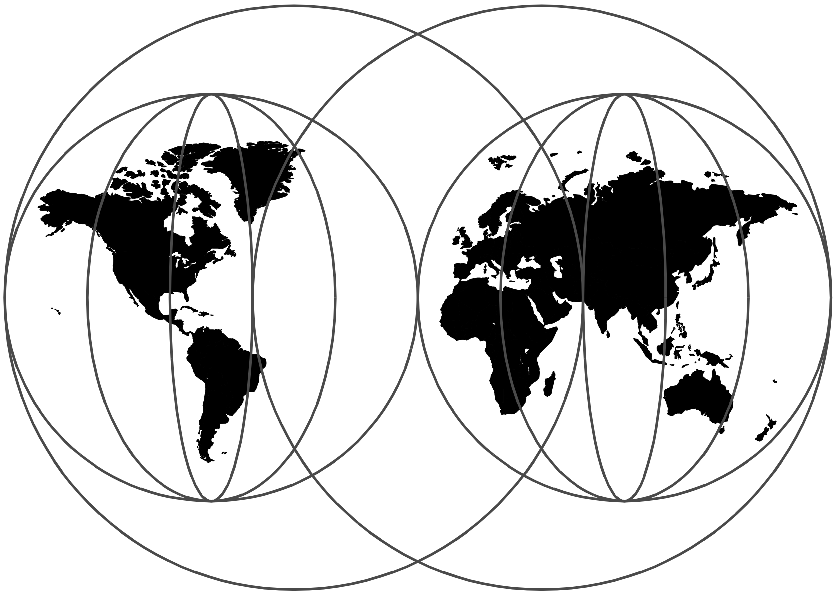
IBM Certification Study Guide AIX HACMP

David Thiessen, Achim Rehor, Reinhard Zettler



International Technical Support Organization

http://www.redbooks.ibm.com

SG24-5131-00

SG24-5131-00

International Technical Support Organization

IBM Certification Study Guide AIX HACMP

May 1999

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix A, “Special Notices” on page 205.

First Edition (May 1999)

This edition applies to HACMP for AIX and HACMP/Enhanced Scalability (HACMP/ES), Program Number 5765-D28, for use with the AIX Operating System Version 4.3.2 and later.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. JN9B Building 003 Internal Zip 2834
11400 Burnet Road
Austin, Texas 78758-3493

Contents

Figures

Tables

Preface

The AIX and RS/6000 Certifications offered through the Professional Certification Program from IBM are designed to validate the skills required of technical professionals who work in the powerful and often complex environments of AIX and RS/6000. A complete set of professional certifications is available. It includes:

- IBM Certified AIX User
- IBM Certified Specialist - RS/6000 Solution Sales
- IBM Certified Specialist - AIX V4.3 System Administration
- IBM Certified Specialist - AIX V4.3 System Support
- IBM Certified Specialist - RS/6000 SP
- IBM Certified Specialist - AIX HACMP
- IBM Certified Specialist - Domino for RS/6000
- IBM Certified Specialist - Web Server for RS/6000
- IBM Certified Specialist - Business Intelligence for RS/6000
- IBM Certified Advanced Technical Expert - RS/6000 AIX

Each certification is developed by following a thorough and rigorous process to ensure the exam is applicable to the job role and is a meaningful and appropriate assessment of skill. Subject Matter Experts who successfully perform the job participate throughout the entire development process. These job incumbents bring a wealth of experience into the development process, thus making the exams much more meaningful than the typical test, which only captures classroom knowledge. These Subject Matter experts ensure the exams are relevant to the *real world* and that the test content is both useful and valid. The result is a certification of value that appropriately measures the skill required to perform the job role.

This redbook is designed as a study guide for professionals wishing to prepare for the certification exam to achieve IBM Certified Specialist - AIX HACMP.

The AIX HACMP certification validates the skills required to successfully plan, install, configure, and support an AIX HACMP cluster installation. The requirements for this include a working knowledge of the following:

- Hardware options supported for use in a cluster, along with the considerations that affect the choices made

- AIX parameters that are affected by an HACMP installation, and their correct settings
- The cluster and resource configuration process, including how to choose the best resource configuration for a customer requirement
- Customization of the standard HACMP facilities to satisfy special customer requirements
- Diagnosis and troubleshooting knowledge and skills

This redbook helps AIX professionals seeking a comprehensive and task-oriented guide for developing the knowledge and skills required for the certification. It is designed to provide a combination of theory and practical experience. It also provides sample questions that will help in the evaluation of personal progress and provide familiarity with the types of questions that will be encountered in the exam.

This redbook will not replace the practical experience you should have, but, when combined with educational activities and experience, should prove to be a very useful preparation guide for the exam. Due to the practical nature of the certification content, this publication can also be used as a desk-side reference. So, whether you are planning to take the AIX HACMP certification exam, or just want to validate your HACMP skills, this book is for you.

For additional information about certification and instructions on How to Register for an exam, call IBM at 1-800-426-8322 or visit our Web site at: `http://www.ibm.com/certify`

The Team That Wrote This Redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization Austin Center.

David Thiessen is an Advisory Software Engineer at the International Technical Support Organization, Austin Center. He writes extensively and teaches IBM classes worldwide on all areas of high availability and clustering. Before joining the ITSO six years ago, David worked in Vancouver, Canada as an AIX Systems Engineer.

Achim Rehor is a Software Service Specialist in Mainz/Germany. He is Team Leader of the HACMP/SP Software Support Group in the European Central Region (Germany, Austria and Switzerland). Achim started working with AIX in 1990, just as AIX Version 3 and the RISC System/6000 were first being introduced. Since 1993, he has specialized in the 9076 RS/6000 Scalable

POWERparallel Systems area, known as the SP1 at that time. In 1997 he began working on HACMP as the Service Groups for HACMP and RS/6000 SP merged into one. He holds a diploma in Computer Science from the University of Frankfurt in Germany. This is his first redbook.

Reinhard Zettler is an AIX Software Engineer in Munich, Germany. He has two years of experience working with AIX and HACMP. He has worked at IBM for two years. He holds a degree in Telecommunication Technology. This is his first redbook.

Thanks to the following people for their invaluable contributions to this project:

Marcus Brewer
International Technical Support Organization, Austin Center

Rebecca Gonzalez
IBM AIX Certification Project Manager, Austin

Milos Radosavljevic
International Technical Support Organization, Austin Center

Comments Welcome

Your comments are important to us!

We want our redbooks to be as helpful as possible. Please send us your comments about this or other redbooks in one of the following ways:

- Fax the evaluation form found in “ITSO Redbook Evaluation” on page 221 to the fax number shown on the form.
- Use the electronic evaluation form found on the Redbooks Web sites:

For Internet users	`http://www.redbooks.ibm.com`
For IBM Intranet users	`http://w3.itso.ibm.com`

- Send us a note at the following address:

  `redbook@us.ibm.com`

Chapter 1. Certification Overview

This chapter provides an overview of the skill requirements for obtaining an IBM Certified Specialist - AIX HACMP certification. The following chapters are designed to provide a comprehensive review of specific topics that are essential for obtaining the certification.

