Bull

HACMP 4.3.1
HANFS Installation & Administration Guide

AIX

ORDER REFERENCE
86 A2 61KX 01

Bull

HACMP 4.3.1
HANFS Installation & Administration Guide

AIX

Software

August 1999

BULL ELECTRONICS ANGERS
CEDOC
34 Rue du Nid de Pie – BP 428
49004 ANGERS CEDEX 01
FRANCE

ORDER REFERENCE
86 A2 61KX 01

Printed in France

Suggestions and criticisms concerning the form, content, and presentation of this book are invited. A form is provided at the end of this book for this purpose.

To order additional copies of this book or other Bull Technical Publications, you are invited to use the Ordering Form also provided at the end of this book.

Trademarks and Acknowledgements

We acknowledge the right of proprietors of trademarks mentioned in this book.

AIX® is a registered trademark of International Business Machines Corporation, and is being used under licence.

UNIX is a registered trademark in the United States of America and other countries licensed exclusively through the Open Group.

Year 2000

The product documented in this manual is Year 2000 Ready.

Contents

About This Guide

This guide provides information necessary to plan, install, configure, maintain, and troubleshoot the High Availability for Network File System for AIX (HANFS for AIX) software.

Who Should Use This Guide

This guide is intended for system administrators and customer engineers responsible for:

- Planning hardware and software resources for an HANFS for AIX cluster
- Installing and configuring an HANFS for AIX cluster
- Maintaining and troubleshooting an HANFS for AIX cluster.

As a prerequisite to installing the HANFS for AIX software, you should be familiar with:

- System components (including disk devices, cabling, and network adapters)
- The AIX operating system, including the Logical Volume Manager subsystem
- The System Management Interface Tool (SMIT)
- Communications, including the TCP/IP subsystem
- The Network File System (NFS).

Before You Begin

Appendix A, Planning Worksheets, contains blank copies of the worksheets referred to in this guide. These worksheets help you plan, install, configure, and maintain an HANFS for AIX cluster. Complete the required worksheets before installing the HANFS for AIX software.

The examples that rely on SMIT assume you are using AIX from an ASCII display. SMIT is also available within the AIXwindows environment.

Highlighting

The following highlighting conventions are used in this guide:

Italic	Identifies variables in command syntax, new terms and concepts, or indicates emphasis.
Bold	Identifies pathnames, commands, subroutines, keywords, files, structures, directories, and other items whose names are predefined by the system. Also identifies graphical objects such as buttons, labels, and icons that the user selects.
`Monospace`	Identifies examples of specific data values, examples of text similar to what you might see displayed, examples of program code similar to what you might write as a programmer, messages from the system, or information that you should actually type.

ISO 9000

ISO 9000 registered quality systems were used in the development and manufacturing of this product.

Related Publications

The following publications provide additional information about the High Availability Cluster Multi-Processing for AIX (HACMP for AIX) software:

- *Release Notes* in **/usr/lpp/cluster/doc/release_notes** describe hardware and software requirements
- *HACMP for AIX, Version 4.3.1: Concepts and Facilities*, order number SC23-4276-01
- *HACMP for AIX, Version 4.3.1: Planning Guide*, order number SC23-4277-01
- *HACMP for AIX, Version 4.3.1: Installation Guide*, order number SC23-4278-01
- *HACMP for AIX, Version 4.3.1: Administration Guide*, order number SC23-4279-01
- *HACMP for AIX, Version 4.3.1: Troubleshooting Guide*, order number SC23-4280-01
- *HACMP for AIX, Version 4.3.1: Programming Locking Applications*, order number SC23-4281-01
- *HACMP for AIX, Version 4.3.1: Programming Client Applications*, order number SC23-4282-01*HACMP for AIX, Version 4.3.1: Enhanced Scalability Installation and Administration Guide*, Volumes I and II, order numbers SC23-4284-01 and SC23-4306
- *HACMP for AIX, Version 4.3.1: Master Index and Glossary*, order number SC23-4285-01
- *IBM International Program License Agreement*, order number S29H-1286

The IBM AIX document set, as well as manuals accompanying machine and disk hardware, also provide relevant information.

Ordering Publications

To order additional copies of this guide, use order number SC23-4283-01.

You can order additional IBM publications from your IBM sales representative or, in the U.S., from IBM Customer Publications Support at 1-800-879-2755. If you believe you are entitled to publications that were not shipped with your HACMP for AIX purchase, contact your IBM sales representative or Customer Publications Support for assistance.

On the World Wide Web, enter the following URL to access an online library of documentation covering AIX, RS/6000, and related products:

`http://www.rs6000.ibm.com/aix/library`

Part 1 HANFS for AIX Concepts

The information contained in this part introduces you to basic concepts about HANFS for AIX, such as required hardware and software components, cluster resources and resource groups, standby configurations, takeover configurations, cluster events, and ways to reduce unscheduled down time.

Chapter 1 Overview of HANFS for AIX

This chapter describes the High Availability for Network File System for AIX (HANFS for AIX) software.

High Availability for Network File System for AIX

The HANFS for AIX software provides a reliable NFS server capability by allowing a backup processor to recover current NFS activity should the primary NFS server fail. HANFS for AIX takes advantage of AIX extensions to the standard NFS functionality that enable it to handle duplicate requests correctly and restore lock state during NFS server fallover and reintegration.

HANFS for AIX is based on the High Availability Cluster Multi-Processing for AIX, Version 4.3.1 (HACMP for AIX) product architecture, which ensures that critical resources, configured as part of a cluster, are highly available for processing. The HANFS for AIX software extends HACMP for AIX by taking advantage of AIX extensions to the standard NFS functionality that enable it to handle duplicate requests correctly and restore lock state during NFS server fallover and reintegration.

Note: A cluster cannot be mixed—that is, have some nodes running the HANFS for AIX software and other nodes running the HACMP for AIX software. A single cluster must either have all nodes running the HANFS for AIX software or all nodes running the HACMP for AIX software. Distinct HANFS and HACMP clusters, however, are allowed on the same physical network.

Handling Duplicate Requests

The HANFS for AIX software provides the NFS duplicate cache (dupcache), which allows NFS to ignore duplicate requests. For example, a duplicate **delete** of a file would fail but a duplicate **read** would succeed. For each entry in the NFS duplicate cache, the software creates a Journaled File System (JFS) log entry so that the NFS duplicate cache can be rebuilt during the log redo portion of a file system consistency check (**fsck**). Since the cache is non-volatile, the backup processor can restore the cache after an NFS server has crashed.

Restoring Lock State

HANFS for AIX provides extensions to the **rpc.statd** daemon so that client lock requests can be tracked and reclaimed after fallover. The **fsck** and log redo operations performed on the file system rebuild the NFS duplicate cache on the peer node so that duplicate requests are handled correctly.

Since the peer node has a record of all clients that held locks on the server (the **rpc.statd** extensions), it informs the clients to resubmit lock requests so that the lock state can be rebuilt. This operation is performed by the NFS **rpc.lockd** daemon during initialization.

HANFS for AIX Cluster Hardware Components

An HANFS for AIX cluster has the following components:

- Nodes
- Shared external disk devices
- Networks
- Network adapters.

Nodes

Nodes form the core of an HANFS for AIX cluster. A node is a processor that runs both AIX and the HANFS for AIX software. The HANFS for AIX software can run on RS/6000 uniprocessors, symmetric multiprocessors (SMPs), or SP processors.

The HANFS for AIX software supports two nodes in a cluster.

Shared External Disk Devices

Each node must have access to one or more shared external disk devices. A *shared external disk device* is a disk physically connected to both nodes in the cluster. The shared disk stores mission-critical data, which is typically mirrored for data redundancy. A node in an HANFS for AIX cluster must also have internal disks that store the operating system and application binaries, which are not shared.

The HANFS for AIX software supports shared external disk configurations that use SCSI-2 Differential disks and disk arrays, IBM 9333 serial disks, and IBM Serial Storage Architecture (SSA) disk subsystems.

Only one connection is active at any given time, and the node with the active connection owns the disk. Disk takeover occurs when the node that currently owns the disk leaves the cluster and the surviving node assumes ownership of the shared disk.

Networks

As an independent, layered component of AIX, the HANFS for AIX software is designed to work with any TCP/IP-based network. The HANFS for AIX software has been tested with Ethernet, Token-Ring, Fiber Distributed Data Interchange (FDDI), Serial Optical Channel Connector (SOCC), and Serial Line Internet Protocol (SLIP) networks.

Types of Networks

The HANFS for AIX software defines three types of networks:

- **Public network**—Connects the nodes and allows clients to access the cluster nodes. Ethernet, Token-Ring, and FDDI networks can be defined as public networks. A SLIP line, which does not provide any client access, can also be defined as a public network.
- **Private Network**—Provides point-to-point communication between two nodes; it does not allow client access. Ethernet, Token-Ring, FDDI, SOCC, and SP Switch networks can be defined as private networks. The High Performance Switch (HPS) used by the SP machine is a special case; it is a private network that can also connect clients.

- **Serial network**—Provides a point-to-point connection between two cluster nodes used by HANFS for AIX for control messages and heartbeat traffic should the TCP/IP subsystem fail. A serial network can be a SCSI-2 Differential bus using target mode SCSI or an RS232 serial line.

Network Adapters

Typically, a node should have at least two network adapters—a service adapter and a standby adapter—for each connected network. For serial or SP Switch networks, however, this requirement does not apply.

Service network adapter—Primary connection between a node and a network. A node has one service network adapter for each physical network to which it connects. This adapter is used for cluster TCP/IP traffic. Its address is published by the Cluster Information Program (Clinfo) to application programs that want to use cluster services.

- **Standby network adapter**—Backs up a service network adapter. The service network adapter can be on the local node or on the remote node (since address takeover must be enabled). If a service network adapter on the local node fails, the HANFS for AIX software swaps the standby network address with the service network address. If the remote node fails, the standby network adapter on the local node assumes the IP address of the service network adapter on the failed node.

Note: In clusters defined on SP systems, you do not need standby adapters for the SP Switch. The SP Switch uses IP address aliasing to permit IP address takeover. For more information, see Chapter C, Installing and Configuring HANFS for AIX on RS/6000 SPs.

Assigning a Boot Adapter Label for IP Address Takeover

IP address takeover (IPAT) is an AIX facility that allows one node to acquire the network address of another node in the cluster. The HANFS for AIX software requires that IPAT be enabled. For IPAT to work correctly, you must assign a boot adapter label to the service adapter on each cluster node. Nodes use the boot adapter label after a system reboot and before the HANFS for AIX software is started.

When HANFS for AIX is started on a node, the node's service adapter is reconfigured to use the service label (address) instead of the boot label. If this node should fail, the takeover node acquires the failed node's service address on its standby adapter, making the failure transparent to clients using that specific service address.

During the reintegration of the failed node, which comes up on its boot address, the takeover node will release the service address it acquired from the failed node. Afterwards, the reintegrating node will reconfigure its boot address to its reacquired service address.

- It is important to realize that the boot address does not use a separate physical adapter, but instead is a second name and IP address associated with a service adapter. All cluster nodes must have this entry in the local **/etc/hosts** file and, if applicable, in the **nameserver** configuration.

HANFS for AIX Software Components

The HANFS for AIX software has the following components:

- Cluster Manager
- Cluster SMUX Peer
- Clinfo.

The Cluster Manager

The Cluster Manager runs on each cluster node. The main task of the Cluster Manager is to monitor nodes and networks in the cluster for possible failures. It is responsible for monitoring local hardware and software subsystems, tracking the state of the peer node, and acting appropriately to maintain the availability of cluster resources when a change in the status of the cluster occurs. The Cluster Managers exchange periodic messages, called *keepalives* or *heartbeats*, that provide this monitoring.

Changes in the state of the cluster are referred to as *cluster events*. The Cluster Manager runs scripts in response to cluster events. Cluster events are described in more detail later in this chapter.

Cluster SMUX Peer

The Cluster SMUX Peer provides Simple Network Management Protocol (SNMP) support to client applications. SNMP is an industry-standard specification for monitoring and managing TCP/IP-based networks. SNMP includes a protocol, a database specification, and a set of data objects. A set of data objects forms a *Management Information Base* (MIB). SNMP provides a standard MIB that includes information such as IP addresses and the number of active TCP connections. The actual MIB definitions are encoded into the agents running on a system. The standard SNMP agent is the **snmpd** daemon.

The Cluster SMUX Peer maintains cluster status information in a special MIB. When the Cluster SMUX Peer daemon starts running on a cluster node, it registers with the SNMP daemon, then continually gathers cluster information from the Cluster Manager daemon. The Cluster SMUX Peer daemon maintains an updated topology map of the cluster in the MIB as it tracks events and resulting states of the cluster.

Cluster Information Program

Clinfo is an SNMP-based monitor application. The **clinfo** daemon running on a cluster node, queries the Cluster SMUX Peer daemon (**clsmuxpd**) for updated cluster information. The **clstat** program uses Clinfo to get information about the state of an HANFS for AIX cluster, nodes, and networks.

Clinfo calls the **/usr/sbin/cluster/etc/clinfo.rc** script whenever a cluster, network, or node event occurs. If you are not using the hardware address swapping facility, a copy of the **clinfo.rc** script must exist on each node and client in the cluster in order for all ARP caches to be updated and synchronized. Flushing the ARP cache typically is not necessary if the HACMP for AIX hardware address swapping facility is enabled because hardware address swapping maintains the relationship between a network address and a hardware address.

Note: In a switched Ethernet network, you may need to flush the ARP cache to ensure that the new MAC address is communicated to the switch. To ensure that the MAC address is communicated correctly, refer to the procedures in the *HACMP for AIX Troubleshooting Guide*.

Cluster Resources and Resource Groups

The HANFS for AIX software provides a highly available environment by:

- Identifying the set of *cluster resources* that are essential to processing.
- Defining the *takeover relationships* between the nodes that access these resources.

By identifying resources and defining takeover relationships, the HANFS for AIX software makes numerous cluster configurations possible, providing tremendous flexibility in defining a cluster environment tailored to individual requirements.

Identifying Cluster Resources and Resource Groups

Cluster resources can include both hardware and software:

- Disks
- Volume groups
- File systems
- Network addresses.

To be made highly available by the HANFS for AIX software, each resource must be included in a *resource group*. Resource groups allow you to combine related resources into a single logical entity for easier configuration and management. An HANFS for AIX cluster can have a maximum of 20 resource groups.

You define the fallover relationship of a resource group by assigning it one of the following type designations:

- Cascading
- Rotating.

Note: The HANFS for AIX software does not support concurrent access resource groups.

Cascading Resource Groups

A *cascading resource group* establishes a direct relationship, or ownership, between a node and a resource group. When the owning node (the node with the higher takeover priority) is active in the cluster, the corresponding resource group is owned by that node. If the owning node fails, the takeover node (the node with the lower takeover priority) assumes control of the resource group. When the owning node rejoins the cluster, it takes back control of the resource group. Use cascading resource groups when you have a strong preference for which node should control a resource group. For example, you may want the node with the more powerful processing capabilities to control the resource group.

Note: Only cascading resource groups support automatic NFS-mounting across servers during fallover.

Rotating Resource Groups

A *rotating resource group* is not associated with a specific node. Rather, it binds to a node at cluster startup and is owned by that node until it leaves the cluster. When the owning node leaves, the other node acquires the resource group. The takeover node may be currently active, or it may be in a standby state. When the detached node rejoins the cluster, however, it does not reacquire the resource group; instead, it rejoins as a standby. Use rotating resource groups when it does not matter as much which cluster node controls a resource group.

Note: Rotating resource groups do not support automatic NFS-mounting across servers during fallover. Instead, you must use additional post events or perform NFS-mounting using normal AIX routines.

Setting the Hostname of an HANFS Server

Because the **rpc.statd** daemon is dependent upon a system's hostname, you must associate the hostname with an IP label that will not change while the system is operational.

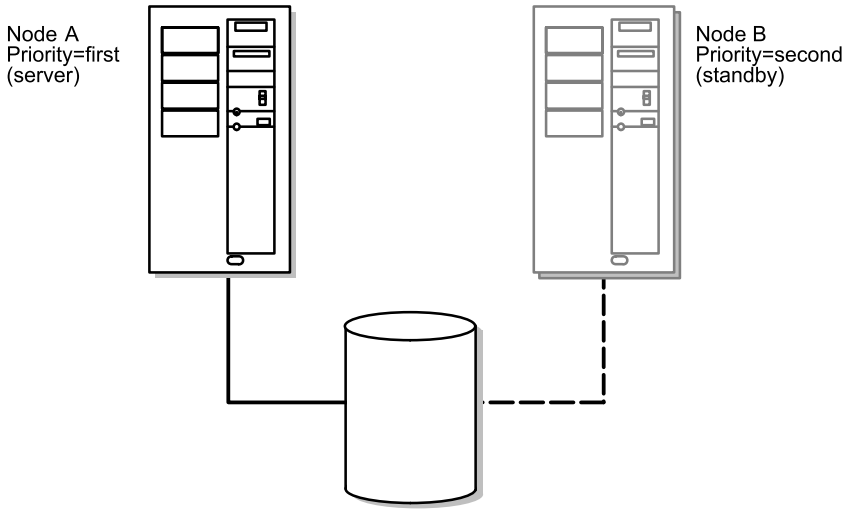
For rotating configurations, IPAT occurs on the service adapter. The hostname, therefore, must be associated with the standby IP address on each node. For cascading configurations, IPAT occurs on the standby adapter. Thus the hostname must be associated with the service IP address. In both configurations, the hostname can be a second label associated with the proper IP address for the service or standby adapter.

Standby Configurations

The standby configuration is a traditional redundant hardware configuration where one node stands idle, waiting for the server node to leave the cluster. The sample standby configurations discussed in this chapter first show how the configuration is defined using cascading resource groups, then how it is defined using rotating resource groups.

Standby Configurations with Cascading Resource Groups

The following figure shows a standby configuration that uses cascading resource groups:



One-for-One Standby with Cascading Resource Groups

In this setup, the cluster resources are defined as part of a single resource group. A resource chain is then defined that consists of both nodes. Node A has first (ownership) priority; Node B has second.

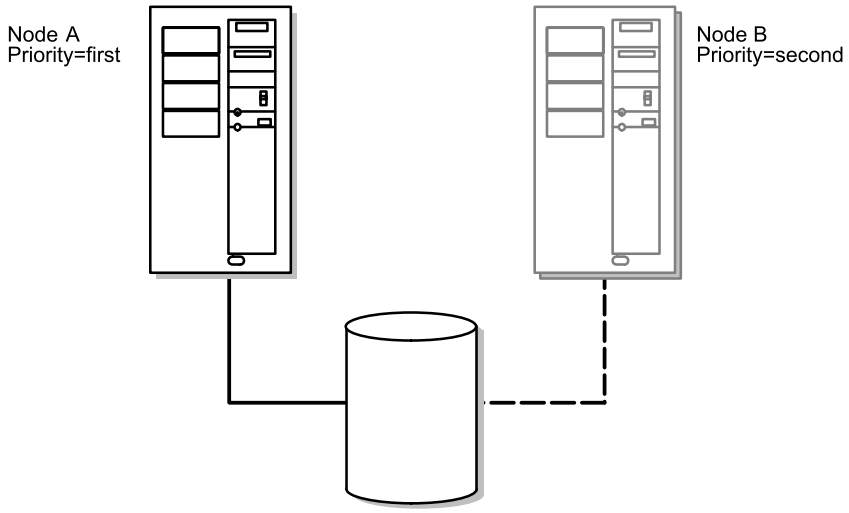
At cluster startup, Node A (which has first priority) assumes ownership of the resource group. Node A is the "server" node. Node B (which has second priority) stands idle, ready should Node A fail or leave the cluster. Node B is, in effect, the "standby."

If the server node leaves the cluster, the standby node assumes control of the resource groups owned by the server, starts the highly available applications, and services clients. The standby node remains active until the node with the higher takeover priority rejoins the cluster. At that point, the standby node releases the resource groups it has taken over, and the server node reclaims them. The standby node then returns to an idle state.

Standby Configurations with Rotating Resource Groups

A standby configuration with rotating resource groups differs from a cascading resource standby configuration in that the ownership of resource groups is not fixed. That is, no strict priority dictating which node owns a resource group exists. Rather, the resource group is associated with an IP address that can rotate among nodes. This makes the role of server and standby fluid, changing over time.

The following figure shows a standby configuration using rotating resource groups:



One-for-One Standby with Rotating Resource Groups

At system startup, the resource group attaches to the node that claims the shared IP address. This node "owns" the resource group for as long as it remains in the cluster. If this node leaves the cluster, the second node assumes the shared IP address and claims ownership of that resource group. Now, this node "owns" the resource group for as long as it remains in the cluster.

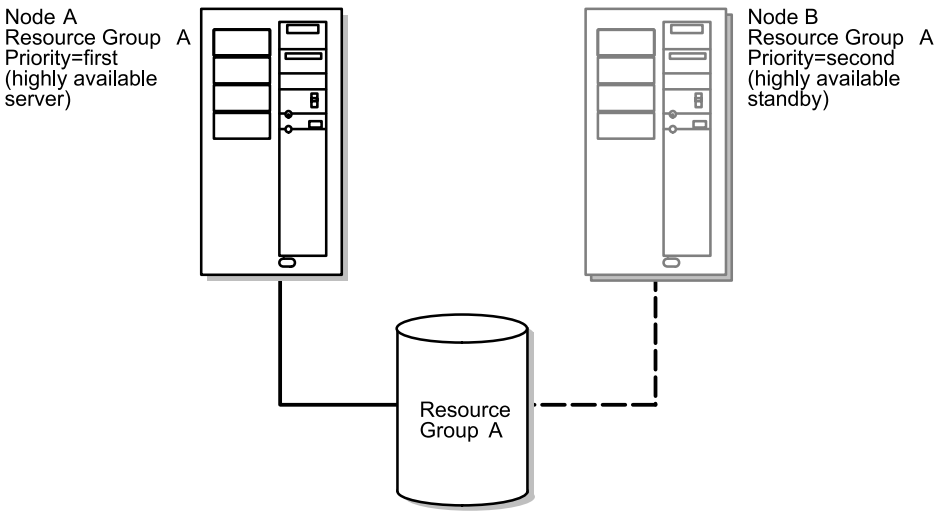
When the node that initially claimed the resource group rejoins the cluster, it does not take it back. Rather, it remains idle for as long as the node currently bound to the shared IP address is active in the cluster. Only if the current owner leaves the cluster does the node that initially "owned" the resource group claim it once again. Thus, ownership of resources rotates between nodes.

Takeover Configurations

Both nodes in a takeover configuration process part of the cluster's workload. There is no standby node. Takeover configurations use hardware resources more efficiently than standby configurations since there is no idle processor. Performance degrades after node failure, however, since the load on the remaining node increases.

One-Sided Takeover Using Cascading Resource Groups

The following figure illustrates a one-sided takeover configuration:



One-sided Takeover Using Cascading Resource Groups

This configuration has both nodes actively processing work, but only one node providing highly available services to cluster members. That is, though there are two sets of resources within the cluster, only one set of resources needs to be highly available. This set of resources is defined as an HANFS for AIX resource group and has a resource chain in which both nodes participate. The second set of resources is not defined as a resource group and, therefore, is not highly available.

At cluster startup, Node A (which has first priority) assumes ownership of Resource Group A. Node A, in effect, "owns" Resource Group A. Node B (which has second priority for Resource Group A) processes its own workload independently of this resource group.

If Node A leaves the cluster, Node B takes control of the shared resources. When Node A rejoins the cluster, Node B releases the shared resources.

If Node B leaves the cluster, however, Node A does not take over any of its resources, since Node B's resources are not defined as part of a highly available resource group in whose chain this node participates.

This configuration is appropriate when a single node is able to run all the critical applications that need to be highly available to cluster clients.

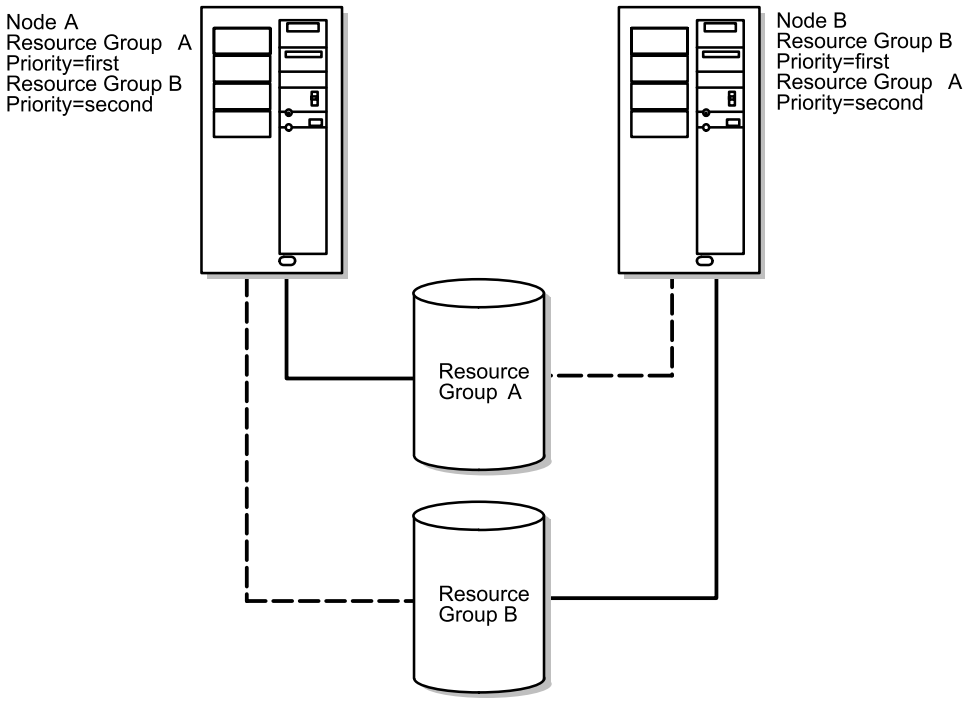
Mutual Takeover Using Cascading Resource Groups

In the mutual takeover configuration, both nodes provide distinct highly available services to cluster clients. For example, each node might run its own instance of a database and access its own disk.

Furthermore, each node has takeover capacity. If one node leaves the cluster, the surviving node takes over the resource groups owned by the departed node.

The mutual takeover configuration is appropriate when each node in the cluster is running critical applications that need to be highly available and when the processors are able to handle the load of more than one node.

The following figure illustrates a two-node mutual takeover configuration:



Mutual Takeover Configuration

The key feature of this configuration is that the cluster's workload is divided, or partitioned, between the nodes. Two separate resource groups exist, along with a separate resource chain for each resource group. The nodes that participate in the resource chains are the same. It is the differing priorities within the chains that designate this configuration as mutual takeover.

The chains for both resource groups consist of Node A and Node B. For Resource Group A, Node A has first takeover priority and Node B has second takeover priority. For Resource Group B, the takeover priorities are reversed. Here, Node B has first takeover priority and Node A has second takeover priority.

At cluster startup, Node A assumes ownership of the Resource Group A, while Node B assumes ownership of Resource Group B.

If either node leaves the cluster, its partner takes control of the departed node's resource group. When the "owner" node for that resource group rejoins the cluster, the takeover node releases it.

Eliminating Single Points of Failure in an HANFS for AIX Cluster

The key facet of a highly available system is its ability to detect and respond to changes that could impair essential services. The HANFS for AIX software allows a cluster to continue to provide application services critical to an installation even though a key system component—a network adapter, for example—is no longer available. When a component becomes unavailable, the HANFS for AIX software is able to detect the loss and shift that component's workload to another component in the cluster.

The HANFS for AIX software enables you to build clusters that are both highly available and scalable by eliminating single points of failure. A *single point of failure* exists when a critical cluster function is provided by a single component. If that component fails, the cluster has no other way to provide that function and essential services become unavailable.

For example, if all the data for a critical application resides on a single disk that is not mirrored, and that disk fails, the disk becomes a single point of failure for the entire system. Clients cannot access that application until the data on the disk is restored.

Potential Single Points of Failure within an HANFS for AIX Cluster

HANFS for AIX provides recovery options for the following cluster components:

- Processors
- Networks and network adapters
- Disks and disk adapters.

To be highly available, a cluster must have no single points of failure. Realize that, while the goal is to eliminate all single points of failure, compromises may have to be made. There is usually a cost associated with eliminating a single point of failure. For example, redundant hardware increases cost. The cost of eliminating a single point of failure should be compared against the cost of losing services should that component fail.

Eliminating Nodes as a Single Point of Failure

Nodes leave the cluster either through a planned transition (a node shutdown or stopping cluster services on a node) or because of a failure.

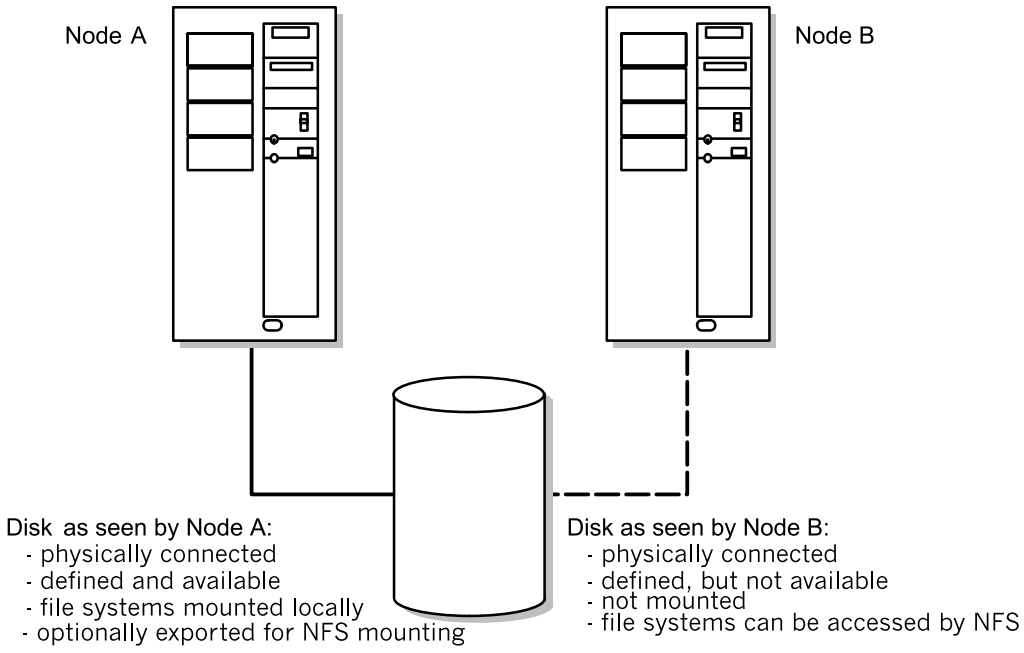
Node failure begins when one node stops receiving keepalive traffic from the other node for a defined period of time. The failing node is removed from the cluster and its resources are taken over by the surviving node, if configured to do so. The surviving node may, for example, take control of the failed node's shared disks. Or, the surviving node may masquerade as the failed node and run that node's processes while still maintaining its own processes. Thus, client applications can switch over to the surviving node for shared-disk and processor services.

The HANFS for AIX software provides two facilities for processing node failure: disk takeover and IP address takeover (with or without hardware address swapping).

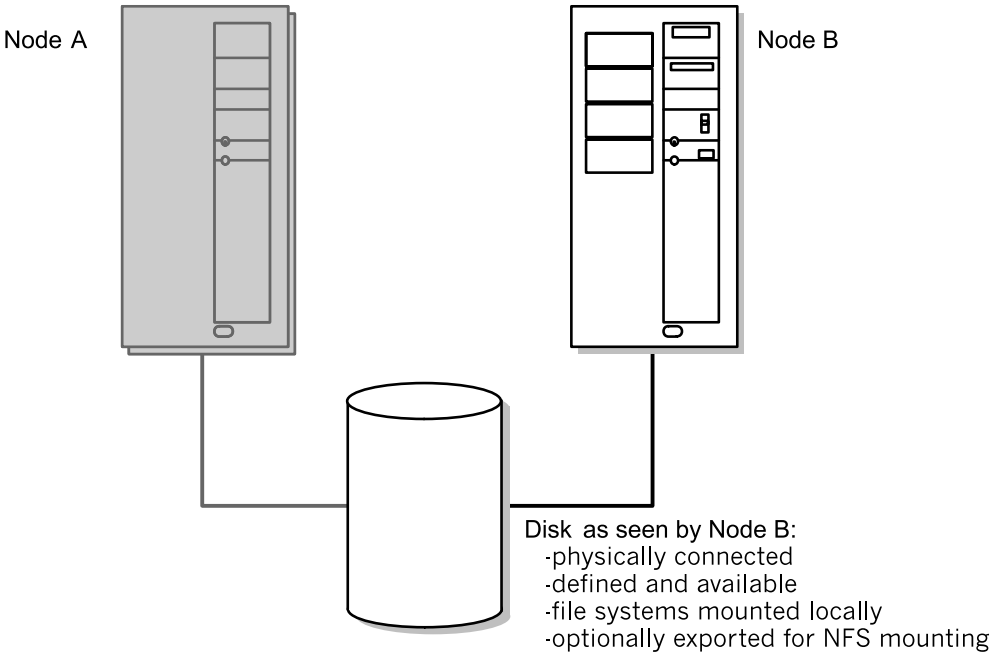
Disk Takeover

In an HANFS for AIX environment, shared disks are physically connected to each node, with only one connection active at any given time, and the node with the active connection owns the disk. *Disk takeover* occurs when the node that currently owns the disk leaves the cluster and the surviving node assumes control of the shared disk so that it remains available.

The following figures illustrate disk takeover:



Before Disk Takeover

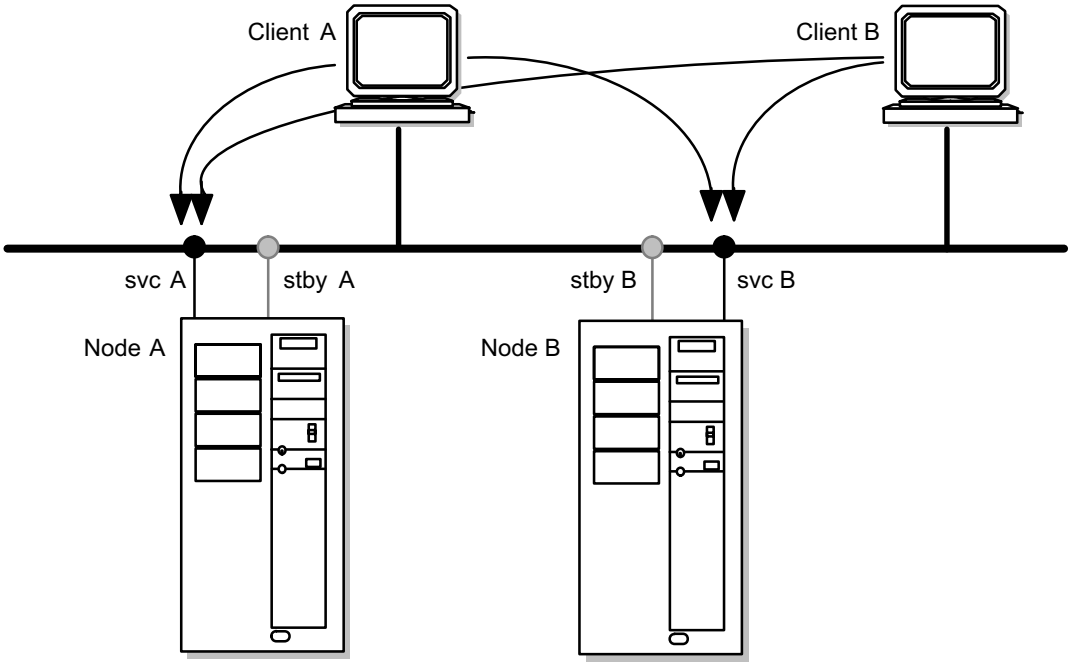


After Disk Takeover

IP Address Takeover

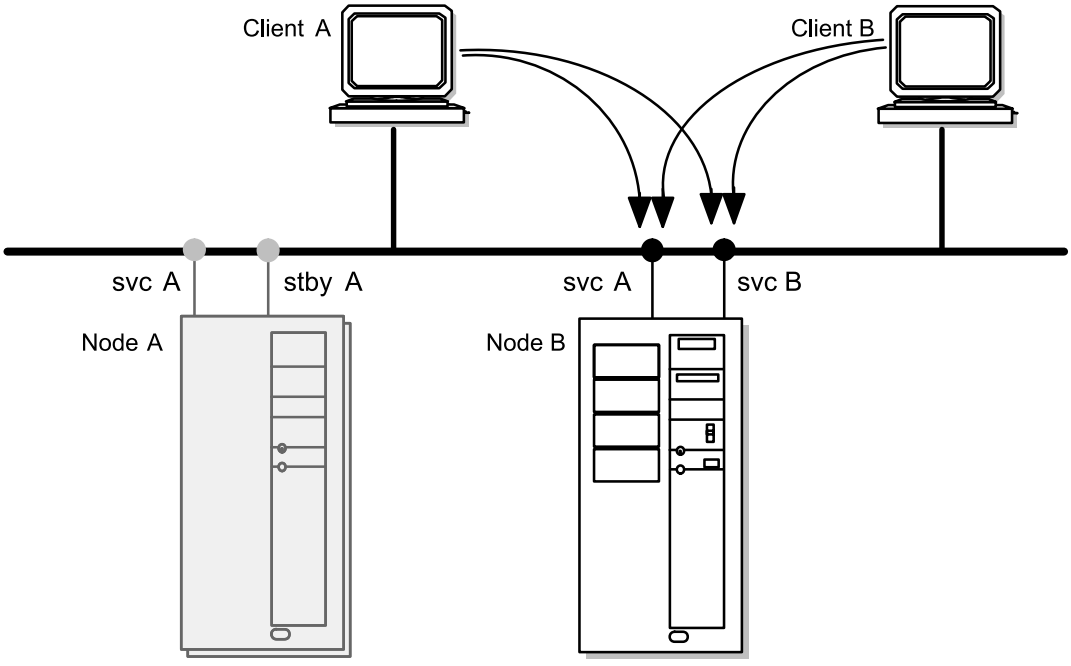
IP address takeover is a networking capability that allows a node to acquire the network address of a node that has left the cluster. IP address takeover is required in an HANFS for AIX cluster.

The following figures illustrate IP address takeover:



Each node provides a separate network service.

Before IP Address Takeover



Node B assumes node A's IP address on its standby interface and provides A's network service to clients.

After IP Address Takeover

Hardware Address Swapping

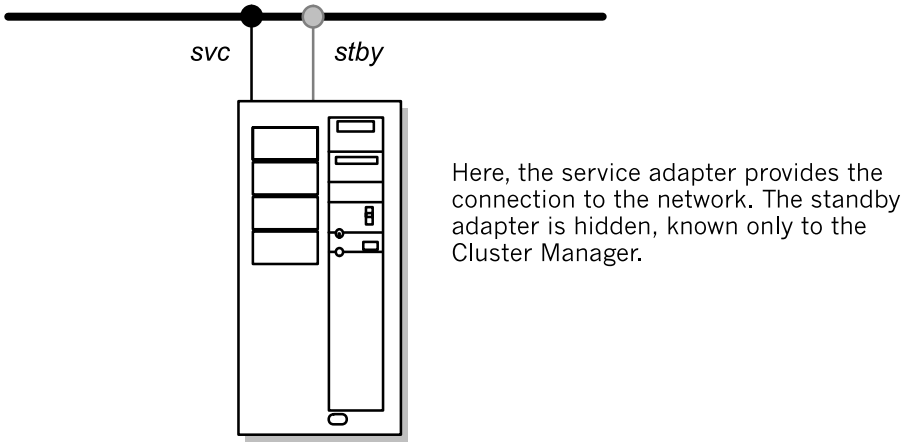
Hardware address swapping works in conjunction with IP address takeover. With hardware swapping enabled, a node also assumes the hardware network address (in addition to the IP address) of a node that has failed so that it can provide the service that the failed node was providing to the cluster's clients.

Without hardware address swapping, TCP/IP clients and routers which reside on the same subnet as the cluster nodes must have their Address Resolution Protocol (ARP) cache updated. The ARP cache maps IP addresses to hardware addresses.

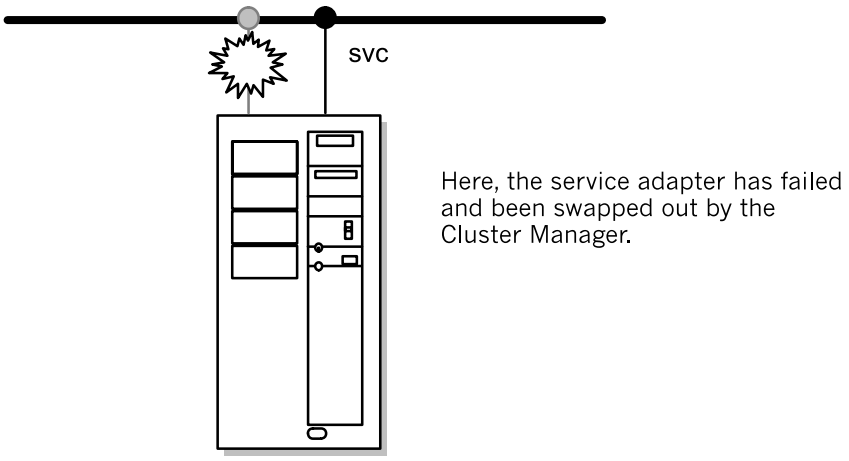
Eliminating Network Adapters as a Single Point of Failure

The HANFS for AIX software handles service and standby network adapter failures. When a service adapter fails, the Cluster Manager swaps the roles of the service and standby adapters on that node. A service adapter failure is transparent other than for a small delay while the system reconfigures itself. While the Cluster Manager does detect a standby adapter failure, it does nothing other than log the event.

The following figures illustrate adapter swapping:



Before Network Adapter Swap



After Network Adapter Swap

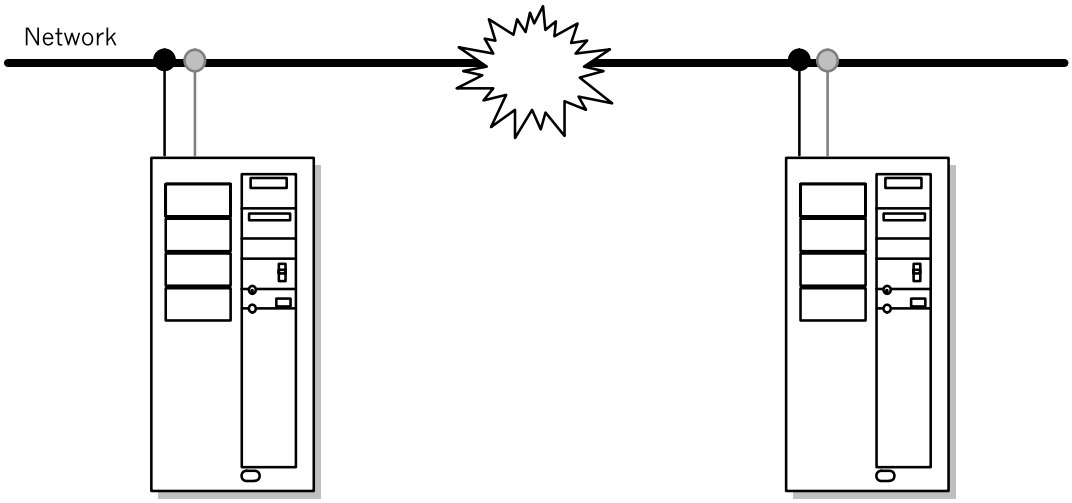
Hardware Address Swapping

Hardware address swapping works in conjunction with adapter swapping (as well as IP address takeover). With hardware address swapping enabled, the standby adapter assumes the hardware network address (in addition to the IP address) of the failed service adapter so that it can provide the service that the failed service adapter was providing to the cluster's clients.

Without hardware address swapping, TCP/IP clients and routers which reside on the same subnet as the cluster nodes must have their Address Resolution Protocol (ARP) cache updated. The ARP cache maps IP addresses to hardware addresses.

Eliminating Networks as a Single Point of Failure

Network failure occurs when the nodes are unable to communicate across that network. The following figure illustrates a network failure:



Here, the network connecting the nodes has failed. The nodes are no longer able to communicate across this network.

Network Failure

The HANFS for AIX software's first line of defense against a network failure is to have the nodes in the cluster connected by multiple networks. If one network fails, the HANFS for AIX software uses a network that is still available for cluster traffic and to monitor the status of the nodes.

The Cluster Manager detects the failure, but takes no action to restore the lost network. You can specify additional actions to process a network failure—for example, rerouting through an alternate network.

Having at least two networks to guard against network failure is highly recommended.

Node Isolation and Partitioned Clusters

Node isolation occurs when all TCP/IP networks connecting the two nodes fail. Each node is then completely isolated from the other. A cluster in which the two nodes are unable to communicate with each other is a *partitioned cluster*.

The problem with a partitioned cluster is that each node interprets the absence of keepalives from its partner to mean that the other node has failed and then generates node failure events. Once this occurs, each node attempts to take over resources from a node that is still active and therefore still legitimately owns those resources. These attempted takeovers can cause unpredictable results in the cluster—for example, data corruption due to a disk being reset.

To guard against the TCP/IP subsystem failure causing node isolation, the nodes should be connected by a point-to-point serial network. This connection reduces the chance of node isolation by allowing the Cluster Managers to communicate even when all TCP/IP-based networks fail.

It is important to understand that the serial network does not carry TCP/IP communication between nodes; it only allows nodes to exchange keepalives and control messages so that the Cluster Manager has accurate information about the status of its partner.

Eliminating Disks and Disk Adapters as a Single Point of Failure

The HANFS for AIX software does not itself directly handle disk and disk adapter failures. Rather, these failures are handled by either AIX mirroring or by internal data redundancy.

For example, if you configure the system with multiple SCSI-2 Differential chains and then mirror the disks across these chains, any single component in the disk subsystem (adapter, cabling, disks) can fail without causing unavailability of data on the disk.

If you are using IBM 7135-110 or 7135-210 RAIDiant Disk Arrays, the disk array itself is responsible for providing data redundancy.

AIX Error Notification Facility

Although the HANFS for AIX software does not monitor the status of disk resources, it does provide a SMIT interface to the AIX Error Notification facility. The AIX Error Notification facility allows you to detect an event not specifically monitored by the HANFS for AIX software—a disk adapter failure, for example—and to program a response to the event.

Permanent hardware errors on disk drives, controllers, or adapters may impact the fault resiliency of data. By monitoring these errors through error notification methods, you can assess the impact of a failure on the cluster's ability to provide high availability. A simple implementation of error notification would be to send a mail message to the system administrator to investigate the problem further. A more complex implementation could include logic to analyze the failure and decide whether to continue processing, stop processing, or escalate the failure to a node failure and have the takeover node make the volume group resources available to clients.

It is strongly recommended that you implement an error notification method for all errors that affect the disk subsystem. Doing so ensures that degraded fault resiliency does not remain undetected.

Automatic Error Notification

You can automatically configure error notification for certain cluster resources. Using this utility, error notification will automatically be turned on or off on the node for particular devices.

Warning: Automatic error notification should be configured only when the cluster is not running.

Select non-recoverable error types are supported by automatic error notification: disk, disk adapter, and SP switch adapter errors. No media errors, recovered errors, or temporary errors are supported.

For more information on automatic error notification, see Chapter 14, Supporting AIX Error Notification.

Error Emulation Functionality

After you have added one or more error notification methods to the AIX Error Notification facility, you can test your methods by emulating an error.

For more information on error emulation, see Chapter 14, Supporting AIX Error Notification.

Reducing Unscheduled Down-Time—Fast Recovery

The HANFS for AIX Version 4.3.1 software provides a fast recovery feature to speed up fallover in large clusters.

Fast Recovery lets you choose a file system consistency check and a file system recovery method:

- If you configure it to do so, it saves time by running **logredo** rather than **fsck** on each file system. If the subsequent **mount** fails, then it runs a full **fsck**.

  If a file system suffers damage in a failure, but can still be mounted, **logredo** may not succeed in fixing the damage, producing an error during data access.
- If you configure it to do so, it saves time by acquiring, releasing, and falling over all resource groups and file systems in parallel, rather than in serial.

  Do not set the system to run these commands in parallel if you have shared, nested file systems. These must be recovered sequentially. (Note that the cluster verification utility, **clverify**, does not report file system and fast recovery inconsistencies.)

Set your choices for these in the **SMIT > Cluster Configuration > Cluster Resources > [name of resource group] > Configure a Resource Group** screen. Your choices affect all the file systems in the resource group. If some file systems need different settings, add them to separate resource groups.

Cluster Events

A *cluster event* is a change in the cluster status that the Cluster Manager detects and processes so that critical resources remain available. A cluster event can be triggered by a change affecting a network adapter, network, or node, or by the cluster reconfiguration process exceeding its time limit.

Detecting Cluster Events

When a cluster event occurs, the Cluster Manager runs the corresponding event script for that event. As the event script is being processed, a series of subevent scripts may be executed. The default scripts are located in the **/usr/sbin/cluster/events** directory.

The Cluster Manager recognizes the following events:

- **node_up** and **node_up_complete** events (a node trying to join the cluster)
- **node_down** and **node_down_complete** events (a node leaving the cluster)
- **network_down** event (a network has failed)
- **network_up** event (a network has connected)
- **swap_adapter** event (a network adapter failed and a new one has taken its place)

Processing Cluster Events

The two primary cluster events that HANFS for AIX software handles are fallover and reintegration. *Fallover* refers to the actions taken by the HANFS for AIX software when a cluster component fails or a node leaves the cluster. *Reintegration* refers to the actions that occur within the cluster when a component that had previously left the cluster returns. Both types of actions are controlled by the event scripts.

Note: Accessing a file system while HANFS for AIX is processing a **node_up** or **node_down** event may cause a delay of several seconds.

Fallover

Nodes leave the cluster either by a planned transition (a node shutdown or stopping cluster services on a node) or by failure. In the former case, the Cluster Manager controls the release of resources held by the exiting node and the acquisition of these resources by the active node. When necessary, you can override the release and acquisition of resources (for example, to perform system maintenance).

Node failure begins when a node monitoring its partner ceases to receive keepalive traffic for a defined period of time. The failing node is removed from the cluster and its resources are taken over by the active node.

If other components fail, such as a network adapter, the Cluster Manager runs an event script to switch network traffic to a standby adapter.

Reintegration

When a node joins a running cluster, the cluster becomes unstable. The currently active node runs event scripts to release any resources the joining node is configured to take over. The joining node then runs an event script to take over these resources. Finally, the joining node becomes a member of the cluster. At this point, the cluster is stable again.

Emulating Events

HANFS for AIX provides an emulation utility to test the effects of running a particular event without modifying the cluster state. The emulation runs on every active cluster node, and the output is stored in an output file on the node from which the emulation was invoked.

For more information on the Event Emulator utility, see the *HACMP for AIX Concepts and Facilities Guide*.

node_up Events

A **node_up** event can be initiated by a node joining the cluster at cluster startup, or rejoining the cluster after having previously left the cluster.

node_down Events

Nodes exchange keepalives so that the Cluster Manager can track the status of the nodes in the cluster. A node that fails or is stopped purposefully no longer sends keepalives. The other Cluster Manager then posts a **node_down** event. Depending on the cluster configuration, the active node takes the necessary actions to get critical applications up and running and to ensure file systems remain available.

A **node_down** event can be initiated by a node:

- Being stopped “gracefully”
- Being stopped “gracefully with takeover”
- Being stopped “forcefully”
- Failing.

Graceful Stop

In a *graceful stop*, the HANFS for AIX software stops on the local node after the **node_down_complete** event releases some or all of the stopped node’s resources. The other node runs the **node_down_complete** event with the “graceful” parameter and does not take over the resources of the stopped node.

Graceful with Takeover Stop

In a *graceful with takeover stop*, the HANFS for AIX software stops after the **node_down_complete** event on the local node releases some or all of its resource groups. The surviving node takes over these resource groups.

Forced Stop

In a *forced stop*, the HANFS for AIX software stops immediately on the local node. The **node_down** event is not run on this node, but it sends a message to the other node to view it as a graceful stop. The Cluster Manager on the remote node processes **node_down** events, but does not take over any resource groups. The stopped node retains control of its resource groups.

Node Failure

When a node fails, the Cluster Manager on that node does not have time to generate a **node_down** event. In this case, the Cluster Manager on the surviving node recognizes a **node_down** event has occurred (when it realizes the failed node is no longer communicating) and triggers **node_down** events.

The **node_down_remote** event initiates a series of subevents that reconfigure the cluster to deal with that failed node. The surviving node will take over the resource groups.

Network Events

Network events occur when the Cluster Manager determines a network has failed or has become available. The default **network_down** event mails a notification to the system administrator, but takes no further action since appropriate actions depend on the local network configuration. The default **network_up** event processing takes no actions since appropriate actions depend on the local network configuration.

Network Adapter Events

The Cluster Manager reacts to the failure, unavailability, or joining of network adapters by initiating one of the following events:

swap_adapter — This event occurs when the service adapter on a node fails. The **swap_adapter** event exchanges or swaps the IP addresses of the service and a standby adapter on the same HANFS for AIX network and then reconstructs the routing table.

swap_adapter_complete — This event occurs only after a **swap_adapter** event has successfully completed. The **swap_adapter_complete** event ensures that the local ARP cache is updated by deleting entries and pinging cluster IP addresses.

swap_address — This event occurs when a user requests to move a service/boot address to an available standby adapter on the same node and network.

swap_address_complete — This event occurs only after a swap_address event has successfully completed. The swap_address_complete event ensures that the local ARP cache is updated by deleting entries and pinging cluster IP addresses.

fail_standby	This event occurs if a standby adapter fails or becomes unavailable as the result of an IP address takeover. The **fail_standby** event displays a console message indicating that a standby adapter has failed or is no longer available.
join_standby	This event occurs if a standby adapter becomes available. The **join_standby** event displays a console message indicating a standby adapter has become available.

Failure of a Single Adapter Does Not Generate Events

Be aware that if you have only one adapter active on a network, the Cluster Manager does not generate a failure event for that adapter. “Single adapter” situations include:

- One-node clusters
- Multi-node clusters with only one node active
- Failure of all but one adapter on a network, one at a time.

For example, starting a cluster with all service or standby adapters disconnected produces results as follows:

1. *First node up*: No failure events are generated.
2. *Second node up*: One failure event is generated.
3. *Third node up*: One failure event is generated.
4. And so on.

Whole-Cluster Status Events

By default, the Cluster Manager recognizes a six-minute time limit for reconfiguring a cluster and processing topology changes. If the time limit is reached, the Cluster Manger initiates one of the following events:

config_too_long	This event occurs when a node has been in reconfiguration for more than six minutes. The event periodically displays a console message.
unstable_too_long	This event occurs when a node has been unstable (processing topology changes) for more than six minutes. The event periodically displays a console message.

Part 2 Planning HANFS for AIX

In this part, you learn about the preliminary decisions you need to make to set up an HANFS for AIX cluster successfully, including planning for your networks, shared disk devices, shared LVM components, and resource groups.

Chapter 2 Planning an HANFS for AIX Cluster

This chapter provides an overview of the recommended planning process.

Design Goal: Eliminating Single Points of Failure

HANFS for AIX software provides numerous facilities you can use to build a two-node cluster that supplies highly available NFS functionality. Adequate planning is the key to building an HANFS for AIX cluster that is easier to install, provides higher availability, performs better, and requires less maintenance than a poorly planned cluster.

Your major goal throughout the planning process is to eliminate single points of failure. Remember: A single point of failure exists when a critical cluster function is provided by a single component.

HANFS for AIX, Version 4.3.1, is designed to recover from a single hardware or software failure. It may not be able to handle multiple failures, depending on the sequence of failures.

The following table summarizes potential single points of failure within an HANFS for AIX cluster and describes how to eliminate them:

Cluster object	Eliminate as single point of failure by using...
Node	Multiple nodes
Power source	Multiple circuits or uninterruptible power supplies
Network adapter	Redundant network adapters
Network	Multiple networks to connect nodes
TCP/IP subsystem	Serial networks to connect adjoining nodes
Disk adapter	Redundant disk adapters
Controller	Redundant disk controllers
Disk	Redundant hardware and disk monitoring

Note: In an HANFS for AIX environment on an SP machine, the SP Switch adapters are potential network single points of failure and should be promoted to node failures. See Chapter 14, Supporting AIX Error Notification for information on handling this.

