High Availability Cluster
Multi-Processing for AIX

Concepts and Facilities

Version 4.4

SC23-4276-02

This edition of *High Availability Cluster Multi-Processing for AIX, Version 4.4: Concepts and Facilities* applies to AIX Version 4.3.3 and to all subsequent releases of this product until otherwise indicated in new releases or technical newsletters.

The following paragraph does not apply to the United Kingdom or any country where such provisions are inconsistent with local law: THIS MANUAL IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

It is not warranted that the contents of this publication or the accompanying source code examples, whether individually or as one or more groups, will meet your requirements or that the publication or the accompanying source code examples are error-free.

This publication could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication.

It is possible that this publication may contain references to, or information about, products (machines and programs), programming, or services that are not announced in your country. Such references or information must not be construed to mean that such products, programming, or services will be offered in your country. Any reference to a licensed program in this publication is not intended to state or imply that you can use only that licensed program. You can use any functionally equivalent program instead.

A reader's comment form is provided at the back of this publication. If the form has been removed address comments to Publications Department, Internal Zip 9561, 11400 Burnet Road, Austin, Texas 78758-3493. To send comments electronically, use this commercial internet address: `aix6kpub@austin.ibm.com`. Any information that you supply may be used without incurring any obligation to you.

Trademarks and Acknowledgments

Contents

About This Guide

This guide introduces High Availability Cluster Multi-Processing for AIX (HACMP for AIX), Version 4.4. Though it is primarily intended to describe the HACMP base system, the product subsystem HACMP/ES is also discussed briefly where appropriate.

Who Should Use This Guide

System administrators, system engineers, and other information systems professionals who want to learn about features and functionality provided by the HACMP for AIX software on RS/6000 machines should read this guide.

How to Use This Guide

The guide contains the following chapters:

- Chapter 1, HACMP for AIX, defines high availability and cluster multi-processing, and describes an HACMP cluster from a functional perspective.
- Chapter 2, Cluster Software, describes the HACMP for AIX software that implements a highly available environment.
- Chapter 3, Ensuring Cluster Availability, describes how the HACMP for AIX software eliminates key system components as single points of failure in a cluster.
- Chapter 4, Cluster Events, describes how the HACMP for AIX software responds to changes in a cluster to maintain high availability.
- Chapter 5, Cluster Configurations, provides examples of the types of cluster configurations supported by the HACMP for AIX software.
- Chapter 6, Administrative Facilities, describes the installation, configuration, and administrative tools supplied with the HACMP for AIX software.

Highlighting

This guide uses the following highlighting conventions:

Italic	Identifies new terms or concepts, or indicates emphasis.
Bold	Identifies routines, commands, keywords, files, directories, menu items, and other items whose actual names are predefined by the system.
`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 books provide additional information about HACMP for AIX:

- *Release Notes* in **/usr/lpp/cluster/doc/release_notes** describe hardware and software requirements
- *HACMP for AIX, Version 4.4: Planning Guide,* order number SC23-4277-02
- *HACMP for AIX, Version 4.4: Installation Guide,* order number SC23-4278-02
- *HACMP for AIX, Version 4.4: Administration Guide,* order number SC23-4279-02
- *HACMP for AIX, Version 4.4: Troubleshooting Guide,* order number SC23-4280-02
- *HACMP for AIX, Version 4.4: Programming Locking Applications,* order number SC23-4281-02
- *HACMP for AIX, Version 4.4: Programming Client Applications,* order number SC23-4282-02
- *HACMP for AIX, Version 4.4: Master Index and Glossary,* order number SC23-4285-02
- *HACMP for AIX, Version 4.4: Enhanced Scalability Installation and Administration Guide, Vol. 1,* order number SC23-4284-02
- *HACMP for AIX, Version 4.4: Enhanced Scalability Installation and Administration Guide, Vol. 2,* order number SC23-4306-01
- *IBM International Program License Agreement,* order number S29H-1286

In addition, if you are configuring an HACMP cluster on an RS/6000 SP system, see Appendix F of the *HACMP for AIX Installation Guide,* or read the *AIX High Availability Cluster Multi-Processing for POWERparallel SP* document for more information.

Ordering Publications

To order additional copies of this guide, use order number SC23-4276-02.

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`

Chapter 1 HACMP for AIX

This chapter first discusses the concepts of high availability and cluster multi-processing, and then describes an HACMP cluster from a functional perspective.

High Availability Cluster Multi-Processing for AIX

IBM's tool for building UNIX-based mission-critical computing platforms is the HACMP for AIX software. The HACMP for AIX software ensures that critical resources are available for processing. HACMP for AIX has two major components: high availability (HA) and cluster multi-processing (CMP).

High Availability

Until recently, the only avenue for achieving high availability in the UNIX realm was through fault tolerant technology. *Fault tolerance* relies on specialized hardware to detect a hardware fault and instantaneously switch to a redundant hardware component—whether the failed component is a processor, memory board, power supply, I/O subsystem, or storage subsystem. Although this cutover is apparently seamless and offers non-stop service, a high premium is paid in both hardware cost and performance because the redundant components do no processing. More importantly, the fault tolerant model does not address software failures, by far the most common reason for down time.

High availability views availability not as a series of replicated physical components, but rather as a set of system-wide, shared resources that cooperate to guarantee essential services. High availability combines software with industry-standard hardware to minimize down time by quickly restoring essential services when a system, component, or application fails. While not instantaneous, services are restored rapidly, often in less than a minute.

The difference between fault tolerance and high availability, then, is this: A fault tolerant environment has no service interruption, while a highly available environment has a minimal service interruption. Many sites are willing to absorb a small amount of down time with high availability rather than pay the much higher cost of providing fault tolerance. Additionally, in most highly available configurations, the backup processors are available for use during normal operation.

High availability systems are an excellent solution for applications that can withstand a short interruption should a failure occur, but which must be restored quickly. Some industries have applications so time-critical that they cannot withstand even a few seconds of down time. Many other industries, however, can withstand small periods of time when their database is unavailable. For those industries, HACMP for AIX can provide the necessary continuity of service without total redundancy.

Cluster Multi-Processing

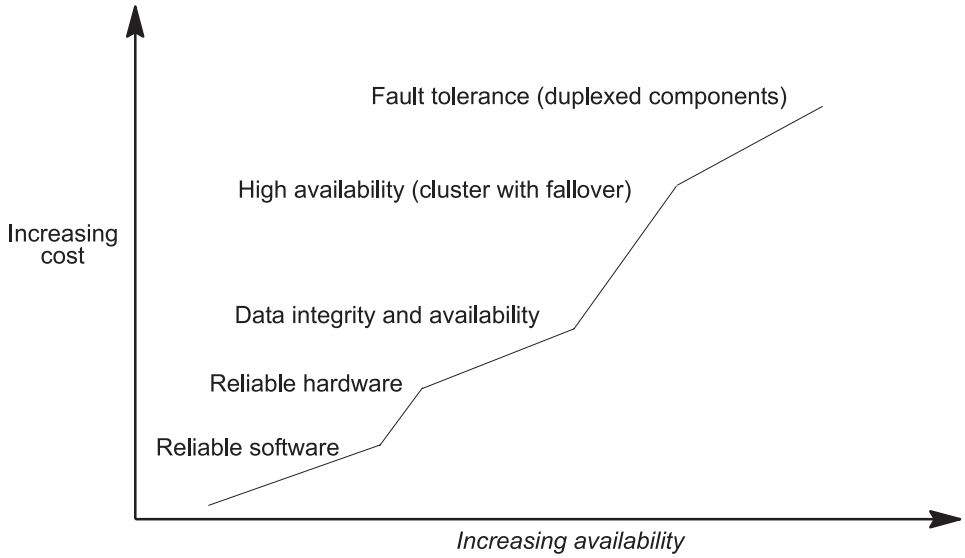
Cluster multi-processing is a group of loosely coupled machines networked together, sharing disk resources. In a cluster, multiple server machines cooperate to provide a set of services or resources to clients.

Clustering two or more servers to back up critical applications is a cost-effective high availability option. You can use more of your site's computing power while ensuring that critical applications resume operations after a minimal interruption caused by a hardware or software failure.

Cluster multi-processing also provides a gradual, scalable growth path. It is easy to add a processor to the cluster to share the growing workload. You can also upgrade one or more of the processors in the cluster to a more powerful model. If you were using a fault tolerant strategy, you must add *two* processors, one as a redundant backup that does no processing during normal operations.

The Availability Costs and Benefits Continuum

The following figure shows the costs and benefits of availability technologies.



Cost and Benefits of Availability Technologies

As you can see, availability is not an all-or-nothing proposition. Think of availability as a continuum. Reliable hardware and software provide the base level of availability. Advanced features such as RAID devices provide an enhanced level of availability. High availability software provides near continuous access to data and applications. Fault tolerant systems ensure the constant availability of the entire system, but at a higher cost.

HACMP Enhanced Scalability for AIX

HACMP for AIX also offers a product subsystem called *Enhanced Scalability.* HACMP/ES takes advantage of RSCT features to bring greater scalability. HACMP/ES clusters can have up to 32 nodes with non-concurrent access, or up to 8 nodes with concurrent access. Concurrent access HACMP/ES clusters require the Concurrent Resource Manager (CRM) feature.

For more detailed information about HACMP/ES, see the *Enhanced Scalability Installation and Administration Guide, Vols. 1 and 2.*

High Availability for Network Filesystem for AIX

Prior to version 4.4, HACMP for AIX included a separate product subsystem called High Availability for Network File System for AIX (HANFS for AIX). HANFS for AIX provided reliable NFS server capability by allowing a backup processor to recover current NFS activity should the primary NFS server fail. The HANFS special functionality extended the HACMP architecture (highly available filesystems and data) to include highly available modifications and locks on NFS filesystems. While HACMP clusters can contain up to eight nodes, HANFS clusters could have a maximum of two nodes.

HANFS Functionality Added to Other HACMP 4.4 Products

In version 4.4, the HANFS extended functionality has been added to the basic HACMP architecture. The following enhancements are included in version 4.4 of the HACMP for AIX product and the HACMP/ES (enhanced scalability) product subsystem:

- You can use the reliable NFS server capability that preserves locks and dupcache (2-node clusters only).
- You can specify a network for NFS mounting.
- You can define NFS exports and mounts at the directory level.
- You can specify export options for NFS-exported directories and filesystems.

For more information on this added functionality, refer to the *HACMP for AIX Planning Guide*, the section Using NFS with HACMP on page 12-11 and in the *HACMP for AIX Installation Guide*,the section NFS Exporting Filesystems and Directories on page 22-5.

In addition, for HACMP/ES, see Chapter 28, Additional Tasks: NFS and Run-Time Parameters of the *Enhanced Scalability Installation and Administration Guide, Vol. 2.*

Goal: Eliminating Scheduled Down-Time

The primary goal of high availability clustering software is to minimize, or ideally, eliminate, the need to take your resources out of service during maintenance and reconfiguration activities.

HACMP for AIX software optimizes availability by allowing for the *dynamic reconfiguration* of running clusters. Most routine cluster maintenance tasks, such as adding or removing a node or changing the priority of nodes participating in a resource group, can be applied to an active cluster without stopping and restarting cluster services. For conceptual information about dynamic reconfiguration, see Chapter 3, Ensuring Cluster Availability. For details about how to use dynamic reconfiguration to change the configuration of a running cluster, see the *HACMP for AIX Administration Guide*.

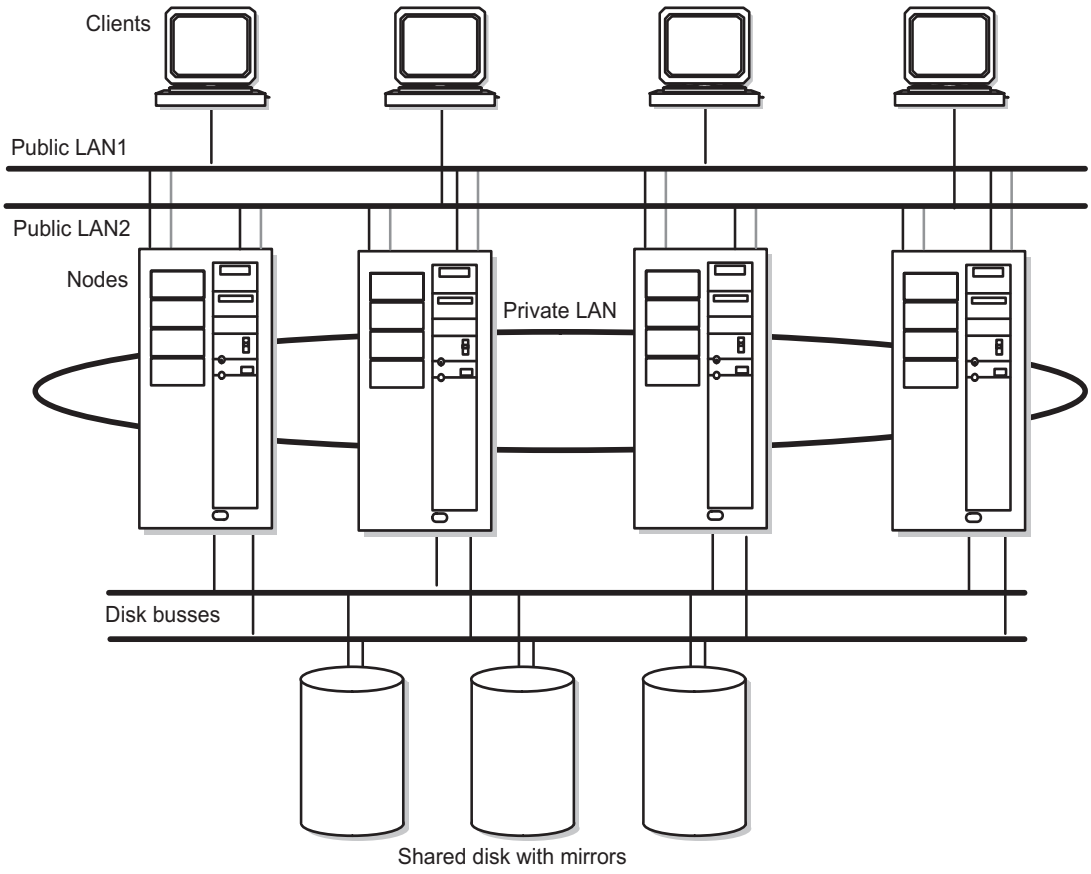
In addition, you can keep an HACMP cluster online while making configuration changes by using the *Cluster Single Point of Control (C-SPOC)* facility. C-SPOC makes cluster management easier, as it allows you to make changes to shared volume groups, users, and groups across the cluster from a single node. The changes are propagated transparently to other cluster nodes.

Components of an HACMP Cluster

An HACMP cluster is made up of the following components:

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

The HACMP for AIX software allows you to combine these components into a wide range of cluster configurations on RS/6000 systems, providing you with flexibility in building a cluster that meets your processing requirements. The following figure shows one example of an HACMP cluster. Other HACMP clusters could look very different—depending on the number of processors, the choice of networking and disk technologies, and so on. Chapter 5, Cluster Configurations, describes examples of the types of cluster configurations supported by the HACMP for AIX software.



An Example of an HACMP Cluster

Nodes

Nodes form the core of an HACMP cluster. A node is a processor that runs both AIX and the HACMP for AIX software. The HACMP for AIX software supports RS/6000 uniprocessor and symmetric multiprocessor (SMP) systems and the Scalable POWERParallel processor (SP) systems as cluster nodes. To the HACMP for AIX software, an SMP system looks just like a uniprocessor. SMP systems provide a cost-effective way to increase cluster throughput. Each node in the cluster can be a large SMP machine, extending an HACMP cluster far beyond the limits of a single system and allowing thousands of clients to connect to a single database.

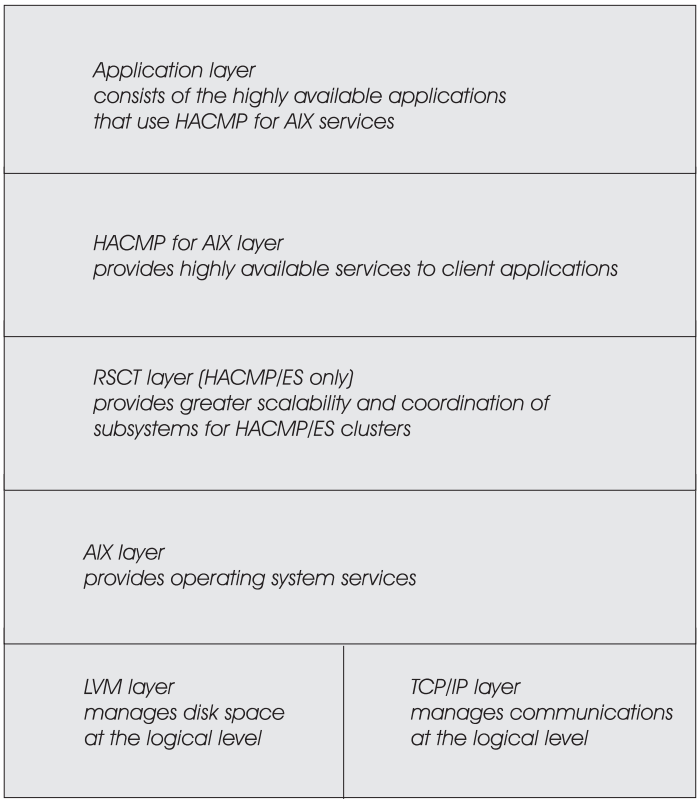
The SP is a parallel processing machine that includes two to 128 processors connected by a high-performance switch. The SP leverages the outstanding reliability of the RS/6000 series by including many standard R/6000 hardware components in its design. The SP's architecture then extends this reliability by enabling processing to continue following the failure of certain components. The SP and its supported disk subsystems provide a robust, stable platform for building highly available clusters. For more information on SP hardware for HACMP, see the *HACMP for AIX: Installation Guide*, Appendix F, or for HACMP/ES, the *Enhanced Scalability Installation and Administration Guide*.

In an HACMP cluster, each node is identified by a unique name. A node may own a set of resources—disks, volume groups, filesystems, networks, network addresses, and applications. Cluster resources are discussed in detail later in this chapter. Typically, a node runs a server or "back end" application that accesses data on the shared external disks. Applications are discussed later in this chapter.