1.1 IBM Certified Specialist - AIX HACMP

This certification demonstrates a proficiency in the implementation skills required to plan, install, and configure AIX High Availability Cluster Multi-Processing (HACMP) systems, and to perform the diagnostic activities needed to support Highly Available Clusters.

Certification Requirement (two Tests):
To attain the IBM Certified Specialist - AIX HACMP certification, candidates must first obtain the AIX System Administration or the AIX System Support certification. In order to obtain one of these prerequisite certifications, the candidate must pass one of the following two exams:

Test 181: AIX V4.3 System Administration

or

Test 189: AIX V4.3 System Support.

Following this, the candidate must pass the following exam:

Test 167: HACMP for AIX V 4.2.

Recommended Prerequisites
A minimum of six to twelve months implementation experience installing, configuring, and testing/supporting HACMP for AIX.

Registration for the Certification Exam
For information about how to register for the certification exam, please visit the following Web site:

`http://www.ibm.com/certify`

1.2 Certification Exam Objectives

The following objectives were used as a basis for what is required when the certification exam was developed. Some of these topics have been regrouped to provide better organization when discussed in this publication.

Section 1 - Preinstallation

The following items should be considered as part of the preinstallation plan:

- Conduct a Planning Session.
  - Set customer expectations at the beginning of the planning session.
  - Gather customer's availability requirements.
  - Articulate trade-offs of different HA configurations.
  - Assist customers in identifying HA applications.
- Evaluate the Customer Environment and Tailorable Components.
  - Evaluate the configuration and identify Single Points of Failure (SPOF).
  - Define and analyze NFS requirements.
  - Identify components affecting HACMP.
  - Identify HACMP event logic customizations.
- Plan for Installation.
  - Develop a disk management modification plan.
  - Understand issues regarding single adapter solutions.
  - Produce a Test Plan.

Section 2 - HACMP Implementation

The following items should be considered for proper implementation:

- Configure HACMP Solutions.
  - Install HACMP Code.
  - Configure an IP Address Takeover (IPAT).
  - Configure non-IP heartbeat paths.
  - Configure a network adapter.
  - Customize/tailor AIX.
  - Set up a shared disk (SSA).
  - Set up a shared disk (SCSI).
  - Verify a cluster configuration.

  - Create an application server.
- Set up Event Notification.
  - Set up event notification and pre/post event scripts.
  - Set up error notification.
- Post Configuration Activities.
  - Configure a client notification and ARP update.
  - Implement a test plan.
  - Create a snapshot.
  - Create a customization document.
- Perform Testing and Troubleshooting.
  - Troubleshoot a failed IPAT failover.
  - Troubleshoot failed shared volume groups.
  - Troubleshoot a failed network configuration.
  - Troubleshoot failed shared disk tests.
  - Troubleshoot a failed application.
  - Troubleshoot failed Pre/Post event scripts.
  - Troubleshoot failed error notifications.
  - Troubleshoot errors reported by cluster verification.

Section 3 - System Management

The following items should be considered for System Management:

- Communicate with the Customer.
  - Conduct a turnover session.
  - Provide hands-on customer education.
  - Set customer expectations of their HACMP solution's capabilities.
- Perform Systems Maintenance.
  - Perform HACMP maintenance tasks (PTFs, adding products, replacing disks, adapters).
  - Perform AIX maintenance tasks.
  - Dynamically update the cluster configuration.
  - Perform testing and troubleshooting as a result of changes.

1.3 Certification Education Courses

Courses and publications are offered to help you prepare for the certification tests. These courses are recommended, but not required, before taking a certification test. At the printing of this guide, the following courses are available. For a current list, please visit the following Web site:

```
http://www.ibm.com/certify
```

Table 1. AIX Version 4 HACMP Installation and Implementation

Course Number	Q1054 (USA) AU54 (Worldwide)
Course Duration	Five days
Course Abstract	This course provides a detailed understanding of the High Availability Clustered Multi-Processing for AIX. The course is supplemented with a series of laboratory exercises to configure the hardware and software environments for HACMP. Additionally, the labs provide the opportunity to: • Install the product. • Define networks. • Create file systems. • Complete several modes of HACMP installations.

The following table outlines information about the next course.

Table 2. AIX Version 4 HACMP System Administration

Course Number	Q1150 (USA); AU50 (Worldwide)
Course Duration	Five days
Course Abstract	This course teaches the student the skills required to administer an HACMP cluster on an ongoing basis after it is installed. The skills that are developed in this course include: • Integrating the cluster with existing network services (DNS, NIS, etc.) • Monitoring tools for the cluster, including HAView for Netview • Maintaining user IDs and passwords across the cluster • Recovering from script failures • Making configuration or resource changes in the cluster • Repairing failed hardware • Maintaining required cluster documentation The course involves a significant number of hands-on exercises to reinforce the concepts. Students are expected to have completed the course AU54 (Q1054) HACMP Installation and Implementation before attending this course.

Chapter 2. Cluster Planning

The area of cluster planning is a large one. Not only does it include planning for the types of hardware (CPUs, networks, disks) to be used in the cluster, but it also includes other aspects. These include resource planning, that is, planning the desired behavior of the cluster in failure situations. Resource planning must take into account application loads and characteristics, as well as priorities. This chapter will cover all of these areas, as well as planning for event customizations and user id planning issues.

2.1 Cluster Nodes

One of HACMP's key design strengths is its ability to provide support across the entire range of RISC System/6000 products. Because of this built-in flexibility and the facility to mix and match RISC System/6000 products, the effort required to design a highly available cluster is significantly reduced.

In this chapter, we shall outline the various hardware options supported by HACMP for AIX and HACMP/ES. We realize that the rapid pace of change in products will almost certainly render any snapshot of the options out of date by the time it is published. This is true of almost all technical writing, though to yield to the spoils of obsolescence would probably mean nothing would ever make it to the printing press.

The following sections will deal with the various:

- CPU Options
- Cluster Node Considerations

available to you when you are planning your HACMP cluster.

2.1.1 CPU Options

HACMP is designed to execute with RISC System/6000 uniprocessors, Symmetric Multi-Processor (SMP) servers and the RS/6000 Scalable POWERparallel Systems (RS/6000 SP) in a *no single point of failure* server configuration. The minimum configuration and sizing of each system CPU is highly dependent on the user's application and data requirements. Nevertheless, systems with 32 MB of main storage and 1 GB of disk storage would be practical, minimum configurations.

Almost any model of the RISC System/6000 POWERserver family can be included in an HACMP environment and new models continue to be added to the list. The following table gives you an overview of the currently supported

RISC System/6000 models as nodes in an HACMP 4.1 for AIX, HACMP 4.2 for AIX, or HACMP 4.3 for AIX cluster.

