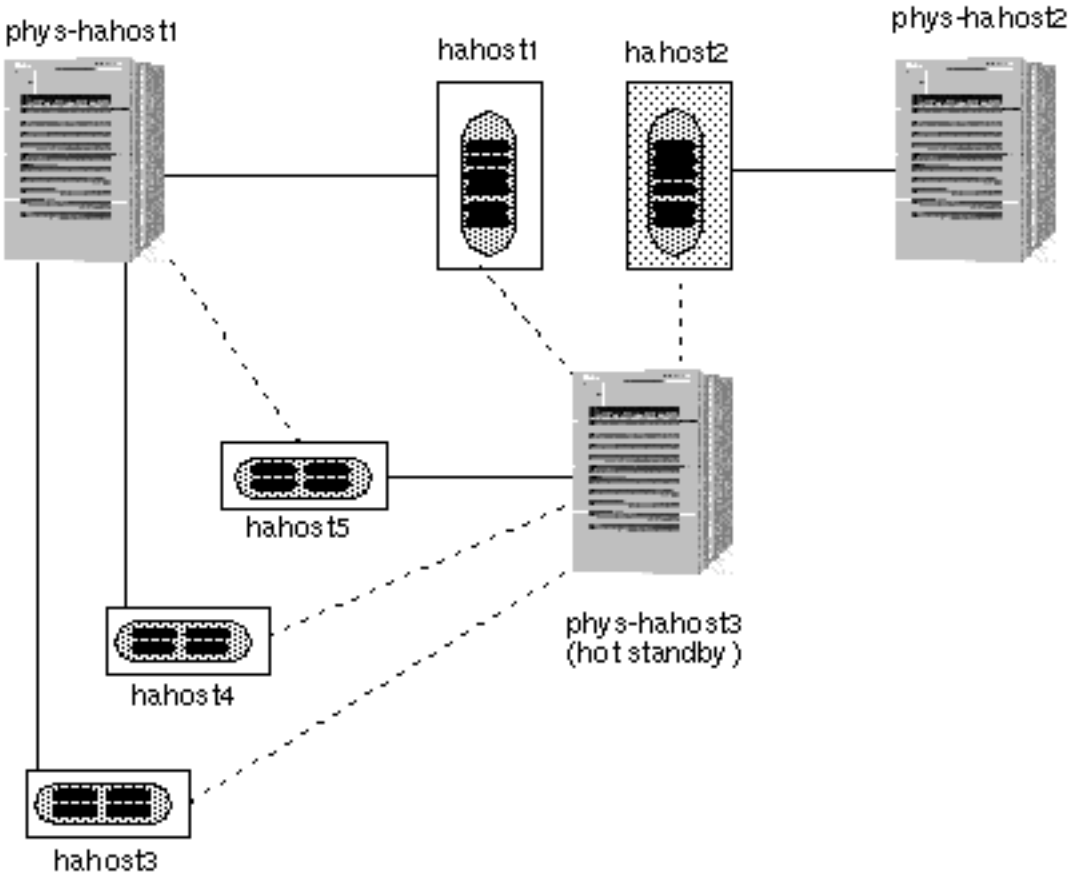


Figure 1–25 After Partial Failover with Multiple Logical Hosts

You can control which data services fail over together by placing them on the same logical host. Refer to Chapter 2, for a discussion of the issues associated with combining or separating data services on logical hosts.

1.5.10.2 Failover With Parallel Databases

In the parallel database environment, there is no concept of a logical host. However, there is the notion of relocatable IP addresses that can migrate between nodes in the event of a node failure. For more information about relocatable IP addresses and failover, see Section 1.5.8 “Logical Hosts” on page 1-38, and Section 1.5.10 “System Failover and Switchover” on page 1-42.

CHAPTER 2

Planning the Configuration

This chapter provides information and procedures for planning your Sun Cluster configuration.

2.1 Configuration Planning Overview

Configuration planning includes making decisions about:

- The Administrative Workstation
- Cluster-specific names and naming conventions
- Network connections
- Volume management
- The Solaris operating environment
- Multihost disk requirements
- File system layout on the multihost disks
- Logical host configuration (HA configurations only)
- Cluster Configuration Database (CCD)

- Quorum device (SSVM and CVM only)
- Data migration strategy
- Multihost backup strategy

Before you develop your configuration plan, consider the reliability issues described in Section 2.5 “Configuration Rules for Improved Reliability” on page 2-23. Also, the Sun Cluster environment imposes some configuration restrictions that you should consider before completing your configuration plan. These are described in Section 2.6 “Configuration Restrictions” on page 2-27.

Appendix A, provides worksheets to help you plan your configuration.

2.2 Configuration Planning Tasks

The following sections describe the tasks and issues associated with planning your configuration. You are not required to perform the tasks in the order shown here, but you should address each task as part of your configuration plan.

2.2.1 Planning the Administrative Workstation

You must decide whether to use a dedicated SPARC™ workstation, known as the *administrative workstation*, for administering the active cluster. The administrative workstation is not a cluster node. The administrative workstation can be any SPARC machine capable of running a `telnet` session to the Terminal Concentrator to facilitate console logins. Alternatively, on E10000 platforms, you must have the ability from the administrative workstation to log into the System Service Processor (SSP) and connect using the `netcon` command.

Sun Cluster does not require a dedicated administrative workstation, but using one provides you these advantages:

- Enables centralized cluster management by grouping console and management tools on the same machine
- Provides potentially quicker problem resolution by Enterprise Services

The administrative workstation must run the same version of the Solaris operating environment (Solaris 2.6 or Solaris 7) as the other nodes in the cluster.

Note - It is possible to use a cluster node as both the administrative workstation and a cluster node. This entails installing a cluster node as both “client” and “server.”

2.2.2 Establishing Names and Naming Conventions

Before configuring the cluster, you must decide on names for the following:

- The cluster itself
- Physical hosts
- Logical hosts
- Disk groups
- Network interfaces

The network interface names (and associated IP addresses) are necessary for each logical host on each public network. Although you are not required to use a particular naming convention, the following naming conventions are used throughout the documentation and are recommended. Use the configuration worksheets included in Appendix A.

Cluster – As part of the configuration process, you will be prompted for the name of the cluster. You can choose any name; there are no restrictions imposed by Sun Cluster.

Physical Hosts – Physical host names are created by adding the prefix `phys-` to the logical host names (for physical hosts that master only one logical host each). For example, the physical host that masters a logical host named `hahost1` would be named `phys-hahost1` by default. There is no Sun Cluster naming convention or default for physical hosts that master more than one logical host.

Caution - If you are using DNS as your name service, do not use an underscore in your physical or logical host names. DNS will not recognize a host name containing an underscore.

Logical Hosts and Disk Groups – Logical host names can be different from disk group names in Sun Cluster. However, using the same names is the Sun Cluster convention and eases administration. Refer to Section 2.2.9 “Planning Your Logical Host Configuration” on page 2-18, for more information.

Public Network – The names by which physical hosts are known on the public network are their primary physical host names. The names by which physical hosts are known on a secondary public network are their secondary physical host names. Assign these names using the following conventions, as illustrated in Figure 2–1:

- For the primary physical host names, simply use the physical host names as described previously; for example, `phys-hahost1` would be used for a physical host associated with logical host `hahost1`.
- For the secondary physical host names, start with the physical host name and add a suffix indicating the secondary network address. For example, the connection to a secondary network with a network address 192.9.201 from physical host phys-`hahost1` would be named `phys-hahost1-201`.

Note - The primary physical host name should be the node name returned by `uname -n`.

Private Interconnect – There is no default naming convention for the private interconnect.

Naming convention examples are illustrated in Figure 2–1.

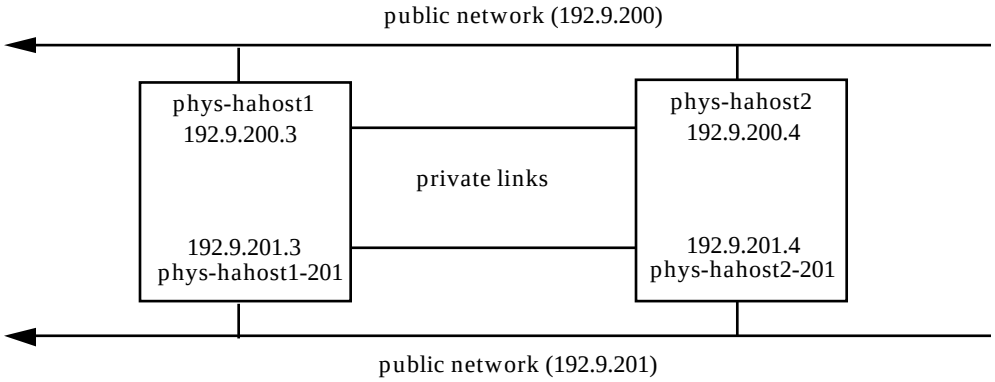


Figure 2–1 Public and Private Network Naming Conventions

2.2.3 Planning Network Connections

You must have at least one public network connection to a local area network and exactly two private interconnects between the cluster nodes. Refer to Chapter 1, for overviews of Sun Cluster network configurations, and to Appendix A, for network planning worksheets.

2.2.3.1 Public Network Connections

Consider these points when planning your public network configuration:

- You must have at least one public network that is attached to all cluster nodes. You can have as many additional public network connections as your hardware configuration allows.
- In configurations providing HA data services, you must provide IP addresses and network interface names for each logical host on each public network. This can lead to many host names. Refer to Section 2.2.2 “Establishing Names and Naming Conventions” on page 2-3, for more information. You must add the logical host names to the `/etc/hosts` files on all cluster nodes.

- Sun Cluster includes a Public Network Management (PNM) component, which enables a public network interface to fail over to another interface within a designated backup group. Refer to the chapter on administering network interfaces in the *Sun Cluster 2.2 System Administration Guide* for more information about PNM.

2.2.3.2 Private Network Connections

Sun Cluster requires two private networks for normal operation. You must decide whether to use 100 Mbit/sec Ethernet or 1 Gbit/sec Scalable Coherent Interface (SCI) connections for the private networks.

In two-node configurations, these networks may be implemented with point-to-point cables between the cluster nodes. In three- or four-node configurations, they are implemented using hubs or switches. Only private traffic between Sun Cluster nodes is transported on these networks.

If you connect nodes by using SCI switches, each node must be connected to the same port number on both switches. During the installation, note that the port numbers on the switches correspond to node numbers. For example, node 0 is the host physically connect to port 0 on the switch, and so on.

A class C network number (204.152.64) is reserved for private network use by the Sun Cluster nodes. The same network number is used by all Sun Cluster systems.

2.2.3.3 Terminal Concentrator and Administrative Workstation Network Connections

The Terminal Concentrator and administrative workstation connect to a public network with access to the Sun Cluster nodes. You must assign IP addresses and host names for them to enable access to the cluster nodes over the public network.

Note - E10000 systems use a System Service Processor (SSP) instead of a Terminal Concentrator. You will need to assign the SSP a host name, IP address, and root password. You will also need to create a user named “ssp” on the SSP and provide a password for user “ssp” during Sun Cluster installation.

2.2.4 Planning Your Solaris Operating Environment Installation

All nodes in a cluster must be installed with the same version of the Solaris operating environment (Solaris 2.6 or Solaris 7) before you can install the Sun Cluster software. When you install Solaris on cluster nodes, follow the general rules in this section.

- Install the Entire Distribution Solaris software packages on all Sun Cluster nodes.

Note - All platforms except the E10000 require at least the Entire Distribution Solaris installation, for both the Solaris 2.6 and Solaris 7 operating environments. E10000 systems require the Entire Distribution + OEM.

- After installing the Solaris operating environment, you must install the latest patches. For the current list of required patches for the Solaris 2.6 or Solaris 7 operating environments, consult your Enterprise Services representative or service provider.
- If you are upgrading from an earlier version of the Solaris operating environment:
  - You must use the upgrade option in the Solaris Interactive Installation program (rather than reinstalling the operating environment) and be prepared to increase the size of your root (`/`) and `/usr` slices to accommodate the Solaris 2.6 or Solaris 7 environment.
  - The upgrade option in the Solaris Interactive Installation program provides the ability to reallocate disk space if the current file systems don't have enough space for the upgrade. By default, an auto-layout feature tries to determine how to reallocate the disk space so the upgrade can succeed. If auto-layout cannot determine how to reallocate disk space, you must specify which file systems can be moved or changed and run auto-layout again.
- If you are installing Sun Cluster for the first time:
  - Set up each Sun Cluster node as stand-alone machine. Do this in response to a question in the Solaris 2.6 or Solaris 7 installation program.
  - Do not define an exported file system. HA-NFS file systems are not mounted on `/export` and only HA-NFS file systems should be NFS-shared on Sun Cluster nodes.
  - Disable the Solaris 2.6 or Solaris 7 power management "autoshutdown" mechanism if it applies to any nodes in your Sun Cluster configuration. See the `pwconfig(1M)` and `power.conf(4)` man pages for details.

2.2.4.1 Using Solaris Interface Groups

A new feature called *interface groups* was added to the Solaris 2.6 operating environment. This feature is implemented as default behavior in Solaris 2.6, but as optional behavior in subsequent releases.

As described in the `ifconfig(1M)` man page, if an interface (logical or physical) shares an IP prefix with another interface, these interfaces are collected into an interface group. IP uses an interface group to rotate source address selection when the source address is unspecified, and in the case of multiple physical interfaces in the same group, to distribute traffic across different IP addresses on a per-IP-destination basis (see `netstat(1M)` for per-IP-destination information).

When enabled, this feature causes a problem with switchover of logical hosts. The system will experience RPC timeouts and the switchover will fail, causing the logical host to remain mastered on its current host.

Interface groups should be disabled on all cluster nodes. The status of interface groups is determined by the value of `ip_enable_group_ifs` in `/etc/system`.

The value for this parameter can be checked with the following `ndd` command:

```
# ndd /dev/ip ip_enable_group_ifs
```

If the value returned is 1 (enabled), disable interface groups by running the following command:

```
set ip:ip_enable_group_ifs=0
```

Caution - Whenever you modify the `/etc/system` file, you must reboot the system.

2.2.4.2 Partitioning System Disks

When Solaris 2.6 or Solaris 7 is installed, the system disk is partitioned into slices for root (/), `/usr`, and other standard file systems. You must change the partition configuration to meet the requirements of Sun Cluster and your volume manager. Use the guidelines in the following sections to allocate disk space accordingly.

File System Slices

Table 2–1 shows the slice number, contents, and suggested space allocation for file systems, swap space, and slices. These values are used as the default when you install Solaris with JumpStart™, but they are not required by Sun Cluster.

TABLE 2–1 File systme slices

Number	Contents	Allocation (Mbytes)
0	root (/)	80
1	swap	50
3	`/var`	remaining free space (varies)

TABLE 2–1 File systme slices *(continued)*

Number	Contents	Allocation (Mbytes)
5	`/opt`	300
6	`/usr`	300

Volume Manager Slices

Additionally, if you will be using Solstice DiskSuite, you must set aside a 10 Mbyte slice on the system disk for metadevice state database replicas. See the Solstice DiskSuite documentation for more information about replicas.

If you will be using SSVM or CVM, you must set aside two partitions and a small amount of free space (1024 sectors) on each multihosted disk that is to be managed by SSVM or CVM, for the disk group `rootdg`. The free space should be located at the beginning or end of each disk and should not be allocated to any slice. Refer to Section 2.2.5.2 “Sun StorEdge Volume Manager and Cluster Volume Manager Considerations” on page 2-10, for more information.

The Root (/) Slice

The root (`/`) slice on your local disk must have enough space for the various files and directories as well as space for the device inodes in `/devices` and symbolic links in `/dev`.

The root slice also must be large enough to hold the following:

- Solaris system software
- Sun Cluster, some components from your volume management software, and any third party software packages
- Data space for symbolic links in `/dev` for the disk units and for volume manager use

---

Note - Sun Cluster uses various shell scripts that run as root processes. For this reason, the `/.cshrc*` and `/.profile` files for user `root` should be empty or non-existent on the cluster nodes.

Your cluster might require a larger root file system if it contains large numbers of disk drives.

Note - If you run out of free space, you must reinstall the operating environment on all cluster nodes to obtain additional free space in the root slice. Make sure at least 20 percent of the total space on the root slice is left free.

The `/usr`, `/var`, *and* `/opt` *Slices*

The `/usr` slice holds the user file system. The `/var` slice holds the system log files. The `/opt` slice holds the Sun Cluster and data service software packages. See the *Solaris Advanced Installation Guide* for details about changing the allocation values as Solaris is installed.

2.2.5 Volume Management

Sun Cluster uses volume management software to group disks into *disk groups* that can then be administered as one unit. Sun Cluster supports Solstice DiskSuite, Sun StorEdge Volume Manager (SSVM), and Cluster Volume Manager (CVM). You can use only one volume manager within a single cluster configuration.

You must install the volume management software after you install the Solaris operating environment. You can install the volume management software either before or after you install Sun Cluster software. Refer to your volume manager software documentation and to Chapter 3, for instructions on installing the volume management software.

Use these guidelines when configuring your disks:

- Mirroring of root disks is recommended, but not required.
- All multihosted disks must be mirrored across arrays. An exception to this is the Sun StorEdge A3000, which is configured to mirror or provide hardware RAID5.
- Use of hot spares is highly recommended, but not required.

See “Volume Manager Slices” on page 2-8 for disk layout recommendations related to volume management, and consult your volume manager documentation for any additional restrictions.

2.2.5.1 Solstice DiskSuite Considerations

Consider these points when planning Solstice DiskSuite configurations:

- Always use trans metadevices for file systems within disksets.
- If you run with only two disk expansion units, you will need to use Solstice DiskSuite mediators.

- When using Solstice DiskSuite mediators with any disksets in your configuration, only two cluster nodes can act as mediator hosts. Those two nodes must be used for all disksets requiring mediators, regardless of which nodes master those disksets. Therefore, in Sun Cluster configurations with more than two nodes, it is possible for a diskset's mediator host to be one that is not actually a potential master of that diskset.

2.2.5.2 Sun StorEdge Volume Manager and Cluster Volume Manager Considerations

Consider these points when planning SSVM and CVM configurations:

- You must create a default disk group (`rootdg`) on each cluster node. The `rootdg` group consists of either one slice (a "simple" disk) or a whole disk. It can be encapsulated.
- If the `rootdg` consists of a simple disk, a one slice table entry and at least two cylinders on the root disk must be left free.
- If the `rootdg` is encapsulated, then two disk slice table entries must be left free. In addition, only the root (`/`), `/usr`, `/var`, and `swap` file systems should be present on the encapsulated root disk.
- Encapsulation makes it more difficult to upgrade the volume manager software; if you anticipate frequent upgrades, encapsulating the `rootdg` is not recommended. However, if you want to mirror the root disk, you must encapsulate the `rootdg`.

Caution - Insufficient disk space and slices prevent encapsulation of the boot disk later and increase installation time because the operating environment might have to be reinstalled.

Note - You will need licenses for Sun StorEdge Volume Manager if you use it with any storage devices other than SPARCstorage Arrays or Sun StorEdge A5000s. SPARCstorage Arrays and Sun StorEdge A5000s include bundled licenses for use with SSVM. Contact the Sun License Center for any necessary SSVM licenses; see `http://www.sun.com/licensing/` for more information.

You do not need licenses to run Solstice DiskSuite or Cluster Volume Manager with Sun Cluster.

2.2.6 File System Logging

One important aspect of high availability is the ability to bring file systems back online quickly in the event of a node failure. This aspect is best served by using a logging file system. Sun Cluster supports three logging file systems; VxFS logging from Veritas, DiskSuite UFS logging, and Solaris UFS logging. Cluster Volume

Manager (CVM), when used with Oracle Parallel Server, uses raw partitions so does not use a logging file system. However, you can also run CVM in a cluster with both OPS and HA data services. In this configuration, the OPS shared disk groups would use raw partitions, but the HA disk groups could use either VxFS or Solaris UFS logging file systems (Solaris UFS logging is supported only under Solaris 7). Excluding the co-existent CVM configuration described above, Sun Cluster supports the following combinations of volume managers and logging file systems:

TABLE 2–2 Supported File System Matrix

Solaris Operating Environment	Volume Manager	Supported File Systems
Solaris 2.6	Sun StorEdge Volume Manager	VxFS, UFS (no logging)
	Solstice DiskSuite	DiskSuite UFS logging
Solaris 7	Solstice DiskSuite	DiskSuite UFS logging, Solaris UFS logging

CVM uses a feature called Dirty Region Logging to aid in fast recovery after a reboot, similar to what the logging file systems provide. For information on CVM, refer to the Sun Cluster *Cluster Volume Manager Administration Guide*. For information on DiskSuite UFS logging, refer to the Solstice DiskSuite documentation. For information on VxFS logging, see the Veritas documentation. Solaris UFS logging is described briefly below. See the `mount_ufs(1M)` for more details.

Solaris UFS logging is a new feature in the Solaris 7 operating environment.

Solaris UFS logging uses a circular log to journal the changes made to a UFS file system. As the log fills up, changes are “rolled” into the actual file system. The advantage of logging is that the UFS file system is never left in an inconsistent state, that is, with a half-completed operation. After a system crash, `fsck` has nothing to fix, so you boot up much faster.

Solaris UFS logging is enabled using the “logging” mount option. To enable logging on a UFS file system, you either add `-o logging` to the mount command or add the word “logging” to the `/etc/opt/SUNWcluster/conf/hanfs/vfstab.`*logicalhost* entry (the rightmost column).

Solaris UFS logging always allocates the log using free space on the UFS file system. The log takes up 1 MByte on file systems less than 1 GByte in size, and 1 MByte per GByte on larger file systems, up to a maximum of 64 MBytes.

Solaris UFS logging always puts the log files on the same disk as the file system. If you use this logging option, you are limited to the size of the disk. DiskSuite UFS

logging allows the log to be separated on a different disk. This has the effect of reducing a bit of the I/O that is associated with the log.

With DiskSuite UFS logging, the trans device used for logging creates a metadevice. The log is yet another metadevice which can be mirrored and striped. Furthermore, you can create up to a 1TByte logging file system with Solstice DiskSuite.

The “logging” mount option will not work if you already have logging provided by Solstice DiskSuite—you will receive a warning message explaining you already have logging on that file system. If you require more control over the size or location of the log, you should use DiskSuite UFS logging.

Depending on the file system usage, Solaris UFS logging gives you performance that is the same or better than running without logging.

There is currently no support for converting from DiskSuite UFS logging to Solaris UFS logging.

2.2.7 Determining Your Multihost Disk Requirements

Unless you are using a RAID5 configuration, all multihost disks must be mirrored in Sun Cluster configurations. This enables the configuration to tolerate single-disk failures. Refer to Section 2.5.1 “Mirroring Guidelines” on page 2-24, and to your volume management documentation, for more information.

Determine the amount of data that you want to move to the Sun Cluster configuration. If you are not using RAID5, double that amount to allow disk space for mirroring. With RAID5, you need extra space equal to 1/(# of devices -1). Use the worksheets in Appendix A, to help plan your disk requirements.

Consider these points when planning your disk requirements:

- Sun Cluster supports several multihost disk expansion units. Consider the size of disks available with each disk expansion unit when you calculate the amount of data to migrate to Sun Cluster.
- With Sun StorEdge A3000 units, you need only one disk expansion unit because each unit has two controllers. With Sun StorEdge MultiPacks, you must have at least two disk expansion units.
- Under some circumstances, there might be an advantage to merging several smaller file systems into a single larger file system. This reduces the number of file systems to administer and might help speed up cluster takeovers.
- The size of the dump media (backup system) might influence the size of the file systems in your configuration.
- With Solstice DiskSuite, if you have only two disk expansion units, then you must configure dual-string mediators. If you have more than two disk expansion units, you need not configure mediators. Sun StorEdge Volume Manager and Cluster Volume Manager do not support mediators. See the chapter on using dual-string

mediators in the *Sun Cluster 2.2 System Administration Guide* for details on the dual-string mediator feature.

- Refer to the *Sun Cluster 2.2 Release Notes* for the list of supported disk types.

2.2.7.1 Disk Space Growth

Consider these points when planning for disk space growth:

- Less administration time is required to configure disks during initial configuration than to add them while the system is in service.
- Leaving empty slots in the multihost disk expansion units during initial configuration allows you to add disks easily later.
- When your site needs additional disk expansion units, you might have to reconfigure your data to prevent mirroring within a single disk expansion unit. Therefore, if all the existing disk expansion units are full, the easiest way to add disk expansion units without reorganizing data is to add them in pairs.

2.2.7.2 Size and Number of Disk Drives

Several sizes of disks are supported in multihost disk expansion units. Consider these points when deciding which size drives to use:

- If you use lower capacity drives, you can have more spindles; this increases potential I/O bandwidth, assuming the disks have the same I/O rates.
- If you use higher capacity disks, then fewer devices are required in the configuration. This can help speed up takeovers because takeover time can be partially dependent on the number of drives being taken over.
- You can determine the number of disks needed by dividing the total disk capacity that you have selected (including mirrors) by the disk size in your disk expansion units.

2.2.8 Planning Your File System Layout on the Multihost Disks

Sun Cluster does not require any specific disk layout or file system size. The requirements for the file system hierarchy are dependent on the volume management software you are using.

Regardless of your volume management software, Sun Cluster requires at least one file system per disk group to serve as the *HA administrative file system*. This administrative file system is generally mounted on */logicalhost*, and must be a minimum of 10 Mbytes. It is used to store private directories containing data service configuration information.

For clusters using Solstice DiskSuite, you need to create a metadevice to contain the HA administrative file system. The HA administrative file system should be configured the same as your other multihost file systems, that is, it should be mirrored and set up as a trans device.

For clusters using SSVM or CVM, Sun Cluster creates the HA administrative file system on a volume named *dg*-stat where *dg* is the name of the disk group in which the volume is created. *dg* is usually the first disk group in the list of disk groups specified when defining a logical host.

Consider these points when planning file system size and disk layout:

- When mirroring, lay out disks so that they are mirrored across disk controllers.
- Partitioning or subdividing all similar disks identically simplifies administration.

2.2.8.1 File Systems With Solstice DiskSuite

Solstice DiskSuite software requires some additional space on the multihost disks and imposes some restrictions on its use. For example, if you are using UNIX file system (UFS) logging under Solstice DiskSuite, one to two percent of each multihost disk must be reserved for metadevice state database replicas and UFS logging. Refer to Appendix B, and to the Solstice DiskSuite documentation for specific guidelines and restrictions.

All metadevices used by each shared diskset are created in advance, at reconfiguration boot time, based on settings found in the `md.conf` file. The fields in `md.conf` file are described in the Solstice DiskSuite documentation. The two fields that are used in the Sun Cluster configuration are `md_nsets` and `nmd`. The `md_nsets` field defines the number of disksets and the `mnd` field defines the number of metadevices to create for each diskset. You should set these fields at install time to allow for all predicted future expansion of the cluster.

Extending the Solstice DiskSuite configuration after the cluster is in production is time consuming because it requires a reconfiguration reboot for each node and always carries the risk that there will not be enough space allocated in the root (`/`) file system to create all of the requested devices.

The value of `md_nsets` must be set to the expected number of logical hosts in the cluster, plus one to allow Solstice DiskSuite to manage the private disks on the local host (that is, those metadevices that are not in the local diskset).

The value of `nmd` must be set to the predicted largest number of metadevices used by any one of the disksets in the cluster. For example, if a cluster uses 10 metadevices in its first 15 disksets, but 1000 metadevices in the 16th diskset, `nmd` must be set to at least 1000.

Caution - All cluster nodes (or cluster pairs in the cluster pair topology) *must* have identical `md.conf` files, regardless of the number of logical hosts served by each node. Failure to follow this guideline can result in serious Solstice DiskSuite errors and possible loss of data.

Consider these points when planning your Solstice DiskSuite file system layout:

- The HA administrative file system cannot be grown using `growfs(1M)`.
- You must create mount points for other file systems at the */logicalhost* level.
- Your application might dictate a file system hierarchy and naming convention. Sun Cluster imposes no restrictions on file system naming, as long as names do not conflict with data service required directories.
- Use the partitioning scheme described in Table 2–3 for the majority of drives.

TABLE 2–3 Solstice DiskSuite disk partitioning

Slice	Description
7	2 Mbytes, reserved for Solstice DiskSuite
6	UFS logs
0	Remainder of the disk
2	Overlaps Slice 6 and 0

- In general, if UFS logs are created, the default size for Slice 6 should be 1 percent of the size of the largest multihost disk found on the system.

Note - The overlap of Slices 6 and 0 by Slice 2 is used for raw devices where there are no UFS logs.

In addition, the first drive on each of the first two controllers in each of the disksets should be partitioned as described in Table 2–4.

TABLE 2–4 Multihost Disk Partitioning for the First Drive on the First Two Controllers

Slice	Description
7	2 Mbytes, reserved for Solstice DiskSuite
5	2 Mbytes, UFS log for HA administrative file systems

TABLE 2–4 Multihost Disk Partitioning for the First Drive on the First Two Controllers *(continued)*

Slice	Description
4	9 Mbytes, UFS master for HA administrative file systems
6	UFS logs
0	Remainder of the disk
2	Overlaps Slice 6 and 0

- Each disk group has an HA administrative file system associated with it. This file system is not NFS-shared. It is used for data service specific state or configuration information.

Partition 7 is always reserved for use by Solstice DiskSuite as the first or last 2 Mbytes on each multihost disk.

2.2.8.2 File Systems With VERITAS VxFS

You can create UNIX File System (UFS) or Veritas File System (VxFS) file systems in the disk groups of logical hosts. When a logical host is mastered on a cluster node, the file systems associated with the disk groups of the logical host are mounted on the specified mount points of the mastering node.

When you reconfigure logical hosts, Sun Cluster must check the file systems before mounting them, by running the `fsck` command. Even though the `fsck` command checks the UFS file systems in non-interactive parallel mode on UFS file systems, this still consumes some time, and this affects the reconfiguration process. VxFS drastically cuts down on the file system check time, especially if the configuration contains large file systems (greater than 500 Mbytes) used for data services.

When setting up mirrored volumes, always add a Dirty Region Log (DRL) to decrease volume recovery time in the event of a system crash. When mirrored volumes are used in clusters, DRL must be assigned for volumes greater than 500 Mbytes.

With SSVM and CVM, it is important to estimate the maximum number of volumes that will be used by any given disk group at the time the disk group is created. If the number is less than 1000, default minor numbering can be used. Otherwise, you must carefully plan the way in which minor numbers are assigned to disk group volumes. It is important that no two disk groups shared by the same nodes have overlapping minor number assignments.

As long as default numbering can be used and all disk groups are currently imported, it is not necessary to use the `minor` option to the `vxdg init` command at

disk group creation time. Otherwise, the `minor` option must be used to prevent overlapping the volume minor number assignments. It is possible to modify the minor numbering later, but doing so might require you to reboot and import the disk group again. Refer to the `vxdg(1M)` man page for details.

2.2.8.3 Mount Information

The `/etc/vfstab` file contains the mount points of file systems residing on local devices. For a multihost file system used for a logical host, all the nodes that can potentially master the logical host should possess the mount information.

The mount information for a logical host's file system is kept in a separate file on each node, named /etc/opt/SUNWcluster/conf/hanfs/vfstab.*logicalhost*. The format of this file is identical to the `/etc/vfstab` file for ease of maintenance, though not all fields are used.

Note - You must keep the vfstab.*loghostname* file consistent among all nodes of the cluster. Use the `rcp` command or file transfer protocol (FTP) to copy the file to the other nodes of the cluster. Alternately, simultaneously edit the file by using `crlogin` or `ctelnet`.

The same file system cannot be mounted by more than one node at the same time, because a file system can be mounted only if the corresponding disk group has been imported by the node. The consistency and uniqueness of the disk group imports and logical host mastery is enforced by the cluster framework logical host reconfiguration sequence.

2.2.8.4 Booting From a SPARCstorage Array

Sun Cluster supports booting from a private or shared disk inside a SPARCstorage Array.

Consider these points when using a boot disk in an SSA:

- Make sure that each cluster node's boot disk is on a different SSA. If nodes share a single SSA for their boot disks, the loss of a single controller would bring down all nodes.
- For SSVM and CVM configurations, do not configure a boot disk and a quorum device on the same tray. This is especially true for a two-node cluster. If you place both on the same tray, the cluster loses one of its nodes as well as its quorum device when you remove the tray. If for any reason a reconfiguration happens on the surviving node during this time, the cluster is lost. If a controller containing the boot disk and the quorum disk becomes faulty, the node that has its boot disk on the bad controller inevitably hangs or crashes, causing the other node to reconfigure and abort, since the quorum device is inaccessible. (This is a likely

scenario in a minimal configuration consisting of two SSAs with boot disks and no root disk mirroring.)

- Mirroring the boot disks in a bootable SSA configuration is recommended. However, there is an impact on software upgrades. Neither Solaris nor the volume manager software can be upgraded while the root disk is mirrored. In such configurations, perform upgrades carefully to avoid corruption of the root file system. Refer to Section 2.5.1.1 “Mirroring Root (/)” on page 2-24, for additional information on mirroring the root file system.

2.2.9 Planning Your Logical Host Configuration

A disk group stores the data for one or more data services. Generally, several data services share a logical host, and therefore fail over together. If you want to enable a particular data service to fail over independently of all other data services, then assign a logical host to that data service alone, and do not allow any other data services to share it.

As part of the installation and configuration, you need to establish the following for each logical host:

- Default master – Each logical host can potentially be mastered by any physical host to which it is connected.
- HA administrative file system – This is a mount point on the logical host for the administrative file system. Refer to Section 2.2.8 “Planning Your File System Layout on the Multihost Disks” on page 2-13, for more information.
- `vfstab` file name – Each logical host needs a separate `vfstab` file to store file system information. This name is generally `vfstab.`*logicalhost*.
- Disk group – Each disk group has its own name. By convention, the disk group name and the logical host name are the same, but you can give the disk group another name.

Use the logical host worksheet in Appendix A, to record this information.

Consider these points when planning your logical host configuration:

- If a data service does not put a heavy load on the CPU or memory, then you will not gain a load-balancing advantage by switching it over separately.
- Use care when load balancing an N+1 configuration. If the data service puts a heavy load on the CPU or memory, then you should limit the workload on the hot-standby node. A large workload on the hot-standby node compromises its ability to take over should any active node fail.
- If the data service software is relatively reliable and starts up quickly, then you will not gain much availability by failing it over separately.
- If the data service uses only a small amount of disk space, you might waste a lot of disk space by putting it on a separate logical host, because you must have at least a mirrored pair of drives per disk group.

- The administrative burden increases as the number of logical hosts grows.

2.2.10 Planning the Cluster Configuration Database Volume

As part of the installation and configuration, you configure a *Cluster Configuration Database* (CCD) volume to store internal configuration data. In a two-node cluster using SSVM or CVM, this volume can be shared between the nodes thereby increasing the availability of the CCD. In clusters with more than two nodes, a copy of the CCD is local to each node. See Section C.5 “Configuring the Shared CCD Volume” on page 14-8, for details on configuring a shared CCD.

Note - You cannot used a shared CCD in a two-node cluster using Solstice DiskSuite.

If each node keeps its own copy of the CCD, then updates to the CCD are disabled by default when one node is not part of the cluster. This prevents the database from getting out of synchronization when only a single node is up.

The CCD requires two disks as part of a disk group for a shared volume. These disks are dedicated for CCD use and cannot be used by any other application, file system, or database.

The CCD should be mirrored for maximum availability. The two disks comprising the CCD should be on separate controllers.

In clusters using CVM or SSVM, the `scinstall(1M)` script will ask you how you want to set up the CCD on a shared volume in your configuration.

Refer to Chapter 1, for a general overview of the CCD. Refer to the chapter on general Sun Cluster administration in the *Sun Cluster 2.2 System Administration Guide* for procedures used to administer the CCD.

Note - Although the installation procedure does not prevent you from choosing disks on the same controller, this would introduce a possible single point of failure and is not recommended.

2.2.11 Planning the Quorum Device (SSVM and CVM Only)

If you are using Cluster Volume Manager or Sun StorEdge Volume Manager as your cluster volume manager, you must configure a quorum device regardless of the number of cluster nodes. During the Sun Cluster installation process, `scinstall(1M)` will prompt you to configure a quorum device.

The quorum device is either an array controller or a disk.

- If it is an array controller, all disks in the array must be part of the cluster applications. No private data (a private file system or disk groups private to a node) can be stored in the array controller.
- If the quorum device is a disk, that disk must be part of the cluster application. The disk cannot be private to either of the nodes.

During the cluster software installation, you will need to make decisions concerning:

- Type of quorum configuration (simple mode or complex mode) – In simple mode, the quorum device is configured automatically. In complex mode, you must configure the quorum device manually.
- Quorum device behavior – If the cluster is partitioned into subsets, you can configure the Cluster Membership Monitor either to automatically select which subset stays up, or to have the system prompt you for action.
- Quorum device policy – If you choose to have the system automatically select which subset stays up, you must configure the policy. You choose either lowest or highest node ID, to specify which subset of nodes automatically becomes the new cluster in the event the quorum device is activated. Refer to Section 1.3 “Quorum, Quorum Devices, and Failure Fencing” on page 1-11, for more information on the quorum device policy.
- Type of device – The quorum device can be a controller or disk in a multihost disk expansion unit.
    - If all the disks in an expansion unit are going to be used for shared disk groups (for OPS) or for HA disk groups, then the array controller can be used for the quorum device.
    - If one or more disks in an array are used for the private storage of a node (either as a file system or as a raw device), then one of the disks belonging either to the shared disk group (for OPS) or to one of the HA disk groups must be used as the quorum device.
    - You can also choose a dedicated disk as the quorum device (one on which no data is stored).

2.2.11.1 Cluster Topology Considerations

Before you select the quorum device for your cluster, be aware of the implications of your selection. Any node pair of the cluster must have a quorum device. That is, one quorum device must be specified for every node set that share multihost disks. Each node in the cluster *must be informed of all quorum devices* in the cluster, not just the quorum device connected to it. The `scinstall`(1M) script offers all possible node pairs in sequence and displays any common devices that are quorum device candidates.

In two-node clusters with dual-ported disks, a single quorum device needs to specified.

In greater than two-node clusters with dual-ported disks, not all of the cluster nodes have access to the entire disk subsystem. In such configurations, you must specify one quorum device for each set of nodes that shares disks.

Sun Cluster configurations can consist of disk storage units (such as the Sun StorEdge A5000) that can be connected to all nodes in the cluster. This allows for applications such as OPS to run on clusters of greater than two nodes. A disk storage unit that is physically connected to all nodes in the cluster is referred to as a *direct attached device*. In this type of cluster a single quorum device needs to be selected from a direct attached device.

In clusters with direct attached devices, if the cluster interconnect fails, one of the following will happen:

- If manual intervention was specified when the quorum device was configured, all nodes will prompt for operator assistance.
- If automatic selection was specified when the quorum device was configured, the highest or lowest node ID will reserve the quorum device and all other nodes will prompt for operator assistance.

In clusters without direct attached devices to all nodes of the cluster, you will, by definition, have multiple quorum devices (one for each node pair that share disks). In this configuration, the quorum device only comes into play where only two nodes are remaining and they share a common quorum device.

In the event of a node failure, the node that is able to reserve the quorum device remains as the sole survivor of the cluster. This is necessary to ensure the integrity of data on the shared disks.

2.2.12 Planning a Data Migration Strategy

Consider these points when deciding how to migrate existing data to the Sun Cluster environment.

- You cannot move data into the Sun Cluster configuration by connecting existing disks that contain data. The volume management software must be used to prepare the disks before moving data to them.
- You can use `ufsdump(1M)` and `ufsrestore(1M)` or other suitable file system backup products to migrate UNIX file system data to Sun Cluster nodes.
- When migrating databases to a Sun Cluster configuration, use the method recommended by the database vendor.

2.2.13 Selecting a Multihost Backup Strategy

Before you load data onto the multihost disks in a Sun Cluster configuration, you should have a plan for backing up the data. Sun recommends using Solstice Backup™ or `ufsdump` to back up your Sun Cluster configuration.

If you are converting your backup method from Online:Backup™ to Solstice Backup, special considerations exist because the two products are not compatible. The primary decision for the system administrator is whether or not the files backed up with Online:Backup will be available online after Solstice Backup is in use. Refer to the Solstice Backup documentation for details on conversion.

2.2.14 Planning for Problem Resolution

The following files should be saved after the system is configured and running. In the unlikely event that the cluster should experience problems, these files can help service providers debug and solve cluster problems.

- `did.conf` `/etc`

  This file contains the disk ID (DID) mappings for Solstice DiskSuite configurations. this information can be used to track and verify the correct DID configurations after a catastrophic failure on one node.

- `ccd.database` `/etc/opt/SUNWcluster/conf`

  This file contains important cluster configuration information. See the instructions for saving this file in the section on troubleshooting the CCD in the *Sun Cluster 2.2 System Administration Guide*.

- *cluster_name*`.cdb` `/etc/opt/SUNWcluster/conf`

  This file contains current information about the cluster configuration. It can be used as a reference to determine what changes have occurred since the original setup.

- `metastat -s` *diskset_name* `-p` > *sav.diskset_name*

  This command saves the current Solstice DiskSuite diskset configuration.

2.3 Selecting a Solaris Install Method

You can install Solaris from a local CD-ROM or from a network install server using JumpStart. If you are installing several Solaris machines, consider a network install. Otherwise, use the local CD-ROM.

Note - Configurations using FDDI as the primary public network cannot be network-installed directly using JumpStart because the FDDI drivers are unbundled and are not available in “mini-unix.” If you use FDDI as the primary public network, you must install Solaris from CD-ROM.

2.4 Licensing

Sun Cluster 2.2 requires no framework or HA data service licenses to run. You do not need licenses to run Solstice DiskSuite or Cluster Volume Manager with Sun Cluster 2.2. However, you will need licenses for Sun Enterprise Volume Manager, if you use it with any storage devices other than SPARCstorage Arrays or StorEdge A5000s. SPARCstorage Arrays and StorEdge A5000s include bundled licenses for use with SSVM. Contact the Sun License Center for any necessary SSVM licenses; see `http://www.sun.com/licensing/` for more information.

You may need to obtain licenses for DBMS products and other third party products. Contact your third party service provider for third party product licenses.

2.5 Configuration Rules for Improved Reliability

The rules discussed in this section help ensure that your Sun Cluster configuration is highly available. These rules also help determine the appropriate hardware for your configuration.

- Although it is not required in some configurations, in general the Sun Cluster nodes should have identical local hardware. This means that if one cluster node is configured with two FC/S cards, then all Sun Cluster nodes in the cluster also should have two FC/S cards.
- Identify “redundant” hardware components on each node and plan their placement to prevent the loss of both components in the event of a single hardware failure. For example, consider the private networks on the E10,000 system. The minimum configuration consists of two I/O boards, each supporting one of the private network connections and one of the multihost disk connections. A localized failure on an I/O board is unlikely to affect both private network connections, or both multihost disk connections.

  This is not always possible—some configurations might contain only one system board—but some of the concerns can still be addressed easily with hardware

options. For example, in an Ultra Enterprise™ 2 Cluster with two SPARCstorage Arrays, one private network can be connected to a Sun Quad FastEthernet™ Controller card (SQEC), while the other private network can be connected to the on-board interface.

2.5.1 Mirroring Guidelines

Unless you are using a RAID5 configuration, all multihost disks must be mirrored in Sun Cluster configurations. This enables the configuration to tolerate single-disk failures.

Consider these points when mirroring multihost disks:

- Each submirror of a given mirror or plex should reside in a different multihost disk expansion unit.
- Mirroring doubles the amount of necessary disk space.
- Three-way mirroring is supported by Solstice DiskSuite, Sun StorEdge Volume Manager, and Cluster Volume Manager. However, only two-way mirroring is required by Sun Cluster.
- Under Solstice DiskSuite, mirrors are made up of other metadevices such as concatenations or stripes. Large configurations might contain a large number of metadevices. For example, seven metadevices are created for each logging UFS file system.
- If you mirror to a disk of a different size, your mirror capacity is limited to the size of the smallest submirror or plex.

2.5.1.1 Mirroring Root (/)

For maximum availability, you should mirror root (/), `/usr`, `/var`, `/opt`, and `swap` on the local disks. Under Sun StorEdge Volume Manager and Cluster Volume Manager, this means encapsulating the root disk and mirroring the generated subdisks. However, mirroring the root disk is not a requirement of Sun Cluster.

You should consider the risks, complexity, cost, and service time for the various alternatives concerning the root disk. There is not one answer for all configurations. You might want to consider your local Enterprise Services representative's preferred solution when deciding whether to mirror root.

Refer to your volume manager documentation for instructions on mirroring root.

Consider the following issues when deciding whether to mirror the root file system.

- Mirroring root adds complexity to system administration and complicates booting in single user mode.
- Regardless of whether or not you mirror root, you also should perform regular backups of root. Mirroring alone does not protect against administrative errors;

only a backup plan can allow you to restore files which have been accidentally altered or deleted.

- Under Solstice DiskSuite, in failure scenarios in which metadevice state database quorum is lost, you cannot reboot the system until maintenance is performed.

  Refer to the discussion on metadevice state database and state database replicas in the Solstice DiskSuite documentation.

- Highest availability includes mirroring root on a separate controller.

- You might regard a sibling node as the "mirror" and allow a takeover to occur in the event of a local disk drive failure. Later, when the disk is repaired, you can copy over data from the root disk on the sibling node.

  Note, however, that there is nothing in the Sun Cluster software that guarantees an immediate takeover. In fact, the takeover might not occur at all. For example, presume some sectors of a disk are bad. Presume they are all in the user data portions of a file that is crucial to some data service. The data service will start getting I/O errors, but the Sun Cluster node will stay up.

- You can set up the mirror to be a bootable root disk so that if the primary boot disk fails, you can boot from the mirror.

- With a mirrored root, it is possible for the primary root disk to fail and work to continue on the secondary (mirror) root disk.

  At a later point the primary root disk might return to service (perhaps after a power cycle or transient I/O errors) and subsequent boots are performed using the primary root disk specified in the OpenBoot™ PROM –`boot-device` field. Note that a Solstice DiskSuite resync has not occurred—that requires a manual step when the drive is returned to service.

  In this situation there was no manual repair task—the drive simply started working "well enough" to boot.

  If there were changes to any files on the secondary (mirror) root device, they would not be reflected on the primary root device during boot time (causing a stale submirror). For example, changes to `/etc/system` would be lost. It is possible that some Solstice DiskSuite administrative commands changed `/etc/system` while the primary root device was out of service.

  The boot program does not know whether it is booting from a mirror or an underlying physical device, and the mirroring becomes active part way through the boot process (after the metadevices are loaded). Before this point the system is vulnerable to stale submirror problems.

- Upgrading to later versions of the Solaris environment while using volume management software to mirror root requires steps not currently outlined in the Solaris documentation. The current Solaris upgrade is incompatible with the volume manager software used by Sun Cluster. Consequently, a root mirror must be converted to a one-way mirror before running the Solaris upgrade. Additionally, all three supported volume managers require that other tasks be

performed to successfully upgrade Solaris. Refer to the appropriate volume management documentation for more information.

2.5.1.2 Solstice DiskSuite Mirroring Alternatives

Consider the following alternatives when deciding whether to mirror root (/) file systems under Solstice DiskSuite. The issues mentioned in this section are not applicable to Sun StorEdge Volume Manager or Cluster Volume Manager configurations.

- For highest availability, mirror root on a separate controller with metadevice state database replicas on three different controllers. This tolerates both disk and controller failures.
- Under Solstice DiskSuite, to tolerate disk media failures only:

  - Mirror the root disk on a second controller and keep a copy of the metadevice state database on a third disk on one of the controllers.

    or

  - Mirror the root disk on the same controller and keep a copy of the metadevice state database on a third disk on the same controller.

  It is possible to reboot the system before performing maintenance in these configurations, because a quorum is maintained after a disk media failure. These configurations do not tolerate controller failures, with the exception that option ‘a’ tolerates controller failure of the controller that contains metadevice state database replicas on a single disk.

  If the controller that contains replicas on two disks fails, quorum is lost.

- Mirroring the root disk on the same controller and storing metadevice state database replicas on both disks tolerates a disk media failure and prevents an immediate takeover. However, you cannot reboot the machine until after maintenance is performed because more than half of the metadevice state database replicas are not available after the failure.
- Do not mirror the root disk, but perform a daily manual backup of the root disk (with `dd(1)` or some other utility) to a second disk which can be used for booting if the root disk fails. Configure the second disk as an alternate boot device in the OpenBoot PROM. The `/etc/vfstab` file might need to be updated after the `dd(1)` operation to reflect the different root partition. Configure additional metadevice state database replicas on Slice 4 of the second disk. In the event of failure of the first disk, these will continue to point to the multihost disk replicas. Do not copy and restore the metadevice state database. Rather, let Solstice DiskSuite do the replication.

2.6 Configuration Restrictions

This section describes Sun Cluster configuration restrictions.

2.6.1 Service and Application Restrictions

Note the following restrictions related to services and applications.

- Sun Cluster can be used to provide service for only those data services that either are supplied with Sun Cluster or set up using the Sun Cluster data services API.
- Do not configure the Sun Cluster nodes as mail servers, because sendmail(1M) is not supported in a Sun Cluster environment. No mail directories should reside on Sun Cluster nodes.
- Do not configure Sun Cluster systems as routers (gateways). If the system goes down, the clients cannot find an alternate router and recover.
- Do not configure Sun Cluster systems as NIS or NIS+ servers. Sun Cluster nodes can be NIS or NIS+ clients, however.
- A Sun Cluster configuration cannot be used to provide a highly available boot or install service to client systems.
- A Sun Cluster configuration cannot be used to provide highly available `rarpd` service.

2.6.2 Sun Cluster HA for NFS Restrictions

Note the following restrictions related to Sun Cluster HA for NFS.