The HACMP for AIX software supports from two to eight nodes in a cluster, depending on the disk technology used for the shared external disks. Shared external disks are discussed later in this chapter. HACMP/ES clusters can have up to 32 nodes in a cluster.

Software Components of an HACMP for AIX Node

The following figure shows the layers of software on a node in an HACMP cluster:



A Model of an HACMP for AIX Cluster Node

The following list describes each layer:

- Applications—Software that uses the services provided by the AIX and HACMP for AIX software, and that can be started and stopped by the HACMP for AIX software. Applications can also incorporate calls to the HACMP for AIX application programming interfaces (APIs) to institute resource locking and to obtain cluster status.
- HACMP for AIX software—Software that recognizes changes within a cluster and coordinates the use of AIX features to create a highly available environment for critical data and applications. This software is described in Chapter 2, Cluster Software.
- RSCT—The IBM RS/6000 Cluster Technology services that are packaged with HACMP/ES provide greater scalability, notify distributed subsystems of software failure, and coordinate recovery and synchronization among all subsystems in the software stack. RSCT includes the Event Manager, Group Services, and Topology Services components. For more information, see the *Enhanced Scalability Installation and Administration Guide*.
- AIX operating system—Software that provides the underlying support for an HACMP cluster, including:
  - Logical Volume Manager (LVM) subsystem, which manages disks at the logical level.
  - TCP/IP subsystem, which provides communications support for an HACMP 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 multiple nodes. The shared disk stores mission-critical data, typically mirrored or RAID-configured for data redundancy. A node in an HACMP cluster must also have internal disks that store the operating system and application binaries, but these are not shared.

The HACMP for AIX software supports shared external disk configurations that use SCSI-2 Differential disks, SCSI-2 Differential Fast/Wide disks, SCSI-2 Differential disk arrays, serial disks, and Serial Storage Architecture (SSA) disks. For more information about supported disks, see the *HACMP for AIX Planning Guide*.

Types of Access to Shared External Disks

Depending on the type of disk used, the HACMP for AIX software supports two types of access to shared external disk devices: non-concurrent access and concurrent access.

In non-concurrent environments, 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 a surviving node assumes ownership of the shared disk.

In concurrent access environments, the shared disks are actively connected to more than one node simultaneously. Therefore, disk takeover is not required when a node fails. The differences between the two methods are explained more fully later in this chapter.

Networks

As an independent, layered component of AIX, the HACMP for AIX software is designed to work with any TCP/IP-based network. Nodes in an HACMP cluster use the network to allow clients to access the cluster nodes, enable cluster nodes to exchange keep-alive messages and, in concurrent access environments, serialize access to data.

The HACMP for AIX software has been tested with Ethernet, Token-Ring, FDDI, ATM, and other networks. For more information about supported networks, see the *HACMP for AIX Planning Guide*.

Types of Networks

The HACMP for AIX software defines three types of networks, characterized by their use: public, private, and serial.

- **Public network**—Connects multiple (two or more) 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 typically does not allow client access. The HACMP for AIX software prefers to use a private network for lock traffic, if one is available. The HACMP for AIX software uses a public network for lock traffic if no private networks are available. Ethernet, Token-Ring, FDDI, Serial Optical Channel Connector (SOCC), and Asynchronous Transfer Mode (ATM) networks can be defined as private networks.

Note: In clusters defined on RS/6000 SP systems, the SP Switch must be defined as a private network.

- **Serial network**—Provides a point-to-point connection between two cluster nodes for HACMP for AIX control messages and heartbeat traffic in the event the TCP/IP subsystem fails. A serial network can be a SCSI-2 Differential bus using target mode SCSI, a target mode SSA connection, 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. However, for serial or SP Switch networks, and for the Administrative Ethernet on the SP, this requirement does not apply.

- **Service network adapter**—Primary connection between an HACMP node and a network. A node can have one or more service network adapters 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, if IP address takeover is enabled, on a remote node. If a service network adapter on the local node fails, the HACMP for AIX software swaps the standby network address with the service network address. If the local node is designated to take over the network address of a peer node should that node fail, the standby network adapter on the local node assumes the IP address of the service network adapter on the failed node.

Assigning a Boot Adapter Label for IP Address Takeover

IP address takeover (IPAT) is an HACMP facility that lets one node acquire the network address of another cluster node. For IPAT to work correctly, however, you must configure the boot adapter to the AIX TCP/IP configuration and then configure both the boot and service IP addresses as part of the HACMP adapter configuration. And both addresses must be assigned to the same network. Cluster nodes use the boot label after a system reboot and before the HACMP for AIX software is started, or after it is stopped gracefully with or without takeover. When a node is forced down, however, the adapter does not revert to its boot address.

When HACMP is started on a node, the node's service adapter is reconfigured to use the service label (address) instead of the boot label. If the node should fail, a 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.

Clients

A client is a processor that can access the nodes in a cluster over a public local area network. Clients each run a "front end" or client application that queries the server application running on the cluster node.

The HACMP for AIX software provides a highly available environment for critical data and applications on cluster nodes. *The HACMP for AIX software does not make the clients themselves highly available.* AIX clients can use the Clinfo services to receive notice of cluster events. Clinfo provides an API that displays cluster status information. The **/usr/sbin/cluster/clstat** utility, a Clinfo client shipped with the HACMP for AIX software, provides information about all cluster service interfaces. See Chapter 2, Cluster Software, for more information about how Clinfo obtains cluster status information.

HACMP for AIX Required and Supported Hardware

For a list of supported hardware, see the *HACMP for AIX Installation Guide*.

Cluster Resources and Resource Groups

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

1. Identifying the set of *cluster resources* that are essential to processing.
2. Defining the *takeover relationships* among the cluster nodes that access these resources.

By identifying resources and defining takeover relationships, the HACMP 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
- Filesystems
- Network addresses
- Application Servers
- Certain applications integrated with HACMP for AIX, including Communications Server for AIX (CS/AIX), AIX Connections, and AIX Fast Connect.

To be made highly available by the HACMP 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.

When defining takeover relationships, keep in mind that a resource group can be taken over by a single node or by several nodes in the cluster. However, any single resource can be taken over by only one node at a time (in *cascading* or *rotating* resource groups).

Defining the Takeover Relationships Among Cluster Nodes

The takeover relationships among cluster nodes determine which cluster nodes control a resource group and which cluster nodes take over control of the resource group when the original node relinquishes control. You define the takeover relationship of a resource group by assigning it one of the following type designations:

- Cascading (with or without the Cascading without Fallback attribute)
- Rotating
- Concurrent

Each type of takeover relationship allows you to specify varying degrees of control over which node, or nodes, control a resource group. The following sections describe each type.

Fallover vs. Fallback

It is important to keep in mind the difference between *fallover* and *fallback*. You will encounter these terms frequently in discussion of the various resource group policies.

Fallover

Fallover refers to the movement of a resource group from the node on which it currently resides (*owner* node) to another active node after its owner node experiences a failure. The new owner is not a reintegrating or joining node.

Fallback

Fallback refers to the movement of a resource group from its owner node specifically to a node that is joining or reintegrating into the cluster; a fallback occurs during a node_up event.

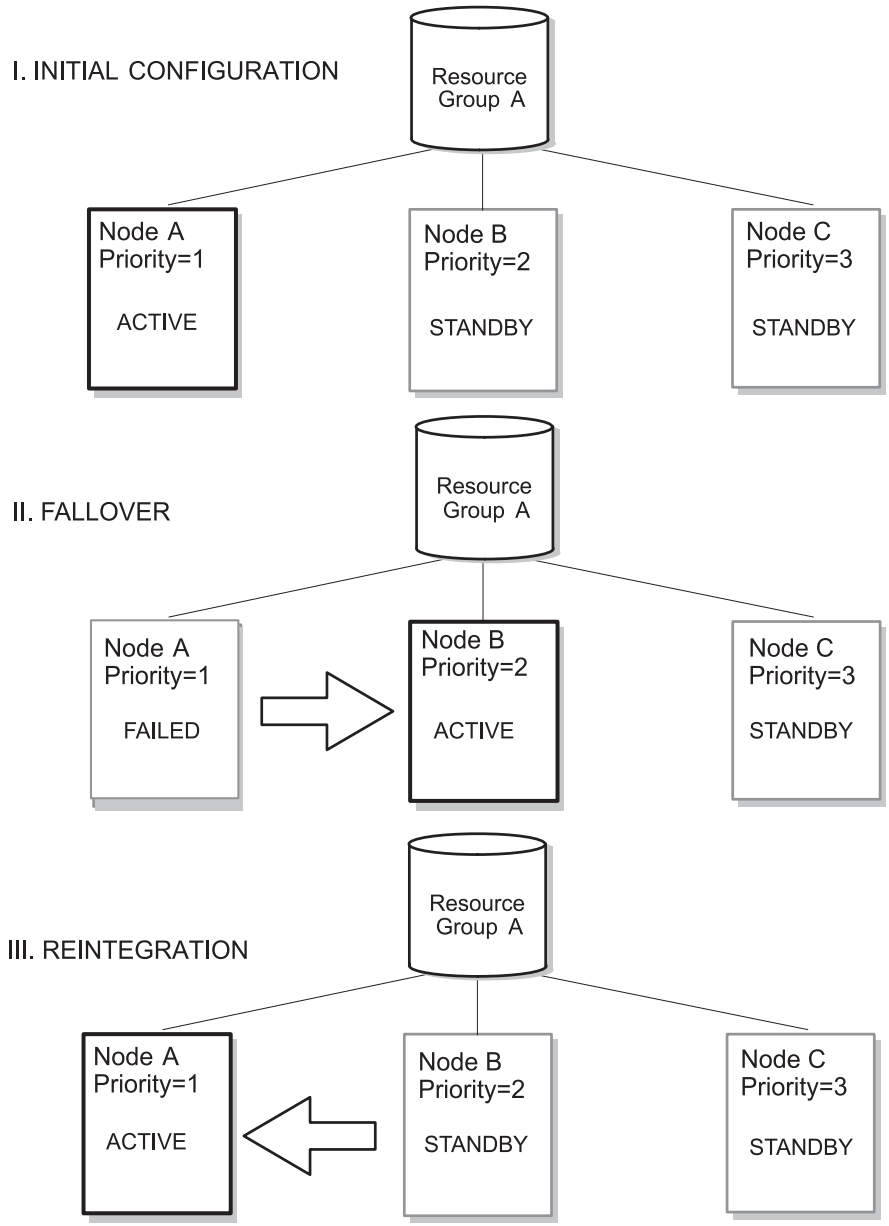
Cascading Resource Groups

A *cascading resource group* defines a list of all the nodes that can control the resource group and then, by assigning a *takeover priority* to each node, specifies a preference for which cluster node controls the resource group. When a fallover occurs, the active node with the highest priority acquires the resource group. If that node is unavailable, the node with the next-highest priority acquires the resource group, and so on.

The list of participating nodes establishes the *resource chain* for that resource group. When a node with a higher priority for that resource group joins or reintegrates into the cluster, it takes control of the resource group. That is, the resource group falls back from nodes with lesser priorities to the higher priority node. Use cascading resource groups with Cascading without Fallback (CWOF) set to **false** when you have a strong preference for which cluster node you want to control a resource group. For example, you may want the cluster node with the highest processing capabilities to control the resource group. Use cascading resource groups with CWOF set to **true** to avoid interruptions of service caused by fallbacks.

The following figure illustrates the takeover relationship among cluster nodes in a cascading resource group with the Cascading without Fallback variable (see below) set to **false** at initial configuration, fallover, and during reintegration. In the figure, lower numbers indicate a higher priority.

.



Interaction of Nodes in Cascading Resource Group with CWOF Set to **False**

Cascading without Fallback Flag

The Cascading without Fallback (CWOF) variable is an attribute of the cascading resource group which defines its fallback behavior.

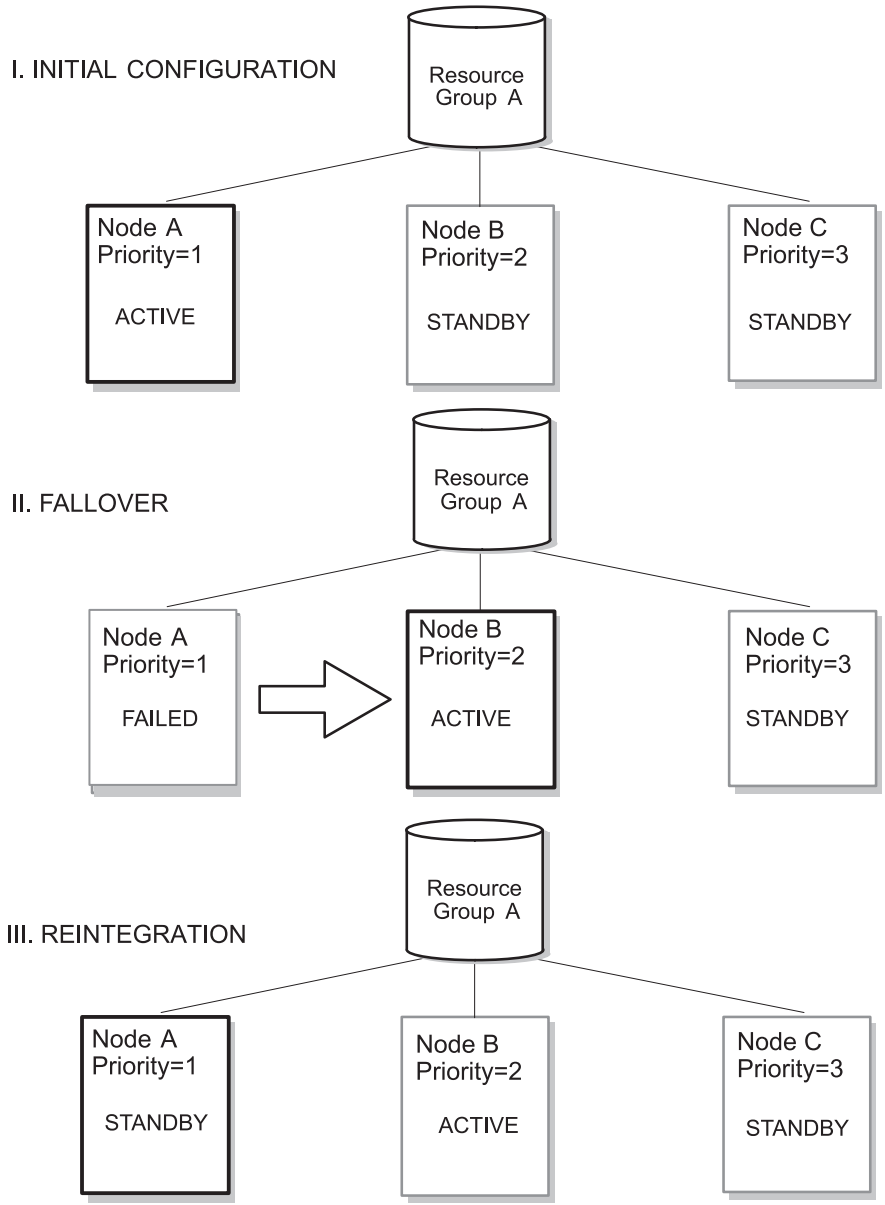
When CWOF is set to **false**, a resource group falls back to any higher priority node when such a node joins or reintegrates into the cluster.

When CWOF is set to **true,** the resource group will not fallback to any node which joins or reintegrates into the cluster.

The CWOF resource group does not require IP Address Takeover to be configured.

Note: A Cascading without Fallback resource group lacks the means of recovery that a fallback may provide. For example, if a cascading group falls over to a node which lacks the resources to optimally support it, you may suffer some loss of service. In a traditional cascading resource group, the resource group subsequently falls back to a higher priority node, and service will likely be restored. Without the possibility for a fallback, a CWOF resource group which falls over to another node will stay there indefinitely until it is manually moved.

The following figure illustrates the takeover relationship among cluster nodes in a cascading resource group with the Cascading without Fallback variable set to **true** at initial configuration, fallover, and during reintegration. Note that in reintegration, the resource group does not fallback to the owner node. In the figure, lower numbers indicate a higher priority.



Interaction of Nodes in Cascading Resource Group with CWOF Set to **True**

Rotating Resource Groups

A *rotating resource group*, like a cascading resource group, defines the list of nodes that can take over control of a resource group and uses priorities to determine the order in which other nodes can take control of the resource. Like cascading resource groups with CWOF set to **true**, control of the resource group does not automatically revert to the node with the highest priority when it reintegrates into the cluster. Use rotating resource groups to avoid the interruption in service caused by a fallback and when it is important that resources remain distributed amongst a number of nodes.

To control the preferred location of rotating resource groups, each group should be assigned a different highest priority node from the list of participating nodes. When the cluster starts, nodes will attempt to acquire the rotating resource group for which it is the highest priority.