Table 3. Hardware Requirements for the Different HACMP Versions

HACMP Version	4.1	4.2	4.3	4.2/ES	4.3/ES
7009 Mod. CXX	yes	yes	yes	no	yes[1]
7011 Mod. 2XX	yes	yes	yes	no	yes[1]
7012 Mod. 3XX and GXX	yes	yes	yes	no	yes[1]
7013 Mod. 5XX and JXX	yes	yes	yes	no	yes[1]
7015 Mod. 9XX and RXX	yes	yes	yes	no	yes[1]
7017 Mod. S7X	yes	yes	yes	no	yes[1]
7024 Mod. EXX	yes	yes	yes	no	yes[1]
7025 Mod. FXX	yes	yes	yes	no	yes[1]
7026 Mod. HXX	yes	yes	yes	no	yes[1]
7043 Mod. 43P, 260	yes	yes	yes	no	yes[1]
9076 RS/6000 SP	yes	yes	yes	yes	yes[1]

[1] AIX 4.3.2 required

For a detailed description of system models supported by HACMP/6000 and HACMP/ES, you should refer to the current Announcement Letters for HACMP/6000 and HACMP/ES.

HACMP/ES 4.3 further enhances cluster design flexibility even further by including support for the RISC System/6000 family of machines and the Compact Server C20. Since the introduction of HACMP 4.1 for AIX, you have been able to mix uniprocessor and multiprocessor machines in a single cluster. Even a mixture of “normal” RS/6000 machines and RS/6000 SP nodes is possible.

2.1.2 Cluster Node Considerations

It is important to understand that selecting the system components for a cluster requires careful consideration of factors and information that may not be considered in the selection of equipment for a single-system environment. In this section, we will offer some guidelines to assist you in choosing and sizing appropriate machine models to build your clusters.

Much of the decision centers around the following areas:

- Processor capacity
- Application requirements
- Anticipated growth requirements
- I/O slot requirements

These paradigms are certainly not new ones, and are also important considerations when choosing a processor for a single-system environment. However, when designing a cluster, you must carefully consider the requirements of the cluster as a total entity. This includes understanding system capacity requirements of other nodes in the cluster beyond the requirements of each system's prescribed normal load. You must consider the required performance of the solution during and after failover, when a surviving node has to add the workload of a failed node to its own workload.

For example, in a two node cluster, where applications running on both nodes are critical to the business, each of the two nodes functions as a backup for the other, in a mutual takeover configuration. If a node is required to provide failover support for all the applications on the other node, then its capacity calculation needs to take that into account. Essentially, the choice of a model depends on the requirements of highly available applications, not only in terms of CPU cycles, but also of memory and possibly disk space. Approximately 50 MB of disk storage is required for full installation of the HACMP software.

A major consideration in the selection of models will be the number of I/O expansion slots they provide. The model selected must have enough slots to house the components required to remove single points of failure (SPOFs) and provide the desired level of availability. A single point of failure is defined as any single component in a cluster whose failure would cause a service to become unavailable to end users. The more single points of failure you can eliminate, the higher your level of availability will be. Typically, you need to consider the number of slots required to support network adapters and disk I/O adapters. Your slot configuration must provide for at least two network adapters to provide adapter redundancy for one service network. If your system needs to be able to take over an IP address for more than one other system in the cluster at a time, you will want to configure more standby network adapters. A node can have up to seven standby adapters for each network it connects to. Again, if that is your requirement, you will need to select models as nodes where the number of slots will accomodate the requirement.

Your slot configuration must also allow for the disk I/O adapters you need to support the cluster's shared disk (volume group) configuration. If you intend to use disk mirroring for shared volume groups, which is strongly recommended, then you will need to use slots for additional disk I/O adapters, providing I/O adapter redundancy across separate buses.

The following table tells you the number of additional adapters you can put into the different RS/6000 models. Ethernet environments can sometimes make use of the integrated ethernet port provided by some models. No such feature is available for token-ring, FDDI or ATM; you must use an I/O slot to provide token-ring adapter redundancy.

Table 4. Number of Adapter Slots in Each Model

RS/6000 Model	Number of Slots	Integrated Ethernet Port
7006	4 x MCA	yes
7009 C10, C20	4x PCI	no
7012 Mod. 3XX and GXX	4 x MCA	yes
7013 Mod. 5XX	7 x MCA	no
7013 Mod. JXX	6 x MCA, 14 x MCA with expansion unit J01	no
7015 Mod. R10, R20, R21	8 x MCA	no
7015 Mod. R30, R40, R50	16 x MCA	no
7017 Mod. S7X	52 x PCI	no
7024 EXX	5 x PCI, 1 x PCI/ISA 2 x ISA	no
7025 F50	6 x PCI, 2 x ISA/PCI	yes
7026 Mod. H50	6 x PCI, 2 x ISA/PCI	yes
7043Mod.	3 x PCI, 2 x ISA/PCI	yes
9076 thin node	4 x MCA	yes
9076 wide node	7 x MCA	no
9076 high node	15 x MCA	no
9076 thin node (silver)	2 x PCI	yes[1]
9076 wide node (silver)	10 x PCI	yes[1]

[1] The switch adapter is onboard and does not need an extra slot.

2.2 Cluster Networks

HACMP differentiates between two major types of networks: TCP/IP networks and non-TCP/IP networks. HACMP utilizes both of them for exchanging heartbeats. HACMP uses these heartbeats to diagnose failures in the cluster. Non-TCP/IP networks are used to distinguish an actual hardware failure from the failure of the TCP/IP software. If there were only TCP/IP networks being used, and the TCP/IP software failed, causing heartbeats to stop, HACMP could falsely diagnose a node failure when the node was really still functioning. Since a non-TCP/IP network would continue working in this event, the correct diagnosis could be made by HACMP. In general, all networks are also used for verification, synchronization, communication and triggering events between nodes. Of course, TCP/IP networks are used for communication with client machines as well.

At the time of publication, the HACMP/ES Version 4.3 product does not use non-TCP/IP networks for node-to-node communications in triggering, synchronizing, and executing event reactions. This can be an issue if you are configuring a cluster with only one TCP/IP network. This limitation of HACMP/ES is planned to be removed in a future release. You would be advised to check on the status of this issue if you are planning a new installation, and to plan your cluster networks accordingly.

2.2.1 TCP/IP Networks

The following sections describe supported TCP/IP network types and network considerations.

2.2.1.1 Supported TCP/IP Network Types

Basically every adapter that is capable of running the TCP/IP protocol is a supported HACMP network type. There are some special considerations for certain types of adapters however. The following gives a brief overview on the supported adapters and their special considerations.

Below is a list of TCP/IP network types as you will find them at the configuration time of an adapter for HACMP. You will find the non-TCP/IP network types in 2.2.2.1, “Supported Non-TCP/IP Network Types” on page 14.

- Generic IP
- ATM
- Ethernet
- FCS

- FDDI
- SP Switch
- SLIP
- SOCC
- Token-Ring

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

Note

ATM and SP Switch networks are special cases of point-to-point, private networks that can connect clients

The HACMP for AIX software supports a maximum of 32 networks per cluster and 24 TCP/IP network adapters on each node. These numbers provide a great deal of flexibility in designing a network configuration. The network design affects the degree of system availability in that the more communication paths that connect clustered nodes and clients, the greater the degree of network availability.