- Do not run, on any Sun Cluster node, any applications that access the Sun Cluster HA for NFS file system locally. For example, on Sun Cluster systems, users should not locally access any Sun Cluster file systems that are NFS exported. This is because local locking interferes with the ability to kill and restart lockd(1M). Between the kill and the restart, a blocked local process is granted the lock, which prevents reclamation of the lock by the client machine.
- Sun Cluster does not support cross-mounting of Sun Cluster HA for NFS resources.
- Sun Cluster HA for NFS requires that all NFS client mounts be “hard” mounts.
- For Sun Cluster HA for NFS, do not use host name aliases for the logical hosts. NFS clients mounting HA file systems using host name aliases for the logical hosts might experience `statd` lock recovery problems.

- Sun Cluster does not support Secure NFS or the use of Kerberos with NFS. In particular, the –`secure` and –`kerberos` options to `share_nfs(1M)` are not supported.

2.6.3 Hardware Restrictions

Note the following hardware-related restrictions.

- A pair of Sun Cluster nodes must have at least two multihost disk enclosures, with one exception: if you use Sun StorEdge A3000 disks, you can use only one such expansion unit.
- The SS1000 and SC2000 hardware platforms are not supported with Sun Cluster 2.2 and Solaris 7. They are supported under Solaris 2.6. This restriction is due to the removal of support for the SFE 1.0 `be(7D)` driver under Solaris 7. That driver is used for the cluster interconnect on the SS1000 and SC2000 machines.
- The following restrictions apply only to Ultra 2 Series configurations:
  - The Sun Cluster node must be reinstalled to migrate from one basic hardware configuration to another. For example, a configuration with three FC/S cards and one SQEC card must be reinstalled to migrate to a configuration with two FC/S cards, one SQEC card, and one SFE or SunFDDI™ card.
  - Dual FC/OMs per FC/S card is supported only when used with the SFE or SunFDDI card.
  - In the SFE or SunFDDI 0 card configuration, the recovery mode of a dual FC/OM FC/S card failure is by a failover, not by mirroring or hot sparing.

2.6.4 Solstice DiskSuite Restrictions

Note the following restrictions related to Solstice DiskSuite.

- In Solstice DiskSuite configurations using mediators, the number of mediator hosts configured for a diskset must be an even number.
- The RAID5 feature in the Solstice DiskSuite product is not supported. RAID5 is supported under Sun StorEdge Volume Manager and Cluster Volume Manager. A hardware implementation of RAID5 is also supported by the Sun StorEdge A3000 disk expansion unit.

2.6.5 Other Restrictions

- In the event of a power failure that brings down the entire cluster, user intervention is required to restart the cluster. The administrator must determine the last node that went down (by examining `/var/adm/messages`) and run

`scadmin startcluster` on that node. Then the administrator must run `scadmin startnode` on the other cluster nodes to bring the cluster back online.

- Sun Cluster does not support the use of the loopback file system (`lofs`) on Sun Cluster nodes.
- Do not run client applications on the Sun Cluster nodes. Because of local interface group semantics, a switchover or failover of a logical host may cause a TCP (`telnet`/`rlogin`) connection to be broken. This includes both connections that were initiated by the server hosts of the cluster, as well as connections that were initiated by client hosts outside the cluster.
- Do not run, on any Sun Cluster node, any processes that run in the real-time scheduling class.
- Do not access the `/`*logicalhost* directories from shells on any nodes. If you have shell connections to any `/`*logicalhost* directories when a switchover or failover is attempted, the switchover or failover will be blocked.
- The Sun Cluster HA administrative file system cannot be grown using the Solstice DiskSuite `growfs(1M)` command.
- File system quotas are not supported in Sun Cluster.
- Logical network interfaces are reserved for use by Sun Cluster.
- Sun Prestoserve is not supported. Prestoserve works within the host system, which means that any data contained in the Prestoserve memory would not be available to the Sun Cluster sibling in the event of a switchover.

CHAPTER 3

Installing and Configuring Sun Cluster Software

This chapter contains guidelines and procedures for installing Sun Cluster 2.2.

This chapter includes the following procedures:

3.1 Installation Overview

This chapter includes the procedures used to install and configure Sun Cluster 2.2.

Before beginning the install procedures, complete the planning exercises described in Chapter 2. These exercises include planning your network connections, logical hosts, disk configuration, and file system layouts. Complete the installation worksheets in Appendix A. You will be prompted for information from the worksheets during the Sun Cluster 2.2 installation process. Then use the procedures in this chapter to install and configure the cluster.

The steps to configure and install Sun Cluster are grouped into three procedures:

1. **Preparing the administrative workstation and installing the client software.**

   This entails installing the Solaris operating environment and Sun Cluster 2.2 client software on the administrative workstation.

2. **Installing the server software.**

   This includes: using the Cluster Console to install the Solaris operating environment and Sun Cluster 2.2 software on all cluster nodes; using `scinstall(1M)` to set up network interfaces, logical hosts, and quorum devices; and selecting data services and volume manager support packages.

3. **Configuring and bringing up the cluster.**

   This includes: setting up paths; installing patches; installing and configuring your volume manager, SCI, PNM backup groups, logical hosts, and data services; and bringing up the cluster.

   If your installation is interrupted or if you make mistakes during any part of the install process, see Section 3.3 “Troubleshooting the Installation” on page 3-28, for instructions on troubleshooting and restarting the install.

3.2 Installation Procedures

This section describes how to install the Solaris operating environment and Sun Cluster client software on the administrative workstation.

▼ How to Prepare the Administrative Workstation and Install the Client Software

After you have installed and configured the hardware, terminal concentrator, and administrative workstation, use this procedure to prepare for Sun Cluster 2.2 Installation. See Chapter 2, and complete the installation worksheets in Appendix A, before beginning this procedure.

Note - Use of an administrative workstation is not required. If you do not use an administrative workstation, perform the administrative tasks from one designated node in the cluster.

These are the high-level steps to prepare the administrative workstation and install the client software:

- Installing the Solaris 2.6 or Solaris 7 operating environment and related patches on the administrative workstation
- Adding the tools directory to the PATH on the administrative workstation
- Using the `scinstall(1M)` command to install the client packages on the administrative workstation
- Changing the default port number used by Sun Cluster SNMP (it conflicts with the default port number used by Solaris SNMP)
- Modifying the `/etc/clusters` and `/etc/serialports` files

These are the detailed steps to prepare the administrative workstation and install the client software.

1. **Install the Solaris 2.6 or Solaris 7 operating environment on the administrative workstation.**

   All platforms except the E10000 require at least the Entire Distribution Solaris installation, for both the Solaris 2.6 and Solaris 7 operating environments. E10000 systems require the Entire Distribution + OEM.

   You can use the following command to verify the distribution loaded:

```
# cat /var/sadm/system/admin/CLUSTER
```

   For details, see Section 2.2.4 “Planning Your Solaris Operating Environment Installation” on page 2-5, and the *Solaris Advanced System Administration Guide.*

Caution - If you install anything less than the Entire Distribution Solaris software set on all nodes, plus the OEM packages for E10000 platforms, your cluster might not be supported by Sun.

2. **Install Solaris patches.**

   Check the patch database or contact your local service provider for any hardware or software patches required to run the Solaris operating environment, Sun Cluster 2.2, or your volume management software.

   Install the patches by following the instructions in the README file accompanying each patch. Reboot the workstation if specified in the patch instructions.

3. **For convenience, add the tools directory** `/opt/SUNWcluster/bin` **to the** PATH **on the administrative workstation.**

4. **Load the Sun Cluster 2.2 CD-ROM on the administrative workstation.**

5. Use `scinstall(1M)` to install the client packages on the administrative workstation.

a. As root, invoke `scinstall(1M)`.

```
# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
# ./scinstall

 Installing: SUNWscins

 Installation of <SUNWscins> was successful.

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

 None of the Sun Cluster software has been installed

    <<Press return to continue>>
```

b. Select the client package set.

```
==== Install/Upgrade Framework Selection Menu ======================
 Upgrade to the latest Sun Cluster Server packages or select package
sets for installation. The list of package sets depends on the Sun
Cluster packages that are currently installed.

 Choose one:
 1) Upgrade             Upgrade to Sun Cluster 2.2 Server packages
 2) Server              Install the Sun Cluster packages needed on a server
 3) Client
Install the admin tools needed on an admin workstation
 4) Server and Client   Install both Client and Server packages

 5) Close               Exit this Menu
 6) Quit                Quit the Program

 Enter the number of the package set [6]:  3
```

c. Choose an install path for the client packages.

Normally the default location is acceptable.

```
What is the path to the CD-ROM image [/cdrom/cdrom0]: /cdrom/suncluster_sc_2_2
```

d. Install the client packages.

Specify automatic installation.

```
Installing Client packages

 Installing the following packages: SUNWscch SUNWccon SUNWccp
 SUNWcsnmp SUNWscsdb

             >>>> Warning <<<<
   The installation process will run several scripts as root.  In
   addition, it may install setUID programs.  If you choose automatic
   mode, the installation of the chosen packages will proceed without
   any user interaction.  If you wish to manually control the install
   process you must choose the manual installation option.

 Choices:
  manual      Interactively install each package
  automatic      Install the selected packages with no user interaction.

 In addition, the following commands are supported:
    list      Show a list of the packages to be installed
    help      Show this command summary
    close      Return to previous menu
    quit      Quit the program


 Install mode [manual automatic] [automatic]:  automatic
```

The `scinstall(1M)` program now installs the client packages. After the packages have been installed, the main `scinstall(1M)` menu is displayed. From the main menu, you can choose to verify the installation, then quit to exit `scinstall(1M)`.

6. Change the port number used by the Sun Cluster SNMP daemon and Solaris SNMP (`smond`).

The default port used by Sun Cluster SNMP is the same as the default port number used by Solaris SNMP; both use port 161. Change the Sun Cluster SNMP port number using the procedure described in the appendix describign Sun Cluster SNMP management solutions in the *Sun Cluster 2.2 System Administration Guide*. You must stop and restart both the `snmpd` and `smond` daemons after changing the port number.

7. Modify the `/etc/clusters` and `/etc/serialports` files.

These files are installed automatically by `scinstall(1M)`. Use the templates included in the files to add your cluster name, physical host names, terminal concentrator name, and serial port numbers, as listed on your installation worksheet. See the `clusters(4)` and `serialports(4)` man pages for details.

Note - The serial port number used in the `/etc/serialports` file is the `telnet(1)` port number, not the physical port number. Determine the serial port number by adding 5000 to the physical port number. For example, if the physical port number is 6, the serial port number should be 5006.

Proceed to the section "How to Install the Server Software" on page 3–6 to install the Sun Cluster 2.2 server software.

▼ How to Install the Server Software

After you have installed the Sun Cluster 2.2 client software on the administrative workstation, use this procedure to install Solaris and the Sun Cluster 2.2 server software on all cluster nodes.

Note - This procedure assumes you are using an administrative workstation. If you are not, then connect directly to the console of each node using a `telnet` connection to the terminal concentrator. Install and configure the Sun Cluster software identically on each node.

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

Caution - If you already have a volume manager installed and a root disk encapsulated, unencapsulate the root disk before beginning the Sun Cluster installation.

These are the high-level steps to install the server software:

- Bringing up the Cluster Control Panel from the administrative workstation and starting the cluster console (`console` mode)
- Installing Solaris 2.6 or Solaris 7, including setting up partitions and configuring the OpenBoot PROM
- Updating naming services
- Installing Solaris patches on all nodes
- Modifying the `/etc/nsswitch.conf` and `/etc/services` files
- Configuring network adapter interfaces for any additional secondary subnets
- Using the `scinstall(1M)` command to:
  - Install the server packages on all nodes

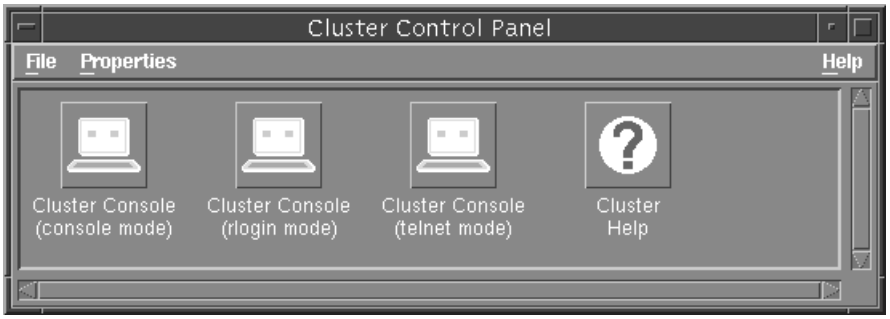
- Select your volume manager
- Specify the cluster name
- Configure private network interfaces
- Set up logical hosts
- Set up primary and secondary public networks and subnets
- Configure failure fencing
- (SSVM and CVM only) Select a quorum device
- Choose Cluster Membership Monitor behavior
- Select data services

These are the detailed steps to install the server software.

1. **Bring up the Cluster Control Panel from the administrative workstation.**

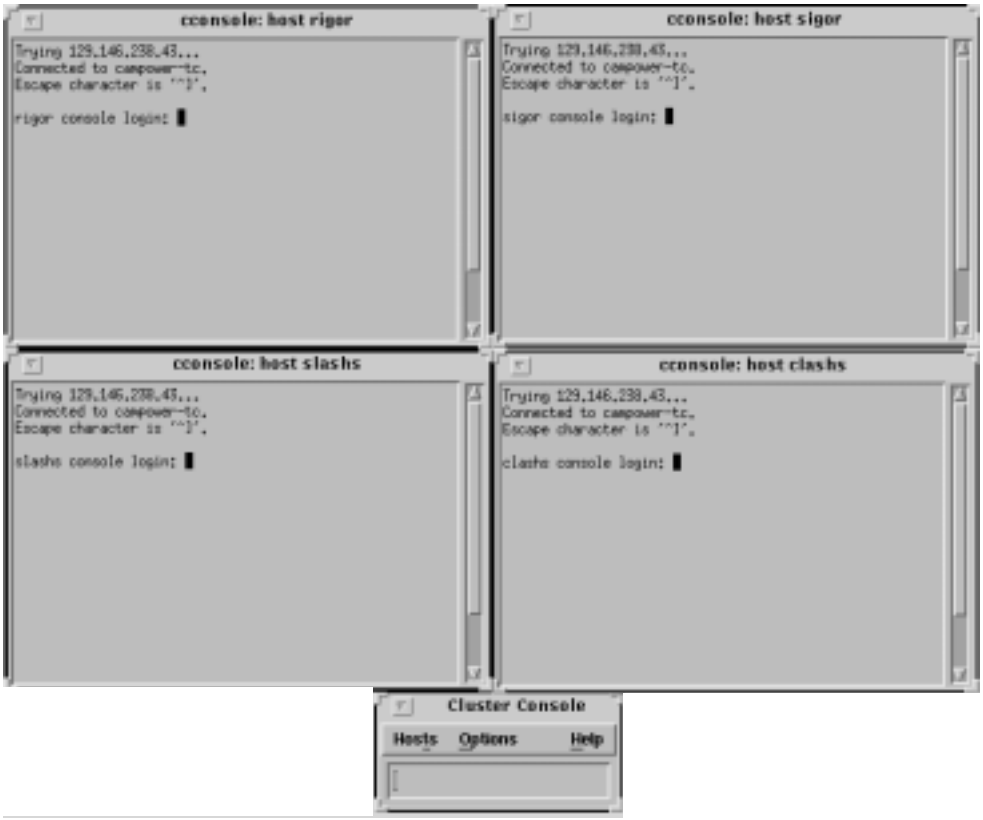
   In this example, the cluster name is `sc-cluster`.

```
# ccp sc-cluster
```



2. **Start the Cluster Console in console mode.**

   From the Cluster Control Panel, select the Cluster Console, console mode. The Cluster Console (CC) will display one window for each cluster node, plus a small common window that you can use to command all windows simultaneously.



Note - Individually, the windows act as vt100 terminal windows. Set your `TERM` type to equal `vt100`.

---

3. **Use the Cluster Console common window to install Solaris 2.6 or Solaris 7 on all nodes.**

   For details, see the *Solaris Advanced System Administration Guide*, and the Solaris installation guidelines described in Chapter 2.

   a. **Partition the local disks on each node to Sun Cluster and volume manager guidelines.**

      For partitioning guidelines, see Section 2.2.4 “Planning Your Solaris Operating Environment Installation” on page 2-5.

   b. **Configure the OpenBoot PROM.**

      If you want to boot from a SPARCstorage Array, you must configure the shared boot device, if you did not do so already during hardware installation. See Section 2.2.8.4 “Booting From a SPARCstorage Array” on page 2-17, for

details about setting up the shared boot device. If your configuration includes copper-connected SCSI storage devices such as Sun StorEdge MultiPacks, Sun StorEdge A1000s, and Sun StorEdge A3x00s, you also need to configure the `scsi-initiator-id`. See your hardware installation manuals for details about configuring the `scsi-initiator-id`.

4. **Update the naming service.**

   If a host name database such as NIS, NIS+, or DNS is used at your site, update the naming service with all logical and physical host names to be used in the Sun Cluster configuration.

5. **Use the Cluster Console common window to log into all nodes.**

6. **Install Solaris patches.**

   Check the patch database or contact your local service provider for any hardware or software patches required to run the Solaris operating environment, Sun Cluster 2.2, and any other software installed on your configuration.

   Install any required patches by following the instructions in the README file accompanying each patch, unless instructed otherwise by the Sun Cluster documentation or your service provider.

   Reboot all nodes if specified in the patch instructions.

7. **Modify the `/etc/nsswitch.conf` file.**

   Ensure that "hosts," "services," and "group" lookups are directed to files first. For example:

```
hosts: files nisplus
 services: files nisplus
 group: files nisplus
```

8. **(Optional) If your cluster serves more than one subnet, configure network adapter interfaces for additional secondary public networks.**

9. **As root, invoke `scinstall(1M)` from the CC common window.**

```
# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
# ./scinstall

 Installing: SUNWscins

 Installation of <SUNWscins> was successful.

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

    <<Press return to continue>>
```

10. Select the server package set.

```
==== Install/Upgrade Framework Selection Menu ============================
 You can upgrade to the latest Sun Cluster packages or select package
 sets for installation, depending on the current state of installation.

 Choose one:
 1) Upgrade         Upgrade to Sun Cluster 2.2
 2) Server         All of the Sun Cluster packages needed on a server
 3) Client         All of the admin tools needed on an admin workstation
 4) Server and Client         All of the Client and Server packages

 5) Close         Exit this Menu
 6) Quit         Quit the Program

 Enter the number of the package set [6]:  2
```

Press Return to continue.

11. Install the server packages.

Specify `automatic` installation. The `scinstall(1M)` program installs the server packages.

```
Installing Server packages

 Installing the following packages: SUNWsclb SUNWsc SUNWccd SUNWcmm SUNWff
 SUNWmond SUNWpnm SUNWscman SUNWsccf SUNWscmgr

             >>>> Warning <<<<
   The installation process will run several scripts as root.  In
   addition, it may install setUID programs.  If you choose automatic
   mode, the installation of the chosen packages will proceed without
   any user interaction.  If you wish to manually control the install
```

(continued)

```
  process you must choose the manual installation option.

Choices:
 manual     Interactively install each package
 automatic     Install the selected packages with no user interaction.

In addition, the following commands are supported:
 list     Show a list of the packages to be installed
 help     Show this command summary
 close     Return to previous menu
 quit     Quit the program

Install mode [manual automatic] [automatic]: automatic
```

The server package set is now installed.

12. Select your volume manager.

In this example, Solstice DiskSuite is specified.

```
Volume Manager Selection

 Please choose the Volume Manager that will be used
 on this node:

 1) Cluster Volume Manager (CVM)
 2) Sun Enterprise Volume Manager (SEVM)
 3) Solstice DiskSuite (SDS)

 Choose the Volume Manager: 3

 Installing Solstice DiskSuite support packages.
  Installing ‘‘SUNWdid’’ ... done
  Installing ‘‘SUNWmdm’’ ... done

      ---------WARNING---------
 Solstice DiskSuite (SDS) will need to be installed before the cluster can
 be started.

         <<Press return to continue>>
```

Note - You will still have to install the volume manager software from the Solstice DiskSuite, SSVM, or CVM media after you complete the cluster installation. This step installs only supporting software (such as drivers).

Caution - If you perform upgrades or package removals with `scinstall(1M)`, `scinstall(1M)` will not remove the `SUNWdid` package. Do NOT remove the `SUNWdid` package manually. Removing the package can cause loss of data.

13. Specify the cluster name.

```
What is the name of the cluster? sc-cluster
```

14. Specify the number of potential nodes and active nodes in your cluster.

You can specify up to four nodes. The active nodes are those you will physically connect and include in the cluster now. You must specify all potential nodes at this time; you will be asked for information such as node names and IP addresses. Later, you can change the status of nodes from potential to active by using the `scconf(1M)` command. See the section on adding and removing cluster nodes in the *Sun Cluster 2.2 System Administration Guide*.

Note - If you want to add a node later that was not already specified as a potential node, you will have to reconfigure the entire cluster.

```
How many potential nodes will sc-cluster have [4]? 3

 How many of the initially configured nodes will be active [3]? 3
```

Note - If your cluster will have two active nodes and only two disk strings and your volume manager is Solstice DiskSuite, you must configure mediators. Do so after you configure Solstice DiskSuite but before you bring up the cluster. See the chapter on using dual-string mediators in the *Sun Cluster 2.2 System Administration Guide* for the procedure.

15. Configure the private network interfaces, using the common window.

Select either Ethernet or Scalable Coherent Interface (SCI).

```
What type of network interface will be used for this configuration?
 (ether|SCI) [SCI]?
```

If you choose SCI, the following screen is displayed. Answer the questions using the information on your installation worksheet. Note that the node name field is case-sensitive; the node names specified here are checked against the `/etc/nodename` file by `scinstall`.

```
What is the hostname of node 0 [node0]? phys-hahost1

 What is the hostname of node 1 [node1]? phys-hahost2
...
```

Note - When nodes are connected through an SCI switch, the connection of the nodes to the switch port determines the order of the nodes in the cluster. The node number must correspond to the port number. For example, if a node named `phys-hahost1` is connected to port 0, then `phys-hahost1` must be node 0. In addition, each node must be connected to the same port on each switch. For example, if `phys-hahost1` is connected to port 0 on switch 0, it also must be connected to port 0 on switch 1.

If you choose Ethernet, the following screen is displayed. Answer the questions using information from the installation worksheet. Complete the network configuration for all nodes in the cluster.

```
What is the hostname of node 0 [node0]? phys-hahost1

 What is phys-hahost1's first private network interface [hme0]? hme0

 What is phys-hahost1's second private network interface [hme1]? hme1

 You will now be prompted for Ethernet addresses of
 the host. There is only one Ethernet address for each host
 regardless of the number of interfaces a host has. You can get
 this information in one of several ways:

 1. use the 'banner' command at the ok prompt,
```

(continued)

(Continuation)

```
2. use the 'ifconfig -a' command (need to be root),
3. use ping, arp and grep commands. ('ping exxon; arp -a | grep exxon')

Ethernet addresses are given as six hexadecimal bytes separated by colons.
ie, 01:23:45:67:89:ab

What is phys-hahost1's ethernet address? 01:23:45:67:89:ab

What is the hostname of node 1 [node1]?
...
```

16. Specify whether the cluster will support any data services and if so, whether to set up logical hosts.

```
Will this cluster support any HA data services (yes/no) [yes]? yes
Okay to set up the logical hosts for those HA services now (yes/
no) [yes]? yes
```

17. Set up primary public networks and subnets.

Enter the name of the network controller for the primary network for each node in the cluster.

```
What is the primary public network controller for ''phys-hahost1''?  hme2
What is the primary public network controller for ''phys-hahost2''?  hme2
```

18. Set up secondary public subnets.

If the cluster nodes will provide data services to more than a single public network, answer **yes** to this question:

```
Does the cluster serve any secondary public subnets (yes/no) [no]?  yes
```

19. Name the secondary public subnets.

Assign a name to each subnet. Note that these names are used only for convenience during configuration. They are not stored in the configuration database and need not match the network names returned by `networks(4)`.

```
Please enter a unique name for each of these additional subnets:

        Subnet name (^D to finish):  sc-cluster-net1
       Subnet name (^D to finish):  sc-cluster-net2
       Subnet name (^D to finish):  ^D

 The list of secondary public subnets is:

        sc-cluster-net1
        sc-cluster-net2

 Is this list correct (yes/no) [yes]?
```

20. Specify network controllers for the subnets.

For each secondary subnet, specify the name of the network controller used on each cluster node.

```
For subnet ‘‘sc-cluster-net1’’ ...
        What network controller is used for ‘‘phys-hahost1’’?  qe0
       What network controller is used for ‘‘phys-hahost2’’?  qe0

 For subnet ‘‘sc-cluster-net2’’ ...
        What network controller is used for ‘‘phys-hahost1’’?  qe1
       What network controller is used for ‘‘phys-hahost2’’?  qe1
```

21. Initialize Network Adapter Failover (NAFO).

You must initialize NAFO, and you must run `pnmset(1M)` later to configure the adapters. See the `pnmset(1M)` man page and the chapter on administering network interfaces in the *Sun Cluster 2.2 System Administration Guide* for more information about NAFO and PNM.

```
Initialize NAFO on ‘‘phys-hahost1’’ with one ctlr per group (yes/no) [yes]?  y
```

22. Set up logical hosts.

```
Enter the list of logical hosts you want to add:

         Logical host (^D to finish):  hahost1
        Logical host (^D to finish):  hahost2
        Logical host (^D to finish):  ^D

 The list of logical hosts is:

         hahost1
        hahost2

 Is this list correct (yes/no) [yes]? y
```

Note - You can add logical hosts or change the logical host configuration after the cluster is up by using `scconf(1M)` or the "Change" option to `scinstall(1M)`. See the `scinstall(1M)` and `scconf(1M)` man pages, and Step 11 on page @–27 in the procedure "How to Configure the Cluster" on page 3–22 for more information.

Note - If you will be using the Sun Cluster HA for SAP data service, do not set up logical hosts now. Set them up with `scconf(1M)` after the cluster is up. See the `scconf(1M)` man page and Chapter 10," for more information.

---

23. Assign default masters to logical hosts.

You must specify the name of a physical host in the cluster as a default master for each logical host.

```
What is the name of the default master for ''hahost1''?  phys-hahost1
```

Specify the host names of other physical hosts capable of mastering each logical host.

```
Enter a list of other nodes capable of mastering ''hahost1'':

         Node name:  phys-hahost2
        Node name (^D to finish): ^D

 The list that you entered is:

         phys-hahost1
```

(continued)

(Continuation)

```
        phys-hahost2

Is this list correct (yes/no) [yes]? yes
```

24. Enable automatic failback.

Answering yes enables the logical host to fail back automatically to its default master when the default master rejoins the cluster.

```
Enable automatic failback for ‘‘hahost1’’ (yes/no) [no]?  yes
```

25. Assign net names and disk group names.

```
What is the net name for ‘‘hahost1’’ on subnet ‘‘sc-cluster-
net1’’? hahost1-pub1
What is the net name for ‘‘hahost1’’ on subnet ‘‘sc-cluster-
net2’’? hahost1-pub2
Disk group name for logical host ‘‘hahost1’’ [hahost1]?
 Is it okay to add logical host ‘‘hahost1’’ now (yes/no) [yes]? yes

 What is the name of the default master for ‘‘hahost2’’?
 ...
```

Continue until all logical hosts are set up.

Note - To set up multiple disk groups on a single logical host, use the `scconf(1M)` command after you have used `scinstall(1M)` to configure and bring up the cluster. See the `scconf(1M)` man page for details.

---

26. If your volume manager is SSVM or CVM and there are more than two nodes in the cluster, configure failure fencing.

This screen will appear only for greater than two-node clusters using SSVM or CVM.

```
Configuring Failure Fencing

 What type of architecture does phys-hahost1 have (E10000|other) [other]?

 What is the name of the Terminal Concentrator connected to the serial
 port of phys-hahost1 [NO_NAME]? sc-tc

 Is 123.456.789.1 the correct IP address for this Terminal Concentrator
 (yes | no) [yes]?

 What is the password for root of the Terminal Concentrator [?]
 Please enter the password for root again [?]

 Which physical port on the Terminal Concentrator is phys-
hahost1 connected to:

 What type of architecture does phys-hahost2 have (E10000|other) [other]?
 Which Terminal Concentrator is phys-hahost2 connected to:

 0) sc-tc    123.456.789.1
 1) Create A New Terminal Concentrator Entry

 Select a device:

 Which physical port on the Terminal Concentrator is phys-
hahost2 connected to:

 What type of architecture does phys-hahost3 have (E10000|other) [other]?
 Which Terminal Concentrator is phys-hahost3 connected to:

 0) sc-tc    123.456.789.1
 1) Create A New Terminal Concentrator Entry

 Select a device:

 Which physical port on the Terminal Concentrator is phys-
hahost3 connected to:

 Finished Configuring Failure Fencing
```

Caution - The SSP password is used in failure fencing. Failure to correctly set the SSP password might cause unpredictable results in the event of a node failure. If you change the SSP password, you must change it on the cluster as well, using `scconf(1M)`. Otherwise, failure fencing will be disabled because the SSP cannot connect to the failed node. See the `scconf(1M)` man page and the *Sun Cluster 2.2 System Administration Guide* for details about changing the SSP password.

27. If your volume manager is SSVM, your cluster has more than two nodes, and you have a direct-attached device, select a nodelock port.

The port you select must be on a terminal concentrator attached to a node in the cluster.

```
Does the cluster have a disk storage device that is
connected to all nodes in the cluster [no]? yes

Which unused physical port on the Terminal Concentrator is to be used for
node locking:
```

28. If your volume manager is SSVM or CVM, select quorum devices.

If your volume manager is SSVM or CVM, you are prompted to select quorum devices. The screen display varies according to your cluster topology. Select a device from the list presented. This example shows a two-node cluster.

```
Getting device information for reachable nodes in the cluster.
 This may take a few seconds to a few minutes...done
 Select quorum device for the following nodes:
 0 (phys-hahost1)
 and
 1 (phys-hahost2)

 1) SSA:000000779A16
 2) SSA:000000741430
 3) DISK:c0t1d0s2:01799413
 Quorum device: 1
...
 SSA with WWN 000000779A16 has been chosen as the quorum device.

 Finished Quorum Selection
```

29. If your cluster has greater than two nodes, select Cluster Membership Monitor behavior.

```
In the event that the cluster is partitioned into two or more subsets of
 nodes, the Cluster Membership Monitor may request input from the operator as
 to how it should proceed (abort or form a cluster) within each subset.  The
 Cluster Membership Monitor can be configured to make a policy-dependent
 automatic selection of a subset to become the next reconfiguration of
 the cluster.

 In case the cluster partitions into subsets, which subset should stay up?
     ask)    the system will always ask the operator.
     select) automatic selection of which subset should stay up.

 Please enter your choice (ask|select) [ask]:
```

If you choose “select,” you are asked to choose between two policies:

```
Please enter your choice (ask|select) [ask]: select

 You have a choice of two policies:

  lowest -- The subset containing the node with the lowest node ID value
      automatically becomes the new cluster.  All other subsets must be
     manually aborted.

  highest -- The subset containing the node with the highest node ID value
     automatically becomes the new cluster.  All other subsets must be
     manually aborted.

 Select the selection policy for handling partitions (lowest|highest)
 [lowest]:
```

The `scinstall(1M)` program now finishes installing the Sun Cluster 2.2 server packages.

```
Installing ethernet Network Interface packages.

 Installing the following packages: SUNWsma
     Installing "SUNWsma" ... done

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

30. Select your data services.

Note that Sun Cluster HA for NFS and Informix-Online XPS are installed automatically with the Server package set.

```
==== Select Data Services Menu ==========================

 Please select which of the following data services are to
 be installed onto this cluster.  Select singly, or in a
 space separated list.
 Note: HA-NFS and Informix Parallel Server (XPS) are
 installed automatically with the Server Framework.

 You may de-select a data service by selecting it a second time.

 Select DONE when finished selecting the configuration.

     1) Sun Cluster HA for Oracle
     2) Sun Cluster HA for Informix
     3) Sun Cluster HA for Sybase
     4) Sun Cluster HA for Netscape
     5) Sun Cluster HA for Netscape LDAP
     6) Sun Cluster HA for Lotus
     7) Sun Cluster HA for Tivoli
     8) Sun Cluster HA for SAP
     9) Sun Cluster HA for DNS
     10) Sun Cluster for Oracle Parallel Server

 INSTALL    11) No Data Services
     12) DONE

 Choose a data service: 3

 What is the path to the CD-ROM image [/cdrom/suncluster_sc_2_2]:

 Install mode [manual automatic] [automatic]:  automatic
...
 Select DONE when finished selecting the configuration.
 ...
```

Note - Do not install the OPS data service unless you are using Cluster Volume Manager. OPS will not run with Sun StorEdge Volume Manager or Solstice DiskSuite.

---

31. Quit `scinstall(1M)`**.**

```
============ Main Menu =================

 1) Install/Upgrade - Install or Upgrade Server Packages or Install Client
        Packages.
 2) Remove  - Remove Server or Client Packages.
 3) Change  - Modify cluster or data service configuration
```

(continued)

(Continuation)

```
 4) Verify  - Verify installed package sets.
 5) List    - List installed package sets.

 6) Quit    - Quit this program.
 7) Help    - The help screen for this menu.

 Please choose one of the menu items: [6]:  6  ...
```

The `scinstall(1M)` program now verifies installation of the packages you selected.

```
==== Verify Package Installation ==========================
 Installation
  All of the install          packages have been installed
 Framework
  All of the client          packages have been installed
  All of the server          packages have been installed
 Communications
  All of the SMA          packages have been installed
 Data Services
  None of the Sun Cluster HA for Oracle packages have been installed
  None of the Sun Cluster HA for Informix packages have been installed
  None of the Sun Cluster HA for Sybase packages have been installed
  None of the Sun Cluster HA for Netscape packages have been installed
  None of the Sun Cluster HA for Lotus packages have been installed
  None of the Sun Cluster HA for Tivoli packages have been installed
  None of the Sun Cluster HA for SAP packages have been installed
  None of the Sun Cluster HA for Netscape LDAP packages have been installed
  None of the Sun Cluster HA for Oracle Parallel Server packages have been
 installed
 #
```

Proceed to the section “How to Configure the Cluster” on page 3–22 to configure the cluster.

▼ How to Configure the Cluster

After installing the Sun Cluster 2.2 client and server packages, complete the following post-installation tasks.

This is the high-level list of steps to perform to configure the cluster:

- Setting up software directory paths on all nodes
- Adding IP addresses to `/.rhosts` files on all nodes
- (SCI only) Modifying the `sm_config` file to comment out nodes specified as potential rather than active
- (SCI only) Configuring SCI private interconnect switches by running the `sm_config(1M)` command
- Installing Sun Cluster patches
- Rebooting all nodes
- (SSVM and CVM only) Installing and configuring SSVM or CVM
- Using the `pnmset(1M)` command to configure NAFO backup groups
- Starting the cluster
- (Solstice DiskSuite only) Installing and configuring Solstice DiskSuite
- (Optional) Configuring additional logical hosts
- (SSVM only) Configuring the shared CCD volume
- Configuring and activating the HA data services
- Set up and start Sun Cluster Manager.

These are the detailed steps to configure the cluster.

1. **Set up the software directory paths on all nodes.**
   a. **On all nodes, set your** `PATH` **to include** `/sbin`**,** `/usr/sbin`**,** `/opt/SUNWcluster/bin`**, and** `/opt/SUNWpnm/bin`**. Set your** `MANPATH` **to include** `/opt/SUNWcluster/man`**.**

   b. **On all nodes, set your** `PATH` **and** `MANPATH` **to include the volume manager specific paths.**

      For SSVM and CVM, set your `PATH` to include `/opt/SUNWvxva/bin` and `/etc/vx/bin`. Set your `MANPATH` to include `/opt/SUNWvxva/man` and `/opt/SUNWvxvm/man`.

      For Solstice DiskSuite, set your `PATH` to include `/usr/opt/SUNWmd/sbin`. Set your `MANPATH` to include `/usr/opt/SUNWmd/man`.

   c. **If you are using Scalable Coherent Interface (SCI) for the private interfaces, set the SCI paths.**

      Set your `PATH` to include `/opt/SUNWsci/bin`, `/opt/SUNWscid/bin`, and `/opt/SUNWsma/bin`. Set your `MANPATH` to include `/opt/SUNWsma/man`.

2. **Add IP addresses to the** `/.rhosts` **file.**

   You must include the following hardcoded private network IP addresses in the `/.rhosts` files on all nodes. For a two node cluster, include only the addresses

specified for nodes 0 and 1 below. For a three node cluster, include the addresses specified for nodes 0, 1, and 2 below. For a four node cluster, include all addresses noted below:

```
# node 0
 204.152.65.33
 204.152.65.1
 204.152.65.17

 # node 1
 204.152.65.34
 204.152.65.2
 204.152.65.18

 # node 2
 204.152.65.35
 204.152.65.3
 204.152.65.19

 # node 3
 204.152.65.36
 204.152.65.4
 204.152.65.20
```

Note - If you fail to include the private network IP addresses in `/.rhosts`, the `hadsconfig(1M)` script will be unable to automatically replicate data service configuration information to all nodes when you configure your data services. You will then need to replicate the configuration file manually as described in the `hadsconfig(1M)` man page.

---

3. **If you are using SCI for the private interfaces and if you specified any potential nodes during server software installation, modify the `sm_config` file.**

   During server software installation with `scinstall(1M)`, you specified active and potential nodes. Edit the `sm_config` file now to comment out the host names of the potential nodes, by prepending the characters "`_%`" to those host names. In this example `sm_config` file, `phys-host1` and `phys-host2` are the active nodes, and `phys-host3` and `phys-host4` are potential nodes to be added to the cluster later.

```
HOST 0 = phys-host1
 HOST 1 = phys-host2
 HOST 2 = _%phys-host3
 HOST 3 = _%phys-host4
```

4. **If you are using SCI for the private interfaces, configure the switches with the `sm_config(1M)` command.**

   You must edit a copy of the `sm_config` template file (`template.sc` located in `/opt/SUNWsma/bin/Examples`) before running the `sm_config(1M)` command. See the `sm_config(1M)` man page and the procedure describing how to add switches and SCI cards in the *Sun Cluster 2.2 System Administration Guide* for details.

Caution - Run the `sm_config(1M)` command on only one node.

```
# sm_config -f templatefile
```

5. **Install Sun Cluster 2.2 patches.**

   Check the patch database or contact your local service provider for any hardware or software patches required to run Sun Cluster 2.2.

   Install any required patches by following the instructions in the `README` file accompanying each patch.

6. **Reboot all nodes.**

   This reboot creates device files for the Sun Cluster device drivers installed by `scinstall(1M)`, and also might be required by some patches you installed in Step 5 on page @–25.

Caution - You must reboot all nodes at this time, even if you did not install SCI or patches.

7. **(SSVM or CVM only) Install and configure SSVM or CVM.**

   Install and configure your volume manager and volume manager patches, using your volume manager documentation.

   This process includes installing the volume manager and patches, creating plexes and volumes, setting up the HA administrative file system (SSVM only), and updating the `vfstab.`*logicalhost* files (SSVM only). Refer to Chapter 2, and to Appendix C, for details. For CVM, refer also to the section on installing Cluster Volume Manager in the *Sun Cluster 2.2 Cluster Volume Manager Guide*.

   Create and populate disk groups and volumes now, but release them before continuing.

8. Configure NAFO backup groups, if you did not do so already.

During initial installation, you can use the `scinstall(1M)` command to install the PNM package, `SUNWpnm`, to configure one controller per NAFO backup group and to initialize PNM.

Note - You must configure a public network adaptor with either `scinstall(1M)` or `pnmset(1M)`, even if you have only one public network connection per node.

Run the `pnmset(1M)` command now if you did not already use `scinstall(1M)` to configure controllers and initialize PNM, or if you want to assign more than one controller per NAFO backup group. The `pnmset(1M)` command runs as an interactive script.

```
# /opt/SUNWpnm/bin/pnmset
```

See the chapter on administering network interfaces in the *Sun Cluster 2.2 System Administration Guide* or the `pnmset(1M)` man page for details.

9. Start the cluster.

Note - If you are using Solstice DiskSuite and you set up logical hosts as part of the server software installation (of the procedure “How to Install the Server Software” on page 3–6), you will see error messages as you start the cluster and it attempts to bring the logical hosts online. The messages will indicate that the Solstice DiskSuite disksets have not been set up. You can safely ignore these messages as you will set up the disksets in Step 10 on page @–27.

---

a. Run the following command on one node.

```
# scadmin startcluster phys-hahost1 sc-cluster
```

Note - If your volume manager is Cluster Volume Manager, you must set up shared disk groups at this point, before the other nodes are added to the cluster.

---

b. Add all other nodes to the cluster by running the following command from each node being added.

```
# scadmin startnode
```

c. Verify that the cluster is running.

From any cluster node, check activity with `hastat(1M)`:

```
# hastat
```

10. (Solstice DiskSuite only) Install and configure Solstice DiskSuite.

This process includes installing the volume manager and patches, creating disksets, setting up the HA administrative file system, and updating the `vfstab.`*logicalhost* files. Refer to Chapter 2, and to Appendix B, for details.

Create and populate disk groups and volumes now, but release them before continuing.

If you have a two-node configuration with only two disk strings, you also must set up mediators. Do so after configuring Solstice DiskSuite. See the chapter on using dual-string mediators in the *Sun Cluster 2.2 System Administration Guide* for instructions.

11. Add logical hosts, if you did not do so already.

Use the "Change" option to `scinstall(1M)` to add and configure logical hosts, if you did not set up all logical hosts during initial installation, or if you want to change the logical host configuration.

To set up multiple disk groups on a single logical host, you must use the `scconf(1M)` command, after you have brought up the cluster. See the `scconf(1M)` man page for details.

See the section on adding and removing logical hosts in the *Sun Cluster 2.2 System Administration Guide*, for more information.

Note - When you use `scinstall(1M)` to add logical hosts initially, you run the command from all hosts before the cluster has been brought up. When you use `scinstall(1M)` to re-configure existing logical hosts, you run the command from only one node while the cluster is up.

---

12. Add logical host names to the `/etc/hosts` files on all nodes.

For example:

```
#
 # Internet host table
 #
 127.0.0.1          localhost
```

(continued)

(Continuation)

```
123.168.65.23    phys-hahost1      loghost
123.146.84.36    123.146.84.36
123.168.65.21    hahost1
123.168.65.22    hahost2
```

13. Bring the logical hosts on line.

Use `haswitch(1M)` to force a cluster reconfiguration that will cause all logical hosts to be mastered by their default masters.

```
# haswitch -r
```

14. (Optional) If your cluster has only two nodes and your volume manager is SSVM, configure the shared CCD volume.

Use the procedures described in Appendix C, to configure a shared CCD volume.

15. Configure and activate the HA data services.

See the relevant data service chapter in this book, and the specific data service documentation for details.

16. Set up and start Sun Cluster Manager.

Sun Cluster Manager is used to monitor the cluster. For instructions, see the *Sun Cluster 2.2 Release Notes* and the section on monitoring the Sun Cluster servers with Sun Cluster Manager in the *Sun Cluster 2.2 System Administration Guide*.

This completes the cluster configuration.

3.3 Troubleshooting the Installation

Table 3–1 describes some common installation problems and solutions.

TABLE 3–1 Common Sun Cluster Installation Problems and Solutions

Problem Description	Solution
When you start a cluster node, it cannot join the cluster because the private net is not configured correctly.	Specify the correct private net interface by running the `scconf(1M)` command with the `–i` option. Then restart the cluster.
When you start a cluster node, it aborts after a failed reservation attempt, because of an incorrectly specified Ethernet address for one of the private nets.	Specify the correct Ethernet address of the node by running the `scconf(1M)` command with the `–N` option. Then restart the cluster.
If the cluster contains an invalid quorum device, the first node is unable to join the cluster because it cannot reserve the quorum device.	Specify a valid quorum device (controller or disk) by running the `scconf(1M)` command with the `–q` option. After configuring a valid quorum device, restart the cluster.
When you try to start the cluster, one node aborts after receiving signals from node 0 to do so.	The problem might be mismatched CDB files (`/etc/opt/SUNWcluster/conf/`*clustername*`.cdb`). Compare the CDB files on the different nodes using `cksum`. If they differ, copy the CDB file from the working node to the other node(s). You also might need to copy over the `ccd.database.init` file from the working node to the other nodes.

3.3.1 Recovering From an Aborted Installation

If your `scinstall(1M)` session did not run to completion during either the client or server installation process, you can re-run `scinstall(1M)` after cleaning up the environment using this procedure.

▼ How to Recover From an Aborted Client Installation

1. **On the administrative workstation, save the `/etc/serialports` and `/etc/clusters` files to a safe location, to be restored later.**

2. **On the administrative workstation, use `pkgrm` to remove the client packages.**

3. **Use `scinstall(1M)` to remove the Sun Cluster 2.2 client packages that have been installed already.**

```
# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
# ./scinstall

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

 1) Install/
Upgrade - Install or Upgrade Server Packages or Install Client Packages.
 2) Remove  - Remove Server or Client Packages.
 3) Change  - Modify cluster or data service configuration
 4) Verify  - Verify installed package sets.
 5) List    - List installed package sets.

 6) Quit    - Quit this program.
 7) Help    - The help screen for this menu.

 Please choose one of the menu items: [6]:  2
```

4. **Rerun** `scinstall(1M)` **using the procedure "How to Prepare the Administrative Workstation and Install the Client Software" on page 3–2.**

5. **Restore the** `/etc/serialports` **and** `/etc/clusters` **files you saved in .**

▼ How to Recover From an Aborted Server Installation

1. **If** `dfstab.`***logicalhost*** **and** `vfstab.`***logicalhost*** **files exist already, save them to a safe location to be restored later.**

   Look for the files in `/etc/opt/SUNWcluster/conf/hanfs`. You will restore these files after re-running `scinstall(1M)` and configuring the cluster.

2. **Use** `scinstall(1M)` **to remove the Sun Cluster 2.2 server packages that have been installed already.**

```
# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
# ./scinstall

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

 1) Install/Upgrade - Install or Upgrade Server
    Packages or Install Client Packages.
 2) Remove  - Remove Server or Client Packages.
 3) Change  - Modify cluster or data service configuration
```

(continued)

```
4) Verify  - Verify installed package sets.
5) List    - List installed package sets.

6) Quit    - Quit this program.
7) Help    - The help screen for this menu.

Please choose one of the menu items: [6]:  2
```

3. **Manually remove the following Sun Cluster 2.2 directories and files from all nodes.**

Caution - The `scinstall(1M)` command will not remove the `SUNWdid` package. Do NOT remove the `SUNWdid` package manually. Removing the package can cause loss of data.

Note that some of these directories might have been removed already by `scinstall(1M)`.

```
# rm /etc/pnmconfig
# rm /etc/sci.ifconf
# rm /etc/sma.config
# rm /etc/sma.ip
# rm -r /etc/opt/SUNWcluster
# rm -r /etc/opt/SUNWpnm
# rm -r /opt/SUNWcluster
# rm -r /opt/SUNWpnm
# rm -r /var/opt/SUNWcluster
```

4. **Restart `scinstall(1M)` to install Sun Cluster 2.2.**

   Return to the procedure “How to Install the Server Software” on page 3–6 and begin at Step 3 on page @–3.

5. **Configure the cluster.**

   Use the procedure “How to Configure the Cluster” on page 3–22.

6. **Restore the `dfstab.`*logicalhost* and `vfstab.`*logicalhost* files you saved in Step 1.**

Before starting the cluster, restore the `dfstab.`*logicalhost* and `vfstab.`*logicalhost* files to `/etc/opt/SUNWcluster/conf/hanfs` on all nodes.

CHAPTER 4

Upgrading Sun Cluster Software

This chapter contains guidelines and procedures for upgrading to Sun Cluster 2.2 from Solstice HA 1.3, Sun Cluster 2.0, and Sun Cluster 2.1.

The software to be upgraded might include the Solaris operating environment, Sun Cluster, and volume management software (Solstice DiskSuite, Sun StorEdge Volume Manager, or Cluster Volume Manager).

This chapter includes the following procedures:

4.1 Upgrade Overview

This section describes the procedures for upgrading to Sun Cluster 2.2 from existing Solstice HA 1.3, Sun Cluster 2.0, and Sun Cluster 2.1 configurations. The upgrade paths documented here preserve the existing cluster configuration and data. Your systems can remain online and available during most of the upgrade, keeping interruption to services minimal.

To upgrade from Solstice HA 1.3:

- Use the procedure “How to Upgrade From Solstice HA 1.3 to Sun Cluster 2.2” on page 4–2.
- If your volume manager is Solstice DiskSuite and you will be upgrading to Solaris 7, you will be required to upgrade to Solstice DiskSuite 4.2.

To upgrade from Sun Cluster 2.0 or 2.1:

- Use the planning information and procedures in Section 4.3 “Upgrading From Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2” on page 4-9.
- If your volume manager is SEVM 2.4 and you will be running Solaris 2.6, you must upgrade your volume manager to SEVM 2.5.

Note - Converting from Solstice DiskSuite to SEVM 2.5, SSVM or CVM is not supported by Sun Cluster 2.2.

If you also want to make configuration changes such as adding disks or services, first complete the upgrade and then make the configuration changes by following the procedures documented in the *Sun Cluster 2.2 System Administration Guide*.

Before starting your upgrade, make sure the versions of any applications you plan to run are compatible with the version of the Solaris operating environment you plan to run.

To upgrade Solaris, you might need to increase the size of your root (/) and `/usr` partitions on the root disks of all Sun Cluster servers in the configuration to accommodate the Solaris 2.6 or Solaris 7 environment. You must install the Entire Distribution Solaris software packages. See the *Solaris Advanced Installation Guide* for details.

4.2 Upgrading From Solstice HA 1.3 to Sun Cluster 2.2

Note - While performing this upgrade, you might see network interface and mediator errors on the console. These messages are side effects of the upgrade and can be ignored safely.

▼ How to Upgrade From Solstice HA 1.3 to Sun Cluster 2.2

These are the high-level steps to upgrade from Solstice HA 1.3 to Sun Cluster 2.2. You can perform the upgrade either from an administrative workstation or from the

console of any physical host in the cluster. Upgrading by using an administrative workstations provides the most flexibility during the process.

Note - This procedure assumes you are using an administrative workstation.

- Installing 2.2 client packages on the administrative workstation
- Stopping Solstice HA on the server to be upgraded first
- If your pre-upgrade cluster was running on Solaris 2.5.1, upgrading to Solaris 2.6 or Solaris 7
- If you are upgrading the Solaris operating environment, updating the Solaris kernel driver configuration files
- Upgrading the volume manager to Solstice DiskSuite 4.2.
- Loading Sun Cluster 2.2 by using the `scinstall(1M)` command
- Installing required patches on the local host
- Rebooting the local host and verifying the installation
- Stopping Solstice HA on the remote host
- Starting Sun Cluster 2.2 on the local host
- Repeating the upgrade steps on the remote host and verifying the upgrade

---

Caution - Back up all local and multihost disks before starting the upgrade. Also, all systems must be operable and robust. Do not attempt to upgrade if systems are experiencing any difficulties.

Caution - On each node, if you customized `hasap_start_all_instances` or `hasap_stop_all_instances` scripts in Solstice HA 1.3 or Sun Cluster 2.1, save them to a safe location before beginning the upgrade to Sun Cluster 2.2. Restore the scripts after completing the upgrade. Do this to prevent loss of your customizations when Sun Cluster 2.2 removes the old scripts.

The configuration parameters implemented in Sun Cluster 2.2 are different from those implemented in Solstice HA 1.3 and Sun Cluster 2.1. Therefore, after upgrading to Sun Cluster 2.2, you will have to re-configure Sun Cluster HA for SAP by running the `hadsconfig(1M)` command. Before starting the upgrade, view the existing configuration and note the current configuration variables. For Solstice HA 1.3, use the `hainetconfig(1M)` command to view the configuration. For Sun Cluster 2.1, use the `hadsconfig(1M)` command to view the configuration. After upgrading to Sun Cluster 2.2, use the `hadsconfig(1M)` command to re-create the instance.

These are the detailed steps to upgrade from Solstice HA 1.3 to Sun Cluster 2.2.

1. Load the Sun Cluster 2.2 client packages onto the administrative workstation.

Refer to Section 3.2 “Installation Procedures” on page 3-2, to set up the administrative workstation, if you have not done so already.

2. Stop Solstice HA on the first server to be upgraded.