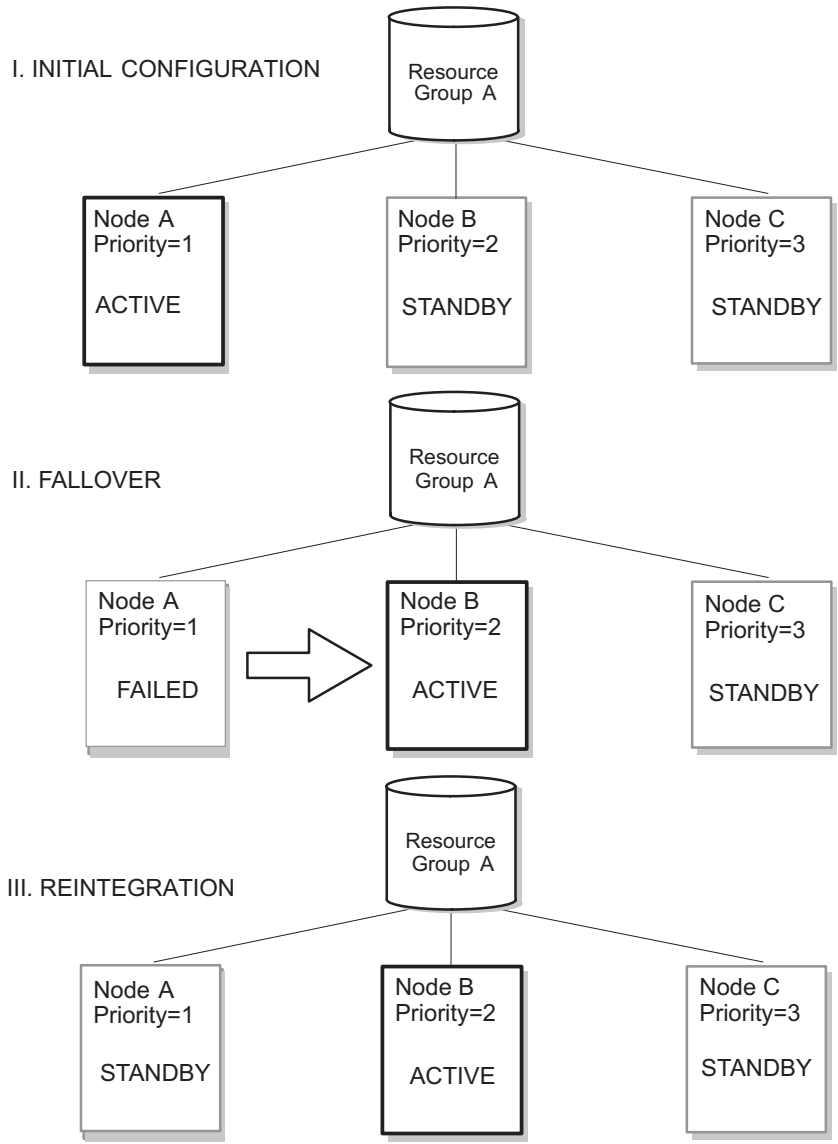
If all rotating resource groups are up, new nodes joining the cluster can join only as standbys for these resource groups. If all rotating groups are not up, a joining node will generally acquire only one of these inactive resource groups. The remaining resource groups will stay inactive. However, if multiple networks exist on which the resource groups can move, a node may acquire multiple rotating groups, one only per network.

When a node managing a resource group fails, the next available node on its boot address (with the highest priority for a resource group) acquires that resource group. The takeover node may be currently managing a resource group, or it may be acting as a standby node. In either case, when a failed node subsequently rejoins the cluster, it does not reacquire any resource groups; instead, it rejoins as a standby node.

For rotating resource groups, the node with the highest priority for a resource group *and* the available connectivity (network, adapter, address) acquires that resource group from a failing node. The HACMP for AIX software assumes that the node that has the rotating resource group's associated service address controls the resource group.

Rotating groups share some similarities with Cascading without Fallback groups. However, there are important differences. Unlike cascading groups, rotating groups interact with one another. Because rotating resource groups require the use of IP address takeover, the nodes in the resource chain must all share the same network connection to the resource group. If several rotating groups share a network, only one of these resource groups can be up on a given node at any time. Thus, rotating groups distribute themselves. Cascading without Fallback groups, however, may “clump” together with multiple CWOF groups on the same node. CWOF does not require an IP address to be associated with the group.

The following figure illustrates the takeover relationship among cluster nodes with a rotating resource group at initial configuration, fallover, and during reintegration. In the figure, lower numbers indicate a higher priority.



Interaction of Nodes in Rotating Resource Group

Concurrent Access Resource Groups

A *concurrent access resource group* may be shared simultaneously by multiple nodes. All nodes concurrently accessing a resource group acquire that resource group when they join the cluster. There are no priorities among nodes.Concurrent access resource groups are supported in clusters with eight or fewer nodes, in HACMP/ES and the HACMP base product subsystem.

The only resources included in a concurrent resource group are volume groups with raw logical volumes, raw disks, and application servers that use the disks. The device on which these logical storage entities are defined must support concurrent access. For more information, see the following section.

Shared External Disk Access

The HACMP for AIX software supports two methods of shared external disk access: non-concurrent and concurrent. Both methods of shared external disk access are described below.

Non-Concurrent Shared External Disk Access

In a non-concurrent environment, only one node has access to a shared external disk at a given time. If this node fails, one of the peer nodes must acquire the disk, mount filesystems defined as resources, and restart applications to restore critical services. Typically, this takes from 30 to 300 seconds, depending on the number and size of the filesystems.

Supported Shared External Disk Types

A non-concurrent configuration can use SCSI-2 Differential disks, SCSI-2 Differential disk arrays, serial disks, and SSA disks as shared external disks. For more information about supported devices, see the *HACMP for AIX Planning Guide*.

Mirroring

To prevent a failed disk from becoming a single point of failure, each logical volume in a shared volume group should be mirrored using the AIX LVM facility. If you are using an IBM 7135 RAIDiant disk array, do not use LVM mirroring. The RAIDiant array provides its own data redundancy.

Applications

Most software that can run in single-machine mode (that is, can run on a single RS/6000 processor) can be managed by the HACMP for AIX software without modification.

Non-concurrent access typically does not require any code changes to server programs (a database management system, for example) or to applications to provide a highly available solution. To end users, node failure looks like a very fast machine reboot. One of the surviving nodes takes ownership of the failed node's resource groups and restarts the highly available applications. The Journaled Filesystem, the native AIX filesystem, guarantees filesystem integrity. The server program guarantees transaction data integrity.

End users simply log onto one of the surviving nodes and restart the application. The logon and application restart procedures can be driven by the HACMP for AIX software. In some HACMP for AIX configurations, users can continue without having to take any action—they simply experience a delay during fallover.

Concurrent Shared External Disk Access

The concurrent access feature enhances the benefits provided by an HACMP cluster. *Concurrent access* allows from two to eight processors to simultaneously access a database or applications residing on shared external disks. Using concurrent access, a cluster can offer nearly continuous availability of resources that rivals fault tolerance, but at a much lower cost. Additionally, concurrent access provides higher performance, eases application development, and allows horizontal growth.

The HACMP for AIX software provides the tools necessary to prepare an application to run in a continuously available mode. These tools are the Cluster Lock Manager and Clinfo programs, described in Chapter 2, Cluster Software.

The benefits of concurrent shared external disk access include the following:

Transparent recovery increases availability

Concurrent access significantly reduces the time for a fallover—sometimes to just a few seconds—because the peer systems already have physical access to the shared disk and are running their own instances of the application.

In a concurrent access environment, fallover basically involves "backing out" in-flight transactions from the failed processor. The server software running on the surviving nodes is responsible for recovering any partial transactions caused by the crash.

Since all nodes have concurrent access to the data, a client/server application can immediately retry a failed request on the surviving nodes, which continue to process incoming transactions.

Harnessing multiple processors increases throughput

Rapid recovery is not the only benefit of the concurrent access environment. Applications are no longer limited to the throughput of a single processor. Instead, multiple instances of an application can run simultaneously on multiple processors. As more processing power is required, more systems can be added to the cluster to increase throughput.

Single database image eases application development and maintenance

In a non-concurrent environment, the only route to improving performance is to partition an application and its data. Breaking code and data into pieces makes both application development and maintenance more complex.

Splitting a database requires a high degree of expertise to make sure that the data and workload are evenly distributed among the processors.

Partitioning code and data is not necessary in a concurrent access environment. To increase throughput, multiple instances of the same application running on different processors can simultaneously access a database on a shared external disk.

Supported Shared External Disk Types

A concurrent configuration can use SCSI-2 Differential disk arrays, serial disks, and SSA disks as shared external disks. For more information about supported devices, see the *HACMP for AIX Planning Guide* and the *HACMP for AIX Installation Guide.*

Mirroring

When creating concurrent access logical volumes, use LVM mirroring to avoid having the disks be a single point of failure.

Note: Do not use LVM mirroring when creating concurrent access logical volumes on IBM 7135-110 and 7135-210 disk arrays. These disk arrays provide their own data redundancy when configured at RAID levels 3 or 5.

Applications

Concurrent access does not support the use of the Journaled File System. Therefore, the database manager must write directly to the raw logical volumes or hdisks in the shared volume group.

An application must use some method (for example, the HACMP for AIX Cluster Lock Manager) to arbitrate all requests for shared data. Most commercial UNIX databases provide a locking model. Using the functions provided in the Cluster Lock Manager API, vendors of specific databases can extend their locking schemes to make them compatible with the HACMP for AIX software. Check with your database vendor to determine whether a specific application supports concurrent access processing.

Chapter 2 Cluster Software

This chapter describes the HACMP for AIX software that implements a highly available environment.

HACMP for AIX Software

The HACMP for AIX software has the following components:

- Cluster Manager
- Cluster SMUX Peer and Cluster Information Program
- Cluster Lock Manager

Cluster Manager

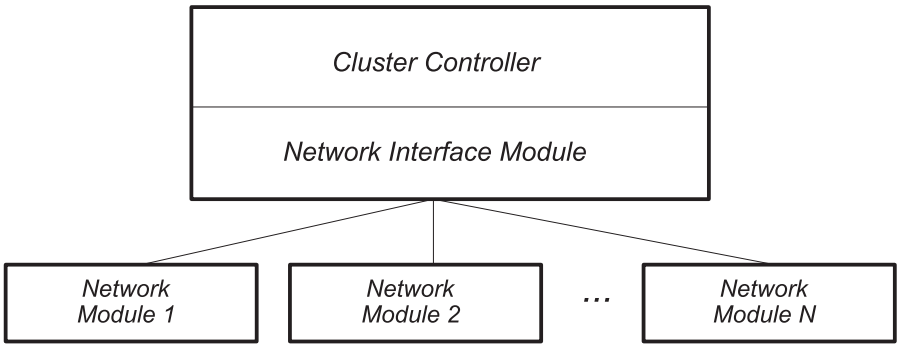
The Cluster Manager runs on each cluster node. The main task of the Cluster Manager is to monitor nodes and networks in the cluster for possible failures. It is responsible for monitoring local hardware and software subsystems, tracking the state of the cluster peers, and acting appropriately to maintain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

Note: In HACMP/ES clusters, the RSCT software components—Group Services, Event Management, and Topology Services—are responsible for most of the monitoring tasks. For more information, see the *HACMP for AIX: Enhanced Scalability Installation and Administration Guide*.

Changes in the state of the cluster are referred to as *cluster events*. The Cluster Manager runs scripts in response to cluster events.

Different modules of the Cluster Manager handle these tasks:

- The *Cluster Controller* performs three tasks. First, it keeps track of cluster membership and synchronizes information among cluster members about the current membership and distribution of cluster resources. Second, the Cluster Controller queues cluster events and runs the appropriate event scripts in response. Third, the Cluster Controller communicates with and coordinates the actions of the cluster's network interface modules.
- The *Network Interface Modules* (NIMs) monitor the nodes and network interfaces associated with a cluster. Each network module monitors one cluster network using one kind of communication protocol (for example, Ethernet or FDDI). Each network module is responsible for maintaining keepalive traffic with neighbor nodes as directed by the Cluster Controller, for providing a link to other nodes on the network it monitors, and for initiating adapter swaps on certain networks.



Cluster Manager Modules in HACMP for AIX

Note: The above diagram illustrates the HACMP for AIX base system architecture. For information on the architecture of the HACMP/ES product subsystem, see the *Enhanced Scalability Installation and Administration Guide.*

Cluster Controller

The Cluster Controller is the Cluster Manager main process. It collects and maintains information about cluster membership, the distribution of resources, and the availability of networks, and synchronizes this information with Cluster Controllers on other nodes. The Cluster Controller then updates this information periodically, based upon script completion information.

The Cluster Controller maintains all event information, including:

- Network joins and failures (information from Cluster Network Modules)
- Remote notification of an event from another Cluster Controller
- User-invoked events.

The Cluster Controller is responsible for:

- Queuing the events and responding intelligently to the situation. For example, it may see multiple network failures in the queue and translate this into a response for a node failure. Thus the Cluster Controller does not simply respond to single events, but may escalate multiple events into more severe failures.
- Communicating the event to all other nodes. Communicating the event about to be processed helps other Cluster Controllers interpret the events in their queues.
- Running the event scripts and handling the event script completion.

Cluster Controller Connection to Other HACMP for AIX Daemons

The Cluster Controller maintains a connection to the local Cluster Lock Manager (CLM), if one exists. It uses this link to determine the state of the Cluster Lock Manager and to send cluster information to it. To determine the state of the Cluster Lock Manager, use the **/usr/sbin/cluster/utilities/clm_stats** command. See the *HACMP for AIX Administration Guide* for a command description.

The Cluster Controller also maintains a connection to the Cluster SMUX Peer daemon, **/usr/sbin/cluster/clsmuxpd**, which gathers cluster information from the Cluster Manager relative to cluster state changes of nodes and interfaces. The Cluster Information Program (Clinfo) gets this information from **clsmuxpd** and allows clients communicating with this program to be aware of changes in a cluster's state. This cluster state information is stored in the HACMP Management Information Base (MIB).

If your system is running TME 10 NetView, the Cluster Controller's connection to the local **clsmuxpd** also allows the HAView utility to obtain cluster state information and to display it graphically through the NetView map. See Chapter 6, Administrative Facilities, for information about how HAView communicates with **clsmuxpd**.

Network Interface Modules

Each supported cluster network in a configured HACMP for AIX system has a corresponding Network Interface Module (NIM). Each network module monitors all I/O on its cluster network.

Each network module maintains a connection to other network modules in the cluster. The Cluster Controllers on nodes in the cluster 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 the remote cluster peers.

The network module is also responsible for determining when a given link actually fails. It does this by sending and receiving periodic keepalive messages to and from other network modules in the cluster.

For a list of the currently supported networks, see the *HACMP for AIX Planning Guide*.

- Configured resource groups and resource group state (HACMP/ES only)
- Resource group location (HACMP/ES only)

Refer to the *HACMP for AIX Administration Guide,* Chapter 3, Monitoring an HACMP Cluster for details about cluster monitoring with Tivoli, and the *HACMP for AIX Installation Guide* for instructions on how to install and configure this functionality.

Cluster SMUX Peer and SNMP Monitoring Programs

An HACMP cluster is dynamic and can undergo various transitions in its state over time. For example, a node can join or leave the cluster, or a standby adapter can replace a service adapter. Each of these changes affects the composition of the cluster, especially when highly available clients and applications must use services provided by cluster nodes.

SNMP Support

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.

SNMP can be extended through the use of the SNMP Multiplexing (SMUX) protocol to include *enterprise-specific* MIBs that contain information relating to a discrete environment or application. A SMUX peer daemon maintains information about the objects defined in its MIB and passes this information on to a specialized network monitoring or network management station.

HACMP for AIX MIB

The Cluster SMUX Peer daemon, **clsmuxpd**, maintains cluster status information in a special HACMP MIB. When **clsmuxpd** starts on a cluster node, it registers with the SNMP daemon, **snmpd**, and 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 HACMP for AIX MIB as it tracks events and resulting states of the cluster.

For more information on the HACMP for AIX MIB, see the manual *HACMP for AIX: Programming Client Applications*.

Cluster Information Program

The Cluster Information Program (Clinfo), the **clinfo** daemon, is an SNMP-based monitor. Clinfo, running on a client machine or on a cluster node, queries the Cluster SMUX Peer for updated cluster information. Through Clinfo, information about the state of an HACMP cluster, nodes, and networks can be made available to clients and applications.

Clients can be divided into two categories: naive and intelligent. A *naive* client views the cluster complex as a single entity. If a cluster node fails, the client must be restarted (or at least must reconnect to the node) if IP address takeover (IPAT) is not enabled. An *intelligent* client, on the other hand, is "cluster-aware"—it reacts appropriately to node failure, connecting to an alternate node and perhaps masking the failure from the user. Such an intelligent client must have knowledge of the cluster state.

The HACMP for AIX software extends the benefits of highly available servers, data, and applications to clients by providing notification of cluster state changes to clients through the **clsmuxpd** and Clinfo API functions.

Responding to Cluster Changes

Clinfo calls the **/usr/sbin/cluster/etc/clinfo.rc** script whenever a cluster, network, or node event occurs. By default, the **clinfo.rc** script flushes the system's ARP cache to reflect changes to network addresses, and it does not update the cache until another address is pinged. 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. Hardware address swapping is described in more detail in Chapter 3, Ensuring Cluster Availability.

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, or use the procedure, "MAC Address Is Not Communicated to the Ethernet Switch," described in the *HACMP for AIX Troubleshooting Guide* to ensure that the hardware address is communicated correctly.

System administrators can add logic to the **clinfo.rc** script if further action is desired.

Clinfo APIs

The Clinfo APIs provide application developers with both a C and a C^{++} language interface for accessing cluster status information. The HACMP for AIX software includes two versions of the Clinfo APIs: one for single-threaded applications and one for multi-threaded applications.

Note: Clinfo and its associated APIs enable developers to write applications that recognize and respond to changes in a cluster. For more information, see the *HACMP for AIX: Programming Client Applications* guide.

Cluster Lock Manager

Ensuring data integrity is of paramount concern in a concurrent access environment where multiple nodes can simultaneously access the same data. It is essential that concurrent requests for the same data do not corrupt shared data. Concurrent access in the HACMP for AIX system is controlled at the application level, not at the operating system level.

The HACMP for AIX software provides a Cluster Lock Manager for this purpose. The Cluster Lock Manager institutes a locking protocol that server processes use to coordinate data access and to ensure data consistency during normal operation as well as in the event of failure. The Cluster Lock Manager must be incorporated into the application to enable concurrent access to shared data. Applications that can benefit from using the Cluster Lock Manager are transaction-oriented, such as a database or a resource controller or manager.

The Cluster Lock Manager provides two distinct lock models:

- CLM lock model, which provides a rich set of locking modes;
- UNIX System V lock model, which supports standard UNIX System V region locking.

Lock Manager APIs

The Cluster Lock Manager provides APIs that applications can use to create a single, unified lock image that is shared among all nodes in the cluster. Cooperating applications running on different nodes in an HACMP cluster can then share common resources without corrupting those resources.

The HACMP for AIX software includes two versions of the Cluster Lock Manager APIs: one for single-threaded applications and one for multi-threaded applications.

For more information about the Cluster Lock Manager, see *HACMP for AIX: Programming Locking Applications*.

Chapter 3 Ensuring Cluster Availability

This chapter describes how the HACMP for AIX software ensures cluster availability by eliminating key system components as single points of failure and by eliminating the need for scheduled down-time for most routine cluster maintenance tasks.

Overview

The key facet of a highly available cluster is its ability to detect and respond to changes that could interrupt the essential services it provides. The HACMP 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 HACMP for AIX software is able to detect the loss and shift that component's workload to another component in the cluster. In planning a highly available cluster, you attempt to ensure that key components do not become *single points of failure*.