2.2.1.2 Special Network Considerations

Each type of interface has different characteristics concerning speed, MAC addresses, ARP, and so on. In case there is a limitation you will have to work around, you need to be aware of the characteristics of the adapters you plan to use. In the next paragraphs, we summarize some of the considerations that are known.

Hardware Address Swapping is one issue. If you enable HACMP to put one address on another adapter, it would need something like a boot and a service address for IPAT, but on the hardware layer. So, in addition to the manufacturers burnt-in address, there has to be an alternate address configured.

The speed of the network can be another issue. Your application may have special network throughput requirements that must be taken into account.

Network types also differentiate themselves in the maximum distance they allow between adapters, and in the maximum number of adapters allowed on a physical network.

- **Ethernet** supports 10 and 100 Mbps currently, and supports hardware address swapping. Alternate hardware addresses should be in the form xxxxxxxxxxyy, where xxxxxxxxxx is replaced with the first five pairs of digits of the original burned-in MAC address and yy can be chosen freely. There is a limit of 29 adapters on one physical network, unless a network repeater is used.
- **Token-Ring** supports 4 or 16 Mbps, but 4 Mbps is very rarely used now. It also supports hardware address swapping, but here the convention is to use 42 as the first two characters of the alternate address, since this indicates that it is a locally set address.
- **FDDI** is a 100 Mbps optical LAN interface, that supports hardware address takeover as well. For FDDI adapters you should leave the last six digits of the burned-in address as they are, and use a 4, 5, 6, or 7 as the first digit of the rest. FDDI can connect as many as 500 stations with a maximum link-to-link distance of two kilometers and a total LAN circumference of 100 kilometers.
- **ATM** is a point-to-point connection network. It currently supports the OC3 and the OC12 standard, which is 155 Mbps or 625 Mbps. You cannot use hardware address swapping with ATM. ATM doesn’t support broadcasts, so it must be configured as a private network to HACMP. However, if you are using LAN Emulation on an existing ATM network, you can use the emulated ethernet or Token-Ring interfaces just as if they were real ones, except that you cannot use hardware address swapping.
- **FCS** is a fiber channel network, currently available as two adapters for either MCA or PCI technology. The Fibre Channel Adapter /1063-MCA, runs up to 1063 Mb/second, and the Gigabit Fibre Channel Adapter for PCI Bus (#6227), announced on October 5th 1998, will run with 100 MBps. Both of them support TCP/IP, but not hardware address swapping.
- **SLIP** runs at up to 38400 bps. Since it is a point-to-point connection and very slow, it is rarely used as an HACMP network. An HACMP cluster is much more likely to use the serial port as a non-TCP/IP connection. See below for details.
- **SOCC** is a fast optical connection, again point-to-point. This is an optical line with a serial protocol running on it. However, the SOCC Adapter (Feature 2860) has been withdrawn from marketing for some years now. Some models, like 7013 5xx, offer SOCC as an option onboard, but these are rarely used today.

- **SP Switch** is a high-speed packet switching network, running on the RS/6000 SP system only. It runs bidirectionally up to 80 MBps, which adds up to 160 MBps of capacity per adapter. This is node-to-node communication and can be done in parallel between every pair of nodes inside an SP. The SP Switch network has to be defined as a private Network, and ARP must be enabled. This network is restricted to one adapter per node, thus, it has to be considered as a Single Point Of Failure. Therefore, it is strongly recommended to use AIX Error Notification to propagate a switch failure into a node failure when appropriate. As there is only one adapter per node, HACMP uses the ifconfig alias addresses for IPAT on the switch; so, a standby address is not necessary and, therefore, not used on the switch network. Hardware address swapping also is not supported on the SP Switch.

For IP Address Takeover (IPAT), in general, there are two adapters per cluster node and network recommended in order to eliminate single points of failure. The only exception to this rule is the SP Switch because of hardware limitations.

2.2.2 Non-TCPIP Networks

Non-TCP/IP networks in HACMP are used as an independent path for exchanging messages or heartbeats between cluster nodes. In case of an IP subsystem failure, HACMP can still differentiate between a network failure and a node failure when an independent path is available and functional. Below is a short description of the three currently available non-TCP/IP network types and their characteristics. Even though HACMP works without one, it is strongly recommended that you use at least one non-TCP/IP connection between the cluster nodes.

2.2.2.1 Supported Non-TCP/IP Network Types

Currently HACMP supports the following types of networks for non-TCP/IP heartbeat exchange between cluster nodes:

- Serial (RS232)
- Target-mode SCSI
- Target-mode SSA

All of them must be configured as **Network Type: serial** in the HACMP definitions.

2.2.2.2 Special Considerations

As for TCP/IP networks, there are a number of restrictions on non-TCP/IP networks. These are explained for the three different types in more detail below.

Serial (RS232)

A serial (RS232) network needs at least one available serial port per cluster node. In case of a cluster consisting of more than two nodes, a ring of nodes is established through serial connections, which requires two serial ports per node. Table 3 shows a list of possible cluster nodes and the number of native serial ports for each:

Table 5. Number of Available Serial Ports in Each Model.

RS/6000 Model	Number of Serial Ports Available
7006	1[1]
7009 C10, C20	1[1]
7012 Mod. 3XX and GXX	2
7013 Mod. 5XX	2
7013 Mod. JXX	3
7015 Mod. R10, R20, R21	3
7015 Mod. R30, R40, R50	3
7013,7015,7017 Mod. S7X	0 [2]
7024 EXX	1[1]
7025 F50	2
7026 Mod. H50	3
7043Mod.	2
9076 thin node	2[3]
9076 wide node	2[3]
9076 high node	3[3]
9076 thin node (silver)	2[3]
9076 wide node (silver)	2[3]

[1] serial port can be multiplexed through a dual-port cable, thus offering two ports

[2] a PCI Multiport Async Card is required in an S7X model, no native ports
[3] only one serial port available for customer use, i.e. HACMP

In case the number of native serial ports doesn’t match your HACMP cluster configuration needs, you can extend it by adding an eight-port asynchronous adapter, thus reducing the number of available MCA slots, or the corresponding PCI Multiport Async Card for PCI Machines, like the S7X model.

Target-mode SCSI

Another possibility for a non-TCP/IP network is a target mode SCSI connection. Whenever you make use of a shared SCSI device, you can also use the SCSI bus for exchanging heartbeats.

Target Mode SCSI is only supported with SCSI-2 Differential or SCSI-2 Differential Fast/Wide devices. SCSI/SE or SCSI-2/SE are not supported for HACMP serial networks.

The recommendation is to not use more than 4 target mode SCSI networks in a cluster.

Target-mode SSA

If you are using shared SSA devices, target mode SSA is the third possibility for a serial network within HACMP. In order to use target-mode SSA, you must use the Enhanced RAID-5 Adapter (#6215 or #6219), since these are the only current adapters that support the Multi-Initiator Feature. The microcode level of the adapter must be 1801 or higher.