```
phys-hahost1# hastop
```

If your cluster is already running Solaris 2.6, and you do not want to upgrade to Solaris 7, skip to Step 5 on page @–5.

3. Upgrade the operating environment to Solaris 2.6 or Solaris 7.

To upgrade Solaris, you must use the `suninstall(1M)` upgrade procedure (rather than reinstalling the operating environment). You might need to increase the size of your root (`/`) and `/usr` partitions on the root disks of all Sun Cluster servers in the configuration to accommodate the Solaris 2.6 or Solaris 7 environment. You must install the Entire Distribution Solaris software packages. See the *Solaris Advanced Installation Guide* for details.

Note - For some hardware platforms, Solaris 2.6 and Solaris 7 attempts to configure power management settings to shut down the server automatically if it has been idle for 30 minutes. The cluster heartbeat is not enough to prevent the Sun Cluster servers from appearing idle and shutting down. Therefore, you must disable this feature when you install Solaris 2.6 or Solaris 7. The dialog used to configure power management settings is shown below. If you do not see this dialog, then your hardware platform does not support this feature. If the dialog appears, you must answer –`n` to the first question and –`y` to the second to configure the server to work correctly in the Sun Cluster environment.

```
************************************************************
 This system is configured to conserve energy.
 After 30 minutes without activity, the system state will be
 saved to disk and the system will be powered off automatically.

 A system that has been suspended in this way can be restored
 back to exactly where it was by pressing the power key.
 The definition of inactivity and the timeout are user
 configurable. The dtpower(1M) man page has more information.
 ************************************************************

 Do you wish to accept this default configuration, allowing
 your system to save its state then power off automatically
 when it has been idle for 30 minutes?  (If this system is used
 as a server, answer n. By default autoshutdown is
 enabled.) [y,n,?] n

 Autoshutdown disabled.
```

(continued)

(Continuation)

```
Should the system save your answer so it won't need to ask
the question again when you next reboot? (By default the
question will not be asked again.) [y,n,?] y
```

4. Update the Solaris 2.6 or Solaris 7 kernel files.

As part of the Solaris upgrade, the files `/kernel/drv/sd.conf` and `/kernel/drv/ssd.conf` will be renamed to `/kernel/drv/sd.conf:2.`*x* and `/kernel/drv/ssd.conf:2.`*x* respectively. New `/kernel/drv/sd.conf` and `/kernel/drv/ssd.conf` files will be created. Run the `diff(1)` command to identify the differences between the old files and the new ones. Copy the additional information that was inserted by Sun Cluster from the old files into the new files. The information will look similar to the following:

```
# Start of lines added by Solstice HA
 sd_retry_on_reservation_conflict=0;
 # End of lines added by Solstice HA
```

5. Upgrade to Solstice DiskSuite 4.2.

a. Upgrade Solstice DiskSuite using the detailed procedure in the *Solstice DiskSuite 4.2 Installation and Product Notes*.

b. On the local host, upgrade the Solstice DiskSuite mediator package, `SUNWmdm`.

```
phys-hahost1# pkgadd -d /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol2_x/ \
 Product/ SUNWmdm

 Processing package instance <SUNWmdm>...

 Solstice DiskSuite (Mediator)
 (sparc) 4.2,REV=1998.23.10.09.59.06
 Copyright 1998 Sun Microsystems, Inc. All rights reserved.

 ## Executing checkinstall script.
    This is an upgrade. Conflict approval questions may be
    displayed. The listed files are the ones that will be
```

(continued)

(Continuation)

```
   upgraded. Please answer "y" to these questions if they are
   presented.
 Using </> as the package base directory.
 ## Processing package information.
 ## Processing system information.
   10 package pathnames are already properly installed.
 ## Verifying package dependencies.
 ## Verifying disk space requirements.
 ## Checking for conflicts with packages already installed.

 The following files are already installed on the system and are
 being used by another package:
   /etc/opt/SUNWmd/meddb
   /usr/opt <attribute change only>
   /usr/opt/SUNWmd/man/man1m/medstat.1m
   /usr/opt/SUNWmd/man/man1m/rpc.metamedd.1m
   /usr/opt/SUNWmd/man/man4/meddb.4
   /usr/opt/SUNWmd/man/man7/mediator.7
   /usr/opt/SUNWmd/sbin/medstat
   /usr/opt/SUNWmd/sbin/rpc.metamedd

 Do you want to install these conflicting files [y,n,?,q] y
## Checking for setuid/setgid programs.

 This package contains scripts which will be executed with super-user
 permission during the process of installing this package.

 Do you want to continue with the installation of <SUNWmdm.2> [y,n,?] y

 Installing Solstice DiskSuite (Mediator) as <SUNWmdm.2>
 ...
```

6. **From the root (/) directory on the local host, use the** `scinstall(1M)` **command to update the cluster packages.**

   Select `Upgrade` from the `scinstall(1M)` menu. Respond to the prompts asking for the location of the Framework packages and cluster name. The `scinstall(1M)` command replaces Solstice HA 1.3 packages with Sun Cluster 2.2 packages.

```
phys-hahost1# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
phys-hahost1# ./scinstall
Installing: SUNWscins

 Installation of <SUNWscins> was successful.
```

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

None of the Sun Cluster software has been installed

        <<Press return to continue>>

==== Install/Upgrade Software Selection Menu ========================
Upgrade to the latest Sun Cluster Server packages or select package
sets for installation. The list of package sets depends on the Sun
Cluster packages that are currently installed.

Choose one:
1) Upgrade            Upgrade to Sun Cluster 2.2 Server packages
2) Server             Install the Sun Cluster packages needed on a server
3) Client             Install the admin tools needed on an admin workstation
4) Server and Client  Install both Client and Server packages

5) Close              Exit this Menu
6) Quit               Quit the Program

Enter the number of the package set [6]: 1

What is the directory where the Framework packages can be found
[/cdrom/cdrom0]: .

** Upgrading from Solstice HA 1.3 **

What is the name of the cluster? sc-cluster
...
```

7. Install the required patches for Sun Cluster 2.2.

Install all applicable Solstice DiskSuite and Sun Cluster patches. If you are using SPARCstorage Arrays, the latest SPARCstorage Array patch should have been installed when you installed the operating environment. Obtain the necessary patches from Sun Enterprise Services. Use the instructions in the patch README files to install the patches.

8. Reboot the machine.

```
phys-hahost1# reboot
```

9. Switch ownership of disks and data services from the remote host to the upgraded local host.

a. Stop Solstice HA 1.3 services on the remote host.

The remote host in this example is `phys-hahost2`.

```
phys-hahost2# hastop
```

b. After Solstice HA 1.3 is stopped on the remote host, start Sun Cluster 2.2 on the upgraded local host.

Since the remote host is no longer running HA, use the `scadmin(1M)` command to start Sun Cluster. This causes the upgraded local host to take over all data services. In this example, `phys-hahost1` is the local physical host name, and `sc-cluster` is the cluster name.

```
phys-hahost1# scadmin startcluster phys-hahost1 sc-cluster
```

c. Verify that the configuration on the local host is stable.

```
phys-hahost1# hastat
```

d. Verify that clients are receiving services from the local host.

10. Repeat Step 2 on page @–4 through Step 8 on page @–7 on the remote host.

11. Return the remote host to the cluster.

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

12. After cluster reconfiguration on the remote host is complete, switch over the data services to the remote host from the local host.

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

13. Verify that the Sun Cluster 2.2 configuration on the remote host is in a stable state, and that clients are receiving services.

```
phys-hahost2# hastat
```

This completes the procedure to upgrade from Solstice HA 1.3 to Sun Cluster 2.2.

4.3 Upgrading From Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2

To upgrade from Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2, you must upgrade the Sun Cluster client software on the administrative workstation or install server, and then upgrade the Sun Cluster server software on all nodes in the cluster. Use the two procedures described in Section 4.3.3 “Performing the Upgrade” on page 4-11.

4.3.1 Planning the Upgrade

If you are working with greater than two-node clusters, consider logical host availability when planning your upgrade. Depending on the cluster configuration, it might not be possible for all logical hosts to remain available during the upgrade process. The following configuration examples illustrate upgrade strategies that minimize downtime of logical hosts.

Two Ring (Cascade) Configuration

Table 4–1 shows a four-node cluster with four logical hosts defined. The table shows which physical nodes can master each of the four logical hosts.

To upgrade this configuration, you can remove nodes 1 and 3 from the cluster and upgrade them without losing access to any logical hosts. After you upgrade nodes 1 and 3 there will be a brief service outage while you take down nodes 2 and 4 and bring up nodes 1 and 3. Nodes 1 and 3 can then provide access to all logical hosts while nodes 2 and 4 are upgraded.

TABLE 4–1 Four Nodes with Four Logical Hosts

	Logical Host 1	**Logical Host 2**	**Logical Host 3**	**Logical Host 4**
Node 1	X	X		
Node 2		X	X	

TABLE 4–1 Four Nodes with Four Logical Hosts *(continued)*

	Logical Host 1	Logical Host 2	Logical Host 3	Logical Host 4
Node 3			X	X
Node 4	X			X

N+1 Configuration

In an N+1 configuration, one node is the backup for all other nodes in the cluster. Table 4–2 shows the logical host distribution for a four-node N+1 configuration with three logical hosts. In this configuration, upgrade node 4 first. After you upgrade node 4, it can provide all services while nodes 1, 2, and 3 are upgraded.

TABLE 4–2 Three Nodes with Three Logical Hosts

	Logical Host 1	Logical Host 2	Logical Host 3
Node 1	X		
Node 2		X	
Node 3			X
Node 4	X	X	X

4.3.2 Using Terminal Concentrator and System Service Processor Monitoring

Sun Cluster 2.2 monitors the Terminal Concentrator (TC), or the System Service Processor (SSP) on E10000 machines, on clusters with greater than two nodes. You can use this feature if you are upgrading from Sun Cluster 2.0 to Sun Cluster 2.2. To enable it, you will need to provide the following information to the `scinstall(1M)` command during the upgrade procedure:

- TC or SSP name(s).
- IP addresses for the TC or SSP.

- The root password for the TC or the user `ssp` password for the SSP.

Caution - The TC and SSP passwords are required for failure fencing to work correctly in the cluster. Failure to correctly set the TC or SSP password might cause unpredictable results in the event of a node failure.

- The physical port numbers to which each server is connected. This is the port number to which the serial line is connected, not the `telnet(1)` port number used in the `/etc/serialports` file. For example, if `/etc/serialports` defines a port connection as 5006, the physical port number is 6.
- The server architecture type. For each cluster node this is either “E10000” or “other.”

4.3.3 Performing the Upgrade

Use the two procedures in this section to perform the upgrade. You should also have available for reference the *Sun StorEdge Volume Manager Installaton Guide.*

Note - If you want to use the Cluster Monitor to continue monitoring the cluster during the upgrade, then upgrade the server software first and the client software last. That is, first perform the procedure “How to Upgrade the Server Software From Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2” on page 4–16 and then perform the procedure “How to Upgrade the Client Software From Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2” on page 4–12.

Caution - Before starting the upgrade, you should have an adequate backup of all configuration information and key data, and the cluster must be in a stable, non-degraded state.

Caution - If you customized `hasap_start_all_instances` or `hasap_stop_all_instances` scripts in Solstice HA 1.3 or Sun Cluster 2.1, save them to a safe location before beginning the upgrade to Sun Cluster 2.2. Restore the scripts after completing the upgrade. Do this to prevent loss of your customizations when Sun Cluster 2.2 removes the old scripts.

Caution - The configuration parameters implemented in Sun Cluster 2.2 are different from those implemented in Solstice HA 1.3 and Sun Cluster 2.1. Therefore, after upgrading to Sun Cluster 2.2, you will have to re-configure Sun Cluster HA for SAP by running the `hadsconfig(1M)` command. Before starting the upgrade, view the existing configuration and note the current configuration variables. For Solstice HA 1.3, use the `hainetconfig(1M)` command to view the configuration. For Sun Cluster 2.1, use the `hadsconfig(1M)` command to view the configuration. After upgrading to Sun Cluster 2.2, use the `hadsconfig(1M)` command to re-create the instance.

▼ How to Upgrade the Client Software From Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2

Upgrading the client software involves removing old client software packages and replacing them with new client software packages, on the administrative workstation.

Upgrading the client software includes:

- Upgrading the Solaris operating environment on the administrative workstation
- Removing the Sun Cluster 2.0 or 2.1 client packages from the administrative workstation
- Installing the Sun Cluster 2.2 client packages on the administrative workstation
- Changing the SNMP port number, if you will be using Sun Cluster SNMP

These are the detailed steps to upgrade the client software from Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2. This procedure assumes you are using an administrative workstation.

1. **Upgrade the Solaris operating environment on the administrative workstation to Solaris 2.6 or 7.**

   For details, see the *Solaris Advanced System Administration Guide* and Chapter 2.

   a. **Partition your local disk according to Sun Cluster guidelines.**

      See Chapter 2, for details.

   b. **Install Solaris patches.**

      Check the patch database or contact your local service provider for any hardware or software patches required to run the Solaris operating environment, Sun Cluster 2.2, or your volume management software.

      Install any required patches by following the instructions in the `README` file accompanying each patch.

   c. **Reboot the administrative workstation.**

2. **Load the Sun Cluster 2.2 CD-ROM on the administrative workstation.**

3. **Use the Sun Cluster 2.0 or 2.1 version of** `scinstall(1M)` **to remove the 2.0 or 2.1 client software packages from the administrative workstation.**

   As root, invoke the Sun Cluster 2.0 or 2.1 version of `scinstall(1M)`. Select "remove" from the `scinstall(1M)` menu and remove the Sun Cluster 2.0 or 2.1 client packages. The screens may vary, depending on your software version.

```
# /opt/SUNWcluster/bin/scinstall

 ================================================
  Sun Cluster package manager
  Version: 2.1,rev 1.9

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

 ================= Package Set Selection ================

 The Sun Cluster software packages can be selected in sets,
 depending on the current state of installation

 Choose the appropriate set from the choices below:
    all          All the client and server packages in this machine
    client         All the admin tools needed on an admin workstation
    server        All the Sun Cluster packages needed on a server
 ...
 Select: [all client server] [all]: client
...

 =============== Sun Cluster Installation Manager ==================

 Current package set: client packages
 ...
 Choices:
    choose     Select the package set to manipulate
    install    Install the selected package sets
    remove     Remove the selected package sets
    obsolete   Remove obsolete packages
    verify     Sanity check the current installation
 ...
 Command: [choose install remove obsolete verify] [install]: remove
...
 Mode [manual automatic] [manual]: automatic
```

4. **Exit the Sun Cluster 2.0 or 2.1 version of** `scinstall(1M)`**.**

5. **Use the Sun Cluster 2.2 version of** `scinstall(1M)` **to install the Sun Cluster 2.2 client software packages on the administrative workstation.**

   a. **From the** `scinstall(1M)` **main menu, select the client package set:**

```
# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
# ./scinstall

 ==== Install/Upgrade Software Selection Menu ====================
 Upgrade to the latest Sun Cluster Server packages or select package
 sets for installation. The list of package sets depends on the Sun
 Cluster packages that are currently installed.

 Choose one:
 1) Upgrade         Upgrade to Sun Cluster 2.2 Server packages
 2) Server         Install the Sun Cluster packages needed on a server
 3) Client         Install the admin tools needed on an admin workstation
 4) Server and Client          Install both Client and Server packages

 5) Close         Exit this Menu
 6) Quit         Quit the Program

 Enter the number of the package set [6]:  3
```

b. Choose an install path for the client packages.

Normally the default location is acceptable:

```
What is the path to the CD-ROM image [/cdrom/cdrom0]:
```

c. Install the client packages.

Note that your packages might differ from those shown in the example.

```
Installing Client packages

 Installing the following packages: SUNWscch SUNWccon SUNWccp
 SUNWcsnmp SUNWscsdb

             >>>> Warning <<<<
   The installation process will run several scripts as root.  In
   addition, it may install setUID programs.  If you choose automatic
   mode, the installation of the chosen packages will proceed without
   any user interaction.  If you wish to manually control the install
   process you must choose the manual installation option.

 Choices:
  manual      Interactively install each package
  automatic      Install the selected packages with no user interaction.

 In addition, the following commands are supported:
    list      Show a list of the packages to be installed
    help      Show this command summary
    close      Return to previous menu
```

(continued)

```
    quit      Quit the program

  Install mode [manual automatic] [automatic]:  automatic
```

The `scinstall(1M)` command now installs the client packages. After the packages have been installed, the main `scinstall(1M)` menu is displayed.

6. Verify the client installation and then quit `scinstall(1M)`.

From the main menu, you can choose to verify the installation. Then quit to exit `scinstall(1M)`.

```
============ Main Menu =================

 1) Install/
Upgrade - Install or Upgrade Server Packages or Install Client Packages.
 2) Remove  - Remove Server or Client Packages.
 3) Verify  - Verify installed package sets.
 4) List    - List installed package sets.

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

 Please choose one of the menu items: [6]:  3

 ==== Verify Package Installation ==========================
 Installation
  All  of the install          packages have been installed
 Framework
  All  of the client          packages have been installed
  None of the server          packages have been installed
 Communications
  None of the SMA          packages have been installed
 ...
```

7. If you will be using Sun Cluster SNMP, change the port number used by the Sun Cluster SNMP daemon and Solaris SNMP (`smond`).

The default port used by Sun Cluster SNMP is the same as the default port number used by Solaris SNMP; both use port 161. Change the Sun Cluster SNMP port number using the procedure described in the appendix on Sun Cluster

SNMP management solutions in the *Sun Cluster 2.2 System Administration Guide*. You must stop and restart both the `snmpd` and `smond` daemons after changing the port number.

▼ How to Upgrade the Server Software From Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2

This procedure describes the steps required to upgrade the server software on a Sun Cluster 2.0 or Sun Cluster 2.1 system to Sun Cluster 2.2, with a minimum of downtime. You should become familiar with the entire procedure before starting the upgrade.

Upgrading the server software includes:

- Stopping Sun Cluster 2.0 or 2.1 on the local host (the host to be upgraded first)
- Unencapsulating the root disk if it is encapsulated
- (Optional) Upgrading your operating environment to Solaris 2.6 or Solaris 7
- (Optional) Upgrading your volume manager
- Re-encapsulating the root disk, if it was encapsulated previously
- Rebooting
- Using the `scinstall(1M)` command to update the software packages and cluster configuration to Sun Cluster 2.2 on the local host
- Installing required patches on the local host
- Rebooting the local host and verifying the installation
- Stopping Sun Cluster on the remote host
- Starting Sun Cluster 2.2 on the host that will master the disks and data services on the local host
- Repeating the upgrade steps on the remote host and verifying the upgrade

---

Note - During the `scinstall(1M)` upgrade procedure, all non-local private link IP addresses will be added, with root access only, to the `/.rhosts` file on every cluster node.

These are the detailed steps to upgrade the server software from Sun Cluster 2.0 or 2.1 to Sun Cluster 2.2. This example assumes an N+1 configuration using SSVM.

1. Stop the first node.

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

2. **If you are upgrading the operating environment and/or SSVM or CVM, run the command** `upgrade_start` **from the SSVM or CVM media.**

   In this example, *CDROM_path* is the path to the tools on the SSVM CD.

```
phys-hahost1# CDROM_path/Tools/scripts/upgrade_start
```

   To upgrade the operating environment, follow the detailed instructions in the appropriate Solaris installation manual and see also Chapter 2.

   If you are upgrading to Solaris 7, you must use SSVM 3.*x*. Refer to the *Sun StorEdge Volume Manager Installaton Guide* for details.

   To upgrade CVM, refer to the *Sun Cluster 2.2 Cluster Volume Manager Guide*.

3. **If you are upgrading the operating environment but not the volume manager, perform the following steps**:

   a. **Remove the volume manager package.**

      Normally, the package name is `SUNWvxvm` for both SSVM and CVM. For example:

```
phys-hahost1# pkgrm SUNWvxvm
```

   b. **Upgrade the operating system.**

      Refer to the Solaris installation documentation for instructions.

   c. **Modify the** `/etc/nsswitch.conf` **file.**

      Ensure that “service,” “group,” and “hosts” lookups are directed to files first. For example:

```
hosts: files nisplus  services: files nisplus
group: files nisplus
```

   d. **Restore the volume manager removed in Step 3 on page 4–17 from the Sun Cluster 2.2 CD-ROM.**

```
phys-hahost1# pkgadd -d CDROM_path/SUNWvxvm
```

4. **If you upgraded SSVM or CVM, run the command** `upgrade_finish` **from the SSVM or CVM media.**

   In this example, *CDROM_path* is the path to the tools on the SSVM CD.

```
phys-hahost1# CDROM_path/Tools/scripts/upgrade_finish
```

5. **Reboot the system.**

6. **Update the cluster software by using the** `scinstall(1M)` **command from the Sun Cluster 2.2 CD-ROM.**

   Invoke the `scinstall(1M)` command and select the `Upgrade` option from the menu presented:

```
phys-hahost1# cd /cdrom/suncluster_sc_2_2/Sun_Cluster_2_2/Sol_2.x/Tools
phys-hahost1# ./scinstall

 Removal of <SUNWscins> was successful.
 Installing: SUNWscins

 Installation of <SUNWscins> was successful.
     Assuming a default cluster name of sc-cluster

 Checking on installed package state.....

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

 1) Install/
Upgrade - Install or Upgrade Server Packages or Install Client Packages.
 2) Remove  - Remove Server or Client Packages.
 3) Change  - Modify cluster or data service configuration
 4) Verify  - Verify installed package sets.
 5) List    - List installed package sets.

 6) Quit    - Quit this program.
 7) Help    - The help screen for this menu.

 Please choose one of the menu items: [7]:  1
...
 ==== Install/Upgrade Software Selection Menu ======================
 Upgrade to the latest Sun Cluster Server packages or select package
 sets for installation. The list of package sets depends on the Sun
 Cluster packages that are currently installed.

 Choose one:
 1) Upgrade            Upgrade to Sun Cluster 2.2 Server packages
 2) Server             Install the Sun Cluster packages needed on a server
 3) Client             Install the admin tools needed on an admin workstation
 4) Server and Client  Install both Client and Server packages

 5) Close              Exit this Menu
```

(continued)

(Continuation)

```
6) Quit               Quit the Program

Enter the number of the package set [6]:  1

What is the path to the CD-ROM image? [/cdrom/cdrom0]:  .

** Upgrading from Sun Cluster 2.1 **
 Removing "SUNWccm" ... done
...
```

7. If the cluster has more than two nodes and you are upgrading from Sun Cluster 2.0, supply the TC/SSP information.

The first time the `scinstall(1M)` command is invoked, the TC/SSP information is automatically saved to a file, `/var/tmp/tc_ssp_info`. Copy this file to the `/var/tmp` directory on all other cluster nodes so the information can be reused when you upgrade those nodes. You can either supply the TC/SSP information now, or do so later by using the `scconf(1M)` command. See the `scconf(1M)` man page for details.

```
SC2.2 uses the terminal concentrator (or system service processor in the
 case of an E10000) for failure fencing. During the SC2.2 installation the
 IP address for the terminal concentrator along with the physical port numbers
 that each server is connected to is requested. This information can be changed
 using scconf.

 After the upgrade has completed you need to run scconf to specify terminal
 concentrator information for each server. This will need to be done on each
 server in the cluster.

 The specific commands that need to be run are:

 scconf clustername -t <nts name> -i <nts name|IP address>
 scconf clustername -H <node 0> -p <serial port for node 0> \
         -d <other|E10000> -t <nts name>

 Repeat the second command for each node in the cluster. Repeat the first
 command if you have more than one terminal concentrator in your
 configuration.

 Or you can choose to set this up now. The information you will need is:

         +terminal concentrator/system service processor names
         +the architecture type (E10000 for SSP or other for tc)
         +the ip address for the terminal concentrator/system service
```

(continued)

(Continuation)

```
          processor (these will be looked up based on the name, you
          will need to confirm)
         +for terminal concentrators, you will need the physical
          ports the systems are connected to (physical ports
          (2,3,4... not the telnet ports (5002,...)

 Do you want to set the TC/SSP info now (yes/no) [no]?  y
```

When the `scinstall(1M)` command prompts for the TC/SSP information, you can either force the program to query the `tc_ssp_info` file, or invoke an interactive session that will prompt you for the required information.

The example cluster assumes the following configuration information:

- Cluster name: sc-cluster
- Number of nodes in the cluster: 2
- Node names: phys-hahost1 and phys-hahost2
- Logical host names: hahost1 and hahost2
- Terminal concentrator name: cluster-tc
- Terminal concentrator IP address: 123.4.5.678
- Physical TC port connected to phys-hahost1: 2
- Physical TC port connected to phys-hahost2: 3

See Section 1.2.7 “Terminal Concentrator or System Service Processor and Administrative Workstation” on page 1-10, for more information on server architectures and TC/SSPs. In this example, the configuration is not an E10000 cluster, so the architecture specified is “other,” and a terminal concentrator is used:

```
What type of architecture does phys-hahost1 have? (E10000|other)
 [other] [?] other
What is the name of the Terminal Concentrator connected to the
 serial port of phys-hahost1 [NO_NAME] [?] cluster-tc
Is 123.4.5.678 the correct IP address for this Terminal
 Concentrator (yes|no) [yes] [?] yes
Which physical port on the Terminal Concentrator is phys-hahost2
 connected to [?] 2
What type of architecture does phys-hahost2 have? (E10000|other)
 [other] [?] other
Which Terminal Concentrator is phys-hahost2 connected to:

 0) cluster-tc       123.4.5.678
 1) Create A New Terminal Concentrator Entry
```

(continued)

```
 Select a device [?] 0
Which physical port on the Terminal Concentrator is phys-hahost2
 connected to [?] 3
The terminal concentrator/system service processor (TC/SSP)
 information has been stored in file /var/tmp/tc_ssp_data. Please
 put a copy of this file into /var/tmp on the rest of the nodes in
 the cluster. This way you don't have to re-enter the TC/SSP values,
 but you will, however, still be prompted for the TC/SSP passwords.
```

8. If you will be using Sun Cluster SNMP, change the port number used by the Sun Cluster SNMP daemon and Solaris SNMP (`smond`).

The default port used by Sun Cluster SNMP is the same as the default port number used by Solaris SNMP; both use port 161. Change the Sun Cluster SNMP port number using the procedure described in the appendix on Sun Cluster SNMP management soutions in the *Sun Cluster 2.2 System Administration Guide.*

9. Reboot the system.

Caution - You must reboot at this time.

10. If your cluster is greater than two nodes and you are using a shared CCD, put all logical hosts into maintenance mode.

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

Note - Greater than two-node clusters do not use a shared CCD. Therefore, for greater than two-node clusters, you do not need to put the data services into maintenance mode before beginning the upgrade.

11. If your configuration includes Oracle Parallel Server (OPS), make sure OPS is halted.

Refer to your OPS documentation for instructions on halting OPS.

12. Stop the cluster software on the remaining nodes running the old version of Sun Cluster.

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

13. Start the upgraded node.

```
phys-hahost1# scadmin startcluster phys-hahost1 sc-cluster
```

Note - As the upgraded node joins the cluster, the system might report several warning messages stating that communication with the terminal concentrator is invalid. At this point these messages are expected and can be safely ignored.

---

14. If you are using a shared CCD and if you upgraded from Sun Cluster 2.0, update the shared CCD now.

Run the `ccdadm(1M)` command only once, on the host that joined the cluster first:

```
phys-hahost1# cd /etc/opt/SUNWcluster/conf
phys-hahost1# ccdadm sc-cluster -r ccd.database_post_sc2.0_upgrade
```

15. If you stopped the data services previously, restart them on the upgraded node.

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

16. Upgrade the remaining nodes.

Repeat Step 2 on page @–17 through Step 9 on page @–21 on the remaining Sun Cluster 2.0 or Sun Cluster 2.1 nodes.

17. After each node is upgraded, add it to the cluster:

```
phys-hahost2# scadmin startnode sc-cluster
```

18. Set up and start Sun Cluster Manager.

Sun Cluster Manager is used to monitor the cluster. For instructions, see the section on monitoring Sun Cluster servers with Sun Cluster Manager in Chapter 2 of the *Sun Cluster 2.2 System Administration Guide*.

This completes the upgrade to Sun Cluster 2.2.

CHAPTER 5

Setting Up and Administering Sun Cluster HA for Oracle

This chapter provides instructions for setting up and administering the Sun Cluster HA for Oracle data service on your Sun Cluster nodes.

This chapter includes the following procedures:

5.1 Preparing to Install Sun Cluster HA for Oracle

Use the following sections to prepare Sun Cluster nodes for Sun Cluster HA for Oracle installation.

5.1.1 Selecting an Install Location for Sun Cluster HA for Oracle

You can install the Oracle binaries either on the local disks of the physical hosts or on the multihost disks. Both locations have advantages and disadvantages. Consider the following points when selecting an install location.

Placing Oracle binaries on the multihost disk eases administration, since there is only one copy to administer. It ensures high availability of the Oracle binaries or server during a cluster reconfiguration. However, it sacrifices redundancy and therefore availability in case of some failures. Note that in cases of switchover or accidental removal of the Oracle binaries from the multihost disk, the data service will be unavailable. Any binaries installed on the multihost disk will be mirrored as part of a disk group, so you must allocate space accordingly.

Alternatively, placing Oracle binaries on the local disk of the physical host increases redundancy and therefore availability in case of failure or accidental removal of one copy, and if a switchover occurs, Oracle will be able to run on the one node. However, placing the Oracle binaries on the local disk increases the administrative overhead, since you must manage multiple copies of the files.

5.1.2 Setting Up the `/etc/nsswitch.conf` File

On each node that can master the logical host running Sun Cluster HA for Oracle, modify the `/etc/nsswitch.conf` file so that “group” lookups are directed to files first. For example:

```
...
 group: files nisplus
 ...
```

Sun Cluster HA for Oracle uses the `su` *user* command when starting and stopping the database server.

Adding these settings will ensure that the `su` *user* command does not refer to NIS or NIS+ when the network information name service is not available due to a failure of the public network on the cluster node.

5.1.3 Setting Up Multihost Disks for Sun Cluster HA for Oracle

If you are using Solstice DiskSuite, you can configure Sun Cluster HA for Oracle to use UFS logging or raw mirrored metadevices. Refer to Appendix B, for details about setting up metadevices.

If you are using Sun StorEdge Volume Manager, you can configure Sun Cluster HA for Oracle to use VxFS logging or raw devices. Refer to Appendix C, for information about configuring VxFS file systems.

5.2 Installing Sun Cluster HA for Oracle

Use the procedures in this section to prepare the Sun Cluster nodes, to install the Oracle software, to create Oracle databases, and to set up Sun Cluster HA for Oracle.

Before setting up Sun Cluster HA for Oracle, you must have configured the Sun Cluster software on each node by using the procedures described in Chapter 3.

▼ How to Prepare the Nodes and Install the Oracle Software

These are the high-level steps to prepare Sun Cluster nodes for Oracle installation and install the Oracle software:

- Choosing a location for the `$ORACLE_HOME` directory
- Creating `/etc/group` and `/etc/passwd` entries for user `oracle`, and running the `pwconv(1M)` command to create an entry in the `/etc/shadow` file
- Installing the Oracle software
- Creating the `/var/opt/oracle/oratab` file
- Verifying the installation

Caution - Perform all steps described in this section on all Sun Cluster nodes.

Consult your Oracle documentation before performing this procedure.

These are the detailed steps to prepare Sun Cluster nodes and install the Oracle software:

1. **Prepare the environment for Oracle installation.**

   Choose a location for the `$ORACLE_HOME` directory, on either a local or multihost disk.

Note - If you choose to install the Oracle binaries on a local disk of a physical host, then mount the Oracle software distribution as a file system on its own separate disk, if possible. This will prevent Oracle binaries from being overwritten if the operating environment is reinstalled.

2. **On each node, create an entry for the database administrator group in the** `/etc/group` **file, and add potential users to the group.**

   This group normally is named `dba`. Verify that `root` and `oracle` are members of the `dba` group, and add entries as necessary for other `dba` users. Make sure that the group IDs are the same on all nodes running Sun Cluster HA for Oracle. For example:

```
dba:*:520:root,oracle
```

   While you can make the name service entries in a network name service (for example, NIS or NIS+) so that the information is available to Sun Cluster HA for Oracle clients, you also should make entries in the local `/etc` files to eliminate dependency on the network name service.

Note - This information must be replicated on each node.

3. **On each node, create an entry for the Oracle user ID (*oracle_id*) in the** `/etc/passwd` **file, and run the** `pwconv(1M)` **command to create an entry in the** `/etc/shadow` **file.**

   This *oracle_id* is normally `oracle`. For example:

```
# useradd -u 120 -g dba -d /oracle oracle
```

   Make sure that the user IDs are the same on all nodes running Sun Cluster HA for Oracle.

4. **Verify that the** `$ORACLE_HOME` **directory is owned by *oracle_id* and is included in the** `dba` **group.**

```
# chown oracle $ORACLE_HOME
# chgrp dba $ORACLE_HOME
```

If $ORACLE_HOME is a symbolic link to the Oracle home directory, then use the following command:

```
# chown oracle $ORACLE_HOME
# chgrp dba $ORACLE_HOME
# chown -h oracle $ORACLE_HOME
# chgrp -h dba $ORACLE_HOME
```

5. Note the requirements for Oracle installation.

Oracle binaries can be installed on either the local disks of the physical hosts, or on the multihost disks. See Section 5.1.1 “Selecting an Install Location for Sun Cluster HA for Oracle” on page 5-2, for more information.

If you plan to install Oracle software on the multihost disks, you first must start Sun Cluster and take ownership of the logical host. See Section 5.2.1 “Creating an Oracle Database and Setting Up Sun Cluster HA for Oracle” on page 5-6, for details.

When first installing Oracle, select the `Install/Upgrade/Patch Software Only` option. This is necessary because database initialization and configuration files must be modified to reflect the logical hosts as the location for the database.

6. Install the Oracle software.

On each node, modify the `/etc/system` files according to standard Oracle installation procedures. Also, on each node, create a `/var/opt/oracle` directory for user `oracle` and group `dba`.

Log in as `oracle` to ensure ownership of the entire directory before performing this step. For complete instructions on installing Oracle software, refer to the ORACLE7 Installation and Configuration Guide and the Oracle7 for Sun SPARC Solaris 2.x Installation and Configuration Guide.

7. Create the `/var/opt/oracle/oratab` file.

As root, run the script `$ORACLE_HOME/orainst/oratab.sh` to create the `/var/opt/oracle/oratab` file with the appropriate permissions.

8. Verify the Oracle installation.

a. Verify that the Oracle kernel, `$ORACLE_HOME/bin/oracle`, is owned by `oracle` and is included in the `dba` group.

b. Verify that the `$ORACLE_HOME/bin/oracle` permissions are set as follows:

```
-rwsr-x--x
```

c. Verify that the listener binaries exist in `$ORACLE_HOME/bin`**.**

d. Verify that the `$ORACLE_HOME/orainst/RELVER` **file exists.**
`$ORACLE_HOME/orainst/RELVER` is created when Oracle Unix Installer is installed on the node. If the `$ORACLE_HOME/orainst/RELVER` file does not exist, then create it manually. Include the correct version number of the Oracle software installed on the node. For example:

```
# cat $ORACLE_HOME/orainst/RELVER
RELEASE_VERSION=8.0.4.0.0
```

5.2.1 Creating an Oracle Database and Setting Up Sun Cluster HA for Oracle

Complete both procedures in this section to create and configure the initial Oracle database in a Sun Cluster configuration. If you are creating and configuring additional databases, perform only the procedure described in "How to Create an Oracle Database" on page 5–7.

▼ How to Prepare Logical Hosts for Oracle Databases

1. **Make sure Sun Cluster is started and the node owns the disk groups.**
   If necessary, as `root`, use the `scadmin(1M)` command.

   ```
   # scadmin startcluster
   ```

   This command causes the node to take disk group ownership.

2. **Configure the disk devices for use by your volume manager.**

Caution - While Oracle supports raw I/O to both raw physical devices and raw metadevices (mirrored or nonmirrored), Sun Cluster only supports raw Oracle I/O on raw mirrored volumes or metadevices. You cannot use devices such as `/dev/rdsk/c1t1d1s2` to contain Oracle data under Sun Cluster.

a. If you are using Solstice DiskSuite, set up UFS logs or raw mirrored metadevices on all nodes that will be running Sun Cluster HA for Oracle.

If you will be using raw mirrored metadevices to contain the databases, change the owner, group, and mode of each of the raw mirrored metadevices. If you are not using raw mirrored metadevices, skip this step. Instructions for creating mirrored metadevices are provided in Appendix B.

If you are creating raw mirrored metadevices, type the following commands for each metadevice:

```
# chown oracle_id /dev/md/disk_group/rdsk/dn
 # chgrp dba_id /dev/md/disk_group/rdsk/dn
 # chmod 600 /dev/md/disk_group/rdsk/dn
```

b. If you are using Sun StorEdge Volume Manager, set up VxFS logs or raw devices on all nodes.

If you will be using raw devices to contain the databases, change the owner, group, and mode of each device. If you are not using raw devices, skip this step. Refer to your Sun StorEdge Volume Manager documentation for information on setting up VxFS logs.

If you are creating raw devices, type the following command for each raw device:

```
# vxvol set owner=oracle_id group=dba_id mode=600 \
 /dev/vx/rdsk/diskgroup_name/volume_name
```

▼ How to Create an Oracle Database

These are the high-level steps to create an Oracle database:

- Preparing the database configuration files
- Creating the database
- Creating the `v$sysstat` view

These are the detailed steps to create an Oracle database.

1. Prepare database configuration files.

Place all parameter files, data files, `redolog` files, and control files on the logical host, that is, the disk group's multihost disks.

Within the `init$ORACLE_SID.ora` or `config$ORACLE_SID.ora` file, you might need to modify the assignments for `control_files` and `background_dump_dest` to specify the location of control files and alert files on the logical host.

Note - If you are using Solaris authentication for database logins, the `remote_os_authent` variable in the `init$ORACLE_SID.ora` file must be set to `TRUE`.

Full path names must be provided for `background_dump_dest`. The special character `?` for specifying `ORACLE_HOME` should not be used.

2. **Create the database.**

   During creation, ensure that all configuration and database files are placed on the logical hosts.

   a. **Start the Oracle installer (`orainst`) and select the `Create New Database Objects` option.**

   b. **During the `orainst` session, place all the database files on the logical hosts.**

      Override the default file locations provided by the Oracle installer.

   c. **Verify that the file names of your control files match the file names in your configuration files.**

      Alternatively, you can create the database using the Oracle `svrmgrl` command, depending on your Oracle version.

3. **Create the `v$sysstat` view.**

   Run the catalog scripts that create the `v$sysstat` view. This view is used by the Sun Cluster fault monitoring scripts.

▼ How to Set Up Sun Cluster HA for Oracle

These are the high-level steps to set up Sun Cluster HA for Oracle:

- Making entries for all of the database instances in the `/var/opt/oracle/oratab` files on all nodes running Sun Cluster HA for Oracle

- Enabling user and password for fault monitoring and, optionally, granting permission for the database to use Solaris authentication

- Configuring SQL*Net to monitor Sun Cluster HA for Oracle
- Verifying installation of the Oracle listener, Sun Cluster software, and the cluster daemon
- Activating the Oracle data service by using the `hareg(1M)` command
- Bringing the Oracle database instance into service

These are the detailed steps to set up Sun Cluster HA for Oracle.

1. **Make `SID` entries for the Sun Cluster HA for Oracle databases of all database instances.**

   You must include the `SID` of the instance associated with your database in the `/var/opt/oracle/oratab` file on all nodes running Sun Cluster HA for Oracle. You must keep this file current on all nodes running Sun Cluster HA for Oracle for a failover to succeed. Update the file manually as are you add or remove a `SID`. If the `oratab` files do not match, an error message will be returned and the `haoracle start` command will fail.

   All entries in the `/var/opt/oracle/oratab` file should have the `–:N` option specified to ensure that the instance will not start automatically on Solaris reboot. For example:

```
oracle_sid:/oracle:N
```

2. **Depending on which authentication method you choose, Oracle authentication or Solaris authentication, perform one of the following steps.**

   a. **Enable access for the user and password to be used for fault monitoring.**

      You must complete this step if you do not enable Solaris authentication, as described in Step 2 on page 5–10.

      In the following examples, the user is `scott` and the password is `tiger`. Note that the user and password pair must agree with those used in Step 6 on page @–13, if you are using Oracle authentication.

      For all supported Oracle releases, enable access by typing the following script into the screen brought up by the `srvmgrl(1M)` command.

```
# svrmgrl

  connect internal;
    grant connect, resource to scott identified by tiger;
    alter user scott default tablespace system quota 1m on
     system;
           grant select on v_$sysstat to scott;
    grant create session to scott;
    grant create table to scott;
  disconnect;
```

(continued)

(Continuation)

```
    exit;
```

b. Grant permission for the database to use Solaris authentication.

You must complete this step if you chose not to complete Step 2 on page @–9. The following sample entry enables Solaris authentication.

```
# svrmgrl

  connect internal;
    create user ops$root identified by externally
     default tablespace system quota 1m on system;
    grant connect, resource to ops$root;
          grant select on v_$sysstat to ops$root;
    grant create session to ops$root;
    grant create table to ops$root;
  disconnect;

    exit;
```

3. Configure SQL*Net V2 for Sun Cluster.

Note - Create and update the `tnsnames.ora` and `listener.ora` files under `/var/opt/oracle` on all nodes.

Sun Cluster HA for Oracle imposes no restrictions on the listener name—it can be any valid Oracle listener name.

The following code sample identifies the lines in `listener.ora` that are updated.

```
LISTENER =
  (ADDRESS_LIST =
    (ADDRESS =
     (PROTOCOL = TCP)
      (HOST = hostname) <- or, use logical host name
     (PORT = 1527)
    )
  )
 .
 .
 SID_LIST_LISTENER =
  .
    .
       (SID_NAME = SID) <- Database name, default is ORCL
```

The following sample identifies the lines in `tnsnames.ora` that are updated.

```
service_name =
  .
    .
       (ADDRESS = <- listener address
         (HOST = server_name) <- logical host name
         (PORT = 1527) <- must match port in LISTENER.ORA
       )
     )
     (CONNECT_DATA =
       (SID = <SID>)) <---database name, default is ORCL
```

Sun Cluster HA for Oracle opens the `/var/opt/tnsnames.ora` file. It scans the file to *service name* by matching SID = *instance name* and *server name*=*logical host*. The *service_name* obtained from `tnsnames.ora` is used by the Sun Cluster HA for Oracle remote fault monitor to connect to the server.

Put the logical host name in place of the host name in the `listener.ora` file in the "`HOST =` *host name*" line.

The following example shows how to update the `listener.ora` and `tnsnames.ora` files given the following Oracle instances.

TABLE 5–1 Example Oracle Configuration

Instance	Logical Host	Listener
ora8	hadbms3	LISTENER-ora8
ora7	hadbms4	LISTENER-ora7

The corresponding `listener.ora` entries would be:

```
LISTENER-ora7 =
  (ADDRESS_LIST =
    (ADDRESS =
     (PROTOCOL = TCP)
     (HOST = hadbms4)
     (PORT = 1530)
    )
   )
 SID_LIST_LISTENER-ora7 =
  (SID_LIST =
    (SID_DESC =
     (SID_NAME = ora7)
    )
   )
 LISTENER-ora8 =
   (ADDRESS_LIST =
     (ADDRESS= (PROTOCOL=TCP) (HOST=hadbms3)(PORT=1806))
   )
 SID_LIST_LISTENER-ora8 =
   (SID_LIST =
      (SID_DESC =
    (SID_NAME = ora8)
    )
   )
```

The corresponding `tnsnames.ora` entries would be:

```
ora8 =
 (DESCRIPTION =
    (ADDRESS_LIST =
    (ADDRESS = (PROTOCOL = TCP)
    (HOST = hadbms3)
    (PORT = 1806))
     )
  (CONNECT_DATA = (SID = ora8))
 )
```

(continued)

(Continuation)

```
ora7 =
(DESCRIPTION =
  (ADDRESS_LIST =
        (ADDRESS =
    (PROTOCOL = TCP)
    (HOST = hadbms4)
    (PORT = 1530))
   )
   (CONNECT_DATA = (
   SID = ora7))
)
```

4. **Verify that Sun Cluster and the cluster daemon are installed and running on all hosts.**

```
# hastat
```

If they are not running already, start them by using the `scadmin startnode` command.

5. **Register and activate Sun Cluster HA for Oracle by using the `hareg(1M)` command.**

Run the `hareg(1M)` command on only one node.

If the Oracle server is not yet registered, use the `hareg(1M)` command to register it. To register the data service only on the logical host, use the `–h` option and supply the logical host name:

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

If the cluster is running already, use the `hareg(1M)` command to activate the Oracle data service:

```
# hareg -y oracle
```

6. **Setup Sun Cluster HA for Oracle configuration data.**

Run the following command so that the instance will be monitored by Sun Cluster:

```
# haoracle insert $ORACLE_SID logicalhost 60 10 120 300 \
  user/password /logicalhost/.../init$ORACLE_SID.ora listener
```

The previous command line includes the following:

- `haoracle insert` – Command and subcommand
- `$ORACLE_SID` – Name of the Oracle database instance
- *logicalhost* – Logical host serving `$ORACLE_SID` (not the physical host)
- `60 10 120 300` – These parameters specify a probe cycle time of 60 seconds, a connectivity probe cycle count of 10, a probe time out of 120 seconds, and a restart delay of 300 seconds.
- *user/password* – These are the user and password to be used for fault monitoring. They must agree with the permission levels granted in Step 2 on page @–9. To use Solaris authentication, enter a slash (/) instead of the user name and password.
- */logicalhost/.../init$ORACLE_SID.ora* – This indicates the `pfile` to use to start up the database. This must be on a logical host's disk group.
- *listener* – The SQL*Net V2 listener. The listener is started and monitored using this name. The default is `LISTENER`. This field is optional.

See the `haoracle(1M)` man page for details on all options to `haoracle(1M)`.

7. Bring the database instance into service.

Bring the Sun Cluster HA for Oracle database into service by executing the `haoracle(1M)` command. Monitoring for that instance will start automatically.

```
# haoracle start $ORACLE_SID
```

Note - If you did not start the Oracle instance before issuing the `haoracle(1M)` command, Sun Cluster will start the Oracle instance for you when you issue the command.

5.2.2 Setting Up Sun Cluster HA for Oracle Clients

Clients must always refer to the database by using the logical host name and not the physical host name.

For example, in the `tnsnames.ora` file for SQL*Net V2, you must specify the logical host as the host on which the database instance is running. See "How to Set Up Sun Cluster HA for Oracle" on page 5–8.

Note - Oracle client-server connections will not survive a Sun Cluster HA for Oracle switchover. The client application must be prepared to handle disconnection and reconnection or recovery as appropriate. A transaction monitor might simplify the application. Further, Sun Cluster HA for Oracle node recovery time is application-dependent.

If your application uses functions from RDBMS dynamic link libraries, you must ensure that these libraries are available in the event of failover. To do so:

- Install the link libraries on the client, or
- If Sun Cluster HA for NFS service is to be provided by the cluster, copy the libraries to the logical host and set the environment variables to the directory path of the link libraries, then NFS share those directories to clients.

5.3 Verifying the Sun Cluster HA for Oracle Installation

Perform the following verification tests to ensure the Sun Cluster HA for Oracle installation was performed correctly.

The purpose of these sanity checks is to ensure that the Oracle instance can be started by all nodes running Sun Cluster HA for Oracle and can be accessed successfully by the other nodes in the configuration. Perform these sanity checks to isolate any problems starting Oracle from the Sun Cluster HA for Oracle data service.

▼ How to Verify the Sun Cluster HA for Oracle Installation

1. **Log in to the node mastering the logical host, and set the Oracle environment variables.**

   Log in as `oracle` to the node that currently masters the logical host, and set the environment variables `ORACLE_SID` and `ORACLE_HOME`.

   a. **Confirm that you can start the Oracle instance from this host.**

   b. **Confirm that you can connect to the Oracle instance.**

      Use the `sqlplus` command with the *tns_service* variable defined in the `tnsnames.ora` file:

```
# sqlplus scott/tiger@tns_service
```

c. Shut down the Oracle instance.