The Planning Process

This section describes the recommended steps for planning an HANFS for AIX cluster. As you go follow these steps, you can record your planning decisions on the worksheets provided in Appendix A, Planning Worksheets. You can then refer to the worksheets as you configure the cluster.

Step 1: Drawing the Cluster Diagram

In this step you plan the core of the cluster—the resource groups (including file systems), the number of nodes, and shared IP addresses. Your goal is to develop a high-level view of the system that serves as a starting point for the cluster design. This step is described in this chapter.

Step 2: Planning TCP/IP and Serial Networks

In this step you plan the TCP/IP and serial network support for the cluster. You first examine issues relating to TCP/IP and serial networks in an HANFS for AIX environment, and then complete the network worksheets. Chapter 3, Planning HANFS for AIX Networks, describes this process.

Step 3: Planning Shared Disk Devices

In this step you plan the shared disk devices for the cluster. You first examine issues relating to different types of disk arrays and subsystems in an HANFS for AIX environment, and then diagram the shared disk configuration. Chapter 4, Planning Shared Disk Devices, describes this process.

Step 4: Planning Shared LVM Components and NFS File Systems

In this step you plan the shared volume groups for the cluster. You first examine issues relating to LVM components in an HANFS for AIX environment, and then fill out the LVM worksheets. This planning includes NFS file system issues. Chapter 5, Planning Shared LVM Components, describes this process.

Step 5: Planning Resource Groups

In this step you plan the resource groups for your HANFS for AIX cluster and complete the resource worksheet. Chapter 6, Planning Resource Groups, describes this process.

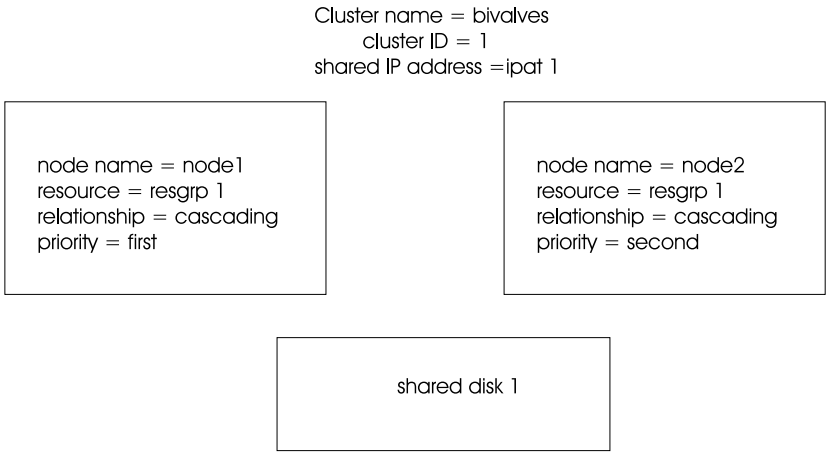
Drawing the Cluster Diagram

The purpose of the cluster diagram is to combine the information from each step in the planning process into one drawing that shows the cluster's function and structure. In this section, you make a first pass at drawing the cluster diagram. Remember, this pass is a starting point. You will refine your diagram throughout the planning process.

The initial pass of the cluster diagram identifies:

- The cluster name and ID
- The nodes within the cluster
- The resource groups (file systems) to be made highly available
- IP addresses shared by the nodes.

Here is an example of a first pass at a cluster diagram:



A First Pass of a Cluster Diagram

This diagram describes a two-node cluster. One node is named *node1*; the other *node2*. The cluster makes a resource group named *resgrp1*, which is a cascading resource group, highly available. The cluster has a single shared IP address that clients will use to access the resource group. A single external disk is shared between the two nodes.

You need to complete the following general steps to begin the cluster diagram (detailed instructions appear in the following sections):

Step	What you do...
1	Name the cluster and assign it an ID.
2	Name the nodes.

Step	What you do...
3	Identify each resource group (including file systems) and indicate whether it cascades or rotates.
4	For each resource group, determine the takeover priority for each node for that resource group.
5	Determine the shared IP addresses.

Naming the Cluster

Name the cluster and give it an ID. The cluster name is an arbitrary string of no more than 31 characters (alphanumeric and underscores only). The cluster ID can be any positive integer less than 99,999. Make sure that the cluster name and ID do not conflict with the names and IDs of other clusters at your site. As shown in the diagram above, the cluster is named *bivalves*; it has an ID of *1*.

Naming the Nodes

The HANFS for AIX software supports two-node clusters. Give each node a name. The name does not have to be the same as the hostname. The node name can be any alphanumeric string up to 31 characters long, and can include underscores. As shown in the preceding diagram, the cluster has two nodes, named *Node1* and *Node2*.

Identifying the Resource Groups

The purpose of the HANFS for AIX software is to ensure that critical file systems and data are available. This guide presumes you have already identified these file systems. For each file system, you need to create a resource group that contains the file system and its associated LVM components, and then assign a takeover strategy to the resource group.

The HANFS for AIX software supports two types of resource groups:

- *Cascading*, where a resource is owned by a specific node whenever that node is active in the cluster, and is taken over by the other node when the owner node fails.
- *Rotating*, where a resource is associated with both nodes and rotates between the nodes. When one node fails, the other node acquires the resource group. The takeover node may be currently active, or it may be in a standby state. When the detached node rejoins the cluster, it does not reacquire resources; instead, it rejoins as a standby.

Keep the following considerations in mind when deciding which type to assign to a resource group:

- If maximizing performance is essential, cascading resources may be the best resource group choice. Using cascading resources ensures that a file system is owned by a particular node whenever that node is active in the cluster.

  If the active node fails, its resources will be taken over by the other node. Note, however, that when the failed node reintegrates into the cluster, it temporarily disrupts availability as it takes back its resources from the takeover node.

- If minimizing downtime is essential, rotating resources may be the best choice. The availability of the resource group is not disrupted during node reintegration because the reintegrating node rejoins the cluster as a standby node only and does not attempt to reacquire its resources.

For each node, indicate the resource group and its corresponding type on the cluster diagram. Each file system can be assigned to only one resource group. A single node, however, can support different resource group types. For applications using cascading resources, also specify the takeover priority for each node.

For example, the cluster in the previous diagram makes the *resgrp1* resource group, which uses cascading resources, highly available. The node *node1* has the higher takeover priority; the node *node2* the lower takeover priority.

Planning Shared IP Addresses

The HANFS for AIX software requires that you define an IP address as part of each resource group. As a resource, an IP address can be acquired by the other node in the cluster should the node with this address fail.

In an HANFS for AIX environment, client/server applications can be "cluster aware" so that the client portion can recognize and submit requests to a shared IP address, or they can be "naive," in which case the client portion only submits requests to a specific network address. In the latter case, you need to designate an IP address as *shared* so that it can be acquired by the fallover node. If an IP address is not shared and a surviving node performed a disk takeover, clients would not be able to obtain network services using their original IP address.

Place a label beneath the cluster name and ID on the cluster diagram for each shared IP address you plan to use. This label is a placeholder until you assign the actual IP address during TCP/IP network planning. In the sample diagram, the cluster has a single shared IP address labeled *ipat1*.

Planning Shared Disk Access

In an HANFS for AIX environment, only one node has access to a shared external disk at a given time. If this node fails, its partner must take over the disk and restart applications to restore critical services to clients. Typically, takeover occurs within 30 to 300 seconds; but this range depends on the number and types of disks being used, the number of volume groups, and the number of file systems.

For the cluster diagram, draw a box representing each shared disk; then label each box with a shared disk name. For example, the cluster in the sample diagram has a single shared external disk named *shared_disk1*.

Where You Go From Here

At this point, you should have a diagram similar to the sample diagram shown earlier in this chapter. In subsequent planning steps, you will expand and refine the diagram by focusing on specific subsystems. In the next chapter, you will plan the cluster's network topology.

Chapter 3 Planning HANFS for AIX Networks

This chapter describes planning TCP/IP and serial network support for an HANFS for AIX cluster.

Planning TCP/IP Networks

In this step, you plan the TCP/IP networking support for the cluster, including:

- Deciding which types of networks and point-to-point connections to use in the cluster
- Designing the network topology
- Defining a network mask for your site
- Defining IP addresses for each node's service and standby adapters
- Defining a boot address for each service adapter that can be taken over
- Defining an alternate hardware address for each service adapter that can have its IP address taken over, if you are using hardware address swapping.

After addressing these issues, add the network topology to the cluster diagram and complete the network worksheets.

Selecting Public and Private Networks

In the HANFS for AIX environment, a *public network* connects the cluster nodes and allows clients to access these nodes. A *private network* is a point-to-point connection that directly links the two nodes. Asynchronous Transfer Mode (ATM) and the SP Switch used by the RS/6000 SP machine are special cases; they are private networks that can also connect clients.

As an independent, layered component of AIX, the HANFS for AIX software works with most TCP/IP-based networks. HANFS for AIX has been tested with standard Ethernet interfaces (en*) but not with IEEE 802.3 Ethernet interfaces (et*), where * reflects the interface number. HANFS for AIX also has been tested with Token-Ring and Fiber Distributed Data Interchange (FDDI) networks, with IBM Serial Optical Channel Converter (SOCC), Serial Line Internet Protocol (SLIP), and Asynchronous Transfer Mode point-to-point connections. HANFS for AIX has been tested with both the Classical IP and LAN Emulation ATM protocols.

See the documentation specific to your network type for detailed descriptions of its features.

Designing a Network Topology

In the HANFS for AIX environment, the *network topology* is the combination of networks and point-to-point connections that link cluster nodes and clients.

The HANFS for AIX software supports an unlimited number of TCP/IP network adapters on each node. Therefore, you have a great deal of flexibility in designing a network configuration. The design affects the degree of system availability such that the more communication paths that connect clustered nodes and clients, the greater the degree of network availability.

When designing your network topology, you must determine the number and types of:

- Networks and point-to-point connections that connect nodes
- Network adapters connected to each node.

An example of a dual-network topology for an HANFS for AIX cluster is shown in the following section.

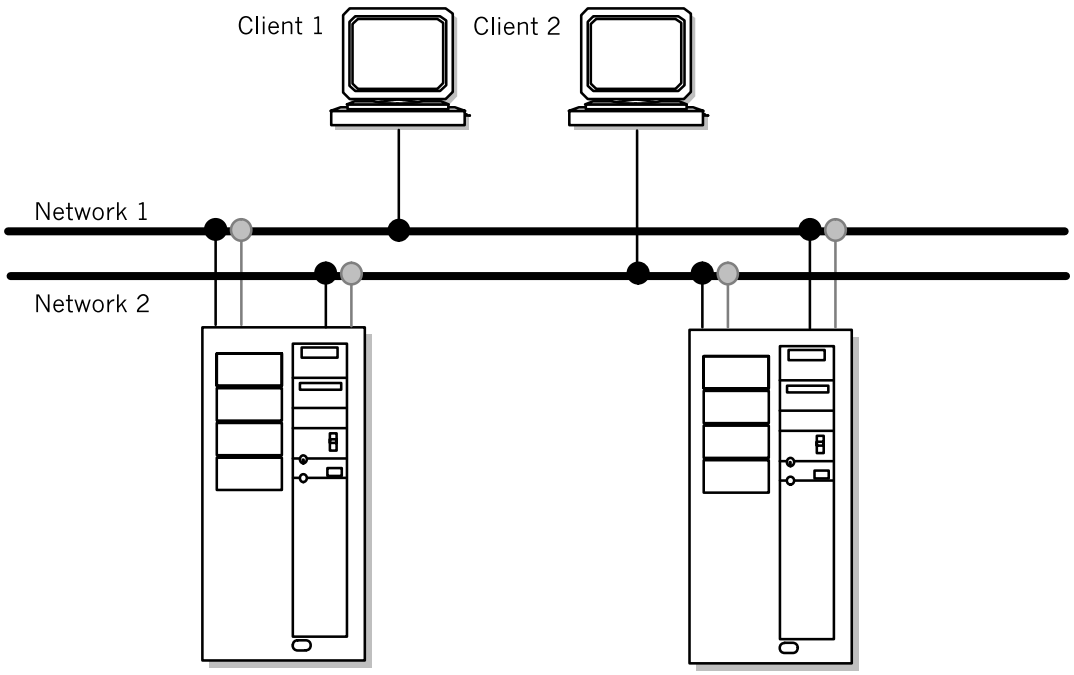
Dual-Network Topology

A dual-network setup is recommended. It has two separate networks for communication. Nodes are connected to two networks, and each node has two service adapters available to clients. If one network fails, the remaining network can still function, connecting nodes and providing access for clients.

In some recovery situations, a node connected to two networks may route network packets from one network to another. In normal cluster activity, however, each network is completely separate—both logically and physically.

Keep in mind that a client, unless it is connected to more than one network, is susceptible to network failure.

The following diagram shows a dual-network setup:



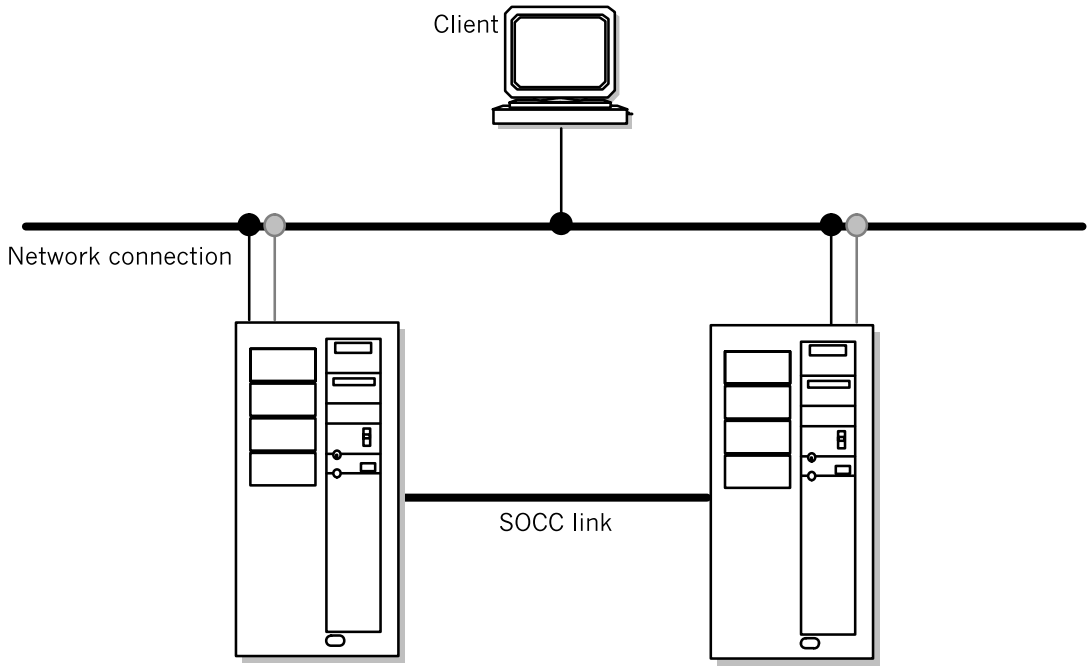
In the dual-network setup, each node is connected to two separate networks. Each node has one service adapter and can have one or more standby adapters per public network.

Dual-Network Setup

Point-to-Point Connection

A point-to-point connection directly links the two cluster nodes. SOCC and SLIP lines are point-to-point connections.

The following diagram shows a cluster consisting of two nodes and a client. A single public network connects the nodes and the client, and the nodes are linked point-to-point by a private high-speed SOCC connection that provides an alternate path for cluster traffic should the public network fail.



A Point-to-Point Connection

Nodes

A node can be either an RS/6000 uniprocessor, symmetric multi-processor (SMP), or an RS/6000 SP system. In an HANFS cluster, a node is identified by a unique node name that does not have to be associated with a hostname as in non-HANFS cluster environments.

In an HANFS for AIX cluster, a node typically can have more than one network interface, and this hostname-to-network interface association does not uniquely identify the node. Instead, an IP label associated with node adapters uniquely identifies the host. At any given time, the hostname corresponds to only one of the node's IP labels.

In a non-HANFS environment, a node typically is identified by a hostname that is associated with a network interface. If a node has only one network interface, the hostname usually also uniquely identifies the node.

Thus, in a HANFS cluster, a node's name should match the adapter label of the primary network's service adapter because other applications may depend on this hostname. Also, because the **rpc.statd** daemon depends on the hostname of the system, you must associate the hostname with an IP label that will not change while the system is running.

See the following section for the definition of an adapter label. Chapter 8, Checking Installed Hardware, provides instructions for setting the node name.

Network Adapters

A network adapter (interface) connects a node to a network. A node typically is configured with at least two network interfaces for each network to which it connects: a service interface that handles cluster traffic, and one or more standby interfaces. A service adapter must also have a boot address defined for it (since IP address takeover is required).

Adapters in an HANFS for AIX cluster have a label and a function (service, standby, or boot). The maximum number of network interfaces per node is 24.

Adapter Label

A network adapter is identified by an adapter label. For TCP/IP networks, the *adapter label* is the name in the **/etc/hosts** file associated with a specific IP address. Thus, a single node can have several adapter labels and IP addresses assigned to it. The adapter labels, however, should not be confused with the hostname.

For example, the following entries show that *nodea* has two Ethernet adapters, labeled *svc* and *stdby*. In this example, the adapter labels reflect separate network interface functions, each associated with a unique IP address.

```
100.100.50.1   nodea_svc
100.100.51.1   nodea_stdby
```

For boot adapters, you can simply add the suffix "boot" to the node name, as in *nodea_boot.* For more information on adapter functions, see the following section.

While an adapter label is typically eight characters long (or less), it can be any alphanumeric string up to 31 characters long, and can include underscores and hyphens.

When deciding on an adapter label, keep in mind that the adapter label also can reflect an adapter interface name. For example, one interface can be labeled *nodea_en0* and the other labeled *nodea_en1*, where *en0* and *en1* indicate the separate adapter names. The label for an RS232 line must end in the characters tty*n*, where *n* is the number of the tty device associated with the serial network (for example, *clam_tty1*). The label for a target mode SCSI-2 bus must end in the characters tmscsi*n*, where *n* is the SCSI device number (for example, *clam_tmscsi2*).

Whichever naming convention you use for adapter labels in an HANFS for AIX cluster, be sure to be consistent.

Note: It is suggested you use the **sl** number assigned to the serial device by AIX when you define it as an adapter to the HANFS for AIX environment. For example, if this device is assigned *sl1*, then the adapter name should end with the characters "sl1" (for example, *clam_sl1*).

Adapter Function

In the HANFS for AIX environment, each adapter has a specific function that indicates the role it performs in the cluster. An adapter's function can be service, standby, or boot.

Note to former HA-NFS Version 3 users: In HA-NFS Version 3, the service adapter was called the primary adapter, and the standby adapter was called the secondary adapter. Note also that in HANFS for AIX you may need a different IP address for the standby adapter than the one you used for the secondary adapter, since the standby adapter must be on a separate logical subnet from the service adapter.

Service Adapter

The *service adapter* is the primary connection between the node and the network. A node has one service adapter for each physical network to which it connects. The service adapter is used for general TCP/IP traffic and is the address the Cluster Information program (Clinfo) makes known to application programs that want to use cluster services.

Note: In configurations using rotating resources, the service adapter on the standby node remains on its boot address until it assumes the shared IP address. Consequently, Clinfo makes known the boot address for this adapter.

Note: In configurations using the Classical IP form of the ATM protocol (ie. *not* ATM LAN Emulation), a maximum of 7 service adapters per cluster is allowed if hardware address swapping is enabled.

Standby Adapter

A *standby adapter* backs up a service adapter. If a service adapter fails, the Cluster Manager swaps the standby adapter's address with the service adapter's address. Using a standby adapter eliminates a network adapter as a single point of failure. A node can have from one to seven standby adapters for each network to which it connects. Your software configuration and hardware constraints determine the actual number of standby adapters that a node can support.

Note: In an HANFS for AIX SP Switch network on the SP, integrated Ethernet adapters cannot be used, and no standby adapters are configured. If a takeover occurs, the service address is aliased onto another node's service address. See the *HACMP for AIX Installation Guide* for complete information on adapter functions in an SP Switch environment.

Boot Adapter (Address)

IP address takeover (IPAT) is an AIX facility that allows one node to acquire the network address of another node in the cluster. HANFS for AIX requires that IPAT be enabled. To enable IPAT, a boot adapter address (label) must be assigned to the service adapter on each cluster node. The boot adapter must be on the same logical network as the service adapter. Nodes use the boot address after a system reboot and before the HANFS for AIX software is started.

When the HANFS for AIX software is started on a node, the node's service adapter is reconfigured to use its service address instead of the boot address assigned to the service adapter. If the node should fail, the takeover node acquires the failed node's service address on its standby adapter, making the failure transparent to clients using that specific service address.

During the reintegration of the failed node, which comes up on its boot address, the takeover node will release the service address it acquired from the failed node. Afterwards, the reintegrating node will reconfigure its boot address to its reacquired service address.

Consider the following scenario: Suppose that Node A fails. Node B acquires Node A's service address and services client requests directed to that address. Later, when Node A is restarted, it comes up on its boot address and attempts to reintegrate into the cluster on its service address by requesting that Node B release Node A's service address. When Node B releases the requested address, Node A reclaims it and reintegrates into the cluster. Reintegration, however, would fail if Node A had not been configured to boot using its boot address. This operation occurs only when the service address is a resource in a cascading configuration.

It is important to realize that the boot address does not use a separate physical adapter, but instead is a second name and IP address associated with a service adapter. It must be on the same subnetwork as the service adapter. All cluster nodes must have this entry in the local **/etc/hosts** file and, if applicable, in the **nameserver** configuration.

Network Interfaces

The network interface is the network-specific software that controls the network adapter. The interface name is a three- or four-character string that uniquely identifies a network interface. The first two or three characters identify the network protocol. For example, *en* indicates a standard Ethernet network.

The network interface identifiers are listed below:

Interface	Identifier
Standard Ethernet	en
Token-Ring	tr
SLIP	sl
FDDI	fi
SOCC	so
SP Switch	css
ATM	at

The last character is the number assigned by AIX to the device. For example, *0*. An interface name of *en0* indicates that this is the first standard Ethernet interface on the system unit.

Networks

An HANFS for AIX cluster can be associated with up to 32 networks. Each network is identified by a name and an attribute.

Network Name

The *network name* is a symbolic value that identifies a network in an HANFS for AIX environment. Cluster processes use this information to determine which adapters are connected to the same physical network. The network name is arbitrary, but must be used consistently. If

several adapters share the same physical network, make sure that you use the same network name when defining these adapters. Network names can contain up to 31 alphanumeric characters, and can include underscores.

Network Attribute

A TCP/IP network's attribute is either public or private.

Public

A *public* network connects the two nodes and allows clients to access cluster nodes. Ethernet, Token-Ring, FDDI, and SLIP are considered public networks. Note that a SLIP line, however, does not provide any client access.

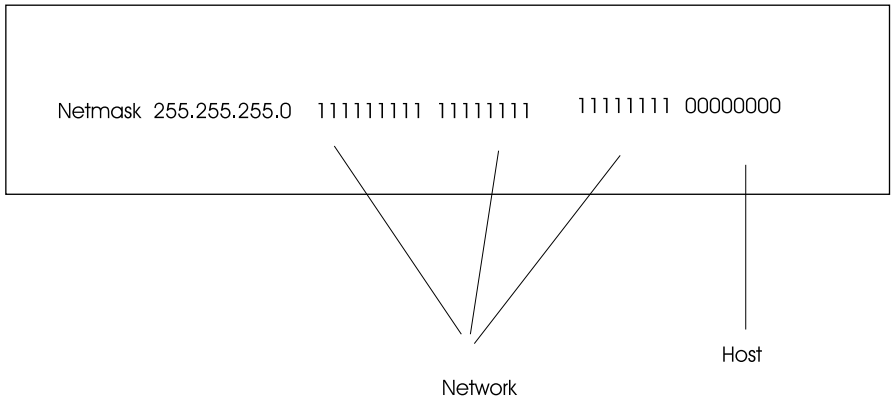
Private

A *private* network provides point-to-point communication between two nodes; it does not allow client access. SOCC lines and ATM networks are private networks. If an SP node is used as a client, the SP Switch network can allow client access.

Defining a Network Mask

The HANFS for AIX software uses the subnet feature of TCP/IP to divide a single physical network into separate logical subnets. In order to use subnets, you must define a network mask for your system.

An IP address consists of 32 bits. Some of the bits form the network address; the remainder form the host address. The *network mask* (or *netmask*) determines which bits in the IP address refer to the network and which bits refer to the host, as shown in the following example:



The preceding figure shows the netmask in both dotted decimal and binary format. A binary 1 indicates that a bit is part of the network address. A binary 0 indicates that a bit is part of the host address. In the example above, the network portion of the address occupies 24 bits; the host portion occupies 8 bits. It is convenient (but not necessary) to define a subnet on an octet boundary.

See the *IBM AIX Communication Concepts and Procedures* manual for more information about classes of addresses. Also, ask your network administrator about the class and subnets used at your site.

Note: In an HANFS for AIX HPS network on the SP, the netmask associated with the HPS "base address (css0)" will be used as the netmask for all HANFS HPS network adapters.

Defining IP Addresses for Standby Adapters

Note: In an HANFS for AIX SP Switch network on the SP, no standby adapters are configured. Service addresses are aliased onto another node's service address. See Chapter C, Installing and Configuring HANFS for AIX on RS/6000 SPs for more information on an SP environment.

The design of the HANFS for AIX software specifies that:

- All client traffic be carried over the service adapter.
- Standby adapters be hidden from client applications and carry only internal Cluster Manager traffic.

To comply with these rules, pay careful attention to the IP addresses you assign to standby adapters. Standby adapters *must* be on a separate subnet from the service adapters, even though they are on the same physical network. Placing standby adapters on a different subnet from the service adapter allows HANFS for AIX to determine which adapter TCP/IP will use to send a packet to a network.

If there is more than one adapter with the same network address, there is no way to guarantee which of these adapters will be chosen by the IP as the transmission route. All choices will be correct, since each choice will deliver the packet to the correct network. To guarantee that only the service adapter handles critical traffic, you must limit the IP's choice of a transmission route to one adapter. This keeps all traffic off the standby adapter so that it is available for adapter swapping and IP address takeover (IPAT). Limiting the IP's choice of a transmission route also facilitates identifying an adapter failure.

Configuring Multiple Adapters on the Same Subnet

When configuring a cluster node with multiple standby adapters, you must configure the standby adapters on the same subnet to avoid the following message that appears after configuring the first standby adapter through the SMIT chinet screen:

```
Method error (/usr/lib/methods/chgif):
       0514-066 Cause not known.
ifconfig: ioctl (SIOCAIFADDR): Do not specify an existing file.
0821-223 chgif: Cannot get records from CuAt.
0821-229 chgif: ifconfig command failed.
The status of "tr2" Interface in the current running system is
uncertain.
```

If this message appears, you can ignore it. To ensure that the standby adapter is configured, however, you should execute the **ping** command and specify the address of the standby adapter.

The following section describes how to place a standby adapter on a separate subnet from the service adapter.

Placing Standby Adapters on a Separate Subnet

Note to former HA-NFS Version 3 users: In HA-NFS Version 3, the primary and secondary adapters were typically on the same subnet. In HANFS for AIX, the service and standby adapters must be on *separate* subnets, with the boot adapter address on the same subnet as the service adapter. So if you had your primary and secondary adapters on the same subnet, you can use your old secondary address as your boot adapter address, and you will need a new address on a separate subnet for your standby adapter.

To place a standby adapter on a different subnet from a service adapter, give it an IP address that has a different network address portion.

Consider the following adapter IP address and network mask:

IP address:

```
100.100.50.121          01100100 01100100 00110010 01111001
```

Netmask:

```
255.255.255.0           11111111 11111111 11111111 00000000
```

In this example, the network address is 100.100.50. The host address is 121. An adapter configured in this way can transmit packets destined to a network whose first three octets are 100.100.50.

Now consider a node with the following two network adapters:

Adapter 1's IP address:

```
100.100.50.121          01100100 01100100 00110010 01111001
```

Adapter 2's IP address:

```
100.100.50.25           01100100 01100100 00110010 00011001
```

Netmask:

```
255.255.255.0           11111111 11111111 11111111 00000000
```

In this example, suppose that a client program on this node wanted to send data to another client at address 100.100.50.112. You cannot predict whether adapter 1 or adapter 2 will transmit the datagram, because the destination IP address (100.100.50.112) can be reached through either adapter.

Now consider a node with the following configuration:

Adapter 1's IP address:

```
100.100.50.121          01100100 01100100 00110010 01111001
```

Adapter 2's IP address:

```
100.100.51.25           01100100 01100100 00110011 00011001
```

Netmask:

```
255.255.255.0           11111111 11111111 11111111 00000000
```

In this case, you can determine that the data destined for the client at address 100.100.50.112 will be sent through adapter 1, since it is the only candidate that can send the packet to the network with address 100.100.50.

Be careful to pay close attention to the IP addresses you assign to standby adapters as you complete the networking worksheets referenced at the end of this chapter.

Note: Although standby adapters are on the same physical network (and must be on the same netmask) as service adapters, standby adapters must be on a different logical subnet from the service adapters. Different HANFS networks may use different netmasks, but the netmask must be the same for all adapters within a given HANFS network.

If you configure multiple standby adapters on cluster nodes, they all must be configured on the same subnet to handle cluster node failures. In addition, keep in mind that with multiple standby adapters configured, a **swap_adapter** event on the standby adapter routing heartbeats on a node may cause all standbys on the node to fail, since only one heartbeat route exists per node for the standbys on that node.

Defining Boot Addresses

You must define a boot address for each service adapter on which IP address takeover might occur. The boot address and the service address must be on the same subnet. That is, the two addresses must have the same value for the network portion of the address; the host portion must be different.

Use a standard formula for assigning boot addresses. For example, the boot address could be the host address plus 64. This formula yields the boot addresses shown below:

```
NODE A service address:       100.100.50.135
NODE A boot address:          100.100.50.199
Network mask:                 255.255.255.0
NODE B service address:       100.100.50.136
NODE B boot address:          100.100.50.200
Network mask:                 255.255.255.0
```

Defining Hardware Addresses

The hardware address swapping facility works in tandem with IP address takeover. Hardware address swapping maintains the binding between an IP address and a hardware address, which eliminates the need to flush the ARP cache of clients after an IP address takeover. This facility, however, is supported only for Ethernet, Token-Ring, FDDI and ATM adapters. It does not work with the SP Switch.

Note that hardware address swapping takes about 60 seconds on a Token-Ring network, and up to 120 seconds on a FDDI network. These periods are longer than the usual time it takes for the Cluster Manager to detect a failure and take action.

Selecting an Alternate Hardware Address

This section provides hardware addressing recommendations for Ethernet, Token Ring, FDDI, and ATM adapters. Note that any alternate hardware address you define for an adapter should be similar in form to the default hardware address the manufacturer assigned to the adapter.

To determine an adapter's default hardware address, use the **netstat -I** command (when the networks are active).

Selecting an Alternate Hardware Address

This section provides hardware addressing recommendations for Ethernet, Token Ring, FDDI, and ATM adapters. Note that any alternate hardware address you define for an adapter should be similar in form to the default hardware address the manufacturer assigned to the adapter.

To determine an adapter's default hardware address, use the **netstat -i** command (when the networks are active).

Using netstat

To retrieve hardware addresses using the **netstat -i** command, enter:

```
netstat –i | grep link
```

which returns output similar to the following:

```
lo0   16896 link#1                          186303     0   186309     0     0
en0   1500  link#2      2.60.8c.2f.bb.93      2925     0     1047     0     0
tr0   1492  link#3      10.0.5a.a8.b5.7b    104544     0    92158     0     0
tr1   1492  link#4      10.0.5a.a8.8d.79     79517     0    39130     0     0
fi0   4352  link#5      10.0.5a.b8.89.4f     40221     0        1     1     0
fi1   4352  link#6      10.0.5a.b8.8b.f4     40338     0        6     1     0
at0   9180  link#7       8.0.5a.99.83.57     54320     0        8     1     0
at2   9180  link#8       8.0.46.22.26.12     54320     0        8     1     0
```

Specifying an Alternate Ethernet Hardware Address

To specify an alternate hardware address for an Ethernet interface, begin by using the first five pairs of alphanumeric characters as they appear in the current hardware address. Then substitute a different value for the last pair of characters. Use characters that do not occur on any other adapter on the physical network.

For example, you could use 10 and 20 for Node A and Node B, respectively. If you have multiple adapters for hardware address swapping in each node, you can extend to 11 and 12 on Node A, and 21 and 22 on Node B.

Specifying an alternate hardware address for adapter interface *en0* in the preceding output thus yields the following value:

Original address `02608c2fbb93`

New address `02608c2fbb10`

Specifying an Alternate Token-Ring Hardware Address

To specify an alternate hardware address for a Token-Ring interface, set the first two digits to **42**, indicating that the address is set locally.

Specifying an alternate hardware address for adapter interface *tr0* in the preceding output thus yields the following value:

Original address `10005aa8b57b`

New address `42005aa8b57b`

Specifying an Alternate FDDI Hardware Address

To specify an alternate FDDI hardware address, enter the new address into the **Adapter Hardware Address** field as follows, *without any decimal separators*:

1. Use 4, 5, 6, or 7 as the first digit (the first nibble of the first byte) of the new address.
2. Use the last 6 octets (3 bytes) of the manufacturer's default address as the last 6 digits of the new address.

Here's a list of some sample valid addresses, shown with decimals for legibility:

```
40.00.00.b8.10.89
40.00.01.b8.10.89
50.00.00.b8.10.89
60.00.00.b8.10.89
7f.ff.ff.b8.10.89
```

Specifying an Alternate ATM Hardware Address

The following procedure applies to ATM Classic IP interface only. Hardware address swapping for ATM LAN Emulation adapters works just like hardware address swapping for the Ethernet and Token-Ring adapters that are being emulated.

Note: An ATM adapter has a hardware address which is 20 bytes in length. The first 13 bytes are assigned by the ATM switch, the next 6 bytes are burned into the ATM adapter, and the last byte represents the interface number (known as the *selector byte).* The above example only shows the burned in 6 bytes of the address. To select an alternate hardware address, you replace the 6 burned in bytes, and keep the last selector byte. The alternate ATM adapter hardware address is a total of 7 bytes.

To specify an alternate hardware address for an ATM Classic IP interface:

1. Use a value in the range of 40.00.00.00.00.00 to 7f.ff.ff.ff.ff.ff for the first 6 bytes.
2. Use the interface number as the last byte.

Here's a list of some sample alternate addresses for adapter interface at2 in the preceding output, shown with decimals for readability:

```
40.00.00.00.00.00.02
40.00.01.00.00.00.02
50.00.00.00.01.00.02
60.00.00.01.00.00.02
7f.ff.ff.ff.ff.ff.02
```

Since the interface number is hard-coded into the ATM hardware address, it must move from one ATM adapter to another during hardware address swapping. This imposes certain requirements and limitations on the HANFS configuration.

HANFS Configuration Requirements for ATM Hardware Address Swapping (Classical IP Only)

- If the hardware address moves to another adapter on the same machine (adapter swapping), the interface will have to be configured on that adapter as well. Likewise, when IP address takeover occurs, the interface associated with the adapter on the remote node will need to be configured on the takeover node.

- There can be *no more than 7 ATM service adapters per cluster* that support hardware address takeover.
- Each of these service interfaces *must have a unique ATM interface number.*
- On nodes that have one standby adapter per service adapter, the standby adapters on *all* cluster nodes will use the eighth possible ATM device *(at7),* so that there is no conflict with the service interface used by any of the nodes. This will guarantee that during IP address takeover with hardware address swapping the interface associated with the hardware address is not already in use on the takeover node.
- On nodes that have more than one standby adapter per service adapter, the total number of available service interfaces is reduced by that same number. For example, if two nodes have two standby adapters each, then the total number of service interfaces is reduced to 5.
- Any ATM adapters that are not being used by HANFS for AIX, but are still configured on any of the cluster nodes that are performing ATM hardware address swapping, will also reduce the number of available ATM interfaces on a one-for-one basis.

Network Configuration Requirements for ATM Hardware Address Swapping (Classical IP and LAN Emulation)

- Hardware address swapping for ATM requires that all adapters that can takeover for a given service address be attached to the same ATM switch.

Avoiding Network Conflicts

Each network adapter is assigned a unique hardware address when it is manufactured, which ensures that no two adapters have the same network address. When defining a hardware address for a network adapter, ensure that the defined address does not conflict with the hardware address of another network adapter in the network. Two network adapters with a common hardware address can cause unpredictable behavior within the network.

To reduce the chance that the chosen address is not a duplicate of another network adapter, consider using the following hardware addresses:

```
0x08007c5d76d1          0x08007c5d6efd
0x08007c5d3c43          0x08007c5d2ea3
0x08007c5d2e44          0x08007c5d26ae
0x08007c5d4d9a          0x08007c5d6484
0x08007c5d774b          0x08007c5d65df
0x08007c5d6455          0x08007c5d6398
```

Afterwards, to confirm that no duplicate addresses exist on your network, bring the cluster up on your new address and ping it from another machine. If you receive two packets for each ping (one with a trailing **DUP!**), you have probably selected an address already in use. Select another and try again. Cycle the Cluster Manager when performing these operations, because the alternate address is used only after the HANFS for AIX software is running and has reconfigured the adapter to use the service IP address the alternate hardware address associated with it.

Note: You cannot use the preceding addresses in a Token-Ring network or ATM Classic IP network. Token-Ring hardware addresses begin with the number 42. Alternate ATM Classic IP hardware addresses use 14 digits.

ATM LAN Emulation

ATM LAN emulation provides an emulation layer between protocols such as Token-Ring or Ethernet and ATM. It allows these protocol stacks to run over ATM as if it were a LAN. You can use ATM LAN emulation to bridge existing Ethernet or Token-Ring networks—particularly switched, high-speed Ethernet—across an ATM backbone network.

LAN emulation servers reside in the ATM switch. Configuring the switch varies with the hardware being used. Once you have configured your ATM switch and a working ATM network, you can configure adapters for ATM LAN emulation.

Note: You must load **bos.atm** from AIX on each machine if you have not already done so.

To configure ATM LAN emulation through SMIT:

1. Enter the SMIT fastpath **atmle_panel**.

   SMIT displays the ATM LAN Emulation menu.

2. Select **Add an ATM LE Client**.
3. Choose one of the adapter types (Ethernet or Token-Ring). A pop-up appears with the adapter selected (Ethernet in this example). Press Enter.
4. SMIT displays the **Add an Ethernet ATM LE Client** screen. Make entries as follows:

Local LE Client's LAN MAC Address (dotted hex)	Assign a hardware address like the burned-in address on actual network cards. The address must be unique on the network to which it is connected.
Automatic Configuration via LECS	**No** is the default. Toggle if you want **yes**.
If no, enter the LES ATM address (dotted hex)	Enter the 20-byte ATM address of the LAN Emulation server.
If yes, enter the LECS ATM address (dotted hex)	If the switch is configured for LAN Emulation Configuration Server, either on the well-known address or on the address configured on the switch, enter that address here.
Local ATM Device Name	Press F4 for a list of available adapters.
Emulated LAN Type	Ethernet/IEEE 802.3 (for this example)
Maximum Frame Size (bytes)	
Emulated LAN name	(optional) Enter a name for this virtual network.

5. Once you make these entries, press Enter. Repeat these steps for other ATM LE clients.

The ATM LE Clients should be visible to AIX as network cards when you execute the **lsdev -Cc adapter** command.

Each virtual adapter has a corresponding interface that must be configured, just like a real adapter of the same type, and it should behave as such.

Defining the ATM LAN Emulation Network to HANFS

After you have installed and tested an ATM LAN Emulation network, you must define it to the HANFS cluster topology as a public network. Chapter 12, Configuring an HANFS for AIX Cluster, describes how to define networks and adapters in an HACMP cluster.

You will define these virtual adapters to HANFS just as if they were real adapters. They have all the functions of Ethernet or Token-Ring adapters, such as hardware address swapping.

Using HANFS with NIS and DNS

HANFS facilitates communication between the nodes of a cluster so that each node can determine if the designated services and resources are available. Resources can include, but are not limited to, IP addresses and names and storage disks.

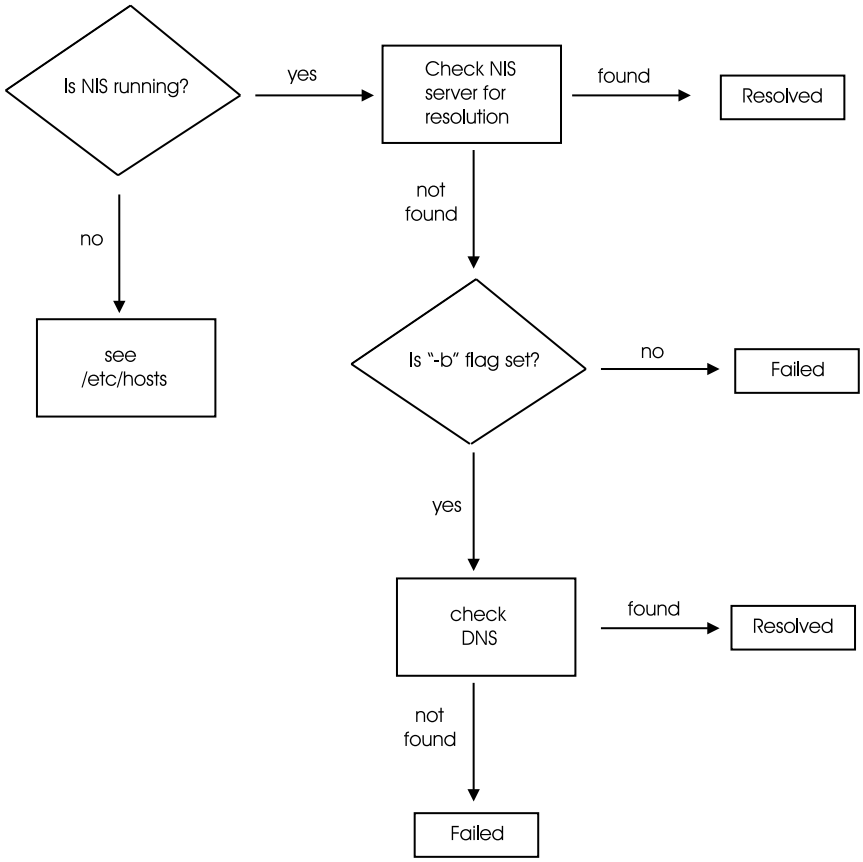
Subnetting the service and standby adapters (using TCP/IP) and having at least two separate networks (Ethernet and RS232) permits HANFS to determine whether a communication problem exists in the adapter, the network, or another common component such as TCP/IP software itself. The ability to subnet the adapters ensures that HANFS can direct the keepalive traffic from service to standby adapters, standby to service adapters, standby to standby adapters, and so on. For example, if both the standby and service adapters of Node A can receive and send to the standby adapter of Node B, but cannot communicate with the service adapter of Node A, it can be assumed that the service adapter is not working properly. HANFS recognizes this and performs the swap between the service and standby adapters.

Some of the commands used to perform the swap require IP lookup. This defaults to a nameserver for resolution if NIS or DNS is operational. If the nameserver was accessed via the adapter that is down, the request will timeout. To ensure that the cluster event (in this case an adapter swap) completes successfully and quickly, HANFS disables NIS or DNS hostname resolution. It is therefore required that the nodes participating in the cluster have entries in the **/etc/hosts** file.

How HANFS Enables and Disables Nameserving

This section provides some additional details on the logic a system uses to perform hostname resolution and how HANFS enables and disables DNS and NIS.

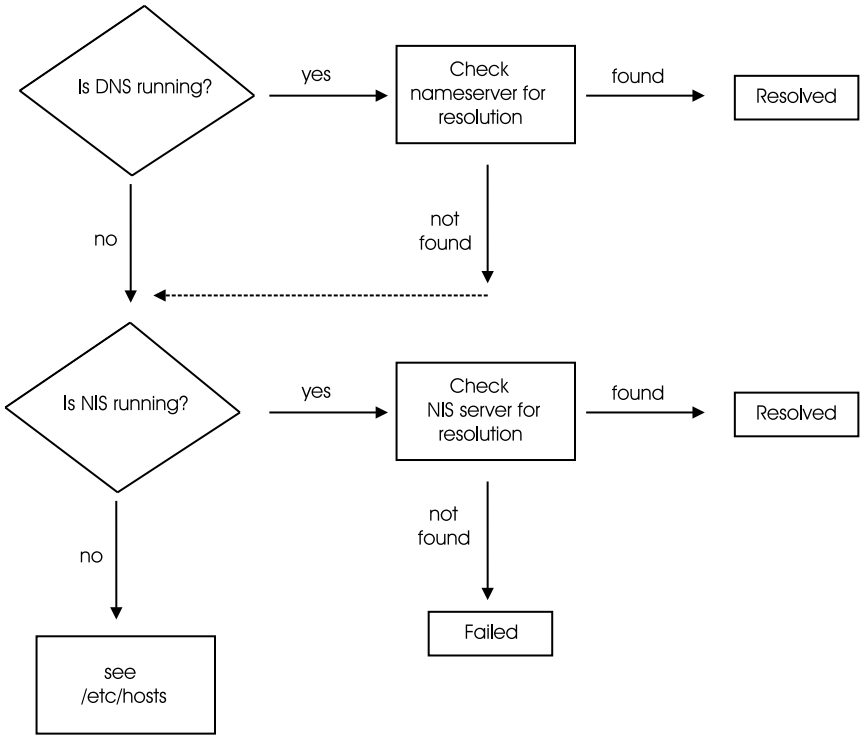
If a node is using either domain nameserving or NIS, the hostname will normally be resolved by contacting a suitable server. This will, at best, cause a time delay, and at worst, never return a response because communication to the server is lost. For example, if NIS alone is running, hostname resolution is performed via the following logic in AIX:



Note: The following applies if the NIS configuration maps include the host map.

As shown, if the NIS host tables were built with the **-b** flag, the NIS Server will continue its search via domain nameserving if that option is available. The key point, however, is that under certain circumstances (for example, an adapter being down, and the NIS server being unavailable), hostname resolution should be localized for quick response time. This infers that the local **/etc/hosts** file should be consulted, and the systems that may be involved in a cluster reconfiguration must be placed within the file in addition to loopback and the local hostname. It also means that the client portion of NIS that is running on that machine must be turned off. This is the case if the cluster node is an NIS client regardless of its status as a server (master or slave). Remember, even an NIS server uses the **ypbind** to contact the **ypserv** daemon running either on the master server or a slave server to perform the lookup within the current set of NIS maps.

Similarly, the logic followed by DNS (AIX) is:



In this situation, if both DNS and NIS were not disabled, a response to a hostname request might be as long as the time required for both NIS and DNS to timeout. This would hinder HANFS system reconfiguration, and increase the takeover time required to maintain the availability of designated resources (IP address). Disabling these services is the only reliable and immediate method of expediting the communication changes HANFS is attempting to accomplish.

The method HANFS uses to cleanly start and stop NIS and DNS is found within the scripts **cl_nm_nis_on** and **cl_nm_nis_off** within the **/usr/sbin/cluster/events/utils** directory.

For NIS, checking the process table for the **ypbind** daemon lets HANFS know that NIS is running, and should be stopped. A check for the existence of the file **/usr/sbin/cluster/hacmp_stopped_ypbind** lets HANFS know that NIS client services must be restarted following the appropriate cluster configuration events. As mentioned earlier, the commands **startsrc -s ypbind** and **stopsrc -s ypbind** are used to start and stop this node from using NIS name resolution. This is effective whether the node is a master or slave server (using NIS client services), or simply an NIS client.

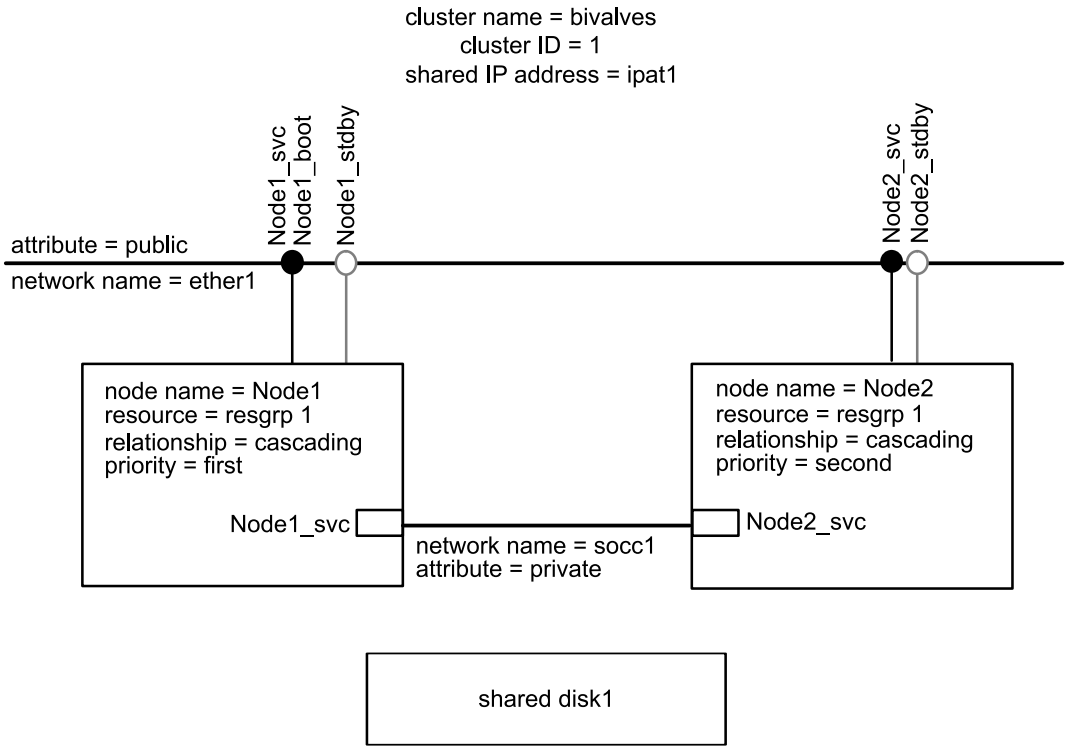
HANFS uses the AIX command **namerslv** to stop and start DNS on the cluster node. After checking for the existence of an **/etc/resolv.conf** file (the method for determining if domain name resolution is in effect), HANFS uses the **namerslv -E** command to stop nameserving by moving the **/etc/resolv.conf** file to a specified file. After reconfiguration, HANFS runs the **namerslv -B** command to restore the domain name configuration file from the specified file. Using this method, it is not necessary to stop and restart the **named** daemon.

Adding the TCP/IP Network Topology to the Cluster Diagram

You can now add the TCP/IP network topology to the cluster diagram:

1. Sketch in the networks, including any point-to-point connections between nodes.
2. Name each network and indicate its attribute. Remember, a network name is an arbitrary string. For example, in the following cluster diagram, the public Ethernet network is named *ether1* and the private SOCC point-to-point connection is named *socc1*.
3. Add the network adapters. Give each adapter a name and indicate its function. Remember to include a boot address for each service adapter that can be taken over. For example, in the following diagram, the service adapter for the Ethernet network on the node *Node1* is named *Node1_svc* and the corresponding boot adapter is named *Node1_boot*.

At this point you should have a diagram similar to the one shown in the following figure. Note that the names and IDs that you define (for example, *ether1*) apply *only* within the specific HANFS for AIX cluster and have no meaning outside of it.



A Cluster Diagram with TCP/IP Network Topology

Completing the TCP/IP Networks Worksheet

The TCP/IP Networks Worksheet helps you organize the networks for an HANFS for AIX cluster. To complete the worksheet:

1. Enter the cluster ID in the **Cluster ID** field.
2. Enter the cluster name in the **Cluster Name** field.

3. In the **Network Name** field, give each network a symbolic name. You use this value during the install process when you configure the cluster. Remember that each SLIP and SOCC line is a distinct network and must be listed separately.
4. Indicate the network's type in the **Network Type** field. For example, it may be an Ethernet, a Token-Ring, and so on.
5. Indicate the network's function in the **Network Attribute** field. That is, the network can be public or private.
6. In the **Netmask** field, provide the network mask of each network. The network mask is site dependent.

Completing TCP/IP Network Adapters Worksheet

The TCP/IP Network Adapters Worksheet helps you define the network adapters connected to each node in the cluster. Complete the following steps for each node on a separate worksheet:

1. Enter the node name in the **Node Name** field.
2. Leave the **Interface Name** field blank. You will enter values in this field after you check the adapter following the instructions in Chapter 3, Planning HANFS for AIX Networks.
3. Enter the symbolic name of the adapter in the **Adapter Identifier** field.
4. Identify the adapter's function as service, standby, or boot in the **Adapter Function** field.
5. Enter the IP address for each adapter in the **Adapter IP Address** field. Note that the SMIT Add an Adapter screen displays an **Adapter Identifier** field that correlates to this field on the worksheet.
6. Enter the name of the network to which this network adapter is connected in the **Network Name** field. Refer to the TCP/IP Networks Worksheet for this information.
7. Identify the network as public or private in the **Network Attribute** field.
8. Enter in the **Boot Address** field the boot address for each service address that can be taken over.
9. *This field is optional*. If you are using hardware address swapping, enter in the **Adapter HW Address** field the hardware address for each service address that has a boot address. The hardware address is a 12-digit hexadecimal value for Ethernet, Token-Ring and FDDI; it is a 14-digit hexadecimal value for ATM. Usually, hexadecimal numbers are prefaced with "0x" (zero x) for readability. *Do not use colons to separate the numbers in the adapter hardware address.*

   Note: Entries in the **Adapter HW Address** field should refer to the locally administered address (LAA), which applies only to the service adapter.

For each node in the cluster, repeat these steps on a separate TCP/IP Network Adapters Worksheet.

Planning Serial Networks

In this step you plan the serial networks for your cluster. This section discusses serial network topology and describes the supported serial network types. You then add the serial network topology to the cluster diagram and complete the Serial Network Worksheet in Appendix A, Planning Worksheets.

Note: In an HANFS for AIX SP Switch network on the SP, an integrated **tty** cannot be used for heartbeat traffic. See Chapter C, Installing and Configuring HANFS for AIX on RS/6000 SPs for more information on the SP Switch environment.

Serial Network Topology

A *serial network* allows the Cluster Managers on each node to communicate even after all TCP/IP-based networks have failed.

The HANFS for AIX software supports a raw RS232 serial line, a SCSI-2 Differential or Differential Fast/Wide bus using target mode SCSI, and target mode SSA on Multi-Initiator RAID adapters (FC 6215 and FC 6219).

RS232 Serial Line

If you are using shared disk devices other than SCSI-2 Differential or SCSI-2 Differential Fast/Wide devices, you must use a raw RS232 serial line as the serial network between the nodes. The RS232 serial line provides a point-to-point connection between nodes in an HANFS for AIX cluster and is classified as a tty device requiring a dedicated serial port at each end.

Note: The 7013-S70, 7015-S70, and 7017-S70 systems do not support the use of the native serial ports in an HACMP RS232 serial network. To configure an RS232 serial network in an S70 system, you must use a PCI multi-port ASync card.

You can label the tty device using any name or characters you choose; however, it is most commonly identified by a four-character string, where the first three characters are *tty* and the fourth is the AIX-assigned device number.

For example, if a node called *clam* is using a tty device as its serial line, you can label its adapter arbitrarily as *clam_serial*, or as *clam_tty1* to reflect the AIX-assigned device number.

Target Mode SCSI

You can configure a SCSI-2 bus as an HANFS for AIX serial network only if you are using SCSI-2 Differential devices that support target-mode SCSI. SCSI-1 Single-Ended and SCSI-2 Single-Ended devices do not support serial networks in an HANFS for AIX cluster. The advantage of using the SCSI-2 Differential bus is that it eliminates the need for a dedicated serial port at each end of the connection, and for associated RS232 cables.

The target mode SCSI device that connects nodes in an HANFS for AIX cluster is identified by a seven-character name string, where the characters *tmscsi* are the first six characters and the seventh character is the number AIX assigns to the device (for example, *tmscsi1*).

Target Mode SSA

You can configure a target-mode SSA connection between nodes sharing disks connected to SSA on Multi-Initiator RAID adapters (FC 6215 and FC 6219). The adapters must be at Microcode Level 1801 or later.

You can define a serial network to HANFS that connects all nodes on an SSA loop. By default, node numbers on all systems are zero. In order to configure the target mode devices, you must first assign a unique non-zero node number to all systems on the SSA loop.