2.3 Cluster Disks

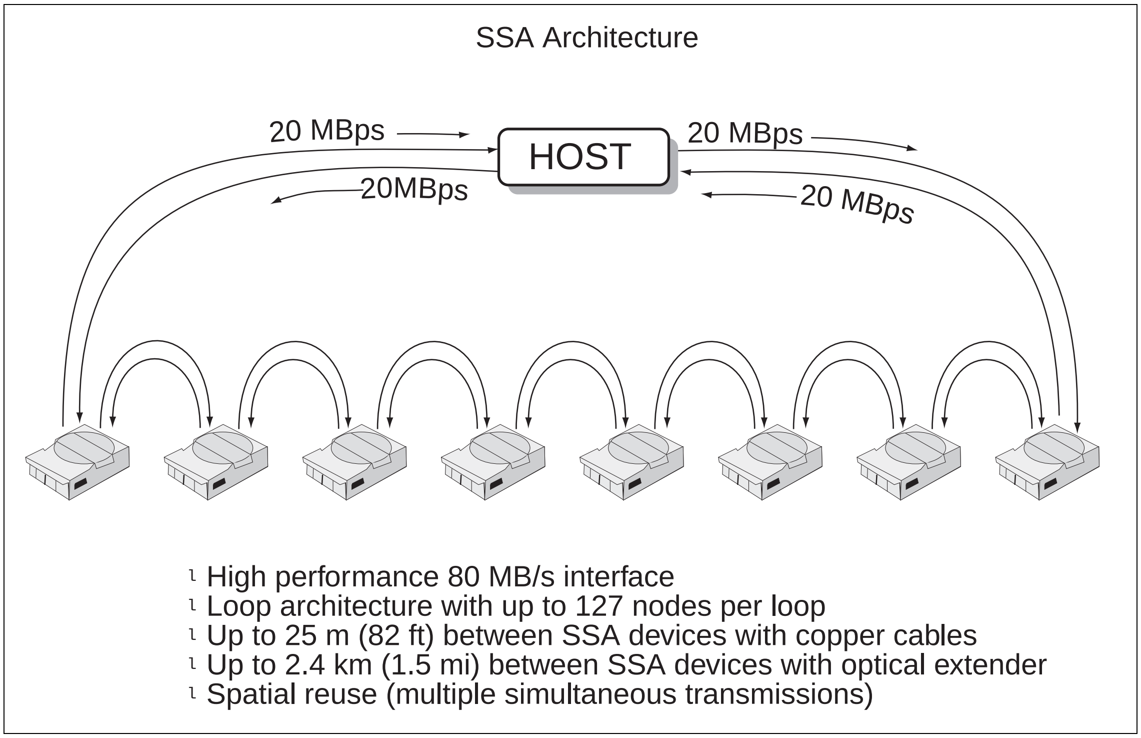
This section describes the various choices you have in selecting the type of shared disks to use in your cluster.

2.3.1 SSA Disks

The following is a brief description of SSA and the basic rules to follow when designing SSA networks. For a full description of SSA and its functionality, please read *Monitoring and Managing IBM SSA Disk Subsystems,* SG24-5251.

SSA is a high-performance, serial interconnect technology used to connect disk devices and host adapters. SSA is an open standard, and SSA specifications have been approved by the SSA Industry Association and also as an ANSI standard through the ANSI X3T10.1 subcommittee.

SSA subsystems are built up from loops of adapters and disks. A simple example is shown in Figure 1.



- High performance 80 MB/s interface
- Loop architecture with up to 127 nodes per loop
- Up to 25 m (82 ft) between SSA devices with copper cables
- Up to 2.4 km (1.5 mi) between SSA devices with optical extender
- Spatial reuse (multiple simultaneous transmissions)

Figure 1. Basic SSA Configuration

Here, a single adapter controls one SSA loop of eight disks. Data can be transferred around the loop, in either direction, at 20 MBps. Consequently, the peak transfer rate of the adapter is 80 MBps. The adapter contains two SSA nodes and can support two SSA loops. Each disk drive also contains a single SSA node. A node can be either an initiator or a target. An *initiator* issues commands, while a *target* responds with data and status information. The SSA nodes in the adapter are therefore initiators, while the SSA nodes in the disk drives are targets.

There are two types of SSA Disk Subsystems for RISC System/6000 available:

- 7131 SSA Multi-Storage Tower Model 405

- 7133 Serial Storage Architecture (SSA) Disk Subsystem Models 010, 500, 020, 600, D40 and T40.

The 7133 models 010 and 500 were the first SSA products announced in 1995 with the revolutionary new Serial Storage Architecture. Some IBM customers still use the Models 010 and 500, but these have been replaced by 7133 Model 020, and 7133 Model 600 respectively. More recently, in November 1998, the models D40 and T40 were announced.

All 7133 Models have redundant power and cooling, which is hot-swappable.

The following tables give you more configuration information about the different models:

Table 6. 7131-Model 405 SSA Multi-Storage Tower Specifications

Item	Specification
Transfer rate SSA interface	80 MB
Configuration	2 to 5 disk drives (2.2 GB, 4.5 GB or 9.1 GB) per subsystem
Configuration range	4.4 to 11 GB (with 2.2 GB disk drives) 9.0 to 22.5 GB (With 4.5 GB disk drives) 18.2 to 45.5 GB (With 9.1 GB disk drives)
Supported RAID levels	5
Supported adapters	6214, 6216, 6217, 6218
Hot-swap disks	Yes

Table 7. 7133 Models 010, 020, 500, 600, D40, T40 Specifications

Item	Specification
Transfer rate SSA interface	80 MB/s
Configuration	4 to 16 disks - 1.1 GB, 2.2 GB, 4.5 GB, for Models 10, 20, 500, and 600 - 9.1 GB for Models 20, 600, D40 and T40 - With 1.1 GB disk drives you must have 8 to 16 disks)
Configuration range	8.8 to 17.6 GB (with 1.1 GB disks) 8.8 to 35.2 GB (with 2.2 GB disks) 18 to 72 GB (with 4.5 GB disks) 36.4 to 145.6 GB (with 9.1 GB disks) 72.8 to 291.2 GB (with 18.2 GB disks)

Item	Specification
Supported RAID level	5
Supported adapters	all
Hot-swappable disk	Yes (and hot-swappable, redundant power and cooling)

2.3.1.1 Disk Capacities

Table 8 lists the different SSA disks, and provides an overview of their characteristics.

Table 8. SSA Disks

Name	Capacities (GB)	Buffer size (KB)	Maximum Transfer rate (MBps)
Starfire 1100	1.1	0	20
Starfire 2200	2.2	0	20
Starfire 4320	4.5	512	20
Scorpion 4500	4.5	512	80
Scorpion 9100	9.1	512	160
Sailfin 9100	9.1	1024	160
Thresher 9100	9.1	1024	160
Ultrastar	9.1, 18.2	4096	160

2.3.1.2 Supported and Non-Supported Adapters

Table 9 lists the different SSA adapters and presents an overview of their characteristics.