2. Transfer the logical host containing the Oracle database to another node in the cluster.

For example:

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

3. Log in to the node now mastering the logical host, and repeat the checks listed in Step 1.

Log in as `oracle` to the new master node and confirm interactions with the Oracle instance.

CHAPTER 6

Setting Up and Administering Sun Cluster HA for Sybase

This chapter provides instructions for setting up and administering the Sun Cluster HA for Sybase data service on your Sun Cluster nodes.

This chapter includes the following procedures:

6.1 Preparing to Install Sun Cluster HA for Sybase

Use the information in this section to prepare Sun Cluster nodes for Sun Cluster HA for Sybase installation.

6.1.1 Selecting an Install Location for Sun Cluster HA for Sybase

You can install the Sybase binaries either on a physical host's local disks or on the shared multihost disks (logical host). Use Table 6–1, which shows the advantages and disadvantages of each install location, to determine the install location that best fits your needs.

TABLE 6–1 Sybase Install Locations Comparison

Install Location	Advantages	Disadvantages
Multihost Disks	■ Eases administration - only one copy must be maintained	■ Sacrifices redundancy ■ In some cases, compromises availability ■ In cases of switchover or accidental removal of Sybase binaries, the server will be unavailable ■ Requires a mirror, so disk space must be doubled
Local Disks	■ Increases redundancy - each node is a copy	■ Increases administrative overhead - multiple copies must be maintained

If you install the Sybase server on the multihost disk, you must install the Sybase clients in `/var/opt/sybase` on all nodes. If you install the Sybase servers on local disks, you are not required to install Sybase clients.

6.1.2 Setting Up the `/etc/nsswitch.conf` File

On each node that can master the logical host running Sun Cluster HA for Sybase, modify the `/etc/nsswitch.conf` file so that "group" lookups are directed to files first. For example:

```
...
 group: files nisplus
 ...
```

Sun Cluster HA for Sybase uses the `su sybase` command when starting and stopping the database server.

Adding these settings will ensure that the `su sybase` command does not refer to NIS/NIS+ when the network information name service is not available due to a failure of the public network on the cluster node.

6.1.3 Setting Up Multihost Disks for Sun Cluster HA for Sybase

If you are using Solstice DiskSuite, you can configure Sun Cluster HA for Sybase to use UFS logging or raw mirrored metadevices. See Appendix B, for details about setting up metadevices.

If you are using Sun StorEdge Volume Manager, you can configure Sun Cluster HA for Sybase to use VxFS logging or raw devices. Refer to your Sun StorEdge Volume Manager documentation for more information.

6.2 Installing Sun Cluster HA for Sybase

Use the procedures in this section to prepare the Sun Cluster nodes, to install the Sybase software, to create Sybase servers and databases, and to set up Sun Cluster HA for Sybase.

Before setting up Sun Cluster HA for Sybase, you must have configured the Sun Cluster software on each node by using the procedures described in Chapter 3.

▼ How to Prepare the Nodes and Install the Sybase Software

These are the high-level steps to prepare Sun Cluster nodes for Sybase installation and install the Sybase software:

- Choosing a location for the `$SYBASE` directories
- Creating `/etc/group` and `/etc/passwd` file entries for Sybase
- Installing the Sybase software

Caution - Perform all steps described in this section on all Sun Cluster nodes.

Consult your Sybase documentation before performing this procedure.

These are the detailed steps to prepare Sun Cluster nodes for Sybase installation.

1. Prepare the environment for Sybase installation.

Choose a location for the $SYBASE directories, on either a local disk or multihost disk (the logical host).

Note - If you choose to install the Sybase binaries on the local disk of a physical host, then mount the Sybase software distribution as a file system on its own separate disk, if possible. This will prevent the Sybase binaries from being overwritten if the operating environment is reinstalled.

---

2. **On each node, create an entry for the database administrator group in the** `/etc/group` **file.**

   This group normally is named `dba`. Verify that `root` and `sybase` are members of the `dba` group. For example:

```
dba:*:520:root,sybase
```

   You can make the name service entries in a network name service (for example, NIS or NIS+) so that the information is available to Sun Cluster HA for Sybase clients, but you also should make entries in the local `/etc` files to eliminate dependency on the network name service. The name service entries should be identical on all nodes.

Note - This information must be replicated on each cluster node.

---

3. **On each node, create an entry for the Sybase login ID in the** `/etc/passwd` **file.**

   This login ID is normally `sybase`. The home directory for this entry should be the Sybase installation directory. For example:

```
# useradd -u 120 -g dba -d /Sybase directory sybase
```

4. **Verify that the Sybase directories are owned by** `sybase` **and are included in the** `dba` **group.**

```
# chown sybase /sybase_directory
# chgrp dba /sybase_directory
```

   If the Sybase home directory is a symbolic link, then use the following command:

```
# chown sybase /sybase_directory
 # chgrp dba /sybase_directory
# chown -h sybase /sybase_directory
# chgrp -h dba /sybase_directory
```

5. **Note the requirements for Sybase installation.**

   Sybase binaries can be installed on either the local disks of the physical host or the multihost disks (the logical host). See Section 6.1.1 “Selecting an Install Location for Sun Cluster HA for Sybase” on page 6-2, for more information.

   If you plan to install Sybase software on the multihost disks, you first must start Sun Cluster and take ownership of the logical host. See Section 6.2.1 “Creating a Sybase SQL Server and Setting Up Sun Cluster HA for Sybase” on page 6-5, for details.

6. **Install the Sybase software.**

   Log in as `sybase` to ensure ownership of the entire directory before performing this step. For complete instructions on installing Sybase software, refer to the Sybase documentation.

   If the Sybase binaries are going to be stored on the multihost disks, install the Sybase Open Client (DB-Library) under `/var/opt/sybase`.

Note - The Sun Cluster HA for Sybase fault monitor requires that the `ctlib.loc` file exist in the `$SYBASE/locales/us_english/iso_1` directory. You can obtain this file by installing a Sybase connectivity tool such as Open Client (DB-Library).

6.2.1 Creating a Sybase SQL Server and Setting Up Sun Cluster HA for Sybase

Use the procedures in this section to create and configure the initial Sybase SQL Server and databases in a Sun Cluster configuration. If you are creating and configuring additional databases, perform only the procedure described in “How to Create a Sybase SQL Server and Databases” on page 6–6.

▼ How to Prepare Multihost Disks for Sybase SQL Servers and Databases

1. **Make sure Sun Cluster is started and the node owns the disk groups.**

If necessary, as `root`, use the `scadmin(1M)` command.

```
# scadmin startcluster
```

This command causes the node to take disk group ownership.

2. Configure the disk devices for use by your volume manager.

Caution - While Sybase supports raw I/O to both raw physical devices and raw metadevices (mirrored or non-mirrored), Sun Cluster only supports raw Sybase I/O on raw mirrored volumes or metadevices. You cannot use devices such as `/dev/rdsk/c1t1d1s2` to contain Sybase data under Sun Cluster.

If you are using Solstice DiskSuite to set up raw mirrored metadevices, perform the following steps:

a. Change the owner, group, and mode of each of the raw mirrored metadevices. (If you are not using raw mirrored metadevices, skip this step.)
Instructions for creating mirrored metadevices are provided in Appendix B.

b. Type the following commands for each metadevice.

```
# chown sybase_id /dev/md/disk_group/rdsk/dn
 # chgrp dba_id /dev/md/disk_group/rdsk/dn
 # chmod 600 /dev/md/disk_group/rdsk/dn
```

If you are using SSVM, refer to Appendix C, and to your SSVM documentation for information on setting up your disk devices.

▼ How to Create a Sybase SQL Server and Databases

These are the high-level steps to create a Sybase SQL Server and databases:

- Preparing the SQL Server configuration files and setting up the SQL Server
- Preparing the Sybase environment
- Creating the database

These are the detailed steps to create a Sybase SQL Server and databases.

1. **Log in as user** `sybase`**.**

   You must be defined as user `sybase` to run the Sybase commands.

2. **Prepare the Sybase environment using the following command. In** `csh`:

```
# setenv SYBASE base_dir
```

3. **Prepare database configuration files. This can be done on either the local disk or the multihost disk.**

   If you will place the Sybase installation directory on the local disk:

   a. **Use the** `sybinit` **command to create the** `RUN_`***sqlserver*** **and (optional)** `RUN_`***backupserver*** **files in the Sybase installation directory.**

   b. **Place the Sybase installation directory on the local disk.**

   c. **Use the** `rcp(1)` **command to copy the** `RUN_` **files to all other potential masters of the logical host.**

   d. **Update the** `$SYBASE/interfaces` **file on those potential masters with entries for the new servers.**

   e. **Place all transaction logs, databases, the** ***server***`.cfg` **file, the** ***server***`.krg` **file, and the** `errorlog` **file on the local disk.**

   f. **Use the** `rcp(1)` **command to copy the files in Step 3 on page 6–7 to the other potential masters.**

   g. **If you use the** `sp_configure` **store procedure to modify configuration settings or to edit the configuration file directly, use** `rcp(1)` **to copy the** ***server***`.cfg` **file to the other potential masters.**

      If you will place the Sybase installation directory on the multihost disk (the logical host):

1. **Use the** `sybinit` **command to create the** `RUN_`***sqlserver*** **and (optional)** `RUN_`***backupserver*** **files in the Sybase installation directory.**

2. **Place the Sybase installation directory on the multihost disk.**

3. **Place all transaction logs, databases, the** ***server***`.cfg` **file, the** ***server***`.krg` **file, and the** `errorlog` **file on the multihost disk.**

4. **Use** `rcp(1)` **to copy the Sybase** `interfaces` **file to** `/var/opt/sybase/interfaces` **on all potential masters.**

5. **Set up the Sybase SQL Server using the** `sybinit` **command.**

Note - With Sun Cluster, there can be no more than one SQL Server for each backup server.

You must use the logical host name when defining the database device. Later, when installing Sybase on the other potential masters, add identical lines to the interfaces file through `sybinit`.

6. **Create the database and place it on the logical host.**

7. **Add the name of your backup server to the** `sysservers` **database.**

   If the name of your backup server is anything other than the default, `SYB_BACKUP`, then you must add it to the `sysservers` database using the following command:

```
# sp_addserver <backup_server_name>
```

   If you do not add a backup server name to the `sysservers` database, then you must use the backup server name `SYB_BACKUP`.

▼ How to Set Up Sun Cluster HA for Sybase

These are the high-level steps to set up Sun Cluster HA for Sybase:

- Making entries for all of the SQL Servers in the `/var/opt/sybase/sybtab` files on nodes running Sun Cluster HA for Sybase
- Starting the SQL Server
- Optionally, creating a Sybase `sa` login and password for fault monitoring
- Verifying the `$SYBASE/interfaces` file
- Verifying installation of the Sun Cluster software and the cluster daemon
- Starting Sun Cluster by using the `scadmin startcluster` and `scadmin startnode` commands
- Activating the Sybase data service by using the `hareg(1M)` command
- Bring Sybase under control of Sun Cluster HA for Sybase
- Bringing the Sybase SQL Server into service by using the `hasybase(1M)` command

These are the detailed steps to set up Sun Cluster HA for Sybase.

1. **Make entries for the names of all SQL Servers.**

   You must include the server names associated with your databases in the `/var/opt/sybase/sybtab` file on all nodes running Sun Cluster HA for Sybase. You must keep this file current on all nodes running Sun Cluster HA for Sybase for a failover to succeed. Update the file manually as SQL Servers are added or removed.

   Entries in the `/var/opt/sybase/sybtab` file have the following format:

   ```
   sql_server:sybase_directory
   ```

   Note - The Sun Cluster HA for Sybase fault monitor does not monitor backup servers. Therefore, do not make separate entries for backup servers in the `/var/opt/sybase/sybtab` file.

2. **Log in with the user ID of the `RUN_`*sqlserver* file and start the SQL Server.**

   If the SQL Server is not running already, start it with the following command:

   ```
   # startserver -f $SYBASE/install/RUN_sqlserver
   ```

3. **(Optional) Create a login and password to be used for fault monitoring.**

   Create a Sybase login "*new_login_name*" with `sa_role` to start and stop the server.

   Note - Skip this stepif you want to use the `sa` login for fault monitoring.

   ```
   # isql -Usa -P -S sql_server_name
    >sp_addlogin new_login_name, password
    >go
     >sp_role ''grant'', ''sa_role'', new_login_name
     >go
     >exit
   ```

4. **Verify the `$SYBASE/interfaces` file.**

If Sybase is installed on the local disks, verify that `$SYBASE/interfaces` refers to a logical host, not a physical host. If Sybase is installed on the multihost disks, verify that the `interfaces` file exists under `/var/opt/sybase` on the local disks and that it is identical to the file `$SYBASE/interfaces` on the multihost disks.

Note - If you install the Sybase SQL server on the multihost disks (logical host), you must install the Sybase clients in the `/var/opt/sybase` directories on all nodes capable of mastering the disks. If you install the Sybase SQL server on the local disks, you do not need to install Sybase clients.

---

5. **Verify that Sun Cluster and the cluster daemon are installed and running on all nodes.**

```
# hastat
```

If they are not running already, start them by using the `scadmin startnode` command.

6. **Register and activate Sun Cluster HA for Sybase using the `hareg(1M)` command.**

Run the `hareg(1M)` command on only one node.

If the Sybase data service is not yet registered, use the `hareg(1M)` command to register it. To register the data service only on the logical host, include the `–h` option and supply the logical host name:

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

If the cluster is running already, use the `hareg(1M)` command to activate the Sybase data service:

```
# hareg -y sybase
```

7. **Bring Sybase under control of Sun Cluster HA for Sybase using the following command.**

```
# hasybase insert sqlserver logicalhost 60 10 120 300 user/password \
  $SYBASE/install/RUN_sqlserver backupserver $SYBASE/install/RUN_backupserver
```

(continued)

The above command line includes the following:

- `hasybase insert` – Command and subcommand
- *sqlserver* – Name of the SQL Server
- *logicalhost* – Logical host serving *sql_server* (not the physical host)
- `–60 –10 –120 –300` – Parameters which specify a probe cycle time of 60 seconds, a connectivity probe cycle count of 10 seconds, a probe timeout of 120 seconds, and a restart delay of 300 seconds
- *user/password* – Login name created in Step 3 on page @–9 and password to be used for fault monitoring
- `$SYBASE/install/RUN_`*sqlserver* – File used to start the SQL Server
- *backupserver* (optional) – Name of the backup server
- `$SYBASE/install/RUN_`*backupserver* (optional) – File used to start the backup server

See the `hasybase(1M)` man page for details on all options to `hasybase(1M)`.

8. Bring the Sybase Server into service.

Bring the SQL Server into service by running the `hasybase(1M)` command. Monitoring for that SQL Server will start automatically. See the `hasybase(1M)` man page for additional details.

```
# hasybase start sql_server
```

Note - If you did not start the Sybase SQL Server before issuing the `hasybase(1M)` command, then issuing it now will cause Sun Cluster to start the SQL Server.

6.2.2 Setting Up Sun Cluster HA for Sybase Clients

Clients must always refer to the Sybase database by using the logical host name, not the physical host name. Except for during start-up, the database should always be available if the logical host is responding on the network.

Note - Sybase client-server connections will not survive a Sun Cluster HA for Sybase switchover. The client application must be prepared to cope with disconnection and reconnection or recovery as appropriate. A transaction monitor might simplify the application. Further, Sun Cluster HA for Sybase server recovery time is application dependent.

If your application uses functions from RDBMS dynamic link libraries, you must ensure that these libraries are available in the event of failover. To do so:

- Install the link libraries on the client, or
- If Sun Cluster HA for NFS service is to be provided by the cluster, copy the libraries to the logical host and set the environment variables to the directory path of the link libraries, then share those directories to clients by using NFS.

6.3 Verifying the Sun Cluster HA for Sybase Installation

Perform the following verification tests to ensure the Sun Cluster HA for Sybase installation was performed correctly.

The purpose of these sanity checks is to ensure that the Sybase SQL Server can be started by all nodes running Sun Cluster HA for Sybase and can be accessed successfully by the other nodes in the configuration.

▼ How to Verify the Sun Cluster HA for Sybase Installation

1. **Log in to the Sun Cluster node mastering the logical host, and set the** SYBASE **environment variable.**

   Log in as `sybase` to the Sun Cluster node that currently masters the logical host, and set the SYBASE environment variable to point to the directory in which the `interfaces` file resides.

   a. **Confirm that you can start the SQL Server from this host.**

   b. **Confirm that you can connect to the Sybase SQL Server from this host.**

   c. **Shut down the SQL Server.**

2. **Transfer the logical host to another Sun Cluster node in the cluster.**

Use the `haswitch(1M)` command to transfer the logical host containing the SQL Server to another Sun Cluster node in the cluster.

3. **Log in to the Sun Cluster node now mastering the logical host, and repeat the checks listed in .**

   Log in as `sybase` to the new master node and confirm interactions with the SQL Server.

CHAPTER 7

Setting Up and Administering Sun Cluster HA for Informix

This chapter provides instructions for setting up and administering the Sun Cluster HA for Informix data service on your Sun Cluster nodes.

This chapter includes the following procedures:

7.1 Preparing to Install Sun Cluster HA for Informix

Use the information in this section to prepare Sun Cluster nodes for Sun Cluster HA for Informix installation.

7.1.1 Selecting an Install Location for Sun Cluster HA for Informix

You can install the Informix binaries either on a physical host's local disks or on the shared multihost disks (logical host). Use Table 7–1, which shows the advantages and disadvantages of each install location, to determine the install location that best fits your needs.

TABLE 7–1 Informix Install Locations Comparison

Install Location	Advantages	Disadvantages
Multihost Disks	■ Eases administration - only one copy must be maintained	■ Sacrifices redundancy ■ In some cases, compromises availability ■ In cases of switchover or accidental removal of Informix binaries, may cause Informix to be unavailable ■ Requires a mirror, so disk space must be doubled
Local Disks	■ Increases redundancy - each node is a copy	■ Increases administrative overhead - multiple copies must be maintained

7.1.2 Setting Up the `/etc/nsswitch.conf` File

On each node that can master the logical host running Sun Cluster HA for Informix, modify the `/etc/nsswitch.conf` file so that "group" lookups are directed to files first. For example:

```
...
 group: files nisplus
 ...
```

Sun Cluster HA for Informix uses the `su informix` command when starting and stopping the database server.

Adding this setting will ensure that the `su informix` command does not refer to NIS or NIS+ when the network information name service is not available due to a failure of the public network.

7.1.3 Setting Up Multihost Disks for Sun Cluster HA for Informix

If you are using Solstice DiskSuite, you can configure Sun Cluster HA for Informix to use UFS logging or raw mirrored metadevices. Refer to Appendix B," for details about setting up metadevices.

If you are using Sun StorEdge Volume Manager, you can configure Sun Cluster HA for Informix to use VxFS logging or raw devices. Refer to your Sun StorEdge Volume Manager documentation for more information.

7.2 Installing Sun Cluster HA for Informix

Use the procedures in this section to prepare the Sun Cluster nodes, to install the Informix software, to create Informix databases, and to set up Sun Cluster HA for Informix.

Before setting up Sun Cluster HA for Informix, you must have configured the Sun Cluster software on each node by using the procedures described in Chapter 3.

If you are running both Sun Cluster HA for NFS and Sun Cluster HA for Informix in your cluster, you can set up the data services in any order.

▼ How to Prepare the Nodes and Install the Informix Software

These are the high-level steps to prepare Sun Cluster nodes for Informix installation and install the Informix software:

- Choosing a location for the `$INFORMIXDIR` directories
- Creating the `/etc/group` and `/etc/passwd` file entries for Informix
- Installing the Informix software

Caution - Perform all steps described in this section on all Sun Cluster nodes.

Consult your Informix documentation before performing this procedure.

These are the detailed steps to prepare Sun Cluster nodes and install the Informix software.

1. **Prepare the environment for Informix installation.**

   Choose a location for the `$INFORMIXDIR` directories, on either a local or multihost disk.

Note - If you choose to install `$INFORMIXDIR` on a local disk, install it as a file system on its own disk, separate from the operating environment, if possible. This prevents Informix from being overwritten if the operating environment is reinstalled.

2. **On each node, create an entry for the `informix` group in the `/etc/group` file.**

   The group normally is named `informix`. Verify that users `root` and `informix` are members of the `informix` group. For example:

```
informix:*:520:root,informix
```

   While you can make the name service entries in a network name service (for example, NIS or NIS+) so that the information is available to Sun Cluster HA for Informix clients, you also should make entries in the local `/etc` files to eliminate dependency on the network name service.

Note - You must replicate this information on each cluster node.

3. **On each node, create an entry for the `informix` user (*informix_id*) group in the `/etc/passwd` file.**

   This entry normally is named `informix`. For example:

```
# useradd -u 135 -g informix -d /informix informix
```

4. **Verify that the `$INFORMIXDIR` directory is owned by *informix_id* and is included in the `informix` group.**

```
# chown informix $INFORMIXDIR
# chgrp informix $INFORMIXDIR
```

   If `$INFORMIXDIR` is a symbolic link to the Informix home directory, then use the following command:

```
# chown informix $INFORMIXDIR
 # chgrp informix $INFORMIXDIR
 # chown -h $INFORMIXDIR
 # chgrp -h $INFORMIXDIR
```

5. **Note the requirements for Informix installation.**

   Informix binaries can be installed on either the local disks of the physical host or the multihost disks. See Section 7.1.1 “Selecting an Install Location for Sun Cluster HA for Informix” on page 7-2,” for more information.

Note - If you plan to install Informix software on the multihost disks, you first must start Sun Cluster and take ownership of the disk group. You also must install the INFORMIX_ESQL Embedded Languages Runtime Facility product in the `/var/opt/informix` file on all nodes running Sun Cluster HA for Informix by using the `installesql` command. See Section 7.2.1 “Creating an Informix Database and Setting Up Sun Cluster HA for Informix” on page 7-5,” for details.

---

6. **Install the Informix software.**

   Log in as `informix` to ensure ownership of the entire directory before performing this step. If Informix is installed on a multihost disk, verify that the `/var/opt/informix/bin` directory is owned by Informix owner *informix_id* on all cluster nodes; this is not necessary if Informix is installed on the local disk.

   For complete instructions on installing Informix, refer to the Informix installation documentation.

7.2.1 Creating an Informix Database and Setting Up Sun Cluster HA for Informix

Use the following procedures to create and set up the initial Informix database in a Sun Cluster configuration. If you are creating and setting up additional databases, perform only the procedures described in “How to Create an Informix Database” on page 7–7 and “How to Set Up Sun Cluster HA for Informix” on page 7–8.

▼ How to Prepare Logical Hosts for Informix Databases

1. **Make sure Sun Cluster is started and the node owns the disk groups.**

   If necessary, as `root`, use the `scadmin(1M)` command.

```
# scadmin startcluster
```

This command causes the node to take disk group ownership.

2. **Configure the disk devices for use by your volume manager.**

   a. **If you are using Solstice DiskSuite, set up UFS logging devices or raw mirrored metadevices on all nodes that will be running Sun Cluster HA for Informix.**

   If you will be using raw mirrored metadevices to contain the databases, change the owner, group, and mode of each of the raw mirrored metadevices. If you are not using raw mirrored metadevices, skip this step.

Caution - While Informix supports raw I/O to both raw physical devices and raw metadevices (mirrored or nonmirrored), Sun Cluster only supports raw Informix I/O on raw mirrored volumes or metadevices. You cannot use devices such as `/dev/rdsk/c1t1d1s2` to contain Informix data under Sun Cluster.

Instructions for creating mirrored metadevices are provided in Appendix B."

If you are creating raw mirrored metadevices, type the following commands for each metadevice.

```
# vxvol set owner=informix_id group=dba_id mode=600 \
  /dev/vx/rdsk/diskgroup_name/volume_name
```

```
# chown informix_id /dev/md/diskset/rdsk/dn
 # chgrp informix_id /dev/md/diskset/rdsk/dn
 # chmod 600 /dev/md/diskset/rdsk/dn
```

   b. **If you are using Sun StorEdge Volume Manager, set up VxFS logs or raw devices on all Sun Cluster nodes.**

   If you will be using raw devices to contain the databases, change the owner, group, and mode of each device. If you are not using raw mirrored metadevices, skip this step. See your Sun StorEdge Volume Manager documentation for further details. Type the following commands for each raw device:

```
# chown informix_id /dev/vx/rdsk/diskgroup_name/volume_name
 # chgrp informix_id /dev/vx/rdsk/diskgroup_name/volume_name
 # chmod 600 /dev/vx/rdsk/diskgroup_name/volume_name
```

▼ How to Create an Informix Database

These are the high-level steps to create an Informix database:

- Preparing the Informix environment
- Creating and customizing the `$ONCONFIG` file
- Creating Informix entries in the `sqlhosts` file
- Configuring the `/etc/services` file

These are the detailed steps to create an Informix database:

1. **Prepare the Informix environment.**

   Prepare the Informix environment using the following commands. In `csh`:

```
# setenv INFORMIXDIR base_dir
 # setenv INFORMIXSERVER server_name
 # setenv ONCONFIG file_name
 # setenv INFORMIXSQLHOSTS $INFORMIXDIR/etc/sqlhosts
```

2. **Create and customize the `$ONCONFIG` file, and copy it to all other nodes in the cluster.**

   Place the `$ONCONFIG` file in the `$INFORMIXDIR/etc` directory. Customize the `ROOTPATH`, `ROOTSIZE`, and `PHYSFILE` variables in the `$ONCONFIG` file, and set `DBSERVERNAME=$INFORMIXSERVER`. Once the `$ONCONFIG` file is customized, copy it to the other nodes in the cluster, if `$INFORMIXDIR` is on the local disk of the physical host.

3. **Create Informix entries in the `$INFORMIXSQLHOSTS` file.**

   Entries in the `sqlhosts` file are composed of four fields:

   - `DBSERVERNAME` – This is the `$INFORMIXSERVER`.
   - `NETTYPE` – Select `ONSOCTCP` or `ONTLITCP`.

- HOSTNAME – This is the logical host name.
- SERVICENAME – This must match the SERVICENAME entry in the /etc/services file.

If you are installing the Informix binaries on the multihost disks, replicate the sqlhosts file on both local disks and on the multihost disks under /var/opt/informix. If you are installing the Informix binaries on the local disks, replicate the sqlhosts file on both local disks.

4. **Configure the** /etc/services **file.**

   You must be root to perform this step. Edit the /etc/services file; add the SERVICENAME and listener port number. The listener port number must be unique.

   When you use the TCP/IP connection protocol, the SERVICENAME entry in the sqlhosts file must correspond to the SERVICENAME entry in the /etc/services file.

▼ How to Set Up Sun Cluster HA for Informix

These are the high-level steps to set up Sun Cluster HA for Informix:

- Updating the /var/opt/informix/inftab file
- Activating the Informix data service
- Enabling user and password for fault monitoring
- Shutting down the Informix database
- Verifying installation of Sun Cluster software and cluster daemon
- Starting Sun Cluster
- Activating the Informix data service by using the hareg(1M) command
- Bringing Informix under control of Sun Cluster HA for Informix
- Bringing the Informix database into service

These are the detailed steps to set up Sun Cluster HA for Informix.

1. **Update the** /var/opt/informix/inftab **file with** $ONCONFIG **information.**

   You must include entries for all $ONCONFIG files associated with your databases in the /var/opt/informix/inftab file on all nodes running Sun Cluster HA for Informix. You must keep this file current on all nodes running Sun Cluster HA for Informix for a failover to succeed. Update the file manually as database servers are added or removed.

   Entries in the /var/opt/informix/inftab file have the following format:

```
$ONCONFIG:$INFORMIXDIR
```

$ONCONFIG is the name of the ONCONFIG file. $INFORMIXDIR is the path to the Informix installation directory. For example, the inftab file might look similar to the following:

```
onconfig.node1:/export/home/informix
```

2. **Activate the Informix data service.**

   Log in as user informix and invoke the oninit command, which formats or "cooks" the raw disk or UFS filespace assigned in the $ONCONFIG file as specified by the ROOTPATH variable.

```
# oninit -iy
```

3. **Create entries in the /etc/hosts.equiv file or the ~informix/.rhosts file that grant permissions to the informix user to access the database from other cluster nodes.**

   These entries have the following format:

```
hostname informix
```

   where *hostname* is the name of the other cluster nodes and informix is the user name.

4. **Enable access for the user and password to be used for fault monitoring.**

   Invoke the dbaccess *dbname* command and add the following lines to the appropriate dbaccess screen.

```
# dbaccess dbname -
...
 grant connect to root;
 grant resource to root;
 grant dba root;
 grant select on sysprofile to root;
```

The database to be probed by the HA fault monitor is identified by the database name (*dbname)*. If that *dbname* is not defined as sysmaster, use the dbaccess *dbname* command to add the following line to the appropriate dbaccess screen:

```
create synonym sysprofile for sysmaster:informix_owner.sysprofile;
```

5. **Shut down the Informix database.**

   As user `informix`, use the `onmode` command to shut down the Informix database:

```
# onmode -ky
```

6. **Verify that Sun Cluster is installed and running on all nodes that will run Sun Cluster HA for Informix.**

   As `root`, verify the configuration with the `hastat` command:

```
# hastat
```

   If the cluster nodes are not running already, start them. The first node must be started using the `scadmin startcluster` command and all other nodes are then started using the `scadmin startnode` command. Refer to the `scadmin(1M)` man page for more information on starting the cluster.

7. **Register and activate the Informix data service by using the `hareg(1M)` command.**

   Run the `hareg(1M)` command on only one host.

   If the Informix data service is not yet registered, use the `hareg(1M)` command to register it. To register the data service only on the logical host, include the `–h` option and logical host name:

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

   If the cluster is running already, use the `hareg(1M)` command to activate the Informix data service:

```
# hareg -y informix
```

8. **Bring Informix under control of Sun Cluster HA for Informix.**

   Run the following command so that the instance will be monitored by Sun Cluster.

```
# hainformix insert $ONCONFIG logicalhost 60 10 120 300 \
  dbname $INFORMIXSERVER
```

The above command line includes the following:

- `hainformix insert` – Command and subcommand
- `$ONCONFIG` – Name of the Informix configuration file
- *logicalhost* – Logical host serving `$ONCONFIG` (not the physical host)
- `–60 –10 –120 –300` – Parameters which specify a probe cycle time of 60 seconds, a connectivity probe cycle count of 10, a probe time out of 120 seconds, and a restart delay of 300 seconds
- *dbname* – Name of the database that Sun Cluster 2.2 is to monitor
- `$INFORMIXSERVER` – Name of the Informix server

See the `hainformix(1M)` man page for details on all options to `hainformix(1M)`.

9. Bring the Sun Cluster HA for Informix database into service.

Bring the Sun Cluster HA for Informix database into service by using the `hainformix(1M)`command.

```
# hainformix start $ONCONFIG
```

Note - If you did not start the Informix OnLine server before this step, then Sun Cluster will start the Informix OnLine Server for you when you issue the `hainformix start` command.

7.2.2 Setting Up Sun Cluster HA for Informix Clients

Clients always must refer to the database using the logical host name and not the physical host name. Except during start-up, the database always should be available if the logical host is responding on the network.

Note - Informix client-server connections will not survive a Sun Cluster HA for Informix switchover. The client application must be prepared to handle disconnection and reconnection or recovery as appropriate. A transaction monitor might simplify the application. Further, Sun Cluster HA for Informix server recovery time is application-dependent.

If your application uses functions from RDBMS dynamic link libraries, you must ensure that these libraries are available in the event of failover. To do so:

- Install the link libraries on the client, or
- If Sun Cluster HA for NFS service is to be provided by the cluster, copy the libraries to the logical host and set the environment variables to the directory path of the link libraries, then NFS share those directories to clients.

7.3 Verifying the Sun Cluster HA for Informix Installation

Perform the following verification tests to ensure the Sun Cluster HA for Informix installation was performed correctly.

The purpose of these sanity checks is to ensure that the Informix OnLine server can be started by all Sun Cluster nodes running Sun Cluster HA for Informix and can be accessed successfully by the other nodes in the configuration.

▼ How to Verify the Sun Cluster HA for Informix Installation

1. **Log in to the Sun Cluster node mastering the logical host, and set the Informix environment variables.**

   Log in as `informix` to the node that currently masters the logical host. Set the environment variables `INFORMIXDIR`, `INFORMIXSERVER`, `ONCONFIG`, and `INFORMIXSQLHOSTS`.

   a. **Confirm that you can start the Informix OnLine server from this host.**

   b. **Confirm that you can connect to the Informix OnLine server from this host:**

   ```
   # dbaccess sysmaster@$INFORMIXSERVER
   ```

   c. **Shut down the Informix OnLine server.**

2. **Transfer the logical host to another node in the cluster.**

   Use the `haswitch(1M)` command to transfer the logical host containing the Informix OnLine server to another node.

3. **Log in to the node now mastering the logical host, and repeat the checks listed in Step 1 on page @–12.**

   Log in as `informix` to the new master node, and confirm interactions with the Informix OnLine server.

CHAPTER 8

Setting Up and Administering Sun Cluster HA for Netscape

This chapter provides instructions for setting up and configuring the Sun Cluster HA for Netscape data services.

This chapter includes the following procedures:

8.1 Sun Cluster HA for Netscape Overview

The Sun Cluster HA for Netscape data services consist of a group of Netscape™ applications that can be made highly available by running them in the Sun Cluster environment. Table 8–1 displays the data service application and its associated highly available data service.

TABLE 8–1 Sun Cluster HA for Netscape Data Services

Netscape Application	Highly Available Data Service Name	Package Name
Netscape News Server	Sun Cluster HA for Netscape News	`SUNWscnew`
Netscape Mail Server	Sun Cluster HA for Netscape Mail	`SUNWscnsm`
Netscape HTTP/Web Server	Sun Cluster HA for Netscape HTTP	`SUNWschtt`
Netscape LDAP Server	Sun Cluster HA for Netscape LDAP	`SUNWscnsl`

See the *Sun Cluster 2.2 Release Notes* for a list of the supported release levels for the data services.

Table 8–2 describes the high-level procedures to set up Netscape data service applications to run with Sun Cluster.

TABLE 8–2 High-Level Steps to Set Up Netscape Data Service Applications

Task	Go To ...
1. Installing the Solaris and Sun Cluster environments, installing the Netscape data service packages, and installing all required patches	Chapter 3
2. Starting the cluster with the `scadmin(1M)` command	Chapter 3, or the `scadmin(1M)` man page
3. (Optional) Installing and setting up DNS for the Netscape data services to use	Chapter 12

TABLE 8–2 High-Level Steps to Set Up Netscape Data Service Applications *(continued)*

Task	Go To ...
4. Installing and configuring Sun Cluster HA for Netscape data services	Section 8.2 “Installing Netscape Services” on page 8-3,” and Section 8.7 “Configuring the Sun Cluster HA for Netscape Data Services” on page 8-20
5. Registering and starting the Sun Cluster HA for Netscape data services	Section 8.7 “Configuring the Sun Cluster HA for Netscape Data Services” on page 8-20

Note - If you are running multiple data services in your Sun Cluster configuration, you can set up the data services in any order, with one exception: if you use Sun Cluster HA for DNS, you must set it up before setting up Sun Cluster HA for NFS. DNS software is included in the Solaris environment. See Chapter 12, for details. If the cluster is to obtain the DNS service from another server, then configure the cluster to be a DNS client first.

After installation, do not manually start and stop the Netscape data services. Once started, they are controlled by Sun Cluster.

The procedures described in this chapter assume that you are familiar with the Sun Cluster concepts of disksets, logical hosts, physical hosts, switchover, takeover, and data services.

8.2 Installing Netscape Services

Before you begin the Netscape services installation, complete the appropriate pre-requisite steps listed in Table 8–2.

▼ How to Install Netscape Services

Consult your Netscape application documentation before performing this procedure. All of the procedures in this chapter must be performed as root.

1. **Make sure each logical host is served by its default master.**

   Each Netscape application will be installed from the physical host that is the logical host’s default master. If necessary, switch over the logical hosts to be served by their respective default masters.

Note - The logical host names you use in your Sun Cluster configuration should be used as the server names when you install and configure the Netscape applications in the following steps. This is necessary for failover of the Netscape server to work properly.

2. **After cluster reconfiguration is complete, install the Netscape application software on the logical hosts by using the Netscape** `ns-setup` **command on the distribution CD.**

   You should install and test the Netscape application software (DNS, Netscape HTTP Server, Netscape News, Netscape Mail, and LDAP) independently of Sun Cluster. Refer to the Netscape application software installation documentation for installation instructions.

Note - Before you install the Netscape application software, refer to the section in this chapter describing the configuration procedures for each Netscape application. These sections describe Sun Cluster-specific configuration information that you must supply when you install the Netscape applications.

Proceed to Section 8.3 “Installing Netscape News” on page 8-4, to install the Sun Cluster HA for Netscape data services.

8.3 Installing Netscape News

Sun Cluster HA for Netscape News is Netscape News running under the control of Sun Cluster. This section describes how to install Netscape News (by using the `ns-setup` command) to enable it to run as the Sun Cluster HA for Netscape News data service. Refer to Netscape documentation for the standard Netscape installation instructions.

There are two prerequisites to installing Netscape News using the `ns-setup` command.

1. A user name and a group name must be configured for the news server. Create these names on all Sun Cluster servers that will be running Sun Cluster HA for Netscape News and verify that they have the same ID numbers on all systems.
2. DNS must be configured and enabled on all servers running Sun Cluster HA for Netscape News. All Sun Cluster servers must have the same `/etc/resolv.conf` file, and the `hosts` entry in the `/etc/nsswitch.conf` file must include `dns`.

Netscape News requires some variation from the default installation parameters, notably:

- Specifying the logical host name rather than the physical host name
- Do not use the default server root disk when prompted, your files must reside on the multihost disk
- When supplying the base install directory pathname, this should be the location of the start and stop scripts

▼ How to Install Netscape News

This procedure shows the user interaction with the `ns-setup` command. Only the sections that are specific to Sun Cluster HA for Netscape News are shown here. For the other sections, choose or change the default value as appropriate.

1. **Run the `ns-setup` command from the Netscape News install directory on the CD.**

   Change directory to the Netscape News distribution location on the CD, and run the `ns-setup` command.

```
phys-hahost1# cd /cdrom/news_server/solaris/news/install
phys-hahost1# ./ns-setup
```

Note - The Netscape directory on the CD might be different from that shown in the example. Check your Netscape documentation for the actual location.

After the licensing agreement you see:

```
Netscape Communications Corporation
 Netscape SuiteSpot Server Installation
 ---------------------------------------------------------------

 This program will extract the server files from the distribution
 media and install them into a directory you specify. This
 directory is called the server root and will contain the server
 programs, the Administration Server, and the server configuration
 files.

 To accept the default shown in brackets, press return.

 Server root [/usr/netscape/suitespot]:
```

2. **Enter the logical host name for the Netscape News Server and the appropriate DNS domain name.**

   A full name is of type *hostname.domainname*, such as `hahost1.sun.com`. You should not accept the default, and you can enter any directory name you might have created for the data service here as well.

   Note - You must use the logical host name rather than the physical host name here and everywhere else you are asked, for Sun Cluster HA for Netscape News to fail over correctly.

   For example:

```
Machine's name [phys-hahost1]:hahost1
```

   Follow the screen instructions (in many cases you may want to accept the default choices) for the server installation.

3. **Enter Server Administrator ID and password when asked.**

   Follow the guidelines for your system.

   Note - The default administration port is not the port that the data services will listen on, so it is an acceptable default. You will however, want to make note of the port number for future use.

   When you see the following, your installation is complete and ready for configuration:

```
Your parameters are now entered into the Administration Server
 configuration files, and the Administration Server will be
 started.

 Writing configuration files ...
```

4. **Bring up the Netscape browser.**

```
Web browser [netscape]:
```

   You should see the Netscape browser.

5. **Enter the URL of the logical host where the server is installed.**

   For example:

```
logicalhostname:admin_port_#
```

   Enter the proper user ID and password when prompted.

   You see the page with the logical host name you assigned and the Administration port number.

6. **Click Create New Netscape Collabra Server 3.5.**

   This should be the second link from the bottom of the page. You should see another page load.

7. **Click OK from the bottom of the following page.**

   You see the Success page with the name of the new server and associated port number.

8. **Run the** `hadsconfig` **command from the physical host:**

```
[phys-hahost1]: hadsconfig
```

9. **Enter the number for the** `nsnews` **menu item at the prompt.**

   Your choices may vary depending on the services installed.

10. **Enter the number for the** `Create a new instance` **item at the prompt.**

    You see something similar to the following:

```
Name of the instance [?] :
 Logical host [?] :
 Base directory of product installation [?] :
 Server Port Number [?] :
 Time between probes (sec) [60]  [?] :
 Time out value for the probe (sec) 60]  [?] :
 Take over flag [y]  [?] :
```

11. **Enter the name for the instance.**

12. **Enter your logical hostname.**

13. **Enter the location to the logical host where the service is installed for the Base directory location.**

    This is the location of the start and stop scripts. For example, `/netscape-1/vol01/nsnews/hahost1`. You may have created other directories between the logical host and service directories.

14. **Enter the server port number you want the server to listen on.**

    For example 119 for News.

15. **Either accept or change the defaults by entering the appropriate information.**

    You can change the defaults now if necessary. After you finish, a confirmation appears.

16. **Add this instance.**

    You see `instance added to workfile`.

17. **Enter the menu item number to go to the `Main Menu`.**

    You see `Configuration has changed in workfile`.

18. **Update the configuration from the workfile.**

19. **Enter the appropriate information when you see the `checking node status...` message.**

20. **Enter the `Quit` menu item number.**

    You are returned to the root prompt of your physical host.

21. **Register and activate the service by using the `hareg(1M)` command.**

    Run the `hareg(1M)` command on only one host.

    If the service is not yet registered, use the `hareg(1M)` command to register it. To register the service only on the logical host, include the –h option and logical host name:

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

    If the cluster is running already, use the `hareg(1M)` command to activate the service:

```
# hareg -y nsnews
```

22. Confirm your News server operation by making a telnet connection to your logical host where the service is listening.

For example:

```
# telnet logicalhost 119
```

After you have confirmed the server's operation, your installation is complete.

8.4 Installing Netscape Web or HTTP Server

Sun Cluster HA for Netscape HTTP is a Netscape Web or HTTP Server running under the control of Sun Cluster. This section describes the steps to take when installing the Netscape Web or HTTP Server (by using the `ns-setup` command) to enable it to run as the Sun Cluster HA for Netscape HTTP data service.

You can install any of a number of Netscape web server products. Refer to Netscape documentation for standard installation instructions.

Note - If you will be running the Sun Cluster HA for Netscape HTTP service and an HTTP server for Sun Cluster Manager (SCM), configure the HTTP servers to listen on different ports. Otherwise, there will be a port conflict between the two servers.

Note - You must follow certain conventions when you configure URL mappings for the web server. For example, when setting the CGI directory, to preserve availability you must locate the mapped directories on the multihost disks associated with the logical host serving HTTP requests for this mapping. In this example, you would map your CGI directory to */logicalhost*`/commerce/ns-home/cgi-bin`.

In situations where the CGI programs access "back-end" data, make sure the data also is located on the multihost disks associated with the logical host serving the HTTP requests.

In situations where the CGI programs access "back-end" servers such as an RDBMS, make sure that the "back-end" server also is controlled by Sun Cluster. If the server is an RDBMS supported by Sun Cluster, use one of the highly available RDBMS packages. If not, you can put the server under Sun Cluster control using the APIs documented in the *Sun Cluster 2.2 API Developer's Guide*.

Netscape Web or HTTP server requires some variation from the default installation parameters, notably:

- Specifying the logical host name rather than the physical host name

- Do not use the default server root disk when prompted, your files must reside on the multihost disk
- When supplying the base install directory pathname, this should be the location of the start and stop scripts

▼ How to Install Netscape Web or HTTP Server

This procedure shows the user interaction with the `ns-setup` command. Only the sections that are specific to Sun Cluster HA for Netscape HTTP are shown here. For the other sections, choose or change the default value as appropriate.

1. **Run the `ns-setup` command from the Netscape Commerce install directory on the CD.**

   From the Netscape Commerce distribution location on the CD, run the `ns-setup` command.

```
phys-hahost1# cd /cdrom/commerce/solaris/us/https/install
phys-hahost1# ./ns-setup
```

Note - The Netscape directory on the CD might be different from that shown in the example. Check your Netscape documentation for the actual location.

After the licensing agreement you see:

```
Netscape Communications Corporation
 Netscape SuiteSpot Server Installation
 ---------------------------------------------------------------

 This program will extract the server files from the distribution
 media and install them into a directory you specify. This
 directory is called the server root and will contain the server
 programs, the Administration Server, and the server configuration
 files.

 Server root [/usr/netscape/suitespot]:
 To accept the default in brackets, press return.
```

2. **Enter the logical host name for the Netscape Web Server and the appropriate DNS domain name.**

A full name is of type *hostname.domainname*, such as `hahost1.sun.com`. You can enter any directory name you might have created for the data service here as well.

Note - You must use the logical host name rather than the physical host name here and everywhere else you are asked, for Sun Cluster HA for Netscape HTTP to fail over correctly.

For example:

```
Machine's name [phys-hahost1]:hahost1
```

Follow the screen instructions (in most cases you may want to accept the default choices) for the server installation.

3. **Enter the Server Administrator ID and password when asked.**

   Follow the guidelines for your system.

Note - The default administration port is not the port where the data services will listen on, so it is an acceptable default. You will however, want to make note of the port number for future use.

When the following message appears, your installation is ready for configuration:

```
Your parameters are now entered into the Administration Server
 configuration files, and the Administration Server will be
 started.

 Writing configuration files ...
```

4. **Bring up the Netscape browser.**

```
Web browser [netscape]:
```

The Netscape browser appears.

5. **Enter the URL of the logical host where the server is installed.**

   For example:

```
logicalhostname:admin_port_#
```

Enter the proper user ID and password when prompted.

You see the page with the logical host name you assigned and the Administration port number.

6. **Enter the proper user ID and password when prompted.**

   You see the page with the logical host name you assigned and the Administration port number.

7. **Click Create New Netscape Enterprise Server 3.5.1.**

   This should be the second link from the bottom of the page. You should see another page load.

8. **Click OK from the bottom of the following page.**

   You see the Success page with the name of the new server and associated port number.

9. **Run the** `hadsconfig` **command from the physical host:**

```
[phys-hahost1]: hadsconfig
```

10. **Enter the number for the** `nshttp` **menu item at the prompt.**

    Your choices may vary depending on the services installed.

11. **Enter the number for the** `Create a new instance` **item at the prompt.**

    You see something similar to the following:

```
Name of the instance [?]
 Logical host [?]
 Base directory of product installation [?]
 Server Port Number [?]
 Time between probes (sec) [60]  [?]
 Time out value for the probe (sec) 60]  [?]
 Take over flag [y]  [?]
```

12. **Enter a name for the instance.**

13. **Enter your logical hostname.**

14. Enter the location to your logical host for the Base directory location.

This is the location of the start and stop scripts. For example, `/netscape-1/vol01/nshttps/hahost1`. You may have created other directories between the logical host and service directories.

15. Enter the server port number you want the server to listen on.

16. Either accept or change the defaults here by entering the appropriate information.

You can change the defaults now, if necessary. After you finish, a confirmation appears.

17. Add this instance.

You see `instance added to workfile`.

18. Enter the menu item number to go to the Main Menu.

You see `Configuration has changed in workfile`.

19. Update the configuration from the workfile.

20. Enter the appropriate information when you see the `checking node status...` message.

21. Enter the `Quit` menu item number.

Return to the root prompt of your physical host.

22. Register and activate the service by using the `hareg(1M)` command.

Run the `hareg(1M)` command on only one host.

If the service is not yet registered, use the `hareg(1M)` command to register it. To register the service only on the logical host, include the –h option and logical host name:

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

If the cluster is running already, use the `hareg(1M)` command to activate the service:

```
# hareg -y nshttp
```

23. Confirm your HTTP server operation by making a telnet connection to your logical host where the service is listening.

For example:

```
# telnet logicalhost port#
```

After you have confirmed the server's operation, your installation is complete.

8.5 Installing Netscape Mail

Sun Cluster HA for Netscape Mail is Netscape Mail running under control of Sun Cluster. This section describes the steps to take when installing Netscape Mail (by using the `ns-setup` command) to enable it to run as the Sun Cluster HA for Netscape Mail data service.

The Sun Cluster HA for Netscape Mail data service is an asymmetric data service. Only one logical host in the cluster provides the mail services.

Note - The Sun Cluster HA for Netscape Mail service fault probing might cause `/var/log/syslog` to fill up quickly. To avoid this, disable logging of `mail.debug` messages in the `/etc/syslog.conf` file by commenting out the `mail.debug` entry and sending a HUP signal to the `syslogd(1M)` daemon.

The following are required on the Sun Cluster servers before configuring Sun Cluster HA for Netscape Mail:

- Each server must have a unique user ID and unique group ID that contains only this unique user ID. These particular IDs will be used by the mail system. The names and numbers must be identical on all servers running Sun Cluster HA for Netscape Mail.
- DNS must be configured and enabled on all servers running Sun Cluster HA for Netscape Mail. All Sun Cluster servers must have the same `/etc/resolv.conf` file, and the `hosts` entry in the `/etc/nsswitch.conf` file must include `dns`.

Because Netscape Mail is installed on one server, Sun Cluster HA for Netscape Mail requires some variation from the default installation parameters, notably:

- Specifying the logical host name rather than the physical host name
- Installing the Netscape Mail software and spool directories on the multihost disks
- Do not use the default server root disk when prompted, your files must reside on the multihost disk
- When supplying the base install directory pathname, this should be the location of the start and stop scripts

▼ How to Install Netscape Mail

This procedure shows the user interaction with the `ns-setup` command and Sun Cluster commands. Only the sections that are specific to Sun Cluster HA for Netscape Mail are shown here. For the other sections, choose or change the default value as appropriate.