In addition, HACMP for AIX software allows a cluster to continue providing application services while routine maintenance tasks are performed using a process called *dynamic reconfiguration*. In dynamic reconfiguration, you can change components in a running cluster, such as adding or removing a node or network adapter, without having to stop and restart cluster services. The changed configuration becomes the active configuration dynamically.

The following sections describe conceptually how to use the HACMP for AIX software to eliminate single points of failure in a cluster, to minimize scheduled down-time in an HACMP cluster with dynamic reconfiguration and C-SPOC, and to minimize unscheduled down-time with the fast recovery feature.

Eliminating Single Points of Failure in an HACMP Cluster

The HACMP 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 has become 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 in an HACMP Cluster

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

- Nodes
- Applications
- Networks and network adapters
- Disks and disk adapters

To be highly available, a cluster must have no single point 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. Again, the purpose of the HACMP for AIX software is to provide a cost-effective, highly available computing environment that can grow to meet future processing demands.

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 a node monitoring a neighbor node ceases to receive keepalive traffic for a defined period of time. If the other cluster nodes agree that the failure is a node failure, the failing node is removed from the cluster and its resources are taken over by the nodes configured to do so. An active node may, for example, take control of the failed node's shared disks. Or, an active node may masquerade as the failed node (by acquiring its service address) and run that node's processes while still maintaining its own processes. Thus, client applications can switch over to a surviving node for shared-disk and processor services.

The HACMP for AIX software provides the following facilities for processing node failure:

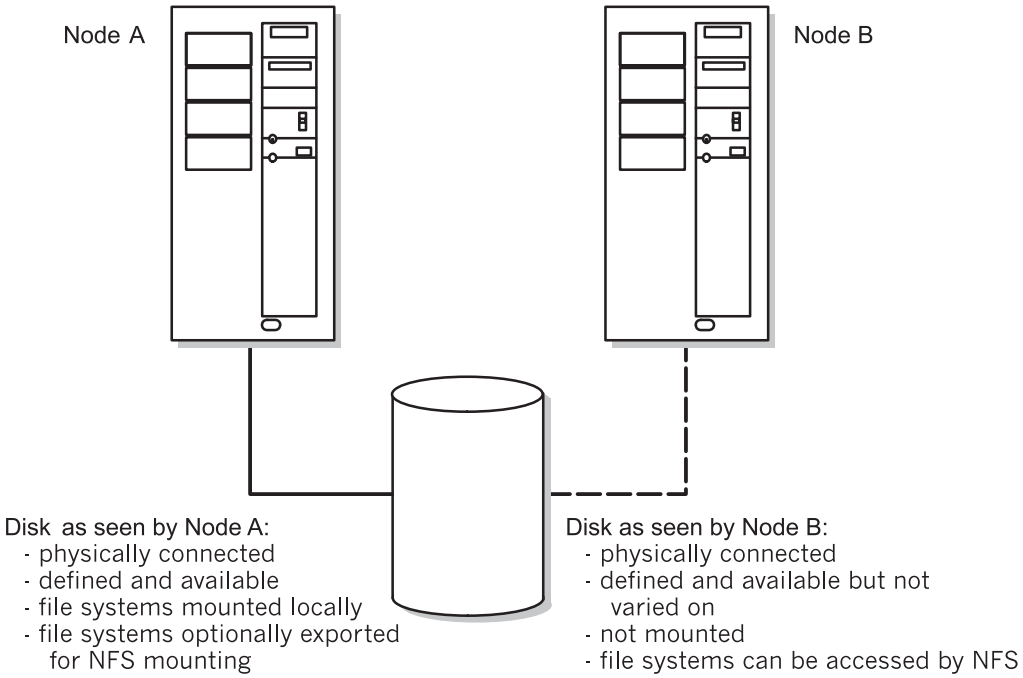
- Disk takeover
- IP address takeover (with or without hardware address swapping)

Disk Takeover

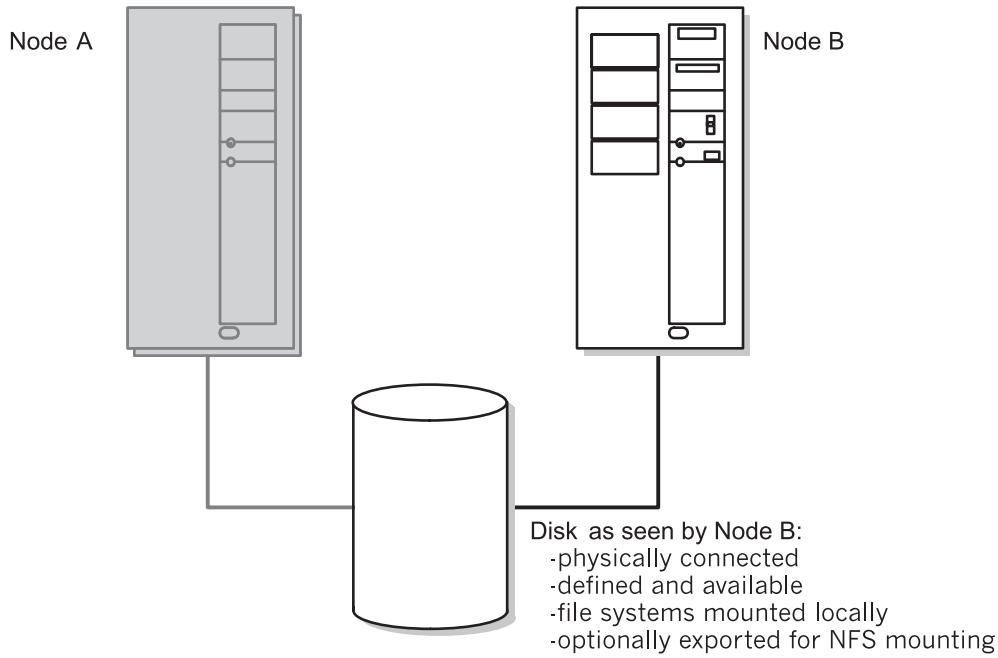
In an HACMP for AIX environment, shared disks are physically connected to multiple nodes. In non-concurrent environments, only one connection is active at any given time, and the node with the active connection owns the disk. *Disk takeover* occurs when the node that currently owns the disk leaves the cluster and an active node assumes control of the shared disk so that it remains available. Note, however, that shared filesystems can be exported and NFS mounted by other cluster nodes under HACMP's control.

In HACMP for AIX version 4.4, the **cl_export_fs** utility can use the optional **/usr/sbin/cluster/etc/exports** file instead of the standard **/etc/exports** file for determining export options. For more information on this capability, see the *HACMP for AIX Installation Guide*, NFS Exporting Filesystems and Directories on page 22-5.

In concurrent access configurations, the shared disks are actively connected to multiple nodes at the same time. Therefore, disk takeover is not required when a node leaves the cluster. The following figures illustrate disk takeover in non-concurrent environments.



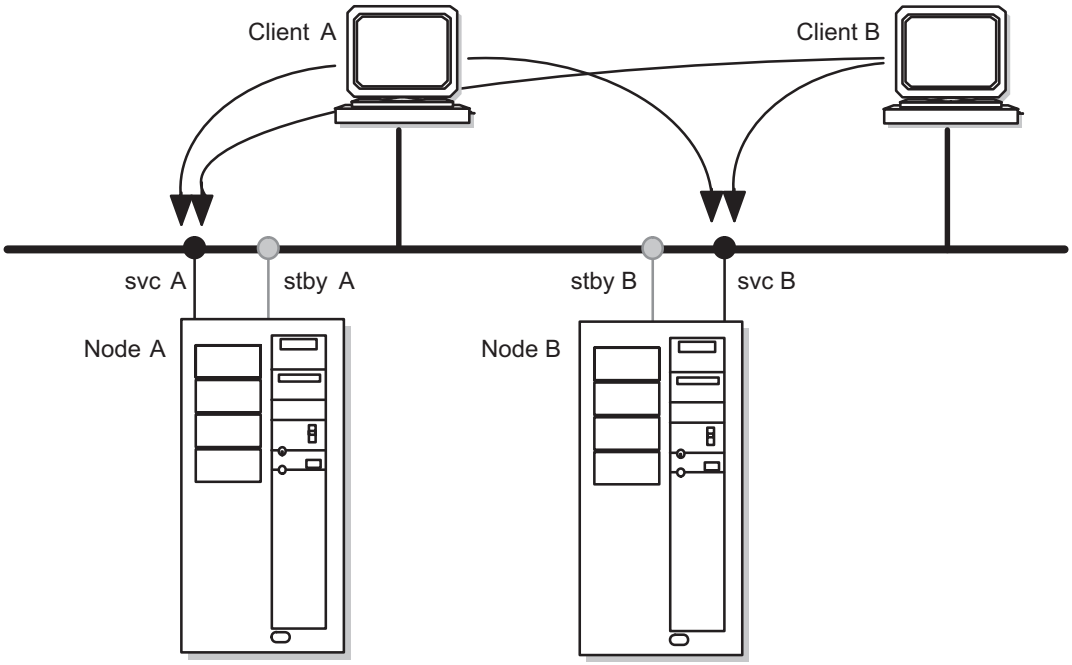
Before Disk Takeover



After Disk Takeover

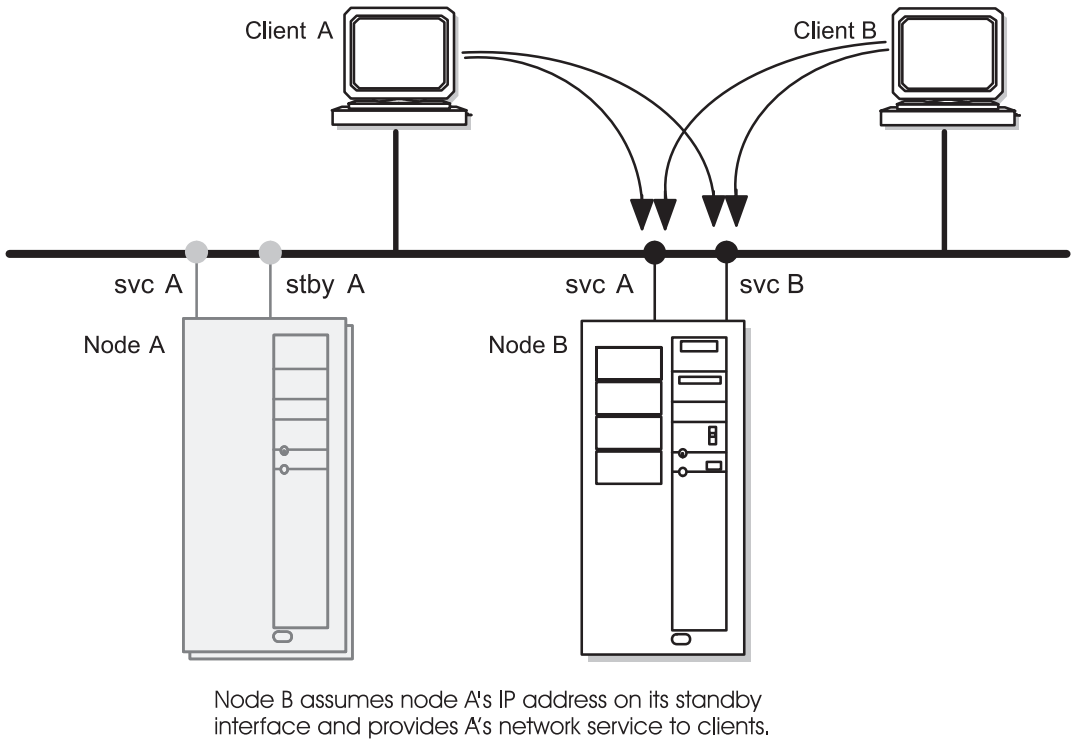
IP Address Takeover

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 necessary in an HACMP cluster when a service being provided to clients is bound to a specific IP address. If a surviving node simply did a disk and application takeover, clients would not be able to continue using the application at the specified server IP address. The following figures illustrate IP address takeover.



Each node provides a separate network service.

Before IP Address Takeover



After IP Address Takeover

Note: In an HACMP for AIX or HACMP/ES environment on the RS/6000 SP, standby adapters are not used by the SP switch, so special considerations apply to IP address takeover. See the *HACMP for AIX Installation Guide* or the *Enhanced Scalability Installation and Administration Guide* for more information.

Hardware Address Swapping and IP Address Takeover

Hardware address swapping works in conjunction with IP address takeover. With hardware address 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 contains a mapping of IP addresses to hardware addresses. The use of hardware address swapping is highly recommended for clients that cannot run the Clinfo daemon (machines not running AIX) or that cannot easily update their ARP cache.

Note: SP Switch networks do not support hardware address swapping. However, note that the SP switch network can be configured such that their IP devices update their ARP caches automatically when IP address takeover occurs. For more information, see the *HACMP for AIX Installation Guide* or the *Enhanced Scalability Installation and Administration Guide.*

Keep in mind that when an adapter swap occurs, the netmask of the service adapter is obtained by the standby adapter; thus, the netmask follows the service address. This means that the netmask for all adapters in an HACMP network must be the same to avoid communication problems between standby adapters after an address swap and during the subsequent release of the address acquired during takeover.

Communication problems occur when the standby adapter releases the service address. The adapter assumes its original address, but retains the netmask of the service address. This address reassignment causes the standby adapter to function on a different subnet from other standby adapters in the network. This netmask change can cause changes in the broadcast address and the routing information such that other standby adapters may now be unable to communicate on the same logical network.

Eliminating Applications as a Single Point of Failure

The primary reason to create HACMP clusters is to provide a highly available environment for mission-critical applications. For example, an HACMP cluster could run a database server program which services client applications. The clients send queries to the server program which responds to their requests by accessing a database, stored on a shared external disk.

In an HACMP for AIX cluster, these critical applications can be a single point of failure. To ensure the availability of these applications, the node configured to take over the resources of the node leaving the cluster should also restart these applications so that they remain available to client processes.

To put the application under HACMP control, you create an *application server* cluster resource that associates a user-defined name with the names of user-provided written scripts to start and stop the application. By defining an application server, HACMP for AIX can start another instance of the application on the takeover node when a fallover occurs. For more information about defining application servers, see the *HACMP for AIX Installation Guide*.

In HACMP/ES, you can also configure an *application monitor* to check for process death or other application failures and automatically take action to restart the application.

Note: Application takeover is usually associated with IP address takeover. If the node restarting the application also acquires the failed node's IP address, the clients only need to reconnect to the same server IP address. If the IP address was not taken over, the client needs to connect to the new server to continue accessing the application.

Additionally, you can use the AIX System Resource Controller (SRC) to monitor for the presence or absence of an application's daemon and to respond accordingly.

For more information and advice on ensuring high availability of applications with HACMP, see the appendix on Applications and HACMP in both the *HACMP for AIX Planning Guide* and in the *Enhanced Scalability Installation and Administration Guide*.

Applications Integrated with HACMP

Normally, you write customized scripts or define *application servers* to make your applications highly available. However, in the case of AIX Fast Connect for Windows, AIX Connections, and Communications Server for AIX (CS/AIX), you can use the SMIT interface to configure

the application as a cluster resource, making it highly available in the event of a node or adapter failure without writing additional scripts. In addition, when you configure these applications as resources, you can verify them using **clverify**, without having to simulate a node or adapter failure.

Note that you cannot configure AIX Fast Connect *and* AIX Connections in a single resource group.

AIX Fast Connect

AIX Fast Connect for Windows is an application that allows Windows PC clients to request files and print services from an AIX server. Fast Connect supports the transport protocol NetBIOS over TCP/IP, and allows file and other resource sharing with PCs running Windows NT, Windows 98, Windows 95, Windows For Workgroups, or OS/2 operating systems.

Through the SMIT interface, you can define Fast Connect services as cluster resources, so HACMP starts and stops the server during fallover, recovery, and resource group migration. For more information on configuring AIX Fast Connect in HACMP, see the *HACMP for AIX Installation Guide,* or the *Enhanced Scalability Installation and Administration Guide* if you have HACMP/ES.

AIX Connections

AIX Connections lets you share files, printers, applications, and other resources with PC and Mac workstations using transport protocols instead of TCP/IP. You still have AIX's multi-user and multi-tasking facilities, scalability, file and record locking features, and other security features. AIX Connections handles the NetWare and AppleTalk protocols as well as NetBIOS over TCP/IP and Net BEUI.

Configuring this application as a resource in a resource group keeps the protocols handled by AIX Connections highly available in the event of node or adapter failure. You configure AIX Connections realm/service pairs through the SMIT interface. For more information on configuring AIX Connections in HACMP, see the *HACMP for AIX Installation Guide,* or the *Enhanced Scalability Installation and Administration Guide* if you have HACMP/ES.

Communications Server for AIX

CS/AIX is a set of communications protocols that enable an AIX computer to participate in an SNA network that includes mainframes, PCs and other workstations. It is typically associated with legacy computer environments, given that a mainframe connection usually exists. CS/AIX is supported over a number of network types, including Token Ring, Ethernet and FDDI.

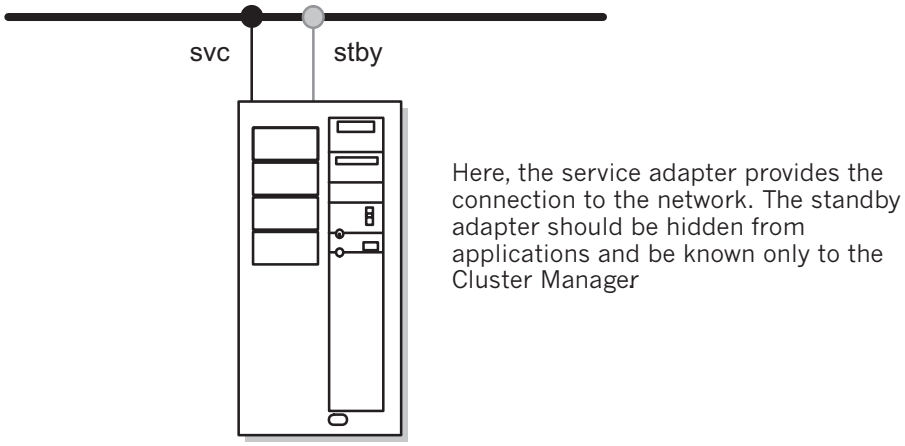
HACMP enables you to designate CS/AIX Data Link Control (DLC) profile(s) as highly available resources. In addition, you can specify associated CS/AIX objects, such as ports, link stations and applications, as highly available. This CS/AIX configuration information is preserved in the event of a node or adapter failure. You configure the highly available CS/AIX communication link(s) through the SMIT interface.