Table 9. SSA Adapters

Feature Code	Adapter Label	Bus	Adapter Description	Number of Adapters per Loop	Hardware Raid Types
6214	4-D	MCA	Classic	2	n/a
6215	4-N	PCI	Enhanced RAID-5	8[1]	5
6216	4-G	MCA	Enhanced	8	n/a
6217	4-I	MCA	RAID-5	1	5
6218	4-J	PCI	RAID-5	1	5

Feature Code	Adapter Label	Bus	Adapter Description	Number of Adapters per Loop	Hardware Raid Types
6219	4-M	MCA	Enhanced RAID-5	8[1]	5

[1] See 2.3.1.3, “Rules for SSA Loops” on page 20 for more information.

The following rules apply to SSA Adapters:

- You cannot have more than four adapters in a single system.
- The MCA SSA 4-Port RAID Adapter (FC 6217) and PCI SSA 4-Port RAID Adapter (FC 6218) are not useful for HACMP, because only one can be in a loop.
- Only the PCI Multi Initiator/RAID Adapter (FC 6215) and the MCA Multi Initiator/RAID EL Adapter (FC 6219) support target mode SSA (for more information about target mode SSA see 3.2.2, “Non TCP/IP Networks” on page 63).

2.3.1.3 Rules for SSA Loops

The following rules must be followed when configuring and connecting SSA loops:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2).
- Only one of the two pairs of connectors on an adapter card can be connected in a single SSA loop.
- A maximum of 48 devices can be connected in a single SSA loop.
- A maximum of two adapters can be connected in a particular loop if one adapter is an SSA 4-Port adapter, Feature 6214.
- A maximum of eight adapters can be connected in a particular loop if all the adapters are Enhanced SSA 4-Port Adapters, Feature 6216.
- A maximum of two SSA adapters, both connected in the same SSA loop, can be installed in the same system.

For SSA loops that include an SSA Four-Port RAID adapter (Feature 6217) or a PCI SSA Four-Port RAID adapter (Feature 6218), the following rules apply:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2).

- A maximum of 48 devices can be connected in a particular SSA loop.
- Only one pair of adapter connectors can be connected in a particular SSA loop.
- Member disk drives of an array can be on either SSA loop.

For SSA loops that include a Micro Channel Enhanced SSA Multi-initiator/RAID EL adapter, Feature 6215 or a PCI SSA Multi-initiator/RAID EL adapter, Feature 6219, the following rules apply:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter (that is, either Connectors A1 and A2, or Connectors B1 and B2).
- A maximum of eight adapters can be connected in a particular loop if none of the disk drives in the loops are array disk drives and none of them is configured for fast-write operations. The adapters can be up to eight Micro Channel Enhanced SSA Multi-initiator/RAID EL Adapters, up to eight PCI Multi-initiator/RAID EL Adapters, or a mixture of the two types.
- A maximum of two adapters can be connected in a particular loop if one or more of the disk drives in the loop are array disk drives that are not configured for fast-write operations. The adapters can be two Micro Channel Enhanced SSA Multi-initiator/RAID EL Adapters, two PCI Multi-initiator/RAID EL Adapters, or one adapter of each type.
- Only one Micro Channel Enhanced SSA Multi-initiator/RAID EL Adapter or PCI SSA Multi-initiator/RAID EL Adapter can be connected in a particular loop if any disk drives in the loops are members of a RAID-5 array, and are configured for fast-write operations.
- All member disk drives of an array must be on the same SSA loop.
- A maximum of 48 devices can be connected in a particular SSA loop.
- Only one pair of adapter connectors can be connected in a particular loop.
- When an SSA adapter is connected to two SSA loops, and each loop is connected to a second adapter, both adapters must be connected to both loops.

For the IBM 7190-100 SCSI to SSA converter, the following rules apply:

- There can be up to 48 disk drives per loop.
- There can be up to four IBM 7190-100 attached to any one SSA loop.

2.3.1.4 RAID vs. Non-RAID

RAID Technology

RAID is an acronym for Redundant Array of Independent Disks. Disk arrays are groups of disk drives that work together to achieve higher data-transfer and I/O rates than those provided by single large drives.

Arrays can also provide data redundancy so that no data is lost if a single drive (physical disk) in the array should fail. Depending on the RAID level, data is either mirrored or striped. The following gives you more information about the different RAID levels.

RAID Level 0

RAID 0 is also known as data striping. Conventionally, a file is written out sequentially to a single disk. With striping, the information is split into chunks (fixed amounts of data usually called blocks) and the chunks are written to (or read from) a series of disks in parallel. There are two main performance advantages to this:

1. Data transfer rates are higher for sequential operations due to the overlapping of multiple I/O streams.
2. Random access throughput is higher because access pattern skew is eliminated due to the distribution of the data. This means that with data distributed evenly across a number of disks, random accesses will most likely find the required information spread across multiple disks and thus benefit from the increased throughput of more than one drive.

RAID 0 is only designed to increase performance. There is no redundancy; so any disk failures will require reloading from backups.

RAID Level 1

RAID 1 is also known as disk mirroring. In this implementation, identical copies of each chunk of data are kept on separate disks, or more commonly, each disk has a twin that contains an exact replica (or mirror image) of the information. If any disk in the array fails, then the mirrored twin can take over.

Read performance can be enhanced because the disk with its actuator closest to the required data is always used, thereby minimizing seek times. The response time for writes can be somewhat slower than for a single disk, depending on the write policy; the writes can either be executed in parallel for speed or serially for safety.

RAID Level 1 has data redundancy, but data should be regularly backed up on the array. This is the only way to recover data in the event that a file or directory is accidentally deleted.

RAID Levels 2 and 3

RAID 2 and RAID 3 are parallel process array mechanisms, where all drives in the array operate in unison. Similar to data striping, information to be written to disk is split into chunks (a fixed amount of data), and each chunk is written out to the same physical position on separate disks (in parallel). When a read occurs, simultaneous requests for the data can be sent to each disk.

This architecture requires parity information to be written for each stripe of data; the difference between RAID 2 and RAID 3 is that RAID 2 can utilize multiple disk drives for parity, while RAID 3 can use only one. If a drive should fail, the system can reconstruct the missing data from the parity and remaining drives.

Performance is very good for large amounts of data but poor for small requests since every drive is always involved, and there can be no overlapped or independent operation.

RAID Level 4

RAID 4 addresses some of the disadvantages of RAID 3 by using larger chunks of data and striping the data across all of the drives except the one reserved for parity. Using disk striping means that I/O requests need only reference the drive that the required data is actually on. This means that simultaneous, as well as independent reads, are possible. Write requests, however, require a read/modify/update cycle that creates a bottleneck at the single parity drive. Each stripe must be read, the new data inserted and the new parity then calculated before writing the stripe back to the disk. The parity disk is then updated with the new parity, but cannot be used for other writes until this has completed. This bottleneck means that RAID 4 is not used as often as RAID 5, which implements the same process but without the bottleneck. RAID 5 is discussed in the next section.

RAID Level 5

RAID 5, as has been mentioned, is very similar to RAID 4. The difference is that the parity information is distributed across the same disks used for the data, thereby eliminating the bottleneck. Parity data is never stored on the same drive as the chunks that it protects. This means that concurrent read and write operations can now be performed, and there are performance increases due to the availability of an extra disk (the disk previously used for parity). There are other enhancements possible to further increase data transfer rates, such as caching simultaneous reads from the disks and transferring that information while reading the next blocks. This can generate data transfer rates that approach the adapter speed.