1. **Run the `ns-setup` command from the Netscape Mail install directory on the CD.**

   Change directory to the Netscape Mail distribution location on the CD, and run the `ns-setup` command:

```
phys-hahost1# cd /cdrom/commerce/solaris/us/https/mail
phys-hahost1# ./ns-setup
```

Note - The Netscape directory on the CD might be different from that shown in the example. Check your Netscape documentation for the actual location.

After the licensing agreement you should see something like the following:

```
Netscape Communications Corporation
 Netscape SuiteSpot Server Installation
 ---------------------------------------------------------------

 This program will extract the server files from the distribution
 media and install them into a directory you specify. This
 directory is called the server root and will contain the server
 programs, the Administration Server, and the server
 configuration files.

 Server root [/usr/netscape/suitespot]:
 To accept the default in brackets, press return.
```

2. **Enter the logical host name for the Netscape Web Server and the appropriate DNS domain name.**

   A full name is of type *hostname.domainname*, such as `hahost1.sun.com`. You can enter any directory name you might have created for the data service here as well.

Note - You must use the logical host name rather than the physical host name here and everywhere else you are asked, for Sun Cluster HA for Netscape Mail to fail over correctly.

For example:

```
Machine's name [phys-hahost1]:hahost1
```


```
Machine's name [phys-hahost1]:hahost1
```


Follow the screen instructions (in many cases you may want to accept the default choices) for the server installation.

3. **Enter Server Administrator ID and password when asked.**

   Follow the guidelines for your system.

Note - The default administration port is not the port that the data services will listen on, so it is an acceptable default.

You see information similar to the following:

```
Attempting to start Netscape Admin Server...
```

4. **Continue with the installation when prompted.**

5. **Specify user, group, and domain names.**

   Enter the user name you configured for the mail server on all Sun Cluster servers running Sun Cluster HA for Netscape Mail.

6. **Specify directories for system components.**

   You are asked the names of directories where the various components of the system will be installed. Enter a location on the logical host, for example, `/hahost1/mail/mailbox`, and `/hahost1/mail/postoffice`.

7. **Specify a Server Identifier name.**

8. **Specify whether to use the NIS module and the Greeting Forms feature.**

You see a confirmation of the information you specified as in the following example:

```
Mail user name:
 Domain name:
 Mailbox directory:
 Post Office directory:
 Server Identifier:
 NIS lookups:
 Greeting forms:

 You may accept these choices or quit the installation.

 Install Netscape Messaging Server? [y]:
```

9. When you are ready, install the Netscape Messaging Server and reply to queries when prompted.

Depending on how you set up your configuration, specify items appropriately. After all changes take effect, you see:

```
Netscape Messaging Server installation complete
```

10. Start the Netscape Messaging Server when prompted.

11. Run the `hadsconfig` command from the physical host:

```
[phys-hahost1]: hadsconfig
```

12. Enter the number for the `nsmail` menu item at the prompt.

Your choices may vary depending on the services installed.

13. Enter the number for the `Create a new instance` item at the prompt.

You see something similar to the following:

```
Name of the instance [nsmail]  [?]
 Logical host [?]
 Take over flag [y]

 Following are the specifications of this instance
 Name of the instance :
 Logical host :
```

(Continuation)

```
Number of times to retry :
Time between retries (sec) :
Configuration File :
Fault probe program :
Time between probes (sec) :
Time out value for the probe (sec) :
Take over flag :
Add this instance ?  (yes/no) [yes]
Instance added to workfile
Press enter to return to main menu
```

14. Enter the name for the instance.

15. Enter your logical hostname.

16. Either accept or change the defaults for the remaining items depending on your configuration.

You can change these defaults at this time if necessary.

17. Add this instance.

You see `Instance added to workfile`.

18. Go to the `–Main –Menu` **when prompted.**

You see `Configuration has changed in workfile`.

19. Update the configuration from the workfile when prompted.

20. Enter the `–Quit` **menu item number.**

Return to the root prompt of your physical host.

21. Register and activate the service by using the `hareg(1M)` **command.**

Run the `hareg(1M)` command on only one host.

If the service is not yet registered, use the `hareg(1M)` command to register it. To register the service only on the logical host, include the `–h` option and logical host name:

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

If the cluster is running already, use the `hareg(1M)` command to activate the service:

```
# hareg -y nsmail
```

This completes the installation of Netscape Mail.

8.6 Installing Netscape Directory Server

Sun Cluster HA for Netscape LDAP is the Netscape Directory Server using the Lightweight Directory Assistance Protocol (LDAP) and running under the control of Sun Cluster. This section describes the steps to take when installing Netscape Directory Server (by using the `ns-setup` command) to enable it to run as the Sun Cluster HA for Netscape LDAP data service.

If not already installed, use `pkgadd` to install the `SUNWhadns` package on each Sun Cluster server.

Netscape Directory server requires some variation from the default installation parameters, notably:

- Specifying the logical host name rather than the physical host name
- Do not use the default server root disk when prompted, your files must reside on the multihost disk
- When supplying the base install directory pathname, this should be the location of the start and stop scripts

▼ How to Install Netscape Directory Server

This procedure shows the user interaction with the `ns-setup` command. Only the sections that are specific to Sun Cluster HA for Netscape LDAP are shown here. For the other sections, choose or change the default values as appropriate. These are the basic steps; consult your Netscape Directory Server documentation for details.

1. **Install Netscape Directory Server.**

   Choose the logical host that will provide directory services for the cluster. Install the Netscape Directory Server product on that logical host's shared disk.

2. **Run the `ns-setup` command from the install directory on the CD.**

Run the `ns-setup` command from the Netscape Directory Server install directory. You must supply the logical host name when `ns-setup` prompts you for the full server name. In this example, the logical host is `hahost1`:

```
phys-hahost1# ./ns-setup
Server root [/usr/netscape/suitespot]: /hahost1/d1/ns-home
Full name [phys-hahost1]: hahost1
```

3. **Use the Netscape admin server to configure and test the Netscape Directory Server.**

   See your Netscape documentation for details.

8.7 Configuring the Sun Cluster HA for Netscape Data Services

After you have installed the Sun Cluster HA for Netscape packages and the Netscape applications, you are ready to configure the individual data services.

Sun Cluster currently supports these Netscape data services: Sun Cluster HA for Netscape News, Sun Cluster HA for Netscape HTTP, Sun Cluster HA for Netscape Mail, and Sun Cluster HA for Netscape LDAP.

Sun Cluster HA for Netscape allows configurable *instances,* which are independent of each other. For example, you can install and configure any number of web servers; each such server is considered an instance.

All Sun Cluster HA for Netscape data services are configured by using the `hadsconfig(1M)` command.

▼ How to Configure the Sun Cluster HA for Netscape Data Services

1. **Run the `hadsconfig(1M)` command to configure your Sun Cluster data service(s).**

   The `hadsconfig(1M)` command is used to create, edit, and delete instances of a Sun Cluster HA for Netscape data service. See the `hadsconfig(1M)` man page for details. Refer to Section 8.7.1 “Configuration Parameters for the Sun Cluster

HA for Netscape Data Services” on page 8-22, for information on the input to supply to `hadsconfig(1M)`.

```
phys-hahost1# hadsconfig
```

Note - Sun Cluster HA for Netscape News and Sun Cluster HA for Netscape HTTP support installation of multiple instances of news and http servers, which can be located anywhere in the cluster. Because the mail protocol listens to a well-known port, only one instance of Sun Cluster HA for Netscape Mail can exist in a cluster.

2. **Register the Sun Cluster HA for Netscape data services.**

   Register the data services by running the `hareg(1M)` command.

   If you installed the data service packages on all potential masters of a logical host but not on all hosts in the cluster, use the –`h` option and specify the logical host name.

TABLE 8–3 Data Service Registration Names and Syntax

Data Service	Registration Syntax
Sun Cluster HA for Netscape HTTP	`hareg -s -r nshttp [-h` *logicalhost*`]`
Sun Cluster HA for Netscape News	`hareg -s -r nsnews [-h` *logicalhost*`]`
Sun Cluster HA for Netscape Mail	`hareg -s -r nsmail [-h` *logicalhost*`]`
Sun Cluster HA for Netscape LDAP	`hareg -s -r nsldap [-h` *logicalhost*`]`

3. **Run the `hareg -Y` command to enable all services and perform a cluster reconfiguration.**

```
phys-hahost1# hareg -Y
```

The configuration is complete.

8.7.1 Configuration Parameters for the Sun Cluster HA for Netscape Data Services

This section describes the information you supply to the `hadsconfig(1M)` command to create configuration files for each Sun Cluster HA for Netscape data service. The `hadsconfig(1M)` command uses templates to create these configuration files. The templates contain some default, some hard coded, and some unspecified parameters. You must provide values for those parameters that are unspecified.

The fault probe parameters, in particular, can affect the performance of Sun Cluster HA for Netscape data services. Tuning the probe interval value too low (increasing the frequency of fault probes) might encumber system performance, and also might result in false takeovers or attempted restarts when the system is simply slow.

Fault probe parameters are configurable for Sun Cluster HA for Netscape HTTP, Sun Cluster HA for Netscape News, and Sun Cluster HA for Netscape LDAP. Fault probe parameters are not configurable for Sun Cluster HA for Netscape Mail.

All Sun Cluster HA for Netscape data services require you to set the takeover flag. This flag specifies how Sun Cluster will handle partial failover. There are two options:

- `–y` (yes) – Sun Cluster will attempt to switch over the logical host to another master, but if the attempt fails the logical host will remain on the original master.
- `–n` (no) – Sun Cluster will not move the logical host to another master, even if it detects problems with the data server, nor will it take any action against the faulty data server or database on the logical host.

8.7.1.1 Configuration Parameters for Sun Cluster HA for Netscape News

Configure the Sun Cluster HA for Netscape News parameters listed in the `hadsconfig(1M)` input form by supplying options described in Table 8–4.

TABLE 8–4 Configuration Parameters for Sun Cluster HA for Netscape News

Parameter	Description
Name of the instance	Nametag used as an identifier for the instance. The log messages generated by Sun Cluster HA for Netscape News refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. For example, if you specify "`nsnews_119`," the `hadsconfig(1M)` command produces "`SUNWscnew_nsnews_119`."
Logical host	Name of the logical host that provides service for this instance of Sun Cluster HA for Netscape News.

TABLE 8–4 Configuration Parameters for Sun Cluster HA for Netscape News *(continued)*

Parameter	Description
Base directory of product installation	Rooted path name specifying the location on the multihost disk of the Netscape News installation. This is the "instance path," for example, `/hahost1/news-hahost1`.
Probe interval	The time, in seconds, between fault probes. The default interval is 60 seconds.
Probe timeout	The time, in seconds, after which a fault probe will time out. The default timeout value is 20 seconds.
Server port number	Unique port for this instance of Sun Cluster HA for Netscape News. This is the "Server Port" value you supplied to the `ns-setup` command.
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the data service instance. Possible values are `–y` (yes), or `–n` (no).

Note - Do not use an HA administrative file system for the Sun Cluster HA for Netscape News installation base directory. Check the `vfstab.`*logicalhost* file to verify that the base directory you have chosen is not an HA administrative file system.

8.7.1.2 Configuration Parameters for Sun Cluster HA for Netscape HTTP

Configure the Sun Cluster HA for Netscape HTTP parameters listed in the `hadsconfig(1M)` input form by supplying options described in Table 8–5.

TABLE 8–5 Configuration Parameters for Sun Cluster HA for Netscape News

Parameter	Description
Name of the instance	Nametag used as an identifier for the instance. The log messages generated by Sun Cluster refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. For example, if you specify "nshttp_80," the `hadsconfig(1M)` command produces "SUNWschtt_nshttp_80."
Logical host	Name of logical host that provides service for this instance of Sun Cluster HA for Netscape HTTP.

TABLE 8–5 Configuration Parameters for Sun Cluster HA for Netscape News *(continued)*

Parameter	Description
Base directory of product installation	This is the base directory of the product installation, plus the server type and server port number. For example, `/hahost1/https-hahost`.
Probe interval	The time, in seconds, between fault probes. The default interval is 60 seconds.
Probe timeout	The time, in seconds, after which a fault probe will time out. The default timeout value is 20 seconds.
Server port number	Unique port for this instance of Sun Cluster HA for Netscape HTTP. This is the "Server Port" value you supplied to the `ns-setup` command.
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the data service instance. Possible values are `–y` (yes), or `–n` (no).

Note - Do not use an administrative file system for the Sun Cluster HA for Netscape HTTP installation base directory. Check the `vfstab` *.logicalhost* file to verify that the base directory you have chosen is not an administrative file system.

8.7.1.3 Configuration Parameters for Sun Cluster HA for Netscape Mail

Configure the Sun Cluster HA for Netscape Mail parameters listed in the `hadsconfig(1M)` input form by supplying options described in Table 8–6.

TABLE 8–6 Configuration Parameters for Sun Cluster HA for Netscape Mail

Parameter	Description
Name of the instance	Nametag used as an identifier for the instance. The log messages generated by Sun Cluster refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. For example, if you specify "nsmail," the `hadsconfig(1M)` command produces "SUNWscnsm_nsmail."
Logical host	Name of logical host that provides service for this instance of Sun Cluster HA for Netscape Mail.
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the data service instance. Possible values are `–y` (yes), `–n` (no).

8.7.1.4 Configuration Parameters for Sun Cluster HA for Netscape LDAP

Configure the Sun Cluster HA for Netscape LDAP parameters listed in the `hadsconfig(1M)` input form by supplying options described in Table 8–7.

TABLE 8–7 Configuration Parameters for Sun Cluster HA for Netscape LDAP

Parameter	Description
Name of the instance	Nametag used as an identifier for the instance. The log messages generated by Sun Cluster refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. For example, if you specify "nsldap," the `hadsconfig(1M)` command produces "SUNWhansm_nsldap."
Logical host	Name of logical host on which the Netscape Directory Server resides.
Base directory of product installation	This is the base directory of the product installation. Include the logical host name prefixed with `ns-slapd_`. For example, `/hahost1/d1/ns-home/ns-slapd_hahost1/`. Make sure the directory you specify includes the `start` script.
Server port number	Unique port for this instance of Sun Cluster HA for Netscape LDAP. This is the "Server Port" value you supplied to the `ns-setup` command. The default value is 389.

TABLE 8–7 Configuration Parameters for Sun Cluster HA for Netscape LDAP *(continued)*

Parameter	Description
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the data service instance. Possible values are –y (yes), or –n (no).
Probe interval	The time, in seconds, between fault probes. The default interval is 60 seconds.
Probe timeout	The time, in seconds, after which a fault probe will time out. The default timeout value is 30 seconds.

CHAPTER 9

Setting Up and Administering Sun Cluster HA for Tivoli

This chapter describes procedures for setting up and administering the Sun Cluster HA for Tivoli data service on your Sun Cluster servers.

This chapter includes the following procedures:

9.1 Overview of Sun Cluster HA for Tivoli

The Sun Cluster HA for Tivoli product consists of a Tivoli Management Environment (TME) server, Tivoli managed nodes, and other components that become highly available when run in the Sun Cluster environment.

You can place Tivoli components inside or outside the Sun cluster; any components you place inside the cluster will be protected by failover. For example, if a Tivoli object dispatcher configured in the cluster fails, it will be restarted automatically or will fail over to another host.

For those Tivoli servers and managed nodes that you place inside the cluster, you must place each one on a separate logical host.

9.2 Installing the Tivoli Server and Managed Nodes

After you have installed and configured the Sun Cluster product, install the Tivoli server and managed nodes. You can use either the Tivoli desktop utility or shell commands to install the Tivoli product. See your Tivoli documentation for detailed Tivoli installation procedures.

▼ How to Install the Tivoli Server and Managed Nodes

Before starting this procedure, you should have already installed and configured Sun Cluster and set up file systems and logical hosts.

1. **Start Sun Cluster and make sure the logical host is mastered by the physical host on which you will install Tivoli.**

   In this example, the physical host is `phys-hahost1` and the logical hosts are `hahost1` and `hahost2`:

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

2. **Run the Tivoli preinstallation script, `WPREINST.SH`.**

   The `WPREINST.SH` script is located on the Tivoli media. The script creates links from an installation directory you specify back to the Tivoli media.

3. **Install the Tivoli server and specify directory locations on the logical host for Tivoli components.**

   Install the Tivoli server on the multihost disk associated with the logical host.

   Note - You can use the Tivoli GUI or Tivoli commands to install the Tivoli server and managed nodes. If you use the Tivoli command line, you must set the environment variable: `DOGUI=no`.

   The following example specifies directory locations on the logical host for the TME binaries and libraries, TME server database, man pages, message catalogs, and X11 resource files:

```
phys-hahost1# ./wserver -c cdrom_path -a $WLOCALHOST -p \
/hahost1/d1/Tivoli! BIN=/hahost1/d1/Tivoli/bin! \
LIB=/hahost1/d1/Tivoli/lib! ALIDB=/hahost1/d1/Tivoli! \
```

```
MAN=/hahost1/d1/Tivoli/man! \
APPD=/hahost1/d1/Tivoli/X11/app-defaults! \
CAT=/hahost1/d1/Tivoli/msg_cat! CreatePaths=1
```

4. Install Tivoli patches.

See your Tivoli documentation or service provider for applicable patches, and install them using instructions in your Tivoli documentation.

5. (Optional) Rename the Tivoli environment directory and copy the directory to all other possible masters of the logical host.

Rename the Tivoli environment directory to prevent it from being overwritten by another installation. Then copy the directory to all other possible masters of the logical host on which the Tivoli server is installed.

```
phys-hahost1# mv /etc/Tivoli /etc/Tivoli.hahost1
phys-hahost1# tar cvf /tmp/tiv.tar /etc/Tivoli.hahost1
phys-hahost1# rcp /tmp/tiv.tar phys-hahost2:/tmp
phys-hahost2# tar xvf /tmp/tiv.tar
```

6. Set up paths and stop and restart the Tivoli daemon.

Use the `setup_env.sh` script to set up paths. The default port number is 94.

```
phys-hahost1# . /etc/Tivoli.hahost1/setup_env.sh
phys-hahost1# odadmin shutdown
phys-hahost1# oserv -H hahost1 -p port_number -k $DBDIR
```

7. (Optional) Install the Tivoli managed node instance on the second logical host.

For example:

```
phys-hahost1# wclient -c cdrom_path -I -p hahost1-region \
BIN=/hahost2/d1/Tivoli/bin! LIB=/hahost2/d1/Tivoli/lib! \
DB=/hahost2/d1/Tivoli! MAN=/hahost2/d1/Tivoli/man! \
APPD=/hahost2/d1/Tivoli/X11/app-defaults! \
CAT=/hahost2/d1/Tivoli/msg_cat! CreatePaths=1 hahost2
```

8. (Optional) Rename the Tivoli environment directory and copy the directory to all other possible masters.

Rename the Tivoli environment directory to prevent it from being overwritten by another installation. Then copy the directory to all other possible masters of the logical host on which the Tivoli server is installed.

```
phys-hahost1# mv /etc/Tivoli /etc/Tivoli.hahost2
phys-hahost1# tar cvf /tmp/tiv.tar /etc/Tivoli.hahost2
phys-hahost1# rcp /tmp/tiv.tar phys-hahost2:/tmp
phys-hahost2# tar xvf /tmp/tiv.tar
```

9. Modify the `/etc/services` file.

Add the following entry to the `/etc/services` file on each physical host that is a possible master of a Tivoli instance. The default port number for Tivoli is 94.

```
objcall    port_number/tcp
```

10. Verify the Tivoli installation.

Before configuring Sun Cluster HA for Tivoli, verify correct installation of the Tivoli server, Tivoli managed node instance, and Tivoli managed nodes used for probing.

```
phys-hahost1# . /etc/Tivoli.hahost1/setup_env.sh
phys-hahost1# odadmin odlist
phys-hahost1# wping hahost1
phys-hahost1# wping hahost2
```

Note - Execute the `setup_env.sh` file from only the first logical host. If you execute the `setup_env.sh` file from the second logical host, the `odadmin` and `wping` commands will fail.

11. Create an administrative user and set permissions correctly on the Tivoli server.

Use the Tivoli user interface to create an administrator with user ID `root` and group ID `root`, and give it `user`, `admin`, `senior`, and `super` authorization. This will enable probing by running the `wping` command.

12. Stop the Tivoli servers or server daemons.

The daemons will be re-started automatically by Sun Cluster when you start the cluster, or when the logical host is switched between masters. The first invocation of `odadmin` shuts down the TMR server. The second invocation shuts down the managed node.

```
phys-hahost1# odadmin shutdown
phys-hahost1# . /etc/Tivoli.hahost2/setup_env.sh
phys-hahost1# odadmin shutdown
```

Proceed to Section 9.3 “Installing and Configuring Sun Cluster HA for Tivoli” on page 9-5, to register and install the Sun Cluster HA for Tivoli data service.

9.3 Installing and Configuring Sun Cluster HA for Tivoli

This section describes the steps to install, configure, register, and start Sun Cluster HA for Tivoli. You must install and set up Sun Cluster and the Tivoli product before configuring Sun Cluster HA for Tivoli.

You will configure Sun Cluster HA for Tivoli by using the `hadsconfig(1M)` command. See the `hadsconfig(1M)` man page for details.

▼ How to Install and Configure Sun Cluster HA for Tivoli

1. **On each Sun Cluster server, install the Tivoli package,** `SUNWsctiv`**, in the default location, if it is not installed already.**

   If the Tivoli package is not installed already, use the `scinstall(1M)` command to install it on each Sun Cluster server that is a potential master of the logical host on which Tivoli is installed.

2. **Run the** `hadsconfig(1M)` **command on one node to configure Sun Cluster HA for Tivoli for both the server and managed node.**

   Use the `hadsconfig(1M)` command to create, edit, and delete instances of the Sun Cluster HA for Tivoli data service for both the server and managed node. Refer to Section 9.3.1 “Configuration Parameters for Sun Cluster HA for Tivoli” on page 9-7, for information on input to supply to `hadsconfig(1M)`. Run the command on one node only.

```
phys-hahost1# hadsconfig
```

   Note - Only the Tivoli server and Tivoli managed node should be configured as instances under the control of Sun Cluster. The Tivoli managed nodes used for probing need not be controlled by Sun Cluster.

3. **Register the Sun Cluster HA for Tivoli data service by running the** `hareg(1M)` **command.**

   Run the command on only one node:

```
phys-hahost1# hareg -s -r tivoli
```

4. **Use the** `hareg(1M)` **command to enable Sun Cluster HA for Tivoli and perform a cluster reconfiguration.**

   Run the command on only one node:

```
phys-hahost1# hareg -y tivoli
```

   The configuration is complete.

9.3.1 Configuration Parameters for Sun Cluster HA for Tivoli

This section describes the information you supply to the `hadsconfig(1M)` command to create configuration files for Sun Cluster HA for Tivoli. The `hadsconfig(1M)` command uses templates to create these configuration files. The templates contain some default, some hard coded, and some unspecified parameters. You must provide values for those parameters that are unspecified.

The fault probe parameters, in particular, can affect the performance of Sun Cluster HA for Tivoli. Tuning the probe interval value too low (increasing the frequency of fault probes) might encumber system performance, and also might result in false takeovers or attempted restarts when the system is simply slow.

Configure Sun Cluster HA for Tivoli by supplying the `hadsconfig(1M)` command with parameters listed in Table 9–1.

TABLE 9–1 Configuration Parameters for Sun Cluster HA for Tivoli

Parameter	Description
Name of the instance	Nametag used as an identifier for the instance. The log messages generated by Sun Cluster HA for Tivoli refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. For example, if you specify “tivoli,” the `hadsconfig(1M)` command produces “SUNWsctiv_tivoli.”
Logical host	Name of the logical host that provides service for this instance of Sun Cluster HA for Tivoli.
Port number	Unique port for Sun Cluster HA for Tivoli. The default port number is 94.
Configuration directory	The directory of the database, that is, the full path of the `$DBDIR`. For example, `/hahost1/d1/Tivoli/<database>.db`.
Local probe flag	Specifies whether the local probe is started automatically at cluster reconfiguration or when the Tivoli service is activated. Possible values are `y` or `–n`.
Probe interval	Time in seconds between successive fault probes. The default is 60 seconds.
Probe timeout	Time out value in seconds for the probe. If the probe has not completed within this amount of time, Sun Cluster HA for Tivoli considers it to have failed. The default is 60 seconds.

TABLE 9–1 Configuration Parameters for Sun Cluster HA for Tivoli *(continued)*

Parameter	Description
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the Tivoli instance. Possible values are –y or –n.
TIV_OSERV_TYPE	This is the TME type. Possible values are `server` or `client`.
TIV_BIN	The path to the TME binaries specified during installation of the instance. This is equivalent to `$BINDIR` without the “Solaris2” suffix. For example, `/hahost1/d1/Tivoli/bin`.
TIV_LIB	The path to the TME libraries specified during installation of the instance. For example, `/hahost1/di/Tivoli/lib`. This is equivalent to `$LIBDIR` without the “Solaris2” suffix.

CHAPTER **10**

Installing and Configuring Sun Cluster HA for SAP

Sun Cluster HA for SAP is SAP components made highly available by running in a Sun Cluster environment. This chapter provides instructions for planning and configuring Sun Cluster HA for SAP on Sun Cluster servers.

This chapter includes the following procedures:

10.1 Sun Cluster HA for SAP Overview

The Sun Cluster HA for SAP data service eliminates single points of failure in a SAP system and also provides fault monitoring and failover mechanisms for the SAP application.

These basic services of the SAP system should be placed within the Sun Cluster framework:

- Database instance
- Central instance, consisting of
  - Message server
  - Enqueue server
  - Dispatcher
- NFS file service

In a Sun Cluster configuration, protection of SAP components is best provided as described in Table 10–1.

TABLE 10–1 Protection of SAP Components

SAP Component	Protected by...
SAP database instance	Sun Cluster HA for Oracle
SAP central instance	Sun Cluster HA for SAP
NFS file service	Sun Cluster HA for NFS
SAP application servers	SAP, through redundant configuration

The Sun Cluster HA for SAP data service can be installed during or after initial cluster installation using `scinstall(1M)`. Sun Cluster HA for SAP requires a functioning cluster that already contains logical hosts and associated IP addresses and disk groups. See Chapter 3, for details about initial installation of clusters and data services. The Sun Cluster HA for SAP data service can be registered after the basic components of the Sun Cluster and SAP software have been installed.

10.2 Configuration Guidelines for Sun Cluster HA for SAP

Consider these general guidelines when designing a Sun Cluster HA for SAP configuration:

- SAP uses a large amount of memory and swap space. Consult the SAP documentation for memory and swap recommendations.
- Be generous in estimating the total possible load on standby servers in case of failover. Allocate ample resources for CPU, swap, shared memory, and I/O bandwidth on the standby server, because in case of failover, the central instance and database instance might co-exist on the standby.
- Limit physical and logical host names to eight characters or less, if possible.

10.2.1 Supported Configurations

See your Enterprise Services representative for the most current information about supported SAP versions. More information on each configuration type is provided in the following sections.

10.2.1.1 Two-Node Cluster With One Logical Host

The simplest SAP cluster configuration is a two-node cluster with one logical host, as illustrated in Figure 10–1. In this asymmetric configuration, the SAP central instance and database instance (collectively called the central system), are both placed on one node. NFS is also be placed on the same node. This configuration is relatively easy to configure and administer. A drawback is that the backup node is underutilized. In case of failover, the central instance, database instance, and NFS service are switched to the backup node.

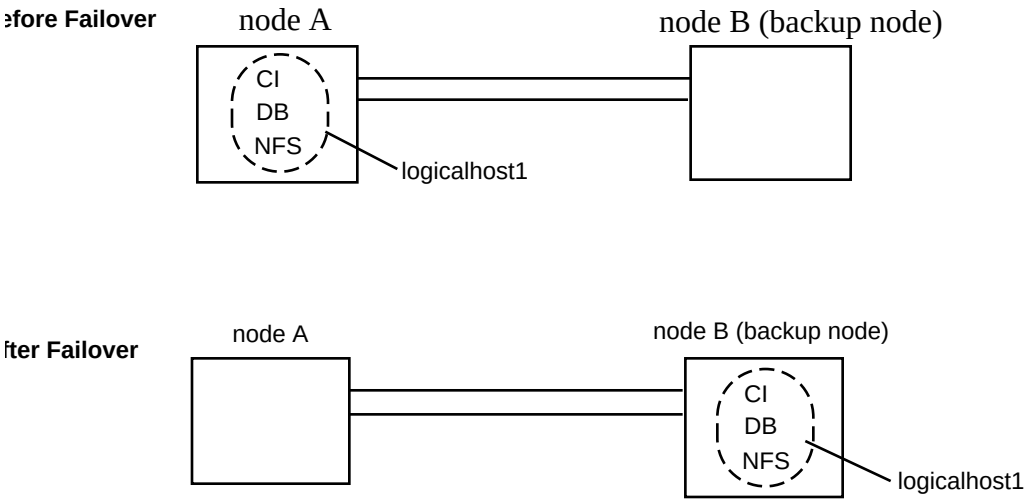


Figure 10–1 Asymmetric SAP Configuration

10.2.1.2 Two-Node Cluster With One Logical Host and Development or Test System

In this configuration, the central system (the central instance and database instance) is placed on one node and a development or test system is placed on a backup node. The development or test system remains running until a failover of the logical host moves the central system to the backup node. This scenario is illustrated in Figure 10–2. In this configuration, you must customize the Sun Cluster HA for SAP `hasap_stop_all_instances` script such that the development or test system is shut down before the SAP central instance is switched over and brought up. See the `hasap_stop_all_instances(1M)` man page and Section 10.2.4 “Configuration Options for Application Servers and Test/Development Systems” on page 10-11, for more information.

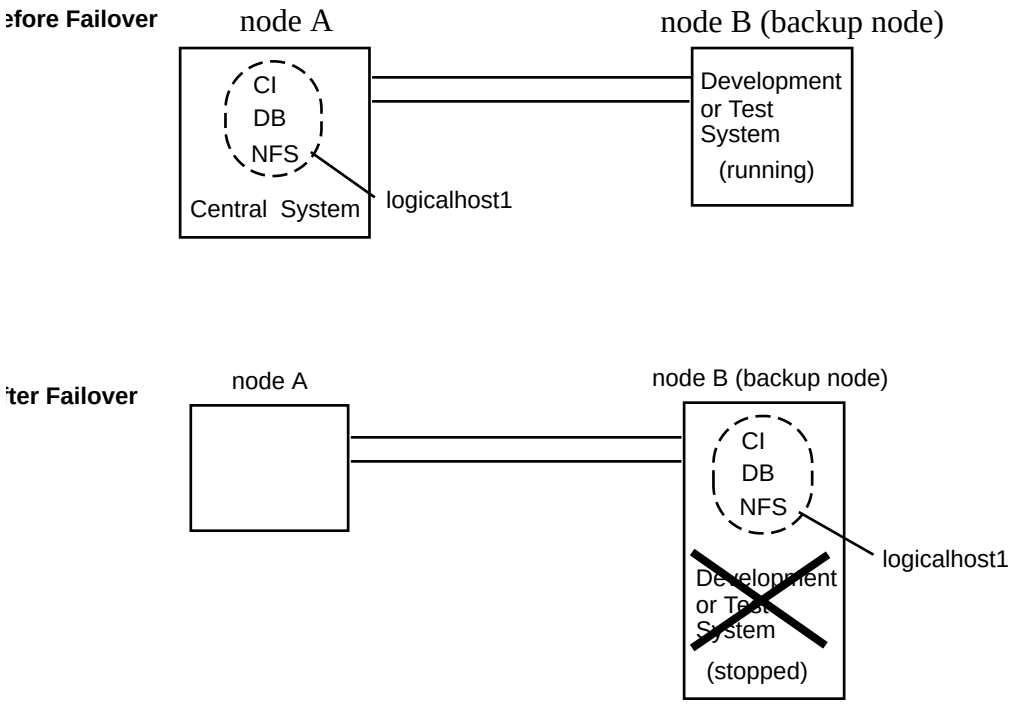


Figure 10–2 Asymmetric SAP Configuration with Development or Test System

10.2.1.3 Two-Node Cluster With One Logical Host, Application Servers, and Separate NFS Cluster

You can also place SAP application servers on one or both physical hosts. In this configuration, you must provide NFS services from a host outside the cluster. Set up the application servers to NFS-mount the file systems from the external NFS cluster, as illustrated in Figure 10–3. In case of failover, the logical host containing the central system (the central instance and database instance) switches to the backup node. The application servers do not migrate with the logical host, but are instead started or shut down depending on where the logical host is mastered. This prevents the application servers from competing for resources with the central instance and database.

NFS is protected by Sun Cluster HA for NFS. For more information, see Section 10.2.5 “Sun Cluster HA for NFS Considerations” on page 10-13.

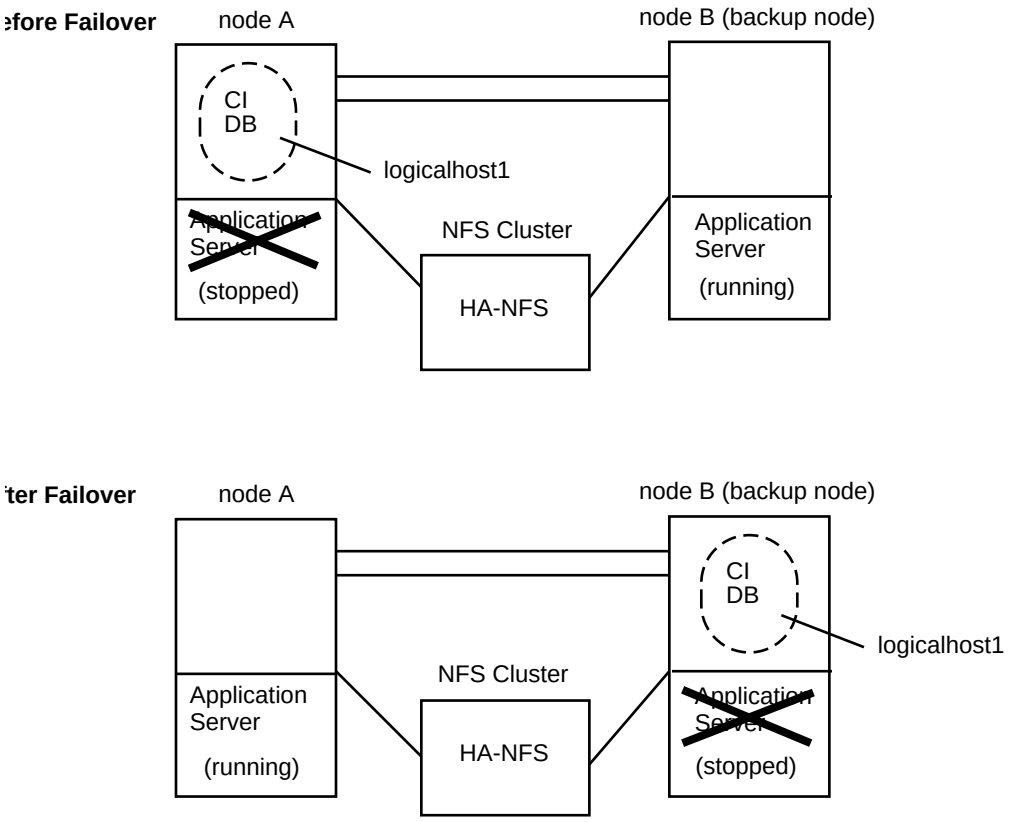


Figure 10–3 Asymmetric SAP Configuration with Application Servers and External HA-NFS

10.2.1.4 Two-Node Cluster With Two Logical Hosts

A two-node cluster with two logical hosts can be configured with the SAP central instance on one logical host and the SAP database instance on the other logical host, as illustrated in Figure 10–4. In this configuration, the nodes are load-balanced and both are utilized. In case of failover, the central instance or database instance is switched to the sibling node.

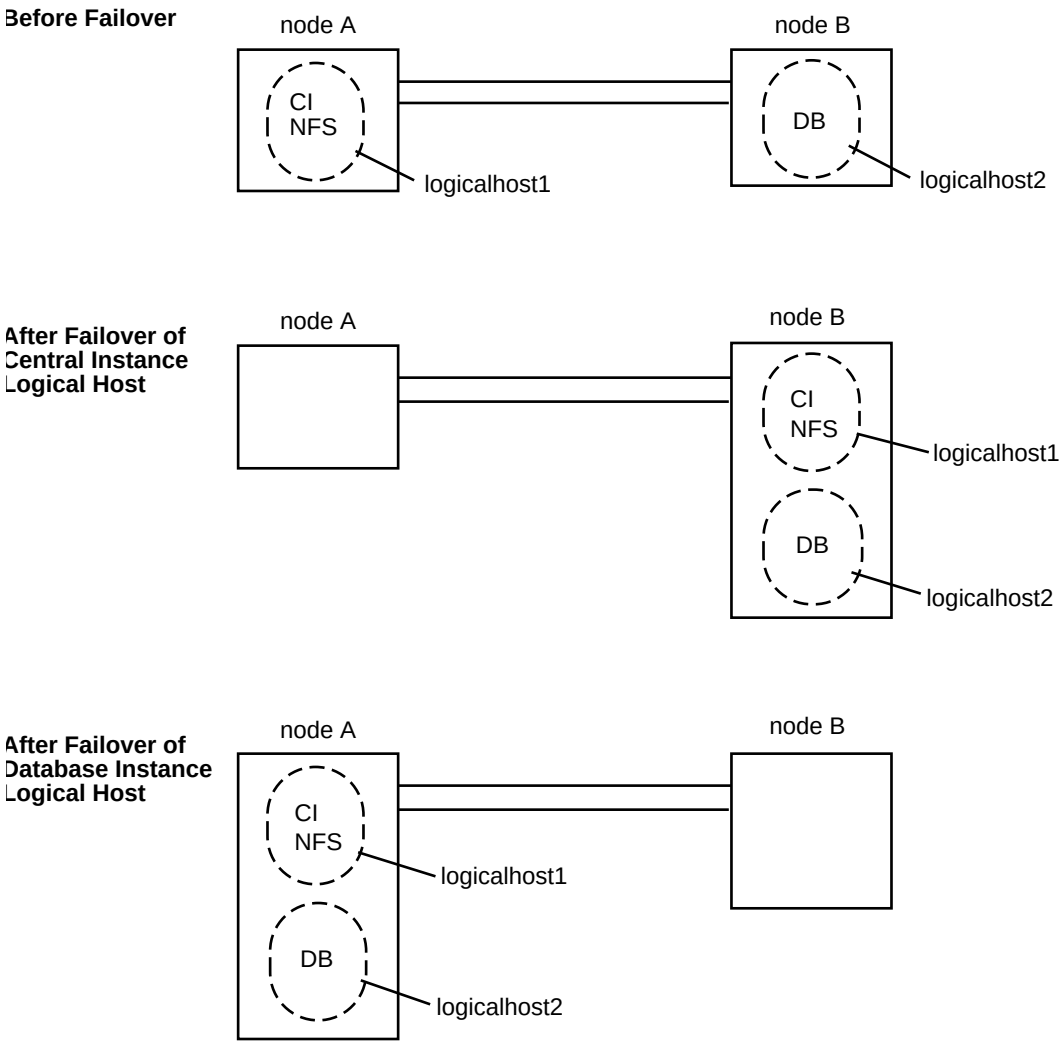


Figure 10–4 Symmetric SAP Configuration With Two Logical Hosts

10.2.1.5 Two-Node Cluster With Two Logical Hosts, Application Servers, and Separate NFS Cluster

A two-node cluster with two logical hosts can be configured with SAP application servers on one or both physical hosts. In this configuration, you must provide NFS services from a host outside the cluster. Set up the application servers to NFS-mount the file systems from the external NFS cluster, as illustrated in Section 10.2.2 “Pre-Installation Considerations” on page 10-10. In this case, both nodes are utilized and load-balanced.

In case of failover, the logical hosts switch over to the sibling node. The application servers do not fail over.

If the central instance logical host fails over, the application server can be shut down through the `hasap_stop_all_instances` script.

There are no customizable scripts to start and stop application servers in case of failover of the database logical host. If the database logical host fails over, the application servers cannot be shut down to release resources for the database logical host. Therefore, you must size your configuration to allow for the possible scenario in which the central instance, database instance, and application server are all running on the same node simultaneously.

In this configuration, NFS is protected by Sun Cluster HA for NFS. For more information, see Section 10.2.5 “Sun Cluster HA for NFS Considerations” on page 10-13.

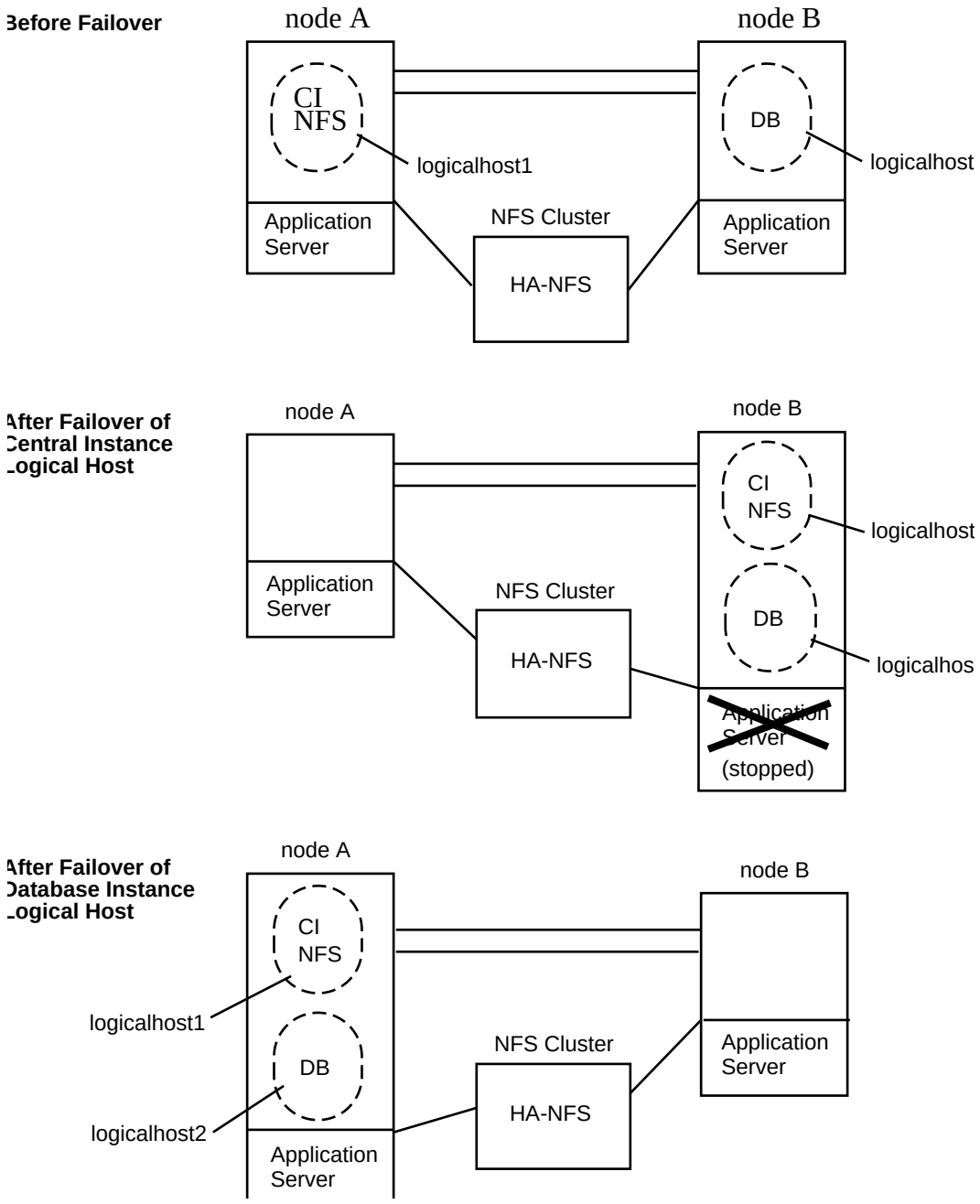


Figure 10–5 Symmetric SAP Configuration With Two Logical Hosts and Application Servers

10.2.2 Pre-Installation Considerations

Before installing Sun Cluster with `scinstall(1M)`, consider the following issues that apply to SAP configurations.

- Use a logging file system:
  - If your volume manager is SSVM, use VxFS and Dirty Region Logging.
  - If your volume manager is Solstice DiskSuite, use either Solaris UFS logging or Solstice DiskSuite UFS logging.
- Configure separate disk groups for SAP software and the database. The `scinstall(1M)` command cannot configure more than one disk group per logical host. Therefore, do not set up logical hosts with `scinstall(1M)` during initial cluster installation. Instead, set up logical hosts with `scconf(1M)`, after the cluster is up. See the `scconf(1M)` man page for details.
- Limit physical host names and logical host names to eight characters or less.
- Do not configure any node to be an NFS client of another node within the same cluster.
- (For SAP 4.0*x* with Oracle) On all potential masters of the central instance logical host, set aside space in `/var/opt/oracle` for the Oracle client binaries. Roughly 140 Mbytes is required. See your SAP documentation for details.

See also Section 10.2.5 “Sun Cluster HA for NFS Considerations” on page 10-13, and Section 10.4 “Preparing the SAP Environment” on page 10-17.

10.2.3 Sun Cluster Software Upgrade Considerations

Note these SAP-related issues before performing an upgrade to Sun Cluster 2.2 from HA 1.3 or Sun Cluster 2.1.

- On each node, if you customized `hasap_start_all_instances` or `hasap_stop_all_instances` scripts in HA 1.3 or Sun Cluster 2.1, save them to a safe location before beginning the upgrade to Sun Cluster 2.2. Restore the scripts after completing the upgrade. Do this to prevent loss of your customizations when Sun Cluster 2.2 removes the old scripts.
- The configuration parameters implemented in Sun Cluster 2.2 are different from those implemented in HA 1.3 and Sun Cluster 2.1. Therefore, after upgrading to Sun Cluster 2.2, you will have to re-configure Sun Cluster HA for SAP by running the `hadsconfig(1M)` command.

  Before starting the upgrade, view the existing configuration and note the current configuration variables. For HA 1.3, use the `hainetconfig(1M)` command to view the configuration. For Sun Cluster 2.1, use the `hadsconfig(1M)` command to view the configuration. After upgrading to Sun Cluster 2.2, use the `hadsconfig(1M)` command to re-create the instance.

- In Sun Cluster 2.2, the `hareg -n` command shuts down the entire Sun Cluster HA for SAP data service, including all instances and fault monitors. In previous releases, the `hareg -n` command, when used with Sun Cluster HA for SAP, shut down only the fault monitors.

  Additionally, before turning on the Sun Cluster HA for SAP data service using `hareg -y`, you must stop the SAP central instance. Otherwise, the Sun Cluster HA for SAP data service will not be able to start and monitor the instance properly.

10.2.4 Configuration Options for Application Servers and Test/Development Systems

Conventionally you stop and restart the application server instances manually after the central instance is restarted. Sun Cluster HA for SAP provides hooks that are called whenever the central instance logical host switches over or fails over. These hooks are provided by calling the `hasap_stop_all_instances` and `hasap_start_all_instances` scripts. The scripts must be idempotent.

If you configure application servers and want to control them automatically when the logical host switches over or fails over, you can create start and stop scripts according to your needs. Sun Cluster provides sample scripts that can be copied and customized:
`/opt/SUNWcluster/ha/sap/hasap_stop_all_instances.sample` and `/opt/SUNWcluster/ha/sap/hasap_start_all_instances.sample`. Customization examples are included in these scripts. Copy the sample scripts, rename them by removing the "`.sample`" suffix, and modify them as appropriate.

After failovers, Sun Cluster HA for SAP will invoke the customized scripts to restart the application servers. The scripts control the application servers from the central instance, and are invoked by the full path name.

If you include a test or development system in your configuration, modify the `hasap_stop_all_instances` script to stop the test or development system in case of failover of the central instance logical host.

During a central instance logical host switchover or failover, the scripts are called in the following sequence:

1. Stopping the application server instances and test or development systems by calling `hasap_stop_all_instances`
2. Stopping the central instance
3. Switching over the logical host(s) and disk group(s)
4. Calling `hasap_stop_all_instances` again to make sure all application servers and test or development systems have stopped
5. Starting the central instance

6. Starting the application server instances by calling `hasap_start_all_instances`

See the `hasap_start_all_instances(1M)` and `hasap_start_all_instances(1M)` man pages for more information

Additionally, you must enable root access to the SAP administrative account (`<sapsid>adm`) on all SAP application servers and test or development systems from all logical hosts and all physical hosts in the cluster. For test or development systems, also grant root access to the database administrative account (`ora<sapsid>`). Accomplish this by creating `.rhosts` files for these users. For example:

```
...
 phys-hahost1  root
 phys-hahost2  root
 phys-hahost3  root
 hahost1         root
 hahost2         root
 hahost3         root
 ...
```

In configurations including several application servers or a test or development system, consider increasing the timeout value of the `STOP_NET` method for Sun Cluster HA for SAP. Increasing the `STOP_NET` timeout value is necessary only if the `hasap_stop_all_instances` script takes longer to execute than 60 seconds, because 60 seconds is the default timeout value for the `STOP_NET` method. If the `hasap_stop_all_instances` script does not finish within the 60 second timeout, then increase the `STOP_NET` timeout value.

Check the timeout value of the `STOP_NET` method by using the following command:

```
# hareg -q sap -T STOP_NET
```

Note - The `hasap_dbms` command can be used only when Sun Cluster HA for SAP is registered but is in the `off` state. Run the command on only one node, while that node is a member of the cluster. See the `hasap_dbms(1M)` man page for more information.

Caution - If the `hasap_dbms(1M)` command returns an error stating that it cannot add rows to or update the CCD, it might be because another cluster utility is also trying to update the CCD. If this occurs, re-run `hasap_dbms(1M)` until it runs successfully. After the `hasap_dbms(1M)` command runs successfully, verify that all necessary rows are included in the resulting CCD by running the command `hareg -q sap`.