This feature is supported with the following two CS/AIX products: Communications Server for AIX, Version 4.2 and eNetwork Communications Server for AIX, Version 5.0. For more information on configuring highly available CS/AIX communication links, see the *HACMP for AIX Installation Guide,* or the *Enhanced Scalability Installation and Administration Guide* if you have HACMP/ES.

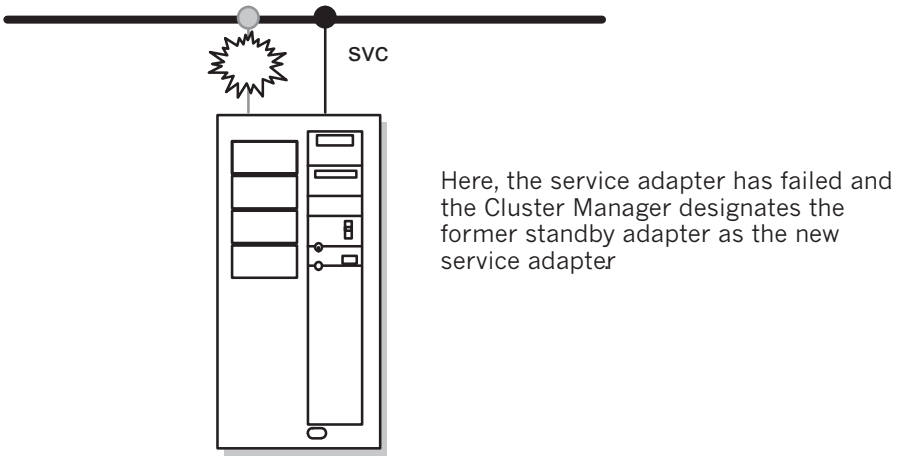
Eliminating Network Adapters as a Single Point of Failure

The HACMP 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 except for a small delay while the system reconfigures the adapter. While the Cluster Manager does detect a standby adapter failure, it only logs the event and sends a message to the system console. If you want additional processing, you can customize the processing for this event.

The following figures illustrate adapter swapping:



Before Network Adapter Swap



After Network Adapter Swap

Hardware Address Swapping and Adapter 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.

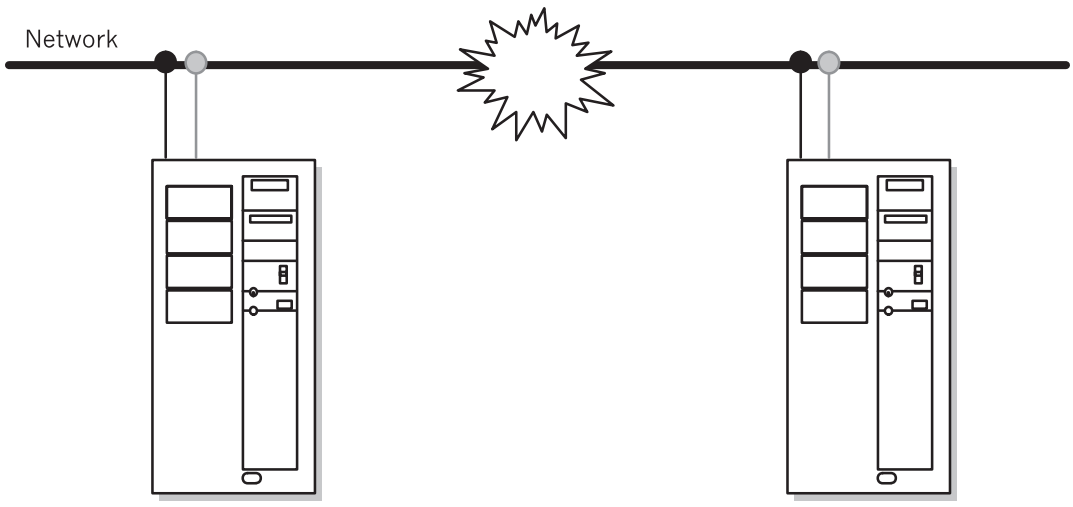
Without hardware address swapping, TCP/IP clients and routers which reside on the same subnet as the cluster nodes must have their Address Resolution Protocol (ARP) cache updated. The ARP cache contains a mapping of IP addresses to hardware addresses. The use of hardware address swapping is highly recommended for clients that cannot run the Clinfo daemon (machines not running AIX) or that cannot easily update their ARP cache.

Note: SP Switch networks do not support hardware address swapping. However, note that the SP switch network can be configured such that their IP devices update their ARP caches automatically when IP address takeover occurs. For more information, see the *HACMP for AIX Installation Guide* or the *Enhanced Scalability Installation and Administration Guide.*

Eliminating Networks as a Single Point of Failure

Network failure occurs when an HACMP network fails for all the nodes in a cluster. This type of failure occurs when none of the cluster nodes can access each other using the service or standby adapters configured for a given HACMP 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.

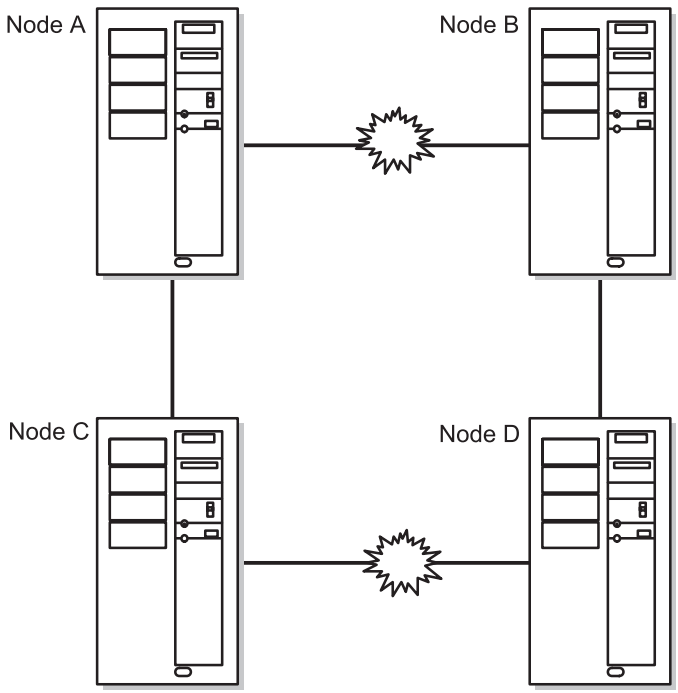
Network Failure

The HACMP 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 HACMP 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, re-routing 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 networks connecting two or more parts of the cluster fail. Each group (one or more) of nodes is completely isolated from the other groups. A cluster in which certain groups of nodes are unable to communicate with other groups of nodes is a *partitioned cluster*.

In the following illustration of a partitioned cluster, Node A and Node C are on one side of the partition and Node B and Node D are on the other side of the partition.



A Partitioned Cluster

The problem with a partitioned cluster is that the nodes on one side of the partition interpret the absence of keepalives from the nodes on the other side of the partition to mean that those nodes have failed and then generate node failure events for those nodes. Once this occurs, nodes on each side of the cluster (if so configured) attempt 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.

Using Serial Networks to Prevent Partitioning

To guard against the TCP/IP subsystem failure causing node isolation, each node in the cluster should be connected by a point-to-point serial network to its neighboring nodes, forming a logical "ring." This logical ring of serial networks reduces the chance of node isolation by allowing neighboring Cluster Managers to communicate even when all TCP/IP-based networks fail.

Serial networks are especially important in concurrent access configurations so that data does not become corrupted when TCP/IP traffic among nodes is lost.

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 Cluster Managers have accurate information about the status of peer nodes.

Using Global Networks to Prevent Partitioning

In the HACMP for AIX Enhanced Scalability subsystem (HACMP/ES), it is possible to configure a global network that groups multiple networks of the same type. Global networks help avoid node isolation when a network fails. For more information, see the *HACMP for AIX Enhanced Scalability Installation and Administration Guide.*

Eliminating Disks and Disk Adapters as a Single Point of Failure

The HACMP for AIX software does not itself directly handle disk and disk adapter failures. Rather, these failures are handled by AIX through LVM mirroring on disks and by internal data redundancy on the IBM 7135-110 and 7135-210 Disk Arrays.

For example, by configuring the system with multiple SCSI-2 Differential chains, serial adapters, and then mirroring 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 the IBM 7135-110 or 7135-210 Disk Arrays, the disk array itself is responsible for providing data redundancy.

The AIX Error Notification Facility

The AIX Error Notification facility allows you to detect an event not specifically monitored by the HACMP for AIX software—a disk adapter failure, for example—and to program a response to the event. For more information about using this facility, see the *HACMP for AIX Installation Guide*.

Permanent hardware errors on disk drives, controllers, or adapters can affect 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 a specific option in SMIT. By choosing this option, error notification will automatically be turned on or off on all nodes in the cluster for particular devices.

Note: 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 by this feature. One of two error notification methods is assigned for all error types supported by automatic error notification.

For more information on Automatic Error Notification, see the chapter on supporting AIX Error Notification in the *HACMP for AIX Installation Guide*.

Error Emulation

The Error Emulation utility allows you to test your error notification methods by simulating an error. When the emulation is complete, you can check whether your customized notify method was exercised as intended. For a full description of this feature, refer to the *HACMP for AIX Installation Guide*.

Minimizing Scheduled Down-Time with HACMP

The HACMP for AIX software enables you to perform most routine maintenance tasks on an active cluster dynamically—without having to stop and then restart cluster services to make the changed configuration the active configuration. Several features contribute to this:

- Dynamic reconfiguration (DARE)
- DARE resource migration
- Cluster Single Point of Control (C-SPOC)
- Dynamic adapter swap for replacing hot-pluggable adapter cards

Dynamic Reconfiguration (DARE)

This process, called *dynamic reconfiguration (DARE)*, is triggered when you synchronize the cluster topology or synchronize the cluster resource configuration after making changes on an active cluster. Applying a cluster snapshot using SMIT or the **xhacmpm** application also triggers a dynamic reconfiguration event. For more information about using SMIT or the **xhacmpm** application, see Chapter 6, Administrative Facilities.

For example, to add a node to a running cluster, you simply connect the node to the cluster, add the node to the cluster topology on any of the existing cluster nodes, and synchronize the cluster topology. The new node is added to the cluster topology definition on all cluster nodes and the changed configuration becomes the currently active configuration. After the dynamic reconfiguration event completes, you can start cluster services on the new node.

HACMP for AIX verifies the modified configuration before making it the currently active configuration to ensure that the changes you make result in a valid configuration.

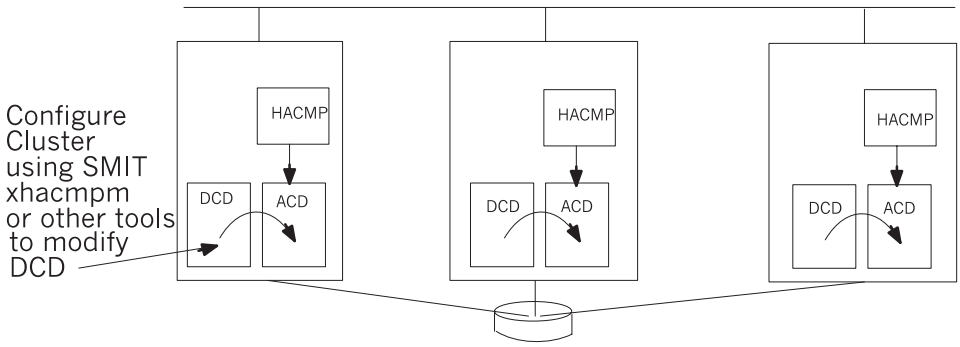
How Dynamic Reconfiguration Works

To allow for the dynamic reconfiguration of a running cluster, HACMP for AIX, whenever it starts, creates a private copy of the HACMP-specific object classes stored in the system default ODM. All the HACMP for AIX daemons, scripts, and utilities on a running node reference the ODM data stored in this private directory, called the Active Configuration Directory (ACD), instead of in the ODM data stored in the system default ODM, stored in the Default Configuration Directory (DCD).

By default, the DCD is the directory named **/etc/objrepos**. This directory contains the default system object classes, such as the customized device database (CuDv) and the predefined device database (PdDv), as well as the HACMP-specific object classes. By default, the ACD is **/usr/sbin/cluster/etc/objrepos/active**.

Note: When you configure a cluster, you modify the ODM data stored in the DCD—not data in the ACD. SMIT, **xhacmpm**, and other HACMP configuration utilities all modify the ODM data in the DCD. In addition, all user commands that display ODM data, such as the **ls** command, read data from the DCD.

The following figure illustrates how the HACMP daemons, scripts, and utilities all reference the ACD when accessing configuration information.



Relationship of HACMP for AIX to ACD at cluster start-up

Reconfiguring a Cluster Dynamically

The HACMP for AIX software depends on the location of certain ODM repositories to store configuration data. The presence or absence of these repositories are sometimes used to determine steps taken during cluster configuration and operation. The ODMPATH environment variable allows ODM commands and subroutines to query locations other than the default location (held in the ODMDIR environment variable) if the queried object does not exist in the default location. You can set this variable, but it must not be set to include the **/etc/objrepos** directory or you will lose the integrity of the HACMP configuration information.

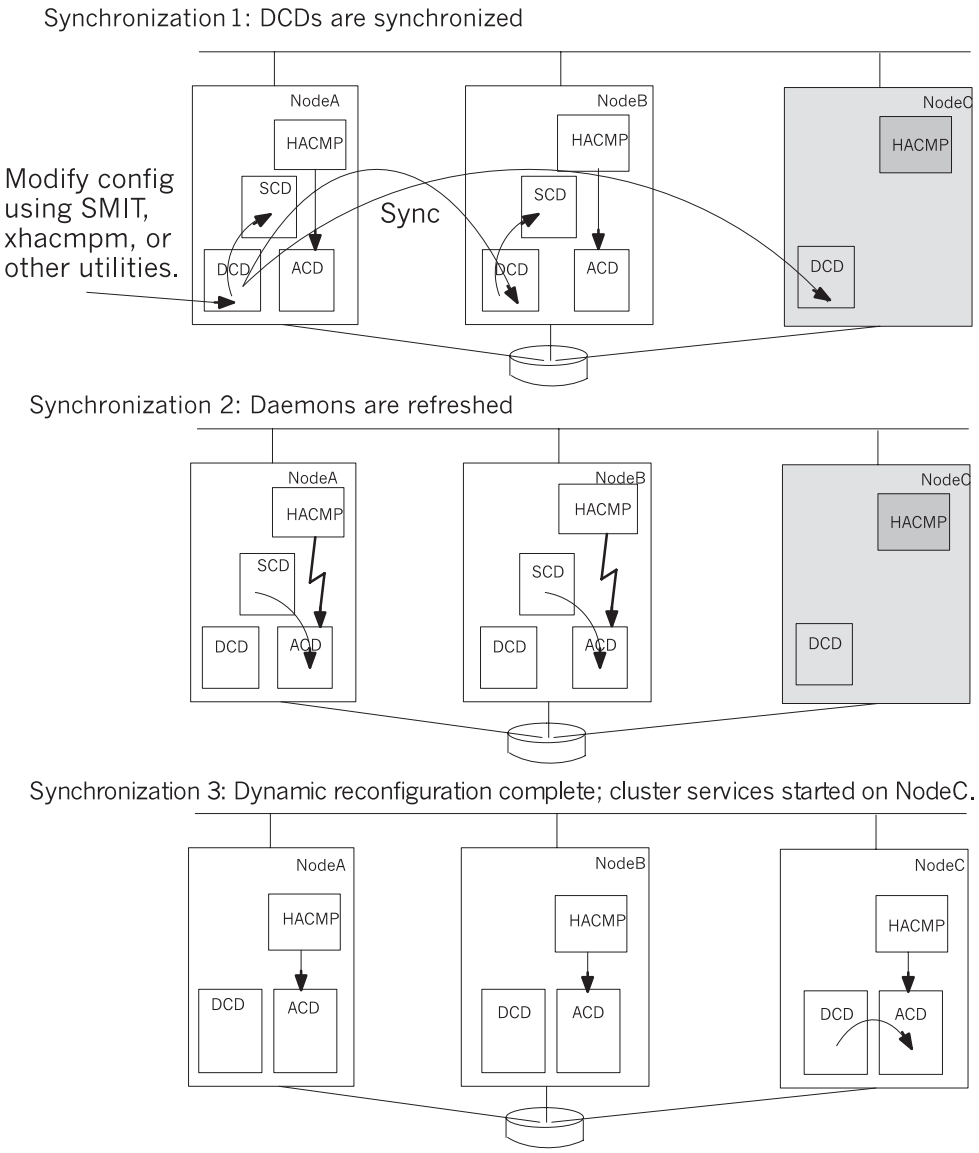
To change the configuration of an active cluster, you modify the cluster definition stored in the HACMP-specific ODM classes stored in the DCD using SMIT or the **xhacmpm** application. When you synchronize your configuration across all cluster nodes, a cluster-wide dynamic reconfiguration event occurs. When HACMP for AIX processes a dynamic reconfiguration

event, it updates the ODM object classes stored in the DCD on each cluster and replaces the ODM data stored in the ACD with the new ODM data in the DCD, in a coordinated, cluster-wide transition. It also refreshes the cluster daemons so that they reference the new configuration data.

After this processing, the cluster heartbeat is suspended briefly and the cluster is in an unstable state. The changed configuration becomes the active configuration. After cluster services are started on the newly added node, it can be integrated into the cluster.

The following figure illustrates the processing involved with adding a node to an active cluster using dynamic reconfiguration. The node to be added is connected to a running cluster but cluster services are inactive on this node. The configuration is redefined on NodeA. When the changes to the configuration are synchronized, the ODM data stored in the DCD on NodeA is copied to the DCDs on other cluster nodes and a dynamic reconfiguration event is triggered. HACMP for AIX copies the new ODM data in the DCD into a temporary location on each node, called the Staging Configuration Directory (SCD). The default location of the SCD is **/usr/sbin/cluster/etc/objrepos/stage**. By using this temporary location, HACMP for AIX allows you to start making additional configuration changes while a dynamic reconfiguration is in progress. Before copying the new ODM data in the SCD over the current ODM data in the ACD, HACMP for AIX verifies the new configuration.