The target mode SSA device that connects nodes in an HANFS for AIX cluster is identified by a six-character name string, where the characters *tmssa* are the first six characters and the seventh character is the unique node number you assign to the device (for example, *tmssa1*).

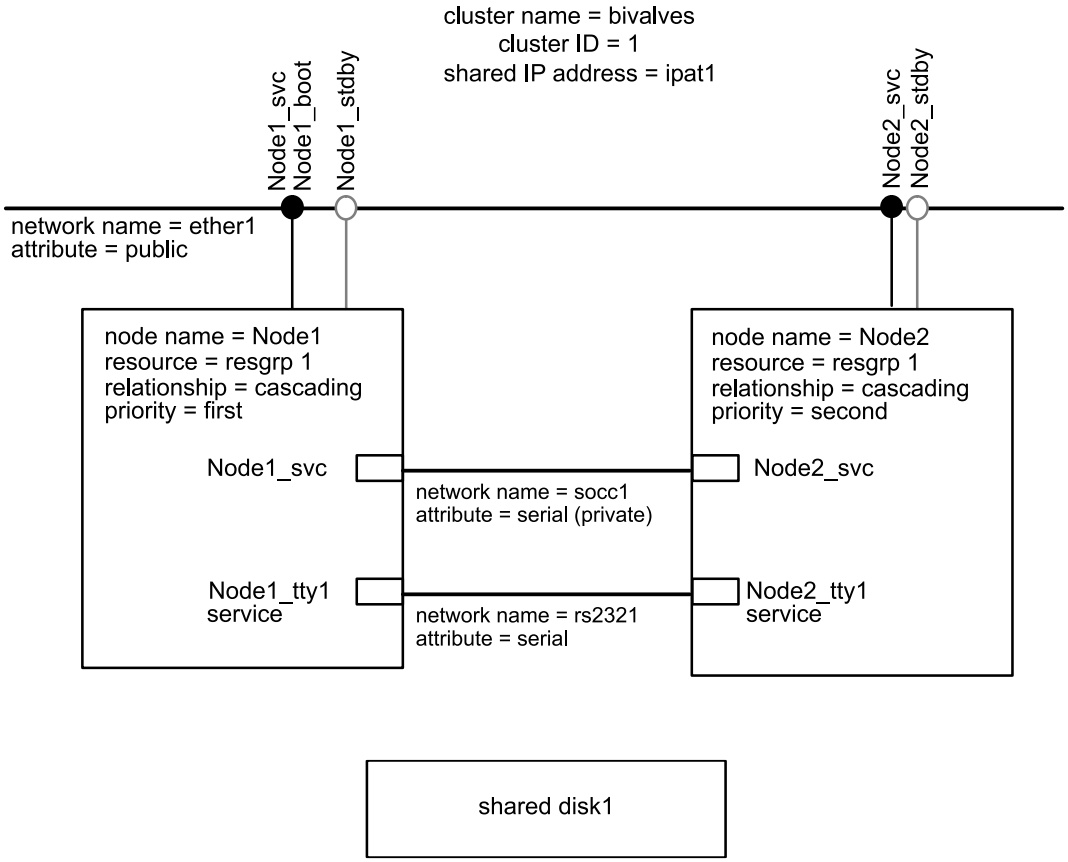
Configuring the target mode connection creates two special files in the **/dev** directory of each node, the **/dev/tmssa#.im** and **/dev/tmssa#.tm** files. The file with the **.im** extension is the initiator, which transmits data. The file with the **.tm** extension is the target, which receives data.

Adding the Serial Network Topology to the Cluster Diagram

You can now add the serial network topology to the cluster diagram:

1. Sketch in the serial networks. Remember, it is recommended that a serial network connect the nodes.
2. Name each network; then indicate that its attribute is serial. A network name is an arbitrary string. For example, you could name an RS232 line *serial1*. (The cluster in the following diagram has a single serial network named *rs2321*.)
3. Add the serial network adapters (if planning for a tmscsi or tmssa connection) or specify a tty port connection, then give each a name. Indicate its function as **service**. For example, the node *Node1* is using tty1 as its serial line connection; its adapter label is *Node1_tty1*. If you plan for a tmscsi connection, its adapter label might be *Node1_tmscsi2*.

At this point, you should have a cluster diagram similar to the following one:



A Cluster Diagram with Serial Network Topology

Completing the Serial Network Worksheet

Record the serial network information for your cluster:

1. Give each network a symbolic name in the **Serial Network Name** field. Remember, each RS232 serial line or target mode SCSI-2 Differential bus is a distinct network and must be listed separately.
2. Indicate the network's type as an RS232 serial line, a target mode SCSI-2 Differential bus, or a target mode SSA loop in the **Network Type** field.
3. Indicate the network's attribute as "serial" in the **Network Attribute** field.
4. List the names of the nodes connected to each serial network in the **Node Names** field. Refer to your cluster diagram.

Note: Because these serial networks do not use the TCP/IP protocol, none of them requires a netmask.

Completing the Serial Network Adapter Worksheet

This information helps you define the serial connections between cluster nodes. Complete the following steps for each node:

1. Enter the node name in the **Node Name** field.
2. Enter the slot number used to make the serial connection.
3. Leave the **Interface Name** field blank. You will enter the device name in this field after you configure the serial adapters during the install.
4. Enter the symbolic name of the adapter in the **Adapter Label** field.
5. Enter the name of the network to which this adapter is connected in the **Network Name** field.
6. Identify the network as serial in the **Network Attribute** field.
7. Identify the adapter's function as "service" in the **Adapter Function** field.

Repeat these steps on a new worksheet for the other node in the cluster.

Customizing Events in Response to Network Failure

By default, HANFS for AIX does not take any action upon network failure. You might want to consider customizing events to make network maintenance easier. Some actions you might consider include:

- Sending mail to the system administrator.

  The **/usr/sbin/cluster/events/network_down** script shows an example of how to do this upon an Ethernet failure. This type of action should be tailored to suit your specific environment.
- Promoting local network failure to node failure.
- Re-establishing the NFS connection on a backup network.
- Having clients remount NFS file systems on the backup network.

Where You Go From Here

You have now planned the TCP/IP and serial networks for the HANFS for AIX cluster. The next step in the planning process is to lay out the shared disk configuration for your cluster, as described in the following chapter.

Chapter 4 Planning Shared Disk Devices

This chapter discusses information you must consider before configuring shared external disks in an HANFS for AIX cluster. Also, you should refer to AIX documentation for the general hardware and software setup for your disks.

Overview

The HANFS for AIX software supports several shared disk configurations. Choosing the best setup for your needs requires careful thought before actually installing the disks. This chapter specifically includes information to help you:

- Choose a disk technology.
- Plan shared and non-shared storage.
- Plan to install any of the following SCSI disk configurations: SCSI-2 Differential, SCSI-2 Differential Fast/Wide, IBM 7137 Disk Array, IBM 7135-110 or 7135-210 RAIDiant Disk Array. Adapter requirements for each configuration are included in the planning considerations. The arrays can appear to the host as multiple SCSI devices. Use the general SCSI disk instructions in this chapter to plan for these arrays.
- Plan to install an IBM 9333 serial disk subsystem. Adapter requirements are included in the planning considerations.
- Plan to use an IBM 7133 Serial Storage Architecture (SSA) disk subsystem in your configuration. Adapter requirements are included in the planning considerations.

After reading this chapter, you will be able to complete the shared disk configuration portion of your cluster diagram. Completing this diagram will make installing disks and arrays easier.

Choosing a Shared Disk Technology

The HANFS for AIX software supports the following disk technologies as shared external disks in a highly available cluster:

- SCSI-2 SE, SCSI-2 Differential, and SCSI-2 Differential Fast/Wide adapters and drives
- IBM 9333 serial-link adapters and serial disk drive subsystems or enclosures
- IBM SSA adapters and SSA disk subsystems.

You can combine these technologies within a cluster. Before choosing a disk technology, however, review the considerations for configuring each technology as described in this section.

SCSI Disks

The HANFS for AIX software supports the following SCSI disk devices and arrays as shared external disk storage in cluster configurations:

- SCSI-2 SE, SCSI-2 Differential, and SCSI-2 Differential Fast/Wide disk devices

- The IBM 7135-110 and 7135-210 RAIDiant Disk Arrays
- The IBM 7137 Disk Array.

These devices and arrays are described in the following section.

SCSI-2 SE, SCSI-2 Differential, and SCSI-2 Differential Fast/Wide Disk Devices

The benefit of the SCSI implementation is its low cost. It provides a shared disk solution that works with most supported RISC System/6000 processor models (including the SP and the SMP) and requires minimal hardware overhead.

In an HANFS cluster, shared SCSI disks are connected to the same SCSI bus for the nodes that share the devices. In a non-concurrent access environment, the disks are "owned" by only one node at a time. If the owner node fails, the cluster node with the next highest priority in the resource chain acquires ownership of the shared disks as part of fallover processing. This ensures that the data stored on the disks remains accessible to client applications. The following restrictions, however, apply to using shared SCSI disks in a cluster configuration:

- SCSI-2 SE, SCSI-2 Differential, and SCSI-2 Differential Fast/Wide disks and disk arrays support non-concurrent shared disk access. The *only* SCSI-2 Differential disk devices that support concurrent shared disk access are the IBM 7135-110 and 7135-210 RAIDiant Arrays and the IBM 7137 Disk Array.
- Different types of SCSI buses can be configured in an HANFS cluster. Specifically, SCSI-2 Differential and SCSI-2 Differential Fast/Wide devices can be configured in clusters of up to four nodes, where all nodes are connected to the same SCSI bus attaching the separate device types. (You cannot mix SCSI-2 SE, SCSI-2 Differential, and SCSI-2 Differential Fast/Wide devices on the same bus.)
- You can connect the IBM 7135-210 RAIDiant Disk Array to *only* High Performance SCSI-2 Differential Fast/Wide adapters, while the 7135-110 RAIDiant Array *cannot* use those High Performance Fast/Wide adapters.
- You can connect up to sixteen devices to a SCSI-2 Differential Fast/Wide bus. Each SCSI adapter and disk is considered a separate device with its own SCSI ID. The SCSI-2 Differential Fast/Wide maximum bus length of 25 meters provides enough length for most cluster configurations to accommodate the full 16 device connections allowed by the SCSI standard.
- Do not connect other SCSI devices, such as CD-ROMs or tape drives, to a shared SCSI bus.
- The IBM High Performance SCSI-2 Differential Fast/Wide Adapter cannot be assigned SCSI IDs 0, 1, or 2; the adapter restricts the use of these IDs. The IBM SCSI-2 Differential Fast/Wide Adapter/A (FC 2416) cannot be assigned SCSI IDs 0 or 1.

IBM 7135 RAIDiant Disk Array Devices

You can use an IBM 7135-110 or 7135-210 RAIDiant Disk Array for shared external disk access in HANFS for AIX cluster configurations. The benefits of using an IBM 7135 RAIDiant Disk Array in an HANFS for AIX cluster are its storage capacity, speed, and reliability. The IBM 7135 RAIDiant Disk Array contains a group of disk drives that work together to provide enormous storage capacity (up to 135 GB of nonredundant storage) and higher I/O rates than single large drives.

Note: You can connect the IBM 7135-210 RAIDiant Disk Array to High Performance SCSI-2 Differential Fast/Wide adapters only, while the 7135-110 RAIDiant Array *cannot* use High Performance Fast/Wide adapters.

RAID Levels

The IBM 7135 RAIDiant Disk Array supports reliability features that provide data redundancy to prevent data loss if one of the disk drives in the array fails. As a RAID device, the array can provide data redundancy through RAID levels. The IBM 7135 RAIDiant Disk Array supports RAID levels 0, 1, and 5. RAID level 3 can be used only with a raw disk.

In RAID level 0, data is striped across a bank of disks in the array to improve throughput. Because RAID level 0 does not provide data redundancy, it is not recommended for use in HANFS for AIX clusters.

In RAID level 1, the IBM 7135 RAIDiant Disk Array provides data redundancy by maintaining multiple copies of the data on separate drives (mirroring).

In RAID level 5, the IBM 7135 RAIDiant Disk Array provides data redundancy by maintaining parity information that allows the data on a particular drive to be reconstructed if the drive fails.

All drives in the array are hot-pluggable. When you replace a failed drive, the IBM 7135 RAIDiant Disk Array reconstructs the data on the replacement drive automatically. Because of these reliability features, you should not define LVM mirrors in volume groups defined on an IBM 7135 RAIDiant Disk Array.

Dual Active Controllers

To eliminate adapters or array controllers as single points of failure in an HANFS for AIX cluster, the IBM 7135 RAIDiant Disk Array can be configured with a second array controller that acts as a backup controller in the event of a fallover. This configuration requires that you configure each cluster node with two adapters. You connect these adapters to the two array controllers using separate SCSI buses.

In this configuration, each adapter and array-controller combination defines a unique path from the node to the data on the disk array. The IBM 7135 RAIDiant Disk Array software manages data access through these paths. Both paths are active and can be used to access data on the disk array. If a component failure disables the current path, the disk array software automatically re-routes data transfers through the other path.

Note: This dual-active path-switching capability is independent of the capabilities of the HANFS for AIX software, which provides protection from a node failure. When you configure the IBM 7135 RAIDiant Disk Array with multiple controllers and configure the nodes with multiple adapters and SCSI busses, the disk array software prevents a single adapter or controller failure from causing disks to become unavailable.

The following restrictions apply to using shared IBM 7135 RAIDiant Disk Array in a cluster configuration:

- You can connect the IBM 7135-210 RAIDiant Disk Array only to High Performance SCSI-2 Differential Fast/Wide adapters, but the 7135-110 RAIDiant Disk Array cannot use those adapters.

- You can include up to two IBM 7135 RAIDiant Disk Arrays per bus.

  Note: Each array controller and adapter on the same SCSI bus requires a unique SCSI ID.

- You may need to configure the drives in the IBM 7135 RAIDiant Disk Array into logical units (LUNs) before setting up the cluster. A standard SCSI disk equates to a single LUN. AIX configures each LUN as a hard disk with a unique logical name of the form *hdiskn*, where *n* is an integer.

  An IBM 7135 RAIDiant Disk Array comes preconfigured with several LUNs defined. This configuration depends on the number of drives in the array and their sizes and is assigned a default RAID level of 5. You can, if desired, use the disk array manager utility to configure the individual drives in the array into numerous possible combinations of LUNs, spreading a single LUN across several individual physical drives.

  For more information about configuring LUNs on an IBM 7135 RAIDiant Disk Array, see the documentation you received with your disk array for the specific LUN composition of your unit.

  Once AIX configures the LUNs into hdisks, you can define the ownership of the disks as you would any other shared disk.

IBM 7137 Disk Array

The IBM 7137 disk array contains multiple SCSI-2 Differential disks. On the IBM 7137 array, these disks can be grouped together into multiple LUNs, with each LUN appearing to the host as a single SCSI device (hdisk).

IBM 9333 Serial Disk Subsystems

The HANFS for AIX software supports IBM 9333 serial disk drive subsystems as shared external disk storage devices. These drives are part of the IBM 9333 High-Performance Disk Drive Subsystem.

If you include the IBM 9333 serial disks in a volume group that uses LVM mirroring, you can replace a failed drive without powering off the entire subsystem.

In an HANFS for AIX cluster, you connect the shared IBM 9333 serial disks to the cluster nodes by cross-linking the IBM 9333 controllers.

A single serial adapter can support up to 16 disks. The number of serial adapters supported by the processor, therefore, determines the number of shared disks that can be connected to a node. The maximum number of shared disks that can be connected to a single node in an IBM 9333 serial configuration is 112, which can be divided among up to 7 adapter cards.

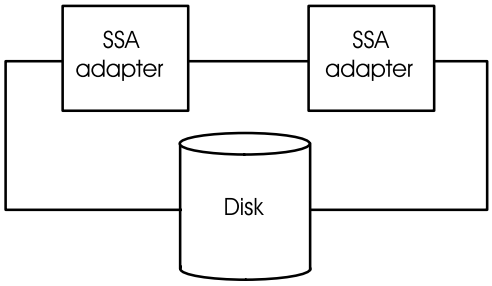
IBM Serial Storage Architecture Disk Subsystems

You can use IBM 7133 or 7131-405 SSA disk subsystems as shared external disk storage devices in an HANFS for AIX cluster. HANFS for AIX only supports SSA disk devices configured in a loop.

If you include SSA disks in a volume group that uses LVM mirroring, you can replace a failed drive without powering off the entire subsystem.

SSA Loop Configuration

An SSA loop is a cyclic network between adapters that access a disk subsystem. If one adapter fails, the other adapter can still access the disk. The following figure shows the basic SSA loop configuration.



Basic SSA Loop Configuration

The SSA adapter card has four ports which always operate as two dual ports. This allows one of the two dual ports to be connected externally to the system unit.

The following restrictions apply to using shared IBM SSA disk subsystems in an HANFS for AIX cluster configuration:

- SSA disks must be connected in loops (SSA string and switch configurations are not supported).
- Only one of the two pairs of connectors on an adapter card can be connected in a given SSA loop.
- No more than 48 disk drives may be connected in the same loop.
- The maximum number of shared disks that can be connected to a single cluster node in an IBM SSA configuration is 96 (two loops).

Power Supply Considerations

Reliable power sources are critical for a highly available cluster. Each node and mirrored disk chain in the cluster should have a separate power source. As you plan the cluster, make sure that the failure of any one power source (a blown fuse, for example) does not disable more than one node or mirrored chain. The following sections discuss specific power supply considerations for supported disk types.

SCSI Configurations

If the cluster site has a multiple-phase power supply, you must ensure that the cluster nodes are attached to the same power phase. Otherwise, the ground will move between the systems across the SCSI bus and cause write errors.

Uninterruptible power supply (UPS) devices are necessary for preventing data loss. The bus and devices shared between two nodes are subject to the same operational power surge restrictions as standard SCSI systems. When power is first applied to a SCSI device, the attached bus may

have active data corrupted. You can avoid such errors by briefly halting data transfer operations on the bus while a device (disk or adapter) is turned on. For example, if cluster nodes are installed on two different power grids and one node has a power surge that causes it to reboot, the surviving node may lose data if a data transfer is active.

IBM 7135 RAIDiant Disk Arrays are not prone to power supply problems because they come with redundant power supplies.

IBM 9333 Serial Disk Subsystem Configurations

Clusters with IBM 9333 serial disks are not prone to power supply problems. Nevertheless, UPS devices are still valuable for maintaining and ensuring that these configurations have no single point of failure.

When cross-linking several rack-mounted IBM 9333 serial disk subsystems, you should be sure to use the Dual Power Control to interconnect the IBM 9333 drawers in each rack to the power distribution unit in the other rack. This interconnection ensures that the IBM 9333 drawer does not power down when the node in the same rack goes down. If the drawer powered down, the fallover would fail when the cross-linked system, in a separate rack, attempted to take over the failed node's disks.

IBM SSA Disk Subsystem Configurations

Clusters with IBM SSA disk subsystems are not prone to power supply problems because they come with redundant power supplies.

Planning for Non-Shared Disk Storage

Keep the following considerations in mind regarding non-shared disk storage:

- The internal disks on each node in a cluster must provide sufficient space for:
  - AIX software (approximately 320 MB)
  - HANFS for AIX software (approximately 15 MB for a server node)
- The root volume group (**rootvg**) for each node must not reside on the shared SCSI bus.
- Use the AIX Error Notification Facility to monitor the **rootvg** on each node. Problems with the root volume group can be promoted to node failures. See Chapter 14, Supporting AIX Error Notification, for more information on using the Error Notification facility.
- Because shared disks require their own adapters, you cannot use the same adapter for both a shared and a non-shared disk. The internal disks on each node require one SCSI adapter apart from any other adapters within the cluster.
- Internal disks must be in a different volume group from external shared disks.
- The executable modules of the highly available applications should be on the internal disks and not on the shared external disks, for the following reasons:
  - **Licensing**—Some vendors require a unique license for each processor or multi-processor that runs an application, and thus license-protect the application by incorporating processor-specific information into the application when it is installed. As a result, it is possible that even though the HANFS for AIX software processes a node failure correctly, it is unable to restart the application on the fallover node because

of a restriction on the number of licenses available within the cluster for that application. To avoid this problem, make sure that you have a license for each processor in the cluster that may potentially run an application.

- **Starting Applications**—Some applications (such as databases) contain configuration files that you can tailor during installation and store with the binaries. These configuration files usually specify startup information, such as the databases to load and log files to open, after a fallover situation.

  If you plan to put these configuration files on a shared file system, they will require additional tailoring. You will need to determine logically which system (node) actually is to invoke the application in the event of a fallover. Making this determination becomes particularly important in fallover configurations where conflicts in the location and access of control files can occur.

  You can avoid much of the tailoring of configuration files by placing slightly different startup files for critical applications on local file systems on either node. This allows the initial application parameters to remain static; the application does not need to recalculate the parameters each time it is invoked.

Planning for Shared Disk Storage

When planning for shared storage, consider the following when calculating disk requirements:

- You need multiple physical disks on which to put the mirrored logical volumes. Putting copies of a mirrored logical volume on the same physical device defeats the purpose of making copies. See Chapter 5, Planning Shared LVM Components, for more information on creating mirrored logical volumes.

  When using an IBM 7135 RAIDiant Disk Array, do not create mirrored logical volumes. You must, however, account for the data redundancy maintained by the IBM 7135 RAIDiant Disk Array when calculating total storage capacity requirements. For example, in RAID level 1, because the IBM 7135 RAIDiant Disk Array maintains two copies of the data on separate drives, only half the total storage capacity is usable. Likewise, with RAID level 5, 20 to 30 percent of the total storage capacity is used to store and maintain parity information.
- Consider quorum issues when laying out a volume group. With quorum enabled, a two-disk volume group puts you at risk for losing quorum and data access. Either build three-disk volume groups or disable quorum.

  In an IBM 7135 RAIDiant Disk Array, where a single volume group can contain multiple LUNs (collections of physical disks) that appear to the host as a single device (hdisk), quorum is not an issue because of the large storage capacity the LUNs provide, and because of the data redundancy capabilities of the array.
- Physical disks containing logical volume copies should be connected to different power supplies; otherwise, loss of a single power supply can prevent access to all copies. In practice, this can mean placing copies in different IBM 9333 subsystem drawers or in different SCSI desk-side units.
- Physical disks containing logical volume copies should be on separate adapters. If all logical volume copies are connected to a single adapter, the adapter is potentially a single point of failure. If the single adapter fails, you must move the volume group to an alternate node. Separate adapters prevent any need for this move.

Planning a Shared SCSI-2 Disk Installation

The following list summarizes the basic hardware components required to set up an HANFS cluster that includes SCSI-2 SE, SCSI-2 Differential, or SCSI-2 Differential Fast/Wide devices as shared storage. Your exact cluster requirements will depend on the configuration you specify. To ensure that you account for all required components, complete a diagram for your system.

Disk Adapters

The HANFS for AIX software supports both the IBM SCSI-2 Differential High Performance External I/O Controller and the IBM High Performance SCSI-2 Differential Fast/Wide Adapter/A. For each IBM 7135 RAIDiant Disk Array, an HANFS for AIX configuration requires that you configure each cluster node with two host adapters.

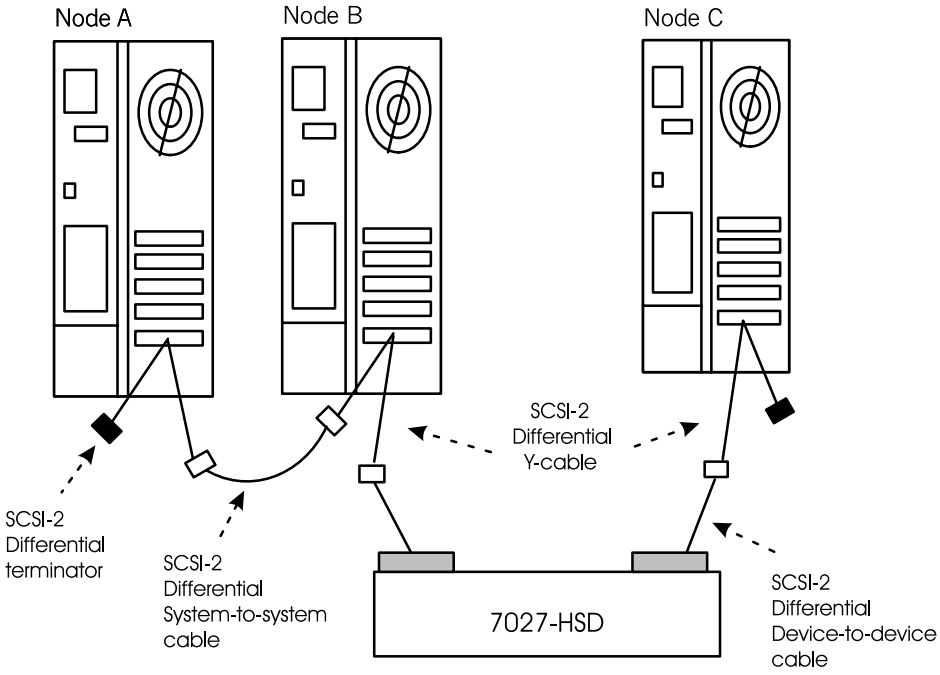
Note: Remove any SCSI terminators on the adapter card. You must use external terminators in an HANFS cluster. If you terminate the shared SCSI bus on the adapter, you lose termination when the cluster node that contains the adapter fails.

Cables

The cables required to connect nodes in your cluster depend on the type of SCSI bus you are configuring. Be sure to choose cables that are compatible with your disk adapters and controllers. For information on the specific cable type and length requirements for SCSI-2 Differential or SCSI-2 Differential Fast/Wide devices, see the hardware documentation that accompanies each device you want to include on the SCSI bus. Examples of SCSI bus configurations using IBM 7027-HSD and IBM 7204-315 enclosures, IBM 7137 Disk Arrays, and IBM 7135-210 RAIDiant Disk Arrays are shown throughout the following pages.

Sample SCSI-2 Differential Configuration

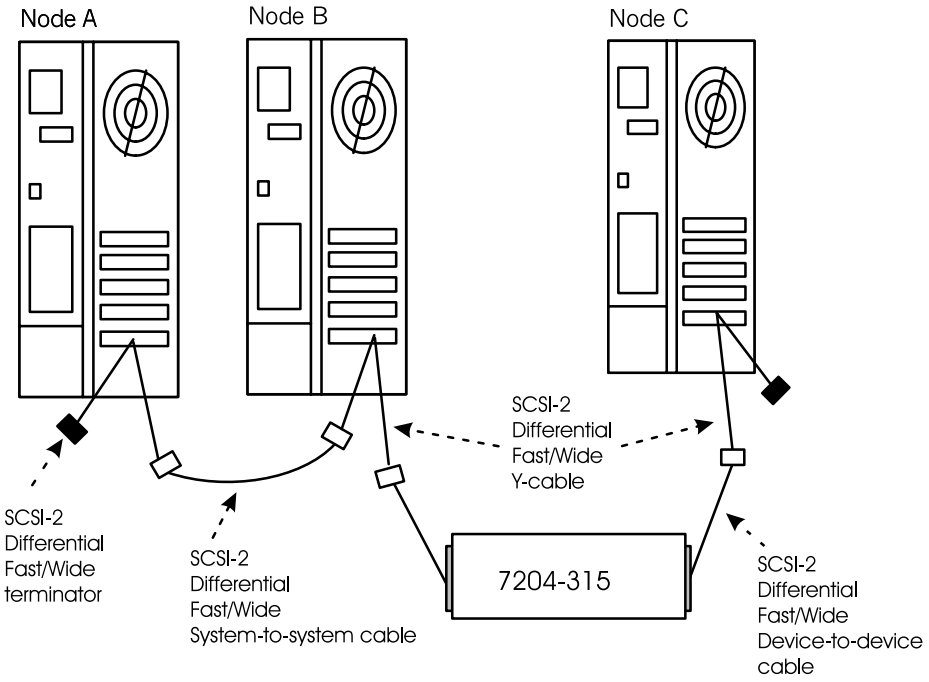
The following figure illustrates a SCSI-2 Differential configuration in which three nodes are attached to four SCSI-2 Differential disks in a IBM 7027-HSD enclosure. Each adapter is connected to a SCSI-2 Differential Y-cable. Other required cables are labeled in the figure.



Shared SCSI-2 Differential Disk Configuration

Sample SCSI-2 Differential Fast/Wide Configuration

The following figure illustrates a SCSI-2 Differential Fast/Wide configuration in which three nodes are attached to four SCSI-2 Differential disks in an IBM 7204-315 enclosure. Each adapter is connected to a SCSI-2 Differential Fast/Wide Y-cable. Other required cables are labeled in the figure.



Shared SCSI-2 Differential Fast/Wide Disk Configuration

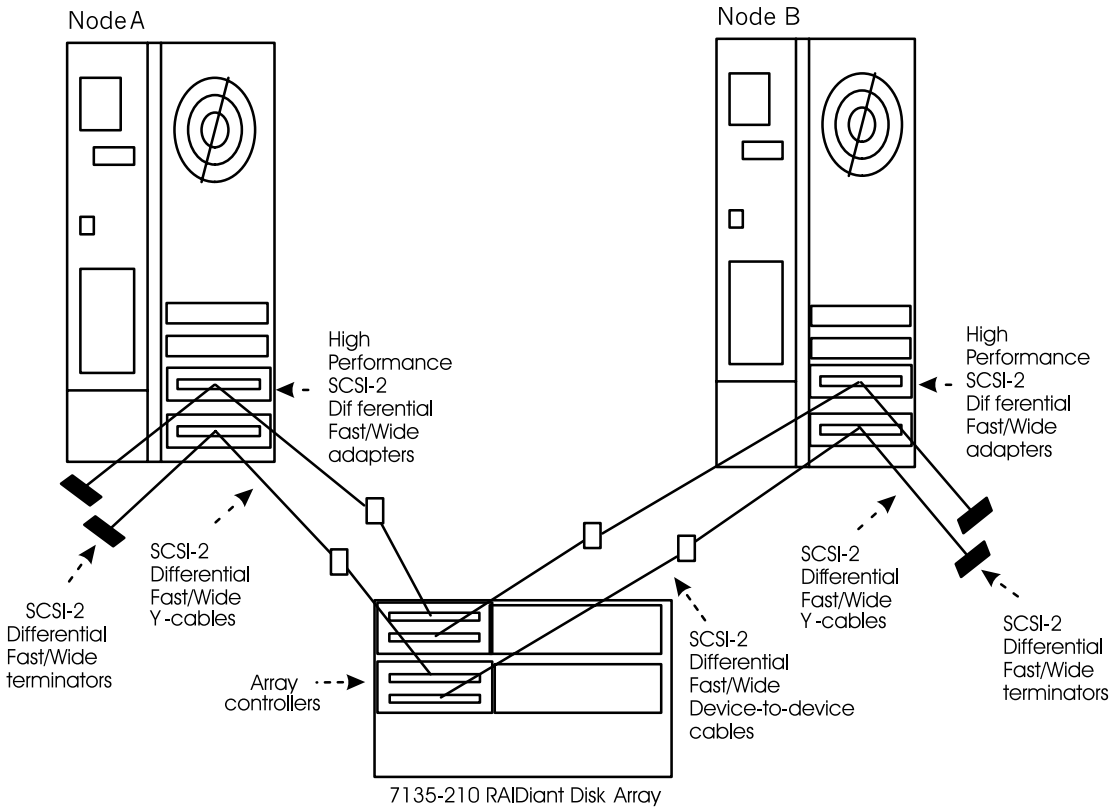
Sample IBM 7135-210 RAIDiant Disk Array Configuration

To take advantage of the path-switching capability of the IBM 7135-210 RAIDiant Disk Array software, you must configure each node with two adapters. In this way, the device driver can define multiple paths between the host and the disk array, eliminating the adapters, the controllers, and the SCSI bus as single points of failure. If a component failure disables one path, the IBM 7135-210 RAIDiant Disk Array device driver can switch to the other path automatically. This switching capability is the dual-active feature of the array.

Note: Although each controller on the IBM 7135-210 RAIDiant Disk Array contains two connectors, each controller requires only one SCSI ID.

SCSI-2 Differential Fast/Wide IBM 7135-210 RAIDiant Disk Array Configuration

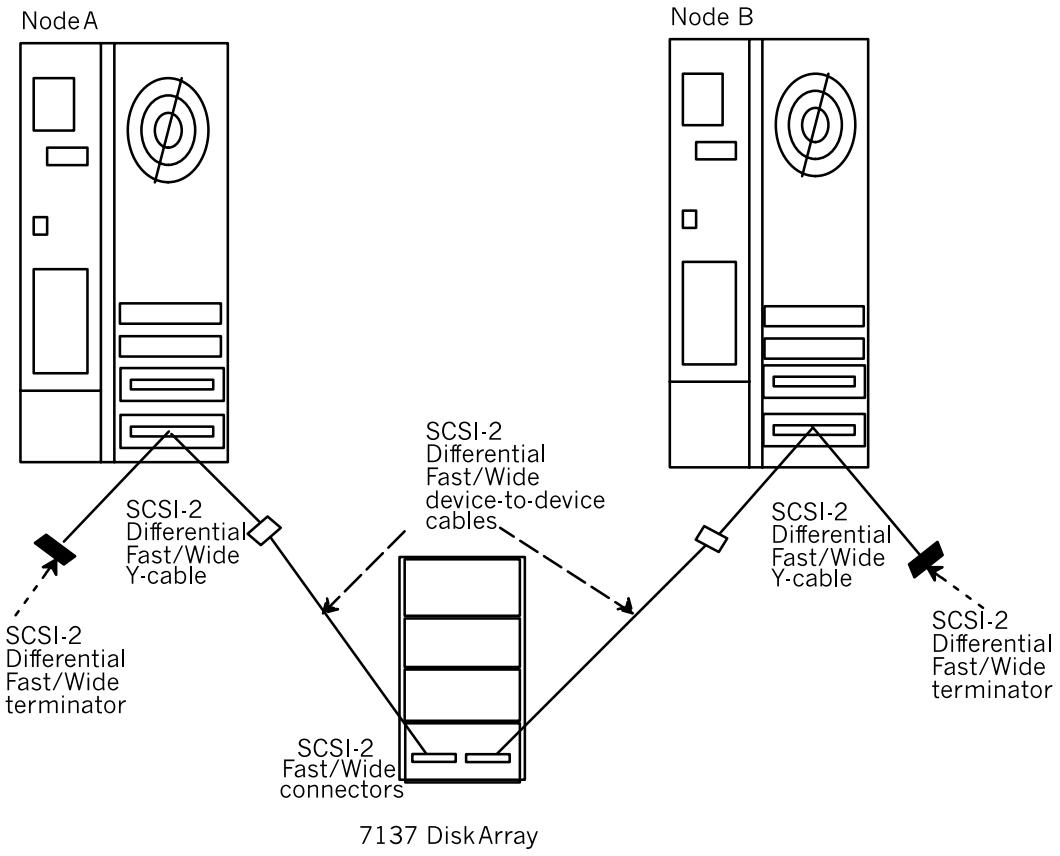
The following figure illustrates an HANFS cluster with an IBM 7135-210 RAIDiant Disk Array connected to two SCSI-2 Differential Fast/Wide buses.



IBM 7135-210 RAIDiant Disk Array Configuration with SCSI-2 Differential Fast/Wide Buses

SCSI-2 Differential Fast/Wide IBM 7137 Disk Array Configuration

The following figure illustrates an HANFS cluster with an IBM 7137 Disk Array connected to a SCSI-2 Differential Fast/Wide bus.



IBM 7137 Disk Array with a SCSI-2 Differential Fast/Wide Bus

Note: SCSI-2 Fast/Wide connectors on an IBM 7137 Disk Array are positioned vertically.

Planning a Shared IBM 9333 Serial Disk Installation

The following list summarizes the basic hardware components required to set up an HANFS for AIX cluster that includes IBM 9333 serial disk subsystems as shared storage. Exact requirements will depend on the configuration you specify. To ensure that you account for all required components, complete a system diagram. You should consider using the following hardware:

Adapters — Used to connect a host to up to four disk drawers or desk-side units. Each disk subsystem can contain up to four disks per drawer or desk-side unit. Apart from cable-length restrictions, this makes it possible to attach an adapter to a disk associated with another node.

Note: You must use the High-Performance Subsystem adapter with the IBM 9333 Model 011 or Model 501.

External serial-link cables — Used to connect the host adapters on each cluster node to the IBM 9333 drawer or desk-side unit.

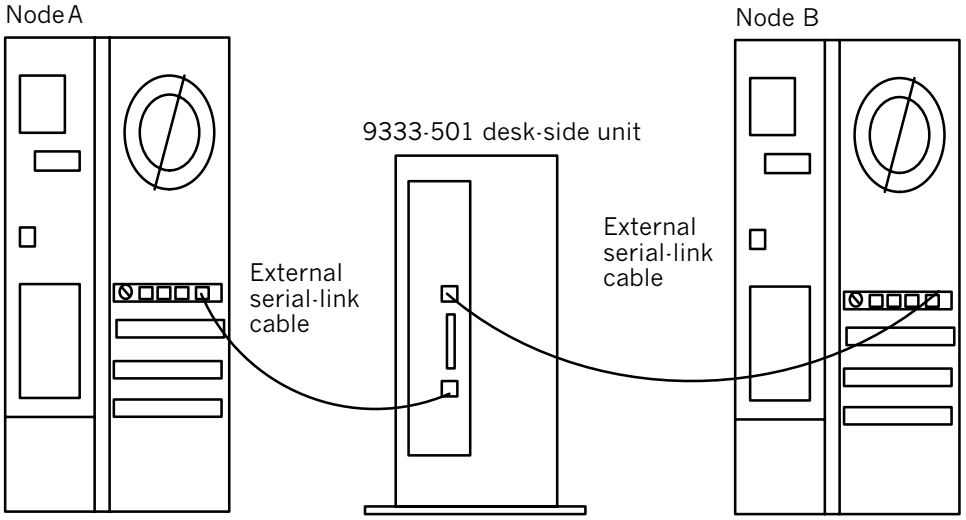
Local serial-link cables — Used as jumpers on IBM 9333 Model 011 or Model 501 disk subsystems to connect the Multiple Systems Attachment cards to the IBM 9333 serial-link controller.

Refer to AIX documentation for the initial hardware and software setup of the disk subsystem. Keep in mind that to make logical volume mirroring effective, the mirrors should be placed on separate disk subsystems. When planning a shared IBM 9333 serial-link disk installation, consult the restrictions described in IBM 9333 Serial Disk Subsystems on page 4-4.

Sample Two-Node Configuration

To support HANFS for AIX cluster configurations, each node requires one IBM 9333 adapter specifically dedicated to the shared IBM 9333 serial disk subsystem.

Take, for example, a two-node cluster consisting of Node A and Node B. Use one external serial-link cable to connect the adapter on Node A to one of connectors on the IBM 9333 serial-link subsystem controller. Use another external serial-link cable to connect the adapter on Node B to the other connector on the same disk subsystem as shown in the following figure:



Two-Node Cluster with Shared IBM 9333 Serial Disk Subsystem

Planning a Shared IBM SSA Disk Subsystem Installation

The following list summarizes the basic hardware components required to set up an HANFS for AIX cluster that includes IBM 7133 or 7131-405 SSA disk subsystems as shared storage devices. Exact requirements will depend on the configuration you specify. To ensure that you account for all required components, complete a system diagram. Consider using the following hardware:

Adapters Used to connect a host to up to 48 disk drives. Each disk subsystem can contain up to 6 disks per drawer.

The 7133 SSA adapter card has four external connectors. This allows one of the two dual ports to be connected externally to the system unit. The ports are numbered 1A, 1B, 2A, and 2B at the connectors, indicating that ports 1A and 2A are paired, and 1B and 2B are paired.

External SSA cables Used to connect the host adapters on each cluster node to the disk subsystem.

Refer to the system's documentation for the initial hardware and software setup of the disk subsystem. Keep in mind that to make logical volume mirroring effective, the mirrors should be placed on separate SSA loops. When planning a shared IBM SSA disk installation, consult the restrictions described in IBM SSA Disk Subsystem Configurations on page 4-6.

SSA Naming Conventions

When setting up an SSA subsystem for shared storage, use the following SSA loop naming scheme for the HANFS for AIX worksheets.

```
Loop #: nodeA – nodeB / nodeB – nodeA
```

Notation for the nodes is **n.a.p**, where

- **n** = the name of the node in the configuration
- **a** = the SSA adapter number on the node
- **p** = the port number on the SSA adapter.

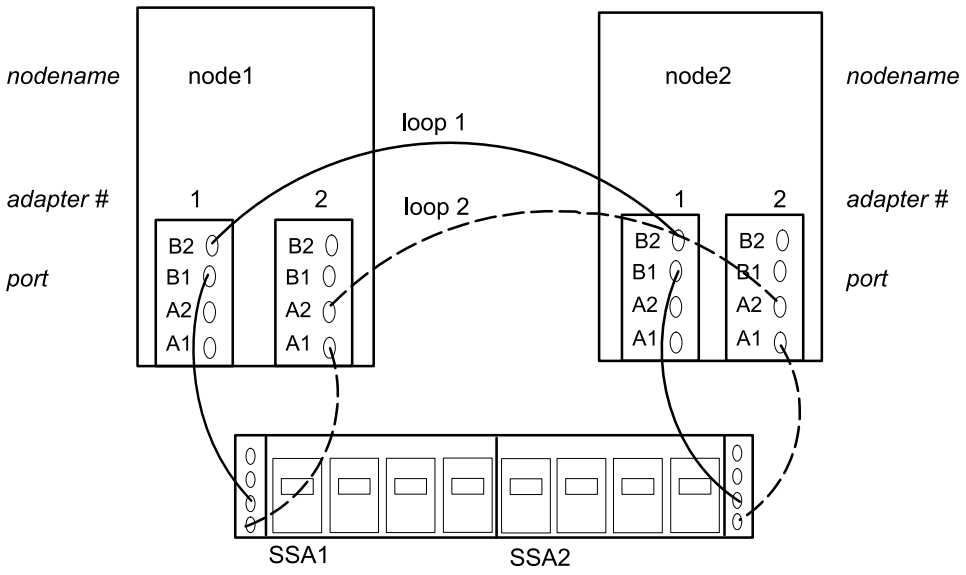
For an example of how to use this naming convention and how to configure loops, see the following section.

Sample SSA Configuration

To support HANFS for AIX cluster configurations, each node requires two IBM SSA adapters specifically dedicated to the shared IBM SSA disk subsystem. The configuration, shown in the following figure, incorporates two SSA loops using the subdivision of the 7133 drawer into SSA1 (four disks) and SSA2 (four disks). The two loops are described using the naming convention as follows:

```
Loop 1: node1.1.B1 – SSA 1 – node2.1.B1 / node2.1.B2 – node1.1.B2
Loop 2: node1.2.A1 – SSA 2 – node2.2.A1 / node2.2.A2 – node1.1.A2
```

In the figure, Loop 1 is illustrated using solid lines and Loop 2 is illustrated using dashed lines.



Two-Loop SSA Configuration in an HANFS for AIX Cluster with IBM 7133 SSA Disk Subsystem

Adding the Disk Configuration to the Cluster Diagram

Once you have chosen a disk technology, draw a diagram that shows the shared disk configuration. Be sure to include adapters and cables in the diagram. You may be able to expand on the cluster diagram you began in Chapter 2, Planning an HANFS for AIX Cluster, or you may need to have a separate diagram (for the sake of clarity) that shows the shared disk configuration for the cluster.

Where You Go From Here

You have now planned your shared disk configuration. The next step is to plan the shared volume groups for your cluster. This step is described in the following chapter.

Chapter 5 Planning Shared LVM Components

This chapter describes planning shared volume groups and file systems for an HANFS for AIX cluster, and discusses LVM issues as they relate to the HANFS for AIX environment. It does not provide an exhaustive discussion of LVM concepts and facilities in general. Refer to the *AIX System Management Guide* for more information on the AIX LVM.

LVM Components in the HANFS for AIX Environment

The LVM controls disk resources by mapping data between physical and logical storage. *Physical storage* refers to the actual location of data on a disk. *Logical storage* controls how data is made available to the user. Logical storage can be discontiguous, expanded, and replicated, and can span multiple physical disks. These features provide improved availability of data.

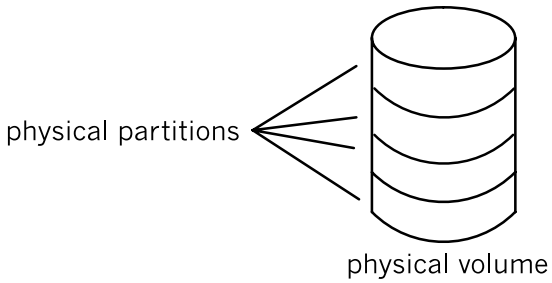
The LVM organizes data into the following components:

- Physical volumes
- Volume groups
- Logical volumes
- File Systems.

Considerations for each component as it relates to planning an HANFS for AIX cluster are discussed in the following sections.

Physical Volumes

A *physical volume* is a single physical disk. The physical volume is partitioned to provide AIX with a way of managing how data is mapped to the volume. The following figure shows how the physical partitions within a physical volume are conventionally diagrammed:



Physical Partitions on a Physical Volume

hdisk Numbers

Physical volumes are known in the AIX operating system by sequential *hdisk* numbers assigned when the system boots. For example, */dev/hdisk0* identifies the first physical volume in the system, */dev/hdisk1* identifies the second physical volume in the system, and so on.

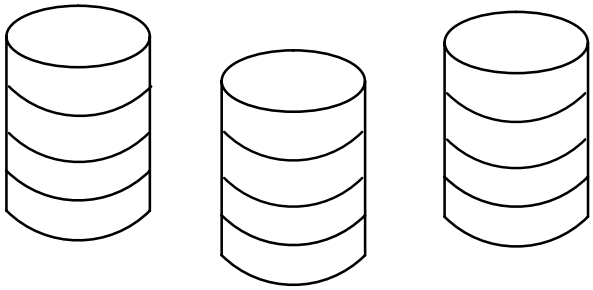
When sharing a disk in an HANFS for AIX cluster, the nodes sharing the disk each assign an hdisk number to that disk. These hdisk numbers may not match, but refer to the same physical volume. For example, each node may have a different number of internal disks, or the disks may have changed since AIX was installed.

The HANFS for AIX software does not require that the hdisk numbers match across nodes (though your system is easier to manage if they do). In situations where the hdisk numbers must differ, be sure that you understand each node's view of the shared disks. Draw a diagram that indicates the hdisk numbers that each node assigns to the shared disks and record these numbers on the appropriate volume group worksheets in Appendix A, Planning Worksheets. When in doubt, use the hdisk's PVID to verify its identity on a shared bus.

Volume Groups

A *volume group* is a set of physical volumes that AIX treats as a contiguous, addressable disk region. You can place from 1 to 32 physical volumes in the same volume group.

The following figure shows a volume group of three physical volumes:



A Volume Group of Three Physical Volumes

Shared Volume Groups

In the HANFS for AIX environment, a *shared volume group* is a volume group that resides entirely on the external disks shared by the cluster nodes. Do not include an internal disk in a shared volume group, since it cannot be accessed by other nodes. If you include an internal disk in a shared volume group, the **varyonvg** command fails.

The name of a shared volume group must be unique. It cannot conflict with the name of an existing volume group on any node in the cluster. Each volume group that has file systems residing on it has a log logical volume (**jfslog**) that must also have a unique name.

In an HANFS for AIX cluster, a shared volume group can be varied on by only one node at a time. As a general rule, the shared volume groups in an HANFS for AIX cluster should not be activated (varied on) automatically at system boot, but by cluster event scripts.

Logical Volumes

A *logical volume* is a set of logical partitions that AIX makes available as a single storage unit—that is, the logical view of a disk. A *logical partition* is the logical view of a physical partition. Logical partitions may be mapped to one, two, or three physical partitions to implement mirroring.

In the HANFS for AIX environment, logical volumes can be used to support a journaled file system.

Shared Logical Volumes

A shared logical volume must have a unique name within an HANFS for AIX cluster. By default, AIX assigns a name to any logical volume that is created as part of a journaled file system (for example, *lv01*). If you rely on the system-generated logical volume name, this name could cause the import to fail when you attempt to import the volume group containing the logical volume into another node's ODM structure, especially if that volume group already exists. Chapter 9, Defining Shared LVM Components, describes how to change the name of a logical volume.

File Systems

A file system is written to a single logical volume. Ordinarily, you organize a set of files as a file system for convenience and speed in managing data.

Shared File Systems

In an HANFS for AIX cluster, a *shared file system* is a journaled file system that resides entirely in a shared logical volume. You need to plan shared file systems to be placed on external disks shared by cluster nodes. Data resides in file systems on these external shared disks in order to be made highly available.

LVM Mirroring

This section does not apply to the IBM 7135-110 or 7135-210 RAIDiant Disk Arrays, which provide their own data redundancy.

LVM mirroring provides the ability to specify more than one copy of a physical partition. Using the LVM mirroring facility increases the availability of the data in your system. When a disk fails and its physical partitions become unavailable, you still have access to the data if there is a mirror on an available disk. The LVM performs mirroring within the logical volume. Within an HANFS for AIX cluster, you mirror:

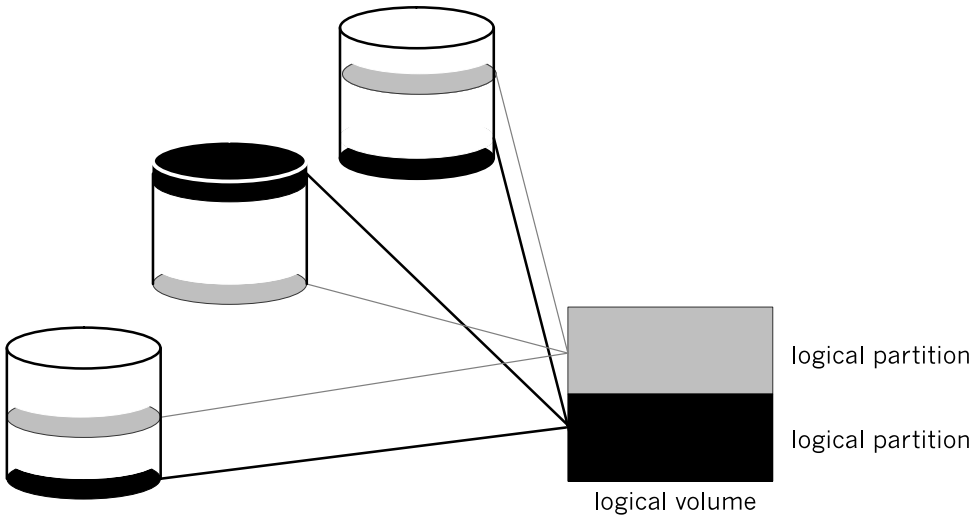
- Logical volume data in a shared volume group
- Log logical volume data for each shared volume group with file systems.

Mirroring Physical Partitions

For a logical volume, you allocate one, two, or three copies of the physical partition that contains data. This allocation lets you mirror data, which improves the availability of the logical volume. If a copy is lost due to an error, the other unaffected copies are accessed, and AIX

continues processing with an accurate copy. After access is restored to the failed physical partition, AIX resynchronizes the contents (data) of the physical partition with the contents (data) of a consistent mirror copy.

The following figure shows a logical volume composed of two logical partitions with three mirrored copies. In the figure, each logical partition maps to three physical partitions. Each physical partition should be designated to reside on a separate physical volume within a single volume group. This provides the maximum number of alternative paths to the mirror copies and, therefore, the greatest availability.



A Logical Volume of Two Logical Partitions with Three Mirrored Copies

The mirrored copies are transparent, meaning that you cannot isolate one of these copies. For example, if a user deletes a file from a logical volume with multiple copies, the deleted file is gone from all copies of the logical volume.

Using mirrored copies improves the availability of data on your cluster. The following considerations also improve data availability:

- Allocating three copies in a logical partition provides greater protection than allocating one or two copies.
- Allocating the copies of a logical partition on different physical volumes provides greater protection than allocating the copies on the same physical volume.
- Allocating the copies of a logical partition on different adapters provides greater protection than allocating the copies on a single adapter.

Keep in mind that anything that improves availability may increase the time necessary for write operations. Nevertheless, using mirrored copies spanning multiple disks (on separate power supplies) together with multiple adapters ensures that no disk is a single point of failure for your cluster.

Mirroring Journal Logs

AIX uses journaling for its file systems. In general, this means that the internal state of a file system at startup (in terms of the block list and free list) is the same state as at shutdown. In practical terms, this means that when AIX starts up, the extent of any file corruption can be no worse than at shutdown.

Each volume group contains a **jfslog**, which is itself a logical volume. This log typically resides on a different physical disk in the volume group than the journaled file system. If access to that disk is lost, however, changes to file systems after that point are in jeopardy.

To avoid the possibility of that physical disk being a single point of failure, you can specify mirrored copies of each **jfslog**. Place these copies on separate physical volumes.

Quorum

This section does not apply to the IBM 7135-110 or 7135-210 RAIDiant Disk Array, which provides its own data redundancy.

Quorum is a feature of the AIX LVM that determines whether or not a volume group can be placed online, using the **varyonvg** command, and whether or not it can remain online after a failure of one or more of the physical volumes in the volume group.

Each physical volume in a volume group has a Volume Group Descriptor Area (VGDA) and a Volume Group Status Area (VGSA).

VGDA Describes the physical volumes (PVs) and logical volumes (LVs) that make up a volume group and maps logical partitions to physical partitions. The **varyonvg** command reads information from this area.

VGSA Maintains the status of all physical volumes and physical partitions in the volume group. It stores information regarding whether a physical partition is potentially inconsistent (stale) with mirror copies on other physical partitions, or is consistent or synchronized with its mirror copies. Proper functioning of LVM mirroring relies upon the availability and accuracy of the VGSA data.

Quorum at Vary On

When a volume group is brought online using the **varyonvg** command, VGDA and VGSA data structures are examined. If more than half of the copies are readable and identical in content, quorum is achieved and the **varyonvg** command succeeds. If exactly half the copies are available, as with two of four, quorum is not achieved and the **varyonvg** command fails.

Quorum after Vary On

If a write to a physical volume fails, the VGSAs on the other physical volumes within the volume group are updated to indicate this physical volume has failed. As long as more than half of all VGDAs and VGSAs can be written, quorum is maintained and the volume group remains varied on. If exactly half or less than half of the VGDAs and VGSAs are inaccessible, quorum is lost, the volume group is varied off, and its data becomes unavailable.

Keep in mind that a volume group can be varied on or remain varied on with one or more of the physical volumes unavailable. However, data contained on the missing physical volume will not be accessible unless the data is replicated using LVM mirroring and a mirror copy of the data is still available on another physical volume. Maintaining quorum without mirroring does not guarantee that all data contained in a volume group is available.

Quorum has nothing to do with the availability of mirrored data. It is possible to have failures that result in loss of all copies of a logical volume, yet the volume group remains varied on because a quorum of VGDAs/VGSAs are still accessible.

Disabling and Enabling Quorum

Quorum checking is enabled by default. In AIX version 3.2.3e and later versions, quorum checking can be disabled using the **chvg -Qn *vgname*** command, or by using the **smit chvg** fastpath.

Quorum Enabled

With quorum enabled, more than half of the physical volumes must be available and the VGDA and VGSA data structures must be identical before a volume group can be varied on with the **varyonvg** command.

With quorum enabled, a volume group will be forced offline if one or more disk failures causes a majority of the physical volumes to be unavailable. Having three or more disks in a volume group avoids a loss of quorum in the event of a single disk failure.

Quorum Disabled

With quorum disabled, *all* the physical volumes in the volume group must be available and the VGDA data structures must be identical for the **varyonvg** command to succeed. With quorum disabled, a volume group will remain varied on until the last physical volume in the volume group becomes unavailable. The following sections summarize the effect quorum has on the availability of a volume group.

Forcing a Varyon

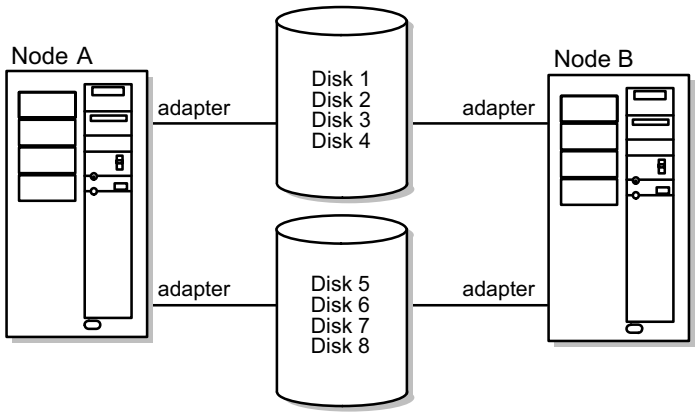
A volume group with quorum disabled and one or more physical volumes unavailable can be “forced” to vary on by using the **-f** flag with the **varyonvg** command. Forcing a varyon with missing disk resources could cause unpredictable results, including a **reducevg** of the physical volume from the volume group. Forcing a varyon should be an overt (manual) action and should only be performed with a complete understanding of the risks involved.