If the `hareg(1M)` command returns an error, then first restore the original method timeouts by running the command `hasap_dbms -f`. Second, restore the default dependencies by running the command `hasap_dbms -r`. After both commands complete successfully, retry the original `hasap_dbms(1M)` command to configure new dependencies and method timeouts. See the `hasap_dbms(1M)` man page for more information.

Increase the `STOP_NET` timeout value by using the following command:

```
# /opt/SUNWcluster/ha/sap/hasap_dbms -t STOP_NET=new_timeout_value
```

If you increase the `STOP_NET` method timeout value, you also must increase the timeouts that the Sun Cluster framework uses when remastering logical hosts during cluster reconfiguration. Use the `scconf(1M)` command to increase logical host timeout values. Refer to the section on configuring timeouts for cluster transition steps in the *Sun Cluster 2.2 System Administration Guide* for details about how to increase the timeouts for the cluster framework. Make sure that the `loghost_timeout` value is at least double the new `STOP_NET` timeout value.

10.2.5 Sun Cluster HA for NFS Considerations

If you have application servers outside the cluster, you must configure Sun Cluster HA for NFS on the central instance logical host. Application servers outside the cluster must NFS-mount the SAP profile directories and executable directories from the SAP central instance. See Chapter 11, for detailed procedures on setting up Sun Cluster HA for NFS, and note the following SAP-specific guidelines:

- Do not configure any node to be an NFS client of another node within the same cluster.
- If you will run application servers within the cluster, you must set up an external cluster running NFS. The application servers and central instance will mount files from this NFS cluster.
- There are start order dependencies among Sun Cluster HA for NFS, HA-DBMS, and Sun Cluster HA for SAP data services. You can use special scripts to manage these dependencies. See Section 10.7 “Setting Data Service Dependencies for SAP” on page 10-40, for more information.
- Usually, you should share the following directories to all SAP instances:

- /usr/sap/trans
- /sapmnt/<SAPSID>/exe
- /sapmnt/<SAPSID>/global
- /sapmnt/<SAPSID>/profile

10.3 Overview of Procedures

Table 10–2 summarizes the tasks you must complete to configure SAP.

TABLE 10–2 High-Level Steps to Install and Configure SAP

Task	Description	For Instructions, Go To...
Plan the SAP installation	- Read through all guidelines and procedures	Section 10.1 “Sun Cluster HA for SAP Overview” on page 10-2 and Section 10.2 “Configuration Guidelines for Sun Cluster HA for SAP” on page 10-3
	- Complete the SAP installation worksheet	Section 10.3.1 “Installation Worksheet for Sun Cluster HA for SAP” on page 10-16
Prepare the environment for SAP	- Perform all pre-requisite installation tasks - Set up Solaris - Set up the volume manager - Create disk groups or disksets - Create volumes and file systems - Install Sun Cluster - Set up PNM - Set up logical hosts and mount points - Set up HA-NFS, if necessary - Adjust kernel parameters - Create swap space - Create user and group accounts	Section 10.4 “Preparing the SAP Environment” on page 10-17 See also: Chapter 3,” Appendix B, and Appendix C

TABLE 10–2 High-Level Steps to Install and Configure SAP *(continued)*

Task	Description	For Instructions, Go To...
Install and configure SAP and the database	- Install the SAP central instance and database instance - Load the database - Load all reports - Install the GUI	"How to Install SAP and the Database" on page 10–22
Enable SAP to run in the cluster	- Set up the SAP central instance admin environment - Modify SAP profile files - Modify the database environment - Update `/etc/services` and create `/usr/sap/tmp` - Test the SAP installation	"How to Enable SAP to Run in the Cluster" on page 10–23
Configure the HA-DBMS	- Shut down SAP and the database - Adjust Oracle alert files and listener files - Register and activate the database - Set up the database instance - Start fault monitoring for the database - Test switchover of the database	"How to Enable SAP to Run in the Cluster" on page 10–23

TABLE 10–2 High-Level Steps to Install and Configure SAP *(continued)*

Task	Description	For Instructions, Go To...
Configure Sun Cluster HA for SAP	- Install and register Sun Cluster HA for SAP	"How to Configure Sun Cluster HA for SAP" on page 10–36
	- Configure Sun Cluster HA for SAP	"How to Configure Sun Cluster HA for SAP" on page 10–36, and Section 10.6.1 "Configuration Parameters for Sun Cluster HA for SAP" on page 10-38
	- Set dependencies, if necessary	Section 10.7 "Setting Data Service Dependencies for SAP" on page 10-40
	- Test switchover of Sun Cluster HA for SAP	"How to Configure Sun Cluster HA for SAP" on page 10–36
	- Customize and test start and stop scripts for the application servers and test / development systems	Section 10.2.4 "Configuration Options for Application Servers and Test / Development Systems" on page 10-11

10.3.1 Installation Worksheet for Sun Cluster HA for SAP

Complete the following worksheet before beginning the Sun Cluster HA for SAP installation.

TABLE 10–3 Installation Worksheet for Sun Cluster HA for SAP

Name of Cluster	
Number of logical hosts	
Name and IP address of all physical hosts that are potential masters of the CI logical host	
Name and IP address of CI logical host	
SAP system ID (<SAPSID>)	

TABLE 10–3 Installation Worksheet for Sun Cluster HA for SAP *(continued)*

SAP system number	
Name and IP address of all physical hosts that are potential masters of the DB logical host	
Name and IP address of DB logical host (In asymmetric configurations, this is identical to the CI logical host)	
Name of NFS logical host (If all application servers are external to cluster, this name is the central instance logical host. If the application servers are inside the cluster, this name is the logical host that provides NFS service from the external NFS cluster.) See Section 10.2.5 “Sun Cluster HA for NFS Considerations” on page 10-13.”	
SAP license for each potential master of the CI logical host	

10.4 Preparing the SAP Environment

Before beginning the SAP or Sun Cluster HA for SAP installation procedures, perform the following prerequisite tasks.

- Load the Solaris operating environment and any required Solaris patches.
- Install Volume Manager software and any required Volume Manager patches.
- Create Solstice DiskSuite disksets or SSVM disk groups (separate disk groups for the central instance and database instance are recommended).
- Create volumes according to Sun Cluster guidelines:
  - Mirror volumes across controllers
  - With SSVM, use Dirty Region Logging for faster mirror resynchronization
  - Use a logging file system for faster logical host failover

  Use Table 10–4 as a worksheet to capture the name of the volume that corresponds to each file system used for the SAP central instance. Refer to the SAP installation guide for the file system sizes recommended for your particular configuration. These are database-independent file systems.

TABLE 10–4 Worksheet: File Systems and Volume Names for the SAP Central Instance

File System Name / Mount Point	Volume Name
/usr/sap/trans	
/sapmnt/<SAPSID>	
/usr/sap/<SAPSID>	

Use Table 10–5 as a worksheet to capture the name of the volume that corresponds to each file system used for the database instance. Refer to the SAP installation guide for the file system sizes recommended for your particular configuration. These are database-dependent file systems.

TABLE 10–5 Worksheet: File Systems and Volume Names for the SAP Database Instance

File System Name / Mount Point	Volume Name
/oracle/<SAPSID>	
/oracle/stage/stage_<version>	
/oracle/<SAPSID>/origlogA	
/oracle/<SAPSID>/origlogB	
/oracle/<SAPSID>/mirrlogA	
/oracle/<SAPSID>/mirrlogB	
/oracle/<SAPSID>/saparch	
/oracle/<SAPSID>/sapreorg	
/oracle/<SAPSID>/sapdata1	
/oracle/<SAPSID>/sapdata2	

TABLE 10–5 Worksheet: File Systems and Volume Names for the SAP Database Instance *(continued)*

File System Name / Mount Point	Volume Name
/oracle/<SAPSID>/sapdata3	
/oracle/<SAPSID>/sapdata4	
/oracle/<SAPSID>/sapdata5	
/oracle/<SAPSID>/sapdata6	

- Install Sun Cluster, Sun Cluster HA for SAP, Sun Cluster HA for Oracle, and any required patches. Use the procedures described in Chapter 3, but do not set up logical hosts with `scinstall(1M)` during this installation. Instead, set up logical hosts with `scconf(1M)` after the cluster is up. Set up two disksets per logical host.
- Configure PNM on all nodes.
- Start the cluster.
- (SSVM only) Verify that all disk groups are deported.
- (Solstice DiskSuite) Release ownership of all disksets.
- Create logical hosts with `scconf(1M)`; the number of logical host depends on your particular configuration. See the section on adding and removing logical hosts in the *Sun Cluster 2.2 System Administration Guide*. You will need:
  - logical host name(s)
  - physical host names of potential masters of logical host(s)
  - names of the primary public network controllers for the potential masters of the logical host(s)
  - disk group name(s)

  When you create logical hosts, disable the automatic failback mechanism by using the `–m` option to `scconf(1M)`.
- (SSVM, two-node configurations only) Configure the shared CCD.
- After creating the logical host(s), create the logical host administrative file system. For detailed procedures, see Appendix B, or Appendix C.
- As part of the logical host configuration, create mount points for the central instance and database instance volumes, and enter them into the respective `vfstab.`*logicalhost* files on all potential masters of each logical host. These files are located in `/etc/opt/SUNWcluster/conf/hanfs`.

Table 10–6 lists the suggested file system mount points for the disk groups (SSVM) or disksets (Solstice DiskSuite) associated with the central instance and database instance. Note that separating the central instance and database instance file systems into separate disk groups or disksets (even if using a single logical host) may provide more configuration flexibility in the future.

TABLE 10–6 File Systems and Mount Points for the SAP Central Instance and Database Instance

Disk Group (SSVM)	Diskset (Solstice DiskSuite)	Volume Name	Mount Point
ci_dg	*CIloghost*	sap	/usr/sap/<SAPSID>
ci_dg	*CIloghost*	saptrans	/usr/sap/trans
ci_dg	*CIloghost*	sapmnt	/sapmnt/<SAPSID>
db_dg	*DBloghost*	oracle	/oracle/<SAPSID>
db_dg	*DBloghost*	stage	/oracle/stage/stage_<version>
db_dg	*DBloghost*	origlogA	/oracle/<SAPSID>/origlogA
db_dg	*DBloghost*	origlogB	/oracle/<SAPSID>/origlogB
db_dg	*DBloghost*	mirrlogA	/oracle/<SAPSID>/mirrlogA
db_dg	*DBloghost*	mirrlogB	/oracle/<SAPSID>/mirrlogB
db_dg	*DBloghost*	saparch	/oracle/<SAPSID>/saparch
db_dg	*DBloghost*	sapreorg	/oracle/<SAPSID>/sapreorg
db_dg	*DBloghost*	sapdata1	/oracle/<SAPSID>/sapdata1
db_dg	*DBloghost*	sapdata2	/oracle/<SAPSID>/sapdata2
db_dg	*DBloghost*	sapdata3	/oracle/<SAPSID>/sapdata3

TABLE 10–6 File Systems and Mount Points for the SAP Central Instance and Database Instance *(continued)*

Disk Group (SSVM)	Diskset (Solstice DiskSuite)	Volume Name	Mount Point
db_dg	*DBloghost*	sapdata4	/oracle/<SAPSID>/sapdata4
db_dg	*DBloghost*	sapdata5	/oracle/<SAPSID>/sapdata5
db_dg	*DBloghost*	sapdata6	/oracle/<SAPSID>/sapdata6

- If SAP application servers will be configured outside the cluster, then configure Sun Cluster HA for NFS and enter the appropriate shared file systems into the `dfstab.`*logicalhost* files on all potential masters of each logical host. These files are located in `/etc/opt/SUNWcluster/conf/hanfs`. See Section 10.2.4 "Configuration Options for Application Servers and Test/Development Systems" on page 10-11, and Chapter 11, for more information.

  Share the following file systems to SAP application servers outside the cluster. These are general guidelines. See the SAP documentation for more information.

TABLE 10–7 File Systems to Share in HA-NFS to External SAP Application Servers

File Systems to Share to External Application Servers
/usr/sap/trans
/sapmnt/<SAPSID>/exe
/sapmnt/<SAPSID>/profile
/sapmnt/<SAPSID>/global

- Test the functionality and mount points of the logical host(s) by switching them between all potential masters. This verifies that all mount points have been created correctly.
- Adjust kernel parameters on all potential masters, as per the "R/3 Installation on UNIX: OS Dependencies" guidelines in the SAP documentation. In configurations

where the central instance and database instance may coexist with each other or with other instances, be sure to size the kernel parameters accordingly.

- Create appropriately sized permanent swap areas on all potential master nodes. See the Installation Requirements Checklist in your SAP documentation for swap guidelines. Use the SAP-supplied `memlimits` utility to assist you in sizing the swap space. See the "R/3 Installation on UNIX" guidelines in the SAP documentation for more information on this utility.
- Stop the cluster and reboot all nodes after adjusting kernel parameters and swap space.
- Create SAP and database user and group accounts on all potential masters of the logical hosts. Refer to the "R/3 Installation on UNIX: OS Dependencies" guidelines in the SAP documentation for details. User and group IDs must be identical on all nodes. Create the home directories for these users on the shared diskset. Table 10–8 shows suggested home directory paths for the user accounts.

TABLE 10–8 Home Directory Paths for SAP User Accounts

User	Home directory
<sapsid>adm	/usr/sap/<SAPSID>/home
ora<sapsid>	/oracle/<SAPSID>

Note - For SAP 4.0b, read OSS note 0100125 for special steps required when creating user home directories outside of the `/home` location.

After all of these prerequisites have been fulfilled, proceed to Section 10.5 "Installing and Configuring SAP and the Database" on page 10-22.

10.5 Installing and Configuring SAP and the Database

This section describes how to install and configure SAP.

▼ How to Install SAP and the Database

1. **Verify that you have completed all tasks listed in Section 10.4 "Preparing the SAP Environment" on page 10-17.**

2. **Verify that all nodes are running in the cluster.**

3. **Switch over all logical hosts to the node from which you will install SAP and the database.**

```
# scadmin switch clustername phys-hahost1 CIloghost DBloghost ...
```

4. **Create the SAP installation directory and begin SAP installation.**

   Refer to the "R/3 Installation on UNIX" guidelines in the SAP documentation for details.

Note - Read all SAP OSS notes prior to beginning the SAP installation.

---

   a. **Install the central instance and database instance on the node currently mastering the central instance and database instance logical host.**

      (For SAP 3.1x only) When installing SAP using `R3INST`, specify the physical host name of the current master of the database logical host when prompted for "Database Server." After the installation is complete, you must manually adjust various files to refer to the logical host where the database resides.

      (For SAP 4.0x only) When installing SAP using `R3SETUP`, select the `CENTRDB.SH` script to generate the installation command file.

   b. **Continue the SAP installation to install the central instance, to create and load the database, to load all reports, and to install the R/3 Frontend (GUI).**

▼ How to Enable SAP to Run in the Cluster

1. **Set up the SAP central instance administrative environment.**

   During SAP installation, SAP creates files and shell scripts on the server on which the SAP central instance is installed. These files and scripts use physical host names. Follow these steps to replace all occurrences of physical host names with logical host names.

Note - Make backup copies of all files before performing the following steps.

   First, shut down the SAP central instance and database using the following command:

```
# su - <sapsid>adm
$ stopsap all
...
 # su - ora<sapsid>
$ lsnrctl stop
```

Note - Become the `<sapsid>adm` user before editing these files.

a. Revise the `.cshrc` file in the `<sapsid>adm` home directory.

On the server on which the SAP central instance is installed, the `.cshrc` file contains aliases that use Sun Cluster physical host names. Replace the physical host names with the central instance logical host name.

(For SAP 3.1x only) The resulting `.cshrc` file should look similar to the following example, in which *CIloghost* is the logical host containing the central instance and *DBloghost* is the logical host containing the database. If the central instance and database are on the same logical host, then use that logical host name for the substitutions.

```
# aliases
 alias   startsap     ''$HOME/startsap_CIloghost_00''
 alias   stopsap     ''$HOME/stopsap_CIloghost_00''

 # RDBMS environment
 if (-e $HOME/.dbenv_DBloghost.csh) then
    source $HOME/.dbenv_DBloghost.csh
 else if (-e $HOME/.dbenv.csh) then
    source $HOME/.dbenv.csh
 endif
```

(For SAP 4.0x only) The resulting `.cshrc` file should look similar to the following example, in which *CIloghost* is the logical host containing the central instance and *DBloghost* is the logical host containing the database. If the central instance and database are on the same logical host, then use that logical host name for the substitutions:

```
if ( -e $HOME/.sapenv_CIloghost.csh ) then
    source $HOME/.sapenv_CIloghost.csh
 else if ( -e $HOME/.sapenv.csh ) then
    source $HOME/.sapenv.csh
 endif

 # RDBMS environment
 if ( -e $HOME/.dbenv_DBloghost.csh ) then
    source $HOME/.dbenv_DBloghost.csh
 else if ( -e $HOME/.dbenv.csh ) then
    source $HOME/.dbenv.csh
 endif
```

b. (For SAP 4.0x only) Rename the file `.sapenv_`***physicalhost***`.csh` **to** `.sapenv_`***CIloghost***`.csh`**, and edit it to replace occurrences of the physical host name with the logical host name.**

First rename the file, replacing the physical host name with the central instance logical host name.

```
$ mv .sapenv_physicalhost.csh .sapenv_CIloghost.csh
```

Then edit the aliases in the file. For example:

```
alias startsap "$HOME/startsap_CIloghost_00"
 alias stopsap "$HOME/stopsap_CIloghost_00"
```

c. Rename the `.dbenv_`***physicalhost***`.csh` **file.**

Rename the `.dbenv_`*physicalhost*`.csh` file to `.dbenv_`*DBloghost*`.csh`. If the central instance and database are on the same logical host, use that logical host name for the substitution.

```
$ mv .dbenv_physicalhost.csh .dbenv_DBloghost.csh
```

d. (For SAP 4.0x only) Edit the `.dbenv_`***DBloghost***`.csh` **file to set the** `ORA_NLS` **environment variable to point to the appropriate subdirectories of** `/var/opt/oracle` **for the database client configuration files. Also, set the** `TNS_ADMIN` **environment variable to point to the** `/var/opt/oracle` **directory.**

The `.dbenv_`*DBloghost*`.csh` file is located in the `<sapsid>adm` home directory.

```
#setenv ORA_NLS /oracle/<SAPSID>/ocommon/NLS_723/admin/data
 setenv ORA_NLS /var/opt/oracle/ocommon/NLS_723/admin/data

 #setenv ORA_NLS32 /oracle/<SAPSID>/ocommon/NLS_733/admin/data
 setenv ORA_NLS32 /var/opt/oracle/ocommon/NLS_733/admin/data

 #setenv ORA_NLS33 /oracle/<SAPSID>/ocommon/NLS_804/admin/data
 setenv ORA_NLS33 /var/opt/oracle/ocommon/NLS_804/admin/data

 ...

 # setenv TNS_ADMIN @TNS_ADMIN@
 setenv TNS_ADMIN /var/opt/oracle
 ...
```

e. Rename and revise the SAP instance `startsap` and `stopsap` shell scripts in the `<sapsid>adm` home directory.

On the server on which the SAP central instance is installed, the `<sapsid>adm` home directory contains shell scripts that include physical host names. Rename these shell scripts by replacing the physical host names with logical host names. In this example, *CIloghost* represents the logical host name of the central instance:

```
$ mv startsap_physicalhost_00 startsap_CIloghost_00
$ mv stopsap_physicalhost_00 stopsap_CIloghost_00
```

The `startsap_`*CIloghost*`_00` and `stopsap_`*CIloghost*`_00` shell scripts specify physical host names in their `START_PROFILE` parameters. Replace the physical host name with the central instance logical host name in the `START_PROFILE` parameters in both files.

```
...
 START_PROFILE=''START_DVEBMGS00_CIloghost''
 ...
```

f. Revise the SAP central instance profile files.

During SAP installation, SAP creates three profile files on the server on which the SAP central instance is installed. These files use physical host names. Use these steps to replace all occurrences of physical host names with logical host names. To revise these files, you must be user <sapsid>adm, and you must be in the profile directory.

- Rename the START_DVEBMGS00_*physicalhost* and <SAPSID>_DVEBMGS00_*physicalhost* profile files.

  In the /sapmnt/<SAPSID>/profile directory, replace the physical host name with the logical host name. In this example, the <SAPSID> is HA1:

```
$ cdpro; pwd
/sapmnt/HA1/profile
 $ mv START_DVEBMGS00_physicalhost START_DVEBMGS00_CIloghost
$ mv HA1_DVEBMGS00_physicalhost HA1_DVEBMGS00_CIloghost
```

- In the START_DVEBMGS00_*CIloghost* profile file, replace occurrences of the physical host name with the central instance logical host name for all 'pf=' arguments.

  In this example, the <SAPSID> is HA1:

```
...
 Execute_00 =local $(DIR_EXECUTABLE)/sapmscsa -n \
 pf=$(DIR_PROFILE)/HA1_DVEBMGS00_CIloghost
Start_Program_01   =local $(_MS) pf=$(DIR_PROFILE)/HA1_DVEBMGS00_CIloghost
Start_Program_02   =local $(_DW) pf=$(DIR_PROFILE)/HA1_DVEBMGS00_CIloghost
Start_Program_03   =local $(_CO) -F pf=$(DIR_PROFILE)/HA1_DVEBMGS00_CIloghost
Start_Program_04   =local $(_SE) -F pf=$(DIR_PROFILE)/HA1_DVEBMGS00_CIloghost
...
```

- Edit the <SAPSID>_DVEBMGS00_*CIloghost* file to add a new entry for the SAPLOCALHOST parameter.

  Add this entry only for the central instance profile. Set the SAPLOCALHOST parameter to be the central instance logical host name. This parameter

allows external application servers to locate the central instance by using the logical host name.

```
...
 SAPLOCALHOST                =CIloghost
...
```

- Edit the DEFAULT.PFL file to replace occurrences of the physical host name with the logical host name.

  For each of the rdisp/ parameters, replace the physical host name with the central instance logical host name. For the SAPDBHOST parameter, enter the logical host name of the database. If the central instance and database are installed on the same logical host, enter the central instance logical host name. If the database is installed on a different logical host, use the database logical host name instead. In this example, *CIloghost* represents the logical host name of the central instance, *DBloghost* represents the logical host name of the database, and HA1 is the <SAPSID>:

```
...
 SAPDBHOST                  =DBloghost
rdisp/mshost                    =CIloghost
rdisp/sna_gateway              =CIloghost
rdisp/vbname                    =CIloghost_HA1_00
 rdisp/enqname                     =CIloghost_HA1_00
 rdisp/btcname                     =CIloghost_HA1_00
 ...
```

g. Revise the TPPARAM **transport configuration file.**

Change to the directory containing the transport configuration file.

```
# cd /usr/sap/trans/bin
```

Replace the database physical host name with the database logical host name. In this example, *DBloghost* represents the database logical host name and HA1 is the <SAPSID>. For example:

```
...
 HA1/dbhost = DBloghost
...
```

h. (For SAP 4.0*x* only) In the `TPPARAM` file, also set `/var/opt/oracle` to be the location for the database client configuration files.

```
...
 HA1/dbconfpath = /var/opt/oracle  ...
```

2. Modify the environment for the SAP database user.

During SAP installation, SAP creates Oracle files that use Sun Cluster physical host names. Replace the physical host names with logical host names using the following steps.

Note - Become the `ora<sapsid>` user before editing these files.

---

a. Revise the `.cshrc` file in the `ora<sapsid>` home directory.

The `.cshrc` file on the server in which SAP was installed contains aliases that use Sun Cluster physical host names. Replace the physical host names with logical host names.

(For SAP 3.1x only) The resulting file should look similar to the following example, in which *CIloghost* represents the central instance logical host and *DBloghost* is the database logical host. If the central instance and database reside on the same logical host, use the central instance logical host name for each of the substitutions:

```
# aliases
 alias   startsap     ''$HOME/startsap_CIloghost_00''
 alias   stopsap     ''$HOME/stopsap_CIloghost_00''

 # RDBMS environment
 if (-e $HOME/.dbenv_DBloghost.csh) then
    source $HOME/.dbenv_DBloghost.csh
 else if (-e $HOME/.dbenv.csh) then
    source $HOME/.dbenv.csh
 endif
```

(For SAP 4.0x only) The resulting `.cshrc` file should look similar to the following example, in which *CIloghost* is the central instance logical host and

DBloghost is the database logical host. If the central instance and database reside on the same logical host, use the central instance logical host name for each of the substitutions:

```
if ( -e $HOME/.sapenv_CIloghost.csh ) then
    source $HOME/.sapenv_CIloghost.csh
 else if ( -e $HOME/.sapenv.csh ) then
    source $HOME/.sapenv.csh
 endif

 # RDBMS environment
 if ( -e $HOME/.dbenv_DBloghost.csh ) then
    source $HOME/.dbenv_DBloghost.csh
 else if ( -e $HOME/.dbenv.csh ) then
    source $HOME/.dbenv.csh
 endif
```

b. (For SAP 4.0x only) Rename the `.sapenv_`***physicalhost***`.csh` **to** `.sapenv_`***CIloghost***`.csh`**.**

In this example, *CIloghost* represents the central instance logical host name.

```
$ mv .sapenv_physicalhost.csh .sapenv_CIloghost.csh
```

c. Rename the `.dbenv_`***physicalhost***`.csh` **file.**

Replace the physical host name with the database logical host name in the `.dbenv_`*physicalhost*`.csh` file name. If the central instance and database are on the same logical host, use the central instance logical host name for the substitution. In this example, *DBloghost* represents the database logical host:

```
$ mv .dbenv_physicalhost.csh .dbenv_DBloghost.csh
```

d. (For SAP 4.0*x* only) Edit the `.dbenv_`***DBloghost***`.csh` **file to set the** `ORA_NLS` **environment variable to point to the appropriate subdirectories of** `/var/opt/oracle` **for the database client configuration files. Also, set the** `TNS_ADMIN` **environment variable to point to the** `/var/opt/oracle` **directory.**

The `.dbenv_`*DBloghost*`.csh` file is located in the `ora<sapsid>` home directory.

```
#setenv ORA_NLS /oracle/<SAPSID>/ocommon/NLS_723/admin/data
 setenv ORA_NLS /var/opt/oracle/ocommon/NLS_723/admin/data

 #setenv ORA_NLS32 /oracle/<SAPSID>/ocommon/NLS_733/admin/data
 setenv ORA_NLS32 /var/opt/oracle/ocommon/NLS_733/admin/data

 #setenv ORA_NLS33 /oracle/<SAPSID>/ocommon/NLS_804/admin/data
 setenv ORA_NLS33 /var/opt/oracle/ocommon/NLS_804/admin/data

 ...

 # setenv TNS_ADMIN @TNS_ADMIN@
 setenv TNS_ADMIN /var/opt/oracle
 ...
```

3. **Edit the Oracle SQL*Net configuration files to replace occurrences of the physical host name with the database logical host name.**

   If the central instance and database instance are on the same logical host, use the central instance logical host name for the substitutions.

4. **Make the SQL*Net configuration files locally accessible on every potential master.**

   Use the following steps to accomplish this.

   a. **Replace all occurrences of physical host names with the database logical host name in the** `listener.ora` **and** `tnsnames.ora` **files.**

      (For SAP 3.1x only) The `listener.ora` file is located at `/etc/listener.ora`. The `tnsnames.ora` file is located at `/usr/sap/trans/tnsnames.ora`.

      (For SAP 4.0x only) The `listener.ora` file is located at `/oracle/<SAPSID>/network/admin/listener.ora`. The `tnsnames.ora` file is located at `/oracle/<SAPSID>/network/admin/tnsnames.ora`.

   b. **Relocate the SQL*Net configuration files on the node where the database is installed.**

      (For SAP 3.1x only) During installation, SAP places the `listener.ora` file in the local `/etc` directory of the node where the installation took place, and creates a soft link in `/usr/sap/trans`. Move the `listener.ora` file to `/var/opt/oracle`. Reset soft links in `/usr/sap/trans` to point to the new location. Move the `tnsnames.ora` and `sqlnet.ora` files to the `/var/opt/oracle` directory.

```
$ su
# mv /etc/listener.ora /var/opt/oracle
# rm /usr/sap/trans/listener.ora
# ln -s /var/opt/oracle/listener.ora /usr/sap/trans
# mv /usr/sap/trans/tnsnames.ora /var/opt/oracle
# ln -s /var/opt/oracle/tnsnames.ora /usr/sap/trans
# mv /usr/sap/trans/sqlnet.ora /var/opt/oracle
# ln -s /var/opt/oracle/sqlnet.ora /usr/sap/trans
```

(For SAP 4.0x only) SAP places the `listener.ora` file in the default directory under `$ORACLE_HOME/network/admin`. Use the steps below to move the `listener.ora` file to `/var/opt/oracle`, and re-set soft links in the original directory to point to the new location. Move all other SQL*Net files to the new location and re-set links to point to the new location.

```
$ su
# mv /oracle/<SAPSID>/network/admin/listener.ora /var/opt/oracle
# ln -s /var/opt/oracle/listener.ora /oracle/<SAPSID>/network/admin
# mv /oracle/<SAPSID>/network/admin/tnsnames.ora /var/opt/oracle
# ln -s /var/opt/oracle/tnsnames.ora /oracle/<SAPSID>/network/admin
# mv /oracle/<SAPSID>/network/admin/sqlnet.ora /var/opt/oracle
# ln -s /var/opt/oracle/sqlnet.ora /oracle/<SAPSID>/network/admin
# mv /oracle/<SAPSID>/network/admin/protocol.ora /var/opt/oracle
# ln -s /var/opt/oracle/protocol.ora /oracle/<SAPSID>/network/admin
```

c. (For SAP 4.0*x* only) Copy the Oracle client configuration files to the common `/var/opt/oracle` directory.

```
# cd /var/opt/oracle; mkdir  rdbms ocommon lib
# cd /var/opt/oracle/rdbms; cp -R /oracle/<SAPSID>/rdbms/mesg .
# cd /var/opt/oracle/ocommon; cp -R /oracle/<SAPSID>/ocommon/NLS* .
# cd /var/opt/oracle/lib; cp /oracle/<SAPSID>/lib/libclntsh.so.1.0 .
```

d. Distribute the `SQL*Net` configuration files to all potential masters of the central instance and database instance.

Copy or transfer the `SQL*Net` configuration files from the node on which the database was initially installed into the local directory `/var/opt/oracle` on all potential central instance and database masters. In this example, *physicalhost2* represents the name of the backup physical host.

```
$ su
# tar cvf - /var/opt/oracle | rsh physicalhost2 tar xvf -
```

Note - As part of the maintenance of HA-DBMS, the configuration files must be synchronized on all potential master nodes, whenever modifications are made.

5. **Update the** `/etc/services` **files on all potential masters to include the new SAP service entries.**

   The `/etc/services` files must be identical on all nodes.

6. **Create the** `/usr/sap/tmp` **directory on all nodes.**

   The `saposcol` program will rely on this directory.

7. **Test the SAP installation.**

   Test the SAP installation by manually shutting down SAP, manually switching the logical host between the potential master nodes, and then manually starting SAP on the backup node. This will verify that all kernel parameters, service port entries, file systems and mount points, and user/group permissions are properly set on all potential masters of the logical hosts.

   a. **Start the central instance and database.**

```
# su - ora<sapsid>
$ lsnrctl start
...
 # su - <sapsid>adm
$ startsap all
```

   b. **Run the GUI and verify that SAP comes up correctly.**

      In this example, the dispatcher port number is 3200.

```
# su - <sapsid>adm
$ setenv DISPLAY your_workstation:0
$ sapgui /H/CIloghost/S/3200
```

c. Verify that SAP can connect to the database.

```
# su - <sapsid>adm
$ R3trans -d
```

d. Run the `saplicense` utility to get a `CUSTOMER KEY` for the current node.

You will need a SAP license for all potential masters of the central instance logical host.

e. Stop SAP and the database.

```
# su - <sapsid>adm
$ stopsap all
...
 # su - ora<sapsid>
$ lsnrctl stop
```

8. For each remaining node that is a potential master of the central instance logical host, switch the central instance logical host to that node and repeat the test sequence described in Step 7 on page @–33.

```
# scadmin switch clustername phys-hahost2 CIloghost
```

▼ How to Configure the HA-DBMS

1. Shut down SAP and the database.

```
# su - <sapsid>adm
$ stopsap all
...
 # su - ora<sapsid>
$ lsnrctl stop
```

2. **(For SAP 3.1*x* only) Adjust the Oracle alert file parameter in the** `init<SAPSID>.ora` **file.**

   By default, SAP uses the prefix "`?/...`" in the `init<SAPSID>.ora` file to denote the relative path from `$ORACLE_HOME`. The Sun Cluster fault monitors cannot parse the prefix, but instead require the full path name to the alert file. Therefore, you must edit the `/oracle/<SAPSID>/dbs/init<SAPSID>.ora` file and define the dump destination parameters as follows:

```
background_dump_dest = /oracle/<SAPSID>/saptrace/background
```

3. **Register and activate the database.**

   Run the `hareg(1M)` command from only one node. For example, for Oracle:

```
# hareg -s -r oracle -h DBloghost
 # hareg -y oracle
```

4. **Set up the database instance.**

   See Chapter 5, for more information.

   For example, for Oracle:

```
# haoracle insert <SAPSID> DBloghost 60 10 120 300 \
  user/password /oracle/<SAPSID>/dbs/init<SAPSID>.ora LISTENER
```

5. **Start fault monitoring for the database instance.**

   For example:

```
# haoracle start <SAPSID>
```

6. **Test switchover of the HA-DBMS.**

   For example:

```
# scadmin switch clustername phys-hahost2 DBloghost
```

10.6 Configuring Sun Cluster HA for SAP

This section describes how to register and configure Sun Cluster HA for SAP.

▼ How to Configure Sun Cluster HA for SAP

1. **If Sun Cluster HA for SAP has not yet been installed, install it now by running `scinstall(1M)` on all nodes and adding the Sun Cluster HA for SAP data service.**

   See Section 3.2 “Installation Procedures” on page 3-2, for details. If the cluster is already running, you must stop it before installing the data service.

2. **Register the Sun Cluster HA for SAP data service by running the `hareg(1M)` command.**

   Run this command on only one node:

```
# hareg -s -r sap -h CIloghost
```

3. **Verify that all nodes are running in the cluster.**

4. **Create a new Sun Cluster HA for SAP instance using the `hadsconfig(1M)` command.**

   The `hadsconfig(1M)` command is used to create, edit, and delete instances of the Sun Cluster HA for SAP data service. The configuration parameters are described in Section 10.6.1 “Configuration Parameters for Sun Cluster HA for SAP” on page 10-38.

   Run this command on only one node, while all nodes are running in the cluster:

```
# hadsconfig
```

5. **If Sun Cluster HA for SAP is dependent upon other data services within the same logical host, set dependencies on those data services.**

   See "How to Set a Data Service Dependency for SAP" on page 10–41. If you do set dependencies, start all services on which SAP depends before proceeding.

6. **Stop the central instance before starting SAP under the control of Sun Cluster HA for SAP.**

```
# su - <sapsid>adm
$ stopsap r3
```

Caution - The SAP central instance must be stopped before Sun Cluster HA for SAP is turned on.

7. **Turn on the Sun Cluster HA for SAP instance.**

```
# hareg -y sap
```

8. **Test switchover of Sun Cluster HA for SAP.**

   For example:

```
# scadmin switch clustername phys-hahost2 CIloghost
```

9. **(Optional) If you have application servers or a test/development system, customize and test the** `hasap_start_all_instances` **and** `hasap_stop_all_instances` **scripts.**

   See Section 10.2.4 "Configuration Options for Application Servers and Test/Development Systems" on page 10-11, for details. Test switchover of Sun Cluster HA for SAP and verify start and stop of application servers. Verify that the test/development system stops when the central instance logical host is switched to the test/development system physical host.

```
# scadmin switch clustername phys-hahost1 CIloghost
```

10.6.1 Configuration Parameters for Sun Cluster HA for SAP

This section describes the information you supply to `hadsconfig(1M)` to create configuration files for the Sun Cluster HA for SAP data service. The `hadsconfig(1M)` command uses templates to create these configuration files. The templates contain some default, some hard coded, and some unspecified parameters. You must provide values for all parameters that are unspecified.

The fault probe parameters, in particular, can affect the performance of Sun Cluster HA for SAP. Tuning the probe interval value too low (increasing the frequency of fault probes) might encumber system performance, and also might result in false takeovers or attempted restarts when the system is simply slow.

Configure Sun Cluster HA for SAP by supplying the `hadsconfig(1M)` command with parameters listed in Table 10–9.

TABLE 10–9 Sun Cluster HA for SAP Configuration Parameters

Name of the Instance	Nametag used internally as an identifier for the instance. The log messages generated by Sun Cluster refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. You can use the `SAPSID` for this nametag. For example, if you specify `HA1`, `hadsconfig(1M)` produces `SUNWscsap_HA1`.
Logical Host	Name of the logical host that provides service for this instance of Sun Cluster HA for SAP. This name should be the logical host name for the central instance.
Time Between Probes	The interval, in seconds, of the fault probing cycle. The default value is 60 seconds.
SAP R/3 System ID	This is the SAP system name or `<SAPSID>`.
Central Instance ID	This is the SAP system number or Instance ID. For example, the CI is normally “00.”
SAP Admin Login Name	The name used by Sun Cluster HA for SAP to log in to the SAP central instance administrative account. This name must exist on all central instance and application server hosts. This is the `<sapsid>adm`. For example, “ha1adm.”
Database Admin Login Name	This is the SAP database administrator’s account. For SAP with Oracle, this is the `ora<sapsid>`. For example, `oraha1`.
Database Logical Host Name	Name of the logical host for the database used by SAP. This might be the same as the logical host name used for the central instance, depending on your configuration.

TABLE 10–9 Sun Cluster HA for SAP Configuration Parameters *(continued)*

Log Database Warnings	Possible values are “y” or “n.” If set to “y” and the Sun Cluster HA for SAP probe detects that it cannot connect to the database during a probe cycle, a warning message appears saying the database is unavailable. For example, this occurs if the database logical host is in maintenance mode or if the database is being relocated to another node in the cluster. If the parameter is set to “n,” then no messages appear if the probe cannot connect to the database.
Central Instance Start Retry Count	This must be an integer greater than or equal to 1. This is the number of times Sun Cluster HA for SAP should attempt to start the central instance before giving up. This value is also the number of times the Sun Cluster HA for SAP fault monitor will probe in grace mode before entering normal probe mode. While in grace mode, the probe will not perform a restart or initiate a failover of the central instance if the probe detects that the central instance is not yet up. Instead, the fault monitor will report the status of all probes and will continue in grace mode until all probes pass, or until the retry count has been exhausted.
Central Instance Start Retry Interval	This is the number of seconds Sun Cluster HA for SAP should wait between each attempt to start the central instance. This value is also the number of seconds that the Sun Cluster HA for SAP fault monitor will sleep (between probe attempts) while in grace mode.
Time Allowed to Stop All Instances Before Central Instance Starts	This must be an integer greater than or equal to 0. This parameter dictates for how much time (in seconds) the `hasap_stop_all_instances` script should be run before starting the central instance. If set to 0, then `hasap_stop_all_instances` is run in the background while the central instance is being started. If set to a positive integer, then `hasap_stop_all_instances` is run for that amount of time in the foreground before the central instance is started.
Allow the Central Instance to Start if Foregrounded Stop All Instances Returns Error	This flag should be set to either “y” or “n”. This value determines whether the central instance should be started in the case where the `hasap_stop_all_instances` script returns a non-zero exit code or does not complete in the time specified by the “Time Allowed to Stop All Instances Before Central Instance Starts” parameter. If set to “n“ *and* the value for “Time Allowed to Stop All Instances Before Central Instance Starts” is greater than 0, *and* if the `hasap_stop_all_instances` script does not complete in the time configured above or the `hasap_stop_all_instances` script returns a non-zero exit status, the central instance will not be started and the fault monitors will take action based on the other configuration parameters. If set to “y,” then the central instance will be started regardless of whether `hasap_stop_all_instances` returns an error code or finishes within the timeout specified above.
Number of Central Instance Restarts on Local Node:	This must be an integer greater than or equal to 0. This dictates how many times the SAP central instance will be restarted on the local node before giving up, after a failure has been detected. When this number of restarts has been exhausted, Sun Cluster HA for SAP either issues a failover request, if permitted by the “Allow Central Instance Failover” parameter, or does nothing to correct the failure detected by the fault monitor.

TABLE 10–9 Sun Cluster HA for SAP Configuration Parameters *(continued)*

Number of Probe Successes to Reset the Restart Count	This parameter should be an integer that is greater than or equal to 0. If set to a positive integer, then after that many consecutive successful probes, the count of restarts done so far on the local node will be reset to 0. For example, if the value for "Number of Central Instance Restarts on Local Node" parameter is 1 and the value for "Number of Probe Successes to Reset the Restart Count" is 60, then after the first failure occurs, the probe will try to restart the central instance on the local node. If this restart succeeds, then after 60 successful probes, the restart count will be reset to 0, allowing the probe to do another restart if it detects another failure. If the parameter "Number of Probe Successes to Reset the Restart Count" is set to 0, then the restart count is never reset. This means that the number of restarts set in the parameter "Number of Central Instance Restarts on Local Node" is the absolute number of restarts that will be done on the local node before failing over.
Allow Central Instance Failover	Possible values are "y" or "n." If set to "y" *and* Sun Cluster HA for SAP detects an error in the SAP instance it is monitoring *and* the "Number of central instance Restarts on Local Node" has been exhausted, then Sun Cluster HA for SAP issues a request to relocate the instance's logical host to another cluster node. If this flag is set to "n," then even if an error is detected and all of the local restarts have been exhausted, Sun Cluster HA for SAP will not cause a relocation of this instance's logical host. When this occurs, the central instance is left in the its failed state, and the probe exits.

10.7 Setting Data Service Dependencies for SAP

Setting a dependency with `hasap_dbms` is only necessary to specify the order that data services are started and stopped within a single logical host. There is no mechanism for setting dependencies for data services configured on two different logical hosts.

If Sun Cluster HA for Oracle or Sun Cluster HA for NFS are configured on the same logical host as Sun Cluster HA for SAP, then you should set a dependency for Sun Cluster HA for SAP on those data services. You can use the `hasap_dbms` command to create or remove such a dependency. These dependencies affect the order that the services are started and stopped. Sun Cluster HA for Oracle and Sun Cluster HA for NFS should always be started before Sun Cluster HA for SAP is started. Similarly, Sun Cluster HA for SAP should always be stopped before the other data services are stopped.

Caution - If Sun Cluster HA for Oracle or Sun Cluster HA for NFS is *not* configured on the same logical host as Sun Cluster HA for SAP, then do not use the `hasap_dbms` command.

▼ How to Set a Data Service Dependency for SAP

To set a data service dependency, issue *one* of the `hasap_dbms` commands described below.

Note - The `hasap_dbms` command can be used only when Sun Cluster HA for SAP is registered but is in the `off` state. Run the command on only one node, while that node is a member of the cluster. See the `hasap_dbms(1M)` man page for more information.

Caution - If the `hasap_dbms(1M)` command returns an error stating that it cannot add rows to or update the CCD, it might be because another cluster utility is also trying to update the CCD. If this occurs, re-run `hasap_dbms(1M)` until it runs successfully. After the `hasap_dbms(1M)` command runs successfully, verify that all necessary rows are included in the resulting CCD by running the command `hareg -q sap`.

If the `hareg(1M)` command returns an error, then first restore the original method timeouts by running the command `hasap_dbms -f`. Second, restore the default dependencies by running the command `hasap_dbms -r`. After both commands complete successfully, retry the original `hasap_dbms(1M)` command to configure new dependencies and method timeouts. See the `hasap_dbms(1M)` man page for more information.

1. **Set the data service dependency using *one* of the following commands.**

   If you are using only Sun Cluster HA for NFS and Sun Cluster HA for SAP on the same logical host, use the following command:

```
# /opt/SUNWcluster/ha/sap/hasap_dbms -d nfs
```

   If you are using only Sun Cluster HA for Oracle and Sun Cluster HA for SAP on the same logical host, use the following command:

```
# /opt/SUNWcluster/ha/sap/hasap_dbms -d oracle
```

   If you are using Sun Cluster HA for Oracle, Sun Cluster HA for NFS, and Sun Cluster HA for SAP on the same logical host, use the following command:

```
# /opt/SUNWcluster/ha/sap/hasap_dbms -d oracle,nfs
```

2. **Check the dependencies set for Sun Cluster HA for SAP using the following command:**

```
# hareg -q sap -D
```

▼ How to Remove a Data Service Dependency for SAP

The dependencies set for Sun Cluster HA for SAP can be removed by running the `hasap_dbms -r` command. Issuing this command causes all of the dependencies set for Sun Cluster HA for SAP to be removed.

Note - The `hasap_dbms` command can be used only when Sun Cluster HA for SAP is registered but is in the `off` state. Run the command on only one node, while that node is a member of the cluster. See the `hasap_dbms(1M)` man page for more information.

Caution - If the `hasap_dbms(1M)` command returns an error stating that it cannot add rows to or update the CCD, it might be because another cluster utility is also trying to update the CCD. If this occurs, re-run `hasap_dbms(1M)` until it runs successfully. After the `hasap_dbms(1M)` command runs successfully, verify that all necessary rows are included in the resulting CCD by running the command `hareg -q sap`.

If the `hareg(1M)` command returns an error, then first restore the original method timeouts by running the command `hasap_dbms -f`. Second, restore the default dependencies by running the command `hasap_dbms -r`. After both commands complete successfully, retry the original `hasap_dbms(1M)` command to configure new dependencies and method timeouts. See the `hasap_dbms(1M)` man page for more information.

---

1. **Remove all of the dependencies set for Sun Cluster HA for SAP, using the following command:**

```
# /opt/SUNWcluster/ha/sap/hasap_dbms -r
```

2. **Check the dependencies set for Sun Cluster HA for SAP, using the following command:**

```
# hareg -q sap -D
```

CHAPTER 11

Setting Up and Administering Sun Cluster HA for NFS

This chapter provides instructions for setting up and administering the Sun Cluster HA for NFS data service.

This chapter includes the following procedures

11.1 Sun Cluster HA for NFS Overview

This chapter describes the steps necessary to configure and run Sun Cluster HA for NFS on your Sun Cluster servers. It also describes the steps necessary to add Sun Cluster HA for NFS to a system that is already running Sun Cluster.

Before beginning the tasks in this chapter, see Chapter 2, for more information on setting up file systems. Refer to Section 2.6 “Configuration Restrictions” on page 2-27, for HA-NFS configuration restrictions.

Note - To avoid any failures due to name service lookup, all logical host names should be present in the server’s and client’s `/etc/hosts` file. Name service mapping on the servers should be configured to look first at the local files before trying to access NIS or NIS+. This prevents timing related errors in this area and ensures that `ifconfig` and `statd` do not fail in resolving logical host names.

Table 11–1 shows the high-level steps to configure Sun Cluster HA for NFS to work with Sun Cluster. The tasks should be performed in the order shown.

TABLE 11–1 High-Level Steps to Configure Sun Cluster HA for NFS

Task	Go To ...
1. Updating the name service with logical host names	Step 4 on page @–9 in the procedure “How to Install the Server Software” on page 3–6. (This should have been done prior to running the `scinstall(1M)` command.)
2. Modifying name service lookups in the `/etc/nsswitch.conf` file to access `/etc` files first	Step 7 on page @–9 in the procedure “How to Install the Server Software” on page 3–6. (This should have been done prior to running the `scinstall(1M)` command.)
3. Initializing NAFO	Step 21 on page @–15in the procedure “How to Install the Server Software” on page 3–6. (This should have been done as part of the `scinstall(1M)` process.)
4. Setting up logical hosts	in the procedure “How to Install the Server Software” on page 3–6. (This should have been done as part of the `scinstall(1M)` process.)
5. Assigning net names and disk groups	Step 25 on page @–17 in the procedure “How to Install the Server Software” on page 3–6. (This should have been done as part of the `scinstall(1M)` process.)
6. Configuring the volume manager	For Solstice DiskSuite configurations, refer to Appendix B. For Sun StorEdge Volume Manager configurations, refer to Appendix C.

TABLE 11–1 High-Level Steps to Configure Sun Cluster HA for NFS *(continued)*

Task	Go To ...
7. Creating NAFO backup groups	Step 8 on page @–26 in the procedure "How to Configure the Cluster" on page 3–22. (This should have been done as part of the installation process.)
8. Creating multihost file systems	For Solstice DiskSuite configurations, refer to Appendix B." For Sun StorEdge Volume Manager configurations, refer to Appendix C."
9. Editing the `dfstab.`*logicalhost* files	Refer to "How to Share NFS File Systems" on page 11–4.
10. Registering Sun Cluster HA for NFS	Refer to "How to Register and Activate NFS" on page 11–5.

The mount points for NFS file systems placed under the control of Sun Cluster HA for NFS must be the same on all nodes that are capable of mastering the logical host containing those file systems.

Note - To avoid "stale file handle" errors on the client during NFS failovers, make sure that the Sun StorEdge Volume Manager `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.

See Appendix C, for pseudo-device major number administrative procedures.

11.2 Sharing NFS File Systems

This section describes the procedures used to set up file systems to be shared by NFS by editing the logical host's `dfstab.`*logicalhost* files.

Note - Before you set up file systems to be shared by NFS, make sure you have configured your logical hosts. When you first configure the cluster, you provide the `scinstall(1M)` command with information about your logical host configuration. Once the cluster is up, you can configure logical hosts by running either the `scinstall(1M)` or `scconf(1M)` commands.

▼ How to Share NFS File Systems

Note - NFS file systems are not shared until you perform a cluster reconfiguration as outlined in "How to Register and Activate NFS" on page 11–5.

---

1. **Create the multihost file systems.**

   Use the procedures described in Appendix B, and in Appendix C, to create the multihost file systems.

2. **Verify that all nodes in the cluster are up and running.**

3. **From a `cconsole(1)` window, use an editor such as `vi` to create and edit the `/etc/opt/SUNWcluster/conf/hanfs/dfstab.`*logicalhost* file.**

   By using a `cconsole(1)` window, you can make changes on all the potential masters of these file systems. You can also update `dfstab.`*logicalhost* on one node and use `rcp(1)` to copy it to all other cluster nodes that are potential masters of the file systems. Add entries for all files systems created in that will be shared.