Note: You can initiate a second reconfiguration while a dynamic reconfiguration is in progress, but you cannot synchronize it. The presence of an SCD on any cluster node acts as a lock, preventing the initiation of a new dynamic reconfiguration.



Dynamic Reconfiguration Processing

DARE Resource Migration

The HACMP for AIX software provides a Dynamic Reconfiguration (DARE) Resource Migration utility that allows you to change the location or status of resource groups (along with their resources—IP addresses, applications, and disks). You can perform DARE resource migrations using the **cldare** command at the command line, or through the SMIT interface. You also can emulate the dynamic migration of resource groups either the **cldare** command or SMIT.

For more information about DARE resource migration, see the overview in Chapter 6, Administrative Facilities. Also see Chapter 7, Changing Resources and Resource Groups, of the *HACMP for AIX Administration Guide* for full instructions on performing resource migrations from the command line and with SMIT.

Dynamic Adapter Swap

The dynamic adapter swap feature lets you swap the IP address of an active service or boot adapter with the IP address of a user-specified active, available standby adapter on the *same* node and network. 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.

The adapter swap feature is invoked through SMIT. The service/boot address is moved from its current adapter to a user-specified standby adapter. The service/boot address then becomes an available standby adapter. When the (hot pluggable) adapter to be replaced is pulled from the node, HACMP for AIX makes the adapter unavailable as a standby. When the new adapter card is placed in the node, the adapter is incorporated into the cluster as an available standby again. The feature can be invoked again through SMIT to swap the IP address from the standby back to the original adapter.

Note: The dynamic adapter swap feature is not supported on the SP switch network.

Note: This type of dynamic adapter swap can only be performed within a single node. You cannot swap the IP address of a service or boot adapter with the IP address of a standby adapter on a different node with this feature. To move an IP address to another node, move its resource group using the DARE Resource Migration utility.

See the *HACMP for AIX Administration Guide* for more information about dynamic adapter swapping.

Cluster Single Point of Control (C-SPOC)

With the C-SPOC utility, you can make changes to the whole cluster from a single cluster node. Instead of performing administrative tasks on each cluster node, you can use the SMIT interface to issue a C-SPOC command once, on a single node, and the change is propagated across all cluster nodes.

Minimizing Unscheduled Down-Time: Fast Recovery

Another important goal with HACMP is to minimize *unscheduled* down-time in response to unplanned cluster component failures. The HACMP software provides a *fast recovery* feature to minimize unplanned down-time.

Fast Recovery

The HACMP fast recovery feature speeds up fallover in large clusters.

Fast recovery lets you choose a filesystems consistency check and a filesystems recovery method:

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

  If a filesystem 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 filesystems in parallel, rather than in serial.

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

  The **varyonvg** and **varyoffvg** commands always run on volume groups in parallel, regardless of the setting of the recovery method.

Set your choices for these in the **SMIT hacmp > Cluster Configuration > Cluster Resources > Change/Show Resources for a Resource Group >** ***specify name of resource group*** **> Configure Resources for a Resource Group** screen. Your choices affect all filesystems in the resource group. If some filesystems need different settings, add them to separate resource groups.

See the *HACMP for AIX Installation Guide* for more information about fast recovery.

Chapter 4 Cluster Events

This chapter describes how the HACMP for AIX software responds to changes in a cluster to maintain high availability.

Cluster Events

An HACMP for AIX cluster environment is event-driven. An event is a change of status within a cluster that the Cluster Manager recognizes and processes. 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. When the Cluster Manager detects a change in cluster status, it executes a script designated to handle the event and its subevents.

The following are some examples of events the Cluster Manager recognizes:

- **node_up** and **node_up_complete** events (a node joining 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)
- dynamic reconfiguration 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 HACMP for AIX software provides a script for each event and subevent. The default scripts are located in the **/usr/sbin/cluster/events** directory.

By default, the Cluster Manager calls the corresponding event script supplied with the HACMP for AIX software for a specific event. You can specify additional processing to customize event handling for your site if needed. For more information, see Customizing Event Processing on page 4-2.

Processing Cluster Events

The two primary cluster events that HACMP for AIX software handles are fallover and reintegration. *Fallover* refers to the actions taken by the HACMP 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 to the cluster. Both types of actions are controlled by event scripts.

During event script processing, cluster-aware application programs see the state of the cluster as unstable.

Fallover

A fallover occurs when a resource group moves from its host node to another node because its host node leaves the cluster.

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 nodes still active in the cluster. When necessary, you can override the release and acquisition of resources (for example, to perform system maintenance).

Node failure begins when a node monitoring a neighboring node ceases to receive keepalive traffic for a defined period of time. If the other cluster nodes agree that the failure is a node failure, the failing node is removed from the cluster and its resources are taken over by the active nodes configured to do so.

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

Fallback

A fallback differs from a fallover in that it occurs specifically during a node up event. That is, a fallback occurs when a resource group moves to a node which has just joined the cluster.

For example, a cascading resource group with cascading without fallback not enabled “falls back” to a higher priority node as it joins a cluster. While a cascading without fallback resource group may fallover to another node, it will not fallback to another node.

Reintegration

When a node joins a running cluster, the cluster becomes temporarily unstable. The member nodes synchronize at the beginning of the join process and then run 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 Cluster Events

HACMP 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.

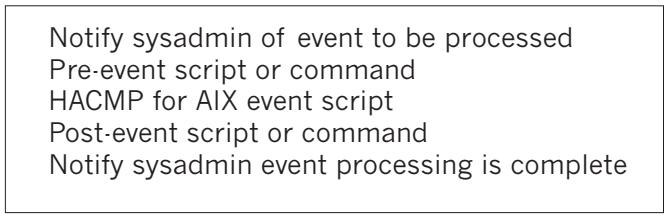
For more information on the Event Emulator utility, see Chapter 6, Administrative Facilities.

Customizing Event Processing

The HACMP for AIX software has an event customization facility you can use to tailor event processing. The Cluster Manager’s ability to recognize a specific series of events and subevents permits a very flexible customization scheme. Customizing event processing allows you to provide the most efficient path to critical resources should a failure occur.

You can define multiple pre- and post-events for each of the events defined in the HACMPevent ODM class.

Customization for an event could include notification to the system administrator before and after the event is processed, as well as user-defined commands or scripts before and after the event processing, as shown in the following figure.



A Customized Event

Use this facility for the following types of customization:

- Pre- and post-event processing
- Event notification
- Event recovery and retry.

Note: In HACMP for AIX, the event customization information stored in the ODM is synchronized across all cluster nodes when the cluster resources are synchronized. Thus, pre-, post-, notify, and recovery event script names must be the same on all nodes, although the actual processing done by these scripts can be different.

Defining New Events

In HACMP/ES, it is possible to define new events as well as tailor the existing ones. For more information on HACMP/ES and events, see the *Enhanced Scalability Installation and Administration Guide*.

Pre- and Post-Event Processing

To tailor event processing to your environment, specify commands or user-defined scripts that execute before and after a specific event is generated by the Cluster Manager. For pre-processing, for example, you may want to send a message to specific users, informing them to stand by while a certain event occurs. For post-processing, you may want to disable login for a specific group of users if a particular network fails.

Event Notification

You can specify a command or user-defined script that provides notification (for example, mail) that an event is about to happen and that an event has just occurred, along with the success or failure of the event.

Event Recovery and Retry

You can specify a command that attempts to recover from an event command failure. If the retry count is greater than zero and the recovery command succeeds, the event script command is rerun. You can also specify the number of times to attempt to execute the recovery command.

Chapter 5 Cluster Configurations

This chapter provides examples of the types of cluster configurations supported by the HACMP for AIX software.

Sample Cluster Configurations

There are two basic types of cluster configurations:

- *Standby configurations*—These are the traditional redundant hardware configurations where one or more standby nodes stand idle, waiting for a server node to leave the cluster.
- *Takeover configurations*—In these configuration, *all cluster nodes do useful work,* processing part of the cluster's workload. There are no standby nodes. Takeover configurations use hardware resources more efficiently than standby configurations since there is no idle processor. Performance can degrade after node detachment, however, since the load on remaining nodes increases.

  Takeover configurations that use *concurrent access* use hardware efficiently and also minimize service interruption during fallover because there is no need for the takeover node to acquire the resources released by the failed node—the takeover node already shares ownership of the resources.

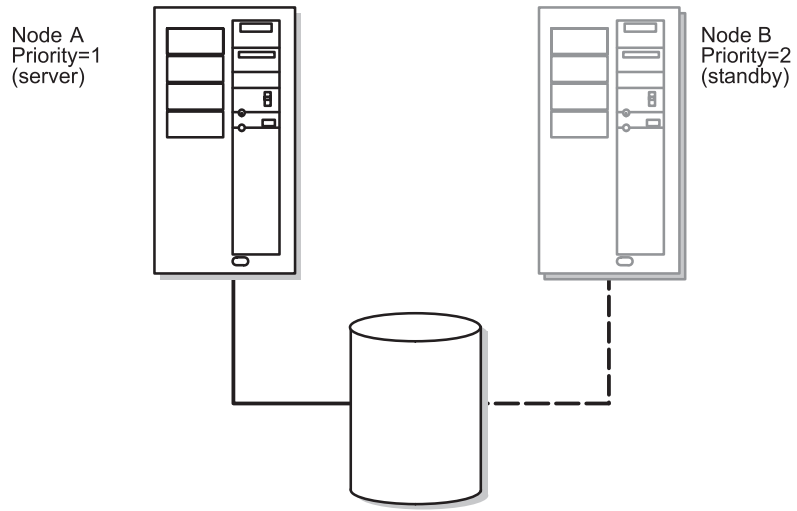
The sample cluster configurations shown in this chapter are by no means an exhaustive catalog of the possible configurations you can define using the HACMP for AIX software. Rather, use them as a starting point for thinking about the cluster configuration best suited to your environment.

Standby Configurations

The standby configuration is a traditional redundant hardware configuration, where one or more standby nodes stand idle, waiting for a 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. Concurrent resource groups, which require all nodes to have simultaneous access to the resource group, cannot be used in a standby configuration.

Standby Configurations with Cascading Resource Groups

The following figure shows a two-node standby configuration that uses cascading resource groups. In the figure, a lower number indicates a higher priority.



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 as consisting of two nodes. The first node, Node A, is assigned a takeover (ownership) priority of 1. The second node, Node B, is assigned a takeover priority of 2.

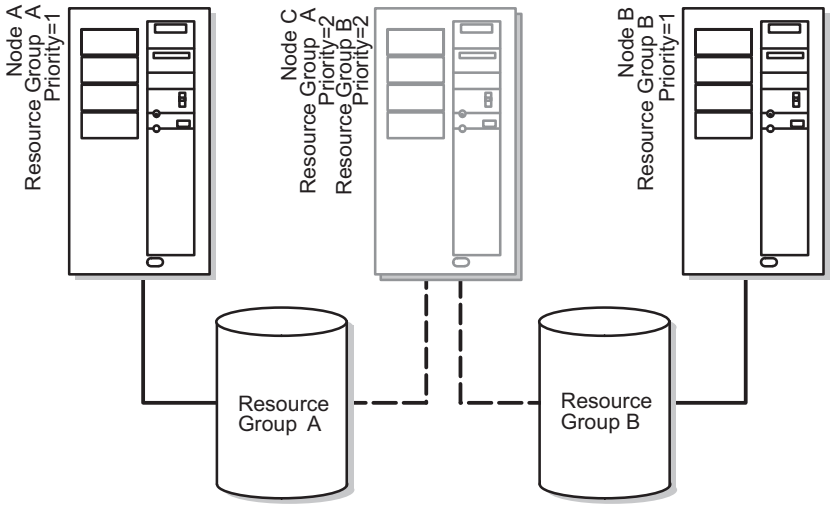
At cluster startup, Node A (which has a priority of 1) assumes ownership of the resource group. Node A is the "server" node. Node B (which has a priority of 2) stands idle, ready should Node A fail or leave the cluster. Node B is, in effect, the "standby."

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.

Extending Standby Configurations with Cascading Resource Groups

The standby with cascading resource groups configuration can be easily extended to larger clusters. The advantage of this configuration is that it makes better use of the hardware. The disadvantage is that the cluster can suffer severe performance degradation if more than one server node leaves the cluster.

The following figure illustrates a three-node standby configuration using cascading resource groups. In the figure, a lower number indicates a higher priority.



One-for-Two Standby Using Cascading Resource Groups

In this configuration, two separate resource groups (A and B) and a separate resource chain for each resource group exist. The chain for Resource Group A consists of Node A and Node C. Node A has a takeover priority of 1, while Node C has a takeover priority of 2. The chain for Resource Group B consists of Node B and Node C. Node B has a takeover priority of 1; Node C again has a takeover priority of 2. (Remember, a resource group can be owned by only a single node in a non-concurrent configuration.)

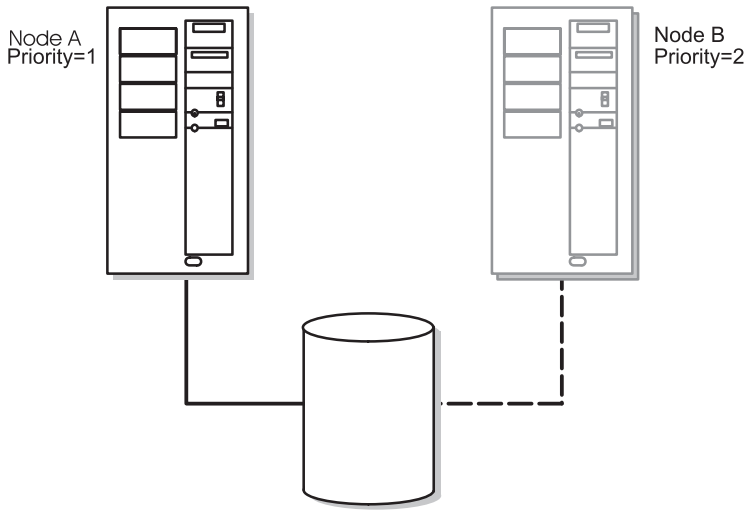
Since each resource group has a different node at the head of its priority chain, the cluster's workload is divided, or partitioned, between these two resource groups. Both resource groups, however, have the same node as the standby in their resource chains. If either server node leaves the cluster, the standby node assumes control of that server node's resource group and functions as the departed node.

In this example, the standby node has three network interfaces (not shown) and separate physical connections to each server node's external disk. Therefore, the standby node can, if necessary, take over for both server nodes concurrently. The cluster's performance, however, would most likely degrade while the standby node was functioning as both server nodes.

Standby Configurations with Rotating Resource Groups

A standby configuration with rotating resource groups differs from a cascading resource standby configuration in that the ownership priority of resource groups is not fixed. Rather, the resource group is associated with an IP address that can rotate among nodes. This makes the roles of server and standby fluid, changing over time.

The following figure shows a one-for-one standby configuration using rotating resource groups. In the figure, a lower number indicates a higher priority.



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 peer node assumes the shared IP address and claims ownership of that resource group. Now, the peer 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 the resource group 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 peer node leaves the cluster does the node that initially "owned" the resource group claim it once again. Thus, ownership of resources rotates between nodes.

Extending Standby Configurations with Rotating Resource Groups

As with cascading resource groups, configurations using rotating resource groups can be easily extended to larger clusters. For example, in a one-for-two standby configuration with rotating resource groups, the cluster could have two separate resource groups, each of which includes a distinct shared IP address.

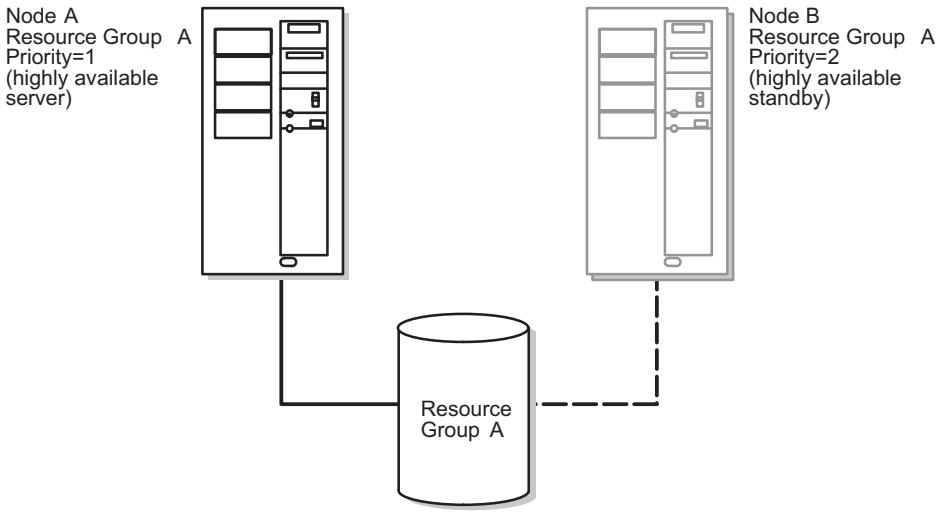
At cluster startup, the first two nodes each claim a shared IP address and assume ownership of the resource group associated with that shared IP address. The third node remains idle. If an active node leaves the cluster, the idle node claims that shared IP address and takes control of that resource group.

Takeover Configurations

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

One-Sided Takeover Using Cascading Resource Groups

The figure below illustrates a two-node, one-sided takeover configuration. In the figure, a lower number indicates a higher priority.



One-sided Takeover Using Cascading Resource Groups

This configuration has two 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 (for example, two server applications that handle client requests), only one set of resources needs to be highly available. This set of resources is defined as an HACMP 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 a priority of 1) assumes ownership of Resource Group A. Node A, in effect, "owns" Resource Group A. Node B (which has a priority of 2 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.