The HANFS for AIX software assumes that a volume group is not degraded and all physical volumes are available when the **varyonvg** command is issued at startup or when a volume group resource is taken over during a fallover. The cluster event scripts provided with the HANFS for AIX software do not “force” varyon with the **-f** flag, which could cause unpredictable results. For this reason, modifying the cluster event scripts to use the **-f** flag is strongly discouraged.

Quorum Considerations for HANFS for AIX

While specific scenarios can be constructed where quorum protection does provide some level of protection against data corruption and loss of availability, quorum provides very little actual protection in HANFS for AIX configurations. In fact, enabling quorum may mask failures by allowing a volume group to varyon with missing resources. Also, designing logical volume configuration for no single point of failure with quorum enabled may require the purchase of additional hardware. In spite of these facts, you must keep in mind that disabling quorum can result in subsequent loss of disks—after varying on the volume group—that go undetected.

Often it is not practical to configure disk resources as shown in the following figure because of the expense. Take, for example, a cluster that requires 8 GB of disk storage (4 GB double mirrored). This requirement could be met with two IBM 9333 or 9334 disk subsystems and two disk adapters in each node. For data availability reasons, logical volumes would be mirrored across disk subsystems.



Quorum in an HANFS for AIX Cluster

With quorum enabled, the failure of a single adapter, cable, or disk subsystem power supply would cause exactly half the disks to be inaccessible. Quorum would be lost and the volume group varied off even though a copy of all mirrored logical volumes is still available. The solution is to turn off quorum checking for the volume group. The tradeoff is that, with quorum disabled, all physical volumes must be available for the **varyonvg** command to succeed.

Major Numbers on Shared Volume Groups

When a node leaves the cluster, NFS clients attached to the cluster operate as they do when a standard NFS server fails and reboots.

To prevent problems with NFS file systems in an HANFS for AIX cluster, make sure that each shared volume group has the same major number on both nodes. If the major numbers are different, client sessions will not be able to recover when the takeover node re-exports an NFS file system after the owner node has left the cluster. Client sessions are not able to recover because the file system exported by the takeover node appears to be different from the one exported by the owner node.

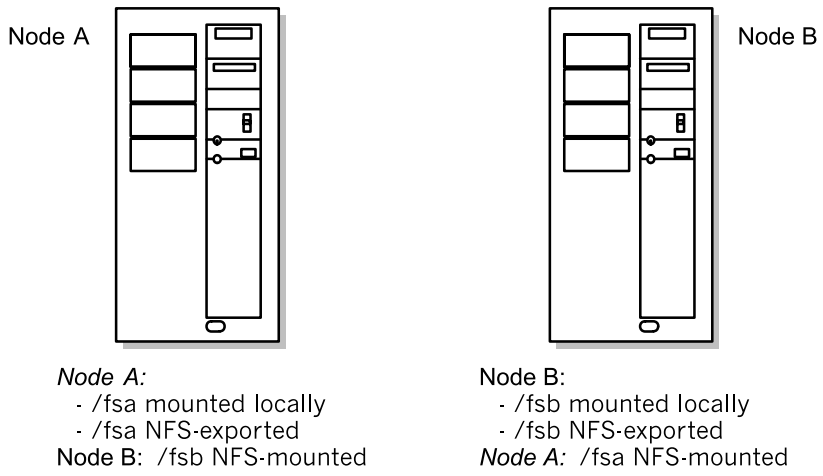
The **lvlstmajor** command lists the free major numbers on a node. Use this command on each node to find a major number that is free on all cluster nodes.

Mount Points for NFS Mounts

If you plan to NFS-mount directories of file systems, you must manually create mount points for those directories on the node which will NFS-mount these directories.

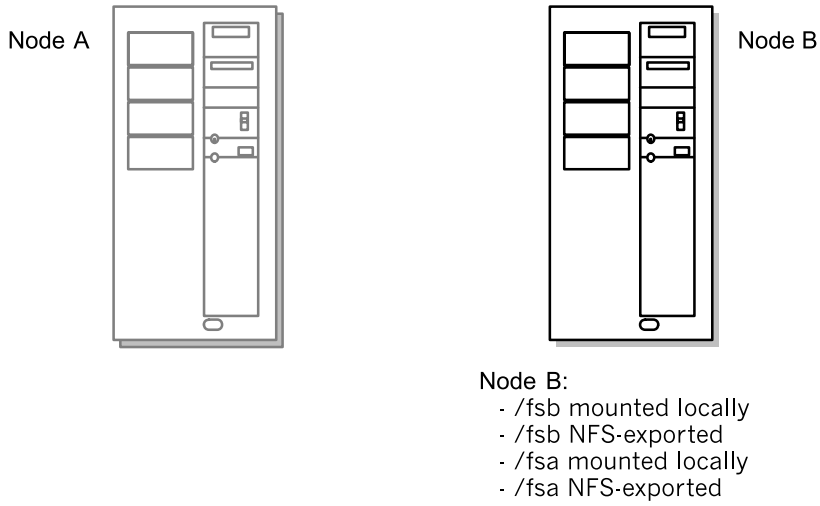
Cross-Mounting File Systems

The HANFS for AIX software allows you to configure a cluster so that nodes can NFS-mount each other's file systems. An example of a cross-mounted NFS configuration is shown in the following figure:



Cross-Mounting during Normal Operation

When Node A fails, Node B closes the open files in the **/fsa** file system, unmounts it, mounts it locally, and re-exports it to waiting clients, as shown in the following figure:



Cross-Mounting after Node Failure

Note: The shared volume groups must have the same major device number on the server nodes, which allows the clients to re-establish the NFS mount transparently after a fallover.

In a cluster like the one shown above, you have to create two resource groups, *NodeA_rg* and *NodeB_rg*. These resource groups are defined as follows:

- Resource Group *NodeA_rg*:
  - Participating node names = Node A Node B
  - File Systems = /fsa (local)
  - File Systems to export = /fsa (local)
  - File Systems to NFS-mount = /fsa (remote)
- Resource Group *NodeB_rg*:
  - Participating node names = Node B Node A
  - File Systems = /fsb (local)
  - File Systems to export = /fsb (local)
  - File Systems to NFS-mount = /fsb (remote)

Setting up your cascading resource groups like this ensures the default server-to-server NFS behavior described above. On reintegration, the **/fsa** file system is passed back to Node A, locally mounted, and exported. Node B mounts it via NFS again.

Note: Make sure that all file systems configured to be NFS-mounted are also configured to be exported.

Planning Guideline Summary

Consider the following guidelines as you plan shared LVM components:

- In general, planning for logical volumes concerns the availability of your data. However, creating logical volume copies is not a substitute for regularly scheduled backups. Backups protect against loss of data regardless of cause; logical volume copies protect against loss of data from physical access failure.
- All operating system files should reside in the root volume group (**rootvg**) and all user data should reside outside that group. This makes it more manageable to update or reinstall the operating system and to back up data.
- Volume groups that contain at least three physical volumes provide the maximum availability when implementing mirroring.
- When using copies, each physical volume containing a copy should get its power from a separate source. If one power source fails, separate power sources maintain the no-single-point-of-failure objective.
- Consider quorum issues when laying out a volume group. With quorum enabled, a two-disk volume group puts you at risk for losing quorum and data access. Either build three-disk volume groups or disable quorum.
- Keep in mind the cluster configurations that you have designed. A node whose resources are not taken over should not own critical volume groups.

Completing the Shared LVM Components Worksheets

Fill out the planned physical and logical storage for your cluster on the shared LVM components worksheets. Refer to the completed worksheet when you define the shared LVM components following the instructions in Chapter 9, Defining Shared LVM Components, and the cluster resource configuration following the instructions in Chapter 12, Configuring an HANFS for AIX Cluster.

The shared LVM components worksheets include the:

- Non-Shared Volume Groups Worksheet
- Shared Volume Group/File Systems Worksheet
- NFS File Systems Worksheet

Completing the Non-Shared Volume Groups Worksheet

For each node in the cluster, complete the information for each volume group residing on a local (non-shared) disk.

1. Fill in the node name in the **Node Name** field.
2. Record the name of the volume group in the **Volume Group Name** field.
3. Enter the name of the logical volume in the **Logical Volume Name** field.
4. List the device names of the physical volumes comprising the volume group in the **Physical Volumes** field.

In the remaining sections of the worksheet, enter the following information for each logical volume in the volume group. Use additional sheets if necessary.

5. Enter the name of the logical volume in the **Logical Volume Name** field.
6. If you are using LVM mirroring, indicate the number of logical partition copies (mirrors) in the **Number Of Copies Of Logical Partition** field. You can specify one or two copies (in addition to the original logical partition, for a total of three).
7. If you are using LVM mirroring, specify whether each copy will be on a separate physical volume in the **On Separate Physical Volumes?** field.
8. Record the full-path mount point of the file system in the **File System Mount Point** field.
9. Record the size of the file system in 512-byte blocks in the **Size** field.

Completing the Shared Volume Group/File Systems Worksheet

Fill out the information for each volume group that will reside on the shared disks. You need a separate worksheet for each shared volume group, so be sure to make sufficient copies of the worksheet before you begin.

1. Fill in the name of each node in the cluster in the **Node Names** field.
2. Record the name of the shared volume group in the **Shared Volume Group Name** field.
3. Record the major number of the shared volume group in the **Major Number** field.
4. Record the name of the log logical volume (**jfslog**) in the **Log Logical Volume Name** field.
5. Pencil-in the planned physical volumes in the **Physical Volumes** field. You will enter exact values for this field after you have installed the disks following the instructions in Chapter 8, Checking Installed Hardware.

   In the remaining sections of the worksheet, enter the following information for each logical volume in the volume group. Use additional sheets as necessary.
6. Enter the name of the logical volume in the **Logical Volume Name** field.
7. *This step does not apply to the IBM 7135-110 or 7135-210 RAIDiant Disk Arrays.* If you are using LVM mirroring, indicate the number of logical partition copies (mirrors) in the **Number Of Copies of Logical Partition** field. You can specify that you want one or two copies (in addition to the original logical partition, for a total of three).
8. *This step does not apply to the IBM 7135-110 or 7135-210 RAIDiant Disk Arrays.* If you are using LVM mirroring, specify whether each copy will be on a separate physical volume in the **On Separate Physical Volumes?** field.
9. Record the full-path mount point of the file system in the **File System Mount Point** field.
10. Record the size of the file system in 512-byte blocks in the **Size** field.

Completing the NFS File Systems Worksheet

Provide the following information for each node:

1. List the shared file systems to be controlled by this node in the **Shared File Systems Controlled By This Node** field. Include all file systems that contain directories that you want to export from this node.
2. Enter the list of shared file systems and directories to be NFS-exported by this node in the **File Systems/Directories to Export** field.
3. This field is optional. Enter the list of shared file systems and directories that you want to be automatically NFS-mounted on the other node in the **File Systems/Directories to NFS-Mount** field. This list should be a subset of the file systems and directories listed in the **File Systems/Directories to Export** field.

Note to former HA-NFS Version 3 users: If you had a server-backup configuration, only the node you chose to be the server node will have a list of shared file systems. If you had a server-server configuration, each node will have a list of shared file systems.

Where You Go From Here

You have now planned the shared LVM components for your cluster. Use this information when you define the volume groups, logical volumes, and file systems during installation. In the next step of the planning process, you address issues relating to planning for your resource groups. The next chapter describes this step.

Chapter 6 Planning Resource Groups

This chapter describes how to plan the resource groups for an HANFS for AIX cluster.

Planning Resource Groups

A *resource group* is a set of resources that you define in such a way that the HANFS for AIX software can treat them as a single unit. A resource group consists of one or more file systems, the underlying volume groups, and a shared IP address. An HANFS for AIX cluster can have a maximum of 20 resource groups.

You made preliminary choices about the type of resource group and the takeover priority for each node in Chapter 1, Overview of HANFS for AIX. In this chapter, you'll review your choices in light of your interim planning.

The more general steps in planning a resource group are described in the following table:

Step	What you do...
1	Decide whether you want your cluster to use cascading or rotating resource groups.
2	Indicate the order in which nodes take over a given resource group.
3	Identify the individual resources that constitute the resource group.

Guidelines

The HANFS for AIX software does not restrict the number of individual resources in a resource group or the number of resource groups in a cluster. The benefit to this approach is that it gives you flexibility in designing your cluster.

Nevertheless, as a general rule you should strive to keep your design as simple as possible. Doing so makes it easier to configure and maintain resource groups; it also makes the takeover of a resource group faster. Use the following guidelines to plan your cluster:

- Every cluster resource must be part of a resource group. If you want a resource to be kept "separate," you define a group for that resource alone. A resource group may have one or more resources defined.
- A resource may not be included in more than one resource group.
- The components of a resource group must be unique down to the physical volume.
- Both a cascading resource group and a rotating resource group must have a service IP label defined for it.
- To avoid errors, you must configure a cascading resource group with IPAT if the resource group is configured with an NFS-mount point.

- When using a rotating resource group configuration, be sure to include the standby IP address labels in the **/.rhosts** file on each cluster node.
- Make sure that each node has the ability to take all resource groups simultaneously in case of a fallover.

Completing the Resource Group Worksheet

For each node, record the resource group's information on the Resource Group Worksheet:

1. Name the resource group in the **Resource Group Name** field. Use no more than 31 characters. You can use alphabetic or numeric characters, dots, dashes, and underscores. Duplicate entries are not allowed.
2. Record the node/resource relationship (that is, the type of resource for the resource group) in the **Node Relationship** field. Indicate whether the resource group is cascading or rotating. You made a preliminary choice about the type of resource groups you want to use in your cluster in Chapter 1, Overview of HANFS for AIX. Review this choice in light of the subsequent planning.
3. List the name of the node assigned to originally control this resource. This is the first name listed in the **Participating Node Names** field. Again, review your earlier choice in light of the subsequent planning.
4. List the file systems to include in this resource group in the **File Systems** field.
5. In the **File Systems/Directories to Export** field, list the pathname of the file systems in this resource group that should be NFS-exported by the node currently holding the resource. These file systems should be a subset of the file systems listed step 4. You will use this list to add directories of the shared file systems to the **/etc/exports** file.
6. List the file systems in the resource group that should be NFS-mounted by the other node in the resource chain (the node not currently holding the resource) in the **File Systems to NFS-Mount** field.
7. List in the **Volume Groups** field the shared volume groups that should be varied on when this resource group is acquired or taken over.
8. *This field applies only to cascading resource groups.* Indicate in the **Inactive Takeover** field how you want to control the initial acquisition of a resource by a node.
   - If you specify that Inactive Takeover is **false**, the first node up will initially acquire the resource only if it is the node with the higher priority for that resource.
   - If you specify that Inactive Takeover is **true**, the first node in the resource chain to join the cluster acquires the resource. Subsequently the resource cascades to the node in the chain with higher priority as it joins the cluster. Note that this causes an interruption in service as resource ownership transfers to the node with higher priority.

Complete the preceding steps for each resource group in your cluster.

Where You Go From Here

You have now planned your HANFS for AIX cluster and are ready to install the software. Use the planning diagrams and worksheets you completed during the planning process to guide you through the installation process. See Part 3, Installing and Configuring HANFS for AIX, for instructions on installing and configuring your HANFS for AIX cluster.

Part 3 Installing and Configuring HANFS for AIX

In this part, you learn the exact tasks you must perform to install and configure an HANFS for AIX cluster, including checking hardware, defining shared LVM components, installing the HANFS for AIX software, configuring a cluster, setting up monitoring scripts and files, setting up support for error notification, and miscellaneous administrative tasks related to AIX.

Chapter 7 Overview of the Installation and Configuration Process

This chapter provides an overview of how to install and configure the HANFS for AIX software.

Prerequisites

Read Part 2, Planning HANFS for AIX, before installing HANFS for AIX software. It contains the necessary worksheets and diagrams to consult as you proceed through the installation and configuration steps listed in this chapter. If you have not completed these worksheets and diagrams, return to the appropriate chapters and do so before continuing.

Steps for Installing and Configuring an HANFS for AIX Cluster

This section identifies the steps required to set up, install, and configure an HANFS for AIX cluster. Steps are divided into the following major areas:

- Preparing AIX for an HANFS cluster—setting up hardware and software in AIX
- Installing and configuring an HANFS cluster—configuring the cluster to handle resources according to your specifications.

Preparing AIX for an HANFS for AIX Cluster

Step 1: Check Installed Hardware

In this step you ensure that network adapters and shared external disk devices are ready to support an HANFS for AIX cluster as described in Chapter 8, Checking Installed Hardware.

Step 2: Define Shared LVM Components

In this step you create the shared volume groups, logical volumes, and file systems for your cluster as described in Chapter 9, Defining Shared LVM Components.

Step 3: Perform Additional AIX Administrative Tasks

In this step you review or edit various AIX files to ensure a proper configuration for I/O pacing, network options, and for various host files as described in Chapter 10, Performing Additional AIX Tasks.

Installing and Configuring HANFS for AIX Software

Step 4: Install HANFS for AIX Software

In this step you install the HANFS for AIX software on each cluster node as described in Chapter 11, Installing HANFS for AIX Software.

Step 5: Configure the HANFS Cluster

In this step you define the components of your HANFS for AIX cluster as described in Chapter 12, Configuring an HANFS for AIX Cluster.

Step 6: Set up the Cluster Information Program

In this step you edit the **/usr/sbin/cluster/etc/clhosts** file and the **/usr/sbin/cluster/etc/clinfo.rc** script as described in Chapter 13, Configuring Monitoring Scripts and Files.

Step 7: Enable AIX Error Notification Facility Support (optional)

In this step you use the AIX Error Notification facility or the Automatic Error Notify method to identify and respond to failures within an HANFS for AIX cluster as described in Chapter 14, Supporting AIX Error Notification.

The installation is complete when you have performed these tasks.

Chapter 8 Checking Installed Hardware

This chapter describes how to verify that network adapters and shared external disk devices are ready to support an HANFS for AIX cluster. The chapter presumes that you have already installed the devices following the instructions in the relevant AIX documentation.

Checking Network Adapters

This section describes how to ensure that network adapters are configured properly to support the HANFS for AIX software. For each node, check the settings of each adapter to make sure they match the values on the completed copies of the TCP/IP Network Adapter Worksheet. Special considerations for a particular adapter type are discussed below.

Ethernet, Token-Ring, and FDDI Adapters

- When using the **smit mktcpip** fastpath to define an adapter, the HOSTNAME field changes the default hostname. For instance, if you configure the first adapter as *clam_svc*, and then configure the second adapter as *clam_sby*, the default hostname at system boot is *clam_sby*.

  To avoid this problem, it is recommended that you configure the adapter with the desired default hostname last. The hostname must be associated with an IP address that will not change while the system is running.
- Use the **smit chinet** or **smit chghfcs** fastpath to configure each adapter for which IP address takeover might occur to boot from the boot adapter address and not from its service adapter address. Refer to your completed copies of the TCP/IP Network Adapter Worksheet.

SOCC Optical Link

Use the **smit mktcpip** fastpath to make sure that the **START Now** field is set to **no**.

SLIP Line

Test communication over the SLIP line:

1. Run the **netstat -i** command to make sure the SLIP line is recognized. You should see the device listed as **sl1**.
2. On the first node, enter the following command:

   ```
   ping IP_address_of_other_node
   ```
   where *IP_address_of_other_node* is the address (in dotted decimal) you configured as the destination address for the other node.
3. Do the same on the second node, entering the destination address of the first node:

   ```
   ping IP_address_of_other_node
   ```

Note to Former HA-NFS Version 3 Users on Service Adapters

Service (primary) adapters must be configured to default to the up state for HANFS for AIX to work properly, rather than in the down state, as they were in HA-NFS Version 3. To make sure all adapters are configured correctly, complete the following steps for the service (primary) network adapter on each node in the cluster:

1. Enter the **smit chinet** fastpath.
2. Select the service (primary) adapter.
3. Make sure the **Current STATE** field is set to **up**.
4. If necessary, press the Enter key to save your changes.
5. Press the F10 key to exit SMIT.

Note to Former HA-NFS Version 3 Users on Hardware Address Swapping

Hardware address takeover is an optional feature in HANFS for AIX, and it is configured differently than it was in HA-NFS Version 3. In HANFS for AIX, the hardware address of a standby (secondary) adapter cannot be changed to that of the other node's service (primary) adapter.

If you use hardware address swapping, make sure that the alternate hardware address is a new address not in use elsewhere on the physical network, as described in Chapter 3, Planning HANFS for AIX Networks.

If you do not use hardware address swapping, complete the following steps for the standby (secondary) network adapter on each node to make sure that all adapters are configured correctly:

1. Enter the **smit chgenet** fastpath.
2. Select the standby (secondary) adapter.
3. Make sure the **Enable ALTERNATE address** field is set to **no**.
4. If necessary, press the Enter key to save your changes.
5. Press the F10 key to exit SMIT.

Completing the TCP/IP Network Adapter Worksheets

After checking all network interfaces for a node, record the network interface names on that node's TCP/IP Network Adapter Worksheet. Enter the following command:

```
lsdev -Cc if
```

This displays a list of "Available" and "Defined" adapters for the node. At this point, all interfaces used by the HANFS for AIX software should be available. List the adapters marked "Available" in the **Interface Name** field on the TCP/IP Network Adapter Worksheet.

Serial Networks

It is strongly recommended that a serial network connect the nodes in an HANFS for AIX cluster. The serial network allows the Cluster Managers to continue to exchange keepalive packets should the TCP/IP-based subsystem, networks, or network adapters fail. Thus, the

serial network prevents the nodes from becoming isolated and attempting to take over shared resources. The HANFS for AIX software supports three types of serial networks: RS232 lines, the SCSI-2 Differential bus using target mode SCSI, and target mode SSA links.

See Chapter 3, Planning HANFS for AIX Networks, for more information on serial networks. See Appendix B, Configuring Serial Networks, for instructions on configuring serial networks in an HANFS for AIX cluster.

Checking an SP Switch

The SP Switch used by an SP node serves as a network device for configuring multiple clusters, and it can also connect clients. This switch is not required for an HANFS for AIX installation. When installed, the SP Switch Network Module default settings are sufficient to allow it to operate effectively in an HANFS for AIX cluster environment.

Basic points to remember about the SP Switch in an HANFS for AIX configuration:

- ARP must be enabled for the SP Switch network so that IP address takeover can work.
- All SP Switch addresses must be defined on a private network.
- HANFS for AIX SP Switch boot and service addresses are alias addresses on the SP Switch css0 IP interface. The css0 base IP address is unused and should not be configured for IP address takeover. Standby adapters are not used for SP Switch IP address takeover. The alias service addresses appear as **ifconfig alias** addresses on the css0 interface.
- The netmask associated with the css0 base IP address will be used as the netmask for all HANFS for AIX SP Switch network adapters.

For more information about the SP Switch and the SP, see Appendix C, Installing and Configuring HANFS for AIX on RS/6000 SPs.

Configuring for Asynchronous Transfer Mode (ATM)

An Asynchronous Transfer Mode (ATM) network is a connection-oriented, point-to-point network. It must be defined as a private network when defining it to the HANFS cluster topology because it does not support broadcasting. It does, however, handle keepalive traffic between Cluster Managers.

ATM networks require the use of another system as an ARP server to handle ATM hardware address-to-IP address resolutions. One ARP server exists for each subnetwork defined; typically, two ARP servers exist in an HANFS cluster—one for the service network and one for the standby networks. An AIX-based machine can serve several ATM subnetworks.

Warning: An ATM ARP server can be an HANFS for AIX client, but it cannot function as an HANFS for AIX server because hardware address swapping is not supported and because ATM ARP clients require the ATM ARP server hardware address.

ATM can support multiple interfaces per ATM adapter; however, adapters used in HANFS for AIX servers must use only one interface per adapter, and interface numbers must match the adapter numbers. HANFS for AIX clients are not restricted in this way.

Note: The current support provided for ATM networks used with HANFS allows a full client network implementation and requires that the interfaces be set up as Switched Virtual Circuits, through the ATM switch, and as an ARP server on a non-cluster node.

After installing the ATM adapters and the network, you can configure the ATM network by completing the following general steps:

Step	What you do...
1	Determine which non-cluster machine will function as the ARP server.
2	Configure the first ATM interface (at0) as an ARP server for the service subnet.
3	Configure an additional interface (at1) as an ARP server for the standby subnet. Configure this interface on the same adapter used to configure the service subnet; for example at0 and at1 on atm0 adapter.
4	Determine both ATM hardware addresses of the ARP server.
5	Configure the service and standby ATM adapters in AIX to use the defined ARP server hardware addresses. Perform this step on each cluster node.
6	Define the ATM adapters to the HANFS cluster topology as a private network.
7	Test the configuration.

See Chapter 12, Configuring an HANFS for AIX Cluster, for information about defining and configuring adapters and their network type and attributes.

Configuring an ATM ARP Server for HANFS

Configuring an ATM network requires that you configure its ARP server for HANFS, which requires that you complete the procedures in this section for the service, standby, or additional interfaces.

Configuring ARP Service for HANFS Service and Other Interfaces

To configure an ARP server for the HANFS service interface (subnetwork), or for other interfaces:

1. Enter:

   ```
   smitty chinet
   ```

   SMIT displays a similar list of supported network interfaces.

2. Select **at0** as the ATM network interface. This interface will serve as the ARP server for the subnetwork 192.168.110.

   Note: The **Connection Type** field is set to **svc_s** to indicate that the interface is used as an ARP server.

3. Press F10 to exit SMIT.

Configuring ARP Service for HANFS Standby Interfaces

To configure an ARP server for an HANFS standby interface (subnetwork):

1. Enter:

   ```
   smitty chinet
   ```
   SMIT displays a similar list of supported network interfaces.

2. Select **at1** as the ATM network interface.

3. Set the remaining fields are set as follows:
   - **Connection Type** should designate the interface as an ARP server, **svc_s**
   - **Alternate Device** field should be set to **atm0**. This setting puts at1 on atm0 with at0.

Obtaining an ARP Server's Hardware Addresses:

To show an ARP server's hardware addresses, use the **arp** command as follows on the ATM ARP server:

```
arp -t atm -a svc
```

Note: The ATM ARP server address is the 20-byte hardware address of the ATM ARP server for the subnet of an Internet address.

Configuring ATM on Cluster Nodes

To configure HANFS nodes (ARP clients) in AIX:

1. Use the **smitty chinet** command, as in the last two procedures, to configure two ATM interfaces, one on each adapter (**at0** on atm0 for "service," and **at1** on atm1 for "standby").

2. Indicate the following differences for these interfaces:

Note: Set the hardware address found previously for this subnet.

Testing Communication Over the Network

To test communication over the network after configuring ARP servers:

1. Run the **netstat -i** command to make sure the ATM network is recognized. You should see the device listed as **at1**.

2. Enter the following command on the first node:

   ```
   ping IP_address_of_other_node
   ```
   where *IP_address_of_other_node* is the address (in dotted decimal) that you configured as the destination address for the other node.

3. Repeat Steps 1 and 2 on the second node, entering the destination address of the first node as follows:

   ```
   ping IP_address_of_other_node
   ```

You must perform the first two steps on each node connected to the ATM network, and then test the ATM network. After configuring the ATM network interfaces for a node, record the interface names on that node's TCP/IP Network Adapter Worksheet.

Defining the ATM Network

After you have installed and tested an ATM network, you must define it to the HANFS cluster topology as a private network. Chapter 12, Configuring an HANFS for AIX Cluster, describes how to define an ATM network in an HANFS cluster.

Checking Shared External Disk Devices

This section describes how to verify that shared external disk devices are configured properly for the HANFS for AIX software. Separate procedures are shown for SCSI-2 Differential disk devices, IBM SCSI-2 Differential Disk Arrays, IBM 9333 serial disk subsystems, and IBM Serial Storage Architecture (SSA) disk subsystems.

Verifying Shared SCSI-2 Differential Disks

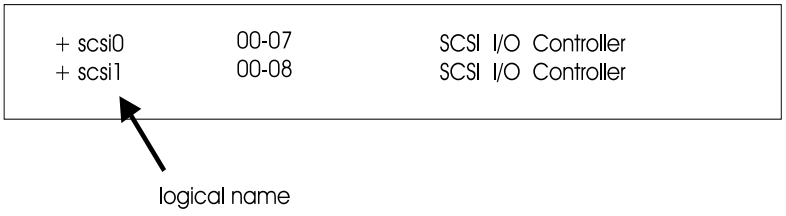
Complete the following steps to verify that a SCSI-2 Differential disk is installed correctly. These steps are valid for both SCSI-2 Differential and SCSI-2 Differential Fast/Wide disks. Differences in procedures are noted as necessary. As you verify the installation, record the shared disk configuration on the Shared SCSI-2 Differential Disk Worksheet. Use a separate worksheet for each set of shared SCSI-2 disks. You will refer to the completed worksheets when you configure the cluster.

1. Note the type of SCSI bus installation, SCSI-2 Differential or SCSI-2 Differential Fast/Wide, in the **Type of SCSI Bus** section of the worksheet.
2. Fill in the node name of each node that connects to the shared SCSI bus in the **Node Name** field.
3. Enter the logical name of each adapter in the **Logical Name** field.

   To determine the logical name, use the command:

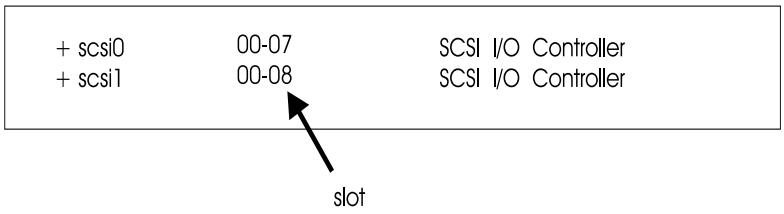
   ```
   lscfg | grep scsi
   ```

   The first column lists the logical name of the SCSI adapters.



4. Record the Microchannel I/O slot that each SCSI adapter uses in the **Slot Number** field.

   The second column of the existing display lists the SCSI-2 Differential adapter's location code in the format AA-BB. The last digit of that value (the last B) is the Microchannel slot number.



5. Record the SCSI ID of each SCSI adapter on each node in the **Adapter** field. To determine the SCSI IDs of the disk adapters, use the **lsattr** command, as in the following example to find the ID of the adapter *scsi1*:

   For SCSI-2 Differential adapters:

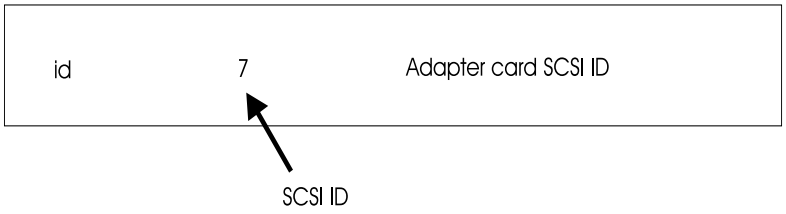
   ```
   lsattr -E -l scsi1 | grep id
   ```

   For SCSI-2 Differential Fast/Wide adapters:

   ```
   lsattr -E -l ascsi1 | grep external_id
   ```

   Do not use wildcard characters or full pathnames on the command line for the device name designation.

   In the resulting display, the first column lists the attribute names. The integer to the right of the **id** (or **external_id**) attribute is the adapter SCSI ID.

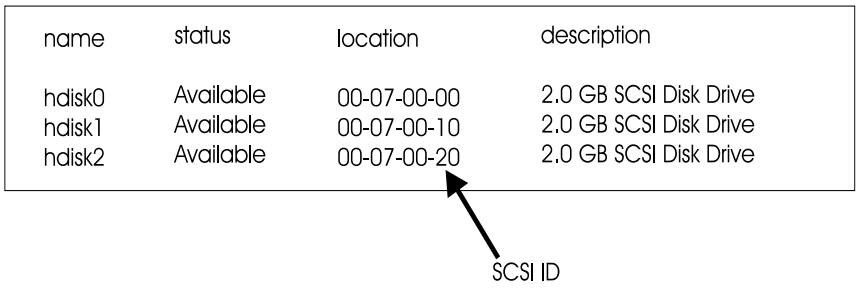


   Note: The IBM High Performance SCSI-2 Differential Fast/Wide Adapter cannot be assigned SCSI IDs 0, 1, or 2. The IBM SCSI-2 Differential Fast/Wide Adapter/A cannot be assigned SCSI IDs 0 or 1.

6. Record the SCSI IDs of the physical disks in the **Shared Drive** fields. Use the command:

   ```
   lsdev -Cc disk -H
   ```

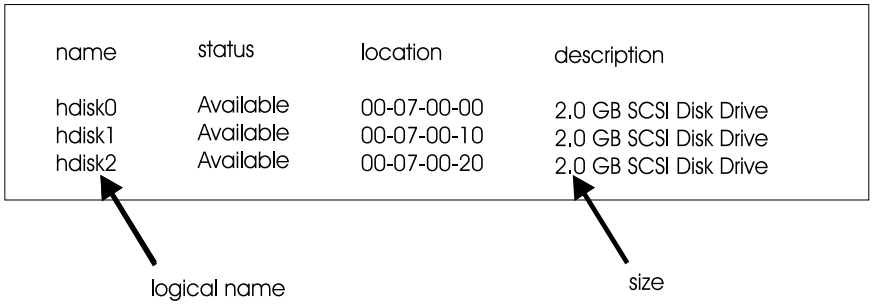
   The third column of the display generated by the **lsdev -Cc disk -H** command is a numeric location with each row in the format AA-BB-CC-DD. The first digit (the first D) of the DD field is the SCSI ID.



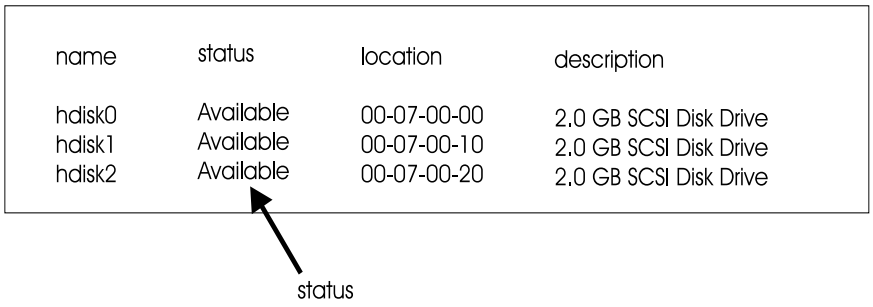
Note: For SCSI-2 Differential Fast/Wide disks, a comma follows the SCSI ID field of the location code since the IDs can require two digits, such as 00-07-00-12,0.

7. At this point, verify that each SCSI device connected to the shared SCSI bus has a unique ID. A common configuration is to set the SCSI ID of the adapters on the nodes to be higher than the SCSI IDs of the shared devices. (Devices with higher IDs take precedence in SCSI bus contention.) For example, the adapter on one node can have SCSI ID 6 and the adapter on the other node can be SCSI ID 7, and the external disk SCSI IDs should be an integer from 0 through 5.
8. Determine the logical names of the physical disks.

   The first column generated by the **lsdev -Cc disk -H** command lists the logical names of the SCSI disks.



   Record the name of each external SCSI disk in the **Logical Device Name** field, and the size in the **Size** field. *Be aware that the nodes can assign different names to the same physical disk. You should note these situations on the worksheet.*
9. Verify that all disks have a status of "Available." The second column of the display generated by the **lsdev -Cc disk -H** command shows the status.



If a disk has a status of "Defined" instead of "Available," check the cable connections and then use the **mkdev** command to make the disk available.

At this point, you have verified that the SCSI-2 Differential disk is configured properly for the HANFS for AIX environment.

Verifying IBM SCSI-2 Differential Disk Arrays

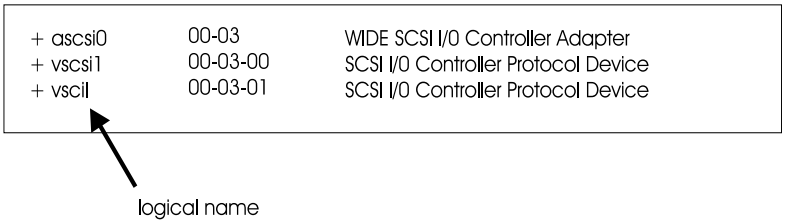
Complete the following steps to verify that an IBM 7135 RAIDiant Disk Array or an IBM 7137 Disk Array is installed correctly. As you verify the installation, record the shared disk configuration on the Shared SCSI-2 Differential Disk Array Worksheet. Use a separate worksheet for each disk array. You will refer to the completed worksheets when you define the cluster topology.

1. Fill in the name of each node connected to this shared SCSI-2 Differential bus in the **Node Name** field.
2. Record the logical device name of each adapter in the **Adapter Logical Name** field.

   To determine the logical device name, use the **lscfg** command, as in the following example:

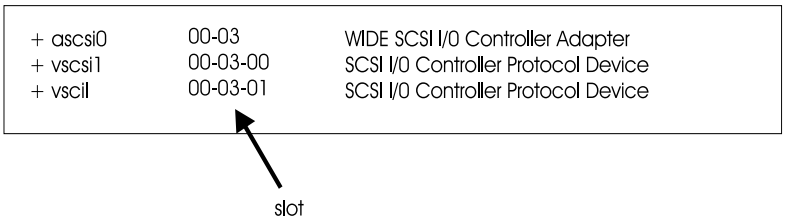
   ```
   lscfg | grep scsi
   ```

   (The following examples display the information for an IBM 7135-110 RAIDiant Disk Array using SCSI-2 Differential Fast/Wide adapters.)



   The first column of the display generated by the **lscfg** command lists the logical name of the SCSI adapters.

3. Record the Microchannel I/O slot of each SCSI-2 Differential adapter used in this shared SCSI bus in the **Slot Number** field of the configuration worksheet.



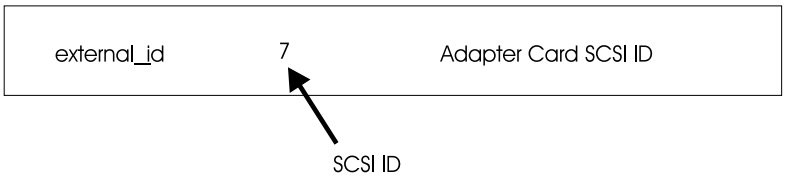
In the existing display, the second column lists the location code of the adapter in the format AA-BB. The last digit of that value (the last B) is the Microchannel slot number.

4. Record the SCSI ID of each SCSI adapter on each node in the **Adapter** field. To determine the SCSI IDs of the disk adapters, use the **lsattr** command, specifying the logical name of the adapter as an argument. In the following example, the SCSI ID of the adapter named *ascsi0* is obtained:

   ```
   lsattr -E -l ascsi0 | grep external_id
   ```

   Do not use wildcard characters or full pathnames on the command line for the device name designation.

   In the resulting display, the first column lists the attribute names. The integer to the right of the **id** (**external_id**) attribute is the adapter SCSI ID.

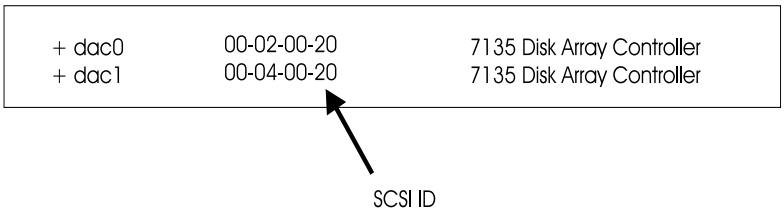


   Note: The IBM High Performance SCSI-2 Differential Fast/Wide Adapter cannot be assigned SCSI IDs 0, 1, or 2. The IBM SCSI-2 Differential Fast/Wide Adapter/A cannot be assigned SCSI IDs 0 or 1.

5. Record the SCSI IDs of the array controllers in the **Array Controller** fields. To determine that AIX has the correct SCSI IDs for the array controllers, obtain a listing of the array controllers using the **lscfg** command, as in the following example:

   ```
   lscfg | grep dac
   ```

   In the display generated by the command, the first column lists the logical names of the array controllers. The second column contains the location code of the array controller in the format AA-BB-CC-DD. The seventh digit of the location code (the first D) is the SCSI ID of the array controller.



Note: Since these controllers are on separate SCSI buses, they can have the same SCSI ID.

The SCSI ID in the display should match the numbers you set for the SCSI ID on each array controller.

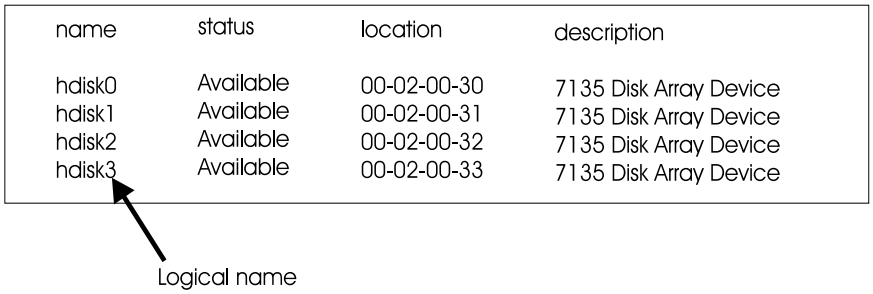
6. At this point, determine that each device connected to a shared SCSI bus has a unique SCSI ID. A common configuration is to let one of the nodes keep the default SCSI ID 7 and assign the adapter on the other cluster node SCSI ID 6. The array controller SCSI IDs should later be set to an integer starting at 0 and going up. Make sure no array controller has the same SCSI ID as any adapter.

7. Verify that AIX created the physical volumes (hdisks) that you expected.

   To determine the logical names of the LUNs on the disk array, use the **lsdev** command, as in the following example:

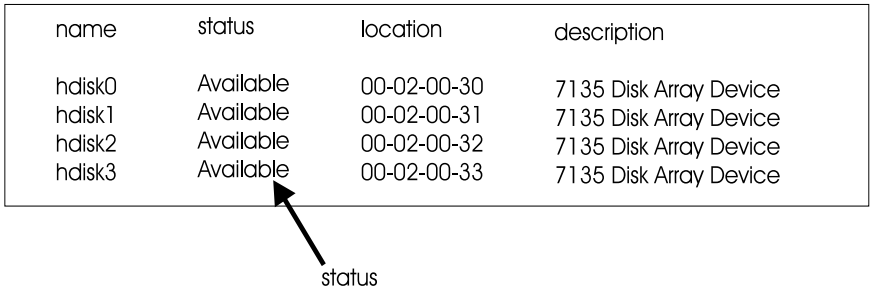
   ```
   lsdev -Cc disk -H
   ```

   The example illustrates how this command lists the hard disks created for a four-LUN 7135-110 RAIDiant Disk Array. The display includes the location code of the hard disk in the form AA-BB-CC-DD. The last digit of the location code included in the display represents the LUN number. The first column lists the logical names of the LUNs on the 7135-110 RAIDiant Disk Array.



   Record the logical name of each LUN in the **Logical Device Name** field. *Be aware that the nodes can assign different names to the same physical disk. You should note these situations on the worksheet.*

8. Verify that all disks have a status of "Available." The second column of the existing display shows the status.



If a disk has a status of "Defined" instead of "Available," check the cable connections and then use the **mkdev** command to make the disk available.

At this point, you have verified that the disk array is configured properly for the HANFS for AIX software.

Target Mode SCSI Connections

It is strongly recommended that a serial network connect the nodes in an HANFS for AIX cluster. The serial network allows the Cluster Managers to continue to exchange keepalive packets should the TCP/IP-based subsystem, networks, or network adapters fail. Thus, the serial network prevents nodes from becoming isolated and attempting to take over shared resources. The HANFS for AIX software supports two types of serial networks: RS232 lines and the SCSI-2 Differential bus using target mode SCSI.

See Chapter 3, Planning HANFS for AIX Networks, for more information on serial networks. See Appendix B, Configuring Serial Networks, for instructions on configuring target mode SCSI connections as serial networks in an HANFS for AIX cluster.

Verifying Shared IBM 9333 Serial Disk Subsystems

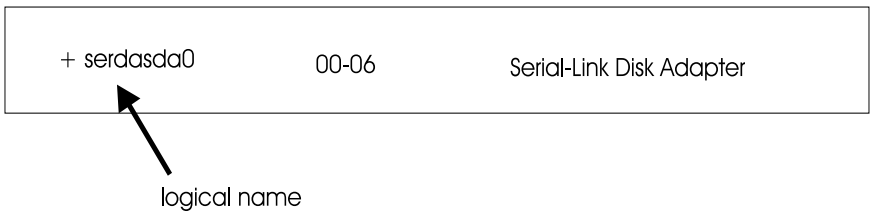
Complete the following steps to verify a shared IBM 9333 serial disk subsystem. As you verify the installation, record the shared disk configuration on the Shared IBM 9333 Serial Disk Subsystem Worksheet. Use a separate worksheet for each set of shared IBM 9333 serial disks. You will refer to the completed worksheets when you configure the cluster.

1. Fill in the node name of each node connected to the shared IBM 9333 serial disk subsystem in the **Node Name** field.
2. Record the logical device name of each adapter in the **Logical Name** field.

   To get the logical device name, enter the following command at each node:

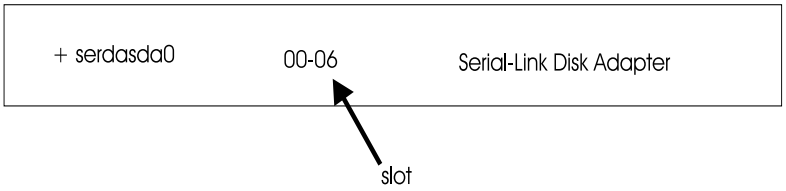
   ```
   lscfg | grep serdasda
   ```

   The first column of the resulting display lists the logical device names of the 9333 adapters.



3. For each node, record the Microchannel I/O slot that each adapter uses in the **Slot Number** field. The slot number value is an integer value from 1 through 16.

   The second column of the existing display lists a value of the form AA-BB. The last digit of that value (the last B) is the Microchannel slot number.

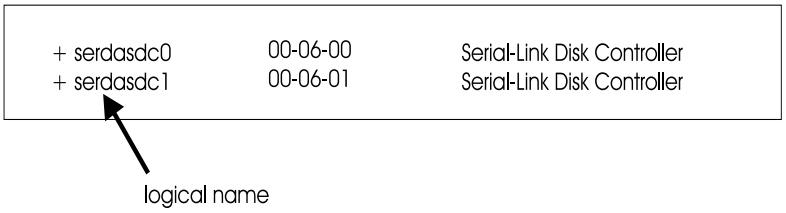


4. For each drawer on each node, record the integer signifying the 9333 adapter I/O connector to which the drawer is connected in the **Adapter I/O Controller** field.

5. For each drawer on each node, record the logical name for the 9333 serial-link controller in the **Controller** field. To get the name, enter:

   ```
   lscfg | grep serdasdc
   ```

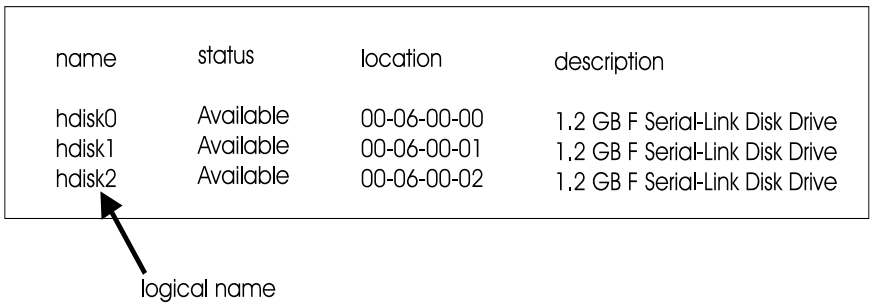
   The first column of the resulting display lists the logical names of the 9333 serial-link controllers.



6. Determine the logical device name and size of each physical volume and record the values on the worksheet. On each node use the command:

   ```
   lsdev -Cc disk -H
   ```

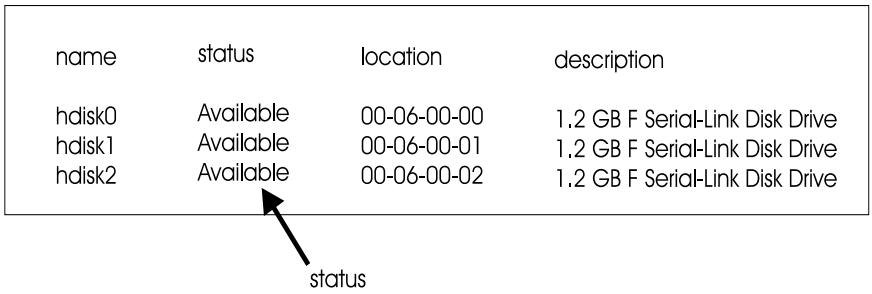
   The first column of the resulting display lists the logical names of the disks.



   Enter the name in the **Logical Device Name** field.

   Record the size of each external disk in the **Size** field.

7. Verify that all disks have a status of "Available." The second column of the existing display indicates the disk status.



If a disk has a status of "Defined" instead of "Available," check the cable connections and then use the **mkdev** command to make the disk available.

At this point, you have verified that the IBM 9333 serial disk is configured properly for the HANFS for AIX software.

Verifying Shared IBM SSA Disk Subsystems

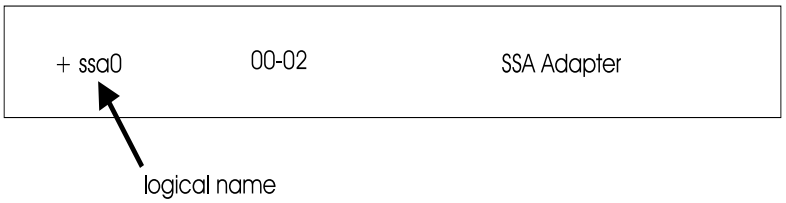
Complete the following steps to verify a shared IBM SSA disk subsystem. As you verify the installation, record the shared disk configuration on the Shared IBM SSA Disk Subsystem Worksheet. Use a separate worksheet for each set of shared IBM SSA disk subsystems. You will refer to the completed worksheets when you configure the cluster.

1. Fill in the node name of each node connected to the shared IBM SSA disk subsystem in the **Node Name** field.
2. Record the logical device name of each adapter in the **Adapter Logical Name** field.

   To get the logical device name, enter the following command at each node:

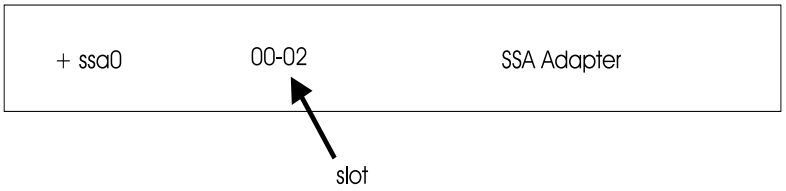
   ```
   lscfg | grep ssa
   ```

   The first column of the resulting display lists the logical device names of the SSA adapters.



3. For each node, record the slot that each adapter uses in the **Slot Number** field. The slot number value is an integer value from 1 through 16.

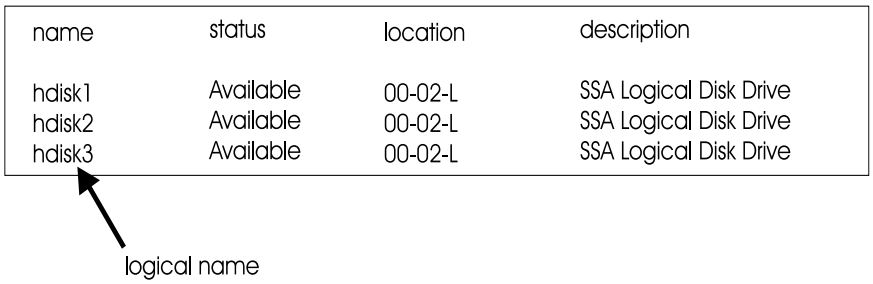
   The second column of the existing display lists a value of the form AA-BB. The last digit of that value (the last B) is the Microchannel slot number.



4. Determine the logical device name and size of each physical volume and record the values on the worksheet. On each node use the command:

   ```
   lsdev -Cc disk | grep -i ssa
   ```

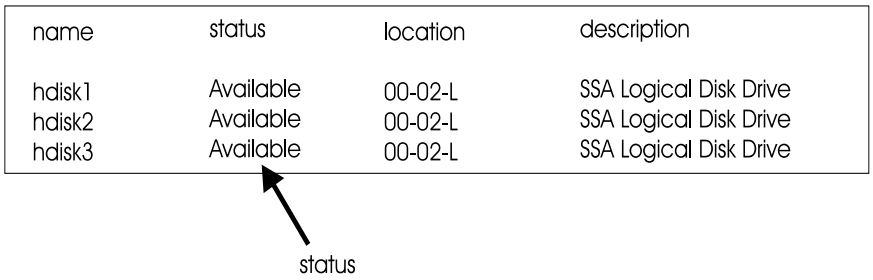
   The first column of the resulting display lists the logical names of the disks.



   Enter the name in the **Logical Device Name** field.

   Record the size of each external disk in the **Size** field.

5. Verify that all disks have a status of "Available." The second column of the existing display indicates the disk status.



   If a disk has a status of "Defined" instead of "Available," check the cable connections and then use the **mkdev** command to make the disk available.

At this point, you have verified that the IBM SSA disk subsystem is configured properly for the HANFS for AIX environment.

Chapter 9 Defining Shared LVM Components

This chapter describes how to define the LVM components shared by the nodes in an HANFS for AIX cluster.

Defining Shared LVM Components

Creating the volume groups, logical volumes, and file systems shared by the nodes in an HANFS for AIX cluster requires that you perform steps on both nodes in the cluster. In general, you define the components on one node (referred to in the text as the source node) and then import the volume group on the other node in the cluster (referred to as destination node). This ensures that the ODM definitions of the shared components are the same on both nodes.

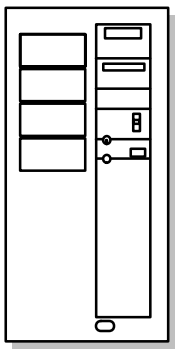
HANFS for AIX environments typically use journaled file systems to manage data. (In some cases, a database application may bypass the journaled file system and access the raw logical volume directly.)

The key consideration, however, is whether a non-concurrent access environment uses mirrors. Shared logical volumes residing on non-RAID disk devices should be mirrored in AIX to eliminate the disk as a single point of failure. Shared volume groups residing on a RAID device should not be AIX mirrored; the disk array provides its own data redundancy.

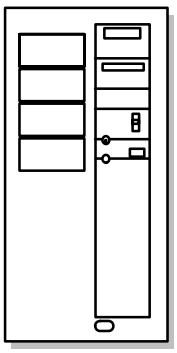
Note: The discussion of the IBM 7135 RAIDiant Disk Array assumes you are using RAID level 1, 3, or 5. RAID level 0 does not provide data redundancy and therefore is not recommended for use in an HANFS for AIX configuration.

The following figures show the tasks you complete to define the shared LVM components with and without mirrors. Each task is described in the sections following the figures. Refer to your completed copies of the shared volume group worksheets as you define the shared LVM components.

Defining Shared LVM Components with Mirrors



Source Node

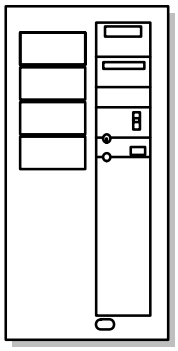


Destination Nodes

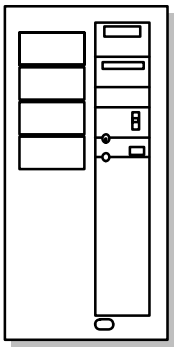
Create volume group
Create journaled file system
Rename jfslog and logical volume
Mirror jfslog and logical volume
Vary off volume group

Import volume group
Change volume group to remain dormant at startup
Vary off volume group

Defining Shared LVM Components without Mirrors



Source Node



Destination Nodes

Create volume group
Create journaled file system
Rename jfslog and logical volume
Vary off volume group

Import volume group
Change volume group to remain dormant at startup
Vary off volume group

Creating a Shared Volume Group on the Source Node

Use the **smit mkvg** fastpath to create a shared volume group. Use the default field values unless your site has other requirements, or unless you are specifically instructed otherwise here.

VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
Activate volume group AUTOMATICALLY at system restart?	Set to **no** so that the volume group can be activated as appropriate by the cluster event scripts.
ACTIVATE volume group after it is created?	Set to **yes**.
Volume Group MAJOR NUMBER	Make sure to use the same major number on all nodes. Use the **lvlstmajor** command on each node to determine a free major number common to all nodes.

Creating a Shared File System on the Source Node

Use the **smit crjfs** fast path to create the shared file system on the source node. When you create a journaled file system, AIX creates the corresponding logical volume. Therefore, you do not need to define a logical volume. You do, however, need to later rename both the logical volume and the log logical volume for the file system and volume group.

Mount AUTOMATICALLY at system restart?	Make sure this field is set to **no**.
Start Disk Accounting	Make sure this field is set to **no**.

Renaming a jfslog and Logical Volumes on the Source Node

AIX assigns a logical volume name to each logical volume it creates. Examples of logical volume names are */dev/lv00* and */dev/lv01*. Within an HANFS for AIX cluster, the name of any shared logical volume *must* be unique. Also, the journaled file system log (**jfslog**) is a logical volume that requires a unique name in the cluster.

To make sure that logical volumes have unique names, rename the logical volume associated with the file system and the corresponding **jfslog** logical volume. Use a naming scheme that indicates the logical volume is associated with a certain file system. For example, *lvsharefs* could name a logical volume for the */sharefs* file system.

1. Use the **lsvg -l *volume_group_name*** command to determine the name of the logical volume and the log logical volume (**jfslog**) associated with the shared volume groups. In the resulting display, look for the logical volume name that has type **jfs**. This is the logical volume. Then look for the logical volume name that has type **jfslog**. This is the log logical volume.
2. Use the **smit chlv** fastpath to rename the logical volume and the log logical volume.

   After renaming the **jfslog** or a logical volume, check the **/etc/filesystems** file to make sure the **dev** and **log** attributes reflect the change. Check the **log** attribute for each file system in the volume group and make sure that it has the new **jfslog** name. Check the **dev** attribute for the logical volume you renamed and make sure that it has the new logical volume name.

Adding Copies to Logical Volume on the Source Node

To add logical volume copies on the source node:

1. Use the **smit mklvcopy** fastpath to add copies to a logical volume. Add copies to both the **jfslog** log logical volume and the logical volumes in the shared file systems. To avoid space problems, first mirror the **jfslog** log logical volume and then the shared logical volumes.

   The copies should reside on separate disks that are controlled by different disk adapters and are located in separate drawers or units, if possible.

   Note: These steps do not apply to IBM 7135-110 and 7135-210 RAIDiant Disk Arrays, which provide their own mirroring of logical volumes. Continue with Testing a File System on page 9-4.

2. Verify the number of logical volume copies. Enter:

   `lsvg -l volume_group_name`

   In the resulting display, locate the line for the logical volume for which you just added copies. Notice that the number in the physical partitions column is *x* times the number in the logical partitions column, where *x* is the number of copies.

3. To verify the placement of logical volume copies, enter:

   `lspv -l hdiskx`

   where *hdiskx* is the name of each disk to which you assigned copies. That is, you enter this command for each disk. In the resulting display, locate the line for the logical volume for which you just added copies. For copies placed on separate disks, the numbers in the logical partitions column and the physical partitions column should be equal. Otherwise, the copies were placed on the same disk and the mirrored copies will not protect against disk failure.

Testing a File System

To run a consistency check on each file system's information:

1. Enter:

   `fsck /filesystem_name`

2. Verify that you can mount the file system by entering:

   `mount /filesystem_name`

3. Verify that you can unmount the file system by entering:

   `umount /filesystem_name`

Varying Off a Volume Group on the Source Node

After completing the previous tasks, use the **varyoffvg** command to deactivate the shared volume group. You vary off the volume group so that it can be properly imported onto the destination node and activated as appropriate by the cluster event scripts. Enter the following command:

`varyoffvg volume_group_name`

Importing a Volume Group onto Destination Nodes

Importing the volume group onto the destination node synchronizes the ODM definition of the volume group on each node on which it is imported.

Use the **smit importvg** fastpath to import the volume group.

VOLUME GROUP name	Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the source node.
PHYSICAL VOLUME name	Enter the name of a physical volume that resides in the volume group. Note that a disk may have a different physical name on different nodes. Make sure that you use the disk name as it is defined on the destination node.
ACTIVATE volume group after it is imported?	Set the field to **yes**.
Volume Group MAJOR NUMBER	Make sure to use the same major number on all nodes. Use the **lvlstmajor** command on each node to determine a free major number common to all nodes.

Changing a Volume Group's Startup Status

By default, a volume group that has just been imported is configured to automatically become active at system restart. In an HANFS for AIX environment, a volume group should be varied on as appropriate by the cluster event scripts. Therefore, after importing a volume group, use the Change a Volume Group SMIT screen to reconfigure the volume group so that it is not activated automatically at system restart.

Use the **smit chvg** fastpath to change the characteristics of a volume group.

Activate volume group Automatically at system restart?	Set this field to **no**.
A QUORUM of disks required to keep the volume group online?	This field is site dependent. See Chapter 5, Planning Shared LVM Components, for a discussion of quorum in an HANFS for AIX cluster.

Varying Off a Volume Group on Destination Nodes

Use the **varyoffvg** command to deactivate the shared volume group so that it can be imported onto another destination node or activated as appropriate by the cluster event scripts. Enter:

```
varyoffvg volume_group_name
```

Chapter 10 Performing Additional AIX Tasks

This chapter discusses several general tasks necessary to ensure that your HANFS for AIX environment works as planned.

AIX Administrative Tasks

Consider the following issues to ensure that AIX works as expected in an HANFS for AIX cluster:

- I/O pacing
- User and group IDs
- Network option settings
- **/etc/hosts** file and nameserver edits
- **cron** and NIS settings
- **/.rhosts** file edits
- **/etc/rc.net** file edits on NFS clients
- Managing applications using the SPX/IPX protocol.

I/O Pacing

By default, AIX is installed with high- and low-water marks set to **zero**, which disables I/O pacing. While enabling I/O pacing may have a slight performance effect on very I/O intensive processes, it is *required* for an HANFS for AIX cluster to behave correctly during large disk writes. If you anticipate heavy I/O on your HANFS for AIX cluster, you should enable I/O pacing.

Use the **smit chgsys** fastpath to set high- and low-water marks on the Change/Show Characteristics of Operating System SMIT screen. Although the most efficient high- and low-water marks vary from system to system, an initial high-water mark of 33 and low-water mark of 24 provide good starting points. These settings only slightly reduce write times and consistently generate correct fallover behavior from the HANFS for AIX software.

See the *AIX Performance Monitoring & Tuning Guide* for more information on I/O pacing.

Checking User and Group IDs

If a node fails, users should be able to log on to the surviving node without experiencing problems caused by mismatches in the user or group IDs. To avoid mismatches, make sure that user and group information is propagated to nodes as necessary. User and group IDs should be the same on both nodes.

Checking Network Option Settings

The default settings for the following network options should be changed as described in this section:

- thewall
- routerevalidate

Changing thewall Network Option

To ensure that HANFS for AIX requests for memory are handled correctly, you can set (on every cluster node) “thewall” network option to be higher than its default value. The suggested value for his option is shown below:

```
thewall = 5120
```

1. To change this default value, add the following line to the end of the **/etc/rc.net** file:

   ```
   no –o thewall=5120
   ```

2. After making this change, monitor mbuf usage using the **netstat -m** command and increase or decrease “thewall” option as needed.

3. To list the values of other network options that are currently set on a node, enter:

   ```
   no –a
   ```

Changing routerevalidate Network Option

Changing hardware and IP addresses within HANFS changes and deletes routes. Due to the fact that AIX caches routes, it is required that you set the “routerevalidate” network option as follows:

```
routerevalidate=1
```

This setting ensures the maintenance of communication between cluster nodes.

- To change the default value, add the following line to the end of the **/etc/rc.net** file:

  ```
  no –o routerevalidate=1
  ```

Editing the /etc/hosts File and nameserver Configuration

Make sure both nodes can resolve all cluster addresses. Edit the **/etc/hosts** file (and the **/etc/resolv.conf** file, if using the **nameserver** configuration) on each node in the cluster to make sure the IP addresses of all clustered interfaces are listed. Make sure that all service, standby, and boot addresses are listed.

For each boot address, make an entry similar to the following:

```
100.100.50.200 crab_boot
```

Also, make sure that the **/etc/hosts** file on each node has the following entry:

```
127.0.0.1 loopback localhost
```

cron and NIS Considerations

If your HANFS cluster nodes use NIS services that include the mapping of the **/etc/passwd** file and if IPAT is enabled, users that are known only in the NIS-managed version of the **/etc/passwd** file will not be able to create crontabs. This is because **cron** is started via the **/etc/inittab** file with run level 2 (for example, when the system is booted), but **ypbind** is started in the course of starting HANFS via the **rcnfs** entry in **/etc/inittab**. When IPAT is enabled in HANFS, the run level of the **rcnfs** entry is changed to -**a** and run via the **telinit -a** command by HANFS.

In order to let those NIS-managed users create crontabs, you can do one of the following:

1. Change the run level of the **cron** entry in **/etc/inittab** to -**a** and make sure it is positioned after the **rcnfs** entry in **/etc/inittab**. This solution is recommended if it is acceptable to start **cron** after HANFS has started.
2. Add an entry to the **/etc/inittab** file like the following script with runlevel -**a.** Make sure it is positioned after the **rcnfs** entry in **/etc/inittab**. The important thing is to kill the **cron** process, which will respawn and know about all of the NIS-managed users. Whether or not you log the fact that **cron** has been refreshed is optional.

```
#! /bin/sh
 # This script checks for a ypbind and a cron process. If both exist and cron was
 # started before ypbind, cron is killed so it will respawn and know about any new
 # users that are found in the passwd file managed as an NIS map
echo "Entering $0 at `date`" >> /tmp/refr_cron.out
 cronPid=`ps -ef |grep "/etc/cron" |grep -v grep |awk \
 '{ print $2 }'`
 ypbindPid=`ps -ef | grep "/usr/etc/ypbind" | grep -v grep | \
 if [ ! -z "${ypbindPid}" ]
 then
     if [ ! -z "${cronPid}" ]
     then
         echo "ypbind pid is ${ypbindPid}" >> /tmp/refr_cron.out
         echo "cron pid is ${cronPid}" >> /tmp/refr_cron.out
         echo "Killing cron(pid ${cronPid}) to refresh user \
               list" >> /tmp/refr_cron.out
         kill -9 ${cronPid}
         if [ $? -ne 0 ]
         then
             echo "$PROGNAME: Unable to refresh cron." \
             >>/tmp/refr_cron.out
             exit 1
        fi
     fi
              fi
              echo "Exiting $0 at `date`" >> /tmp/refr_cron.out
              exit 0
```

Editing the /.rhosts File

Make sure that each node's service adapters and boot addresses are listed in the **/.rhosts** file on both nodes in the cluster. Doing so allows the **/usr/sbin/cluster/utilities/clruncmd** command and the **/usr/sbin/cluster/godm** daemon to run. (The **/usr/sbin/cluster/godm** daemon is used when nodes are configured from a central location.)

For security reasons, IP label entries you add to the **/.rhosts** file to identify cluster nodes should be deleted when you no longer need to log on to a remote node from these nodes. The cluster synchronization and verification functions use **rcmd** and **rsh** and thus require these **/.rhosts** entries. These entries are also required to use C-SPOC commands in a cluster environment. The **/usr/sbin/cluster/clstrmgr** daemon, however, does not depend on **/.rhosts** file entries.

Editing the /etc/rc.net File on NFS Clients

By default, AIX 4.3.2 NFS clients do not respond to a ping on the broadcast address. Therefore, on any clients that use NFS, add the following line to the **/etc/rc.net** file:

```
/usr/sbin/no -o bcastping=1
```

Managing Applications That Use the SPX/IPX Protocol

If running applications or hardware that use the SPX/IPX protocol stack to establish a STREAMS connection through an adapter to another node, be aware that the SPX/IPX stack that is loaded to support applications like netware (or hardware, like the 7318 model P10 terminal server) is not unloaded by the **ifconfig** detach command, as is the IP stack, during a takeover or when attempting a hardware address swap. Thus, the **rmdev** command used in **cl_swap_HW_address** fails with a "device busy" error prior to changing the adapter hardware address.

To avoid this error, use the following command to unload the SPX/IPX stack:

```
strload -uf /etc/dlpi.conf
```

This command unloads the dlpi device driver.

Likewise, when using 7318 P10-style ports, you may want to unload other drivers using commands similar to the following:

```
strload -uf /etc/xtiso.conf
strload -uf /etc/netware.conf
```

These commands must be issued after the SPX/IPX stack is loaded and before HANFS runs **cl_swap_HW_address**. For example, run these commands before the **cl_swap_HW_address** script is called by **release_service_address** or **acquire_takeover_address**.

You can run these commands in pre- or post-event scripts that run prior to **cl_swap_HW_address**, or you can add them to the **stop_server** script that stops the application using the protocol. The specific scripts you use depend on when SPX/IPX-related drivers are loaded at your site.

Chapter 11 Installing HANFS for AIX Software

This chapter describes how to install the HANFS for AIX software.

Prerequisites

If you had previously used the HA-NFS software, be sure to remove that software before installing HANFS for AIX. SMIT will not work properly if both HA-NFS and HANFS for AIX are installed on a node.

The HANFS for AIX software has the following prerequisites:

- Each cluster node must have AIX installed.
- The following AIX optional bos components are mandatory for HANFS for AIX:
  - bos.net.nfs.client
  - bos.net.nfs.server
  - bos.net.tcp.client
  - bos.net.tcp.server
  - bos.sysmgt.trace
- Each cluster node requires its own HANFS for AIX software license.
- The installation process must be performed by the root user.
- The **/usr** directory must have 40 MB of free space to install the base version of HANFS.
  - Add an additional 1.6 MB for the Japanese message catalogs, if you require them. (You should choose to install only the message catalogs for the language you will be using rather than all message catalogs.)
  - If you are planning to install HAView, you will need an additional 14 MB, plus 28 KB for the Japanese message catalogs.
- HAView requires that you install the following NetView filesets:
  - nv6000.base.obj 4.1.2.0
  - nv6000.database.obj 4.1.2.0
  - nv6000.Features.obj 4.1.2.0
  - nv6000.client.obj 4.1.2.0

  The first three—base, database, and features—must be installed on the NetView server. The client fileset is installed on the NetView client.

  You install the HAView server image, or both server and client images, on the NetView server. Install only the HAView client image on the NetView client.
- The **/** (**root**) directory must have 500 KB of free space (beyond any need to extend the **/usr** directory).

If you are installing HANFS for AIX on a RS/6000 SP node, refer to Appendix C, Installing and Configuring HANFS for AIX on RS/6000 SPs.

Installing the HANFS for AIX Software

Note: The HANFS for AIX software and the HACMP for AIX software cannot be installed on the same machine. If HACMP for AIX is already installed, HANFS for AIX installation fails with a message similar to the following:

```
HACMP for AIX is installed on the system.
You must remove HACMP for AIX prior to installation of HANFS.
Aborting software installation.
```

HAView Installation Notes

HAView requires TME 10 NetView for AIX. Install NetView before installing HAView.

The HAView fileset includes a server image and a client image. If NetView is installed using a client/server configuration, the HAView server image should be installed on the NetView server, and the client image on the NetView client. Otherwise, you can install both the HAView client and server images on the NetView server.

Note: It is recommended that you install the HAView components on nodes outside the cluster, even though you can install it on any node that has a NetView server installed. If HAView is configured on a cluster node, and that node fails, you will lose your view of the cluster configuration, and all HAView monitoring functions will be lost until another HAView node can be brought online.

Installing HAView outside the cluster minimizes the probability of losing monitoring capabilities during a cluster node failure.

For more information on using HAView to monitor your cluster, see Chapter 15, Maintaining an HANFS for AIX Environment.

Installation Media

The HANFS for AIX software is distributed on installation media. HANFS for AIX files are installed in the **/usr/sbin/cluster** and **/usr/lpp/cluster** directories.

Installing HANFS for AIX

To install the HANFS software, complete the following steps on *both* nodes:

Note: You can use this procedure to install HANFS for the first time or to upgrade from an earlier version.

1. Insert the installation medium and enter:

   ```
   smit install_selectable_all
   ```

2. Enter the device name of the installation medium or Install Directory in the **INPUT device / directory for software** field and press Enter.

If you are unsure about the input device name or about the Install Directory, press F4 to list available devices. Then select the proper drive or directory and press Enter. The correct value is entered into the **INPUT device/directory** field as the valid input device.

3. Press Enter to display the next screen.
4. Enter field values as follows:

SOFTWARE to install	Change the value to **cluster.hanfs**.
PREVIEW ONLY?	Change the value to **no**.
OVERWRITE same or newer versions?	Leave this field set to **no**. Set it to **yes** if you are reinstalling or reverting to Version 4.3.1 from a newer version of the HANFS for AIX software.
AUTOMATICALLY install requisite software	Set this field to **no** if the prerequisite software for Version 4.3.1 is installed or if the **OVERWRITE same or newer versions?** field is set to **yes**; otherwise, set this field to **yes** to install required server and client software.

5. Enter values for the other fields as appropriate for your site.
6. When you are satisfied with the entries, press Enter. SMIT responds:

   ```
   ARE YOU SURE?
   ```

7. Press Enter again.

   SMIT instructs you to read the HANFS 4.3.1 **release_notes** file in the **/usr/lpp/cluster/doc** directory for further information.

8. Are you upgrading from HANFS version 4.2.2 with this installation?

   If *no*, go to step 9.

   If *yes*, you need to uninstall the **cluster.cspoc** filesets, as the C-SPOC facility is no longer supported in HANFS for AIX. To uninstall the filesets, type:

   ```
   smit remove
   ```

   and enter the following field values:

SOFTWARE NAME	Enter **cluster.cspoc***. Be sure to include the asterisk (*) so that all the C-SPOC filesets are removed.
Preview?	Change the value to **no**.

9. Reboot the node.

Changes to AIX /etc/services File

The HANFS for AIX installation script adds the following information to the AIX **/etc/services** file:

clm_keepalive	6000/udp
cllockd	6100/udp
clm_pts	6255/tcp

clsmuxpd	6270/tcp
clm_lkm	6150/tcp
clm_smux	6175/tcp
clinfo_deadman	6176/tcp
godm	6177/tcp

Changes to AIX /etc/rc.net File

The **/etc/rc.net** file is called by **cfgmgr** to configure and start TCP/IP during the boot process. It sets hostname, default gateway and static routes. The following entry is added at the beginning of the file for a node on which IP address takeover is enabled:

```
# HACMP for AIX
# HACMP for AIX These lines added by HACMP for AIX software
[ "$1" = "-boot" ] && shift || { ifconfig lo0 127.0.0.1 up; exit 0; } #HACMP for AIX
# HACMP for AIX
```

The entry prevents **cfgmgr** from reconfiguring boot and service addresses while HANFS for AIX is running.

Problems During the Installation

If you encounter problems during the installation, the installation program automatically performs a cleanup process. If, for some reason, the cleanup is not performed after an unsuccessful installation:

1. Enter the **smit install** software to display the Installation and Maintenance menu.
2. Select **Install and Update Software > Clean Up After a Interrupted Installation**.
3. Review the SMIT output (or examine the **/smit.log** file) for the interruption's cause.
4. Fix any problems and repeat the installation process.

Verifying Cluster Software

This section describes how to verify that the software installed on a cluster node is compatible with the installed version of the HANFS for AIX software.

Using the /usr/sbin/cluster/diag/clverify Utility

Use the **/usr/sbin/cluster/diag/clverify** utility on each node in the cluster to verify that the software installed on the system is compatible with the installed version of the HANFS for AIX software. This utility tells you if the LPP installation itself is incorrect.

You should run this utility before starting up HANFS for AIX. You can use the **/usr/sbin/cluster/diag/clverify** utility in either of two modes: interactive or command line. Using the interactive mode, you step through the list of valid options until you get to the specific

program you want to run. The interactive mode also includes a help facility. If you know the complete syntax the utility needs for verifying a given feature, you can enter the command and its required options at the system prompt to run the program immediately.

The following sections describe how to use the utility in interactive mode.

Verifying Cluster Software

To ensure that an HANFS for AIX cluster works properly, you must verify that the correct HANFS-for-AIX-specific modifications to AIX system files exist. The cluster software verification procedure automates this task for you.

Note: You should run the software verification procedure before verifying the cluster configuration.

To run the software verification procedure interactively:

1. Type:

   ```
   clverify
   ```

   The command returns a list of command options and the `clverify` prompt.

   ```
   ------------------------------------------------------------
   To get help on a specific option, type: help <option>
   To return to previous menu, type: back
   To quit the program, type: quit
   ------------------------------------------------------------
   Valid Options are:
   software
   cluster

   clverify>
   software
   ```

   The following options and an updated prompt appear:

   ```
   Valid Options are:
   lpp

   clverify.software>
   ```

2. Verify that the HANFS for AIX files are installed correctly, and log the results in a file called **verify_hacmp** by using the **-R** option with **lpp**:

   ```
   lpp -R verify_hacmp
   ```

   When the program completes, read the **verify_hacmp** file. If no problems exist, no messages are logged. If the HANFS for AIX software is not installed correctly, you receive a message similar to the following that informs you of errors:

   ```
   inetd is not configured for HACMP.
   The /.rhost file does not exist.
   ```

   Note: If you receive messages about configuration errors but have not yet configured the cluster, ignore them. Also note that the **clverify** utility does not detect netmask inconsistencies.

3. Type CTRL-C or **quit** to return to the system prompt.

Chapter 12 Configuring an HANFS for AIX Cluster

This chapter describes how to configure an HANFS for AIX cluster.

Note: The default locations of HANFS log files are used in this chapter. If you redirected any of these logs, check the appropriate location.

Overview

Complete the following procedures to define the cluster configuration.

- Define the cluster topology
- Configure the **/usr/sbin/cluster/etc/exports** file
- Configure resources
- Customize cluster log files
- Verify the cluster environment.

Each procedure is described in the following sections.

Making the Cluster Configuration Active

The cluster configuration becomes active the next time the Cluster Manager starts. At startup, the Cluster Manager initializes the environment with the values defined in the HANFS for AIX ODM. See Chapter 15, Maintaining an HANFS for AIX Environment, for more information on starting and stopping cluster services.

Defining the Cluster Topology

Complete the following procedures to define the cluster topology. You only need to perform these steps on one node. When you synchronize the cluster topology, its definition is copied to the other node.

1. Give the cluster an ID and a name on the Add a Cluster Definition SMIT screen.
2. Define the nodes on the Add Cluster Nodes screen.
3. Define the network adapters on the Add an Adapter screen.
4. *(optional)* View or change the network modules on the Configure Network Modules screen. Changes or additions are unlikely. When configuring a cluster, you do not need to enter any information in the Network Module screens. Network modules are pre-loaded when you install the HANFS for AIX software.

   Note: The SP Switch network module can support multiple clusters; therefore, its settings should remain at their default values to avoid affecting HANFS for AIX event scripts.

5. Copy the HANFS for AIX ODM entries to each node in the cluster, using the **Synchronize Cluster Topology** option on the Configure Cluster menu.

Each procedure is described below.

Defining the Cluster ID and Name

The cluster ID and name uniquely identify each cluster in an HANFS for AIX cluster environment. If more than one cluster on the same network has the same cluster ID, the Cluster Manager fails and generates an error message. Thus, make sure that all clusters on the same network have unique cluster IDs.

To define the cluster ID and name, refer to your completed network planning worksheets for the values. Then do the following:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **Cluster Configuration > Cluster Topology > Configure Cluster > Add a Cluster Definition**.
3. Enter field values as follows:

Cluster ID	Enter a positive integer unique to your site (in the range 1 to 999999).
Cluster Name	Enter an ASCII text string that identifies the cluster. The cluster name can include alphabetic and numeric characters and underscores. Use no more than 31 characters.

4. Press Enter.

   The HANFS for AIX software uses this information to create the cluster entries for the ODM.
5. Press F3 until you return to the Cluster Topology screen, or press F10 to exit SMIT.

Defining Nodes

After defining the cluster name and ID, define the cluster nodes.

To define the cluster nodes:

1. From the Cluster Topology menu, select **Configure Nodes > Add Cluster Nodes**.
2. Enter the name for each cluster node in the **Node Names** field.

   Names cannot exceed 31 characters. They can include alphabetic and numeric characters and underscores. Leave a space between names. The node name can be different from the hostname.If you duplicate a name, you will get an error message. The information is effective when you synchronize both nodes.
3. Press F3 until you return to the Cluster Topology screen, or press F10 to exit SMIT.

Defining Adapters

To define the adapters, first consult your planning worksheets for both TCP/IP and serial networks, and then complete the following steps:

1. From the Cluster Topology menu, select **Configure Adapters > Add an Adapter**.

2. Enter field values as follows:

Adapter Label Enter the IP label (the name) of the adapter you have chosen as the service address for this adapter. Adapter labels can contain up to 31 alphanumeric characters, and can include underscores and hyphens.

Each adapter that can have its IP address taken over must have a boot adapter (address) label defined for it. Use a consistent naming convention for boot adapter labels. (You will choose the Add an Adapter option again to define the boot adapter when you finish defining the service adapter.)

Network Type Indicate the type of network to which this adapter is connected. Pre-installed network modules are listed on the pop-up pick list.

See the section below, Network Modules Supported on page 12-4 for information on possible network types.

Network Name Enter the network name connected to this adapter.

The network name is arbitrary, but must be used consistently. If several adapters share the same physical network, make sure you use the same network name for each of these adapters.

Network Attribute Indicate whether the network is **public**, **private**, or **serial**. Press TAB to toggle the values.

Ethernet, Token-Ring, FDDI, and SLIP are public networks. SOCC and HPS and ATM are private networks. RS232 lines, target mode SCSI-2 busses, and TMSSA links are serial networks.

Adapter Function Indicate whether the adapter's function is **service**, **standby**, or **boot**. Press TAB to toggle the values.

A node has a single service adapter for each public or private network. A node has a only a single service adapter for a serial network. A node can have one or more standby adapters for each public network. Serial and HPS networks do not have standby adapters. ATM networks (which are always private) and Token-Ring, FDDI, and Ethernet networks (which may be defined as private) allow standby adapters.

Note: In an HANFS for AIX HPS network on the SP, integrated Ethernet adapters cannot be used, and standby adapters are not used. If a takeover occurs, the service address is aliased onto another node's service address. See Appendix C, Installing and Configuring HANFS for AIX on RS/6000 SPs for complete information on adapter functions in an HPS environment.

Adapter Identifier Enter the IP address in dotted decimal format or a device file name.

IP address information is required for non-serial network adapters only if the node's address cannot be obtained from the domain name server or the local **/etc/hosts** file (using the adapter IP label given).

You must enter device file names for serial network adapters. RS232 serial adapters must have the device file name **/dev/tty*n***. Target mode SCSI serial adapters must have the device file name **/dev/tmscsi*n***.Target mode SSA serial adapters must have the device file name **/dev/tmssa*n***.

Adapter Hardware Address *This field is optional.* Enter a hardware address for the adapter. The hardware address must be unique within the physical network. Enter a value in this field only if you are currently defining a service adapter, the adapter has a boot address, and you want to use hardware address swapping. The hardware address is 12 digits for Ethernet, Token-Ring and FDDI; and 14 digits for ATM.

Node Name Define a node name for all adapters except for those whose addresses may be shared by nodes participating in the resource chain for a rotating resource configuration. These adapters are rotating resources. The event scripts use the user-defined configuration to associate these service addresses with the proper node. In all other cases, addresses are associated with a particular node (service, boot, and standby).

3. Press Enter. The system adds these values to the HANFS for AIX ODM and displays the Configure Adapters menu.
4. Define all the adapters, then press F3 until you return to the Cluster Topology screen, or press F10 to exit SMIT.

Note: Although it is possible to have only one physical network adapter (no standby adapters), this constitutes a potential single point of failure condition and is not recommended for an HANFS for AIX configuration. The instructions listed here assume you have at least one standby adapter for each public network.

Network Modules Supported

Each supported cluster network in a configured HACMP cluster has a corresponding cluster network module. Each network module monitors all I/O to its cluster network.

Note: The Network Modules are pre-loaded when you install the HANFS for AIX software.

Each network module maintains a connection to other network modules in the cluster. The Cluster Managers on cluster nodes send messages to each other through these connections. Each network module is responsible for maintaining a working set of service adapters and for verifying connectivity to cluster peers. The network module also is responsible for reporting when a given link actually fails. It does this by sending and receiving periodic heartbeat messages to or from other network modules in the cluster.

Currently, network modules support communication over the following types of networks:

- Serial (RS232)
- Target-mode SCSI
- Target-mode SSA
- IP
- Ethernet
- Token-Ring
- FDDI
- SOCC
- SLIP
- SP Switch
- ATM.

Synchronizing the Cluster Definition Across Nodes

The HANFS for AIX ODM entries must be the same on both nodes. If the definitions are not synchronized across nodes, the HANFS for AIX software generates a run-time error at cluster startup. Also, before synchronizing a cluster definition across nodes, both nodes must be powered on, the HANFS for AIX software must be installed, and the **/etc/hosts** and **/.rhosts** files must include all HANFS for AIX IP labels.

Note: The **/.rhosts** file is not required on SP systems running HANFS Enhanced Security. This feature removes the requirement of TCP/IP access control lists (that is, the **/.rhosts** file) on remote nodes during HANFS configuration.

Complete the following steps to synchronize a cluster definition across nodes:

1. Enter the **smit hacmp** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **Cluster Configuration > Cluster Topology > Synchronize Cluster Topology** to copy the cluster definition as defined on the local node to the remote node.

   SMIT responds:

   ```
   ARE YOU SURE?
   ```

3. Press Enter.

   The cluster definition (including all node, adapter, and network module information) is copied to the other node.

4. Press F10 to exit SMIT.

Before you synchronize a cluster, it is always a good idea to run the cluster verification utility, **clverify**, to make sure that cluster is configured as you intended. For more information about **clverify**, see Verifying the Cluster Environment on page 12-11.

Note: The **clverify** utility checks that each log file has the same pathname on every node in the cluster and reports an error if this is not the case.

Configuring the exports File

In this step you add the directories of the shared file systems to the **exports** file.

Note for former HA-NFS Version 3 users: You can copy your **/etc/exports.hanfs** file to **/usr/sbin/cluster/etc/exports** and continue with Creating Resource Groups on page 12-6.

Complete the following steps for each directory you want to add to the **exports** file. Refer to your NFS File Systems Worksheet.

1. Enter the **smit mknfsexp** fastpath to display the Add a Directory to Exports List screen.
2. In the **EXPORT directory now, system restart or both** field, enter **restart**.
3. In the **PATHNAME of alternate Exports file** field, enter **/usr/sbin/cluster/etc/exports**.
4. Add values for the other fields as appropriate for your site, and press Enter.
   HANFS for AIX uses this information to update the **/usr/sbin/cluster/etc/exports** file.
5. Press F3 to return to the Add a Directory to Exports List screen, or F10 to exit SMIT.

Repeat steps 1 through 4 for each directory listed in the **File Systems/Directories to Export** field on your planning worksheets.

Configuring Resources

You now configure resources (file systems) and set up the node environment. This involves:

- Configuring resource groups and node relationships to behave as desired
- Adding individual resources to each resource group
- Setting up run-time parameters per node
- Synchronizing the node environment.

Each step is explained in the following sections.

Creating Resource Groups

Notes for former HA-NFS Version 3 users: If you have an HA-NFS Version server-backup configuration (where one node is the “active” server and the other node only acts as a server if the active server fails), you can create one resource group and put all the file systems that you want to be able to NFS-mount into that resource group.

If you have an HA-NFS Version 3 server-server configuration (where both nodes are “active” servers, each being the backup for the other in case of failure), you should create two resource groups. Put the file systems to be exported by one node in one resource group, and the file systems to be exported by the second node in the other resource group.

To create resource groups:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **Cluster Configuration > Cluster Resources > Define Resource Groups > Add a Resource Group**.

3. Enter the field values as follows:

Resource Group Name	Enter the desired name. Use no more than 31 characters. You can use alpha or numeric characters and underscores. Duplicate entries are not allowed.
Node Relationship	Toggle the entry field between **Cascading** and **Rotating**.
Participating Node Names	Enter the names of the nodes that can own or take over this resource group. Enter the node with the higher priority first, followed by the node with the lower priority. Leave a space between node names.

Notes for former HA-NFS Version 3 users: If you have a server-backup configuration, you only need one resource group. List both nodes, with the "active" server first.

If you have a server-server configuration, you need two resource groups. List both nodes, with each node first in the resource group which will define file systems mounted locally on that node.

4. Press Enter to add the resource group information to the HANFS for AIX ODM.
5. Press F3 after the command completes until you return to the Cluster Resources screen, or press F10 to exit SMIT.

Configuring (Assigning) Resources for Resource Groups

Once you have defined a resource group, you assign resources to it. You can configure a resource group even if a node is powered down; however, SMIT cannot list possible shared resources for the node (making configuration errors likely).

Note: You cannot configure a resource group until you have completed the information on the Add a Resource Group screen.

To assign resources to a resource group:

1. On the Define Resource Groups menu, select **Change/Show a Resource Group** and press Enter to display a list of defined resource groups.

   Note: If you configure a cascading resource group with an NFS mount point, you must also configure the resource to use IP Address Takeover. If you do not do this, takeover results are unpredictable. You should also set the field value **File Systems Mounted Before IP Configured** to **true** so that the takeover process proceeds correctly.

2. Select the resource group you want to configure and press Enter. SMIT displays a screen with the **Resource Group Name**, **Node Relationship**, and **Participating Node Names** fields filled in.

   If the participating nodes are powered on, you can press F4 to list the shared resources. If a resource group/node relationship has not been defined, or if a node is not powered on, F4 displays the appropriate warnings.

3. Enter the field values as follows:

Resource Group Name	Reflects the choice you made on the previous screen; this field is not editable in this screen.
Node Relationship	Reflects the fallover strategy entered when you created the resource group; this field is not editable in this screen.
Participating Node Names	Reflects the names of the nodes that you entered as members of the resource chain for this resource group. Node names are listed in order from highest to lowest priority (left to right), as you designated them. This field is not editable in this screen.
Service IP Label	List the IP label to be taken over when this resource group is taken over. Press F4 to see a list of valid IP labels. These include addresses which rotate or may be taken over.
HTY Service Label	NTX adapters are not supported for HANFS for AIX 4.3.1.
File Systems	Identify the file systems to include in this resource group. Press F4 to see a list of the file systems. **Note for former HA-NFS Version 3 users:** Be sure to list the file systems for all the directories listed in your **/etc/exports.hanfs** file.
File Systems Consistency Check	Identify the method of checking consistency of file systems, **fsck** (default) or **logredo** (for fast recovery).
File Systems Recovery Method	Identify the recovery method for the file systems, **parallel** (for fast recovery) or **sequential** (default). Do *not* set this field to **parallel** if you have shared, nested file systems. These must be recovered sequentially. (Note that the cluster verification utility, **clverify**, does not report file system and fast recovery inconsistencies.)
File Systems to Export	Identify the file systems to be exported to include in this resource group. These should be a subset of the file systems listed above. Press F4 for a list. **Note for former HA-NFS Version 3 users:** Be sure to list the file systems for all the directories listed in your **/etc/exports.hanfs** file.
File Systems to NFS-Mount	Identify the subset of file systems to NFS-mount. All nodes in the resource chain that do not currently hold the resource will attempt to NFS-mount these file systems while the owner node is active in the cluster.

Inactive Takeover Activated

Set this variable to control the *initial acquisition* of a resource group by a node when the node/resource relationship is cascading. This variable does not apply to rotating resource groups.

If you specify that Inactive Takeover is **true**, the first node in the resource group to join the cluster acquires the resource group, regardless of the node's designated priority. Subsequently the resource group cascades to the node in the chain with higher priority as it joins the cluster.

If you specify that Inactive Takeover is **false**, the first node to join the cluster acquires only those resource groups for which it has been designated the highest priority node. Each subsequent node that joins the cluster acquires all resource groups for which the node has a higher priority than any other node currently up in the cluster. Depending on the order in which the nodes are brought up, this may result in resource groups cascading to higher priority nodes as those nodes join the cluster.

The default is **false**.

4. Press Enter to add the values to the HANFS for AIX ODM.
5. Press F3 until you return to the Cluster Resources menu, or press F10 to exit SMIT.

Configuring Run-Time Parameters

To define the run-time parameters for a node:

1. On the Cluster Resources menu, select **Change/Show Run-Time Parameters** to list the node names. You define run-time parameters individually for each node.
2. Select a node name and press Enter. SMIT displays the **Change/Show Run-Time Parameters** screen with the node name displayed.
3. Enter field values as follows:

Debug Level

Cluster event scripts have two levels of logging. The **low** level logs errors encountered while the script executes. The **high** level logs all actions performed by the script. The default is **high**.

Host uses NIS or Name Server

If the cluster uses Network Information Services (NIS) or name serving, set this field to **true**. The HANFS for AIX software then disables these services before entering reconfiguration, and enables them after completing reconfiguration. The default is **false**.

Cluster Security Mode

To enable Kerberos authentication using the **/usr/kerberos/bin/rsh** command, set this field to **Enhanced**. HANFS Enhanced Security can be run only on SP systems. This field should be set to **Standard** if your cluster nodes are not on an SP.

4. Press Enter to add the values to the HANFS for AIX ODM.
5. Press F3 until you return to the Cluster Resources menu, or press F10 to exit SMIT.

Synchronizing the Node Environment

Synchronizing the node environment sends the information contained on the local node to the remote node.

Note: Both nodes must be on their boot addresses when a cluster has been configured and the nodes are synchronized for the first time. Any node not on its boot address will not have its **/etc/rc.net** file updated with the HANFS for AIX entry; this causes problems when reintegrating the node into the cluster.

If a node attempts to join a cluster with a node environment which is out of sync with the active node, it will be denied. You must synchronize the node environment to the joining member.

To synchronize the node environment:

1. Select **Synchronize Cluster Resources** from the Cluster Resources menu. SMIT prompts you to confirm that you want to synchronize cluster resources.
2. Press Enter to synchronize the resource group configuration and node environment across the cluster.
3. Press F3 until you return to the HANFS for AIX menu, or press F10 to exit SMIT.

Customizing Cluster Log Files

You can redirect a cluster log from its default directory to a directory of your choice (directory on a locally attached disk). Should you redirect a log file to a directory of your choice, keep in mind that the requisite (upper limit) disk space for most cluster logs is 2MB. 14MB is recommended for **hacmp.out**. To redirect a cluster log from its default directory to another destination, take the following steps:

1. Enter

   ```
   smitty hanfs
   ```

2. Select **Cluster System Management -> Cluster Log Management -> Change/Show Cluster Log Directory**

   SMIT displays a picklist of cluster log files with a short description of each.

Log Name	Description
cluster.mmdd	Cluster history files generated daily
cm.log	Generated by clstrmgr activity
dms_loads.out	Generated by deadman switch activity
hacmp.out	Generated by event scripts and utilities
cluster.log	Generated by cluster scripts and daemons

3. Select a log that you want to redirect.

SMIT displays a screen with the selected log's name, description, default pathname, and current directory pathname. The current directory pathname will be the same as the default pathname if you do not elect to change it.

The example below shows the **cluster.mmdd** log file screen. Edit the final field to change the default pathname.

Custom Log Name	**cluster.mmdd**
Cluster Log Description	Cluster history files generated daily
Default Log Destination Directory	**/usr/sbin/cluster/history**
Log Destination Directory	The default directory name appears here. To change the default, enter the desired directory pathname.

4. Press F3 to return to the screen to select another log to redirect, or return to the Cluster System Management screen to proceed to the screen for synchronizing cluster resources.
5. After you change a log directory, a prompt appears reminding you to synchronize cluster resources from this node (Cluster log ODMs must be identical across the cluster). The cluster log destination directories as stored on this node will be synchronized to all nodes in the cluster.

Log destination directory changes will take effect when you synchronize cluster resources, or if the cluster is not up, the next time cluster services are restarted.

Note: Logs should not be redirected to shared filesystems or NFS filesystems. Though this may be desirable in rare cases, such action may cause problems if the filesystem needs to unmount during a fallover event.

Verifying the Cluster Environment

This section describes how to verify the cluster environment, including the cluster and node configurations. This process ensures that both nodes agree on cluster topology and assignment of resources. Synchronization of log file pathnames is also mentioned here.

Verifying Cluster and Node Environment

After defining the cluster and node environment, run the cluster verification procedure on one node to check that both nodes agree on the assignment of HANFS for AIX cluster resources.

To verify the cluster and node configuration:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **Cluster Configuration > Cluster Verification > Verify Cluster**.

3. Fill in the fields as follows:

Base HACMP Verification Methods	By default, *both* the cluster topology and resources verification programs are run. You can toggle this entry field to run either program, or you can select **none** to specify a custom-defined verification method in the **Define Custom Verification Method** field.
Define Custom Verification Method	Enter the name of a custom-defined verification method. You can also press F4 for a list of previously defined verification methods. By default, if no methods are selected, the **clverify** utility checks the topology, resources, and all custom verification methods in alphabetical order. The order in which verification methods are listed determines the sequence in which selected methods are run. This sequence remains the same for subsequent verifications until different methods are selected.
Error Count	By default, the program runs to completion no matter how many errors it finds. If you want to cancel the program after a specific number of errors, enter the number in this field.
Log File to store output	Enter the name of an output file in which to store verification output. By default, verification output is stored the **smit.log** file.

4. Press Enter.

If you receive error messages, make the necessary corrections and run the verification procedure again.

The **clverify** utility also checks that each log file has the same pathname on every node in the cluster and reports an error if this is not the case.

Checking Cluster Topology

Run the following command to verify that all nodes agree on the cluster topology:

```
clverify cluster topology check
```

When the program finishes, check the output. If a problem exists with the cluster topology, a message similar to the following appears:

```
ERROR: Could not read local configuration
ERROR: Local Cluster ID XXX different from Remote Cluster ID XXX.
ERROR: Nodes have different numbers of networks
```

Synchronizing Cluster Topology

If both nodes do not agree on the cluster topology, and you are sure you want to define the cluster as it is defined on the local node, you can force agreement on cluster topology.

To synchronize a cluster definition across nodes:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.

2. On the HANFS for AIX menu, select **Cluster Configuration > Cluster Topology > Synchronize Cluster Topology** to copy the cluster definition as defined on the local node to the other cluster nodes.

   SMIT responds:

   ```
   ARE YOU SURE?
   ```

3. Press Enter.

   The cluster definition (including all node, adapter, and network method information) is copied to the other node.

4. Press F10 to exit SMIT.

See the **/usr/sbin/cluster/diag/clverify** man page for details on the cluster verification utility.

Chapter 13 Configuring Monitoring Scripts and Files

This chapter describes how to edit files and scripts used by Clinfo and HAView.

Editing the /usr/sbin/cluster/etc/clhosts File

For the Clinfo daemon (**clinfo**) and HAView to get the information they need, you must edit the **/usr/sbin/cluster/etc/clhosts** file on each cluster node. This file should contain hostnames (addresses) of any HANFS for AIX nodes with which **clinfo** or HAView can communicate, including servers from clusters accessible through logical connections.

As installed, the **/usr/sbin/cluster/etc/clhosts** file on an HANFS for AIX node contains a loopback address. The **clinfo** daemon or HAView first attempts to communicate with a **clsmuxpd** process locally. If it succeeds, **clinfo** or HAView then acquires an entire cluster map, including a list of all HANFS for AIX server interface addresses. From then on, **clinfo** or HAView uses this list rather than the provided loopback address to recover from a **clsmuxpd** communication failure.

If **clinfo** or HAView does not succeed in communicating with a **clsmuxpd** process locally, however, it only can continue trying to communicate with the local address. For this reason, you should replace the loopback address with all HANFS for AIX IP addresses or IP labels of all the service adapters and boot adapters that are accessible through logical connections to this node. The loopback address is provided only as a convenience.

An example **/usr/sbin/cluster/etc/clhosts** file follows:

```
cowrie_en0_cl83        # cowrie service
140.186.91.189         # limpet service
floyd_en0_cl83         # floyd service
squid_en0_cl83         # squid service
```

Warning: Do *not* include standby addresses in the **clhosts** file, and do *not* leave this file empty. If either of these two conditions exist, neither **clinfo** nor HAView nor the **/usr/sbin/cluster/clstat** utility will work properly.

Editing the /usr/sbin/cluster/etc/clinfo.rc Script

The **/usr/sbin/cluster/etc/clinfo.rc** script, executed whenever a cluster event occurs, updates the system's ARP cache. If you are not using the hardware address swapping facility, a copy of the **/usr/sbin/cluster/etc/clinfo.rc** script must be present on each cluster node for all ARP caches to be updated and synchronized, and you must run **clinfo**.

If you choose not to use hardware address takeover, edit the **/usr/sbin/cluster/etc/clinfo.rc** file on each server node by adding the IP label or IP address of each system that will be accessing shared file systems managed by HANFS to the **PING_CLIENT_LIST** list. Then run the **clinfo** daemon. Here's an example of how you might edit the **clinfo.rc** file:

```
PING_CLIENT_LIST="client1 client2 1.1.1.3"
```

When you start the cluster on each cluster node using the **smitty clstart** command, enter **true** in the **Startup Cluster Information Daemon?** field to start the **clinfo** daemon.

The HANFS for AIX software is distributed with a template version of the **/usr/sbin/cluster/etc/clinfo.rc** script. You can use the script as distributed, you can add new functionality to the script, or you can replace it with a custom script.

Note: If you are not using hardware address swapping, the ARP functionality must remain.

The format of the **clinfo** call to **clinfo.rc**:

```
clinfo.rc {join,fail,swap} interface_name
```

When **clinfo** gets a **cluster_stable** event, or when it connects to a new **clsmuxpd**, it receives a new map. **clinfo** then checks for changed states of interfaces.

- If a new state is UP, **clinfo** calls:

  ```
  clinfo.rc join interface_name
  ```
- If a new state is DOWN, **clinfo** calls:

  ```
  clinfo.rc fail interface_name
  ```
- If **clinfo** receives a **node_down_complete** event, it calls **clinfo.rc** with the fail parameter for each interface currently UP.
- If **clinfo** receives a **fail_network_complete** event, it calls **clinfo.rc** with the fail parameter for all associated interfaces.
- If **clinfo** receives a **swap_complete** event, it calls:

  ```
  clinfo.rc swap interface_name
  ```

Chapter 14 Supporting AIX Error Notification

This chapter describes how to use both the AIX Error Notification facility and the Automatic Error Notification utility to identify and respond to failures in an HANFS for AIX cluster. It also discusses using the Error Log Emulation functionality to test customized failure responses.

Using Error Notification in an HANFS for AIX Environment

You can optionally use the Error Notification facility to add an additional layer of high availability to the HANFS for AIX software. Take the example of a cluster where an owner node and a takeover node share a SCSI disk. The owner node is using the disk. If the SCSI adapter on the owner node fails, an error may be logged, but neither the HANFS for AIX software nor the AIX Logical Volume Manager responds to the error. If the error has been defined to the Error Notification facility, however, an executable that shuts down the node with the failed adapter could be run, allowing the surviving node to take over the disk.

Defining an Error Notification Object and Notify Method

To define an error notification object and its corresponding notify method:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support > Error Notification > Add a Notify Method**.

   The Add a Notify Method screen appears, on which you define the notification object and its associated notify method.
3. Enter values for the following fields:

Notification Object Name	Enter a user-defined name that identifies the error.
Persist across system restart?	Set this field to **yes** if you want this notification object to survive a system reboot. If not, set the field to **no**.
Process ID for use by Notify Method	Specify a process ID for the notify method to use. Objects that have a process ID specified should have the **Persist across system restart?** field set to **no**.
Select Error Class	Identify the class of error log entries to match. Valid values are: **None**—No error class **All**—All error classes **Hardware**—Hardware error class **Software**—Software error class **Errlogger**—Messages for the **errlogger** command

Select Error Type	Identify the severity of error log entries to match. Valid values are: **None**—No entry types to match **All**—Match all error types **PEND**—Impending loss of availability **PERM**—Permanent **PERF**—Unacceptable performance degradation **TEMP**—Temporary **UNKN**—Unknown
Match Alertable Errors?	Indicate whether the error is alertable. This descriptor is provided for use by alert agents associated with network management applications. Valid alert descriptor values are: **None**—No errors to match **All**—Match all alertable errors **TRUE**—Matches alertable errors **FALSE**—Matches non-alertable errors
Select Error Label	Enter the label associated with a particular error identifier as defined in the **/usr/include/sys/errids.h** file. If you are unsure about an error label, press F4 for a listing. Specify **All** to match all error labels.
Resource Name	Indicate the name of the failing resource. For the hardware error class, a resource name is a device name. For the software error class, the resource name is the name of the failing executable. Specify **All** to match all resource names.
Resource Class	Indicate the class of the failing resource. For the hardware error class, the resource class is the device class. The resource error class does not apply to software errors. Specify **All** to match all resource classes.
Resource Type	Enter the type of failing resource. For the hardware error class, a resource is the device type by which a resource is known in the devices object. Specify **All** to match all resource types.

Notify Method Enter the name of the executable that should run whenever an error matching the defined selection criteria occurs. The following keywords are automatically expanded by the error notification daemon as arguments to the notify method:

$1—Sequence number from the error log entry

$2—Error ID from the error log entry

$3—Error class from the error log entry

$4—Error type from the error log entry

$5—Alert flag values from the error log entry

$6—Resource name from the error log entry

$7—Resource type from the error log entry

$8—Resource class from the error log entry

4. Press Enter to create the notification object and method.
5. Press F10 to exit SMIT.

You now have defined a notification object and a corresponding notify method. The next time the error occurs, the system detects the error and responds as directed.

Examples

The following examples suggest how the Error Notification facility can be used in an HANFS for AIX cluster.

Example 1: Permanent Software Errors

In this example, the notification object is any permanent software error. The notify method is a mail message to the system administrator indicating that an error has occurred. The message includes the error ID and resource name.

Notification Object Name	SOFT_ERR.
Persist across system restart?	**yes**
Process ID for use by Notify Method	
Select Error Class	**Software**
Select Error Type	**PERM**
Match Alertable Errors?	
Select Error Label	.
Resource Name	
Resource Class	
Resource Type	

Notify Method	echo Permanent error occurred. Error ID + $2 Resource Name = $6 | mail sysadmin

Example 2: Permanent Hardware Errors

In this example, the notification object is a permanent hardware error on the SCSI adapter device driver. The notify method performs a **halt -q** to shut down the node.

Notification Object Name	SCSI_adapter
Persist across system restart?	**yes**
Process ID for use by Notify Method	
Select Error Class	**Hardware**
Select Error Type	**PERM**
Match Alertable Errors?	**None**
Select Error Label	**SCSI_ERR1**
Resource Name	**scsi0**
Resource Class	**adapter**
Resource Type	**adapter**
Notify Method	`halt -q`

Automatic Error Notification

Using HANFS 4.3.1 SMIT screen options, you can configure error notification automatically for the cluster resources listed below, list currently defined automatic error notify entries for the same cluster resources, or remove previously configured automatic error notify methods. Before you configure Automatic Error Notification, you must have a valid HACMP configuration.

Warning: Automatic error notification should be configured only when the cluster is not running.

Choosing to add error notify methods automatically runs the **cl_errnotify** utility which turns on error notification on the node for the following devices:

- All disks in the rootvg volume group
- All disks in HANFS volume groups and filesystems (To avoid single points of failure, the JFS log must be included in an HACMP volume group.)
- All disks defined as HANFS resources
- The SP switch adapter (for clusters with SP nodes).

Automatic error notification applies to selected hard, non-recoverable error types: disk, disk adapter, and SP switch adapter errors. No media errors, recovered errors, or temporary errors are supported by this utility.

Executing automatic error notification assigns one of two error notification methods for all the error types noted:

- **cl_failover** is assigned if a disk or an adapter (including an SP switch adapter) in a relevant device is determined to be a single point of failure and its failure should cause the cluster resources to fall over. In case of a failure of any of these devices, this method logs the error to **hacmp.out** and shuts down the cluster software on the node. It first tries to do a graceful shutdown with takeover; if this fails, it calls **cl_exit** to shut down the node.
- **cl_logerror** is assigned for all other error types. In case of a failure of any of these devices, this method logs the error to **hacmp.out**.

You can also use the utility to list all currently defined auto error notification entries in your HACMP cluster configuration and to delete all automatic error notify methods.

Configuring Automatic Error Notification

To configure automatic error notification, take the following steps:

1. Be sure that the cluster is not running.
2. Open the SMIT main HANFS menu by typing `smit hanfs`.
3. From the main menu, choose **RAS Support > Error Notification > Configure Automatic Error Notification.**
4. Select the **Add Error Notify Methods for Cluster Resources** option from the following list:

List Error Notify Methods for Cluster Resources	Lists all currently defined auto error notify entries for certain cluster resources: HACMP defined volume groups, filesystems, and disks; rootvg; SP switch adapter (if present). The list is output to the screen.
Add Error Notify Methods for Cluster Resources	Error notification methods are automatically configured on all relevant cluster nodes.
Delete Error Notify Methods for Cluster Resources	Error notification methods previously configured with the **Add Error Notify Methods for Cluster Resources** option are deleted on all relevant cluster nodes.

5. (optional) Since error notification is automatically configured for all the listed devices on all nodes, you must make any modifications to individual devices or nodes manually, after running this utility. To do so, choose the **Error Notification** option in the **RAS Support** SMIT screen. See the earlier section in this chapter.

   Note: If you make any changes to cluster topology or resource configuration, you may need to reconfigure automatic error notification. When you run **clverify** after making any change to the cluster configuration, you will be reminded to reconfigure error notification if necessary.

Listing Error Notify Methods

To see the automatic error notify methods that currently exist for your cluster configuration, take the following steps:

1. From the main HACMP menu, choose **RAS Support > Error Notification > Configure Automatic Error Notification.**
2. Select the **List Error Notify Methods for Cluster Resources** option. The utility lists all currently defined automatic error notification entries with these HACMP components: HACMP defined volume groups, concurrent volume groups, filesystems, and disks; rootvg; SP switch adapter (if present). The list is output to a screen similar to that shown below, in which the cluster nodes are named *sioux* and *quahog*:

```
                        COMMAND STATUS

Command: OK             stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

sioux:
sioux: HACMP Resource           Error Notify Method
sioux:
sioux: hdisk0           /usr/sbin/cluster/diag/cl_failover
sioux: hdisk1           /usr/sbin/cluster/diag/cl_failover
sioux: scsi0            /usr/sbin/cluster/diag/cl_failover
quahog:
quahog: HACMP Resource          Error Notify Method
quahog:
quahog: hdisk0          /usr/sbin/cluster/diag/cl_failover
quahog: scsi0           /usr/sbin/cluster/diag/cl_failover
```

Deleting Error Notify Methods

To delete automatic error notification entries previously assigned using this utility, take the following steps:

1. From the main menu, choose **RAS Support > Error Notification > Configure Automatic Error Notification.**
2. Select the **Delete Error Notify Methods for Cluster Resources** option. Error notification methods previously configured with the **Add Error Notify Methods for Cluster Resources** option are deleted on all relevant cluster nodes.

Error Log Emulation

After you have added one or more notify methods, you can test your methods by emulating an error. This shows you whether your pre-defined notify method carries out the intended action.

To emulate an error log entry:

1. From the main HANFS menu, choose **RAS Support > Error Notification > Emulate Error Log Entry**.

   The **Select Error Label** box appears, showing a picklist of the notification objects for which notify methods have been defined.

2. Select a notification object and press return to begin the emulation. As soon as you press the return key, the emulation process begins: the emulator inserts the specified error into the AIX error log, and the AIX error daemon runs the notification method for the specified object.

When the emulation is complete, you can view the error log by typing the **errpt** command to be sure the emulation took place. The error log entry has either the resource name EMULATOR, or a name as specified by the user in the **Resource Name** field during the process of creating an error notify object.

You will now be able to determine whether the specified notify method was carried out.

Note: Remember that the actual notify method will be run. Whatever message, action, or executable you defined will occur. Depending on what it is, you may need to take some action, for instance, to restore your cluster to its original state.

Only the root user is allowed to run an error log emulation.