   The `dfstab.`*logicalhost* file is in `dfstab` format. The file contains `share(1M)` commands in this syntax.

```
share [-F fstype] [-o options] [-d ''<text>''] <pathname> [resource]
```

   If you use the `rw`, `rw=`, `ro`, or `ro=` options to the `share -o` command, grant access to all hostnames that Sun Cluster uses. This enables Sun Cluster HA for NFS fault monitoring to operate most efficiently. Include all physical and logical hostnames that are associated with the Sun Cluster, as well as the hostnames on all public networks to which the Sun Cluster is connected.

   If you use `netgroups` in the `share` command (rather than names of individual hosts), add all those cluster hostnames to the appropriate `netgroup`.

Note - Do not grant access to the hostnames on the private nets.

   Grant read and write access to all the hosts' hostnames, to enable the HA-NFS monitoring to do a thorough job. However, you can restrict write access to the file system, or make the file system entirely read-only. In this case, Sun Cluster HA for NFS fault monitoring will still be able to perform monitoring without having write access.

   The resulting file will look similar to this example, which shows the logical host name (`hahost1`), the file system type (`nfs`), and the mount point names (`/hahost1/1` and `/hahost1/2`).

```
share  -F nfs  -d ‘‘hahost1 fs 1’’ /hahost1/1
 share  -F nfs  -d ‘‘hahost1 fs 2’’ /hahost1/2
```

Note - When constructing share options, generally avoid using the –`root` option, and avoid mixing –`ro` and –`rw` options.

4. **(Optional) Create the file** `.probe_nfs_file` **in each directory to be NFS-shared.**

   For enhanced fault monitoring, each directory exported by Sun Cluster HA for NFS (that is, each directory listed in the `dfstab` files for Sun Cluster HA for NFS) should contain the file `.probe_nfs_file`. For each such directory, `cd` to the directory and create the file using the `touch(1)` command:

   Do this on the physical host that currently masters the logical host for that `dfstab` file.

```
phys-hahost1# touch .probe_nfs_file
```

   After completing these steps, register and activate NFS using the procedure “How to Register and Activate NFS” on page 11–5.

▼ How to Register and Activate NFS

After setting up and configuring NFS, you must activate Sun Cluster HA for NFS by using the `hareg(1M)` command to start the Sun Cluster monitor.

1. **Register Sun Cluster HA for NFS.**

   Use the `hareg(1M)` command to register the Sun Cluster HA for NFS data service on all hosts in the Sun Cluster. Run the command on only one node.

```
# hareg -s -r nfs
```

   The following command registers the Sun Cluster HA for NFS data service only on logical hosts `hahost1` and `hahost2`. Run the command on only one node.

```
# hareg -s -r nfs -h hahost1,hahost2
```

2. **Activate the NFS service by invoking the `hareg(1M)` command on one host.**

```
# hareg -y nfs
```

3. **Execute a membership reconfiguration.**

```
# haswitch -r
```

Refer to the *Sun Cluster 2.2 System Administration Guide* for more information on forcing a cluster reconfiguration.

The process of setting up, registering, and activating Sun Cluster HA for NFS on your Sun Cluster servers is now complete.

▼ How to Add NFS to a System Already Running Sun Cluster

1. **Create and edit the `/etc/opt/SUNWcluster/conf/dfstab.`*logicalhost* file.**

   Follow the instructions in "How to Share NFS File Systems" on page 11–4 to edit the `dfstab` file.

2. **Register and activate NFS.**

   Follow the instructions in "How to Register and Activate NFS" on page 11–5.

11.3 Administering NFS in Sun Cluster Systems

This section describes the procedures used to administer NFS in Sun Cluster systems.

11.3.1 Adding an Existing File System to a Logical Host

After Sun Cluster is running, use the following procedures to add an additional file system to a logical host.

Note - Use caution when manually mounting multihost disk file systems that are not listed in the Sun Cluster `vfstab.`*logicalhost* and `dfstab.`*logicalhost* files. If you forget to unmount that file system, a subsequent switchover of the logical host containing that file system will fail because the device is busy. However, if that file system is listed in the appropriate Sun Cluster `vfstab.`*logicalhost* files, the software can forcefully unmount the file system, and the volume manager disk group release commands will succeed.

▼ How to Add an Existing File System to a Logical Host

1. **From a `cconsole(1)` window, use an editor such as `vi` to add an entry for the file system to the `/etc/opt/SUNWcluster/conf/hanfs/vfstab.`*logicalhost* file.**

   By using a `cconsole(1)` window, you can make changes on all potential masters of these file systems.

2. **Run the `mount(1M)` command to mount the new file system.**

   Specify the device and mount point. Alternatively, you can wait until the next membership reconfiguration for the file system to be automatically mounted.

   Here is an example for Solstice DiskSuite.

```
# mount -F ufs /dev/md/hahost1/dsk/d2 /hahost1/2
```

   Here is an example for Sun StorEdge Volume Manager.

```
# mount -F vxfs /dev/vx/dsk/dg1/vol1 /vol1
```

3. **Add the Sun Cluster HA for NFS file system to the logical host.**

   a. **From a `cconsole(1)` window, use an editor such as `vi` to make the appropriate entry for each file system that will be shared by NFS to the `vfstab.`*logicalhost* and `dfstab.`*logicalhost* files.**

      By using a `cconsole(1)` window, you can make changes on all potential masters of these file systems.

   b. **Execute a membership reconfiguration of the servers.**

```
# haswitch -r
```

Refer to the *Sun Cluster 2.2 System Administration Guide* for more information on forcing a cluster reconfiguration.

Alternatively, the file system can be shared manually. If the procedure is performed manually, the fault monitoring processes will not be started either locally or remotely until the next membership reconfiguration is performed.

11.3.2 Removing a File System From a Logical Host

Use the following procedure to remove a file system from a logical host running Sun Cluster HA for NFS.

▼ How to Remove a File System From a Logical Host

1. **From a `cconsole(1)` window, use an editor such as `vi` to remove the entry for the file system from the `/etc/opt/SUNWcluster/conf/hanfs/vfstab`*.logicalhost* file.**

   By using a `cconsole(1)` window, you can make changes on all the potential masters of these file systems.

2. **Run the `umount(1M)` command to unmount the file system.**

3. **(Optional) Clear the associated trans device.**

   a. **If you are running Solstice DiskSuite, clear the trans metadevice and its mirrors using either the `metaclear -r` command or the `metatool(1M)` GUI.**

   b. **If you are running Sun StorEdge Volume Manager, dissociate the log subdisk from the plex.**

      Refer to your volume manager documentation for more information on clearing logging devices.

11.3.3 Adding an NFS File System to a Logical Host

Use this procedure to add an NFS file system to a logical host.

▼ How to Add an NFS File System to a Logical Host

1. **From a `cconsole(1)` window, use an editor such as `vi` to make the appropriate entry for each file system that will be shared by NFS to the `vfstab.`*logicalhost* and `dfstab.`*logicalhost* files.**

By using a `cconsole(1)` window, you can make changes on all potential masters of these file systems.

2. **Execute a membership reconfiguration of the servers.**

```
# haswitch -r
```

Refer to the *Sun Cluster 2.2 System Administration Guide* for more information on forcing a cluster reconfiguration.

Alternatively, the file system can be shared manually. If the procedure is performed manually, the fault monitoring processes will not be started either locally or remotely until the next membership reconfiguration is performed.

11.3.4 Removing an NFS File System From a Logical Host

Use this procedure to remove an Sun Cluster HA for NFS file system from a logical host.

▼ How to Remove an NFS File System From a Logical Host

1. **From a `cconsole(1)` window, use an editor such as `vi` to remove the entry for the file system from the `/etc/opt/SUNWcluster/conf/hanfs/dfstab.`*logicalhost* file.**

   By using a `cconsole(1)` window, you can make changes on all potential masters of these file systems.

2. **Run the `unshare(1M)` command.**

   The fault monitoring system will try to access the file system until the next membership reconfiguration. Errors will be logged, but a takeover of services will not be initiated by the Sun Cluster software.

3. **(Optional) Remove the file system from the logical host. If you want to retain the UFS file system for another purpose, such as a highly available DBMS file system, skip to Step 4 on page @–9.**

   To perform this task, use the procedure described in Section 11.3.2 “Removing a File System From a Logical Host” on page 11-8.

4. **Execute a membership reconfiguration of the servers.**

```
# haswitch -r
```

Refer to the *Sun Cluster 2.2 System Administration Guide* for more information on forcing a cluster reconfiguration.

11.3.5 Changing Share Options on an NFS File System

If you use the –rw, –rw=, –ro, or –ro= options to the `share –o` command, NFS fault monitoring will work best if you grant access to all the physical host names or `netgroups` associated with all Sun Cluster servers.

If you use `netgroups` in the `share(1M)` command, add all of the Sun Cluster host names to the appropriate `netgroup`. Ideally, you should grant both read and write access to all the Sun Cluster host names to enable the NFS fault probes to do a complete job.

Note - Before you change share options, read the `share_nfs(1M)` man page to understand which combinations of options are legal. When modifying the share options, execute your proposed new `share(1M)` command, interactively, as root, on the Sun Cluster server that currently masters the logical host. This will give you immediate feedback as to whether your combination of share options is legal. If the new `share(1M)` command fails, immediately execute another `share(1M)` command with the old options. After you have a new working `share(1M)` command, change the `dfstab.`*logicalhostname* file to incorporate the new `share(1M)` command.

▼ How to Change Share Options on an NFS File System

1. **From a `cconsole(1)` window, use an editor such as `vi` to make the appropriate changes to the `dfstab.`*logicalhost* files.**

   By using a `cconsole(1)` window, you can make changes on all the potential masters of these file systems.

2. **Execute a membership reconfiguration of the servers.**

```
# haswitch -r
```

   Refer to the *Sun Cluster 2.2 System Administration Guide* for more information on forcing a cluster reconfiguration.

   If a reconfiguration is not possible, you can run the `share(1M)` command with the new options. Some changes can cause the fault monitoring subsystem to issue messages. For instance, a change from read-write to read-only will generate messages.

CHAPTER 12

Setting Up and Administering Sun Cluster HA for DNS

This chapter provides instructions for setting up and administering the Sun Cluster HA for DNS data service.

This chapter includes the following procedures:

12.1 Installing DNS

Sun Cluster HA for DNS is DNS running under the control of Sun Cluster. This section describes the steps to take when installing DNS to enable it to run as the Sun Cluster HA for DNS data service.

▼ How to Install DNS

Refer to DNS documentation by using your AnswerBook software for how to set up DNS. The differences in a Sun Cluster configuration are:

- The database is located on the multihost disks rather than on the private disks.
- The DNS server is a logical host rather than a physical host.

1. Decide which logical host is to provide DNS service.

2. **Choose a location on the logical host for the DNS database.**

   Place the `named.boot` file and the rest of the files constituting the database here. For example, */logicalhost*`/dns`.

3. **In the `/etc/resolv.conf` file on all Sun Cluster servers that will be running Sun Cluster HA for DNS, specify the IP address of the logical host providing DNS service.**

4. **In the `/etc/nsswitch.conf` file, add the string `dns` after `files`.**

   Make sure the logical host name for the HA-DNS server is present in the `/etc/inet/hosts` file on all cluster nodes.

5. **Test DNS outside the Sun Cluster environment.**

   For example:

```
# cd /<logicalhost>/dns
/usr/sbin/in.named -b /<logicalhost>/dns/named.boot
 # nslookup <physicalhost>
```

Once you have installed and set up DNS, configure, register, and start Sun Cluster HA for DNS as described in Section 12.2 “Installing and Configuring Sun Cluster HA for DNS” on page 12-2.

12.2 Installing and Configuring Sun Cluster HA for DNS

This section describes the steps used to install, configure, register, and start Sun Cluster HA for DNS. You must install and set up DNS itself and Sun Cluster before performing this procedure.

You will configure Sun Cluster HA for DNS by using the `hadsconfig(1M)` command. See the `hadsconfig(1M)` man page for details.

▼ How to Install and Configure Sun Cluster HA for DNS

Once you have installed Sun Cluster and DNS, follow the procedures described in this section to install, configure, register, and activate Sun Cluster HA for DNS.

1. **On each Sun Cluster server, install the `SUNWscdns` packages in the default location.**

   If not already installed, use the `pkgadd` command to install the `SUNWscdns` package on each Sun Cluster server.

2. **Run the `hadsconfig(1M)` command to configure Sun Cluster HA for DNS.**

   The `hadsconfig(1M)` command is used to create, edit, and delete instances of a Sun Cluster HA for DNS data service. Refer to Section 12.2.1 “Configuration Parameters” on page 12-4, for information on the input you will need to supply to the `hadsconfig(1M)`command.

```
phys-hahost1# hadsconfig
```

3. **Set Up HA-DBMS for ORACLE7.**

   HA-DBMS for ORACLE7 is required to run the HA-COMMUNITY data services.

   Follow the instructions in Chapter 5.

4. **Register the Sun Cluster HA for DNS data service by running the `hareg(1M)` command.**

   If you installed Sun Cluster HA for DNS on all potential masters of a logical host but not on all hosts in the cluster, use the –h option to specify the logical host name. Run the `hareg(1M)` command on only one node.

```
phys-hahost1# hareg -s -r dns [-h logicalhost]
```

5. **Use the `hareg(1M)` command to enable the Sun Cluster HA for DNS data service and perform a cluster reconfiguration.**

   Run the `hareg(1M)` command on only one node.

```
phys-hahost1# hareg -y dns
```

   The configuration is complete.

12.2.1 Configuration Parameters

This section describes the information you supply to the `hadsconfig(1M)` command to create configuration files Sun Cluster HA for DNS. The `hadsconfig(1M)` command uses templates to create these configuration files. The templates contain some default, some hardcoded, and some unspecified parameters. Accept the default values where possible. You must provide values for all parameters that are unspecified.

The fault probe parameters, in particular, can affect the performance of Sun Cluster HA for DNS. Tuning the probe interval value too low (increasing the frequency of fault probes) might encumber system performance, and also might result in false takeovers or attempted restarts when the system is simply slow.

The Sun Cluster HA for DNS data service requires you to set the takeover flag. This flag specifies how Sun Cluster will handle partial failover. There are two options:

- `–y` (yes) – Sun Cluster will attempt to switch over the logical host to another master, but if the attempt fails the logical host will remain on the original master.
- `–n` (no) – Sun Cluster will not move the logical host to another master, even if it detects problems with the data server, nor will it take any action against the sick data server or database on the logical host.

Configure the Sun Cluster HA for DNS parameters listed in the `hadsconfig(1M)` input form by supplying options described in Table 12–1.

TABLE 12–1 Configuration Parameters for Sun Cluster HA for DNS

Parameter	Description
Name of the instance	Nametag used as an identifier for the instance. The log messages generated by Sun Cluster refer to this nametag. The `hadsconfig(1M)` command prefixes the package name to the value you supply here. For example, if you specify “nsdns_119,” `hadsconfig(1M)` produces “SUNWscdns_nsdns_119.”
Logical host	Name of logical host that provides Sun Cluster HA for DNS service.
Configuration directory	Rooted path name specifying the directory of DNS configuration files and database on multihost disk.
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the data service instance. Possible values are `–y` (yes) and `–n` (no).

TABLE 12–1 Configuration Parameters for Sun Cluster HA for DNS *(continued)*

Parameter	Description
Time between probes	The interval, in seconds, of the fault probing cycle. Accept the default value of 60 seconds.
Probe timeout	The time, in seconds, after which a fault probe will time out. The default timeout value is 60 seconds.

CHAPTER 13

Setting Up and Administering Sun Cluster HA for Lotus

This chapter provides instructions for setting up and administering Sun Cluster HA for Lotus.

This chapter includes the following procedures:

13.1 Sun Cluster HA for Lotus Overview

The Sun Cluster HA for Lotus product consists of the Lotus Domino server made highly available by running it in the Sun Cluster environment. Sun Cluster 2.2 does not support the Lotus Partitioned Server or Lotus Cluster products.

To run Sun Cluster HA for Lotus under Sun Cluster, you must:

- Complete the Lotus Domino pre-installation tasks described in your Lotus Domino documentation
- Install and configure Lotus Domino
- Configure Lotus Domino to run under Sun Cluster using the `hadsconfig(1M)` command
- Verify the Sun Cluster HA for Lotus configuration

The procedures described in this chapter assume that you are familiar with the Sun Cluster concepts of disksets, logical hosts, physical hosts, switchover, takeover, and data services.

Before you install and configure Sun Cluster HA for Lotus, you first must install and configure the Sun Cluster framework. Then use the procedures in the following sections to install and configure Sun Cluster HA for Lotus.

13.1.1 Sun Cluster HA for Lotus Installation Notes

Lotus Domino servers can be set up as HTTP, POP3, IMAP, NNTP and LDAP servers. Some restrictions exist when you include Lotus Domino servers and Sun Cluster HA for Netscape servers in the same cluster. Use the general guidelines outlined in Table 13–1 to determine which Lotus Domino server tasks to specify during installation.

TABLE 13–1 Lotus Domino Server Options - General Guidelines

Server Task	Client Types Supported	Limitations	Default Port
HTTP	Web browsers (Netscape Navigator, Microsoft Internet Explorer, etc.)	Do not install Sun Cluster HA for Netscape HTTP and the Lotus HTTP server on the same logical host, or on logical hosts mastered by the same physical host.	80
IMAP	Internet mail clients using Post Office Protocol 3 (POP3) or Internet Message Access Protocol	Do not install Sun Cluster HA for Netscape Mail and the Lotus IMAP server on the same logical host, or on logical hosts mastered by the same physical host.	110 (POP3) 143 (IMAP)
LDAP	Internet Directory Clients using Lightweight Directory Access Protocol (LDAP)	Do not install Sun Cluster HA for Netscape LDAP and the Lotus LDAP server on the same logical host, or on logical hosts mastered by the same physical host.	389
NNTP	Internet news readers using Network News Transfer Protocol (NNTP)	Do not install Sun Cluster HA for Netscape News and the Lotus NNTP server on the same logical host, or on logical hosts mastered by the same physical host.	119
SMTP/ MIME	Internet mail clients using Simple Mail Transfer Protocol (SMTP)	Do not install HA-nsmail and the Lotus SMTP/MIME on the same logical host, or on logical hosts mastered by the same physical host.	n/a

TABLE 13–1 Lotus Domino Server Options - General Guidelines *(continued)*

13.2 Installing and Configuring Lotus Domino

This section describes the steps to take when installing Lotus Domino to enable it to run as the Sun Cluster HA for Lotus data service.

▼ How to Install and Configure Lotus Domino

Consult your Lotus Domino documentation before performing this procedure.

1. **On each node that can master the logical host running Sun Cluster HA for Lotus, modify the `/etc/nsswitch.conf` file.**

   Modify the `/etc/nsswitch.conf` file so that "group" lookups are directed to files first. For example:

```
...
group: files nisplus
...
```

   Sun Cluster HA for Lotus uses the `su` *user* command when starting and stopping the database server.

2. **Install the Solaris and the Sun Cluster environments.**

   Refer to Chapter 3. Use the `scinstall(1M)` command to install all of the Sun Cluster HA for Lotus packages that you will be using. Complete the post-installation procedures to install any required patches.

   Note - At this time, do not install any patches that are not required by Sun Cluster.

3. **Start Sun Cluster by using the `scadmin(1M)`command.**

   Start the first node. From the administrative workstation:

```
# scadmin startcluster localhostname clustername
```

Then add each node to the cluster. From each node:

```
# scadmin startnode
```

After completing this step, the cluster should be up and running and the HA file systems should be mounted on their default masters.

4. **Make sure each logical host is served by its default master.**

   Sun Cluster HA for Lotus will be installed from the physical host that is the logical host's default master. If necessary, switch over the logical hosts to be served by their respective default masters.

   The logical host names you use in your Sun Cluster configuration should be used as the server names when you install and configure Sun Cluster HA for Lotus applications. This is necessary for failover of the Lotus server to work properly.

5. **On each Sun Cluster server that will be running Lotus Domino, specify user and group names for Lotus Domino.**

   Create a Lotus group, normally named `notes`. Create a user account, also normally named `notes`, and make it a member of the `notes` group. The group ID and user ID should be identical on all nodes.

```
# groupadd notes
# useradd -u notes -g notes -d /opt/lotus/bin notes
```

6. **On each Sun Cluster server that will be running HA-Lotus, install the Lotus Domino software.**

   Log in as root to ensure ownership of the entire directory before performing this step. From the install directory, copy the Lotus install program to your local disk and install the software.

   By default, the Lotus Domino software is installed in the `/opt/lotus` directory, but you can select a different directory on the local or logical disk. The install program will create a symbolic link between the default install directory and the install directory you specify. Run the `install` command as root.

```
# cd /cdrom/notes_r4/uni
# ./install
```

Note - The Lotus Domino installation directory on the Lotus CD-ROM might vary from the directory shown here. Check your Lotus Domino installation documentation for the actual path.

7. **On each Sun Cluster server that will be running HA-Lotus, set up a** `$PATH` **variable for Lotus Domino.**

```
# set PATH = /opt/lotus/bin $PATH .
```

8. **On each Sun Cluster server that will be running Sun Cluster HA for Lotus, set up the Lotus Domino server.**

   Use the Lotus Domino set-up program to set up Lotus Domino. Log in as user `notes` to ensure access to the Domino server data files. You must place the Domino server data files on the logical host.

```
# ./opt/lotus/bin/notes
```

This completes the installation and set up of Lotus Domino. Proceed to Section 13.3 “Installing and Configuring Sun Cluster HA for Lotus” on page 13-5.

13.3 Installing and Configuring Sun Cluster HA for Lotus

This section describes the steps used to install, configure, register, and start Sun Cluster HA for Lotus.

▼ How to Install and Configure Sun Cluster HA for Lotus

1. **On each Sun Cluster server that will be running HA-Lotus, run the** `hadsconfig(1M)` **command to configure Sun Cluster HA for Lotus.**

Use the `hadsconfig(1M)` command to create, edit, and delete instances of Sun Cluster HA for Lotus. Refer to Section 13.3.1 “Configuration Parameters for Sun Cluster HA for Lotus” on page 13-7, for information on the input you will need to supply to the `hadsconfig(1M)` command. See the `hadsconfig(1M)` man page for details.

```
# hadsconfig
```

Caution - Configure only one Lotus instance per cluster. Activation of Lotus instances is “all or nothing”, that is, you cannot activate only a subset of instances in a cluster. Therefore, multiple instances can conflict with each other.

2. Register and activate Sun Cluster HA for Lotus using the `hareg(1M)` command.

The `hareg(1M)` command adds the Sun Cluster HA for Lotus data service to the Cluster Configuration Database, performs a cluster reconfiguration, and starts all of your Lotus Domino servers. Run this command on only one node:

```
# hareg -s -r lotus
...
 # hareg -y lotus
```

3. Verify the Sun Cluster HA for Lotus configuration.

Log in as `notes` and verify the configuration by starting and stopping the Lotus Domino server on one of the Sun Cluster servers:

```
phys-hahost1# /opt/lotus/bin/server
...
 phys-hahost1# /opt/lotus/bin/server -q
```

You can test more of the configuration by starting the cluster, mastering the logical hosts from various physical hosts, and then starting and stopping the Lotus Domino server from those physical hosts. For example:

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

Log in as user `notes`, and stop and start the Lotus Domino server from the Domino data directory. For example:

```
phys-hahost2# cd /hahost1/domino_data_dir
 phys-hahost2# /opt/lotus/bin/server
 ...
 phys-hahost2# /opt/lotus/bin/server -q
```

This completes the configuration and activation of Sun Cluster HA for Lotus.

13.3.1 Configuration Parameters for Sun Cluster HA for Lotus

This section describes the information you supply to the `hadsconfig(1M)` command to create configuration files for each Sun Cluster HA for Lotus data service. The `hadsconfig(1M)` command uses templates to create these configuration files, and stores the files in the `/etc/opt/SUNWsclts` directory. The templates contain some default, some hardcoded, and some unspecified parameters. You must provide values for the parameters that are unspecified.

The fault probe parameters, in particular, can affect the performance of Sun Cluster HA for Lotus. Tuning the probe interval value too low (increasing the frequency of fault probes) might encumber system performance, and also might result in false takeovers or attempted restarts when the system is simply slow.

You must set the takeover flag for Sun Cluster HA for Lotus. This flag specifies how Sun Cluster will handle partial failover. There are two options:

- `–y` (yes) – Sun Cluster will switch over the logical host to another master. If the attempt fails, Sun Cluster will switch over all logical hosts to the target master anyway, and also might halt or reboot the original master. This flag is the default.
- `–n` (none) – Sun Cluster will not move the logical host to another master, even if it detects problems with the data server, nor will it take any action against the sick data server or database on the logical host.

13.3.1.1 Configuration Parameters for Sun Cluster HA for Lotus

Configure the Sun Cluster HA for Lotus parameters listed in the `hadsconfig(1M)` input form by supplying options described in Table 13–2.

TABLE 13–2 Configuration Parameters for Sun Cluster HA for Lotus

Parameter	Description
Name of the instance	Logical host name used as an identifier for the instance. The log messages generated by Sun Cluster HA for Lotus refer to this identifier. The `hadsconfig(1M)` command prefixes the package name to the logical host name you supply. For example, if you specify "hahost1," `hadsconfig(1M)` produces "SUNWsclts_hahost1."
Logical host	Name of the logical host that provides service for this instance of Sun Cluster HA for Lotus.
Base directory of product installation	Rooted path name specifying the location on the multihost disk of the HA Lotus installation. This is the "instance path," for example, `/hahost1/lotus-home/lotus_1`.
Configuration directory	The directory of the database, for example, `/hahost1/d1/Lotus/`*database*`.db`.
Remote probe	Specifies whether the Lotus fault probe will probe the remote host. Default value is `–n`.
Local probe	Specifies whether the Lotus fault probe will probe the local host. Default value is `–y`.
Probe interval	The time, in seconds, between fault probes. The default interval is 60 seconds.
Probe timeout	The time, in seconds, after which a fault probe will time out. The default timeout value is 60 seconds.
Server port number	Unique port for this instance of Sun Cluster HA for Lotus. The default port number is 1352.
Takeover flag	Specifies whether a failure of this instance will cause a takeover or failover of the logical host associated with the data service instance. Possible values are `–y` (yes) or `–n` (no). Default value is `–y`..

CHAPTER 14

Setting Up and Administering Parallel Database Systems

This chapter provides instructions for setting up and configuring the following Parallel Database Systems on your Sun Cluster servers:

- Oracle Parallel Server (OPS)
- Informix-Online XPS
- Section 14.1 “General Information for Parallel Database Systems” on page 14-1
- Section 14.2 “Installing OPS” on page 14-5
- Section 14.3 “Installing Informix-Online XPS” on page 14-6

14.1 General Information for Parallel Database Systems

14.1.1 Shared Disk Architecture

The shared disk configuration of Sun Cluster is used by OPS. In this configuration, a single database is shared among multiple instances of OPS, which access the database concurrently. Conflicting access to the same data is controlled by means of a distributed lock manager (the Oracle UNIX Distributed Lock Manager (DLM)). If a process or a node crashes, the DLM is reconfigured to recover from such failure.

14.1.2 Shared Nothing Architecture

The shared nothing disk configuration of Sun Cluster is used by Informix-Online XPS. The database server instance(s) on each node has sole access to its own database partition.

A database query from a client is analyzed by the servers for its table partitions and forwarded across the private network to the appropriate servers. The results are merged across the private network and returned to the client.

14.1.3 SMA Shared Memory Issues

Some applications (OPS, for example) sometimes require modification of the `/etc/system` file so that the minimum amount of shared memory that may be requested is unusually high. For example, 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, as it ends up requesting a smaller number of bytes than the minimum the system can allocate. As a result, the `shmget`(2) system call made by the `sm_config(1M)` command fails, thus aborting `sm_config(1M)`.

To work around this problem, edit the `/etc/system` file and set the value of *shmsys:shminfo_shmmin* to 200. The value of *shmsys:shminfo_shmmax* should be greater than 200. Then reboot the machine for the new values to take effect.

If you encounter `semsys` warnings and core dumps, it could mean that the semaphore values contained in the `semsys:seminfo_*` fields in the `/etc/system` file do not match the actual physical limits of the machine.

14.1.4 OPS and IP Failover

In the event of a node failure in an OPS environment, Oracle SQL*Net clients may be configured to reconnect to the surviving server without the use of IP failover.

In an OPS environment, multiple Oracle instances co-operate to provide access to the same shared database. The Oracle clients may access the database using any of the instances. Thus if one or more instance have failed, clients may continue to access the database by connecting to a surviving instance.

There are many ways to accomplish the task of reconnecting to a surviving instance transparently to the end user:

- Design the application such that if the Oracle client loses the connection to the Oracle instance, it reestablishes the connection to an alternate instance. This implies the client is aware that it is operating in a multi-instance (OPS) environment.

However, such a solution is seldom used. Instead, most implementations use middleware such as the Tuxedo transaction monitor (TM), to implement the reconnection logic. The Oracle client connects to the TM, which in turn connects to one of the many database instances. The TM hides the failure of a particular database instance from the clients by reconnecting to alternate instances. The advantage of the TM approach is existing Oracle client applications need not be rewritten to take advantage of the multiple instances in an OPS environment. The disadvantage is the cost of integrating with a TM.

- Design the application (Oracle client) such that when it loses the connection to the database instance, it retries the connection to the same server. Thus Oracle client applications designed for a non-parallel environment can be moved into an OPS environment without redesign. The "infrastructure" is then designed to ensure the connection is routed to the surviving server.

  One solution used to accomplish this is to use the IP failover features of Sun Cluster in conjunction with the OPS data services. The rest of this document describes a simpler alternative through the use of high availability features of Oracle SQL*Net. IP failover is not required to implement this functionality.

14.1.4.1 High Availability Features of Oracle SQL*Net

From the Oracle client perspective the model is simple, when the server crashes the client sees a broken connection. The client reconnects to the server, and resubmits the transaction. Oracle SQL*Net provides features and capabilities to incorporate multiple instances running on different hosts under the same service. Hence, when the client reconnects it is automatically connected through to the surviving instance. The reconnection is not automatic. The client typically incorporates the code to reconnect broken connections (to the same service as before).

Note - With a node or instance failure, the surviving instance(s) must first recover the failed instances state. During this recovery time clients will see a lack of response from the instance. This recovery has nothing do with the Sun Cluster framework. Recovery is totally dependent on Oracle, the transaction volume, and recovery mechanism for OPS.

14.1.4.2 Configuring Oracle SQL*Net

Two ways to configure Oracle SQL*Net on the client (the `TNSNAMES.ORA` file) are shown below. The client reconnection time to the surviving instance is not influenced by the method used to configure Oracle SQL*Net.

- Configure the same "connect string" for multiple instances to run on different hosts with the same ORACLE SID.

```
ora =
 (DESCRIPTION =
   (ADDRESS_LIST =
    (ADDRESS =
      (PROTOCOL = TCP)
      (HOST = erlan)
      (PORT = 1526) <- instance 1
    )
    (CONNECT_DATA= (SID=ora))
   )
  (ADDRESS_LIST =
    (ADDRESS =
     (PROTOCOL = TCP)
      (HOST = weibull)
      (PORT = 1526) <- instance 2
    )
    (CONNECT_DATA= (SID=ora))
  )
 )
```

- Configure the same “connect string” for instances to run on different hosts with different ORACLE SIDs.

```
ora =
 (DESCRIPTION_LIST =
  (DESCRIPTION =
       (ADDRESS_LIST =
     (ADDRESS =
       (PROTOCOL = TCP)
       (HOST = erlang)
       (PORT = 1526))
     (CONNECT_DATA = (SID = ora)(GLOBAL_NAME = ora))
  )
  (DESCRIPTION =
      (ADDRESS_LIST =
             (ADDRESS =
       (PROTOCOL = TCP)
       (HOST = weibull)
       (PORT = 1526))
       (CONNECT_DATA = (SID = ora1)(GLOBAL_NAME = ora))
  )
 )
```

This configuration has listeners running for each of the instances.

14.2 Installing OPS

If you are installing Oracle7 Parallel Server, refer to the *Oracle7 for Sun SPARC Solaris 2.x Installation and Configuration Guide, Release 7.x*. If you are installing Oracle8 Parallel Server, refer to the *Oracle8 Parallel Server Concepts and Administration, Release 8.0*.

▼ How to Install OPS

In a Sun Cluster configuration, perform the following steps on all nodes when you install OPS.

1. **Edit the `/etc/system` file by using your favorite text editor, for example:**

```
# vi /etc/system
```

2. **Look for a line similar to the following:**

```
set shmsys:shminfo_shmmax=10000000
```

3. **If your Oracle documentation explains this line, then skip Step 4 on page @–5 and Step 5 on page @–5. Otherwise, perform Step 4 on page @–5 and Step 5 on page @–5.**

Note - The Oracle7 and Oracle8 documentation always supersedes the information in this section.

---

4. **If the number is less than 10 million, change the number to 10 million (`–10000000`). Otherwise, leave the number as is.**

   Depending on the size of the database, you may have to change this number later as part of the OPS installation.

5. **Reboot all nodes.**

14.3 Installing Informix-Online XPS

For information about installing Informix-Online XPS, refer to the *Installation Notes INFORMIX-OnLine XPS* document. Any changes made in the `/etc/system` file to support Informix-Online XPS must be based on the information contained in your Informix documentation.

Note - The Informix-Online XPS documentation always supersedes the information in this section.

▼ How to Install Informix-Online XPS

In a Sun Cluster configuration, perform the following steps to install Informix-Online XPS.

1. **Edit the `/etc/hosts` file using your favorite text editor.**

   For example:

```
# vi /etc/hosts
```

2. **Add lines similar to the following to the `/etc/hosts` file.**

```
204.152.65.1    ssa28-scid0
 204.152.65.17   ssa28-scid1
 204.152.65.2    ssa28a-scid0
 204.152.65.18   ssa28a-scid1
```

   In this example, 204.152.65.1 and 204.152.65.17 are the IP addresses assigned by the Sun Cluster system to the SCI cards on the primary node. IP addresses 204.152.65.2 and 204.152.65.18 are the addresses assigned to the SCI cards on the secondary node.

   From node 1, you can use `ssa28a-scid0` or `ssa28a-scid1` to communicate with node 2 in the cluster. This choice enables you to select the connection (SCI 0 or SCI 1) that is used to carry the message.

   The Informix `onconfig` configuration file uses these names to set up communication between the two nodes.

Note - Network Security—Informix requires the Private Interconnect IP address of the cluster nodes to be available in the `/etc/hosts` file during installation. If any of the nodes of the cluster are configured to run as NIS or DNS name servers, this requirement may present a security problem because the name servers may make the private addresses available to unauthorized hosts. In the interest of security, if Informix is installed on your cluster, you may not want to configure the nodes as NIS or DNS name servers.

APPENDIX **A**

Configuration Worksheets and Examples

A.1 Configuration Worksheets

This appendix provides worksheets for planning your:

- Network connections
- Host names and IP addresses
- Disks configurations
- Logical hosts
- Metadevices

Node 1

Private Networks (Ethernet only)

First private net interface _____

Second private net interface _____

Node 2

Private Networks (Ethernet only)

First private net interface _____

Second private net interface _____

Node 3

Private Networks (Ethernet only)

First private net interface _____

Second private net interface _____

Node 4

Private Networks (Ethernet only)

First private net interface _____

Second private net interface _____

Administrative Workstation

Host name _____

IP address _____

Terminal Concentrator

Host name _____

IP address _____

System Service Processor (E10000 only)

Host name _____

IP address _____

Figure A–1 Installation Worksheet: Private Network Interfaces, Administrative Workstation, Terminal Concentrator, and System Service Processor

Primary Public Network _____

Node 1

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Node 2

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Node 3

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Node 4

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Figure A–2 Installation Worksheet: Primary Public Network Names and IP Addresses

Secondary Public Network _____

Node 1

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Node 2

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Node 3

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Node 4

Physical host name _____
Ethernet address _____
Physical port of TC/SSP connected to ____

Logical Hosts

Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____
Logical host name_____
Logical IP address_____

Figure A–3 Installation Worksheet: Secondary Public Network Names and IP Addresses

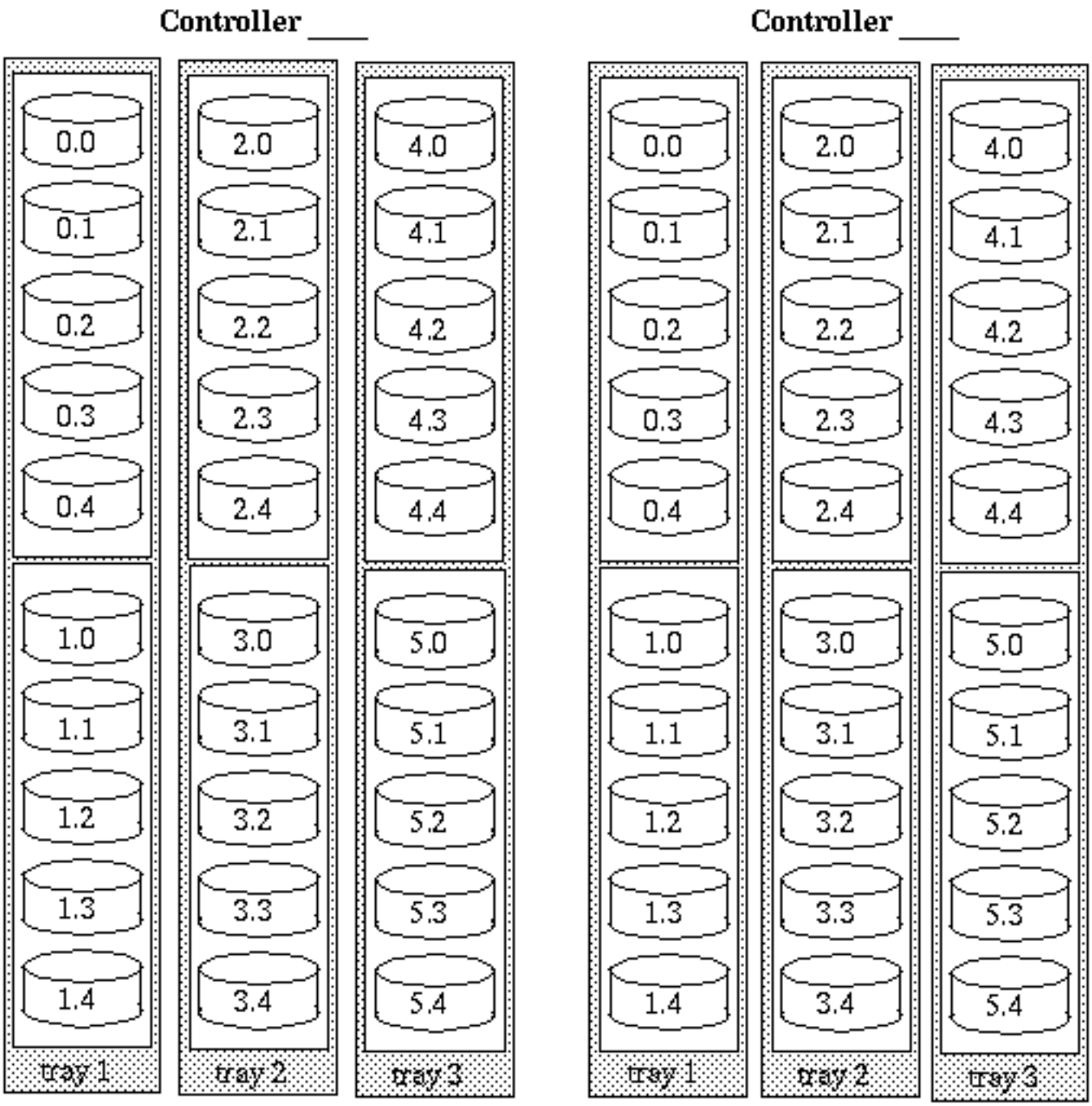


Figure A–4 SPARCStorage Array Model 100 Disk Setup Worksheet - Part 1

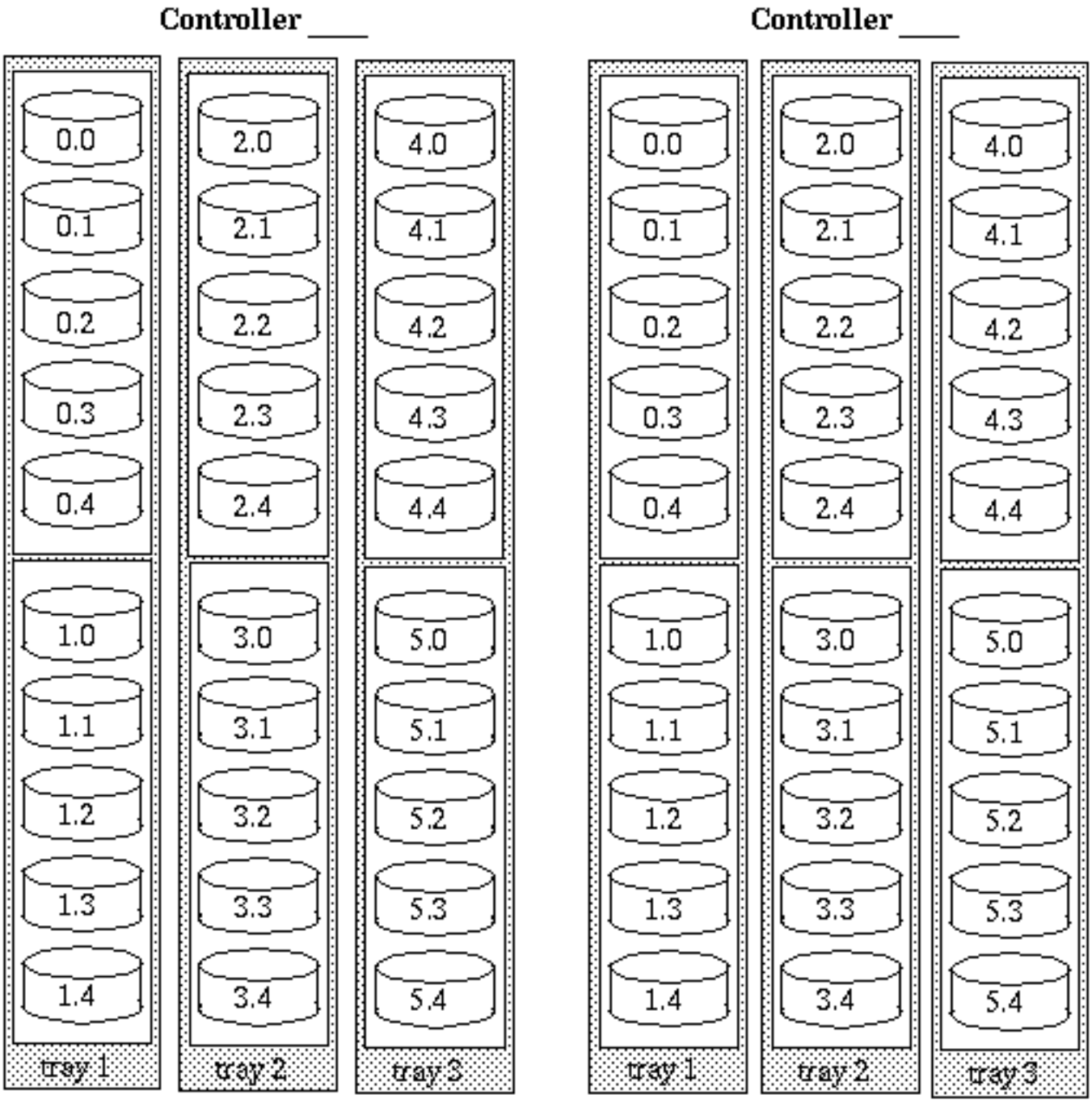


Figure A–5 SPARCStorage Array Model 100 Disk Setup Worksheet - Part 2

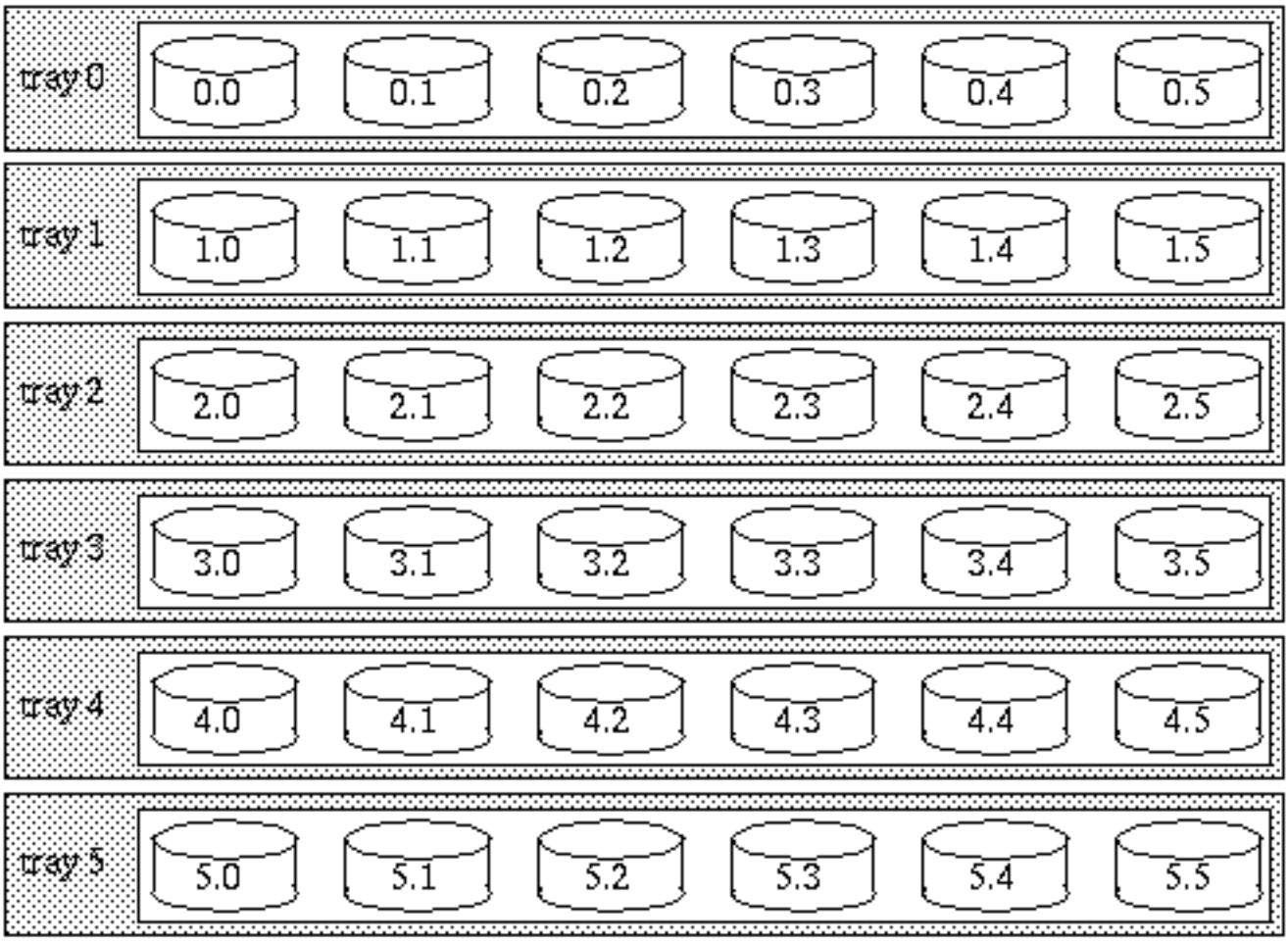


Figure A–6 SPARCstorage Array Model 200 Disk Setup Worksheet - Part 1

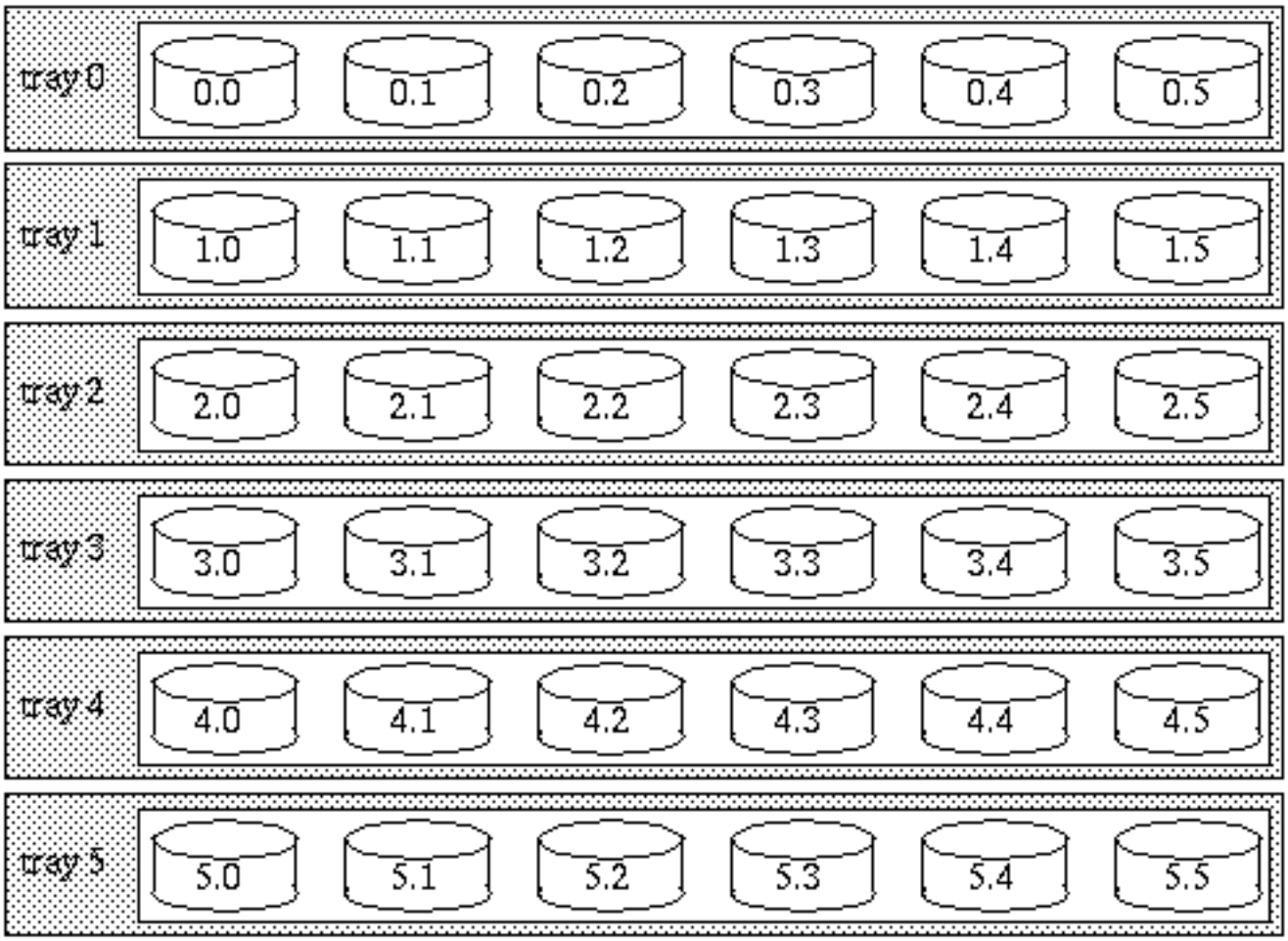


Figure A–7 SPARCstorage Array Model 200 Disk Setup Worksheet - Part 2

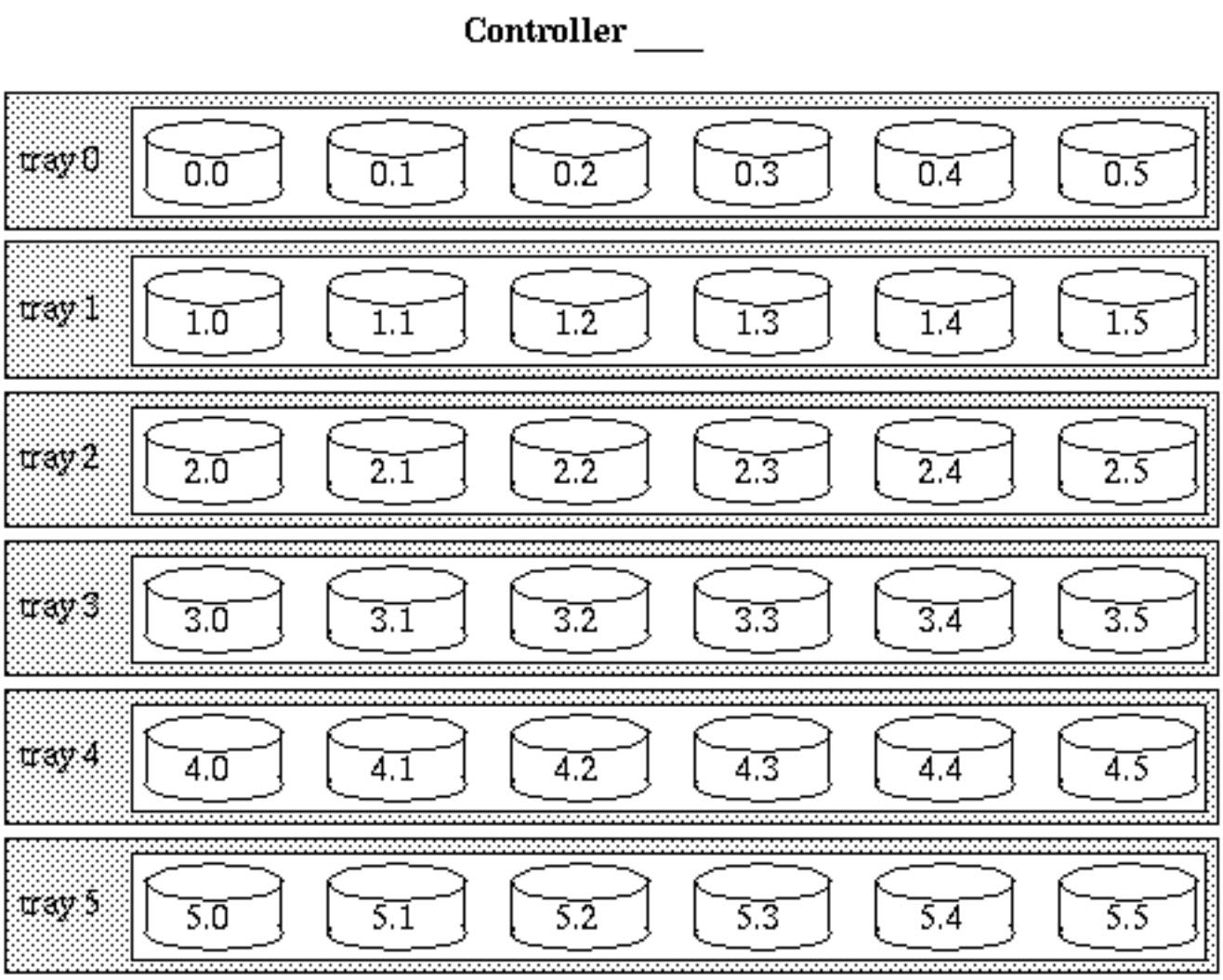


Figure A–8 SPARCstorage Array Model 200 Disk Setup Worksheet - Part 3

TABLE A–1 Metadevice Planning Worksheet

File System	Metatrans	Metamirrors (Data)	Metamirrors (Log)	Submirrors (Data)	Submirrors (Log)	Hot Spare Pool	Physical Device (Data)	Physical Device (Log)
A	d10	d11		d12		hsp000	c1t0d0s0	
				d13		hsp000	c2t0d1s0	
			d14		d15	hsp006		c1t0d1s6
					d16	hsp006		c2t1d0s6

TABLE A–1 Metadevice Planning Worksheet *(continued)*

File System	Metatrans	Metamirrors (Data)	Metamirrors (Log)	Submirrors (Data)	Submirrors (Log)	Hot Spare Pool	Physical Device (Data)	Physical Device (Log)

TABLE A–2 Logical Host Information Worksheet

Logical Host Name	Default Master	Administrative File System	vfstab File Name	Diskset Name
hahost1	phys-hahost1	/hahost1	vfstab.hahost1	hahost1

TABLE A–2 Logical Host Information Worksheet *(continued)*

Logical Host Name	Default Master	Administrative File System	vfstab File Name	Diskset Name

APPENDIX **B**

Configuring Solstice DiskSuite

Configure your local and multihost disks for Solstice DiskSuite by using the guidelines in this chapter along with the information in Chapter 2. Then create your `md.tab` file using the guidelines and examples in this chapter. Refer to the Solstice DiskSuite documentation for additional details on creating an `md.tab` file. It is easiest to create the `md.tab` file as you plan and design your metadevice configuration, then copy the `md.tab` file to each of the Sun Cluster nodes after installing the Sun Cluster and Solstice DiskSuite software.