As with RAID 3, in the event of disk failure, the information can be rebuilt from the remaining drives. RAID level 5 array also uses parity information, though it is still important to make regular backups of the data in the array. RAID level 5 stripes data across all of the drives in the array, one segment at a time (a segment can contain multiple blocks). In an array with n drives, a stripe consists of data segments written to n-1 of the drives and a parity segment written to the nth drive. This mechanism also means that not all of the disk space is available for data. For example, in an array with five 2 GB disks, although the total storage is 10 GB, only 8 GB are available for data.

The advantages and disadvantages of the various RAID levels are summarized in the following table:

Table 10. The Advantages and Disadvantages of the Different RAID Levels

RAID Level	Availability Mechanism	Capacity	Performance	Cost
0	none	100%	high	medium
1	mirroring	50%	medium/high	high
3	parity	80%	medium	medium
5	parity	80%	medium	medium

RAID on the 7133 Disk Subsystem

The only RAID level supported by the 7133 SSA disk subsystem is RAID 5. RAID 0 and RAID 1 can be achieved with the striping and mirroring facility of the Logical Volume Manager (LVM).

RAID 0 does not provide data redundancy, so it is not recommended for use with HACMP, because the shared disks would be a single point of failure. The possible configurations to use with the 7133 SSA disk subsystem are RAID 1 (mirroring) or RAID 5. Consider the following points before you make your decision:

- Mirroring is more expensive than RAID, but it provides higher data redundancy. Even if more than one disk fails, you may still have access to all of your data. In a RAID, more than one broken disk means that the data are lost.
- The SSA loop can include a maximum of two SSA adapters if you use RAID. So, if you want to connect more than two nodes into the loop, mirroring is the way to go.
- A RAID array can consist of three to 16 disks.

- Array member drives and spares must be on same loop (cannot span A and B loops) on the adapter.
- You cannot boot (ipl) from a RAID.

2.3.1.5 Advantages

Because SSA allows SCSI-2 mapping, all functions associated with initiators, targets, and logical units are translatable. Therefore, SSA can use the same command descriptor blocks, status codes, command queuing, and all other aspects of current SCSI systems. The effect of this is to make the type of disk subsystem transparent to the application. No porting of applications is required to move from traditional SCSI I/O subsystems to high-performance SSA. SSA and SCSI I/O systems can coexist on the same host running the same applications.

The advantages of SSA are summarized as follows:

- Dual paths to devices.
- Simplified cabling - cheaper, smaller cables and connectors, no separate terminators.
- Faster interconnect technology.
- Not an arbitrated system.
- Full duplex, frame multiplexed serial links.
- 40 MBps total per port, resulting in 80 MBps total per node, and 160 MBps total per adapter.
- Concurrent access to disks.
- Hot-pluggable cables and disks.
- Very high capacity per adapter - up to 127 devices per loop, although most adapter implementations limit this. For example, current IBM SSA adapters provide 96 disks per Micro Channel or PCI slot.
- Distance between devices of up to 25 meters with copper cables, 10km with optical links.
- Auto-configuring - no manual address allocation.
- SSA is an open standard.
- SSA switches can be introduced to produce even greater fan-out and more complex topologies.

2.3.2 SCSI Disks

After the announcement of the 7133 SSA Disk Subsystems, the SCSI Disk subsystems became less common in HACMP clusters. However, the 7135 RAIDiant Array (Model 110 and 210) and other SCSI Subsystems are still in use at many customer sites. We will not describe other SCSI Subsystems such as 9334 External SCSI Disk Storage. See the appropriate documentation if you need information about these SCSI Subsystems.

The 7135 RAIDiant Array is offered with a range of features, with a maximum capacity of 135 GB (RAID 0) or 108 GB (RAID-5) in a single unit, and uses the 4.5 GB disk drive modules. The array enclosure can be integrated into a RISC System/6000 system rack, or into a deskside mini-rack. It can attach to multiple systems through a SCSI-2 Differential 8-bit or 16-bit bus.

2.3.2.1 Capacities

Disks

There are four disk sizes available for the 7135 RAIDiant Array Models 110 and 210:

- 1.3 GB
- 2.0 GB
- 2.2 GB (only supported by Dual Active Software)
- 4.5 GB (only supported by Dual Active Software)

Subsystems

The 7135-110/210 can contain 15 Disks (max. 67.5 GB) in the base configuration and 30 Disks (max. 135 GB) in an extended configuration.You can for example only use the full 135 GB storage space for data if you configure the 7135 with RAID level 0. When using RAID level 5, only 108 GB of the 135 GB are available for data storage.

2.3.2.2 How Many in a String?

HACMP supports a maximum of two 7135s on a shared SCSI bus. This is because of cable length restrictions.

2.3.2.3 Supported SCSI Adapters

The SCSI adapters that can be used to connect RAID subsystems on a shared SCSI bus in an HACMP cluster are:

- SCSI-2 Differential Controller (MCA, FC: 2420, Adapter Label: 4-2)
- SCSI-2 Differential Fast/Wide Adapter/A (MCA, FC: 2416, Adapter Label: 4-6)

- Enhanced SCSI-2 Differential Fast/Wide Adapter/A (MCA, FC: 2412, Adapter Label: 4-C); not usable with 7135-110
- SCSI-2 Fast/Wide Differential Adapter (PCI, FC: 6209, Adapter Label: 4-B)
- DE Ultra SCSI Adapter (PCI, FC: 6207, Adapter Label: 4-L); not usable with 7135-110

2.3.2.4 Advantages - Disadvantages

The 7135 RAIDiant Array incorporates the following high availability features:

- Support for RAID-1, RAID-3 (Model 110 only) and RAID-5

  You can run any combination of RAID levels in a single 7135 subsystem. Each LUN can run its own RAID level.

- Multiple Logical Unit (LUN) support

  The RAID controller takes up only one SCSI ID on the external bus. The internal disks are grouped into logical units (LUNs). The array will support up to six LUNs, each of which appears to AIX as a single hdisk device. Since each of these LUNs can be configured into separate volume groups, different parts of the subsystem can be logically attached to different systems at any one time.

- Redundant Power Supply

  Redundant power supplies provide alternative sources of power. If one supply fails, power is automatically supplied by the other.

- Redundant Cooling

  Extra cooling fans are built into the RAIDiant Array to safeguard against fan failure.

- Concurrent Maintenance

  Power supplies, cooling fans, and failed disk drives can be replaced without the need to take the array offline or to power it down.

- Optional Second Array Controller

  This allows the array subsystem to be configured with no single point of failure. Under the control of the system software, the machine can be configured in *Dual Active* mode, so that each controller controls the operation of specific sets of drives. In the event of failure of either controller, all I/O activity is switched to the remaining active controller.