Part 4 Maintaining HANFS for AIX

This part describes the ongoing tasks involved in maintaining an HANFS for AIX environment.

Chapter 15 Maintaining an HANFS for AIX Environment

This chapter explains how to maintain an HANFS for AIX environment by performing the following maintenance tasks:

- Starting and stopping cluster services
- Performing intentional fallover in a cluster
- Monitoring a cluster
- Maintaining exported file systems.

Note: The default locations of cluster log files are used in this chapter. If you redirected any logs, check the appropriate location.

Starting and Stopping Cluster Services on Nodes

Several maintenance tasks require that you first stop and later restart the HANFS for AIX software on one or more nodes. *Starting cluster services* refers to the sequence of steps that start HANFS for AIX and its associated processes on each node in the cluster. *Stopping cluster services* refers to the steps taken to halt the HANFS for AIX software on one or both nodes in a controlled situation. Starting and stopping cluster services requires a thorough understanding of node interaction, the available tools, and the impact on your system's availability.

The cluster services on the cluster nodes consist of the following HANFS for AIX daemons:

- The Cluster Manager (**clstrmgr**) daemon maintains the heartbeat protocol between the nodes in the cluster, monitors the status of the nodes and their interfaces, and invokes the appropriate scripts in response to node or network events. The **clstrmgr** daemon is required on cluster nodes.
- The Cluster Information Program (**clinfo**) daemon provides status information about the cluster to the nodes and invokes the **/usr/sbin/cluster/etc/clinfo.rc** script in response to a cluster event.
- The Cluster SMUX Peer (**clsmuxpd**) daemon maintains status information about cluster objects. This daemon works in conjunction with the Simple Network Management Protocol (**snmpd**) daemon.

The following sections describe how to start and stop cluster services using the SMIT interface. You can also use the System Resource Controller (SRC) commands **startsrc** and **stopsrc** to start and stop cluster services. All cluster daemons are part of the cluster group; you specify the **-g** flag with the cluster group identifier.

Scripts and Files Involved in Starting and Stopping HANFS for AIX

Various AIX and HANFS scripts and files are involved in starting and stopping cluster services. To fully understand this process, you need to know which scripts and files are involved and what they do:

/etc/inittab file

The following entry is added to this file if the **Start on system restart** option is chosen on the Start Cluster Services screen:

```
hacmp:2:wait:/usr/sbin/cluster/etc/rc.cluster -boot> /dev/console 2>&1
# Bring up Cluster
```

During the system boot, the **/etc/inittab** file calls the **/usr/sbin/cluster/etc/rc.cluster** script to start HANFS for AIX.

/usr/sbin/cluster/etc/rc.cluster script

This script is run when cluster services are started from SMIT or from **/etc/inittab**). The script does some necessary initialization and then calls the **/usr/sbin/cluster/utilities/clstart** script to start HANFS for AIX. See the **rc.cluster** man page for more information.

/etc/rc.net script

This script is called by the **/usr/sbin/cluster/etc/rc.cluster** script with a **-boot** option to configure and start the TCP/IP interfaces and to set the required network options.

/usr/sbin/cluster/utilities/clstart script

This script, which is called by the **/usr/sbin/cluster/etc/rc.cluster** script, invokes the SRC facility to start cluster daemons. The **clstart** script starts HANFS for AIX with the options (daemons) you specify on the Start Cluster Services screen.

Note that the **clsmuxpd** daemon cannot be started unless the **snmpd** daemon is running. The **/usr/sbin/cluster/utilities/clstart** script checks for this condition and invokes the **snmpd** daemon if necessary. See the **clstart** man page for more information.

/usr/sbin/cluster/utilities/clstop script

This script, which is called by the Stop Cluster Services screen, invokes the SRC facility to stop the cluster daemons with the options you specify. See the **clstop** man page for more information.

/usr/sbin/cluster/utilities/clexit.rc script

If the SRC detects that the **clstrmgr** daemon has exited abnormally (without being shut down using the **clstop** command), it executes this script to halt the system. If the SRC detects that any other HANFS for AIX daemon exits abnormally, it also executes the **clexit.rc** script to stop these processes. See the **clexit.rc** man page for additional information.

Starting Cluster Services

The steps below describe the procedure for starting HANFS for AIX. Start the nodes in the logical order for the correct distribution of the resources you have defined (cascading or rotating).

Note: If a node has a tty console, it must be powered on for HANFS for AIX to be started on that node. If this is not desirable, find the line in the **/rc.cluster** script that ends with `2>/dev/console` and change this line to reflect whatever behavior is desired. For example, you can redirect output to another tty. Be aware that this redirects the startup messages on the node.

Note: For cold boots, let one node fully boot before powering up the other node.

1. Boot the first node and wait until the login screen appears.
2. Boot the second node and wait until the login screen appears.
3. Start the cluster services on the first node. As the root user, enter the **smit clstart** fastpath to display the Start Cluster Services screen.
4. Enter field values as follows:

Start now, on system restart or both	Indicate whether you want the **clstrmgr** and **clsmuxpd** daemons to start when you commit the values on this screen (**now**), when the operating system reboots (**restart**), or on **both** occasions. Choosing **on system restart** or **both** means that the **/etc/inittab** file is altered so that the cluster services are always brought up automatically after a system reboot.
BROADCAST message at startup?	Indicate whether you want to send a broadcast message to users when the cluster services start.
Startup Cluster Information Services?	Indicate whether you want to start the **clinfo** daemon. If your application uses the Clinfo Information Program (Clinfo) or if you use the **clstat** monitor, set this field to **yes**. Otherwise, set it to **no**. The value that you enter in the **Startup Cluster Information Services?** field works in conjunction with the value you enter in the **Start now, on system restart or both** field. If you set the Clinfo startup field to **yes** and the **Start now, on system restart or both** field to **both**, then this daemon is also started whenever the **clstrmgr** and **clsmuxpd** daemons are started.

5. Press Enter.

   The system starts the cluster services, activating the configuration that you have defined.

After a few seconds, a message appears on the console indicating that the **/usr/sbin/cluster/events/node_up** script has begun. The time that it takes this script to run depends on your configuration (that is, the number of disks, the number of file systems to mount, and the number of applications being started). When the **/usr/sbin/cluster/events/node_up_complete** script completes, a second message should be displayed. At this point, HANFS for AIX is up and running on the first node.

6. Start the cluster services on the second node.

   Check for any fatal or non-fatal error messages. You can also use the **/usr/sbin/cluster/clstat** utility, described in Monitoring an HANFS for AIX Cluster on page 15-10, to verify that the nodes are up.

7. Ensure that the HANFS-controlled applications are available.

   Access the applications controlled by HANFS for AIX to verify that they behave as expected.

Ensuring that Network Daemons Start as Part of HANFS Start-Up

Clusters that have both rotating and cascading resource groups, where the cascading resource group does not have IP address takeover configured (that is, the cascading resources are not tied to a service label), can experience problems starting the network daemons. At cluster startup, the first node to join the cluster acquires the rotating address and runs the **telinit -a** command. The second node to join the cluster, since it does not migrate from a boot to a service address, does not run the **telinit -a** command. As a result, the network-related daemons (NFS and NIS for example) do not get started. If your cluster requires this combination of resource groups, customize the pre- and post-event processing to issue the **telinit -a** command.

Stopping Cluster Services

You can stop HANFS for AIX on a node in three different ways.

Graceful	In a *graceful* stop, the HANFS for AIX software shuts down its applications and releases its resources. The other node does not take over the resources of the stopped node.
Graceful with Takeover	In a *graceful with takeover* stop, the HANFS for AIX software shuts down its applications and releases its resources. The surviving node takes over these resources.
Forced	In a *forced* stop, the HANFS for AIX software shuts down immediately. The **/usr/sbin/cluster/events/node_down** script is not run on this node. The other node views it as a graceful shutdown and, therefore, does not take over any resources. The node that is shutting down retains control of all its resources. Applications that require no HANFS for AIX daemons continue to run.

The steps below describe the procedure for stopping HANFS for AIX on a node. If both nodes need to be stopped, follow the same steps on both nodes, but stop them sequentially, not in parallel.

Warning: Never use the **kill -9** command on the **clstrmgr** daemon. The **clstrmgr** daemon is monitored by the SRC, which runs **/usr/sbin/cluster/utilities/clexit.rc** whenever the **clstrmgr** daemon exits abnormally. This script halts the system immediately. The surviving node then initiates fallover.

1. Minimize activity on the system.

   Highly available services become unavailable when the node is stopped. You should notify users of your intentions if their applications will be unavailable, and let them know when services will be restored. Applications that use NFS-mounted file systems should be stopped.

2. Stop the cluster services on the node. As the root user, enter the **smit clstop** fastpath to display the Stop Cluster Services screen.

3. Enter field values as follows:

Stop now, on system restart or both	Indicate whether you want the cluster services to stop **now**, at **restart** (when the operating system reboots), or on **both** occasions. If you choose **restart** or **both**, the **/etc/inittab** file will be modified so that cluster services will no longer come up automatically after a reboot.
BROADCAST cluster shutdown?	Indicate whether you want to send a broadcast message to users before the cluster services stop.
Shutdown mode	Indicate the type of shutdown: - **graceful**—Shut down after the **/usr/sbin/cluster/events/node_down_complete** script is run on this node to release its resources. The other node does not take over the resources of the stopped node. - **graceful with takeover**—Shut down after the **/usr/sbin/cluster/events/node_down_complete** script runs to release its resources. The other node takes over the resources of the stopped node. - **forced**—Shut down immediately. The **/usr/sbin/cluster/events/node_down** script is not run on this node. This node retains control of all its resources. The forced shutdown is intended to be used when a cluster is unstable or is in reconfiguration. You can also use this option to bring down a node while you make a change to the cluster configuration, such as adding an adapter. The node's applications remain available (but without the services of HANFS for AIX daemons).

4. Press Enter.

   The system stops the cluster services according to your specifications.

Note: When you stop the cluster and restart it immediately, there is a two-minute delay before the **clstrmgr** daemon becomes active.

Using the stopsrc Command

You can also use the standard AIX **stopsrc** command with the **-g** flag, specifying the cluster group, to initiate a graceful stop:

```
stopsrc -g cluster
```

Note: Using the **stopsrc** command to stop an individual HANFS for AIX daemon (by specifying the **-s** flag with the name of the daemon) is not recommended and may cause unexpected behavior.

Reintegrating Nodes after a Forced Down

When a node that owns a resource group is forced down, the resource group remains with that node. If the second node that participates in the resource group joins the cluster while the owning node is down, the second node attempts to acquire the resource group, causing possible resource contention.

To avoid this situation, integrate the owning node before any other nodes, if possible. If this is not possible, release the resources owned by the downed node before integrating a different node.

Performing Intentional Fallover

HANFS for AIX makes maintaining a system easier, by allowing applications to continue to function while a node is serviced. This process is called *intentional fallover*, where you purposefully have a node with assigned resources fall over and release its resources to the other node. For example, you would perform intentional fallover in either of the following situations:

- Before any scheduled maintenance.

  Stop HANFS for AIX on a node before you make any hardware or software changes or other scheduled node shutdowns or reboots. Failing to do so may cause unintended cluster events to be triggered on the other node.

- Before any reconfiguration activity.

  Any change to the cluster information within the ODM requires stopping and restarting the cluster services on *both* nodes for the changes to become active.

To perform intentional fallover, complete the following steps:

1. Stop cluster services in graceful with takeover mode on the node to be serviced, as explained in the previous section.
2. Service the inactive node.
3. Reintegrate the node into the cluster.

To reintegrate a node, you start cluster services on that node by entering the **clstart** fastpath, as described in Starting Cluster Services on page 15-3; however, the Cluster Manager is already running on the takeover node. The cluster's fallover configuration determines whether the node rejoining the cluster takes back the shared resources. If the resources are defined as rotating, the node joining the cluster (the original owner) does not take back the shared resources.

You can check the error logs to see if the reintegration has succeeded and the cluster is up and stable. You can also verify that the nodes are up by using the **/usr/sbin/cluster/clstat** utility, described in Monitoring an HANFS for AIX Cluster on page 15-10.

After reintegration, ensure that the HANFS-controlled applications are available. Access the file systems controlled by HANFS for AIX to verify that they behave as expected:

- On each node that is configured to export file systems, execute the **mount** command to make sure the file system can be mounted and execute the **exportfs** command to make sure the file systems can be exported.
- On each node that is configured to NFS-mount the file systems exported by other cluster nodes, execute the **mount** command to make sure the file system can be NFS-mounted.

Remember, if you made any changes to cluster information stored in the ODM, the changes are not active until HANFS for AIX has been stopped and restarted on all nodes in the cluster.

Swapping a Network Adapter Dynamically

As a system administrator, you may at some point experience a problem with a network adapter card on one of the HANFS for AIX cluster nodes. If this occurs, the dynamic adapter swap feature can be used to swap the IP address if an active service or boot adapter with the IP address of an active, available standby adapter on the same node and network. You can choose which standby adapter to which the IP address will be moved. Cluster services do not have to be stopped to perform the swap.

This feature can be used to move an IP address off of an adapter that is behaving erratically, to another standby adapter, without shutting down the node. It can also be used if a hot pluggable adapter device is being replaced on the node. Hot pluggable adapters can be physically removed and replaced without powering off the node.

If hardware address swapping is enabled, the hardware address will be swapped along with the IP address.

Note: A dynamic adapter swap can be performed only between adapters on a single node. You cannot swap the IP address of a service or boot address with the IP address of a standby adapter on a different node. To move an IP address to another node, move its resource group using the DARE Resource Migration utility.

To dynamically swap a network adapter, perform the following procedure:

1. Make sure that no other HANFS for AIX events are running before swapping an adapter.
2. From the main HANFS for AIX SMIT screen, select **Swap Network Adapter**.

   SMIT displays a list of available service/boot adapters.

3. Select the service/boot adapter you want to remove from cluster use, and press Enter.

   SMIT displays a list of available standby adapters.

4. Select the standby adapter you want, and press Enter.

   The Swap Network Adapter menu appears.

5. Verify the service/boot IP label, and the standby IP label you have chosen. If these are correct, press Enter.

   A pop-up message asks if you are sure you want to do this operation. Press Enter again *only* if you are sure you want to swap the network adapter.

After the adapter swap, the service/boot address becomes an available standby adapter. At this point, you can take action to repair the faulty adapter. If you have a hot pluggable adapter, you can replace the adapter while the node and cluster services are up. Otherwise, you will have to stop cluster services and power down the node to replace the adapter.

If you have a hot pluggable adapter, HANFS for AIX will make the adapter unavailable as a standby when you pull it from the node. When the new adapter card is placed in the node, the adapter is incorporated into the cluster as an available standby again. You can then use the dynamic adapter swap feature again to swap the IP address from the standby back to the original adapter.

If you need to power down the node to replace the faulty adapter, HANFS for AIX will configure the service/boot and standby addresses on their original adapters when cluster services are restarted. You do not need to use the dynamic adapter swap feature again to swap the adapters. HANFS for AIX does not record the swapped adapter information in the AIX ODM. Therefore, the changes are not persistent across system reboots or cluster restarts.

Note: All nodes must be at HANFS for AIX Version 4.3.1 or higher to use this feature.

Maintaining Exported File Systems

As part of maintaining an HANFS for AIX cluster, you may have to remove file systems you have exported for use as NFS-mounted file systems or change the export attributes of these exported file systems.

Removing a Directory from the Exports List

If you need to remove a directory from the **exports** file, take the following steps for each directory to remove.

1. To remove a directory from the **exports** file, enter the **smit rmnfsexp** fastpath to display the Remove a Directory from Exports List screen.
2. Enter the full pathname of the directory you want to remove from the **exports** list and press Enter. SMIT displays the Remove a Directory from Exports List screen.

3. Enter field values as follows:

Pathname of exported directory to be removed from the NFS exports list	The full pathname of the exported directory you want to remove from the **exports** list.
REMOVE export now, system restart or both	Enter **restart**.
PATHNAME of alternate exports file	You must enter **/usr/sbin/cluster/etc/exports** in this field for the directory to be properly removed from the HANFS for AIX **exports** list.

4. Press Enter.

   HANFS for AIX uses this information to update the **/usr/sbin/cluster/etc/exports** file. The following is an example of an **exports** file.

```
/delta
/source/master -root=pearl
/source/master/bldenv -root=pearl
/temp -root=chowdereye:steamereye
```

5. Press F3 to return to the **Remove a Directory from Exports List** screen, or F10 to exit SMIT.

   Repeat steps 2 through 5 for each directory you want to remove.

6. Change the appropriate resource groups that include the directories you are removing. See Chapter 12, Configuring an HANFS for AIX Cluster, for instructions about configuring resources for resource groups.
7. Synchronize the cluster topology and node configuration.

Changing the Export Options for a Directory

Because the Change/Show Attributes of an Exported Directory SMIT screen provided by AIX does not allow you to specify an alternate pathname for the **exports** file, you cannot change the export options for a directory using SMIT.

If you use SMIT, you must first remove the directory and then add it with the desired options.

As an alternative, you can change the export options of a file system by editing the **/usr/sbin/cluster/etc/exports** file directly.

Monitoring an HANFS for AIX Cluster

By design, HANFS for AIX compensates for various failures that occur within a cluster. For example, HANFS for AIX compensates for a network adapter failure by swapping in a standby adapter. As a result, it is possible that a component in the cluster has failed and that you are unaware of the fact. The danger here is that, while HANFS for AIX can survive one or possibly several failures, *a failure that escapes your notice threatens a cluster's ability to provide a highly available environment*.

To avoid this situation, you should customize your system by adding event notification to the scripts designated to handle the various cluster events. You can specify a command that sends you mail indicating that an event is about to happen (or that an event has just occurred), along with information about the success or failure of the event. The mail notification system enhances the standard event notification methods.

Use the AIX Error Notification facility to add an additional layer of high availability to an HANFS for AIX environment. Although the combination of HANFS for AIX and the inherent high availability features built into the AIX system keeps single points of failure to a minimum, there are still failures that, although detected, are not handled in any useful way. See Chapter 14, Supporting AIX Error Notification, for instructions on customizing error notification.

Tools for Monitoring an HANFS for AIX Cluster

HANFS for AIX provides the following tools for monitoring a cluster:

- The **/usr/sbin/cluster/clstat** utility, which reports the status of the key cluster components—the cluster itself, the nodes in the cluster, and the network adapters connected to the nodes.
- The HAView utility, which monitors HANFS clusters through the TME 10 NetView for AIX graphical network management interface. It enables you to monitor multiple HANFS clusters and cluster components across a network from a single node.
- The Show Cluster Services screen, which shows the status of the HANFS for AIX daemons.
- The **/usr/adm/cluster.log** file, which tracks cluster events, the **/tmp/hacmp.out** file, which records the output generated by configuration scripts as they execute, and the **/usr/sbin/cluster/history/cluster.*mmdd*** log file, which logs the daily cluster history.

When you monitor a cluster, use the **/usr/sbin/cluster/clstat** or HAView utility to examine the cluster and its components. Also, constantly monitor the **hacmp.out** file. Use the Show Cluster Services screen to make sure that the necessary HANFS for AIX daemons are running on each node. Finally (if necessary) examine the other cluster log files to get a more in-depth view of the cluster status.

See Chapter 16, Troubleshooting HANFS for AIX Clusters, for more information about diagnosing a problem with an HANFS cluster.

Using the clstat Utility to Monitor Cluster Status

HANFS for AIX provides the **/usr/sbin/cluster/clstat** utility for monitoring a cluster and its components. The **clstat** utility is a Clinfo client program that uses the Clinfo API to retrieve information about the cluster. Both Clinfo and the **clstat** utility must be running on a node for **clstat** to work properly.

The **/usr/sbin/cluster/clstat** utility runs on both ASCII and X Window Display clients in either single-cluster or multi-cluster mode. The client display automatically corresponds to the capability of the system; however, you can run an ASCII display on an X-capable machine.

The **clstat** utility reports whether the cluster is up, down, or unstable. It also reports whether a node is up, down, joining or leaving, and the number of nodes in the cluster. For each node, the utility displays the IP label and address of each network interface attached to the node, and whether that interface is up or down. See the **clstat** man page for additional information about this utility.

Single-Cluster ASCII Display Mode

In single-cluster ASCII display mode, the **clstat** utility displays information about only one cluster.

To invoke the **clstat** utility in single-cluster (non-interactive) mode, enter:

```
clstat
```

A screen similar to the following appears:

```
              clstat - HACMP for AIX Cluster Status Monitor
              ---------------------------------------------

Cluster: Cluster1       (40)            Date: July 15, 1998 (3:01 PM)
                State: UP               Nodes: 2
                SubState: STABLE

        Node: beavis            State: UP
           Interface: beavis_boot (0)           Address: 140.186.150.181
                                                State:   UP
           Interface: beavis_tty (1)            Address: 0.0.0.0
                                                State:   UP

        Node: buzzcut           State: UP
           Interface: buzzcut_boot (0)          Address: 140.186.150.156
                                                State:   UP
           Interface: buzzcut_tty (1)           Address: 0.0.0.0
                                                State:   UP


      ************ f/forward, b/back, r/refresh, q/quit **********
```

The clstat Single-Cluster ASCII Display Mode

The cluster information displayed shows both a cluster name and an ID. In this example, the cluster is up and nodes are up. Each node has two network adapters. Note that the *forward* and *back* menu options apply when more than one page of information is available to display.

If more than one cluster exists when you run the **clstat** command, the utility notifies you of this fact and requests that you retry the command specifying one of the following options:

```
clstat [-c cluster ID | -r seconds | -i ]
```

-c *cluster ID* Displays information about the cluster with the specified ID. If the cluster is not available, the **clstat** utility continues looking for it until it is found or until the program is cancelled. Note that this option cannot be used if the **-i** option (for multi-cluster mode) is used.

-r *seconds* Updates the cluster status display at the specified number of seconds. The default is 1 second; however, the display is updated only if the cluster state changes.

-i Displays information about clusters interactively.

To see cluster information about a specific cluster, enter:

```
clstat [-c cluster ID]
```

Multi-Cluster ASCII Display Mode

The multi-cluster (interactive) mode lets you monitor all clusters that Clinfo can access from the list of active service IP labels or addresses found in the **/usr/sbin/cluster/etc/clhosts** file. In multi-cluster mode, the **clstat** utility displays this list of recognized clusters and their IDs, allowing you to select a specific cluster to monitor. For example, to invoke the **clstat** utility in multi-cluster mode, enter:

```
clstat -i
```

where the **-i** indicates multi-cluster (interactive) mode. A screen similar to the following appears:

```
              clstat - HACMP for AIX Cluster Status Monitor
              ---------------------------------------------

Number of clusters active: 1

              ID      Name            State

              40      Cluster1        UP



Select an option:
        # - the Cluster ID                       q- quit
```

The clstat Multi-Cluster Mode Menu

This screen displays the ID, name, and state of each active cluster accessible by the local node. You can either select a cluster for which you want to see detailed information, or quit the **clstat** utility. When you enter a cluster ID, a screen similar to the following appears:

```
              clstat - HACMP for AIX Cluster Status Monitor
              ---------------------------------------------

Cluster: Cluster1       (40)            Date: July 15, 1998 (3:58 PM)
                State: UP               Nodes: 2
                SubState: STABLE

        Node: beavis            State: UP
           Interface: beavis_boot (0)           Address: 140.186.150.181
                                                State:   UP
           Interface: beavis_tty (1)            Address: 0.0.0.0
                                                State:   UP
```

```
Node: buzzcut           State: UP
   Interface: buzzcut_boot (0)          Address: 140.186.150.156
                                        State:   UP
   Interface: buzzcut_tty (1)           Address: 0.0.0.0
                                        State:   UP


********** f/forward, b/back, r/refresh, q/quit **********
```

The clstat Multi-Cluster ASCII Display Mode

After viewing this screen, press “q” to exit the display. The multi-cluster mode returns you to the cluster list so you can select a different cluster. Note that you can use all menu options displayed. The *forward* and *back* options allow you to scroll through displays of active clusters without returning to the previous screen.

Multi-Cluster X Window System Display

The **/usr/sbin/cluster/clstat** utility can run on X-capable clients. A client’s DISPLAY environment variable, however, must be set to run the X Window System. Once the DISPLAY environment variable is set, the **clstat** utility shows node and interface information for a selected cluster. To invoke the **clstat** utility X Window System display, enter the **clstat** command:

```
clstat [-c ID | -n name | -r interval]
```

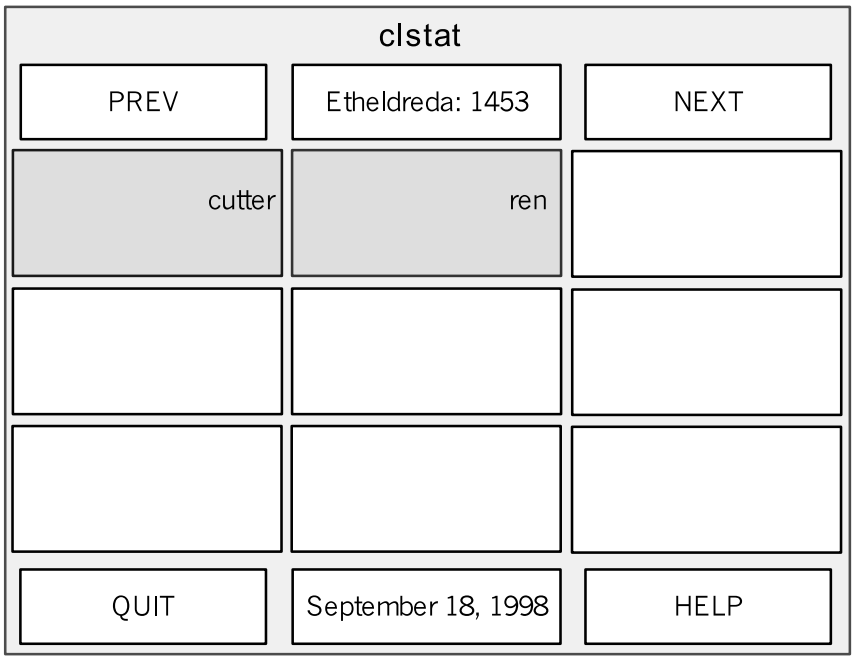
where:

-c *ID* Displays information about the cluster with the specified ID if that cluster is active. This option cannot be used with the **-n** option.

-n *name* Displays information about the cluster with the specified name if that cluster is active. This option cannot be used with the **-c** option.

-r *interval* The interval, in seconds, at which the **clstat** utility updates the display. The default is 10 seconds.

The **clstat** utility generates a window similar to the one shown in the following figure:



The clstat X Window System Display

The middle box in the top row indicates the cluster name and ID. If the cluster is stable, this box appears green. If the cluster destabilizes for any reason, this box changes to red. The large boxes in other rows represent nodes. A node name appears in a box for each active node in the cluster. Nodes that are up are shown in green, nodes that are down are shown in red, nodes that are joining or leaving the cluster are shown in yellow, and nodes that are undefined are shown in the background color. Colors are configured in the **Xclstat** resource file.

On a monochrome display, gray shading represents the colors as follows:

dark gray: red

gray: yellow

light gray: green

Five buttons are available on the **clstat** display:

PREV Displays the previous cluster (loops from end to start).

NEXT Displays the next cluster (loops from start to end).

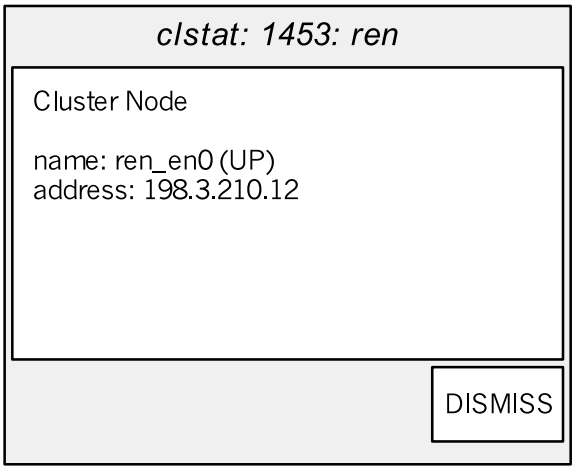
cluster:ID The refresh bar. Pressing this bar updates the status display.

QUIT Cancels the **clstat** utility.

HELP Displays help information.

Viewing Network Interface Information on an X Window System Display

To view information about network interfaces attached to a node, click mouse button 1 on the appropriate node box in the **clstat** display. A pop-up window similar to the following appears. The title in the example shows that you are viewing node *ren* in cluster *1453*.



The clstat Node Information Display

Press the DISMISS button to close the pop-up window and to return to the **clstat** display window.

Monitoring a Cluster With HAView

HAView is a cluster monitoring utility that allows you to monitor HACMP clusters using TME 10 NetView for AIX. Using NetView, you can monitor clusters and cluster components across a network from a single management station.

HAView creates and modifies NetView objects that represent clusters and cluster components. It also creates submaps that present information about the state of all nodes, networks, and network interfaces associated with a particular cluster. This cluster status and configuration information is accessible through NetView's menu bar.

HAView monitors cluster status using the Simple Network Management Protocol (SNMP). It combines periodic polling and event notification through traps to retrieve cluster topology and state changes from the HACMP management agent, the Cluster SMUX peer daemon (**clsmuxpd**).

You can view cluster event history using the HACMP Event Browser and node event history using the Cluster Event Log. Both browsers can be accessed from the NetView menu bar. The **/usr/sbin/cluster/history/cluster.*mmdd*** file contains more specific event history. This information is helpful for diagnosing and troubleshooting fallover situations. See Cluster Message Log Files on page 16-2 for more information about this log file.

HAView Installation Considerations

HAView has a client/server architecture. You must install both an HAView server image and an HAView client image, on the same machine or on separate server and client machines. For more information on installation considerations, see Chapter 11, Installing HANFS for AIX Software, in particular, the section HAView Installation Notes on page 11-2.

HAView File Modification Considerations

Certain files may need to be modified in order for HAView to monitor your cluster properly. When configuring HAView, check and/or edit the following files:

- **haview_start**
- **clhosts**
- **snmpd.conf**

The haview_start File

The **haview_start** file must be edited by the user so that it includes the name of the node that has the HAView server executable installed. This is how the HAView client knows where the HAView server is located. Regardless of whether the HAView server and client are on the same node or different nodes, you are required to specify the HAView server node in the **haview_start** file.

The **haview_start** file is loaded when the haview client is installed and is stored in **/usr/haview**. Initially, the **haview_start** file contains only the following line:

```
"${HAVIEW_CLIENT:-/usr/haview/haview_client}" $SERVER
```

Add the following line to the file:

```
SERVER="${SERVER:-<your server name>}"
```

For example, if the HAView server is installed on *mynode*, the edited **haview_start** file appears as follows:

```
SERVER="${SERVER:-mynode}"
"${HAVIEW_CLIENT:-/usr/haview/haview_client}" $SERVER
```

where *mynode* is the node that contains the HAView server executable.

The clhosts File

HAView monitors a cluster's state within a network topology based on cluster-specific information in the **/usr/sbin/cluster/etc/clhosts** configuration file. The **clhosts** file must be present on the NetView management node. Make sure this file contains the IP address or IP label of the service and/or boot adapters of the nodes in each cluster that HAView is to monitor.

Note: Make sure the hostname and the service label of your NetView nodes are exactly the same. (If they are not the same, add an alias in the **/etc/hosts** file to resolve the name difference.)

Warning: If an invalid IP address exists in the **clhosts** file, HAView will fail to monitor the cluster. Make sure the IP addresses are valid, and there are no extraneous characters in the **clhosts** file.

The snmpd.conf File

The NetView management node must also be configured in the list of trap destinations in the **snmpd.conf** files on the cluster nodes of all clusters you want it to monitor. This makes it possible for HAView to utilize traps in order to reflect cluster state changes in the submap in a timely manner. Also, HAView can discover clusters not specified in the **clhosts** file on the nodes in another cluster.

The format for configuring trap destinations is as follows:

```
trap  <community name> <IP address of NetView management node>1.2.3 fe
```

For example, enter:

```
trap          public                  140.186.131.121 1.2.3 fe
```

Note the following:

- You can specify the name of the management node instead of the IP address.
- You can include multiple trap lines in the snmpd.conf file.

NetView Hostname Requirements for HAView

The following hostname requirements apply to using HAView in a NetView environment. If you change the hostname of an adapter, the NetView database daemons and the default map are affected as follows:

Hostname Effect on the NetView Daemon

The hostname required to start NetView daemons must be associated with a valid interface name or else NetView fails to start.

Hostname Effect on the NetView Default Map

If you change the hostname of the NetView client, the new hostname does not match the original hostname referenced in the NetView default map database and NetView will not open the default map. Using the NetView **mapadmin** command, you need to update the default map (or an invalid map) to match the new hostname.

See the *NetView for AIX Administrator's Guide* for more information about updating or deleting an invalid NetView map.

Starting HAView

Once you've installed the HAView client and server, HAView is started and stopped when you start or stop NetView, but there are some conditions to verify on the management node before starting HAView.

Before starting NetView/HAView, check the management node as follows:

- Make sure both client and server components of HAView are installed. See Chapter 11, Installing HANFS for AIX Software, for more information.
- Make sure access control has been granted to remote nodes by running the **xhost** command with the plus sign (+) or with specified nodes:

  `xhost +` (to grant access to all computers)

  or, to grant access to specific nodes only:

  `xhost <computers to be given access>`

- Make sure the DISPLAY variable has been set to the monitoring node and to a label that can be resolved by and contacted from remote nodes:

  ```
  export DISPLAY=<monitoring node>:0.0
  ```

These actions allow you to access the HAView Cluster Administration option.

After ensuring these conditions are set, type the following to start NetView:

```
/usr/OV/bin/nv6000.
```

(Refer to the *NetView for AIX User's Guide for Beginners* for further instructions about starting NetView.)

When NetView starts, HAView creates objects and symbols to represent a cluster and its components. Through submaps, you can view detailed information about these components.

HAView places the Clusters symbol on the NetView map after NetView starts. As shown in the following figure, the Clusters symbol is placed alongside the NetView Collections, Manager Submap, and IP Internet symbols.

HAView Clusters Symbol

Viewing Clusters and Components

To see which clusters HAView is currently monitoring, double-click the Clusters symbol. The Clusters submap appears. You may see one or more symbols that represent specific clusters. Each symbol is identified by a label indicating the cluster's name. Double-click a cluster symbol to display symbols for nodes and networks within that cluster.

Note that the cluster status symbol may remain unknown until the next polling cycle, even though the status of its cluster components is known. See Polling Intervals on page 15-21 for more information about the default intervals and how to change them using SMIT.

Note: You can view component details at any time using the shortcut **ctrl-o**. See Obtaining Component Details on page 15-20 for information and instructions.

Read-Write and Read-Only NetView Maps

Normally, you have one master monitoring station for NetView/HAView. This station is supplied with new information as cluster events occur, and its map is updated so it always reflects the current cluster status.

In normal cluster monitoring operations, you will probably not need to open multiple NetView stations on the same node. If you do, and you want the additional stations to be updated with current cluster status information, you must be sure they use separate maps with different map names. For more information on multiple maps and changing map permissions, see the *NetView for AIX Administrators Guide*.

Interpreting Cluster Topology States

When using HAView to view cluster topology, symbols for clusters and cluster components such as nodes and networks are displayed in various colors depending on the object's state. The following table summarizes colors you may see when monitoring a cluster.

Status	Meaning	Symbol Color	Connection Color (network submap)
Critical	The object has failed or is not functioning. If the symbol is a node or network, the node or network is DOWN.	Red	Red
Normal	The object is functioning correctly. If the symbol is a node object, the node is UP.	Green	Black
Marginal	Some object functions are working correctly; others are not.	Yellow	Red
Unknown	The object's state cannot be determined. It may not be currently monitored by HAView.	Blue	Blue

Note: You can select **Legend** at any time from the Help pull-down menu to view NetView and HAView symbols and to understand their associative colors.

The Navigation Tree and Submap Windows

In addition to the submap window, the NetView Navigation Tree Window can help you keep track of your current location in the HAView hierarchy. Press the Tree button to see the Navigation Tree Window. In the Navigation Tree, the blue outline indicates which submap you are in

The Symbols Legend

At any time, you can select **Legend** from the Help pull-down menu to view all NetView and HAView symbols and the meanings of the symbols and their various colors.

The Help Menu

To view help topics, select **Help > Index > Tasks > HAView Topics**.

Viewing Networks

To view the state of the nodes and addresses connected to a network associated with a specific cluster, double-click a network symbol in the specific Cluster submap. A network submap appears displaying symbols for all nodes connected to the network. The symbols appear in a color that indicates the nodes' current state. The vertical line representing a network is called the network connection. Its color indicates the status of the connection between the node and the network.

Viewing Nodes

To view the state of nodes associated with a particular network, double-click a network symbol. A submap appears displaying all nodes connected to the network. Each symbol's color indicates the associated node's current state.

You can also view the state of any individual node associated with a cluster by double-clicking on that node's symbol in the specific cluster submap.

Viewing Addresses

To view the status of addresses serviced by a particular node, double-click a node symbol from either a cluster or network submap. A submap appears displaying symbols for all addresses configured on a node. Each symbol's color indicates the associated address's current state.

Note: When viewing adapters in a node submap from a network submap, all adapters relating to that node are shown, even if they are not related to a particular network.

Obtaining Component Details

NetView dialog boxes allow you to view detailed information about a cluster object. A dialog box can contain information about a cluster, network, node, network adapter, or resource group, or about cluster events. You can access an object's dialog box using the NetView menu bar or the Object Context menu, or by pressing **ctrl-o** at any time:

To view details about a cluster object using the NetView menu bar:

1. Click on an object in any submap.
2. Select the **Modify/Describe** option from the NetView Edit menu.
3. Select the **Object** option.

   An Object Description dialog window appears.
4. Select **HAView for AIX** and click on **View/Modify Object Attributes**.

   An Attributes dialog window appears.

Note: You can view dialog boxes for more than one object simultaneously by either clicking the left mouse button and dragging to select multiple objects, or by pressing the **Alt** key and clicking on all object symbols for which you want more information.

To view details about a cluster object using the Object Context menu:

1. Click on an object in any submap.
2. Click on the symbol you have highlighted to display the object context menu, using:
   - Button 3 on a three-button mouse
   - Button 2 on a two-button mouse.
3. Select **Edit** from the object context menu.
4. Select **Modify/Describe** from the Edit cascade menu.
5. Select the **Object** option.

   An Object Description dialog window appears.
6. Select **HAView for AIX** and click on **View/Modify Object Attributes**.

   An Attributes dialog window appears.

Polling Intervals

To ensure that HAView is optimized for system performance and reporting requirements, you can customize these two parameters:

- The polling interval (in seconds) at which HAView polls the HACMP clusters to determine if cluster configuration or object status has changed. The default is 60 seconds.
- The polling interval (in minutes) at which HAView polls the **clhosts** file to determine if new clusters have been added. The default for Cluster Discovery polling is 120 minutes.

You can change the HAView polling intervals using the SMIT interface as follows:

1. On the HAView server node, open a SMIT screen by typing:

   ```
   smitty haview
   ```

   The Change/Show Server Configuration window opens.
2. Enter the polling interval numbers you want (between 1 and 32000) and press OK.

Note: If the **snmpd.conf** file is not properly configured to include the NetView server as a trap destination, HAView can detect a trap that occurs as a result of a cluster event, but information about the network topology may not be timely. Refer back to HAView File Modification Considerations on page 15-16 for more information on the snmpd.conf file.

Removing a Cluster

If a cluster does not respond to status polling, you can use the Remove Cluster option to remove the cluster from the database. To remove a cluster, it must be in an UNKNOWN state, represented by a blue cluster symbol. If the cluster is in any other state, the Remove Cluster option is disabled.

Warning: The **Remove Cluster** option is the only supported way to delete HAView objects from submaps. Do not delete an HAView symbol (cluster or otherwise) through the Delete Object or Delete Symbol menu items. If you use these menu items, HAView continues to poll the cluster.

When you remove a cluster, the following actions occur:

- The cluster name is removed from the NetView object database and HAView stops polling the cluster.
- The symbol for the cluster is deleted.
- The symbols for all child nodes, networks, and addresses specific to that cluster are deleted.

If you are removing the cluster permanently, remember to remove the cluster addresses from the **/usr/sbin/cluster/etc/clhosts** file. If you do not remove the cluster addresses from the clhosts file, New Cluster Discovery polling continues to search for the cluster.

To remove a cluster:

1. Click on the cluster symbol you wish to remove. The cluster must be in an UNKNOWN state, represented by a blue cluster symbol.
2. Select **HAView** from the Tools pull-down menu.
3. Select **Remove Cluster** from the HAView cascade menu.

Using the HAView Cluster Administration Utility

HAView allows you to start a **SMIT hacmp** session to perform cluster administration functions from within the NetView session. The administration session is run on an aixterm opened on the chosen node through a remote shell. You can open multiple sessions of SMIT hacmp while in HAView.

Note: You can start an administration session for any node that is in an UP state (the node symbol is green). If you attempt to start an administration session when the state of the node is DOWN or UNKNOWN, no action occurs.

When bringing a node up, the HAView node symbol may show green before all resources are acquired. If you select the node symbol and attempt to open an administration session before all resources are acquired, you may receive an error.

Opening and Closing a Cluster Administration Session

To open a cluster administration session:

1. Click on an available node symbol (one that is green).

2. Select the **Tools > HAView > Cluster Administration**.
3. Proceed with your tasks in SMIT.
4. Select **F10** to exit the Cluster Administration session. The aixterm session will also close when using either of these choices.

Cluster Administration Notes and Requirements

Keep the following considerations in mind when using the Cluster Administration option:

- Be sure you have run the **xhost** command prior to starting NetView, so that a remote node can start an aixterm session on your machine.
- Be sure you have set the DISPLAY variable to a label that can be resolved and contacted from remote nodes.
- For the cluster administration session to proceed properly, the current NetView user (the account that started NetView) must have sufficient permission and be authenticated to perform an rsh to the remote node in the **.rhosts** file or through Kerberos if you have HANFS on an RS/6000 SP. For more information on editing the .rhosts file, see Chapter 10, Performing Additional AIX Tasks (section on editing the .rhosts file). For information on configuring Kerberos, see Appendix C, Installing and Configuring HANFS for AIX on RS/6000 SPs.
- If an IP Address Takeover (IPAT) occurs while a cluster administration session is running, the route between the remote node that the HAView monitoring node may be lost.

HAView Browsers

HAView provides two browsers which allow you to view the event history of a cluster, the Cluster Event Log and the HACMP Event Browser.

Cluster Event Log

Using the Cluster Event Log you can view the event history for a cluster as recorded by a specific node. The Log browser is accessible through the NetView Tools menu, and is only selectable if an active node symbol is highlighted.

For more detailed information on a node's event history, log onto the specific node and check the Cluster Message Log Files. See the Chapter 16, Troubleshooting HANFS for AIX Clusters for more information on Cluster Message Log Files.

Note: To ensure that the header for the Cluster Event Log displays properly, install all the NetView fonts on your system.

1. Click on the node symbol for which you wish to view a Cluster Event Log.
2. Select **HAView** from the Netview Tools menu.
3. Select the **Cluster Event Log** option.
4. Set the **number of events to view** field. You can use the up and down arrows to change this number or you can enter a number directly into the field. The possible range of values is 1 to 1000 records. The default value is 100.
5. Press the **Issue** button to generate the list of events. The message area at the bottom of the dialog box indicates when the list is done generating.

When the list is done generating, the dialog box displays the following view-only fields:

Event ID	This field displays a numeric identification for each event which occurred on the cluster.
Node Name	The name of the node on which the event occurred.
Time	The date and time the event occurred. This field is in the format MM DD hh:mm:ss.
Description	A description of the event.

6. Press the **Dismiss** button to close the dialog box.

HACMP Event Browser

HAView provides a NetView browser which allows you to view the accumulative event history of a cluster. The browser shows the history of all nodes in the cluster, broadcast through an assigned primary node. If the primary node fails, another node will assume the primary role and continue broadcasting the event history.

The HACMP Event Browser provides information on cluster state events. A filter is used to block all redundant traps.

The HACMP Event Browser is available through the NetView menu bar. The menu item is always active, and when selected will start a NetView browser showing the event history for all active clusters. Note that you can access only one instantiation of the Event Browser at a time.

To view the HACMP Event Browser:

1. Select **HAView** from the Netview Tools menu.
2. Select the **HACMP Event Browser** option.

   The HACMP Event Browser appears. Note that only one instantiation of the Event Browser can be accessed at a time. See the *NetView for AIX User's Guide for Beginners* for more information on the NetView browser functions.
3. Select the **Close** option from the File menu of the HACMP Event Browser menu bar to close the browser.

Note: When you exit the Event Browser, the HAView application restarts. At this time, the HACMP cluster icon turns blue, disappears, and then reappears.

Monitoring Cluster Services

After checking cluster, node, and network interface status, you can check the status of the HANFS for AIX daemons on cluster nodes. Use the Show Cluster Services screen to check the status of the HANFS for AIX daemons on a node.

1. Enter the **smit clshow** fastpath to view cluster services on a node.

   A screen similar to the one shown in the following figure appears:

```
                        COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Subsystem         Group            PID     Status
 clstrmgr         cluster                  inoperative
 clinfo           cluster                  inoperative
 clsmuxpd         cluster                  inoperative
 cllockd          lock                     inoperative
```

The Command Status Display of Node Information

The screen indicates the:

- HANFS for AIX subsystem
- Group to which the subsystem belongs
- Process ID number of the subsystem, if it is running
- Status of the subsystem.

Note: If the **clsmuxpd** daemon is not active, check the status of the SNMP (**snmpd**) daemon.

You can also use the **ps** command and **grep** for the name of the daemon. For example, to find the status of the **clinfo** daemon, enter the following:

```
ps -aux | grep clinfo
```

HANFS for AIX Log Files

HANFS for AIX writes the messages it generates to the system console and to several log files. Because each log file contains a different subset of the types of messages generated by HANFS for AIX, you can get different views of cluster status by viewing different log files. For more detailed information about the log files, see Chapter 16, Troubleshooting HANFS for AIX Clusters.

Part 5 Troubleshooting HANFS for AIX

In this part, you learn about how to handle problems that may arise occasionally with an HANFS for AIX cluster.

Chapter 16 Troubleshooting HANFS for AIX Clusters

This chapter describes how to diagnose a problem with an HANFS for AIX cluster. The chapter describes how to view cluster log files and obtain trace information on HANFS for AIX daemons.

Note: The default locations of cluster log files are used in this chapter. If you redirected any logs, check the appropriate location.

Viewing HANFS for AIX Cluster Log Files

Your first approach to diagnosing a problem affecting your cluster should be to examine the cluster log files for messages output by the HANFS for AIX subsystems. These messages can provide invaluable information toward understanding the current state of the cluster. The following sections describe the types of messages output by the HANFS for AIX software and the log files into which the system writes these messages.

Types of Cluster Messages

The HANFS for AIX software generates several types of messages:

- **Event notification messages**—Cluster events cause HANFS for AIX scripts to execute. When scripts start, complete, or encounter error conditions, the HANFS for AIX software generates a message. For example, the following fragment from a cluster log file illustrates the start and completion messages for several HANFS for AIX scripts. The messages include any parameters passed to the script.

```
Feb 25 11:02:46 EVENT START: node_up 2
Feb 25 11:02:46 EVENT START: node_up_local
Feb 25 11:02:47 EVENT START: acquire_service_addr
Feb 25 11:02:56 EVENT COMPLETED: acquire_service_addr
```

- **Verbose script output**—In addition to the start, completion, and error messages generated by scripts, the HANFS for AIX software can also generate a detailed report of each step of script processing. In verbose mode, which is the default, the shell generates a message for each command executed in the script, including the values of all arguments to these commands. Verbose mode is recommended. The following fragment from a cluster log file illustrates the verbose output of the **node_up** script. The verbose messages are prefixed with a plus (+) sign.

```
Feb 25 11:02:46 EVENT START: node_up 2
+ set -u
+ [ 2 = 2 ]
+ /usr/sbin/cluster/events/cmd/clcallev node_up_local
Feb 25 11:02:46 EVENT START: node_up_local
+ set -u
+ rm -f /usr/sbin/cluster/server.status
+ /usr/sbin/cluster/events/cmd/clcallev acquire_service_addr

Feb 25 11:02:47 EVENT START: acquire_service_addr
+ set -u
+ +grep : boot + cut -d: -f1
/usr/sbin/cluster/utilities/cllsif -cSi 2
```

- **Cluster state messages**—When an HANFS for AIX cluster starts, stops, or goes through other state changes, it generates messages. These messages may be informational, such as a warning message, or they may report a fatal error condition that causes an HANFS for AIX subsystem to terminate. In addition to the **clstart** and **clstop** commands, the following HANFS for AIX subsystems generate status messages:
  - The Cluster Manager daemon (**clstrmgr**)
  - The Cluster Information Program daemon (**clinfo**)
  - The Cluster SMUX Peer daemon (**clsmuxpd**).

  The following example illustrates cluster state messages output by the Cluster Manager, the Clinfo daemon, and several HANFS for AIX scripts:

```
Feb 25 11:02:30 limpet HACMP for AIX: Starting execution of /etc/rc.cluster with
parameters: --
Feb 25 11:02:32 limpet HACMP for AIX: clstart: called with flags -sm
Feb 25 11:02:36 limpet clstrmgr[18363]: CLUSTER MANAGER STARTED
Feb 25 11:02:40 limpet HACMP for AIX: Completed execution of /etc/rc.cluster with
parameters: --. Exit status = 0
Feb 25 11:02:46 limpet HACMP for AIX: EVENT START: node_up 2
Feb 25 11:02:47 limpet HACMP for AIX: EVENT START: node_up_local
Feb 25 11:02:47 limpet HACMP for AIX: EVENT START: acquire_service_addr
Feb 25 11:02:53 limpet HACMP for AIX: EVENT COMPLETED: acquire_service_addr
Feb 25 11:02:54 limpet HACMP for AIX: EVENT START: get_disk_vg_fs
Feb 25 11:02:55 limpet HACMP for AIX: EVENT COMPLETED: get_disk_vg_fs
Feb 25 11:03:35 limpet clinfo[6543]: read_config: node address too long, ignoring.
```

Cluster Message Log Files

The HANFS for AIX software writes the messages it generates to the system console and to several log files. Each log file contains a different subset of messages generated by the HANFS for AIX software. When viewed as a group, the log files provide a detailed view of all cluster activity. The following list describes the log files into which the HANFS for AIX software writes messages and the types of cluster messages they contain. The list also provides recommendations for using the different log files.

/usr/adm/cluster.log The main HANFS for AIX log file. Contains time-stamped, formatted messages generated by HANFS for AIX scripts and daemons. For information about viewing this log file and interpreting its messages, see Understanding the cluster.log File on page 16-4.

Recommended Use: Because this log file provides a high-level view of current cluster status, it is a good place to look first when diagnosing a cluster problem.

/tmp/hacmp.out

Contains time-stamped, formatted messages generated by HANFS for AIX configuration and startup scripts on the current day. The **/tmp/hacmp.out** log file does not contain state messages.

In verbose mode (recommended), this log file contains a line-by-line record of every command executed by scripts, including the values of all arguments to each command. To receive verbose output, the **Debug Level** run-time parameter should be set to **high** (the default). For information about viewing this log file and interpreting its messages, see Understanding the hacmp.out Log File on page 16-7. For information about setting run-time parameters, see Chapter 12, Configuring an HANFS for AIX Cluster.

Recommended Use: Because the information in this log file supplements and expands upon the information in the **/usr/adm/cluster.log** file, it is an important source of information when investigating a problem.

system error log

Contains time-stamped, formatted messages from all AIX subsystems, including HANFS for AIX scripts and daemons. Cluster events are logged as operator messages (error id: AA8AB241) in the system error log. For information about viewing this log file and interpreting the messages it contains, see Understanding the System Error Log on page 16-11.

Recommended Use: Because the system error log contains time-stamped messages from many other system components, it is a good place to correlate cluster events with system events.

/usr/sbin/cluster/history/cluster.*mmdd*

Contains time-stamped, formatted messages generated by HANFS for AIX scripts. The system creates a cluster history file every day, identifying each file by its filename extension, where *mm* indicates the month and *dd* indicates the day. For information about viewing this log file and interpreting its messages, see Understanding the Cluster History Log File on page 16-13.

Recommended Use: Use the cluster history log files to get an extended view of cluster behavior over time. While it is more likely that you will use these files during troubleshooting, you should look at them occasionally to get a more detailed picture of the activity within a cluster.

/tmp/cm.log Contains time-stamped, formatted messages generated by **clstrmgr** activity. By default, the messages are short. IBM Support personnel may have you turn on **clstrmgr** debug options (for verbose, detailed information) to help them understand a particular problem. With debugging turned on, this file grows quickly. You should clean up the file and turn off debug options as soon as possible. The **/tmp/cm.log** file is reset automatically every time Cluster Services are started.

Recommended Use: Information in this file is for IBM Support personnel.

/tmp/emuhacmp.out Contains time-stamped, formatted messages generated by the HACMP for AIX Event Emulator. The messages are collected from output files on each node of the cluster, and cataloged together into the **/tmp/emuhacmp.out** log file.

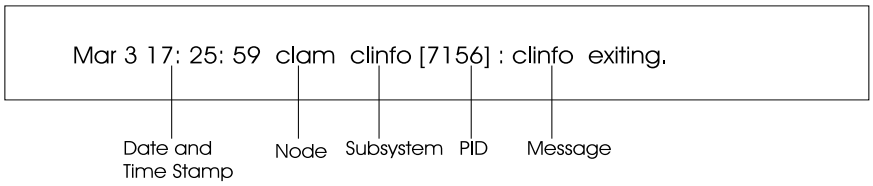
In verbose mode (recommended), this log file contains a line-by-line record of every event emulated. Customized scripts within the event are displayed, but commands within those scripts are not executed. For more information, see Understanding the /tmp/emuhacmp.out File on page 16-14.

Understanding the cluster.log File

The **/usr/adm/cluster.log** file is a standard text file. When checking this file, first find the most recent error message associated with your problem. Then read back through the log file to the first message relating to that problem. Many error messages cascade from an initial error that usually indicates the problem source.

Format of Messages in the cluster.log File

The entries in the **/usr/adm/cluster.log** file use the following format:



Each entry has the following information:

Date and Time stamp The day and time on which the event occurred.

Node The node on which the event occurred.

Subsystem	The HANFS for AIX subsystem that generated the event. The subsystems are identified by the following abbreviations: **clstrmgr**—The Cluster Manager daemon **clinfo**—The Cluster Information Program daemon **clsmuxpd**—The Cluster SMUX Peer daemon **HACMP for AIX**—Startup and reconfiguration scripts.
PID	The process ID of the daemon generating the message. (Not included for messages output by scripts.)
Message	The message text.

The entry in the preceding example indicates that the Cluster Information program (**clinfo**) stopped running on the node named *clam* at 5:25 P.M. on March 3.

Because the **/usr/adm/cluster.log** file is a standard ASCII text file, you can view it using standard AIX file commands, such as the **more** or **tail** commands. However, you can also use the SMIT interface or the HANFS for AIX **cldiag** diagnostic utility. The following sections describe each of the options.

Viewing the cluster.log File Using SMIT

To view the **/usr/adm/cluster.log** file using SMIT:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support** and press Enter.
3. On the RAS Support menu, select **View HANFS Log Files** and press Enter.
4. On the View HANFS Log Files screen, select **Scan the HANFS System Log** and press Enter. This option references the **/usr/adm/cluster.log** file.

Note: You can choose to either *scan* the contents of the **/usr/adm/cluster.log** file as it exists, or you can *watch* an active log file as new events are appended to it in real time. Typically, you *scan* the file to try to find a problem that has already occurred; you *watch* the file as you test a solution to a problem to determine the results.

Viewing cluster.log File Using the cldiag Utility

To view the **/usr/adm/cluster.log** file using the **cldiag** utility, you must include the **/usr/sbin/cluster/diag** directory in your PATH environment variable. Then to run the utility from any directory:

1. Enter the **cldiag** fastpath.

   The utility returns a list of options and the **cldiag** prompt:

```
-------------------------------------------------------
To get help on a specific option, type: help <option>
To return to previous menu, type: back
To quit the program, type: quit
-------------------------------------------------------
```

```
valid options:
debug
logs
vgs
error
trace

cldiag>
```

The **cldiag** utility **help** subcommand provides a brief synopsis of the syntax of the option specified. For more information about the command syntax, see the **cldiag** man page.

2. Enter the **logs** option at the **cldiag** prompt:

   ```
   cldiag> logs
   ```

   The **cldiag** utility displays the following options and prompt. Note that the prompt changes to reflect the last option selection.