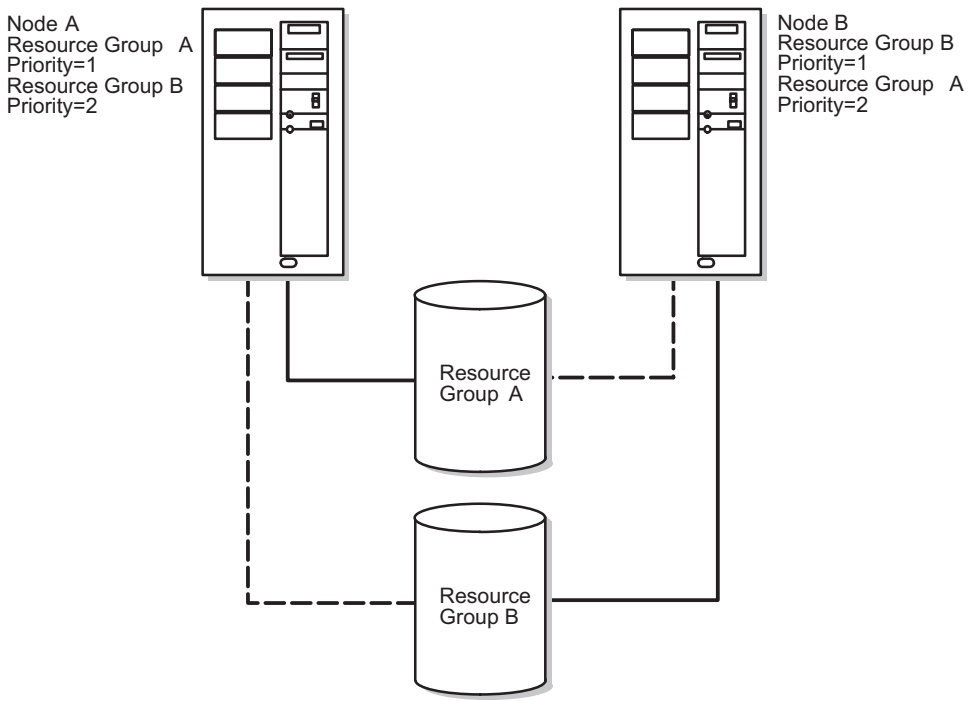
Mutual Takeover Using Cascading Resource Groups

The mutual takeover for non-concurrent access configuration has multiple nodes, each of which provides 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 a node leaves the cluster, a surviving node takes over the resource groups owned by the departed node.

The mutual takeover for non-concurrent access configuration is appropriate when each node in the cluster is running critical applications that need to be highly available and when each processor is able to handle the load of more than one node.

The following figure illustrates a two-node mutual takeover configuration for non-concurrent access. In the figure, a lower number indicates a higher priority.



Mutual Takeover for Non-Concurrent Access

The key feature of this configuration is that the cluster's workload is divided, or partitioned, between the nodes. Two resource groups exist, in addition to 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.

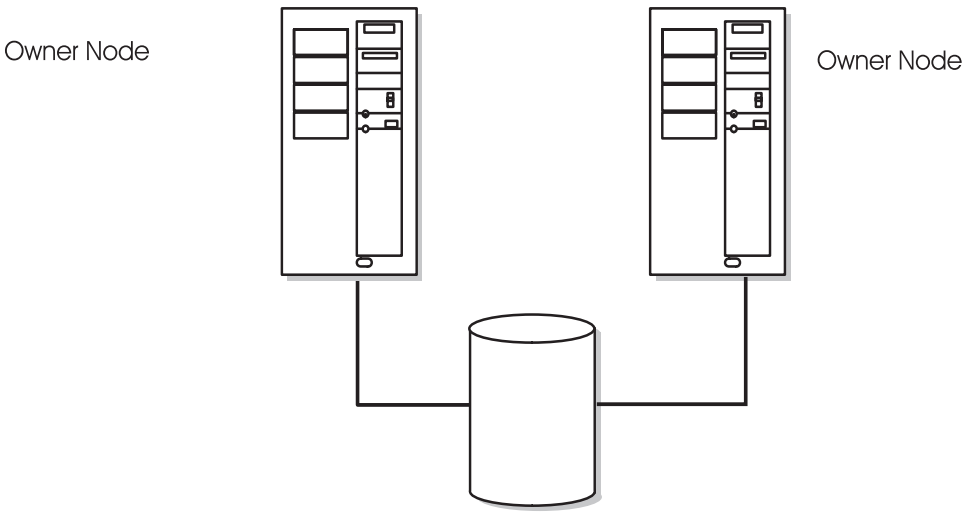
The chains for both resource groups consist of Node A and Node B. For Resource Group A, Node A has a takeover priority of 1 and Node B has a takeover priority of 2. For Resource Group B, the takeover priorities are reversed. Here, Node B has a takeover priority of 1 and Node A has a takeover priority of 2.

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 peer node takes control of the departed node's resource group. When the "owner" node for that resource group rejoins the cluster, the takeover node relinquishes the associated resources; they are reacquired by the higher-priority, reintegrating node.

Two-Node Mutual Takeover Configuration for Concurrent Access

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



Two-Node Mutual Takeover for Concurrent Access

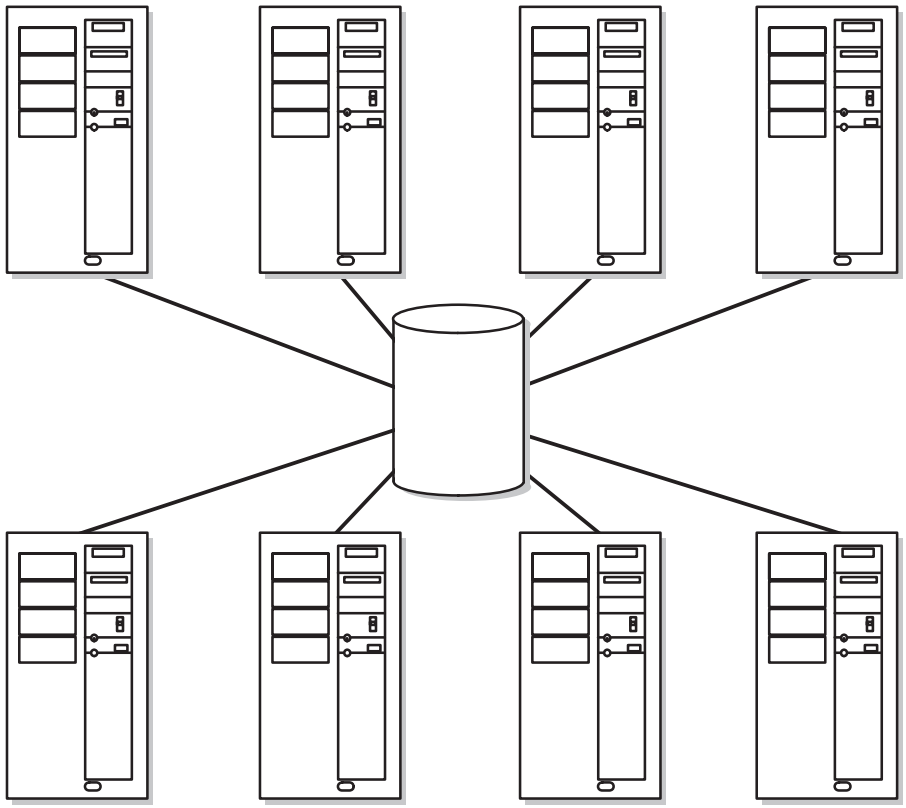
In this configuration, both nodes have simultaneous access to the shared disks and own the same disk resources. There is no "takeover" of shared disks if a node leaves the cluster, since the peer node already has the shared volume group varied on.

In this example both nodes are running an instance of a server application that accesses the database on the shared disk. The Cluster Lock Manager is used to arbitrate application requests for disk resources (which requires modifications to the application's code).

Running multiple instances of the same server application allows the cluster to distribute the processing load. As the load increases, additional nodes can be added to further distribute the load.

Eight-Node Mutual Takeover Configuration for Concurrent Access

The following figure illustrates an eight-node mutual takeover configuration for concurrent access:



Eight-Node Mutual Takeover for Concurrent Access

In this configuration, as in the previous configuration, all nodes have simultaneous—but not concurrent—access to the shared disks and own the same disk resources. Here, however, each node is running a different server application. Clients query a specific application at a specific IP address. Therefore, each application server and its associated IP address must be defined as part of a cascading resource group, and all nodes that are potential owners of that cascading resource group must be included in a corresponding resource chain.

Chapter 6 Administrative Facilities

This chapter describes the administrative tools supplied with the HACMP for AIX software for installing, configuring, and monitoring a cluster.

Overview

The HACMP for AIX software provides you with the following administrative facilities:

Installation and Configuration Tools

- System Management Interface Toolkit (SMIT)
- Cluster Single Point of Control (C-SPOC) utility to manage all cluster nodes from one node using a single command
- Quick Configuration utility for easy installation of predefined configurations
- Cluster Snapshot utility for saving existing cluster configurations
- TaskGuide graphical interface for creating shared volume groups
- Customized event processing, including multiple pre- and post-event definitions
- DARE Resource Migration utility, accessible through the command line and SMIT
- Online planning worksheets to help plan and configure your cluster
- Visual System Management (VSM) graphical configuration application

Monitoring and Diagnostic Tools

- HAView utility
- Cluster Monitoring with Tivoli functionality
- Cluster Status (**clstat**) utility
- Cluster Verification (**clverify**) utility
- Cluster Diagnostic utility
- HACMP for AIX log files
- HACMP for AIX cluster status information (**.info**) file, produced by the Cluster Snapshot utility
- Application Monitoring (HACMP/ES only)
- Enhanced Security for the secure execution of HACMP commands and scripts on remote nodes (on SP systems)
- Automatic Error Notification to configure a set of AIX Error Notification methods automatically in SMIT.

Emulation Tools

- HACMP for AIX Event Emulator
- Emulation of Error Log Driven Events

Installation and Configuration Tools

HACMP for AIX includes the tools described in the following sections for installing, configuring, and managing clusters.

Planning Worksheets

Along with your HACMP software and documentation set, you have two types of worksheets to aid in planning your cluster topology and resource configuration.

Online Worksheets

HACMP now provides online planning worksheets, installable on a PC with an appropriate web browser, that allow you to enter your configuration preferences as you plan. With the online worksheet program, after you have completed the planning process, you can press a button to create a script that actually applies the configuration to your cluster.

For more information on the browser requirements and instructions for using the online planning worksheets, see Appendix B of the *HACMP for AIX: Planning Guide* or the *Enhanced Scalability Installation and Administration Guide*.

Paper Worksheets

The HACMP documentation includes a set of planning worksheets to guide your entire cluster planning process, from cluster topology to resource groups and application servers. You can use these worksheets as guidelines when installing and configuring your cluster. You may find these paper worksheets useful in the beginning stages of planning, when your team might be around a conference table discussing various configuration decisions. The planning worksheets are found in appendix sections of the *HACMP for AIX: Planning Guide* and the *Enhanced Scalability Installation and Administration Guide*.

SMIT Interface

You can use the SMIT screens supplied with the HACMP for AIX software to perform the following tasks:

- Configure clusters, nodes, resources, and events
- Capture and restore snapshots of cluster configurations
- Read log files
- Diagnose cluster problems
- Manage a cluster using the C-SPOC utility
- Perform DARE migration events
- Configure Automatic Error Notification
- Perform dynamic adapter swap
- Configure cluster performance tuning.

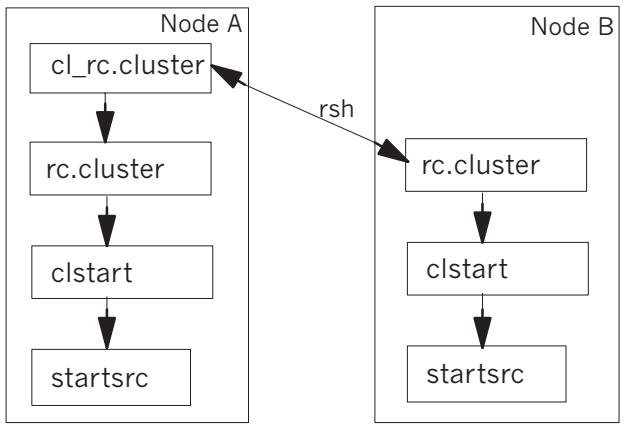
Cluster Single Point of Control (C-SPOC) Utility

The Cluster Single Point of Control (C-SPOC) utility lets system administrators perform administrative tasks on all cluster nodes from any node in the cluster. These tasks are based on commonly performed AIX system administration commands that let you:

- Maintain user and group accounts
- Maintain shared Logical Volume Manager (LVM) components
- Control HACMP services on a cluster-wide basis.

Without C-SPOC functionality, the system administrator must spend time executing administrative tasks individually on each cluster node. For example, to add a user you usually must perform this task on each cluster node. Using the C-SPOC utility, a command executed on one node is also executed on other cluster nodes. Thus C-SPOC minimizes administrative overhead and reduces the possibility of inconsistent node states. Using C-SPOC, however, you issue a C-SPOC command once on a single node, and the user is added to all specified cluster nodes.

C-SPOC also makes managing logical volume components and controlling cluster services more efficient. You can use the C-SPOC utility to start or stop cluster services on nodes from a single node. The following figure illustrates a two-node configuration and the interaction of commands, scripts, and nodes when starting cluster services from a single cluster node. Note the prefix **cl_** begins all C-SPOC commands.



Flow of Commands Used at Cluster Startup by C-SPOC Utility

Note: After the C-SPOC command executes on Node A and the node is stable, the **rsh** command is executed on the next node in sequence.

C-SPOC provides this functionality through its own set of cluster administration commands, accessible through SMIT menus and screens. To use C-SPOC, select the **Cluster System Management** option from the HACMP for AIX menu. See the *HACMP for AIX Administration Guide* for information on using C-SPOC SMIT options.

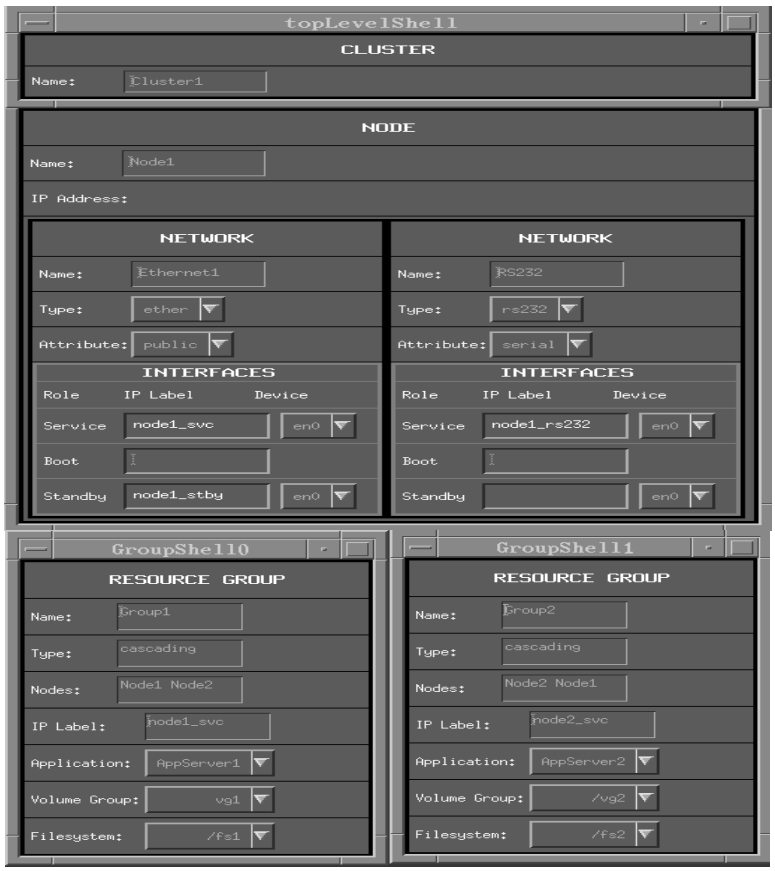
Quick Configuration Utility

The Quick Configuration utility, **xclconfig**, is an X Window System application designed to simplify the task of configuring an HACMP cluster. You use the Quick Configuration utility to configure one of the following four predefined two-node cluster configurations:

- Configuration 1: Two IBM 7204 SCSI-2 disk subsystems, rotating resource groups
- Configuration 2: Two IBM 7204 SCSI-2 disk subsystems, cascading resource groups
- Configuration 3: Two IBM 7137 SCSI-2 disk arrays, concurrent access resource groups
- Configuration 4: One IBM 7133 SSA disk subsystem, cascading resource groups
- Configuration 5: Two IBM 7131-105 SCSI-2 disk subsystems, cascading resource groups

The **xclconfig** utility automates the configuration process for you by displaying Customization Windows for the cluster, nodes, networks, interfaces, and resource groups.

As shown in the following figure, the **xclconfig** utility displays default values for a chosen configuration in these windows, which may appear on your display simultaneously after choosing a configuration. You can find detailed information about each configuration's values in the **/usr/sbin/cluster/snapshots** directory. Each configuration file contains a .odm extension.



HACMP for AIX Quick Configuration—Customization Windows

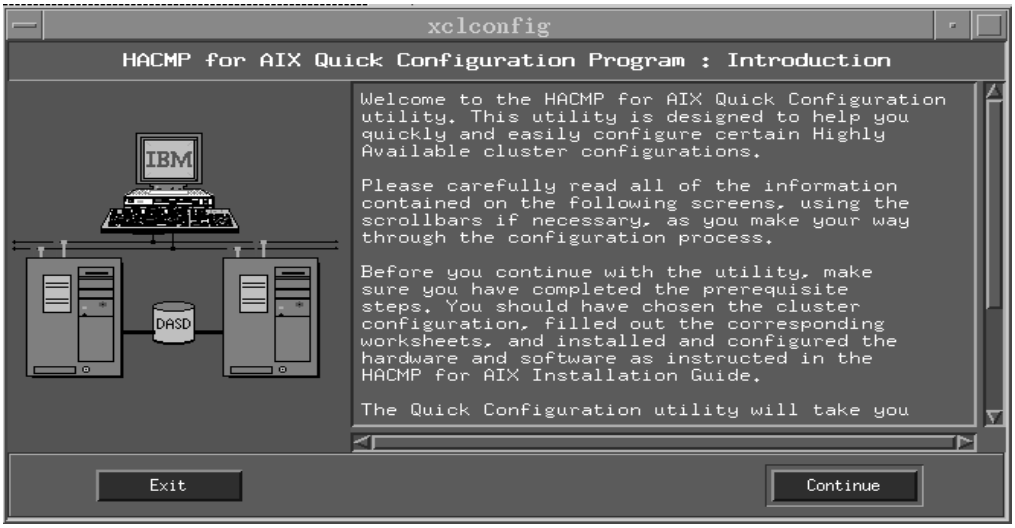
You can customize a configuration by changing the default names for the cluster, nodes, networks, and resource groups. You can change these values using a text editor or through the HACMP for AIX Quick Configuration—Customization windows. You also can verify that your changes are valid through the HACMP for AIX Quick Configuration—Verification screen. This automated approach to configuring two-node clusters makes configuring a cluster less time consuming for system administrators with limited time and equipment.