This appendix includes the following procedures:

B.1 Overview of Configuring Solstice DiskSuite for Sun Cluster

Table B–1 shows the high-level steps to configure Solstice DiskSuite to work with Sun Cluster. The tasks should be performed in the order shown.

TABLE B–1 High-Level Steps to Configure Solstice DiskSuite

Task	Go To ...
1. Planning your Solstice DiskSuite configuration	Chapter 2
2. Calculating the quantity of metadevice names needed	"How to Calculate the Number of Metadevice Names" on page B–4
3. Preparing the disk IDs by running the `scdidadm(1M)` command	"How to Prepare the Configuration to Use the DID Driver" on page B–5
4. Creating metadevice state database replicas on the local (private) disks by running the `metadb(1M)` command	"How to Create Local Metadevice State Database Replicas" on page B–9
5. (Optional) Mirroring the root (`/`) file system	Section B.2.6 "Mirroring the root (`/`) File System" on page B-10
6. Creating the disksets by running the `metaset(1M)` command	"How to Create a Diskset" on page B–11
7. Adding drives to the diskset by running the `metaset(1M)` command	"How to Add Drives to a Diskset" on page B–11
8. (Optional) Repartitioning drives in a diskset	"How to Repartition Drives in a Diskset" on page B–13
9. Setting up the `md.tab` file to create metadevices on disksets	Section B.2.9 "Using the `md.tab` File to Create Metadevices in Disksets" on page B-14

TABLE B–1 High-Level Steps to Configure Solstice DiskSuite *(continued)*

Task	Go To ...
10. "Running" the `md.tab` file by using the `metainit(1M)` command	"How to Initialize the `md.tab` File" on page B–17
11. Configuring file systems for each logical host	"How to Create Multihost UFS File Systems" on page B–18

The following sections describe all the procedures necessary to configure Solstice DiskSuite with Sun Cluster.

Note - If your cluster has only two disk storage units ("two drive strings"), you must configure Solstice DiskSuite Mediators. Refer to the chapter on using dual-string mediators in the *Sun Cluster 2.2 System Administration Guide* for details on configuring and administering mediators.

B.2 Configuring Solstice DiskSuite for Sun Cluster

Use the procedures in this section to configure the following:

- (Optional) Number of metadevice names
- Disk IDs
- Local metadevice state database replicas
- (Optional) Mirrored root (`/`) file system
- Disksets
- Drives in a diskset
- (Optional) Drive partitions
- `md.tab` file
- File systems

Note - For convenience, modify your `PATH` variable to include `/usr/opt/SUNWmd/sbin`.

B.2.1 Calculating the Number of Metadevice Names

You must calculate the number of Solstice DiskSuite metadevice names needed for your configuration before you set up the configuration. The default number of metadevice names is 128. Many configurations will need more than the default. Increasing this number before implementing a configuration will save administration time later on.

▼ How to Calculate the Number of Metadevice Names

1. **Calculate the quantity of metadevice names needed by determining the largest of the metadevice names to be used in each diskset.**

   This requirement is based on the metadevice name value rather than on the actual quantity. For example, if your metadevice names range from `d950` to `d1000`, Solstice DiskSuite will require one thousand names, not fifty.

2. **If the calculated quantity exceeds 128, you must edit the `/kernel/drv/md.conf` file.**

   Set the `nmd` field in `/kernel/drv/md.conf` to the largest metadevice name value used in a diskset.

   Changes to the `/kernel/drv/md.conf` file do not take effect until a reconfiguration reboot is performed. The `md.conf` files on each cluster node must be identical.

   Refer to Appendix A, for worksheets to help you plan your metadevice configuration.

Note - The Solstice DiskSuite documentation states that the only modifiable field in the `/kernel/drv/md.conf` file is the `nmd` field. However, you can modify the `md_nsets` field as well if you want to configure additional disksets.

B.2.2 Using the Disk ID Driver

All new installations running Solstice DiskSuite require a Disk ID (DID) pseudo driver to make use of disk IDs. Disk IDs enable metadevices to locate data independent of the device name of the underlying disk. Configuration changes or hardware updates are no longer a problem because the data is located by Disk ID and not the device name.

To create a mapping between a disk ID and a disk path, you run the `scdidadm(1M)` command from node 0. The `scdidadm(1M)` command sets up three components:

- Disk ID (DID) – This is a “short-hand” number assigned to the physical disk, such as “1.”

- DID Instance Number –This is the full path to the raw disk device, such as phys-hahost3:/dev/rdsk/c0t0d0.
- DID Full Name – This is the full path of the DID, such as /dev/did/rdsk/d1.

The Solstice HA 1.3 release supported two-node clusters only. In this two-node configuration, both nodes were required to be configured identically on identical platforms, so the major/minor device numbers used by the Solstice DiskSuite device driver were the same on both systems. In greater than two-node configurations, it is difficult to cause the minor numbers of the disks to be identical on all nodes within a cluster. The same disk might have different major/minor numbers on different nodes. The DID driver uses a generated DID device name to access a disk that might have different major/minor numbers on different nodes.

Although use of the DID driver is required for clusters using Solstice DiskSuite with more than two nodes, the requirement has been generalized to all new Solstice DiskSuite installations. This enables future conversion of two-node Solstice DiskSuite configurations to greater than two-node configurations.

Note - If you are upgrading from HA 1.3 to Sun Cluster 2.2, you do not need to run the `scdidadm(1M)` command.

▼ How to Prepare the Configuration to Use the DID Driver

To set up a Solstice DiskSuite configuration using the DID driver, complete this procedure.

Note - If you have a previously generated `md.tab` file to convert to use disk IDs, you can use the script included in Section B.2.4 “DID Conversion Script” on page B-8, to help with the conversion.

---

1. **Run the `scdidadm(1M)` command to create a mapping between a disk ID instance number and the local and remote paths to the disk.**

   Perform this step after running the `scinstall(1M)` command with the cluster up. In order to maintain one authoritative copy of the DID configuration file, you can run the script only on node 0 while all nodes are up; otherwise it will fail. The `get_node_status(1M)` command includes the node ID number as part of its output. Refer to the `scdidadm(1M)` man page for details.

   ```
   phys-hahost1# scdidadm -r
   ```

   Note - You must run the `scdidadm(1M)` command from cluster node 0.

If the `scdidadm(1M)` command is unable to discover the private links of the other cluster nodes, run this version of the command from node 0.

```
phys-hahost1# scdidadm -r -H hostname1,hostname2,...
```

Make sure the appropriate host name for node 0 is in the `/.rhosts` files of the other cluster nodes when using this option. Do not include the host name of the cluster node from which you run the command in the `hostname` list.

2. Use the DID mappings to update your `md.tab` file.

Refer to Section B.2.3 “Troubleshooting DID Driver Problems” on page B-7, if you receive the error message:

```
The did entries in name_to_major must be the same on all nodes.
```

Correct the problem, then rerun the `scdidadm(1M)` command.

Once the mapping between DID instance numbers and disk IDs has been created, use the full DID names when adding drives to a diskset and in the `md.tab` file in place of the lower level device names (c*X*t*X*d*X*). The `-l` option to the `scdidadm(1M)` command shows a list of the mappings to help generate your `md.tab` file. In the following example, the first column of output is the DID instance number, the second column is the full path (physical path), and the third column is the full name (pseudo path):

```
phys-hahost1# scdidadm -l
60        phys-hahost3:/dev/rdsk/c4t5d2     /dev/did/rdsk/d60
 59        phys-hahost3:/dev/rdsk/c4t5d1     /dev/did/rdsk/d59
 58        phys-hahost3:/dev/rdsk/c4t5d0     /dev/did/rdsk/d58
 57        phys-hahost3:/dev/rdsk/c4t4d2     /dev/did/rdsk/d57
 56        phys-hahost3:/dev/rdsk/c4t4d1     /dev/did/rdsk/d56
 55        phys-hahost3:/dev/rdsk/c4t4d0     /dev/did/rdsk/d55
 ...
 6         phys-hahost3:/dev/rdsk/c0t1d2     /dev/did/rdsk/d6
 5         phys-hahost3:/dev/rdsk/c0t1d1     /dev/did/rdsk/d5
 4         phys-hahost3:/dev/rdsk/c0t1d0     /dev/did/rdsk/d4
 3         phys-hahost3:/dev/rdsk/c0t0d2     /dev/did/rdsk/d3
 2         phys-hahost3:/dev/rdsk/c0t0d1     /dev/did/rdsk/d2
 1         phys-hahost3:/dev/rdsk/c0t0d0     /dev/did/rdsk/d1
```

Proceed to Section B.2.5 “Creating Local Metadevice State Database Replicas” on page B-9 to create local replicas.

If you have problems with the DID driver, refer to Section B.2.3 “Troubleshooting DID Driver Problems” on page B-7.

B.2.3 Troubleshooting DID Driver Problems

In previous releases, Solstice DiskSuite depended on the major number and instance number of the low-level disk device being the same on the two nodes connected to the disk. With this release of Sun Cluster, Solstice DiskSuite requires that the DID major number be the same on all nodes and that the instance number of the DID device be the same on all nodes. The `scdidadm(1M)` command checks the DID major number on all nodes. The value recorded in the `/etc/name_to_major` file must be the same on all nodes.

If the `scdidadm(1M)` command finds that the major number is different, it will report this and ask you to fix the problem and re-run the `scdidadm(1M)` command. The DID driver uses major number 149; if there is a numbering conflict, you must choose another number for the DID driver. The following procedure enables you to make the necessary changes.

▼ How to Resolve Conflicts With the DID Major Number

1. **Choose a number that does not conflict with any other entry in the `/etc/name_to_major` file.**

2. **Edit the `/etc/name_to_major` file on each node and change the DID entry to the number you chose.**

3. **On each node where the `/etc/name_to_major` file was updated, execute the following commands.**

```
phys-hahost1# rm -rf /devices/pseudo/did* /dev/did
phys-hahost1# reboot -- -r
...
```

4. **On the node used to run the `scdidadm(1M)` command, execute the following commands.**

```
phys-hahost3# rm -f /etc/did.conf
phys-hahost3# scdidadm -r
```

This procedure resolves mapping conflicts and reconfigures the cluster with the new mappings.

B.2.4 DID Conversion Script

If you have a previously generated `md.tab` file to convert to use disk IDs, you can use the script in Code Example B–1 to help with the conversion. The script checks the `md.tab` file for physical device names, such as `/dev/dsk/c0t0d0` or `c0t0d0`, and converts these names to the full DID name, such as `/dev/did/rdsk/d60`.

CODE EXAMPLE B–1 `md.tab` Conversion Script

```
more phys_to_did
 #! /bin/sh
 #
 # ident "@(#)phys_to_did        1.1     98/05/07 SMI"
 #
 # Copyright (c) 1997-1998 by Sun Microsystems, Inc.
 # All rights reserved.
 #
 # Usage: phys_to_did <md.tab filename>
 # Converts $1 to did-style md.tab file.
 # Writes new style file to stdout.

 MDTAB=$1
 TMPMD1=/tmp/md.tab.1.$$
 TMPMD2=/tmp/md.tab.2.$$
 TMPDID=/tmp/didout.$$

 # Determine whether we have a "physical device" md.tab or a "did" md.tab.
 # If "physical device", convert to "did".
 grep "\/dev\/did" $MDTAB > /dev/null 2>&1
 if [ $? -eq 0 ]; then
         # no conversion needed
         lmsg=‘gettext "no conversion needed"‘
         printf "${lmsg}\n"
         exit 0
 fi

 scdidadm -l > $TMPDID
 if [ $? -ne 0 ]; then
         lmsg=‘gettext "scdidadm -l failed"‘
         printf "${lmsg}\n"
         exit 1
 fi

 cp $MDTAB $TMPMD1

 ...
 ...
 # Devices can be specified in md.tab as /dev/rdsk/c?t?d? or simply c?t?d?
 # There can be multiple device names on a line.
 # We know all the possible c.t.d. names from the scdidadm -l output.

 # First strip all /dev/*dsk/ prefixes.
 sed -e ’s:/dev/rdsk/::g’ -e ’s:/dev/dsk/::g’ $TMPMD1 > $TMPMD2

 # Next replace the resulting physical disk names "c.t.d." with
```

(continued)

(Continuation)

```
# /dev/did/rdsk/<instance>
exec < $TMPDID
while read instance fullpath fullname
do
        old=‘basename $fullpath‘
        new=‘basename $fullname‘
        sed -e ’s:’$old’:/dev/did/rdsk/’$new’:g’ $TMPMD2 > $TMPMD1
        mv $TMPMD1 $TMPMD2
done

cat $TMPMD2
rm -f $TMPDID $TMPMD1 $TMPMD2

exit 0
```

B.2.5 Creating Local Metadevice State Database Replicas

Before you can perform any Solstice DiskSuite configuration tasks, such as creating disksets on the multihost disks or mirroring the root (/) file system, you must create the metadevice state database replicas on the local (private) disks on each cluster node. The local disks are separate from the multihost disks. The state databases located on the local disks are necessary for the operation of Solstice DiskSuite.

▼ How to Create Local Metadevice State Database Replicas

Perform this procedure on each node in the cluster.

1. **As root, use the** `metadb(1M)` **command to create local replicas on each cluster node’s system disk.**

   For example, this command creates three metadevice state database replicas on Slice 7 of the local disk.

   ```
   # metadb -afc 3 c0t0d0s7
   ```

   The –c option creates the replicas on the same slice. This example uses Slice 7, but you can use any free slice.

2. **Use the** `metadb(1M)` **command to verify the replicas.**

```
# metadb
flags         first blk      block count
     a        u         16              1034            /dev/dsk/c0t0d0s7
     a        u         1050            1034            /dev/dsk/c0t0d0s7
     a        u         2084            1034            /dev/dsk/c0t0d0s7
```

B.2.6 Mirroring the root (/) File System

You can mirror the root (/) file system to prevent the cluster node itself from going down due to a system disk failure. Refer to Chapter 2, for more information.

The high-level steps to mirror the root (/) file system are:

- Using the `metainit(1M)` `–f` command to put the root slice in a single slice (one-way) concatenation (submirror1)
- Creating a second concatenation (submirror2)
- Using the `metainit(1M)` command to create a one-way mirror with submirror1
- Running the `metaroot(1M)` command
- Running the `lockfs(1M)` command
- Rebooting
- Using the `metattach(1M)` command to attach submirror2
- Recording the alternate boot path

For more information, refer to the `metainit(1M)`, `metaroot(1M)`, and `metattach(1M)` man pages and to the Solstice DiskSuite documentation.

B.2.7 Creating Disksets

A diskset is a set of multihost disk drives containing Solstice DiskSuite objects that can be accessed exclusively (but not concurrently) by multiple hosts. To create a diskset, root must be a member of Group 14.

When creating your disksets, use the following rules to ensure correct operation of the cluster in the event of disk enclosure failure:

1. If exactly two “strings” are being used, the diskset should have the same number of physical disks on the two strings.

Note - For the two-string configuration, mediators are required. Refer to the *Sun Cluster 2.2 System Administration Guide* for details on setting up mediators.

---

1. If more than two strings are being used, for example three strings, then you must ensure that for any two strings S1 and S2, the sum of the number of disks on those strings exceeds the number of disks on the third string S3. As a formula: count(S1) + count(S2) > count(S3).

▼ How to Create a Diskset

Perform this procedure for each diskset in the cluster. All nodes in the cluster must be up. Creating a diskset involves assigning hosts and disk drives to the diskset.

1. **Make sure the local metadevice state database replicas exist.**

   If necessary, refer to the procedure "How to Create Local Metadevice State Database Replicas" on page B–9.

2. **As root, create the disksets by running the** `metaset(1M)` **command from one of the cluster nodes.**

   For example, this command creates two disksets, `hahost1` and `hahost2`, consisting of nodes `phys-hahost1` and `phys-hahost2`.

```
phys-hahost1# metaset -s hahost1 -a -h phys-hahost1 phys-hahost2
phys-hahost1# metaset -s hahost2 -a -h phys-hahost1 phys-hahost2
```

3. **Check the status of the new disksets by running the** `metaset(1M)` **command.**

```
phys-hahost1# metaset
```

   You are now ready to add drives to the diskset, as explained in the procedure "How to Add Drives to a Diskset" on page B–11.

▼ How to Add Drives to a Diskset

When a drive is added to a diskset, Solstice DiskSuite repartitions it as follows so that the metadevice state database for the diskset can be placed on the drive.

- A small portion of each drive is reserved in Slice 7 for use by Solstice DiskSuite. The remainder of the space on each drive is placed into Slice 0.
- Drives are repartitioned when they are added to the diskset only if Slice 7 is not set up correctly.
- Any existing data on the disks is lost by the repartitioning.

- If Slice 7 starts at Cylinder 0, and the disk is large enough to contain a state database replica, the disk is not repartitioned.

After adding a drive to a diskset, you may repartition it as necessary, with the exception that Slice 7 is not altered in any way. Refer to "How to Repartition Drives in a Diskset" on page B–13, and to Chapter 2, for recommendations on how to set up your multihost disk partitions.

Caution - If you repartition a disk manually, create a Partition 7 starting at Cylinder 0 that is large enough to hold a state database replica (approximately 2 Mbytes). The –`Flag` field in Slice 7 must have –`V_UNMT` (unmountable) set and must not be set to read-only. Slice 7 must not overlap with any other slice on the disk. Do this to prevent the `metaset(1M)` command from repartitioning the disk.

Use this procedure to add drives to a diskset.

1. **Make sure you have prepared the configuration to use the DID driver, and that the disksets have been created.**

   If necessary, refer to "How to Prepare the Configuration to Use the DID Driver" on page B–5 and "How to Create a Diskset" on page B–11.

2. **As root, use the `metaset(1M)` command to add the drives to the diskset.**

   Use the DID driver name for the disk drives rather than the character device name. For example:

```
phys-hahost1# metaset -s hahost1 -a /dev/did/dsk/d1 /dev/did/dsk/d2
phys-hahost1# metaset -s hahost2 -a /dev/did/dsk/d3 /dev/did/dsk/d4
```

3. **Use the `metaset(1M)` command to verify the status of the disksets and drives.**

```
phys-hahost1# metaset -s hahost1
phys-hahost1# metaset -s hahost2
```

4. **(Optional) Refer to Section B.2.8 "Planning and Layout of Disks" on page B-13," to optimize multihost disk slices.**

B.2.8 Planning and Layout of Disks

The `metaset(1M)` command repartitions drives in a diskset so that a small portion of each drive is reserved in Slice 7 for use by Solstice DiskSuite. The remainder of the space on each drive is placed into Slice 0. To make more effective use of the disk, use the procedure in this section to modify the disk layout.

▼ How to Repartition Drives in a Diskset

1. **Use the `format(1M)` command to change the disk partitioning for the majority of drives as shown in Table B–2.**

TABLE B–2 Multihost Disk Partitioning for Most Drives

Slice	Description
7	2 Mbytes, reserved for Solstice DiskSuite
6	UFS logs
0	remainder of the disk
2	overlaps Slices 6 and 0

In general, if UFS logs are created, the default size for Slice 6 should be 1 percent of the size of the largest multihost disk found on the system.

Note - The overlap of Slices 6 and 0 by Slice 2 is used for raw devices where there are no UFS logs.

2. **Partition a drive on each of the first two controllers in each of the disksets as shown in Table B–3.**

   In the following table, we partition the first drive on the first two controllers as shown. You are not required to use the first drives or the first two controllers, if you have more than two.

TABLE B–3 Multihost Disk Partitioning–First Drive, First Two Controllers

Slice	Description
7	2 Mbytes, reserved for Solstice DiskSuite
5	2 Mbytes, UFS log for HA administrative file systems
4	9 Mbytes, UFS master for HA administrative file systems
6	UFS logs
0	remainder of the disk
2	overlaps Slices 6 and 0

Partition 7 should be reserved for use by Solstice DiskSuite as the first 2 Mbytes on each multihost disk.

B.2.9 Using the `md.tab` File to Create Metadevices in Disksets

This section describes how to use the `/etc/opt/SUNWmd/md.tab` file to configure metadevices and hot spare pools.

Note - If you have a previously generated `md.tab` file to convert to use disk IDs, you can use the script in Section B.2.4 “DID Conversion Script” on page B-8, to help with the conversion.

B.2.9.1 Creating an `md.tab` File

The `/etc/opt/SUNWmd/md.tab` file can be used by the `metainit(1M)` command to configure metadevices and hot spare pools in a batch-like mode. Solstice DiskSuite does not store configuration information in the `md.tab` file. The only way information appears in the `md.tab` is through editing it by hand.

When using the `md.tab` file, each metadevice or hot spare pool in the file must have a unique entry. Entries can include simple metadevices (stripes, concatenations, and concatenations of stripes); mirrors, trans metadevices, and RAID5 metadevices; and hot spare pools.

Note - Because `md.tab` only contains entries that are manually included in it, you should not rely on the file for the current configuration of metadevices, hot spare pools, and replicas on the system at any given time.

Tabs, spaces, comments (preceded by a pound sign, #), and continuation of lines (preceded by a backslash-newline), are allowed.

B.2.9.2 `md.tab` File Creation Guidelines

Follow these guidelines when setting up your disk configuration and the associated `md.tab` file.

- It is advisable to maintain identical `md.tab` files on each node in the cluster to ease administration.
- A multihost disk and all the partitions found on that disk can be included in no more than one diskset.
- All metadevices used by data services must be fully mirrored. Two-way mirrors are recommended, but three-way mirrors are acceptable.
- No components of a submirror for a given mirror should be found on the same controller as any other component in any other submirror used to define that mirror.
- If more than two disk strings are used, each diskset must include disks from at least three separate controllers. If only two disk strings are used, each diskset must include disks from the two controllers and *mediators* will be configured. See the *Sun Cluster 2.2 System Administration Guide* for more information about using dual-string mediators.
- Hot spares are recommended, but not required. If hot spares are used, configure them so that the activation of any hot spare drive will not result in components of a submirror for a given metamirror sharing the same controller with any other component in any other submirror used to define that given metamirror.
- If you are using Solaris UFS logging, you only need to set up mirrored metadevices in `md.tab` files, transdevices are not necessary.
- If your are using Solstice DiskSuite logging, create multihost file systems on trans metadevices only. Both the logging and master device components of each trans metadevice must be mirrored.
- If your are using Solstice DiskSuite logging, in consideration of performance, do not share spindles between logging and master devices of the same trans metadevice, unless the devices are striped across multiple drives.
- Each diskset has a small “administrative file system” associated with it. This file system is not NFS shared. It is used for data service-specific state or configuration information.

B.2.9.3 Sample `md.tab` File

The ordering of lines in the `md.tab` file is not important, but construct your file in the "top down" fashion described below. The following sample `md.tab` file defines the metadevices for the diskset named `green`. The # character can be used to annotate the file:

```
# administrative file system for logical host mounted under /green
 green/d0 -t green/d1 green/d4
  green/d1 -m green/d2 green/d3
      green/d2 1 1 /dev/did/rdsk/d1s4
      green/d3 1 1 /dev/did/rdsk/d2s4
  green/d4 -m green/d5 green/d6
      green/d5 1 1 /dev/did/rdsk/d3s5
      green/d6 1 1 /dev/did/rdsk/d4s5

 # /green/web
 green/d10 -t green/d11 green/d14
  green/d11 -m green/d12 green/d13
      green/d12 1 1 /dev/did/rdsk/d1s0
      green/d13 1 1 /dev/did/rdsk/d2s0
  green/d14 -m green/d15 green/d16
      green/d15 1 1 /dev/did/rdsk/d3s6
      green/d16 1 1 /dev/did/rdsk/d4s6

 #/green/home to be NFS-shared
 green/d20 -t green/d21 green/d24
  green/d21 -m green/d22 green/d23
      green/d22 1 1 /dev/did/rdsk/d3s0
      green/d23 1 1 /dev/did/rdsk/d4s0
  green/d24 -m green/d25 green/d26
      green/d25 1 1 /dev/did/rdsk/d1s6
      green/d26 1 1 /dev/did/rdsk/d2s6
```

The first line defines the administrative file system as the trans metadevice `d0` to consist of a master (UFS) metadevice `d1` and a log device `d4`. The `-t` signifies this is a trans metadevice; the master and log devices are implied by their position after the `-t` flag.

The second line defines the master device as a mirror of the metadevices. The `-m` in this definition signifies a mirror device.

```
green/d1 -m green/d2 green/d3
```

The fifth line similarly defines the log device, `d4`, as a mirror of metadevices.

```
green/d4 -m green/d5 green/d6
```

The third line defines the first submirror of the master device as a one-way stripe.

```
green/d2 1 1 /dev/did/rdsk/d1s4
```

The next line defines the other master submirror.

```
green/d3 1 1 /dev/did/rdsk/d2s4
```

Finally, the log device submirrors are defined. In this example, simple metadevices for each submirror are created.

```
green/d5 1 1 /dev/did/rdsk/d3s
 green/d6 1 1 /dev/did/rdsk/d4s5
```

Similarly, mirrors are created for two other applications: `d10` will contain a Web server and files, and `d20` will contain an NFS-shared file system.

If you have existing data on the disks that will be used for the submirrors, you must back up the data before metadevice setup and restore it onto the mirror.

▼ How to Initialize the `md.tab` File

This procedure assumes that you have ownership of the diskset on the node where the command is executed. It also assumed that you have configured identical `md.tab` files on each node in the cluster. These files must be located in the `/etc/opt/SUNWmd` directory.

1. **As root, initialize the `md.tab` file by running the `metainit(1M)` command.**
   a. **Take control of the diskset:**

```
phys-hahost1# metaset -s hahost1 -t
```

   b. **Initialize the `md.tab` file. The `–a` option activates all metadevices defined in the `md.tab` file. For example, this command initializes the `md.tab` file for diskset `hahost1`.**

```
phys-hahost1# metainit -s hahost1 -a
```

   c. **Repeat this for each diskset in the cluster.**
   If necessary, run the `metainit(1M)` command from another node that has connectivity to the disks. This is required for clustered pair and ring topologies, where the disks are not accessible by all nodes.

2. **Use the** `metastat(1M)` **command to check the status of the metadevices.**

```
phys-hahost1# metastat -s hahost1
```

B.2.10 Creating File Systems Within a Diskset

You can create logging UFS multihost file systems in the Sun Cluster/Solstice DiskSuite environment by using either of these methods:

- Creating a metatrans device, consisting of a master device and a logging device
- Using the logging feature in the Solaris 7 operating environment

▼ How to Create Multihost UFS File Systems

This procedure explains how to create multihost UFS file systems, including the administrative file system that is a requirement for each diskset.

1. **For each diskset, identify or create the metadevices to contain the file systems.**

   It is recommended that you create a trans metadevice for the administrative file system consisting of these components:

   - Master device: mirror using two 2-Mbyte slices on Slice 4 on Drive 1 on the first two controllers
   - Logging device: mirror using two 2-Mbyte slices on Slice 6 on Drive 1 on the first two controllers

2. **Make sure you have ownership of the diskset.**

   If you are creating multihost file systems as part of your initial setup, you should already have diskset ownership. If necessary, refer to the `metaset(1M)` man page for information on taking diskset ownership.

3. **Create the HA administrative file system.**

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

   This example creates the file system on the trans metadevice `d11`.

```
phys-hahost1# newfs /dev/md/hahost1/rdsk/d11
```

Caution - The process of creating the file system destroys any data on the disks.

   b. **Create the directory mount point for the HA administrative file system.**

This example uses the logical host name as the mount point.

```
phys-hahost1# mkdir /hahost1
```

c. Mount the HA administrative file system.

```
phys-hahost1# mount /dev/md/hahost1/dsk/d11 /hahost1
```

4. Create the multihost UFS file systems.

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

This example creates file systems on trans metadevices `d1`, `d2`, `d3`, and `d4`.

```
phys-hahost1# newfs /dev/md/hahost1/rdsk/d1
phys-hahost1# newfs /dev/md/hahost1/rdsk/d2
phys-hahost1# newfs /dev/md/hahost1/rdsk/d3
phys-hahost1# newfs /dev/md/hahost1/rdsk/d4
```

Caution - The process of creating the file system destroys any data on the disks.

b. Create the directory mount points for the multihost UFS file systems.

```
phys-hahost1# mkdir /hahost1/1
phys-hahost1# mkdir /hahost1/2
phys-hahost1# mkdir /hahost1/3
phys-hahost1# mkdir /hahost1/4
```

5. Create the `/etc/opt/SUNWcluster/conf/hanfs` **directory.**

6. Edit the `/etc/opt/SUNWcluster/conf/hanfs/vfstab.`***logicalhost*** **file to update the administrative and multihost UFS file system information.**

Make sure that all cluster nodes' `vfstab.`*logicalhost* files 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    -
```

7. Release ownership of the diskset.

Unmount file systems first, if necessary.

Because the node performing the work on the diskset takes implicit ownership of the diskset, it needs to release this ownership when done.

```
phys-hahost1# metaset -s hahost1 -r
```

8. (Optional) To make file systems NFS-sharable, refer to Chapter 11."

B.3 Solstice DiskSuite Configuration Examples

The following example helps to explain the process for determining the number of disks to place in each diskset when using Solstice DiskSuite. It assumes that you are using three SPARCstorage Arrays as your disk expansion units. In this example, existing applications are running over NFS (two file systems of five Gbytes each) and two Oracle databases (one 5 Gbytes and one 10 Gbytes).

Table B–4 shows the calculations used to determine the number of drives needed in the sample configuration. If you have three SPARCstorage Arrays, you would need 28 drives that would be divided as evenly as possible among each of the three

arrays. Note that the five Gbyte file systems were given an additional Gbyte of disk space because the number of disks needed was rounded up.

TABLE B–4

Use	Data	Disk Storage Needed	Drives Needed
nfs1	5 Gbytes	3x2.1 Gbyte disks * 2 (Mirror)	6
nfs2	5 Gbytes	3x2.1 Gbyte disks * 2 (Mirror)	6
oracle1	5 Gbytes	3x2.1 Gbyte disks * 2 (Mirror)	6
oracle2	10 Gbytes	5x2.1 Gbyte disks * 2 (Mirror)	10

Table B–5 shows the allocation of drives among the two logical hosts and four data services.

TABLE B–5 Division of Disksets

Logical host (diskset)	Data Services	Disks	SPARCstorage Array 1	SPARCstorage Array 2	SPARCstorage Array 3
`hahost1`	nfs1/oracle1	12	4	4	4
`hahost2`	nfs2/oracle2	16	5	6	5

Initially, four disks on each SPARCstorage Array (so a total of 12 disks) are assigned to `hahost1` and five or six disks on each (a total of 16) are assigned to `hahost2`. In Figure B–1, the disk allocation is illustrated. The disks are labeled with the name of the diskset (`1` for `hahost1` and `2` for `hahost2`.)

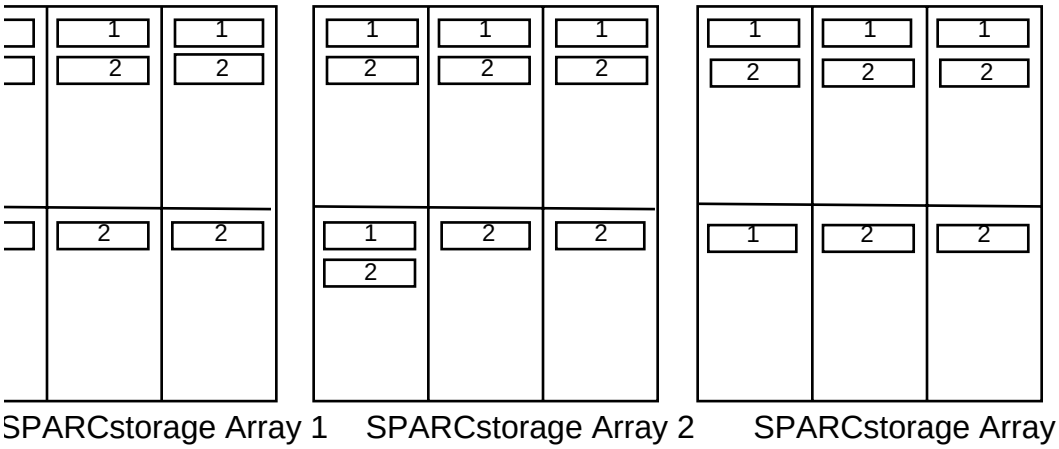


Figure B–1 Sample Diskset Allocation

No hot spares have been assigned to either diskset. A minimum of one hot spare per SPARCstorage Array per diskset enables one drive to be hot spared (restoring full two-way mirroring).

APPENDIX **C**

Configuring Sun StorEdge Volume Manager and Cluster Volume Manager

Configure your local and multihost disks for Sun StorEdge Volume Manager (SSVM) and Cluster Volume Manager (CVM) using the guidelines in this chapter along with the information in Chapter 2. Refer to your SSVM or CVM documentation for additional details.

This appendix includes the following procedures:

C.1 Volume Manager Checklist

Verify that the items listed below are in place before configuring the volume manager:

- The volume manager and VxFS are installed and licensed on each cluster node.

- The volume manager has been installed using the custom install option.

After configuring the volume manager, verify that:

- Only the private disks are included in the root disk group (`rootdg`).
- Disk groups have been deported from all nodes, then imported to the default master node.
- All volumes have been started.

C.2 Configuring SSVM for Sun Cluster

Use this procedure to configure your disk groups, volumes, and file systems for the logical hosts.

Note - This procedure is only applicable for high availability (HA) configurations. If you are using Oracle Parallel Server and Cluster Volume Manager, refer to the *Sun Cluster 2.2 Cluster Volume Manager Guide* for configuration information.

▼ How to Configure SSVM for Sun Cluster

1. **Format the disks to be administered by the volume manager.**

   Use the `fmthard(1M)` command to create a VTOC on each disk with a single Slice 2 defined for the entire disk.

2. **Initialize each disk for use by the volume manager.**

   Use the `vxdiskadd(1M)` command to initialize each disk.

3. **Add each initialized disk to a disk group.**

   Use the `vxdg(1M)` command to add disks to a disk group. You must designate at least one disk to the `rootdg` disk group on each node. When you configure SSVM, you have the option of creating a `rootdg` by encapsulating the boot disk or by creating a simple `rootdg` using a few cylinders of the boot disk. To encapsulate the boot disk, refer to your SSVM documentation. To configure the `rootdg` using part of the boot disk, perform the following steps.

   a. **Create a 10-MByte partition on the boot disk.**

   b. **Add the SSVM packages by using the `pkgadd(1M)` command.**

   c. **Execute the following commands to create the root disk group.**

      In this example, `c0t0d0s7` is the target partition.

```
# vxconfigd -m disable
# vxdctl init
# vxdg init rootdg
# vxdctl add disk c0t0d0s7 type=simple

 vxvm:vxdctl: WARNING: Device c0t0d0s7: Not currently in the configuration

 # vxdisk -f init c0t0d0s7 type=simple
# vxdg -g rootdg adddisk c0t0d0s7
# vxdctl enable
# rm /etc/vx/reconfig.d/state.d/install-db
```

Note - The error message

```
vxvm:vxdctl: WARNING: Device c0t0d0s7:
Not currently in the configuration
```

can be ignored safely at this point.

---

4. **(Optional) Assign hot spares.**

   For each disk group, use the `vxedit(1M)` command to assign one disk as a hot spare for each disk controller.

5. **Reboot all nodes on which you installed SSVM.**

6. **For each disk group, create a volume to be used for the HA administrative file system on the multihost disks.**

   The HA administrative file system is used by Sun Cluster for data service specific state or configuration information.

   Use the `vxassist(1M)` command to create a 10-Mbyte volume mirrored across two controllers for the HA administrative file system. Name this volume *diskgroup*`-stat`.

7. **For each disk group, create the other volumes to be used by HA data services.**

   Use the `vxassist(1M)` command to create these volumes.

8. **Start the volumes.**

   Use the `vxvol(1M)` command to start the volumes.

9. **Create file systems on the volumes.**

Refer to Section C.3 “Configuring VxFS File Systems on the Multihost Disks” on page 14-4, for details on creating the necessary file systems.

C.3 Configuring VxFS File Systems on the Multihost Disks

This section contains procedures to configure multihost VxFS file systems. To configure file systems to be shared by NFS, refer to Chapter 11.

▼ How to Configure VxFS File Systems on the Multihost Disks

1. **Use the `mkfs(1M)` command to create file systems on the volumes.**

   Before you can run the `mkfs(1M)` command on the disk groups, you might need to take ownership of the disk group containing the volume. Do this by importing the disk group to the active node using the `vxdg(1M)` command.

   ```
   phys-hahost1# vxdg import diskgroup
   ```

   a. **Create the HA administrative file systems on the volumes.**

      Run the `mkfs(1M)` command on each volume in the configuration.

      ```
      phys-hahost1# mkfs /dev/vx/rdsk/diskgroup/diskgroup-stat
      ```

   b. **Create file systems for all volumes.**

      These volumes will be mounted by the logical hosts.

      ```
      phys-hahost1# mkfs -F vxfs /dev/vx/rdsk/diskgroup/volume
      ```

2. **Create a directory mount point for the HA administrative file system.**

   ```
   phys-hahost1# mkdir /logicalhost
   ```

3. **Mount the HA administrative file system.**

```
phys-hahost1# mount /dev/vx/dsk/diskgroup/diskgroup-stat/logicalhost
```

4. **Create mount points for the data service file systems created in Step 1b.**

```
phys-hahost1# mkdir /logicalhost/volume
```

5. **Create the** `/etc/opt/SUNWcluster/conf/hanfs` **directory.**

6. **Create and edit the** `/etc/opt/SUNWcluster/conf/hanfs/vfstab.`***logicalhost*** **file to update the administrative and multihost VxFS file system information.**

   Make sure that entries for each disk group appear in the `vfstab.`*logicalhost* files on each node that is a potential master of the disk group. Make sure the `vfstab.`*logicalhost* files contain the same information. Use the `cconsole(1)` facility to make simultaneous edits to `vfstab.`*logicalhost* files on all nodes in the cluster.

   Here is a sample `/etc/vfstab.`*logicalhost* file showing the administrative file system and two other VxFS file systems. In this example, `dg1` is the disk group name and `hahost1` is the logical host name.

```
/dev/vx/dsk/dg1/dg1-stat     /dev/vx/rdsk/dg1/dg1-stat     /hahost1 vxfs - yes -
/dev/vx/dsk/dg1/vol_1        /dev/vx/rdsk/dg1/vol_1        /hahost1/vol_1 vxfs - yes -
/dev/vx/dsk/dg1/vol_2        /dev/vx/rdsk/dg1/vol_2        /hahost1/vol_1 vxfs - yes -
```

7. **Unmount the HA administrative file systems that you mounted in Step 3 on page C–5.**

```
phys-hahost1# umount /logicalhost
```

8. **Export the disk groups.**

   If you took ownership of the disk groups on the active node by using the `vxdg(1M)` command before creating the file systems, release ownership of the disk groups once file system creation is complete.

```
phys-hahost1# vxdg deport diskgroup
```

9. Import the disk groups to their default masters.

It is most convenient to create and populate disk groups from the active node that is the default master of the particular disk group.

Each disk group should be imported onto the default master node using the `-t` option. The `-t` option is important, as it prevents the import from persisting across the next boot.

```
phys-hahost1# vxdg -t import diskgroup
```

10. (Optional) To make file systems NFS-sharable, refer to Chapter 11.

C.4 Administering the Pseudo-Device Major Number

To avoid "Stale File handle" errors on the client on NFS failovers, the `vxio` driver must have 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. Use the following procedures to verify and change the pseudo-device major numbers.

▼ How to Verify the Pseudo-Device Major Number (SSVM)

1. Verify the pseudo-device major number on all nodes.

For example, enter the following:

```
# grep vxio /etc/name_to_major
vxio 45
```

2. **If the pseudo-device number is not the same on all nodes, stop all activity on the system and edit the** `/etc/name_to_major` **file to make the number identical on all cluster nodes.**

   Be sure that the number is unique in the `/etc/name_to_major` file for each node. A quick way to do this is to find, by visual inspection, the maximum number assigned on each node in the `/etc/name_to_major` file, compute the maximum of these numbers, add one, then assign the sum to the `vxio` driver.

3. **Reboot the system immediately after the number is changed.**

4. **(Optional) If the system reports disk group errors and the cluster will not start, you might need to perform these steps.**

   a. **Use the** `vxedit(1M)` **command to change the "failing" field to "off" for affected subdisks. Refer to the** `vxedit(1M)` **man page for more information.**

   b. **Make sure all volumes are enabled and active.**

▼ How to Change the Pseudo-Device Major Number (SSVM)

1. **Unencapsulate the root disk using the SSVM** `upgrade_start` **script.**

   Find the script in the `/Tools/scripts` directory on your SSVM media. Run the script from only one node. In this example, *CDROM_path* is the path to the tools on the SSVM media.

```
phys-hahost1# CDROM_path/Tools/scripts/upgrade_start
```

2. **Reboot the node.**

3. **Edit the** `/etc/name_to_major` **file and remove the appropriate entry, for example,** `/dev/vx/{dsk,rdsk,dmp,rdmp}`**.**

4. **Reboot the node.**

5. **Run the following command:**

```
phys-hahost1# vxconfigd -k -r reset
```

6. **Re-encapsulate the root disk using the SSVM** `upgrade_finish` **script.**

   Find the script in the `/Tools/scripts` directory on your SSVM media. Run the script from only one node.

```
phys-hahost1# CDROM_path/Tools/scripts/upgrade_finish
```

7. **Reboot the node.**

C.5 Configuring the Shared CCD Volume

You use the `confccdssa(1M)` command to create a disk group and volume to be used to store the CCD database. This is supported only on two-node clusters using Sun StorEdge Volume Manager or Cluster Volume Manager as the volume manager. This is not supported on clusters using Solstice DiskSuite.

Note - The root disk group (`rootdg`) must be initialized before you run the `confccdssa(1M)` command.

▼ How to Configure the Shared CCD Volume

1. **Make sure you have configured a volume for the CCD.**

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

```
# scconf clustername -S ccdvol
```

2. **Run the** `confccdssa(1M)` **command on only one node, and use it to select disks for the CCD.**

   Select two disks from the shared disk expansion unit on which the shared CCD volume will be constructed:

```
# /opt/SUNWcluster/bin/confccdssa clustername

 On a 2-node configured cluster you may select two disks
that are shared between the 2 nodes to store the CCD
```

(continued)

```
database in case of a single node failure.

 Please, select the disks you want to use from the following list:

 Select devices from list.
 Type the number corresponding to the desired selection.
 For example: 1<CR>

 1) SSA:00000078C9BF
 2) SSA:00000080295E
 3) DISK:c3t32d0s2:9725B71845
 4) DISK:c3t33d0s2:9725B70870
 Device 1: 3

 Disk c3t32d0s2 with serial id 9725B71845 has been selected
 as device 1.

 Select devices from list.
 Type the number corresponding to the desired selection.
 For example: 1<CR>

 1) SSA:00000078C9BF
 2) SSA:00000080295E
 3) DISK:c3t33d0s2:9725B70870
 4) DISK:c3t34d0s2:9725B71240
 Device 2: 4

 Disk c3t34d0s2 with serial id 9725B71240 has been selected
 as device 2.

 newfs: construct a new file system /dev/vx/rdsk/sc_dg/ccdvol:
 (y/n)? y
...
```

The two disks selected can no longer be included in any other disk group. Once selected, the volume is created and a file system is laid out on the volume. See the `confccdssa(1M)` man page for more details.