   ```
   valid options:
   scripts
   syslog

   cldiag.logs>
   ```

3. To view the **/usr/adm/cluster.log** file, enter:

   ```
   cldiag.logs> syslog
   ```

   By default, the **cldiag** utility displays all messages in the log file for every cluster process on the local node. However, you can optionally view only those messages associated with a specific process or processes.

   To view specific messages, quit the **cldiag** utility and use the **lssrc -g cluster** command at the system prompt to obtain the name of cluster processes. Then restart the **cldiag** utility and specify the name of the process whose messages you want to view. If you want to view more than one process, separate multiple names with spaces.

   For example, to view only those messages generated by the Cluster Manager and Clinfo, specify the names as in the following example:

   ```
   cldiag.logs> syslog clstrmgr clinfo
   ```

   Using flags associated with the **syslog** option, you can specify the types of messages you want to view, the time period covered by the messages, and the file in which you want the messages stored.

   The following table lists the optional command line flags and their functions:

Flag	Function
-h ***hostname***	View messages generated by a particular cluster node.
-e	View only error-level messages.
-w	View only warning-level messages.
-d ***days***	View messages logged during a particular time period. Specify the time period in days.
-R ***filename***	Store the messages in the file specified. By default, the **cldiag** utility writes the messages to **stdout**.

For example, to obtain a listing of all Cluster Manager error-level messages recorded in the last two days and have the listing written to a file named **cm_errors.out**, enter the following command:

```
cldiag logs syslog -d 2 -e -Rcm_errors.out clstrmgr
```

This example illustrates how to execute a **cldiag** function directly without traversing the menu hierarchy.

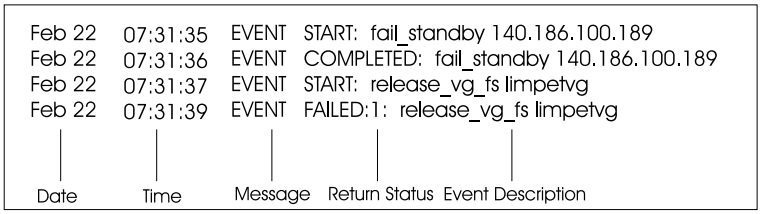
Understanding the hacmp.out Log File

The **/tmp/hacmp.out** file is a standard text file. The system creates a new **hacmp.out** log file every day and retains the last seven copies. Each copy is identified by a number appended to the filename. The most recent log file is named **/tmp/hacmp.out**; the oldest version of the file is named **/tmp/hacmp.out.7**.

When checking the **/tmp/hacmp.out** file, search for EVENT FAILED messages. Starting from the first failure message, read back through the log file to determine exactly what went wrong. The **/tmp/hacmp.out** log file provides the most important source of information when investigating a problem.

Format of Messages in the hacmp.out Log File

In non-verbose mode, the **/tmp/hacmp.out** log contains the start, completion, and error notification messages output by all HANFS for AIX scripts. The following example illustrates the start of the script executed in response to the **node_up** cluster event as it appears in a **/tmp/hacmp.out** file:



Each entry contains the following information:

Date and Time Stamp The day and time on which the event occurred.

Message Text that describes the cluster activity.

Return Status Messages that report failures include the status returned from the script. This information is not included for scripts that complete successfully.

Event Description The specific action attempted or completed on a node, file system, or volume group.

In verbose mode, the **/tmp/hacmp.out** file also includes the values of arguments and flag settings passed to the scripts and commands. These lines are prefixed with a plus sign (+). The following example illustrates the flags and arguments passed to the **release_vg_fs** script in the previous example:

```
Feb 22 07:31:37 EVENT START: release_vg_fs limpetvg

+ [ -n ]
+ [ -n limpetvg ]
+ echo limpetvg
+ awk -F {for(i=1;i<=NF;i++) a[$i]=$i}
END {for(i in a) print a[i]}
+ sort -r
VG=limpetvg
+ /usr/sbin/cluster/events/utils/cl_deactivate_vgs limpetvg
+ set -u
+ [ 1 -ne 0 ]
+ fgrep -s -x limpetvg
+ lsvg -o
+ [ 0 -ne 0 ]
+ varyoffvg limpetvg
0516-012 lvaryoffvg: Logical volume must be closed. If the logical
  volume contains a filesystem, the umount command will close
  the LV device.
0516-942 varyoffvg: Unable to vary off volume group limpetvg.
+ [ 1 -ne 0 ]
+ cl_log 28 /usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of
limpetvg. /usr/sbin/cluster/events/utils/cl_deactivate_vgs limpetvg
+ set -u
+ [ 4 -lt 2 ]
MSG_ID=28
DEFAULT_MSG=/usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of
limpetvg.
+ [ 4 -gt 2 ]
+ shift 2
+ dspmsg scripts.cat 28 /usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed
varyoff of limpetvg. /usr/sbin/cluster/events/utils/cl_deactivate_vgs limpetvg

MSG=/usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of limpetvg.
+ [ /usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of limpetvg. = ]
+ logger -t HACMP /usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of
limpetvg.
+ echo /usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of limpetvg.
/usr/sbin/cluster/events/utils/cl_deactivate_vgs: Failed varyoff of limpetvg.
+ exit 0
STATUS=1
+ exit 1
+ [ 1 -ne 0 ]
STATUS=1
+ exit 1
Feb 25 07:31:52 EVENT FAILED:1: release_vg_fs limpetvg
```

Because the **/tmp/hacmp.out** file is a standard ASCII text file, you can view it using standard AIX file commands, such as the **more** or **tail** commands. However, you can also use the SMIT interface or the HANFS for AIX **cldiag** diagnostic utility. The following sections describe each of the options.

Viewing the hacmp.out File Using SMIT

To view the **/tmp/hacmp.out** file using SMIT:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support** and press Enter.
3. On the RAS Support screen, select **View HANFS Log Files** and press Enter.

   On the View HANFS Log Files menu, you can choose to either *scan* the contents of the **/tmp/hacmp.out** file or *watch* as new events are appended to the log file. Typically, you will scan the file to try to find a problem that has already occurred and then watch the file as you test a solution to the problem. In the menu, the **/tmp/hacmp.out** file is referred to as the HANFS Scripts Log File.
4. Select **Scan the HANFS Script Log File** and press Enter.
5. Select a script log file and press Enter.
6. Press F10 to exit SMIT.

Viewing hacmp.out File Using the cldiag Utility

To view the **/tmp/hacmp.out** file using the **cldiag** utility, you must include the **/usr/sbin/cluster/diag** directory in your PATH environment variable. Then to run the utility from any directory:

1. Enter the **cldiag** fastpath.

   The utility returns a list of options and the **cldiag** prompt:

   ```
   -------------------------------------------------------
   To get help on a specific option, type: help <option>
   To return to previous menu, type: back
   To quit the program, type: quit
   -------------------------------------------------------

   valid options:
   debug
   logs
   vgs
   error
   trace

   cldiag>
   ```

   The **cldiag** utility **help** subcommand provides a brief synopsis of the syntax of the option specified. For more information about the command syntax, see the **cldiag** man page.
2. Enter the **logs** option at the **cldiag** prompt:

   ```
   cldiag> logs
   ```

   The **cldiag** utility displays the following options and prompt. Note that the prompt changes to reflect the current option selection:

   ```
   valid options:
   scripts
   syslog

   cldiag.logs>
   ```
3. To view the **/tmp/hacmp.out** file, enter:

   ```
   cldiag.logs> scripts
   ```

By default, the **cldiag** utility writes the entire contents of **/tmp/hacmp.out** file to **stdout**. However, you can view only messages related to one or more specific events, such as **node_up** or **node_up_local**. Separate multiple events by spaces. The following example views only those messages associated with the **node_up** and **node_up_local** events:

```
cldiag.logs> scripts node_up node_up_local
```

By using flags associated with the **scripts** options, you can specify the types of messages you want to view, the time period covered by the messages, and file in which you want the messages stored. The following table lists the optional command line flags and their functions:

Flag	Function
-h ***hostname***	View messages generated by a particular cluster node. By default, the **scripts** subcommand displays only messages generated by the local node.
-s	View only start and completion messages.
-f	View only failure messages.
-d ***days***	View messages logged during a particular time period. You can specify a time period of up to seven days. (The HANFS for AIX software only keeps the latest seven copies of the **/tmp/hacmp.out** file.) By default, the current day's log, **/tmp/hacmp.out**, is displayed.
-R ***filename***	Store the messages in the file specified. By default, the **cldiag** utility writes the messages to **stdout**.

For example, to obtain a listing of all failure messages associated with the **node_up** event recorded in the last two days, and have the listing written to a file named **script_errors.out**, enter the following:

```
cldiag logs scripts -d 2 -f -R script_errors.out node_up
```

Setting the Level of Information Recorded in the hacmp.out File

To set the level of information recorded in the **/tmp/hacmp.out** file:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **Cluster Configuration** and press Enter.
3. On the Cluster Configuration screen, select **Cluster Resources** and press Enter.
4. On the Cluster Resources screen, select **Change/Show Run Time Parameters** and press Enter. SMIT prompts you to specify the node name of the cluster node you want to modify. (Note that run-time parameters are configured on a per-node basis.)
5. To obtain verbose output, make sure the value of the **Debug Level** field is **high**. If necessary, press Enter to record a new value. The Command Status screen appears.
6. Press F10 to exit SMIT.

Changing the Name or Placement of the hacmp.out Log File

If you want to change the name or placement of the **/tmp/hacmp.out** file, use the following command, where *filename* is a fully qualified pathname:

```
chssys -s clstrmgr -a "-o filename"
```

Restart the cluster to make this change effective. Also, update the LOGFILE environment variable in the **/usr/sbin/cluster/utilities/clcycle** script to reflect the new pathname.

Understanding the System Error Log

The HANFS for AIX software logs script messages to the system error log whenever a script starts, stops, or encounters an error condition, or whenever a daemon generates a state message.

Format of Messages in the System Error Log

The HANFS for AIX messages in the system error log follow the same format as that used by other AIX subsystems. You can view the messages in the system error log in short or long format.

In short format, also called summary format, each message in the system error log occupies a single line. The following figure illustrates the short format of the system error log:

ERROR_ID	TIMESTAMP	T	CL	RESOURCE_NAME	ERROR_DESCRIPTION
DB3E3DFD	0709092293	P	H	ent1	CMSA/CD LAN Communic
ABB81CD5	0709092293	T	H	ent1	COMMUNICATION PROTOCOL
OF27AAE5	0706073993	P	S	SRC	SOFTWARE PROGRAM ERROR
OF27AAE5	0811073993	P	S	SYSPROC	SOFTWARE PROGRAM ABNORMALLY TERMINATED
AA8AB241	0906273935	T	O	clstrmgr	OPERATOR NOTIFICATION
AA8AB241	0906273935	T	O	clstrmgr	OPERATOR NOTIFICATION
AA8AB241	0906273935	T	O	clstrmgr	OPERATOR NOTIFICATION
AA8AB241	0906273935	T	O	clstrmgr	OPERATOR NOTIFICATION
AA8AB241	0906273935	T	O	clstrmgr	OPERATOR NOTIFICATION

Error_ID A unique error identifier.

Timestamp The day and time on which the event occurred.

T Error type: permanent (P), unresolved (U), or temporary (T).

CL Error class: hardware (H), software (S), or informational (O).

Resource_name A text string that identifies the AIX resource or subsystem that generated the message. HANFS for AIX messages are identified by the name of their daemon or script.

Error_description A text string that briefly describes the error.

In long format, a page of formatted information is displayed for each error. See the AIX InfoExplorer facility for a detailed description of this format.

Unlike the HANFS for AIX log files, the **system error log** is not a text file.

Using the AIX Error Report Command

The AIX **errpt** command generates an error report from entries in the system error log. See the **errpt** man page for information on using this command.

Viewing the System Error Log Using SMIT

To view the system error log using SMIT:

1. Enter the **smit problem** fastpath to display the main AIX System Management SMIT screen.
2. On the AIX System Management screen, select **Problem Determination** and press Enter.
3. On the Problem Determination screen, select **Error Log** and press Enter.
4. On the Error Log screen, select **Change / Show Characteristics of the Error Log** and press Enter.
5. Press F10 to exit SMIT.

For more information on this log file, refer to your AIX documentation.

Viewing the System Error Log Using the cldiag Utility

To view the system error log using the **cldiag** utility, you must include the **/usr/sbin/cluster/diag** directory in your PATH environment variable. Then to run the utility from any directory:

1. Enter the **cldiag** fastpath.

   The utility returns a list of options and the **cldiag** prompt:

   ```
   -------------------------------------------------------
   To get help on a specific option, type: help <option>
   To return to previous menu, type: back
   To quit the program, type: quit
   -------------------------------------------------------

   valid options:
   debug
   logs
   vgs
   error
   trace

   cldiag>
   ```

   The **cldiag** utility **help** subcommand provides a brief synopsis of the syntax of the option specified. For more information about command syntax, see the **cldiag** man page.

2. To view the system error log, enter at the **cldiag** prompt the **error** option with the type of error display you want. For example, to view a listing of the system error log in short format, enter the following command:

   ```
   cldiag> error short
   ```

   To obtain a listing of system error log messages in long format, enter the **error** option with the **long** type designation. To view only those messages in the system error log generated by the HANFS for AIX software, enter the **error cluster** option. When you request a listing of cluster error messages, the **cldiag** utility displays system error log messages in short format.

By default, the **cldiag** utility displays the system error log from the local node. Using flags associated with the **error** option, however, you can choose to view the messages for any other cluster node. In addition, you can specify a file into which the **cldiag** utility writes the error log. The following list describes the optional command line flags and their functions:

Flag	Function
-h ***hostname***	View messages generated by a particular cluster node. By default, only messages on the local node are displayed.
-R ***filename***	Store the messages in the file specified. By default, the **cldiag** utility writes the messages to **stdout**.

For example, to obtain a listing of all cluster-related messages in the system error log and have the listing written to a file named **system_errors.out**, enter the following:

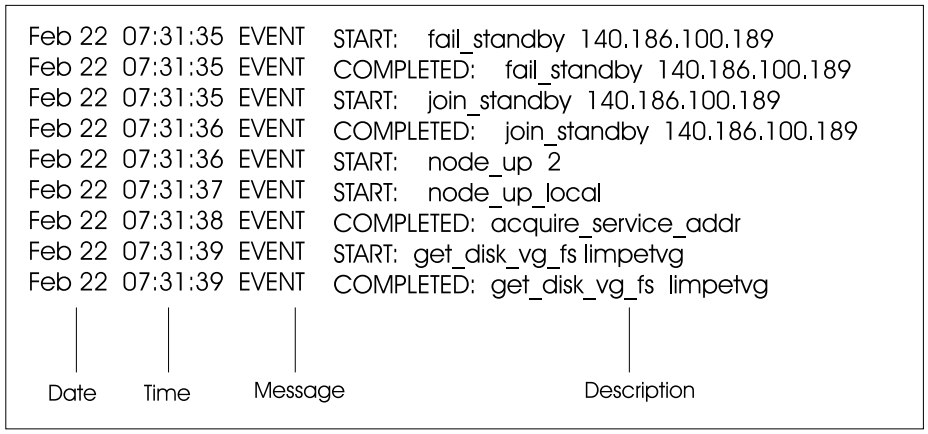
```
cldiag error cluster -R system_errors.out
```

Understanding the Cluster History Log File

The cluster history log file is a standard text file with the system-assigned name **/usr/sbin/cluster/history/cluster.*mmdd***, where *mm* indicates the month and *dd* indicates the day in the month. You should decide how many of these log files you want to retain and purge the excess copies on a regular basis to conserve disk storage space. You may also decide to include the cluster history file in your regular system backup procedures.

Format of Messages in the Cluster History Log File

Entries in the cluster history log file use the following format:



Date and Time Stamp	The date and time at which the event occurred.
Message	Text of the message.
Description	Name of the event script.

Viewing the Cluster History Log File

Because the cluster history log file is a standard text file, you can view its contents using standard AIX file commands, such as **cat**, **more**, and **tail**. You cannot view this log file using SMIT or the **cldiag** utility.

Understanding the /tmp/emuhacmp.out File

The **/tmp/emuhacmp.out** file is a standard text file that resides on the node from which the HANFS Event Emulator was invoked. The file contains information from log files generated by the Event Emulator on all nodes in the cluster. When the emulation is complete, the information in these files is transferred to the **/tmp/emuhacmp.out** file on the node from which the emulation was invoked, and all other files are deleted.

Using the EMUL_OUTPUT environment variable, you can specify another name and location for this output file. The format of the file does not change.

Format of Messages in the /tmp/emuhacmp.out File

The entries in the **/tmp/emuhacmp.log** file use the following format:

```
************************************************************************
******************START OF EMULATION FOR NODE buzzcut***************
************************************************************************

Jul 21 17:17:21 EVENT START: node_down buzzcut graceful

+ [ buzzcut = buzzcut -a graceful = forced ]
+ [ EMUL = EMUL ]
+ cl_echo 3020 NOTICE >>>> The following command was not executed <<<<  \n
NOTICE >>>> The following command was not executed <<<<
+ echo /usr/sbin/cluster/events/utils/cl_ssa_fence down buzzcut\n
/usr/sbin/cluster/events/utils/cl_ssa_fence down buzzcut

+ [ 0 -ne 0 ]
+ [ EMUL = EMUL ]
+ cl_echo 3020 NOTICE >>>> The following command was not executed <<<< \n
NOTICE >>>> The following command was not executed <<<<
+ echo /usr/sbin/cluster/events/utils/cl_9333_fence down buzzcut graceful\n
/usr/sbin/cluster/events/utils/cl_9333_fence down buzzcut graceful

******************** END OF EMULATION FOR NODE BUZZ *****************
```

The output of emulated events is presented as in the **/tmp/hacmp.out** file described earlier in this chapter. The **/tmp/emuhacmp.out** file also contains the following information:

Header — Each node's output begins with a header that signifies the start of the emulation and the node from which the output is received.

Notice — The Notice field identifies the name and path of commands or scripts that are echoed only. If the command being echoed is a customized script, such as a pre- or post-event script, the contents of the script are displayed. Syntax errors in the script are also listed.

ERROR — The error field contains a statement indicating the type of error and the name of the script in which the error was discovered.

Footer Each node's output ends with a footer which signifies the end of the emulation and the node from which the output was received.

Viewing the /tmp/emuhacmp.out File

You can view the **/tmp/emuhacmp.out** file using standard AIX file commands, such as the **more** or **tail** commands. You cannot view this log file using the **cldiag** utility or the SMIT interface. See the **more** or **tail** man pages for information on using these commands.

Tracing HANFS for AIX Daemons

The trace facility helps you isolate a problem within an HANFS for AIX environment by allowing you to monitor selected events. Using the trace facility, you can capture a sequential flow of time-stamped system events that provide a fine level of detail on the activity within an HANFS for AIX cluster.

The trace facility is a low-level debugging tool that augments the troubleshooting facilities described earlier in this book. While tracing is extremely useful for problem determination and analysis, interpreting a trace report typically requires IBM support.

The trace facility generates large amounts of data. The most practical way to use the trace facility is for short periods of time—from a few seconds to a few minutes. This should be ample time to gather sufficient information about the event you are tracking and to limit use of space on your storage device.

The trace facility runs efficiently and has a negligible impact on system performance.

Use the trace facility to track the operation of the following HANFS for AIX daemons:

- The Cluster Manager daemon (**clstrmgr**)
- The Cluster Information Program daemon (**clinfo**)
- The Cluster SMUX Peer daemon (**clsmuxpd**).

The **clstrmgr**, **clinfo**, and **clsmuxpd** daemons are user-level applications under the control of the SRC. Before you can start a trace on one of these daemons, you must first enable tracing for that daemon. *Enabling* tracing on a daemon adds that daemon to the master list of daemons for which you want to record trace data.

You can initiate a trace session using either SMIT or the HANFS for AIX **cldiag** utility. Using SMIT, you can enable tracing in the HANFS for AIX daemons, start and stop a trace session in the daemons, and generate a trace report. Using the **cldiag** utility, you can activate tracing in any HANFS for AIX daemon without having to perform the enabling step. The **cldiag** utility performs the enabling procedure, if necessary, and generates the trace report automatically. The following sections describe how to initiate a trace session using either SMIT or the **cldiag** utility.

Using SMIT to Obtain Trace Information

To initiate a trace session using the SMIT interface:

1. Enable tracing on the daemon or daemons you specify.

   Use the Enable/Disable Tracing of HANFS Daemons screen to indicate that the selected daemons should have trace data recorded for them.

2. Start the trace session.

   Use the Start/Stop/Report Tracing of HANFS Services screen to trigger the collection of data.

3. Stop the trace session.

   You must stop the trace session before you can generate a report. The tracing session stops either when you use the Start/Stop/Report Tracing of HANFS Services screen to stop the tracing session or when the log file becomes full.

4. Generate a trace report.

   Once the trace session is stopped, use the Start/Stop/Report Tracing of HANFS Services screen to generate a report.

Each step is described in the following sections.

Enabling Tracing on HANFS for AIX Daemons

To enable tracing on the following HANFS daemons (**clstrmgr**, **clinfo**, or **clsmuxpd**):

1. Enter the **smit hanfs** fastpath to display the HANFS menu.
2. On the HANFS menu, select **RAS Support** and press Enter.
3. On the RAS Support screen, select **Trace Facility** and press Enter.
4. On the Trace Facility screen, select **Enable/Disable Tracing of HANFS Daemons** and press Enter.
5. On the Trace Subsystem screen, select **Start Trace** and press Enter. SMIT displays the following screen. Note that even though the Start Trace screen appears, you use this screen only to *enable* tracing. Enabling tracing from this screen does not *start* a trace session. Rather, it indicates that you want events related to this particular daemon captured the next time you start a trace session. See Starting a Trace Session on page 16-17 for more information.
6. Enter the PID of the daemon whose trace data you want to capture in the **Subsystem PROCESS ID** field.
   a. Press F4 to see a list of all processes and their PIDs.
   b. Select the daemon and press Enter. Note that you can select *only* one daemon at a time. Repeat these steps for each additional daemon that you want to trace.
7. Indicate whether you want a short or long trace event in the **Trace Type** field.

   A *short* trace contains terse information. For the **clstrmgr** daemon, a short trace produces messages only when topology events occur. A *long* trace contains detailed information on time-stamped events.

8. Press Enter to enable the trace.

SMIT displays a screen indicating that tracing for the specified process is enabled.

Disabling Tracing on HANFS for AIX Daemons

To disable tracing on the **clstrmgr**, **clinfo**, or **clsmuxpd** daemons:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support** and press Enter.
3. On the RAS Support screen, select **Trace Facility** and press Enter.
4. On the Trace Facility screen, select **Enable/Disable Tracing of HANFS Daemons** and press Enter.
5. On the Trace Subsystem screen, select **Stop Trace** and press Enter. SMIT displays the following screen. Note that even though the Stop Trace screen appears, you use this screen only to *disable* tracing. Disabling tracing from this screen does not stop the current trace session. Rather, it indicates that you do not want events related to this particular daemon captured the next time you start a trace session.
6. Enter the PID of the process for which you want to disable tracing in the **Subsystem PROCESS ID** field.
   a. Press F4 to see a list of all processes and their PIDs.
   b. Select the process for which you want to disable tracing and press Enter. Note that you can disable only one daemon at a time. To disable more than one daemon, repeat these steps.
7. Press Enter to disable the trace.

   SMIT displays a screen indicating that tracing for the specified daemon has been disabled.

Starting a Trace Session

Starting a trace session triggers the actual recording of data on system events into the system trace log from which you can later generate a report.

Remember, you can start a trace on the **clstrmgr**, **clinfo**, and **clsmuxpd** daemons only if you have previously enabled tracing for them. You do not need to enable tracing on the **cllockd** daemon; it is a kernel extension.

To start a trace session:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support** and press Enter.
3. On the RAS Support screen, select **Trace Facility** and press Enter.
4. On the Trace Facility screen, select **Start/Stop/Report Tracing of HANFS Services** and press Enter.
5. On the Start/Stop/Report Tracing of HANFS Services screen, select **Start Trace** and press Enter.
6. Enter the trace IDs of the daemons that you want to trace in the **Additional IDs of events to include in trace** field.

Press F4 to see a list of the trace IDs. (Press Ctrl-v to scroll through the list.) Move the cursor to the first daemon whose events you want to trace and press F7 to select it. Repeat this process for each daemon that you want to trace. When you are done, press Enter. The values that you selected are displayed in the **Additional IDs of events to include in trace** field.

The HANFS for AIX daemons have the following trace IDs:

clstrmgr	910
clinfo	911
clsmuxpd	913

7. Enter values as necessary into the remaining fields and press Enter. See the AIX InfoExplorer facility for additional information on this screen.

   SMIT displays a screen indicating that the trace session has started.

Stopping a Trace Session

You need to stop a trace session before you can generate a trace report. A trace session ends when you actively stop it or when the log file is full.

To stop a trace session:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support** and press Enter.
3. On the RAS Support screen, select **Trace Facility** and press Enter.
4. On the Trace Facility screen, select **Start/Stop/Report Tracing of HANFS Services** and press Enter.
5. Select **Stop Trace** and press Enter.

   SMIT displays a screen indicating that the trace session has stopped.

Generating a Trace Report

A trace report formats the information stored in the trace log file and displays it in a readable form. The report displays text and data for each event according to the rules provided in the trace format file.

When you generate a report, you can specify:

- Events to include (or omit)
- The format of the report.

To generate a trace report:

1. Enter the **smit hanfs** fastpath to display the HANFS for AIX menu.
2. On the HANFS for AIX menu, select **RAS Support** and press Enter.
3. On the RAS Support screen, select **Trace Facility** and press Enter.
4. On the Trace Facility screen, select **Start/Stop/Report Tracing of HANFS for AIX Services** and press Enter.

5. On the Trace screen, select **Generate a Trace Report** and press Enter.
6. Indicate the destination and press Enter.
7. Enter the trace IDs of the daemons whose events you want to include in the report in the **IDs of events to INCLUDE in Report** field.

   Press F4 to see a list of the trace IDs. (Press Ctrl-v to scroll through the list.) Move the cursor to the first daemon whose events you want to include in the report and press F7 to select it. Repeat this procedure for each daemon that you want to include in the report. When you finish, press Enter. The values that you selected are displayed in the **IDs of events to INCLUDE in Report** field.

   The HANFS daemons have the following trace IDs:

clstrmgr	910
clinfo	911
clsmuxpd	913

8. Enter values as necessary in the remaining fields and press Enter. See the AIX InfoExplorer facility for additional information on this screen.
9. When the screen is complete, press Enter to generate the report. The output is sent to the specified destination. For an example of a trace report, see Sample Trace Report on page 16-21.

Using the cldiag Utility to Obtain Trace Information

When using the **cldiag** utility, you must include the **/usr/sbin/cluster/diag** directory in your PATH environment variable. Then you can run the utility from any directory. You do not need to enable tracing on any of the HANFS for AIX daemons before starting a trace session.

To start a trace session using the **cldiag** utility:

1. Enter the **cldiag** fastpath.

   The utility returns a list of options and the **cldiag** prompt:

```
-------------------------------------------------------
To get help on a specific option, type: help <option>
To return to previous menu, type: back
To quit the program, type: quit
-------------------------------------------------------

Valid options are:
debug
logs
vgs
error
trace

cldiag>
```

   The **cldiag** utility **help** subcommand provides a brief synopsis of the syntax of the option specified. For more information about the command syntax, see the **cldiag** man page.

2. To activate tracing, enter the **trace** option at the **cldiag** prompt. You must specify (as an argument to the **trace** option) the name of the HANFS for AIX daemons for which you want tracing activated. Use spaces to separate the names of the daemons. For example, to activate tracing in the Cluster Manager and Clinfo daemons, enter the following:

   ```
   cldiag> trace clstrmgr clinfo
   ```

   For a complete list of the HANFS daemons, see Tracing HANFS for AIX Daemons on page 16-15.

   By using flags associated with the trace option, you can specify the duration of the trace session, the level of detail included in the trace (short or long), and the name of a file in which you want the trace report stored. The following table describes the optional command line flags and their functions:

Flag	Function
-l	Obtains a long trace. A *long* trace contains detailed information about specific time-stamped events. By default, the **cldiag** utility performs a short trace. A *short* trace contains terse information. For example, a short trace of the **clstrmgr** daemon generates messages only when topology events occur.
-t *time*	Specifies the duration of the trace session. You specify the time period in seconds. By default, the trace session lasts 30 seconds.
-R *filename*	Stores the messages in the file specified. By default, the **cldiag** utility writes the messages to **stdout**.

   For example, to obtain a 15-second trace of the Cluster Manager daemon and have the trace report written to the file **cm_trace.rpt**, enter:

   ```
   cldiag trace -t 15 -R cm_trace.rpt clstrmgr
   ```

   For an example of the default trace report, see Sample Trace Report on page 16-21.

Sample Trace Report

The sample trace report shown below was obtained by entering the following command:

```
cldiag trace -R clinfo_trace.rpt clinfo
Wed Nov 15 13:01:37 1995
System: AIX steamer Node: 3
Machine: 000040542000
Internet Address: 00000000 0.0.0.0

trace -j 011 -s -a

ID PROCESS NAME I SYSTEM CALL ELAPSED APPL SYSCALL KERNEL INTERRUPT

001 trace 0.000000 TRACE ON channel 0
Fri Mar 10 13:01:38 1995
011 trace 19.569326 HACMP for AIX:clinfo Exiting Function: broadcast_map_request
011 trace 19.569336 HACMP for AIX:clinfo Entering Function: skew_delay
011 trace 19.569351 HACMP for AIX:clinfo Exiting Function: skew_delay, amount:
718650720
011 trace 19.569360 HACMP for AIX:clinfo Exiting Function: service_context
011 trace 19.569368 HACMP for AIX:clinfo Entering Function: dump_valid_nodes
011 trace 19.569380 HACMP for AIX:clinfo Entering Function: dump_valid_nodes
011 trace 19.569387 HACMP for AIX:clinfo Entering Function: dump_valid_nodes
011 trace 19.569394 HACMP for AIX:clinfo Entering Function: dump_valid_nodes
011 trace 19.569402 HACMP for AIX:clinfo Waiting for event
011 trace 22.569933 HACMP for AIX:clinfo Entering Function: service_context
011 trace 22.569995 HACMP for AIX:clinfo Cluster ID: -1
011 trace 22.570075 HACMP for AIX:clinfo Cluster ID: -1
011 trace 22.570087 HACMP for AIX:clinfo Cluster ID: -1
011 trace 22.570097 HACMP for AIX:clinfo Time Expired: -1
011 trace 22.570106 HACMP for AIX:clinfo Entering Function: broadcast_map_request
002 trace 23.575955 TRACE OFF channel 0
 Wed Nov 15 13:02:01 1995
```

Part 6 Appendixes

The appendixes in this part help you plan all aspects of an HANFS for AIX environment, configure serial networks, and install and configure HANFS for AIX on RS/6000 SP systems.

Appendix A Planning Worksheets

This appendix contains the following worksheets:

Worksheet	Purpose	Page
TCP/IP Networks	Use this worksheet to record the TCP/IP network topology for a cluster. Complete one worksheet per cluster.	A-3
TCP/IP Network Adapter	Use this worksheet to record the TCP/IP network adapters connected to each node. You need a separate worksheet for each node defined in the cluster, so begin by photocopying a worksheet for each node and filling in a node name on each worksheet.	A-5
Serial Networks	Use this worksheet to record the serial network topology for a cluster. Complete one worksheet per cluster.	A-7
Serial Network Adapter	Use this worksheet to record the serial network adapters connected to each node. You need a separate worksheet for each node defined in the cluster, so begin by photocopying a worksheet for each node and filling in the node name on each worksheet.	A-9
Shared SCSI-2 Differential or Differential Fast/Wide Disks	Use this worksheet to record the shared SCSI-2 Differential or Differential Fast/Wide disk configuration for the cluster. Complete a separate worksheet for each shared bus.	A-11
Shared IBM SCSI Disk Arrays	Use this worksheet to record the shared IBM SCSI disk array configurations for the cluster. Complete a separate worksheet for each shared SCSI bus.	A-13
Shared IBM 9333 Serial Disk	Use this worksheet to record the IBM 9333 shared disk configuration for the cluster. Complete a separate worksheet for each shared serial bus.	A-15
Shared IBM Serial Storage Architecture Disk Subsystem	Use this worksheet to record the IBM 7131-405 or 7133 SSA shared disk configuration for the cluster.	A-17
Non-Shared Volume Group (Non-Concurrent Access)	Use this worksheet to record the volume groups and file systems that reside on a node's internal disks in a non-concurrent access configuration. You need a separate worksheet for each volume group, so begin by photocopying a worksheet for each volume group and filling in a node name on each worksheet.	A-19
Shared Volume Group/File System (Non-Concurrent Access)	Use this worksheet to record the shared volume groups and file systems in a non-concurrent access configuration. You need a separate worksheet for each shared volume group, so begin by photocopying a worksheet for each volume group and filling in the names of the nodes sharing the volume group on each worksheet.	A-21

Worksheet	Purpose	Page
NFS-Exported File System	Use this worksheet to record the file systems NFS-exported by a node in a non-concurrent access configuration. You need a separate worksheet for each node defined in the cluster, so begin by photocopying a worksheet for each node and filling in a node name on each worksheet.	A-23
Resource Group	Use this worksheet to record the resource groups for a cluster.	A-25
Cluster Event	Use this worksheet to record the planned customization for an HACMP/ES cluster event.	A-27

TCP/IP Networks Worksheet

Cluster ID _____

Cluster Name _____

Network Name	Network Type	Network Attribute	Netmask	Node Names

Sample TCP/IP Networks Worksheet

Cluster ID 1

Cluster Name bivalves

Network Name	Network Type	Network Attribute	Netmask	Node Names
ether1	Ethernet	public	255.255.255.0	clam, mussel, oyster
token1	Token-Ring	public	255.255.255.0	clam, mussel, oyster
fddi1	FDDI	public	255.255.255.0	clam, mussel
socc1	SOCC	private	255.255.255.0	clam, mussel
atm1	ATM	private	255.255.255.0	clam, mussel

TCP/IP Network Adapter Worksheet

Node Name _____

Interface Name	Adapter IP Label	Adapter Function	Adapter IP Address	Network Name	Network Attribute	Boot Address	Adapter HW Address

Note: The SMIT Add an Adapter screen displays an **Adapter Identifier** field that correlates with the **Adapter IP Address** field on this worksheet.

Also, entries in the **Adapter HW Address** field should refer to the locally administered address (LAA), which applies only to the service adapter.

Sample TCP/IP Network Adapter Worksheet

Node Name nodea$_{at0}$

Interface Name	Adapter IP Label	Adapter Function	Adapter IP Address	Network Name	Network Attribute	Boot Address	Adapter HW Address
en0	nodea_en0	service	100.10.1.10	ether1	public		0x08005a7a7610
en0	nodea_boot1	boot	100.10.1.74	ether1	public		
en1	nodea_en1	standby	100.10.11.11	ether1	public		
tr0	nodea_tr0	service	100.10.2.20	token1	public		0x42005aa8b57b
tr0	nodea_boot2	boot	100.10.2.84	token1	public		
fi0	nodea_fi0	service	100.10.3.30	fddi1	public		
sl0	nodea_sl0	service	100.10.5.50	slip1	public		
css0	nodea_svc	service		hps1	private		
css0	nodea_boot3	boot		hps1	private		
at0	nodea_at0	service	100.10.7.10	atm1	private		0x0020481a396500
at0	nodea_boot1	boot	100.10.7.74	atm1	private		

Note: The SMIT Add an Adapter screen displays an **Adapter Identifier** field that correlates with the **Adapter IP Address** field on this worksheet.

Also, entries in the **Adapter HW Address** field should refer to the locally administered address (LAA), which applies only to the service adapter.

Serial Networks Worksheet

Cluster ID _____

Cluster Name _____

Network Name	Network Type	Network Attribute	Node Names
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____
_____	_____	serial	_____

Note: RS232 serial lines, target mode SCSI-2 buses, and tmssa serial links do not use the TCP/IP protocol and do not require a netmask or an IP address.

Sample Serial Networks Worksheet

Cluster ID 1

Cluster Name clus1

Network Name	Network Type	Network Attribute	Node Names
rs232a	RS232	serial	nodea, nodeb
tmscsi1	Target Mode SCSI	serial	nodea, nodeb

Note: RS232 serial lines, target mode SCSI-2 buses, and tmssa serial links do not use the TCP/IP protocol and do not require a netmask or an IP address.

Serial Network Adapter Worksheet

Node Name _____

Slot Number	Interface Name	Adapter Label	Network Name	Network Attribute	Adapter Function
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service
_____	_____	_____	_____	serial	service

Note: Serial networks do not carry TCP/IP traffic; therefore, no boot addresses, adapter identifiers (IP addresses), or adapter hardware addresses are required to maintain keepalives and control messages between nodes.

Sample Serial Network Adapter Worksheet

Node Name nodea

Slot Number	Interface Name	Adapter Label	Network Name	Network Attribute	Adapter Function
SS2	/dev/tty1	nodea_tty1	rs232a	serial	service
08	scsi2	nodea_tmscsi2	tmscsi1	serial	service

Note: Serial networks do not carry TCP/IP traffic; therefore, no boot addresses, adapter identifiers (IP addresses), or adapter hardware addresses are required to maintain keepalives and control messages between nodes.

Shared SCSI-2 Differential or Differential Fast/Wide Disks Worksheet

Note: Complete a separate worksheet for each shared SCSI-2 Differential bus or Differential Fast/Wide bus. Keep in mind that the IBM SCSI-2 Differential High Performance Fast/Wide adapter cannot be assigned SCSI IDs 0, 1, or 2. The SCSI-2 Differential Fast/Wide adapter cannot be assigned SCSI IDs 0 or 1.

Type of SCSI-2 Bus

SCSI-2 Differential _____ **SCSI-2 Differential Fast/Wide** _____

Host and Adapter Information

	Node A	Node B	Node C	Node D
Node Name	_____	_____	_____	_____
Slot Number	_____	_____	_____	_____
Logical Name	_____	_____	_____	_____

SCSI Device IDs on Shared Bus

	Node A	Node B	Node C	Node D
Adapter	_____	_____	_____	_____
First Shared Drive	_____			
Second Shared Drive	_____			
Third Shared Drive	_____			
Fourth Shared Drive	_____			
Fifth Shared Drive	_____			
Sixth Shared Drive	_____			

Shared Drives

Disk	**Size**	**Logical Device Name**			
		Node A	**Node B**	**Node C**	**Node D**
First	_____	_____	_____	_____	_____
Second	_____	_____	_____	_____	_____
Third	_____	_____	_____	_____	_____
Fourth	_____	_____	_____	_____	_____
Fifth	_____	_____	_____	_____	_____
Sixth	_____	_____	_____	_____	_____

Sample Shared SCSI-2 Differential or Differential Fast/Wide Disks Worksheet

Note: Complete a separate worksheet for each shared SCSI-2 Differential bus or Differential Fast/Wide bus. Keep in mind that the IBM SCSI-2 Differential High Performance Fast/Wide adapter cannot be assigned SCSI IDs 0, 1, or 2. The SCSI-2 Differential Fast/Wide adapter cannot be assigned SCSI IDs 0 or 1.

Type of SCSI-2 Bus

SCSI-2 Differential		**SCSI-2 Differential Fast/Wide**	X

Host and Adapter Information

	Node A	Node B	Node C	Node D
Node Name	nodea	nodeb		
Slot Number	7	7		
Logical Name	ascsi1	ascsi1		

SCSI Device IDs on Shared Bus

	Node A	Node B	Node C	Node D
Adapter	6	5		
First Shared Drive	3			
Second Shared Drive	4			
Third Shared Drive	5			
Fourth Shared Drive				
Fifth Shared Drive				
Sixth Shared Drive				

Shared Drives

Disk	**Size**	**Logical Device Name**			
		Node A	**Node B**	**Node C**	**Node D**
First	670	hdisk2	hdisk2		
Second	670	hdisk3	hdisk3		
Third	670	hdisk4	hdisk4		
Fourth					
Fifth					
Sixth					

Shared IBM SCSI Disk Arrays Worksheet

Note: Complete a separate worksheet for each shared SCSI disk array.

Host and Adapter Information

	Node A	Node B	Node C	Node D
Node Name	_____	_____	_____	_____
Slot Number	_____	_____	_____	_____
Logical Name	_____	_____	_____	_____

SCSI Device IDs on Shared Bus

	Node A	Node B	Node C	Node D
Adapter	_____	_____	_____	_____
First Array Controller	_____	_____	_____	_____
Second Array Controller	_____	_____	_____	_____
Third Array Controller	_____	_____	_____	_____
Fourth Array Controller	_____	_____	_____	_____

Shared Drives		Shared LUNs			
Size	**RAID Level**	**Logical Device Name**			
		Node A	**Node B**	**Node C**	**Node D**
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____

Array Controller and Path Information

	Array 1	Array 2
Array Controller Logical Name	_____	_____
Array Controller Logical Name	_____	_____
Disk Array Router Logical Name	_____	_____

Sample Shared IBM SCSI Disk Arrays Worksheet

This sample worksheet shows an IBM 7135 RAIDiant Disk Array configuration.

Note: Complete a separate worksheet for each shared SCSI disk array.

Host and Adapter Information

	Node A	Node B	Node C	Node D
Node Name	nodea	nodeb		
Slot Number	2	2		
Logical Name	ascsi1	ascsi1		

SCSI Device IDs on Shared Bus

	Node A	Node B	Node C	Node D
Adapter	14	15		
First Array Controller	3			
Second Array Controller	4			
Third Array Controller				
Fourth Array Controller				

Shared Drives		Shared LUNs			
Size	**RAID Level**	**Logical Device Name**			
		Node A	**Node B**	**Node C**	**Node D**
2GB	5	hdisk2	hdisk2		
2GB	3	hdisk3	hdisk3		
2GB	5	hdisk4	hdisk4		
2GB	5	hdisk5	hdisk5		

Array Controller and Path Information

	RAIDiant 1
Array Controller Logical Name	dac0
Array Controller Logical Name	dac1
Disk Array Router Logical Name	dar0

Shared IBM 9333 Serial Disk Worksheet

	Node A	Node B	Node C	Node D
Node Name	_____	_____	_____	_____
Slot Number	_____	_____	_____	_____
Logical Name	_____	_____	_____	_____

IBM 9333 Drawer/Desk Label _____

Adapter I/O Connector	_____	_____	_____	_____
Controller Logical Name	_____	_____	_____	_____

IBM 9333 Shared Drives in Node Name _____

Drive	Size (MB)	Logical Device Name			
1	_____	_____	_____	_____	_____
2	_____	_____	_____	_____	_____
3	_____	_____	_____	_____	_____
4	_____	_____	_____	_____	_____

IBM 9333 Drawer/Desk Label _____

Adapter I/O Connector	_____	_____	_____	_____
Controller Logical Name	_____	_____	_____	_____

IBM 9333 Shared Drives in Node Name _____

Drive	Size (MB)	Logical Device Name			
1	_____	_____	_____	_____	_____
2	_____	_____	_____	_____	_____
3	_____	_____	_____	_____	_____
4	_____	_____	_____	_____	_____

Sample IBM 9333 Serial Disk Worksheet

	Node A	Node B	Node C	Node D
Node Name	clam	mussel		
Slot Number	serdasda0	serdasda0		
Logical Name	4	5		

IBM 9333 Drawer/Desk Label drawer1

Adapter I/O Connector	0	1		
Controller Logical Name	serdasdc0	serdasdc0		

IBM 9333 Shared Drives in Node Name _____

Drive	Size (MB)	Logical Device Name			
1	857	hdisk12	hdisk14		
2	857	hdisk13	hdisk15		
3					
4					

IBM 9333 Drawer/Desk Label _____

Adapter I/O Connector	_____	_____	_____	_____
Controller Logical Name	_____	_____	_____	_____

IBM 9333 Shared Drives in Node Name _____

Drive	Size (MB)	Logical Device Name			
1	_____	_____	_____	_____	_____
2	_____	_____	_____	_____	_____
3	_____	_____	_____	_____	_____
4	_____	_____	_____	_____	_____

Shared IBM Serial Storage Architecture Disk Subsystems Worksheet

Host and Adapter Information

	Node A	Node B	Node C	Node D
Node Name	_____	_____	_____	_____
SSA Adapter Label	_____	_____	_____	_____
Slot Number	_____	_____	_____	_____
Dual-Port Number	_____	_____	_____	_____

SSA Logical Disk Drive

Logical Device Name

Node A	Node B	Node C	Node D
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____

SSA Logical Disk Drive

Logical Device Name

Node A	Node B	Node C	Node D
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____

Sample Shared IBM Serial Storage Architecture Disk Subsystems Worksheet

Host and Adapter Information

	Node A	Node B	Node C	Node D
Node Name	clam	mussel		
SSA Adapter Label	ha1, ha2	ha1, ha2		
Slot Number	2, 4	2, 4		
Dual-Port Number	a1, a2	a1, a2		

SSA Logical Disk Drive

	Logical Device Name		
Node A	**Node B**	**Node C**	**Node D**
hdisk2	hdisk2		
hdisk3	hdisk3		
hdisk4	hdisk4		
hdisk5	hdisk5		

SSA Logical Disk Drive

	Logical Device Name		
Node A	**Node B**	**Node C**	**Node D**
hdisk2	hdisk2		
hdisk3	hdisk3		
hdisk4	hdisk4		
hdisk5	hdisk5		

Non-Shared Volume Group Worksheet (Non-Concurrent Access)

Node Name _____

Volume Group Name _____

Physical Volumes _____

_____

_____

Logical Volume Name _____

Number of Copies of Logical Partition _____

On Separate Physical Volumes? _____

File System Mount Point _____

Size (in 512-byte blocks) _____

Logical Volume Name _____

Number of Copies of Logical Partition _____

On Separate Physical Volumes? _____

File System Mount Point _____

Size (in 512-byte blocks) _____

Sample Non-Shared Volume Group Worksheet (Non-Concurrent Access)

Node Name	clam
Volume Group Name	localvg
Physical Volumes	hdisk1
	_____
	_____
Logical Volume Name	locallv
Number of Copies of Logical Partition	1
On Separate Physical Volumes?	no
File System Mount Point	/localfs
Size (in 512-byte blocks)	100000
Logical Volume Name	_____
Number of Copies of Logical Partition	_____
On Separate Physical Volumes?	_____
File System Mount Point	_____
Size (in 512-byte blocks)	_____

Shared Volume Group/File System Worksheet (Non-Concurrent Access)

	Node A	Node B	Node C	Node D
Node Names	_____	_____	_____	_____
Shared Volume Group Name	_____			
Major Number	_____	_____	_____	_____
Log Logical Volume Name	_____			
Physical Volumes	_____	_____	_____	_____
	_____	_____	_____	_____
	_____	_____	_____	_____

Logical Volume Name _____

Number of Copies of Logical Partition _____

On Separate Physical Volumes? _____

File System Mount Point _____

Size (in 512-byte blocks) _____

Logical Volume Name _____

Number of Copies of Logical Partition _____

On Separate Physical Volumes? _____

File System Mount Point _____

Size (in 512-byte blocks) _____

Sample Shared Volume Group/File System Worksheet (Non-Concurrent Access)

	Node A	Node B	Node C	Node D
Node Names	trout	guppy		
Shared Volume Group Name		bassvg		
Major Number	24	24		
Log Logical Volume Name		bassloglv		
Physical Volumes	hdisk6	hdisk6		
	hdisk7	hdisk7		
	hdisk13	hdisk16		

Logical Volume Name	basslv
Number of Copies of Logical Partition	3
On Separate Physical Volumes?	yes
File System Mount Point	/bassfs
Size (in 512-byte blocks)	200000

Logical Volume Name	_____
Number of Copies of Logical Partition	_____
On Separate Physical Volumes?	_____
File System Mount Point	_____
Size (in 512-byte blocks)	_____

NFS-Exported File System Worksheet (Non-Concurrent Access)

Node Name ____________

Exported Directory ____________

Hostname List ____________ ____________ ____________ ____________ ____________

Root Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Client Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Exported Directory ____________

Hostname List ____________ ____________ ____________ ____________ ____________

Root Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Client Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Exported Directory ____________

Hostname List ____________ ____________ ____________ ____________ ____________

Root Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Client Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Exported Directory ____________

Hostname List ____________ ____________ ____________ ____________ ____________

Root Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Client Access ____________ ____________ ____________ ____________ ____________

____________ ____________ ____________ ____________ ____________

Sample NFS-Exported File System Worksheet (Non-Concurrent Access)

Node Name nodea

Exported Directory _____

Hostname List	_____	_____	_____	_____	_____
Root Access	_____	_____	_____	_____	_____
	_____	_____	_____	_____	_____
Client Access	_____	_____	_____	_____	_____
	_____	_____	_____	_____	_____

Resource Group Worksheet

Cluster ID _____

Cluster Name _____

Resource Group Name _____

Node Relationship _____

Participating Node Names _____

File Systems _____

File Systems to Export _____

File Systems to NFS-Mount _____

Volume Groups _____

Raw Disks _____

Application Servers _____

Inactive Takeover _____

Resource Group Name _____

Node Relationship _____

Participating Node Names _____

File Systems _____

File Systems to Export _____

File Systems to NFS-Mount _____

Volume Groups _____

Raw Disks _____

Application Servers _____

Inactive Takeover _____

Sample Resource Group Worksheet

Cluster ID	**1**
Cluster Name	**clus1**
Resource Group Name	rotgrp1
Node Relationship	rotating
Participating Node Names	clam, mussel, oyster
File Systems	/sharedfs1
File Systems to Export	
File Systems to NFS-Mount	/sharedvg1
Volume Groups	
Raw Disks	
Application Servers	imagedemo
Inactive Takeover	false

Resource Group Name	_____
Node Relationship	_____
Participating Node Names	_____
File Systems	_____
File Systems to Export	_____
File Systems to NFS-Mount	_____
Volume Groups	_____
Raw Disks	_____
Application Servers	_____
Inactive Takeover	_____

Cluster Event Worksheet

Note: Use full pathnames for all user-defined scripts.

Cluster ID _____

Cluster Name _____

Cluster Event Name _____

Event Command _____

Notify Command _____

Pre-Event Command _____

Post-Event Command _____

Event Recovery Command _____

Recovery Counter _____

Cluster Event Name _____

Event Command _____

Notify Command _____

Pre-Event Command _____

Post-Event Command _____

Event Recovery Command _____

Recovery Counter _____

Sample Cluster Event Worksheet

Note: Use full pathnames for all user-defined scripts.

Cluster ID	1
Cluster Name	bivalves
Cluster Event Name	node_down_complete
Event Command	
Notify Command	
Pre-Event Command	
Post-Event Command	/usr/local/wakeup
Event Recovery Command	
Recovery Counter	

Cluster Event Name	_____
Event Command	_____
Notify Command	_____
Pre-Event Command	_____
Post-Event Command	_____
Event Recovery Command	_____
Recovery Counter	_____

Appendix B Configuring Serial Networks

This appendix describes how to configure either a SCSI-2 bus or a raw RS232 serial line as a serial network in an HANFS for AIX cluster. See Chapter 3, Planning HANFS for AIX Networks, for more information on serial networks.

Supported Serial Networks

The HANFS for AIX software supports two types of serial networks: the SCSI-2 Differential bus (using target mode SCSI), the raw RS232 serial line, and a target mode SSA loop.

Target Mode SCSI

You can configure a SCSI-2 bus as an HANFS for AIX serial network only if you are using SCSI-2 Differential devices that support target mode SCSI. SCSI-1 Single-Ended and SCSI-2 Single-Ended devices do not support serial networks in an HANFS for AIX cluster. The advantage of using the SCSI-2 Differential bus is that it eliminates the need for a dedicated serial port at each end of the connection, and for associated RS232 cables. See Configuring Target Mode SCSI Connections on page B-2 for more information.

RS232 Serial Line

If you are using shared disk devices other than SCSI-2 Differential devices, you must use a raw RS232 serial line as the serial network. Note that each point-to-point RS232 serial network requires a dedicated serial port at each end. See Configuring RS232 Serial Lines on page B-6 for more information.

Note: The 7013-S70, 7015-S70, and 7017-S70 systems do not support the use of the native serial ports in an HACMP RS232 serial network. To configure an RS232 serial network in an S70 system, you must use a PCI multi-port ASync card.

Target Mode SSA

You can configure a target mode SSA connection between nodes sharing disks connected to SSA on Multi-Initiator RAID adapters (FC 6215 and FC 6219). The adapters must be at Microcode Level 1801 or later.

You can define a serial network to HANFS that connects all nodes on an SSA loop.

Configuring Target Mode SCSI Connections

This section describes how to configure a target mode SCSI-2 Differential bus as a serial network. The SCSI disks must be installed, cabled to the processors, and powered on.

Checking the Status of SCSI Adapters and Disks

To define a target mode SCSI connection, each SCSI adapter (controller) on nodes that will share disks on the SCSI bus must have a unique ID and must be "Defined," known to the system but not yet available. Additionally, all disks assigned to an adapter must also be "Defined" but not yet available.

Note: The uniqueness of adapter SCSI IDs ensures that tmscsi devices created on a given node do not reflect the SCSI IDs of adapters on other nodes connected to the same bus.

To check the status of SCSI adapters you intend to use, enter:

```
lsdev -C | grep scsi
```

If an adapter is "Defined," see Enabling Target Mode SCSI Devices in AIX on page B-3 to configure the target mode connection.

To check the status of SCSI disks on the SCSI bus, enter:

```
lsdev -Cc disk
```

Returning Adapters and Disks to a Defined State

If either an adapter or a disk is "Available," follow these steps to return both the adapter (and its disks) to a defined state so that the adapters can be configured for target mode SCSI and made available:

1. Use the following command to make "Defined" each available disk associated with an adapter:

   ```
   rmdev -l hdiskx
   ```
   where *hdiskx* is the hdisk to be made "Defined."

   For example:

   ```
   rmdev -l hdisk3
   ```

2. Run the following command to return the SCSI adapter to a "Defined" state:

   ```
   rmdev -l scsix
   ```
   where *scsix* is the adapter to be made "Defined".

3. *If using an array controller,* run the **rmdev** command to make a disk and a controller "Defined," as follows:

   ```
   rmdev -l darx
   rmdev -l dacx
   ```

When all controllers and disks are "Defined," see the following section to enable the target mode connection.

Note: Target mode SCSI is automatically configured if you are using the SCSI-2 Differential Fast/Wide Adapter/A. Skip ahead to Defining the Target Mode Connection as a Serial Network on page B-5.

Enabling Target Mode SCSI Devices in AIX

The general steps in defining a target mode SCSI device are:

Step	What you do...
1	Enable the target mode interface for the SCSI adapter.
2	Configure (make available) the devices.

You need to complete both steps on one node, then on the remaining nodes.

Enabling the Target Mode Interface

To enable the target mode interface:

1. Enter the **smit devices** fastpath to display a list of devices.
2. Select **SCSI Adapter > Change/Show Characteristics of a SCSI Adapter**.

   SMIT prompts you to identify the SCSI adapter.
3. Select the appropriate adapter and press Enter to display the Change/Show Characteristics of a SCSI Adapter screen.
4. Set the **Enable TARGET MODE interface** field to **yes** to enable the target mode interface on the device (the default value is **no**).

   At this point, a target mode SCSI device is generated that points to the other cluster nodes that share the SCSI bus. For example, given that Node A and Node B have SCSI IDs of 6 and 7, respectively, running the **lsdev -Cc tmscsi** command on either node returns the following output:

   On Node A:

   ```
   lsdev -Cc tmscsi

   tmscsi0 Available 00-04-00-05 SCSI I/O Controller Initiator Device
   tmscsi1 Available 00-04-00-07 SCSI I/O Controller Initiator Device
   ```

   On Node B:

   ```
   lsdev -Cc tmscsi

   tmscsi0 Available 00-04-00-06 SCSI I/O Controller Initiator Device
   tmscsi1 Available 00-04-00-05 SCSI I/O Controller Initiator Device
   ```

   Note: The adapter SCSI ID on the node from which you enabled the interface will not be listed.
5. Press Enter to commit the value.
6. Press F10 to exit SMIT.

Configuring the Target Mode SCSI Device

After enabling the target mode interface, you must run **cfgmgr** to create the initiator and target devices and make them available:

1. Enter the **smit devices** fastpath to display a list of devices.
2. Select **Install/Configure Devices Added After IPL** and press Enter twice.
3. Press F10 to exit SMIT after the **cfgmgr** command completes.
4. Run the following command to ensure that the devices are paired correctly:

   ```
   lsdev -Cc tmsci
   ```

Repeat the above procedure for the other node to be connected to the SCSI-2 bus.