To start **xclconfig**, enter:

```
/usr/sbin/cluster/xclconfig
```

Note: Do not run xclconfig in the background, as the utility may hang when it reaches the IP address selection stage.

The main window appears at startup, as shown in the following figure. The Introduction screen shows an HACMP cluster. The Help Information provides introductory information on the application's function. Online help information is also provided with the utility.



The HACMP for AIX Quick Configuration Utility (xclconfig)

See the *HACMP for AIX Installation Guide* for more information on using the Quick Configuration utility.

Cluster Snapshot Utility

The Cluster Snapshot utility allows you to save cluster configurations you would like to later restore. You also can save additional system and cluster information that can be useful for diagnosing system or cluster configuration problems. You can create your own custom snapshot methods to store additional information about your cluster.

For more information about using the Cluster Snapshot utility, see the *HACMP for AIX Administration Guide*.

TaskGuide for Creating Shared Volume Groups

The TaskGuide is a web-based graphical interface that simplifies the task of creating a shared volume group within an HACMP cluster configuration. The TaskGuide presents a series of panels that guide the user through the steps of specifying initial and sharing nodes, disks, concurrent or non-concurrent access, volume group name and physical partition size, and cluster settings. The TaskGuide can reduce errors, as it does not allow a user to proceed with steps that conflict with the cluster's configuration. Online help panels give additional information to aid in each step.

The TaskGuide for creating shared volume groups was introduced in HACMP 4.3.0. In version 4.4, the TaskGuide has two enhancements: it creates a JFS log automatically, as you would do manually when creating a shared volume group without the TaskGuide, and it now displays the physical location of available disks.

You start the TaskGuide from the command line or through SMIT. For more information on starting and using the TaskGuide program, see the *HACMP for AIX Installation Guide* or the *HACMP for AIX Administration Guide.*

If you have the HACMP/ES product subsystem, see the *Enhanced Scalability Installation and Administration Guide*.

Customized Event Processing

You can define multiple pre- and post-events to tailor your event processing for your site's unique needs. For more information about writing your own scripts for pre- and post-events, see Chapter 4, Cluster Events, in this book, and the chapter on Tailoring Cluster Event Processing in the *HACMP for AIX: Planning Guide*.

VSM Graphical Configuration Application

The HACMP for AIX Visual System Management (VSM) application, **xhacmpm**, allows you to configure your cluster environment through its graphical interface. Using **xhacmpm**, a Common Desktop Environment (CDE) application, you can define attributes for clusters and cluster objects like nodes, resource groups, or application servers, and simultaneously and dynamically configure your cluster environment. You also can save snapshots of cluster configurations and restore them when needed.

The VSM application is included in the samples directory. For more information, see the *HACMP for AIX Administration Guide, Appendix D, VSM Graphical Configuration Application.*

DARE Resource Migration Utility

The HACMP software provides a Dynamic Reconfiguration (DARE) Resource Migration utility that allows a system administrator to change the status and location of one or more resource groups (along with their resources—IP addresses, applications, and disks) without bringing a node down. You can start, stop, or move resource groups in a running cluster. DARE migration is performed with the **cldare** command at command line, or through the SMIT interface.

Dynamic resource group movement helps you manage your cluster more effectively, giving you better use of your cluster hardware resources. Dynamic resource migration also enables you to perform selective maintenance without rebooting the cluster or disturbing operational nodes.

Using the DARE Resource Migration utility does not affect other resource groups currently owned by a node. The node that currently owns the resource group to be moved releases it as it would during a "graceful shutdown with takeover," and the node to which the resource group is being moved acquires the resource group just as it would during a node fallover.

See Chapter 7 of the *HACMP for AIX Administration Guide* for more information about performing DARE resource migrations, from the command line and through SMIT.

Monitoring and Diagnostic Tools

HACMP for AIX includes the following tools for monitoring clusters and diagnosing problems.

HAView Cluster Monitoring Utility

The HAView cluster monitoring utility makes use of the TME 10 NetView for AIX graphical interface to provide a set of visual maps and submaps of HACMP clusters. HAView extends NetView services to allow you to monitor HACMP clusters and cluster components across a network from a single node. HAView creates symbols that reflect the state of all nodes, networks, and network interface objects associated in a cluster. If you have HACMP/ES, you can also monitor resource groups and their resources through HAView.

The following figure shows the HAView Clusters symbol in the NetView Root map window.

The NetView Root Map with the Clusters Symbol

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 Information Base (MIB). The MIB is maintained by the Cluster SMUX peer daemon (**clsmuxpd**), the HACMP management agent. HAView allows you to:

- View maps and submaps of cluster symbols showing the location and status of nodes, networks, and addresses. (In HACMP/ES, HAView can monitor resource groups and resources as well.)
- View detailed information in NetView dialog boxes about a cluster, network, IP address, and cluster events.
- View cluster event history using the HACMP Event Browser
- View node event history using the Cluster Event Log
- Open a SMIT hacmp session for an active node and perform cluster administration functions from within HAView, using the HAView Cluster Administration facility.

For complete information about installing, configuring, and using HAView, see the *HACMP for AIX Installation Guide* and the chapter on monitoring your cluster in the *HACMP for AIX Administration Guide.* If you have HACMP/ES, see the chapters on installation and monitoring in the *Enhanced Scalability Installation and Administration Guide.*

Cluster Monitoring with Tivoli

You can monitor the state of an HACMP cluster and its components through your Tivoli Framework enterprise management system. Using various windows of the Tivoli Desktop, you can monitor the following aspects of your cluster:

- Cluster state and substate
- Configured networks and network state
- Participating nodes and node state
- Configured resource groups and resource group state (HACMP/ES only)
- Resource group location (HACMP/ES only)

For complete information about installing, configuring, and using the cluster monitoring through Tivoli functionality, see the *HACMP for AIX Installation Guide* and the chapter on monitoring a cluster in the *HACMP for AIX Administration Guide.* If you have HACMP/ES, see the chapters on installation and monitoring in the *Enhanced Scalability Installation and Administration Guide.*

Cluster Status Utility (clstat)

The Cluster Status utility, **/usr/sbin/cluster/utilities/clstat**, runs on both ASCII and X terminals and monitors cluster status. For the cluster as a whole, **clstat** indicates the cluster state and the number of cluster nodes. For each node, **clstat** displays the IP label and address of each service network interface attached to the node, and whether that interface is up or down.

Note: To use **clstat**, you must have Clinfo running on the client machine.

Cluster Verification Utility (clverify)

The Cluster Verification utility, **/usr/sbin/cluster/diag/clverify**, verifies that HACMP-specific modifications to AIX system files are correct, that the cluster and its resources are configured correctly, and that Kerberos security, if set up, is configured correctly. **clverify** also indicates whether custom cluster snapshot methods exist and whether they are executable on each cluster node.

Software and Cluster Verification

clverify has two categories of verification options. Running the *software* option ensures that the HACMP-specific modifications to the AIX software are correct. Verifying the *cluster* ensures that all resources used by the HACMP for AIX software are properly configured, and that ownership and takeover of those resources are assigned properly and are in agreement across all cluster nodes.

Custom Verification Methods

Through SMIT you also can add, change, or remove custom-defined verification methods that perform specific checks on your cluster configuration. See the *HACMP for AIX Administration Guide* for more information about performing these tasks.

You can perform verification from the command line or through SMIT. See the chapter on verifying a cluster configuration in the *HACMP for AIX Administration Guide* or the *Enhanced Scalability Installation and Administration Guide* for detailed information about running **clverify** from the command line or through the SMIT interface.

Cluster Diagnostic Utility

The Cluster Diagnostic utility, **cldiag**, provides a common interface to several HACMP and AIX diagnostic tools you can use to troubleshoot an HACMP cluster. Use **cldiag** to perform the following tasks:

- View the cluster log files for error and status messages.
- Activate and deactivate Cluster Manager debug mode.
- Obtain a listing of all locks in the Cluster Lock Manager's lock resource table.
- Check volume group definitions.
- Activate and deactivate tracing of HACMP for AIX daemons.

Log Files

The HACMP for AIX software writes the messages it generates to the system console and to several log files. Because each log file contains a different level of detail, system administrators can focus on different aspects of HACMP for AIX processing by viewing different log files. The HACMP for AIX software writes messages into the log files described in the following sections.

Note that each log file has a default directory. You may redirect a log file to a storage location other than its default directory if you choose. For more information about specifying custom destination directories for log files, see the chapter on customizing events and log files in the *HACMP for AIX Installation Guide*.

/usr/adm/cluster.log File

The **/usr/adm/cluster.log** file contains time-stamped, formatted messages generated by HACMP for AIX scripts and daemons.

/tmp/hacmp.out File

The **/tmp/hacmp.out** file contains messages generated by HACMP for AIX scripts.

In verbose mode, this log file contains a line-by-line record of every command executed by these scripts, including the values of all arguments to these commands.

System Error Log File

The system error log contains time-stamped, formatted messages from all AIX subsystems, including HACMP for AIX scripts and daemons.

/usr/sbin/cluster/history/cluster.*mmdd* File

The **/usr/sbin/cluster/history/cluster.*mmdd*** file contains time-stamped, formatted messages generated by HACMP for AIX scripts. The system creates a cluster history file every day, identifying each file by the file name extension, where *mm* indicates the month and *dd* indicates the day.

/tmp/cm.log File

Contains time-stamped, formatted messages generated by HACMP for AIX **clstrmgr** activity.

/tmp/cspoc.log File

Contains time-stamped, formatted messages generated by C-SPOC commands.

/tmp/emuhacmp.out File

Contains time-stamped, formatted messages generated by the HACMP for AIX event emulator scripts.

HACMP for AIX Cluster Status Information File

When you use the HACMP for AIX Cluster Snapshot utility to save a record of a cluster configuration (as seen from each cluster node), you cause the utility to run many standard AIX commands and HACMP for AIX commands to obtain status information about the cluster. This information is stored in a file, identified by the **.info** extension, in the snapshots directory. The snapshots directory is defined by the value of the SNAPSHOTPATH environment variable. By default, the cluster snapshot utility includes the output from the commands, such as **cllssif**, **cllsnw**, **df**, **ls**, and **netstat**. You can create custom snapshot methods to specify additional information you would like stored in the .info file.

Enhanced Security Utility

Enhanced security lets you execute HACMP commands on remote nodes more securely, removing the requirement for TCP/IP access control lists (for example, the **/.rhosts** file) during HACMP configuration.

Enhanced security requires the global ODM utilities and remote command execution commands (**rsh**, **rcp**) to use Kerberos Version 4 authentication to grant services. HACMP queries the local ODM about the security mode. If you set the Cluster Security Mode to **Enhanced** in the Change/Show Cluster Security SMIT screen, HACMP uses Kerberos authenticating commands. If you set the mode to **Standard**, HACMP uses **/.rhosts** files for authentication.

Note: IBM supports Kerberos Version 4 on IBM Scalable POWERParallel (RS/6000 SP) systems only.

For more information on configuring Kerberos 4 security (on an RS/6000 SP system), see the *HACMP for AIX Installation Guide,* Appendix J, Installing and Configuring HACMP for AIX on RS/6000 SPs.

DCE Authentication

As of AIX 4.3.1, Kerberos Version 5 (DCE authentication) can be used on SP and non-SP RS/6000 systems. However, if Kerberos 5 is the *only* method of authentication, you will not be able to alter, synchronize, or verify the HACMP configuration. Note that you will not be able to explicitly move a resource group, since that is a type of reconfiguration. If DCE (i.e. only Kerberos V5) is enabled as an authentication method for the AIX remote commands, you can still use HACMP, but must perform additional steps to ensure proper HACMP functioning.

For more information, see Chapter 22, Additional AIX Administrative Tasks, in the *HACMP for AIX Installation Guide* or the *Enhanced Scalability Installation and Administration Guide.*

Automatic Error Notification

You can make use of the AIX Error Notification facility to detect events not specifically monitored by the HACMP for AIX software—a disk adapter failure, for example—and specify a response to take place if the event occurs.

Normally, you must define error notify methods manually, one by one. HACMP provides a set of pre-specified notify methods for important errors that you can automatically “turn on” in one step through the SMIT interface, saving considerable time and effort by not having to define each notify method manually.

Emulation Tools

HACMP for AIX includes the Event Emulator for running cluster event emulations and the Error Emulation functionality for testing notify methods.

HACMP for AIX Event Emulator

The HACMP for AIX Event Emulator is a utility that emulates cluster events and dynamic reconfiguration events by running event scripts that produce output but that do not affect the cluster configuration or status. Emulation allows you to predict a cluster’s reaction to a particular event just as though the event actually occurred.

The Event Emulator follows the same procedure used by the Cluster Manager given a particular event, but does not execute any commands that would change the status of the Cluster Manager. For descriptions of cluster events and how the Cluster Manager processes these events, see the *HACMP for AIX Planning Guide.*

You can run the Event Emulator through SMIT or from the command line. The Event Emulator runs the events scripts on every active node of a stable cluster, regardless of the cluster's size. The output from each node is stored in an output file on the node from which the event emulator is invoked. You can specify the name and location of the output file using the environment variable **EMUL_OUTPUT**, or use the default output file, **/tmp/emuhacmp.out**.

Note: The Event Emulator requires that both the Cluster SMUX peer daemon (**clsmuxpd**) and the Cluster Information Program (**Clinfo**) be running on your cluster.

The events emulated are categorized in two groups:

- Cluster events
- Dynamic reconfiguration events.

Emulating Cluster Events

The cluster events that can be emulated are:

- Node up
- Node down
- Network up
- Network down
- Fail standby
- Join standby
- Swap adapter

For information on emulating cluster events using SMIT, see the *HACMP for AIX Administration Guide*. For information on emulating cluster events from the command line, see the **cl_emulate** man page.

Emulating Dynamic Reconfiguration Events

The dynamic reconfiguration events that can be emulated are:

- Synchronize Cluster Topology
- Synchronize Cluster Resources.

For more information about emulating dynamic reconfiguration events, see the *HACMP for AIX Administration Guide*, the sections on dynamic reconfiguration and also Appendix B. For information on emulating dynamic reconfiguration events from the command line, also see the **cldare** man page.

Restrictions on Event Emulation

The Event Emulator has the following restrictions:

- You can only run one instance of the event emulator at a time. If you attempt to start a new emulation in a cluster while an emulation is already running, the integrity of the results cannot be guaranteed.
- **clinfo** must be running.
- You cannot run successive emulations. Each emulation is a standalone process; one emulation cannot be based on the results of a previous emulation.
- When you run an event emulation, the Emulator's outcome may be different from the cluster manager's reaction to the same event under certain conditions:
  - The Event Emulator will not change the configuration of a cluster device. Therefore, if your configuration contains a process that makes changes to the Cluster Manager (disk fencing, for example), the Event Emulator will not show these changes. This could lead to a different output, especially if the hardware devices cause a fallover.
  - The Event Emulator runs customized scripts (pre- and post-event scripts) associated with an event, but does not execute commands within these scripts. Therefore, if these customized scripts change the cluster configuration when actually run, the outcome may differ from the outcome of an emulation.
- When emulating an event that contains a customized script, the Event Emulator uses the **ksh** flags **-n** and **-v**. The **-n** flag reads commands and checks them for syntax errors, but does not execute them. The **-v** flag indicates verbose mode. When writing customized scripts that may be accessed during an emulation, be aware that the other **ksh** flags may not be compatible with the **-n** flag and may cause unpredictable results during the emulation. See the **ksh** man page for flag descriptions.

Emulation of Error Log Driven Events

Although the HACMP for AIX software does not monitor the status of disk resources, it does provide a SMIT interface to the AIX Error Notification facility, as described on page 3-11.

The AIX Error Notification facility allows you to detect an event not specifically monitored by the HACMP for AIX software—a disk adapter failure, for example—and to program a response (notification method) to the event.

HACMP provides a utility for testing your error notify methods. After you have added one or more error notify methods with the AIX Error Notification facility, you can test your methods by emulating an error. By inserting an error into the AIX error device file (/dev/error), you cause the AIX error daemon to run the appropriate pre-specified notify method. This allows you to determine whether your pre-defined action is carried through, without having to actually cause the error to occur.

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 then be able to determine whether the specified notify method was carried out.

For more information on Error Emulation functionality, see the chapter on supporting AIX Error Notification, in the *HACMP for AIX Installation Guide,* or in the *Enhanced Scalability Installation and Administration Guide*.

Index

D

E

O

P

Q

R

S

T

V

W

XYZ

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

High Availability Cluster Multi-Processing for AIX:

Concepts and Facilities

SC23-4276-02

Overall, how satisfied are you with the information in this book?

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
Overall satisfaction	☐	☐	☐	☐	☐

How satisfied are you that the information in this book is:

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
Accurate	☐	☐	☐	☐	☐
Complete	☐	☐	☐	☐	☐
Easy to find	☐	☐	☐	☐	☐
Easy to understand	☐	☐	☐	☐	☐
Well organized	☐	☐	☐	☐	☐
Applicable to your tasks	☐	☐	☐	☐	☐

Please tell us how we can improve this book:

Thank you for your responses. May we contact you?

Yes ☐ No ☐

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

Name _____ Address _____

Company or Organization _____ _____

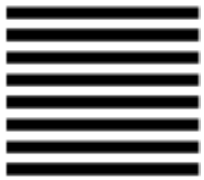
Phone No. _____ _____

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

SC23-4276-02

Fold and Tape | Please do not staple | Fold and Tape

NO POSTAGE
NECESSARY
IF MAILED IN THE
UNITED STATES



BUSINESS REPLY MAIL

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

POSTAGE WILL BE PAID BY ADDRESSEE

IBM CORPORATION
Publications Department
Internal Zip 9561
11400 Burnet Road
Austin, TX
78758-3493

Cut or Fold Along Line

Fold and Tape | Please do not staple | Fold and Tape