Target Mode Files

After you have configured the target mode connection on each adapter, two special files for each target mode interface exist in the **/dev** directory: the **/dev/tmscsi*xx*.im** and **/dev/tmscsi*xx*.tm** files, where ***xx*** is a device number the system assigns sequentially for each tmscsi connection. For example:

```
/dev/tmscsi0.im, /dev/tmscsi0.tm
/dev/tmscsi1.im, /dev/tmscsi1.tm
```

The file with the **.im** extension is the initiator, which transmits data. The file with the **.tm** extension is the target, which receives data.

Testing the Target Mode Connection

For the target mode connection to work, initiator and target devices must be paired correctly. To ensure that devices are paired and that the connection is working after enabling the target mode connection on both nodes:

1. Enter the following command on one node connected to the bus:

   ```
   cat < /dev/tmscsixx.tm
   ```

2. Enter the following command on the other node connected to the bus:

   ```
   cat filename > /dev/tmscsixx.im
   ```

   where *filename* is a file. The contents of the specified file are displayed on the node on which you entered the first command above; however, the tmscsi device numbers will not necessarily be the same. You must rely on the **lsdev -Cc tmscsi** output to determine the pair connection.

Note: After a system reboot, the first execution of the **cat** command will not work. Instead, the target device receives a unit attention that notifies the initiator that the device has been reset. This notification puts the devices in sync. Afterwards, the **cat** command will work.

Note: If the SCSI bus is disconnected while running as a target mode SCSI network, the network will not properly reintegrate when the bus is reconnected. HANFS for AIX should be shut down on a node that has had the SCSI bus detached from it before reattaching the SCSI bus to that node.

Defining the Target Mode Connection as a Serial Network

After configuring and testing the target mode connection, you must define it as a serial network to the HANFS for AIX environment. The following steps describe how to use the **Configure Adapters** option on the Cluster Topology screen to define a target mode connection in the HANFS for AIX cluster.

To define the target mode connection (on each adapter) as a serial network to the HANFS for AIX software:

1. From the Cluster Topology menu, select **Configure Adapters > Add an Adapter**.

   The Add an Adapter screen appears.

2. Enter field values as follows:

Adapter Label	Enter the name of the target mode adapter. The label for a target mode adapter must be unique. For example, *clam_tmscsi2*.
Network Type	Pick **tmscsi** on the pop-up pick list. Use the **lsdev -Cc tmscsi** command to identify the proper target mode SCSI device number, which is not necessarily the same number at both ends. You must define both ends of the **tmscsi** network. For example, define tmscsi2 to Node A, with the tmscsi device pointing to Node B, and vice versa.
Network Name	Enter the name of the network that refers to the target mode connection. This name is arbitrary but must be used consistently. For example, *tmscsi2*.
Network Attribute	Set this field to **serial**.
Adapter Function	Set this field to **service**.
Adapter Identifier	Enter the device name: **/dev/tmscsi*x***. The device number *x* should match the number entered for the adapter label.
Adapter Hardware Address	Leave this field blank.
Node Name	Enter the name of the node connected to the adapter.

3. Press Enter. The system adds these values to the HANFS for AIX ODM and returns you to the Configure Adapters menu.

   Note that when you run the **clverify** utility, it checks to see that defined **/dev/tmscsi*x*** devices exist in the ODM for all devices defined to the HANFS for AIX environment.

Repeat the above procedure to define the other adapter connected by the SCSI-2 bus.

Configuring RS232 Serial Lines

This section describes how to configure an RS232 serial line as a serial network in an HANFS for AIX cluster. Before configuring the RS232 serial line, however, you must have physically installed the line between the two nodes. The HANFS for AIX serial line (a 25-pin null modem, serial-to-serial cable) can be used to connect the nodes. The cable is available in the following lengths:

- 3.7 meter serial-to-serial port cable (FC3124)
- 8 meter serial-to-serial port cable (FC3125).

Checking the Status of the Serial Ports

After you have installed the RS232 serial line, use the **lsdev** command to check the status of each serial port you intend to use:

```
lsdev -Cc tty
```

If the tty device is neither defined nor available, use the **smit tty** fastpath to define the device as described in the following section.

If the tty device is defined but not available, or if you have questions about its settings, use the **rmdev** command to delete the tty device:

```
rmdev -l ttyx -d
```

where *ttyx* is replaced by the targeted tty device (for example, *tty1*). Then use the **smit tty** fastpath to define the device. See the following section.

Removing and then defining the tty device makes it available with the default settings appropriate for the communication test described in Testing the Serial Connection on page B-7.

Defining the tty Device

To create a tty device on each node to be connected to the RS232 line:

1. Enter the **smit tty** fastpath to display the TTY screen.
2. Select **Add a TTY** and press Enter. SMIT prompts you for a tty type.
3. Select **tty rs232 Asynchronous Terminal** and press Enter. SMIT prompts you to identify the parent adapter.
4. Select the parent adapter and press Enter. The parent adapter you select is the adapter to which the RS232 cable is connected.
5. Enter field values as follows:

   PORT number — Press F4 to list the available port numbers. Select the appropriate port number and press Enter. The port that you select is the port to which the RS232 cable is connected.

   ENABLE login — Make sure this field is set to **disable** to prevent any **getty** processes from spawning on this device.

6. Press Enter to commit the values.
7. Press F10 to exit SMIT.

Repeat this procedure for the other node that will be connected to the RS232 line.

Testing the Serial Connection

To test communication over the serial line after creating the tty device on both nodes:

1. On the first node, enter the following command:

   `stty < /dev/`*`ttyx`*
   where *ttyx* is the newly added tty device. The command line on the first node should hang.

2. On the second node, enter the following command:

   `stty < /dev/`*`ttyx`*
   where *ttyx* is the newly added tty device.

If the nodes are able to communicate over the serial line, both nodes display their tty settings and return to the prompt.

Note: This is a valid communication test of a newly added serial connection before the **/usr/sbin/cluster/clstrmgr** daemon has been started. This test yields different results after the **/usr/sbin/cluster/clstrmgr** daemon starts, since this daemon changes the initial settings of the tty devices and applies its own settings.

Defining the RS232 Serial Line as a Serial Network

After you have installed and tested the RS232 serial line, define it as a serial network to the HANFS for AIX cluster. The following steps describe how to use the Add an Adapter screen to define an RS232 serial line to the HANFS for AIX environment.

To associate a network adapter with a cluster node:

1. From the Cluster Topology menu, select **Configure Adapters > Add an Adapter**.

   The Add an Adapter screen appears.

2. Enter field values as follows:

Adapter Label	Enter the name of the serial adapter. The label for a serial adapter must be unique (for example, *caviar_tty1*). Also, make sure that the serial adapters connected to the same RS232 line have different names.
Network Type	Pick the type **RS232** from the pop-up pick list.
Network Name	Enter the name of the network connected to this adapter. Refer to the Serial Network Adapter Worksheet.
Network Attribute	Set this field to **serial**.
Adapter Function	Set this field to **service**.
Adapter Identifier	Enter the full pathname of the tty device (for example, */dev/tty1*).
Adapter Hardware Address	Leave this field blank.
Node Name	Enter the name of the node connected to this adapter.

3. Press Enter.

   The system adds these values to the HANFS for AIX ODM and displays the Configure Adapters menu.

Repeat this procedure to define the other adapter connected by the RS232 line.

Configuring Target Mode SSA Connections

This section describes how to configure a target mode SSA connection between nodes sharing disks connected to SSA on Multi-Initiator RAID adapters (FC 6215 and FC 6219). The adapters must be at Microcode Level 1801 or later.

You can define a serial network to HANFS that connects all nodes on an SSA loop.

Changing Node Numbers on Systems in an SSA Loop

By default, node numbers on all systems are zero. In order to configure the target mode devices, you must first assign a unique node number to all systems on the SSA loop.

To change the node number use the following command:

```
chdev -l ssar -a node_number=#
```

To show the system's node number use the following command:

```
lsattr -El ssar
```

Configuring Target Mode SSA Devices

After enabling the target mode interface, you must run **cfgmgr** to create the initiator and target devices and make them available.

To configure the devices and make them available:

1. Enter:

   ```
   smit devices
   ```

   SMIT displays a list of devices.
2. Select **Install/Configure Devices Added After IPL** and press Enter.
3. Press F10 to exit SMIT after the **cfgmgr** command completes.
4. Enter the following command to ensure that the devices are paired correctly:

   ```
   lsdev -Cc tmssa
   ```

Repeat the above procedure (enabling and configuring the target mode SSA device) for other nodes connected to the SSA adapters.

Target Mode Files

Configuring the target mode connection creates two special files in the **/dev** directory of each node, the **/dev/tmssa#.im** and **/dev/tmssa#.tm** files. The file with the **.im** extension is the initiator, which transmits data. The file with the **.tm** extension is the target, which receives data.

Testing the Target Mode Connection

For the target mode connection to work, initiator and target devices must be paired correctly. To ensure that devices are paired and that the connection is working after enabling the target mode connection on both nodes:

1. Enter the following command on a node connected to the SSA disks:

   ```
   cat < /dev/tmssa#.tm
   ```

   where # must be the number of the target node. (This command hangs and waits for the next command.)

2. On the target node, enter the following command:

   ```
   cat filename > /dev/tmssa#.im
   ```

   where # must be the number of the sending node and *filename* is a file.

   The contents of the specified file are displayed on the node on which you entered the first command.

3. You can also check that the tmssa devices are available on each system using the following command:

   ```
   lsdev -C | grep tmssa
   ```

Defining the Target Mode SSA Serial Network to HANFS

Take the following steps to configure the target mode SSA serial network in the HANFS cluster:

1. From the Cluster Topology menu, select **Configure Adapters > Add an Adapter**.

   Enter the fields as follows.

Adapter Label	Unique label for adapter. For example, *adp_tmssa_1*.
Network Type	Pick **tmssa** from the pop-up pick list.
Network Name	Arbitrary name used consistently for all adapters on this network. For example, *tmssa_1*.
Network Attribute	Set this field to **serial**.
Adapter Function	Set this field to **service**.
Adapter Identifier	Enter the device name */dev/tmssa#*.
Adapter Hardware Address	Leave this field blank.
Node Name	Enter the name of the node the adapter is connected to.

2. Press Enter. The system adds these values to the HANFS for AIX ODM and returns you to the Configure Adapters menu.

Repeat this procedure to define the other adapters connected on the SSA loop.

Appendix C Installing and Configuring HANFS for AIX on RS/6000 SPs

This appendix describes installation and configuration considerations for using HANFS for AIX, Version 4.3.1 software on RS/6000 SP Systems.

Overview

The HANFS for AIX software provides high-availability functions on RS/6000-based products, including the SP platform. Before attempting to install and configure HANFS for AIX on the SP, you should be familiar with manuals in the SP documentation set.

Related Publications

The following publications provide additional information about SP systems:

- *SP Site Planning*
- *SP Library Guide*
- *SP Administration Guide*
- *SP Installation and Migration Guide*
- *SP Diagnosis and Messages Guide*
- *SP Command and Technical Reference*
- *SP Parallel System Support Programs*
- *SP System Planning Guide*
- *SP MIM Volume 1 - Installation*
- *SP MIM Volume 2 - MAPs and Part*
- *SP Installation Aid LPS*

You can find more information about these books in an online library of documentation covering AIX, RS/6000, and related products on the World Wide Web. Enter the following URL:

`http://www.rs6000.ibm.com/aix/library`

Installing HANFS for AIX on an SP System

Although there are no unique HANFS for AIX filesets (install images) for the SP, refer to the instructions in Chapter 11, Installing HANFS for AIX Software, to determine which HANFS for AIX filesets you need.

Additionally, the following software must be installed on the SP control workstation and nodes:

- RS/6000 SP version 3, release 1.0, (or higher) of the AIX Parallel System Support Programs (PSSP).
- Latest service level of the PSSP software you plan to install on your system.

You must install the HANFS for AIX software on all nodes of the SP system that will form an HANFS cluster, and the HANFS for AIX client image on the control workstation, if the workstation will be used as a client to monitor cluster status.

Once the HANFS for AIX install images are available to each node and to the control workstation (either via NFS mount or a local copy), log onto each node and install the HANFS for AIX software by following the instructions in Chapter 11, Installing HANFS for AIX Software. After installing the software, read the HANFS for AIX release notes in the **/usr/lpp/cluster.hanfs/doc** directory.

Assuming all necessary HANFS for AIX filesets are in the **/spdata/sys1/install/lppsource** directory on the SP control workstation (and that you have a **.toc** file created in this directory), perform the following procedure on the control workstation:

Note: If you do not have a **.toc** file in the **/spdata/sys1/install/lppsource** directory, enter either the **inutoc** command or the following command to create it:

```
installp -ld/spdata/sys1/install/lppsource
```

1. Create a file called **/HANFSHOSTS** that contains hostnames of nodes in the SP frame that will have the HANFS for AIX software installed.
2. Export the Working Collective (WCOLL) environment variable using the following command**:**

   ```
   export WCOLL=/HANFSHOSTS
   ```
3. Ensure that all hosts listed in the **/HANFSHOSTS** file are up (that is, each host responds) by entering the following command:

   ```
   /usr/lpp/spp/bin/SDRGetObjects host_responds
   ```
   where **SDRGetObjects** is an SP command that retrieves information from the SP System Data Repository to the SP database. A host response of 1 indicates that the node on which the HANFS for AIX software is installed and responding properly to the **HATS** daemon.
4. Enter the following command to mount the file system (from the control workstation) onto all nodes:

   ```
   dsh -ia /etc/mount CWNAME:/usr/sys/inst.images/ssp /mnt
   ```
   where *CWNAME* is the hostname of the control workstation.
5. Enter the following command to install the HANFS for AIX software on the nodes:

   ```
   dsh -ia "/etc/installp -Xagd /mnt LPP_NAME"
   ```

where *LPP_NAME* is the name of the product/fileset you need to install. This must be done for each fileset you need to install. Note that you can install "all" filesets.

6. Enter the following command on the control workstation:

   ```
   dsh -ia "/etc/umount cwname:/usr/sys/inst.images/ssp /mnt"
   ```

7. Enter the following command to verify that HANFS was successfully installed on each node:

   ```
   dsh -ia "/etc/installp -s | grep cluster"
   ```

8. Are you upgrading from HANFS version 4.2.2 with this installation?

   If *no*, go to step 9.

   If *yes*, you need to uninstall the **cluster.cspoc** filesets, as the C-SPOC facility is no longer supported in HANFS for AIX. To uninstall the filesets, type:

   ```
   smit remove
   ```
   and enter the following field values:

SOFTWARE NAME	Enter **cluster.cspoc***. Be sure to include the asterisk (*) so that all the C-SPOC filesets are removed.
Preview?	Change the value to **no**.

9. Reboot the nodes on which you installed the HANFS for AIX software.

Configuring the RS/6000 SP for HANFS for AIX

The following sections describe SP system changes that should be done for HANFS for AIX. Consult the *SP Installation Guide* and the *SP Administration Guide* for more information about SP management.

Network Options

Consult the *SP Administration Guide* and add the proper network options in the **tuning.cust** file on all SP HANFS for AIX nodes for the SP Switch network. Here's an example of the options for the SP nodes with HANFS for AIX:

```
no -o ipforwarding=0
no -o ipsendredirects=0
no -o thewall=16834
no -o routerevalidate=1
```

Consult the *SP Administration Guide* for more information about changing other network options for maximizing performance based on your expected SP worktype and workload.

Configuring Cluster Security

Kerberos is a network authentication protocol used on the SP. Based on a secret-key encryption scheme, Kerberos offers a secure authentication mechanism for client/server applications.

By centralizing command authority via one authentication server, normally configured to be the SP control workstation, Kerberos eliminates the need for the traditional TCP/IP access control lists (**.rhosts** files) that were used in earlier HANFS security implementations. Rather than storing hostnames in a file (the **.rhosts** approach), Kerberos issues dually encrypted

authentication tickets. Each ticket contains two encryption keys: One key is known to both the client user and to the ticket-granting service, and one key is known to both the ticket-granting service and to the target service that the client user wants to access. By setting up all network IP labels in your HANFS configuration to use Kerberos authentication, you reduce the possibility of a single point of failure.

For a more detailed explanation of Kerberos and the security features of the SP system, refer to the *IBM Parallel System Support Programs for AIX Administration Guide*.

To configure Kerberos on the SPs within an HANFS cluster, you must perform these general steps (detailed procedures appear in the following sections):

Step	What you do...
1	Make sure that HANFS has been properly installed on all nodes in the cluster. For more information, see Chapter 11, Installing HANFS for AIX Software.
2	Configure the HANFS cluster topology information on one node in the cluster. Be sure to include the SP Ethernet as part of the configuration. Note that on the SP **setup_authent** is usually used to configure Kerberos on the entire SP system. **setup_authent** creates **rcmd** (used for **rsh** and **rcp**) service principals for all network IP labels listed in the System Data Repository (SDR). The SDR does not allow multiple IP labels to be defined on the same interface. However, HANFS requires that multiple IP labels be defined for the same interface during IPAT configurations. HANFS also requires that **godm** (Global ODM) service principals be configured on all IP labels for remote ODM operations. For these reasons, each time the nodes are customized after the SP **setup_authent** script is run (via **setup_server** or alone), you must manually reconfigure the systems to use Kerberos.
3	Create new Kerberos service principals and configure all IP labels for Kerberos authentication (see Configuring Kerberos Manually on page C-5).
4	Set the cluster security mode to **Enhanced**, then synchronize the cluster topology. See Setting a Cluster's Security Mode on page C-7.
5	Delete (or at least edit) the **cl_krb_service** file, which contains the Kerberos service principals password you entered during the configuration process. At the very least, you should edit this file to prevent unauthorized users from obtaining the password and possibly changing the service principals.
6	Consider removing unnecessary **.rhosts** files. With Kerberos configured, HANFS does not require the traditional TCP/IP access control lists provided by these files (but other applications might). You should consult your cluster administrator before removing any version of this file.

Configuring Kerberos Manually

To properly configure Kerberos on all HANFS-configured networks, you must perform the following general steps:

Step	What you do...
1	Add an entry for each new Kerberos service principal to the Kerberos Authentication Database. See Adding New Service Principals to the Authentication Database on page C-5.
2	Update the **krb-srvtab** file by extracting each newly added instance from the Kerberos Authentication Database. See Updating the krb-srvtab File on page C-6.
3	Add the new service principals to each node's **/.klogin** file. See Adding Kerberos Principals to Each Node's .klogin File on page C-6.
4	Add the new service principals to each node's **/etc/krb.realms** file. See Adding Kerberos Principals to Each Node's /etc/krb.realms File on page C-7.

Adding New Service Principals to the Authentication Database

To add new service principals to the Kerberos Authentication Database for each network interface:

1. On the control workstation, start the **kadmin** utility

   ```
   kadmin
   ```

   A welcome message appears.

2. At the `admin:` prompt type the **add_new_key** command with the name and instance of the new principal:

   ```
   admin: ank service_name.instance
   ```

   where

   service_name is the service (**godm** or **rcmd**) and *instance* is the address label to be associated with the service. Thus, using the service **godm** and address label **i1_sw** the command is:

   ```
   admin: ank godm.i1_sw
   ```

3. When prompted, enter the Kerberos Administration Password.

   ```
   Admin password: password
   ```

4. When prompted, enter a Kerberos password for the new principal.

   ```
   Password for service_name.instance: password
   ```

   Note: The password can be the same as the Kerberos Administration Password, but doesn't have to be. Follow your site's password security procedures.

5. Verify that you have indeed added the new principals to the Kerberos database.

   ```
   kdb_util dump /tmp/testdb
   cat /tmp/testdb
   ```

   Remove this copy of the database when you have finished examining it.

   ```
   rm /tmp/testdb
   ```

Updating the krb-srvtab File

To update the **krb-srvtab** file and propagate new service principals to the HANFS cluster nodes:

1. Extract each new service principal for each instance you added to the Kerberos Authentication Database for those nodes you want to update. (This operation creates a new file in the current directory for each instance extracted.)

   ```
   usr/lpp/ssp/kerberos/etc/ext_srvtab -n i1_sw i1_en i1_tr
   ```

2. Combine these new files generated by the **ext_srvtab** utility into one file called *node_name-new-srvtab*:

   ```
   cat i1_sw-new-srvtab i1_en-new-srvtab i1_tr-new-srvtab
   > node_name-new-srvtab
   ```

   The new file appears in the directory where you typed the command.

   Note: Shared labels (used for rotating resource groups) need to be included in every **krb-srvtab** file (for nodes in that rotating resource group), so you must concatenate each shared-label srvtab file into each *node_name-new-srvtab* file.

3. Copy each *node_name-new-srvtab* file to its respective node.

4. Make a copy of the current **/etc/krb-srvtab** file so that it can be reused later if necessary:

   ```
   cp /etc/krb-srvtab /etc/krb-srvtab-date
   ```

   (where *date* is the date you made the copy).

5. Replace the current **krb-srvtab** file with the new *node_name-new-srvtab* file:

   ```
   cp node_name-new-srvtab /etc/krb-srvtab
   ```

6. Verify that the target node recognizes the new principals by issuing the following command on it:

   ```
   ksrvutil list
   ```

   You should see all the new principals for each network interface on that node; if not, repeat this procedure.

Adding Kerberos Principals to Each Node's .klogin File

To add the new Kerberos principals to the **/.klogin** file on each HANFS cluster node:

1. Edit the **/.klogin** file on the control workstation to add the principals that were created for each network instance:

   ```
   vi /.klogin
   ```

   Here is an example of the **/.klogin** file for two nodes, i and j. ELVIS_IMP is the name of the realm that will be used to authenticate service requests. Each node has the SP Ethernet, a Token Ring service, and an Ethernet service adapter.

   ```
   root.admin@ELVIS_IMP
   rcmd.i1@ELVIS_IMP
   rcmd.i1_ensvc@ELVIS_IMP
   rcmd.i1_trsvc@ELVIS_IMP
   rcmd.j1@ELVIS_IMP
   rcmd.j1_ensvc@ELVIS_IMP
   rcmd.j1_trsvc@ELVIS_IMP
   godm.i1@ELVIS_IMP
   godm.i1_ensvc@ELVIS_IMP
   godm.i1_trsvc@ELVIS_IMP
   godm.j1@ELVIS_IMP
   ```

```
godm.j1_ensvc@ELVIS_IMP
godm.j1_trsvc@ELVIS_IMP
```

2. Copy the **/.klogin** file from the control workstation to each node in the cluster.

To verify that you set this up correctly, issue a Kerberized **rsh** command on all nodes using one of the newly defined interfaces. For example:

```
/usr/lpp/ssp/rcmd/bin/rsh i1_ensvc date
```

To eliminate single points of failure, you should add Kerberos **rcmd** and **godm** principals for every interface configured in HANFS.

Adding Kerberos Principals to Each Node's /etc/krb.realms File

To add the new Kerberos principals to the **/etc/krb.realms** file on each HANFS cluster node:

1. Edit the **/etc/krb.realms** file on the control workstation and add the principals that were created for each network instance.

   ```
   vi /etc/krb.realms
   ```

   Here is an example of the **krb.realms** file for two nodes, i and j. ELVIS_IMP is the name of the realm that will be used to authenticate service requests. Each node has the SP Ethernet, a Token-Ring service, and an Ethernet service adapter.

   ```
   root.admin ELVIS_IMPHANFS
   i1 ELVIS_IMP
   i1_ensvc ELVIS_IMP
   i1_trsvc ELVIS_IMP
   j1 ELVIS_IMP
   j1_ensvc ELVIS_IMP
   j1_trsvc ELVIS_IMP
   i1 ELVIS_IMP
   i1_ensvc ELVIS_IMP
   i1_trsvc ELVIS_IMP
   j1 ELVIS_IMP
   j1_ensvc ELVIS_IMP
   j1_trsvc ELVIS_IMP
   ```

2. Copy the **/etc/krb.realms** file from the control workstation to each node in the cluster.

Setting a Cluster's Security Mode

To change an entire cluster's security mode, you first need to change the security settings for each node, then synchronize the cluster topology from one node. Here's an example of the process, using a two-node cluster:

Step	What you do...
1	On Node A, select **Cluster Configuration > Cluster Resources > Change/Show Run-Time Parameters**.
2	When the Run-Time Parameters screen appears, select Node A and set its security mode to **Enhanced**.
3	Select Node B and set its security mode to **Enhanced**.

Step	What you do...
4	On Node B, select **Cluster Configuration > Cluster Resources > Change/Show Run-Time Parameters**.
5	When the Run-Time Parameters screen appears, select Node B and set its security mode to **Enhanced**.
6	On Node A, synchronize the cluster topology. See Chapter 12, Configuring an HANFS for AIX Cluster, for more information.

Automount Daemon

For SP installations that require the automount daemon (AMD) on HANFS nodes, a modification is needed to ensure that AMD starts properly (with NFS available and running) on node bootup. This is due to the way HANFS for AIX manages the **inittab** file and run levels upon startup.

To enable AMD on nodes that have HANFS for AIX installed, add the following line as the last line of the file **/usr/sbin/cluster/etc/harc.net**:

```
startsrc –s nfsd
```

Cluster Verification and Synchronization

When running HANFS for AIX verification and synchronization, make sure the SP Switch network is up; otherwise, you receive network down error messages.

Configuring the SP Switch Network

The process for configuring the HANFS for AIX Version 4.3.1 software on SP nodes is similar to configuring it on an RS/6000 system but with additional information needed for configuring the SP Switch.

Note: You must define an additional network besides the SP Switch, in order to avoid synchronization errors. Since the SP Switch uses IP aliasing, defining the network does not update the CuAt ODM database with an HANFS-defined adapter. During synchronization, HANFS looks for entries for adapters in the CuAt ODM database. If it finds none, synchronization fails.

Since only one SP Switch adapter per SP node is present, it is a single point of failure for that node. It is recommended that you use AIX error notification to promote adapter failures to node failures. Errors that should be provided to AIX error notification are “HPS_FAULT9_ER” and “HPS_FAULT3_RE”. To test the notify methods produced by these errors, use the error log emulation utility. You can also use the Automatic Error Notification utility.

For more information on adding a notify method and on the error log emulation utility, see Chapter 14, Supporting AIX Error Notification.

HANFS for AIX Eprimary Management for the SP Switch

HANFS for AIX 4.2.1 and 4.2.2 allowed either the HPS (older version) or the SP (newer version) of the switch. HANFS for AIX 4.3.0 and 4.3.1 support only the SP switch.

The Eprimary node is the designated node for the switch initialization and recovery.

The SP switch cannot be managed by HANFS. The SP switch can configure a secondary Eprimary and reassign the Eprimary automatically on failure. This is handled by the SP software, outside of HANFS.

Upgrading From HPS Switch to SP Switch

You must upgrade to the SP Switch before installing HANFS/ES 4.3.1. If you are currently running HANFS Eprimary management with an HPS switch, you should run the HANFS for AIX script to unmanage the Eprimary BEFORE upgrading the switch.

To check whether the Eprimary is set to be managed:

```
odmget -q'name=EPRIMARY' HACMPsp2
```

If the switch is set to MANAGE, before changing to the new switch, run the script:

```
/usr/sbin/cluster/events/utils/cl_HPS_Eprimary unmanage
```

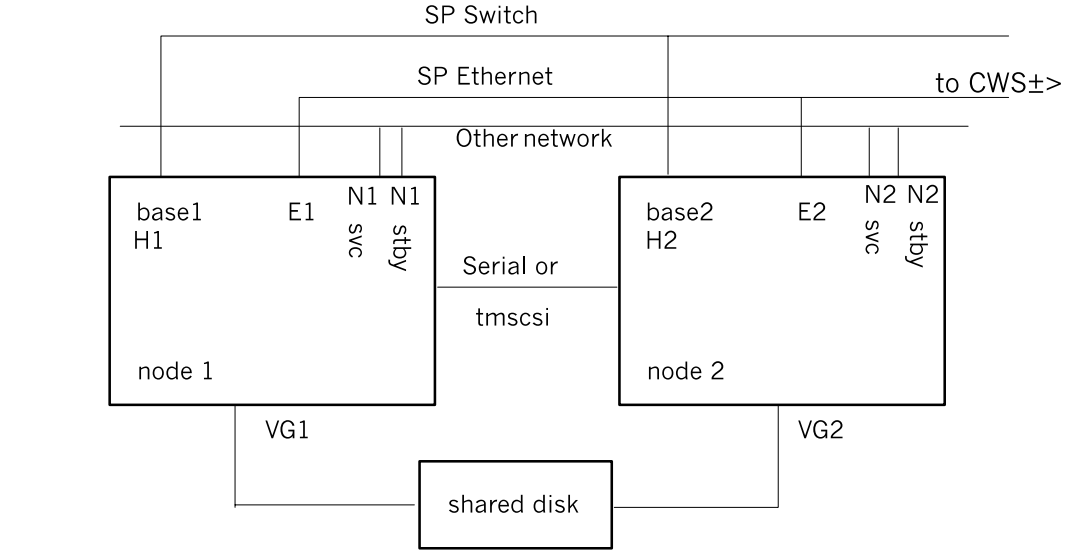
Configuring the SP Switch Base IP Address as Service Adapter

The SP Switch adapter css0 base IP address can be used as a service adapter as long as IP Address Takeover is *not* configured for the switch. It must not be used in an IP Address Takeover configuration.

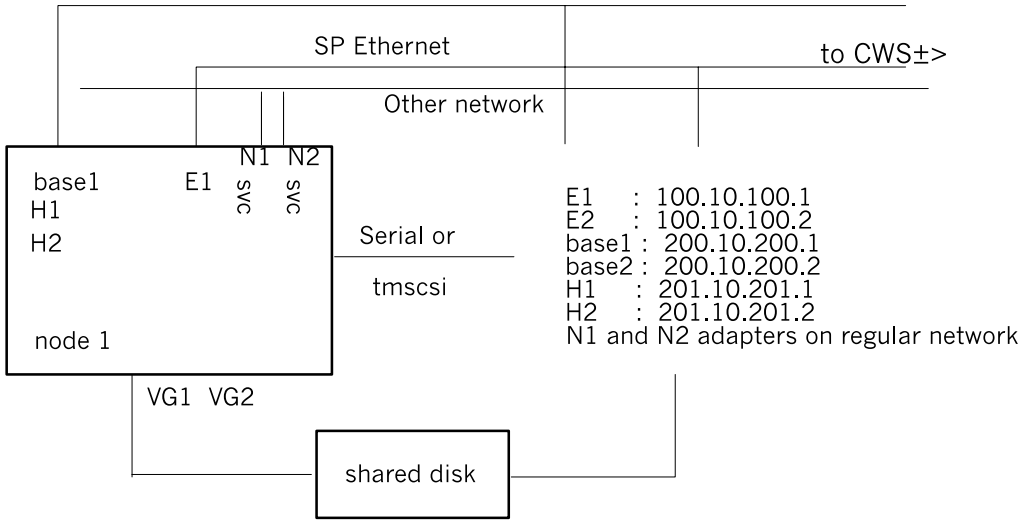
The base address cannot be modified.

Configuring HANFS for AIX for IP Address Takeover on an SP Switch Network

The SP Switch is the first network to make use of IP address aliasing with HANFS for AIX to permit IP address takeover (IPAT). See the following figure for a general illustration of IP address takeover on the SP Switch. In the figure, Node 1 and Node 2 are SP nodes. E1 and E2 are SP Ethernet (Reliable Ethernet) IP addresses. The base1 and base2 labels reflect SP Switch base IP addresses. H1 and H2 are SP Switch alias HANFS for AIX service addresses. VG1 and VG2 are volume groups. N1 and N2 are adapters on other (Ethernet, FDDI) networks.



Cluster after node 2 fails:



Sample HANFS for AIX Two-Node Cluster Configuration on the SP Machine.

Note: The HANFS boot addresses are not included in this figure. Boot addresses are also aliases, different from the SP switch base IP address.

Considerations for IPAT with the SP Switch

Keep the following points in mind when configuring the HANFS for AIX 4.3.1 software for the SP Switch using aliasing for an IPAT configuration:

- HANFS for AIX SP Switch boot and service addresses must be alias addresses on the SP Switch css0 IP interface. The css0 interface can have more than one alias IP address; therefore, it can support IP takeover addresses. At present, only one boot address can be defined per SP Switch css0 interface.

You can configure HANFS to have the switch adapter take over up to seven additional node IP addresses using aliasing. These HANFS for AIX "alias HPS service addresses" appear as "ifconfig alias" addresses on the css0 interface when viewing the node interfaces.

- SP Switch boot and service addresses must be different from the css0 base IP address in order to configure IP Address takeover.
- Address Resolution Protocol (ARP) must be enabled for the SP Switch network in order for IPAT to work on the SP Switch. ARP can be configured by an SP customize operation, or during initial SP setup. A method to update the SP Switch to use ARP is presented in SP Switch Address Resolution Protocol (ARP) on page C-11.
- Standby adapter addresses are not used for SP Switch IP address takeover.
- The SP Switch alias addresses for IPAT can be configured as a part of a cascading or rotating resource group.

Note: In the case of a major SP Switch failure, the aliases HANFS needs for switch IP address takeover may be deleted when the **Eclock** command runs **rc.switch.** For this reason, if you are configuring IPAT with the SP Switch, you should create an event script for either the network_down or the network_down_complete event to add back the aliases for css0.

SP Switch Address Resolution Protocol (ARP)

If your SP nodes are already installed and the switch network is up on all nodes, you can verify whether ARP is enabled. On the control workstation, enter the following command:

```
dsh -av "/usr/lpp/ssp/css/ifconfig css0"
```

If NOARP appears as output from any of the nodes, you must enable ARP to use IP takeover on the SP Switch. ARP must be enabled on all SP nodes connected to the SP Switch.

Warning: Before you perform the following steps, be sure to back up **CuAt**. If user error causes **CuAt** to become corrupt, the SP nodes may be corrupted and will have to be re-installed. You will need to copy your backup of **CuAt** to **/etc/objrepos/CuAt** prior to rebooting the system. Be careful! If you feel this is too risky, customize the nodes to turn ARP on (see the *SP Administration Guide* for help with this procedure).

To enable ARP on all the nodes, follow these steps carefully. Enter all commands from the control workstation. Ensure all nodes are up. The quotation marks shown in the commands must be typed.

1. Create a copy of the **CuAt** file on all nodes:

```
dsh -av "cp /etc/objrepos/CuAt /etc/objrepos/CuAt.save"
dsh -av "odmget -q 'name=css and attribute=arp_enabled' CuAt |
       sed s/no/yes/ > /tmp/arpon.data"
dsh -av "odmchange -o CuAt -q'name=css and attribute=arp_enabled'
       /tmp/arpon.data"
```

2. Verify that the previous commands worked:

```
dsh -av "odmget -a 'name=css and name=arp_enabled' CuAt | grep value"
```

You should see an entry reporting "value=yes" from every node.

3. Remove the temporary file from all nodes:

   ```
   dsh -av rm /tmp/arpon.data
   ```

4. Shut down and reboot the nodes:

   ```
   dsh -av "shutdown -Fr"
   ```

Handling Global Network Failure

The SP Switch is a highly available network. All nodes have four paths to each other through the switch network. Fault isolation and recovery is automatic. However, extremely rare failures will result in SP Switch outage on all nodes, or global network failure. The following section is intended to help in dealing with this situation.

Global Network Failure Detection and Action

Several options exist for detecting failure and invoking user defined scripts to verify the failure and recover.

The switch power off will be seen as a HPS_FAULT9_ER recorded on each node, followed by HPS_FAULT6_ER (fault service daemon terminated). By modifying the AIX error notification strategies, it is possible to call a user script to detect the global switch failure and perform some recovery action. The user script would have to do the following:

- Detect global network failure (switch power failure or fault service daemon terminated on all nodes).
- Take recovery action, such as moving workload to another network, or reconfiguring a backup network.
- In order to recover from a major switch failure (power off, for example), you must issue **Eclock** and **Estart** commands to bring the switch back on-line. The **Eclock** command runs **rc.switch,** which deletes the aliases HANFS needs for SP Switch IP address takeover. It is recommended to create an event script for either the network_down or the network_down_complete event to add back the aliases for css0.

See Chapter 14, Supporting AIX Error Notification for more information.

Other Network Issues

You should give thought to the following issues when using an RS/6000 SP machine:

- No IPAT support on the SP administrative Ethernet
- Serial or non-IP network considerations.

IP Address Takeover Not Supported on the SP Administrative Ethernet

Since some of the SP software requires that an IP address on the SP administrative Ethernet (en0 adapter) is associated with a specific node, IP address takeover, as defined in cascading or rotating resource groups, cannot be configured on this network. You should configure this adapter so the network will be monitored by HANFS for AIX (this is done by configuring the SP Ethernet adapter as part of a cascading resource group where the adapter labels are not part of the resource group), but it must not be configured for IP address takeover (do not configure a boot address). In addition, no owned or takeover resources can be associated with this adapter.

Serial or Non-IP Network Considerations

It is strongly recommended (but not required) that a non-IP network be present between nodes that share resources, in order to eliminate TCP/IP (**inetd)** on one node as a single point of failure. At present, target mode SCSI (tmscsi) target mode SSA (tmssa) or serial (tty) networks are supported by HANFS for AIX.

- On the SP there are no serial ports available on thin or wide nodes. Therefore, any HANFS for AIX configurations that require a tty network need to make use of a serial adapter card (8-port async EIA-232 adapter, FC/2930), available on the SP as an RPQ.
- For 7135 and SCSI configurations, tmscsi or tmssa can be used with I/O pacing to provide the serial network, as described in this book.

Appendix D HANFS for AIX Commands

This appendix provides a quick reference to commands commonly used to obtain information about the cluster environment or to execute a specific function. The chapter lists syntax diagrams and provides examples for using each command.

Overview of Contents

As system administrator, you often must obtain information about your cluster to determine if it is operating correctly. The commands you need to obtain this information are listed in alphabetical order in this chapter.

Highlighting

The following highlighting conventions are used in this appendix:

Bold	Identifies command words, keywords, files, directories, and other items whose actual names are predefined by the system.
Italics	Identifies parameters whose actual names or values are supplied by the user.
`Monospace`	Identifies examples of specific data values, examples of text similar to what you might see displayed, examples of program code similar to what you might write as a programmer, messages from the system, or information you should actually type.

Reading Syntax Diagrams

Usually, a command follows this syntax:

[]	Material within brackets is optional.
{ }	Material within braces is required.
|	Indicates an alternative. Only one of the options can be chosen.
...	Indicates that one or more of the kinds of parameters or objects preceding the ellipsis can be entered.

Note: Flags listed in syntax diagrams throughout this appendix are those recommended for use with the HANFS for AIX software. Flags used internally by SMIT are not listed.

Related Information

For complete information on a command's capabilities and restrictions, see the online man page and the relevant chapter in this guide. Man pages for HANFS for AIX, Version 4.3.1 commands and utilities are installed in the **/usr/share/man/cat1** directory. Use the following syntax to read man page information:

```
man [command-name]
```

where *command-name* is the actual name of the HANFS command or script. For example, type **man clstart** to obtain information about the HANFS cluster startup command.

HANFS for AIX Commands

cldiag debug clstrmgr -l *level* -R *file*

Enables real-time debugging of the Cluster Manager..

clstrmgr Allows viewing of Cluster Manager debug information.

-l *level* The level of debugging to be performed. The levels range from 1 to 9 in increasing amounts of information. The default (0) turns debugging off.

-R *file* Saves output in the specified file.

Example

```
cldiag debug clstrmgr -l2 -R foo
```

Enables Cluster Manager debugging at debug level 2 and saves output to the file named *foo*.

cldiag logs {scripts [-s] [-f] [event...] | syslog [-e] [-w] } [-h *hostname*] [-d *#_of_days*] [-R *file*]

Allows for selected viewing and parsing of HANFS for AIX process and script output files.

-h *hostname* Hostname of the system from which to gather data.

-s Captures all "Start" and "Complete" events (scripts option).

-f Captures all "Fail" events (scripts option).

-d *#_of_days* The number of days preceding the present day from which information will be gathered.

event... Captures all lines containing the cluster event (such as **node_up_local**). This can be a list of events (scripts option).

-e Captures all "Error" events (**syslog** option).

-w Captures all "Warning" events (**syslog** option).

-R *file* Saves output in the specified file.

Example
```
cldiag logs scripts -f -d3
```

Captures all failed script events that occurred within the last three days.

cldiag vgs -h *hostnames* [-v *vgnames*]

Checks for consistencies among volume groups on various hosts, ODMs, and disks.

-h *hostnames* List of 2 – 8 hostnames, separated by commas (no space).

-v *vgnames* List of 2 – 8 volume group names, separated by commas (no space).

Example 1
```
cldiag vgs -h jaws,kelp
```

Checks all common volume groups for hosts *jaws* and *kelp*.

Example 2
```
cldiag vgs -h limpet,cowrie -v vgck1
```

Checks volume group *vgck1* for hosts *limpet* and *cowrie*.

cldiag error { [short | long | cluster] } [-h *hostname*] [-R file]

Allows for parsing the system error log for errors that occurred in a cluster.

short Short error report.

long Long error report.

cluster **clstrmgr** and HANFS for AIX error report.

-h *hostname* Hostname of the system from which to gather data.

-R *file* Saves output in the specified file.

Example
```
cldiag error long -h steamer -R err_rep
```

Generates a long error report for host steamer and sends it to the file **err_rep**.

cldiag trace [-t *time*] [-R *file*] [-l] *daemon ...*

Allows for tracing HANFS daemons (**clstrmgr**, **clsmuxpd**, **clinfo**).

-t *time* Number of seconds to perform the trace.

-R *file* Redirects output to this file.

-l Chooses more detailed trace daemon.

daemon List of cluster daemons to trace.

Example

```
cldiag trace -t 45 clinfo
```

Traces the **clinfo** daemon for 45 seconds; writing to **stdout**.

clgetaddr [-o *odmdir*] *nodename*

Returns a **ping**-able address for the specified node name.

-o Specifies an alternate ODM directory.

Example

To get a **ping**-able address for the node seaweed, enter:

```
clgetaddr seaweed
```

The following address is returned: `2361059035`

clgetgrp { -g *group* } [-o] [-c | -h -f *field*]

Retrieves and displays HANFSgroup class objects.

-g *group* Specifies name of resource group to list.

-o *objdir* Specifies an alternate ODM directory to **/etc/objrepos**.

-c Specifies a colon output format.

-h Specifies to print a header.

-f *field* Specifies the field in object to list.

Example

```
clgetgrp -g grp3
```

Lists information for resource group *grp3*.

clgetif { [-a | -n | -d] *IPlabel* | *IPaddress* }

Prints the interface name and/or netmask associated with a specified IP label or address.

-a	Prints interface name (for example, *en0*).
-n	Prints interface netmask in dotted decimal.
-d	Prints interface device name.
IPlabel	Specifies the name of the adapter to show.
IPaddress	Specifies the address of the adapter to show.

Example 1

```
clgetif –a clam
```

Prints interface name for adapter with IP label *clam*.

Example 2

```
clgetif –n 1.1.1.22
```

Prints netmask for adapter name with IP address *1.1.1.22*.

Example 3

```
clgetif –an clam_svc clam_stby
```

Prints interface names and netmasks for adapters with IP labels *clam_svc* and *clam_stby*.

clgodmget [-c] [-q *criteria*] { -n *nodename class* }

Lists ODM class remotely.

-c	Specifies a colon output format.
-q *criteria*	Specifies search criteria for ODM retrieve. See the **odmget** man page for information on search criteria. If no criteria are entered, all objects in the class are retrieved.
-n *nodename*	Searches the ODM from the specified node.
class	ODM object class.

Example

```
clgodmget –n abalone HANFSadapter
```

Prints adapter information for the node *abalone*.

cllscf

Lists complete cluster topology information.

cllsclstr [-i *id*]

Shows cluster name and ID in the cluster configuration ODM object class. If no cluster ID is included, shows the information for the cluster where the local node is configured.

-i *id* Cluster ID to show.

Example
```
cllsclstr -i 2
```

Shows the cluster name and ID as configured for cluster *2*.

cllsdisk { -g *resource_group* }

Lists PVIDs of accessible disks in a specified resource chain.

-g resource group

Specifies name of resource group to list.

Example
```
cllsdisk -g grp3
```

Lists PVIDs of disks accessible in resource group *grp3*.

cllsfs { -g *resource_group* } [-n]

Lists shared file systems contained in a resource group.

-g *resource_group* Specifies name of resource group for which to list file systems.

-n Lists the nodes that share the file system in the resource group.

Note: Do not run the cllsfs command from the command line. Use the SMIT interface to retrieve file system information, as explained in Chapter 5, Planning Shared LVM Components.

cllsgrp

Lists names of all resource groups configured in the cluster.

cllsnim [-d *odmdir*] [-c] [-n *nimname*]

Lists contents of HANFSnetwork interface module ODM class.

-d *odmdir* Specifies an alternate ODM directory to **/etc/objrepos**.

-c Specifies a colon output format.

-n *nimname* Name of the network interface module for which to list information.

Example 1
```
cllsnim
```

Shows information for all configured network modules.

Example 2
```
cllsnim -n ether
```

Shows information for all configured Ethernet network modules.

cllsnode [-i *nodename*]

Shows node information retrieved from adapter and network ODM object classes.

-i *nodename* Node name of target node.

Example 1
```
cllsnode
```

Shows information for all configured nodes.

Example 2
```
cllsnode -i seaweed
```

Shows information for node *seaweed*.

cllsnw [-n *name*]

Shows network information defined in the adapter and network configuration ODMs.

-n *name* Name of network for information to display.

Example 1
```
cllsnw
```

Shows all networks.

Example 2
```
cllsnw -n ether
```

Shows network named *ether*.

cllsparam {-n *nodename*} [-c] [-s] [-d *odmdir*]

Lists run-time parameters.

-n *nodename* Specifies a node for which to list the information.

-c Specifies a colon output format.

-s Used along with the **-c** flag, specifies native language instead of English.

-d *odmdir* Specifies an alternate ODM directory.

Example
```
cllsparam -n abalone
```

Shows run-time parameters for node *abalone*.

cllsres [-g *group*] [-c] [-s] [-f *field*] [-d *odmdir*] [-q *query*]

Sorts HANFS for AIX ODM resource data by name and arguments.

-g *group*	Specifies name of resource group to list.
-c	Specifies a colon output format.
-s	Used with the **-c** flag, specifies native language instead of English.
-f *field*	Specifies the field in object to list. Fields are the resource group name, the type of resource group (cascading, concurrent, or rotating), and participating nodes.
-d *odmdir*	Specifies an alternate ODM directory.
-q *query*	Specifies search criteria for ODM retrieve. See the **odmget** man page for information on search criteria.

Example 1

```
cllsres
```

Lists resource data for all resource groups.

Example 2

```
cllsres -g grp1
```

Lists resource data for resource group *grp1*.

Example 3

```
cllsres -g grp1 -q"name = FILESYSTEM"
```

Lists file system resource data for resource group *grp1*.

cllsvg { -g *resource_group* } [-n]

List shared volume groups in a specified resource chain.

-g *resource_group*	Specifies name of resource group for which to list volume groups.
-n *nodes*	Specifies all nodes participating in a resource group.

Example

```
cllsvg -g grp1
```

Lists all shared volume groups in resource group *grp1*.

clnodename [-d *odmdir*] [-o *oldname* -n *newname*] [-V *high/low*][-N *true/false*]

clnodename [-d *odmdir*] [-V *high/low*] [-N *true/false*] -a *name*

clnodename [-d *odmdir*] -r *name*]

Lists, adds, changes, or removes a cluster node name.

-d *odmdir*	Specifies an alternate ODM directory other than **/etc/objrepos**.
-o *oldname*	Specifies the old name when changing a node name.
-n *newname*	Specifies the new name when changing a node name.
-a *name*	Specifies the node name or names to be added to the cluster.
-r *name*	Specifies the node name to be removed
-V high/*low*	Verbose logging on high or low.
-N *true/false*	Nameserving is used: true or false.

Example
To list all node names configured for a cluster, enter:

```
clnodename
```

The command returns the list of node names.

clshowres [-g *group*] [-n *nodename*] [-d *odmdir*]

Shows resource group information for a cluster or a node.

-g *group*	Name of resource group to show.
-n *nodename*	Searches the resources ODM from the specified node.
-d *odmdir*	Specifies ***odmdir*** as the ODM object repository directory instead of the default **/etc/objrepos**.

Example 1
```
clshowres
```

Lists all the resource group information for the cluster.

Example 2
```
clshowres -n clam
```

Lists the resource group information for node *clam*.

clstat [-c *id* | -i] [-r *seconds*] [-a]

Cluster Status Monitor (ASCII mode).

-c *id* Displays cluster information only about the cluster with the specified ID. If the specified cluster is not available, **clstat** continues looking for the cluster until the cluster is found or the program is cancelled. May not be specified if the **-i** option is used.

-i Runs ASCII clstat in interactive mode. Initially displays a list of all clusters accessible to the system. The user must select the cluster for which to display the detailed information. A number of functions are available from the detailed display.

-r *seconds* Updates the cluster status display at the specified number of seconds. The default is 1 second; however, the display is updated only if the cluster state changes.

-a Causes **clstat** to display in ASCII mode.

clstat [-a] [-c *id*] [-r *tenths-of-seconds*]

Cluster Status Monitor (X Windows mode).

-a Runs **clstat** in ASCII mode.

-c *id* Displays cluster information only about the cluster with the specified ID. If the specified cluster is not available, **clstat** continues looking for the cluster until the cluster is found or the program is cancelled. May not be specified if the **-n** option is used.

-r *tenths-of-seconds* The interval at which the **clstat** utility updates the display. For the graphical interface, this value is interpreted in tenths of seconds. By default, **clstat** updates the display every 0.10 seconds.

Example 1

```
clstat -c 10
```

Displays the cluster information about the cluster whose ID is *10*.

Example 2

```
clstat -i
```

Runs ASCII **clstat** in interactive mode, allowing multi-cluster monitoring.

Example 3

```
clstat -n waves
```

Displays information about the cluster named *waves.*

Buttons on X Window System Display

Prev	Displays previous cluster.
Next	Displays next cluster.
Name:Id	Refresh bar, pressing bar causes **clstat** to refresh immediately.
Quit	Exits application.
Help	Pop-up help window shows the **clstat** manual page.

clverify cluster { topology check | topology sync | config networks | config resources | config all} [-e *num*] [-R *file*]

Verifies cluster installation and configuration.

topology check	Checks that all nodes agree on cluster topology.
topology sync	Forces all nodes to agree on cluster topology as defined on the local node.
config resources	Verifies ownership of all disks and other resources.
config networks	Verifies configuration of network adapters.
config all	Runs both resources and networks programs.
-e *num*	Aborts the program after ***num*** errors (not available for use with topology options).
-R *file*	Redirects output to a file (not available for use with topology options).

Example

```
clverify cluster config networks -R verify_nw
```

Verifies the cluster networks configuration and logs the results in a file called **verify_nw**.

clverify software {lpp} [-R *file*]

Verifies proper software installation.

lpp	Verifies that HANFS-specific modifications to AIX system files are correct.
-R *file*	Redirects output to a file.

Example

```
clverify software lpp -R verify_lpp
```

Verifies that the proper HANFS-specific modifications exist and are correct, and logs the output in a file called **verify_lpp**.

cl_nfskill

Syntax

```
cl_nfskill [-k] [-t] [-u] directory ...
```

Description

Lists the process numbers of local processes using the specified NFS directory.

Find and kill processes that are executables fetched from the NFS-mounted file system. Only the root user can kill a process of another user.

If you specify the **-t** flag, all processes that have certain NFS module names within their stack will be killed.

Warning: When using the **-t** flag it is not possible to tell which NFS filesystem the process is related to. This could result in killing processes which belong to NFS-mounted filesystems other than those which are cross-mounted from another HACMP node and under HACMP control. This could also mean that the processes found could be related to filesystems under HACMP control but not part of the current resources being taken. This flag should therefore be used with caution and only if you know you have a specific problem with unmounting the NFS filesystems.

To help to control this, the **cl_deactivate_nfs** script contains the normal calls to **cl_nfskill** with the **-k** and **-u** flags and commented calls using the **-t** flag as well. If you choose to use the **-t** flag, you should uncomment those calls and comment the original calls.

Parameters

-k	Sends the SIGKILL signal to each local process,
-u	Provides the login name for local processes in parentheses after the process number.
-t	Finds and kills processes that are just opening on NFS file systems.
directory	Lists of one or more NFS directories to check.

Return Values

None.

get_local_nodename

Returns the name of the local node.

Index

B

C

D

E

F

G

H

M

N

O

P

Q

R

S

T

U

V

W

X

Vos remarques sur ce document / Technical publication remark form

Titre / Title : Bull HACMP 4.3.1 HANFS Installation & Administration Guide

Nº Reférence / Reference Nº : 86 A2 61KX 01

Daté / Dated : August 1999

ERREURS DETECTEES / ERRORS IN PUBLICATION

AMELIORATIONS SUGGEREES / SUGGESTIONS FOR IMPROVEMENT TO PUBLICATION

Vos remarques et suggestions seront examinées attentivement.
Si vous désirez une réponse écrite, veuillez indiquer ci-après votre adresse postale complète.

Your comments will be promptly investigated by qualified technical personnel and action will be taken as required.
If you require a written reply, please furnish your complete mailing address below.

NOM / NAME : _____ Date : _____

SOCIETE / COMPANY : _____

ADRESSE / ADDRESS : _____

_____

Remettez cet imprimé à un responsable BULL ou envoyez-le directement à :

Please give this technical publication remark form to your BULL representative or mail to:

BULL ELECTRONICS ANGERS
CEDOC
34 Rue du Nid de Pie – BP 428
49004 ANGERS CEDEX 01
FRANCE

Technical Publications Ordering Form

Bon de Commande de Documents Techniques

To order additional publications, please fill up a copy of this form and send it via mail to:
Pour commander des documents techniques, remplissez une copie de ce formulaire et envoyez-la à :

BULL ELECTRONICS ANGERS
CEDOC
ATTN / MME DUMOULIN
34 Rue du Nid de Pie – BP 428
49004 ANGERS CEDEX 01
FRANCE

Managers / Gestionnaires :
Mrs. / Mme : **C. DUMOULIN** +33 (0) 2 41 73 76 65
Mr. / M : **L. CHERUBIN** +33 (0) 2 41 73 63 96

FAX : +33 (0) 2 41 73 60 19
E–Mail / Courrier Electronique : srv.Cedoc@franp.bull.fr

Or visit our web site at: / Ou visitez notre site web à:
`http://www-frec.bull.com` (PUBLICATIONS, Technical Literature, Ordering Form)

CEDOC Reference # Nº Référence CEDOC	**Qty** Qté	**CEDOC Reference #** Nº Référence CEDOC	**Qty** Qté	**CEDOC Reference #** Nº Référence CEDOC	**Qty** Qté
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
__ __ ____ _ [__]		__ __ ____ _ [__]		__ __ ____ _ [__]	
[_ _] : **no revision number means latest revision** / pas de numéro de révision signifie révision la plus récente					

NOM / NAME : ____________________ Date : __________

SOCIETE / COMPANY : ____________________

ADRESSE / ADDRESS : ____________________

____

PHONE / TELEPHONE : ____________________ FAX : ____________________

E–MAIL : ____________________

For Bull Subsidiaries / Pour les Filiales Bull :
Identification: ____________________

For Bull Affiliated Customers / Pour les Clients Affiliés Bull :
Customer Code / Code Client : ____________________

For Bull Internal Customers / Pour les Clients Internes Bull :
Budgetary Section / Section Budgétaire : ____________________

For Others / Pour les Autres :
Please ask your Bull representative. / Merci de demander à votre contact Bull.

PLACE BAR CODE IN LOWER LEFT CORNER

BULL ELECTRONICS ANGERS
CEDOC
34 Rue du Nid de Pie – BP 428
49004 ANGERS CEDEX 01
FRANCE

ORDER REFERENCE
86 A2 61KX 01

Bull

Utiliser les marques de découpe pour obtenir les étiquettes.
Use the cut marks to get the labels.

AIX
HACMP 4.3.1
HANFS Installation
& Administration
Guide
86 A2 61KX 01

AIX
HACMP 4.3.1
HANFS Installation
& Administration
Guide
86 A2 61KX 01

AIX
HACMP 4.3.1
HANFS Installation
& Administration
Guide
86 A2 61KX 01