In the last few years, the 7133 SSA Subsystems have become more popular than 7135 RAIDiant Systems due to better technology. IBM decided to withdraw the 7135 RAIDiant Systems from marketing because it is equally possible to configure RAID on the SSA Subsystems.

2.4 Resource Planning

HACMP provides a highly available environment by identifying a set of cluster-wide resources essential to uninterrupted processing, and then defining relationships among nodes that ensure these resources are available to client processes.

When a cluster node fails or detaches from the cluster for a scheduled outage, the Cluster Manager redistributes its resources among any number of the surviving nodes.

HACMP considers the following as resource types:

- Volume Groups
- Disks
- File Systems
- File Systems to be NFS mounted
- File Systems to be NFS exported
- Service IP addresses
- Applications

The following paragraphs will tell you what to consider when configuring resources to accomplish the following:

- IP Address Takeover
- Shared LVM Components
- NFS Exports

and the options you have when combining these resources to a resource group.

2.4.1 Resource Group Options

Each resource in a cluster is defined as part of a resource group. This allows you to combine related resources that need to be together to provide a particular service. A resource group also includes the list of nodes that can acquire those resources and serve them to clients.

A resource group is defined as one of three types:

- Cascading
- Rotating
- Concurrent

Each of these types describes a different set of relationships between nodes in the cluster, and a different set of behaviors upon nodes entering and leaving the cluster.

Cascading Resource Groups: All nodes in a cascading resource group are assigned priorities for that resource group. These nodes are said to be part of that group's resource chain. In a cascading resource group, the set of resources cascades up or down to the highest priority node active in the cluster. When a node that is serving the resources fails, the surviving node with the highest priority takes over the resources.

A parameter called *Inactive Takeover* decides which node takes the cascading resources when the nodes join the cluster for the first time. If this parameter is set to *true*, the first node in a group's resource chain to join the cluster acquires all the resources in the resource group. As successive nodes join the resource group, the resources cascade up to any node with a higher priority that joins the cluster. If this parameter is set to *false*, the first node in a group's resource chain to join the cluster acquires all the resources in the resource group only if it is the node with the highest priority for that group. If the first node to join does not acquire the resource group, the second node in the group's resource chain to join acquires the resource group, if it has a higher priority than the node already active. As successive nodes join, the resource group cascades to the active node with the highest priority for the group. The default is *false*.

Member nodes of a cascading resource chain always release a resource group to a reintegrating node with a higher priority.

Rotating Resource Groups: A rotating resource group is associated with a group of nodes, rather than a particular node. A node can be in possession of a maximum of one rotating resource group per network.

As participating nodes join the cluster for the first time, they acquire the first available rotating resource group per network until all the groups are acquired. The remaining nodes maintain a standby role.

When a node holding a rotating resource group leaves the cluster, either because of a failure or gracefully while specifying the takeover option, the node with the highest priority and available connectivity takes over. Upon reintegration, a node remains as a standby and does not take back any of the resources that it had initially served.

Concurrent Resource Groups: A concurrent resource group may be shared simultaneously by multiple nodes. The resources that can be part of a concurrent resource group are limited to volume groups with raw logical volumes, raw disks, and application servers.

When a node fails, there is no takeover involved for concurrent resources. Upon reintegration, a node again accesses the resources simultaneously with the other nodes.

The Cluster Manager makes the following assumptions about the acquisition of resource groups:

Cascading The active node with the highest priority controls the resource group.

Concurrent All active nodes have access to the resource group.

Rotating The node with the rotating resource group's associated service IP address controls the resource group.

2.4.2 Shared LVM Components

The first distinction that you need to make while designing a cluster is whether you need a non-concurrent or a concurrent shared disk access environment.

2.4.2.1 Non-Concurrent Disk Access Configurations

The possible non-concurrent disk access configurations are:

- Hot-Standby
- Rotating Standby
- Mutual Takeover
- Third-Party Takeover

Hot-Standby Configuration

Figure 2 illustrates a two node cluster in a hot-standby configuration.

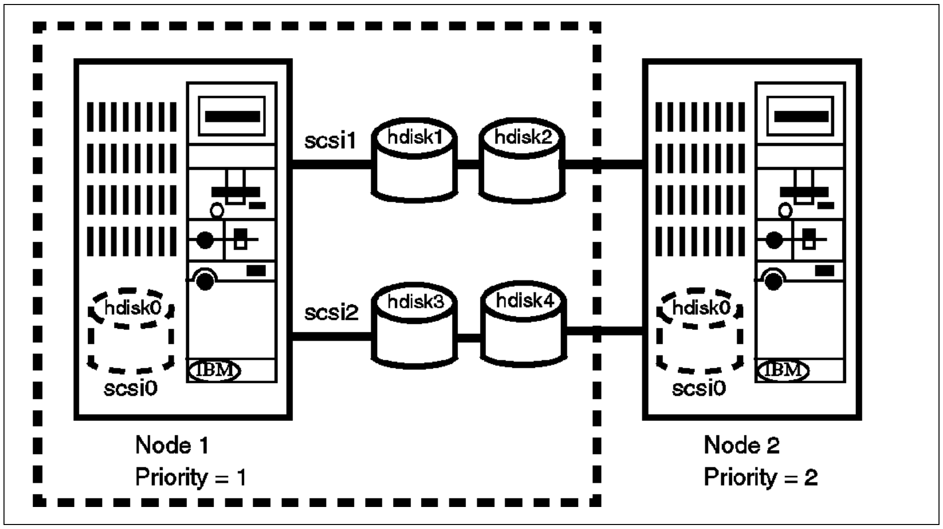


Figure 2. Hot-Standby Configuration

In this configuration, there is one cascading resource group consisting of the four disks, hdisk1 to hdisk4, and their constituent volume groups and file systems. Node 1 has a priority of 1 for this resource group while node 2 has a priority of 2. During normal operations, node 1 provides all critical services to end users. Node 2 may be idle or may be providing non-critical services, and hence is referred to as a hot-standby node. When node 1 fails or has to leave the cluster for a scheduled outage, node 2 acquires the resource group and starts providing the critical services.

The advantage of this type of a configuration is that you can shift from a single-system environment to an HACMP cluster at a low cost by adding a less powerful processor. Of course, this assumes that you are willing to accept a lower level of performance in a failover situation. This is a trade-off that you will have to make between availability, performance, and cost.

Rotating Standby Configuration

This configuration is the same as the previous configuration except that the resource groups used are rotating resource groups.

In the hot-standby configuration, when node 1 reintegrates into the cluster, it takes back the resource group since it has the highest priority for it. This implies a break in service to the end users during reintegration.

If the cluster is using rotating resource groups, reintegrating nodes do not reacquire any of the resource groups. A failed node that recovers and rejoins the cluster becomes a standby node. You must choose a rotating standby configuration if you do not want a break in service during reintegration.

Since takeover nodes continue providing services until they have to leave the cluster, you should configure your cluster with nodes of equal power. While more expensive in terms of CPU hardware, a rotating standby configuration gives you better availability and performance than a hot-standby configuration.

Mutual Takeover Configuration

Figure 3 illustrates a two node cluster in a mutual takeover configuration.

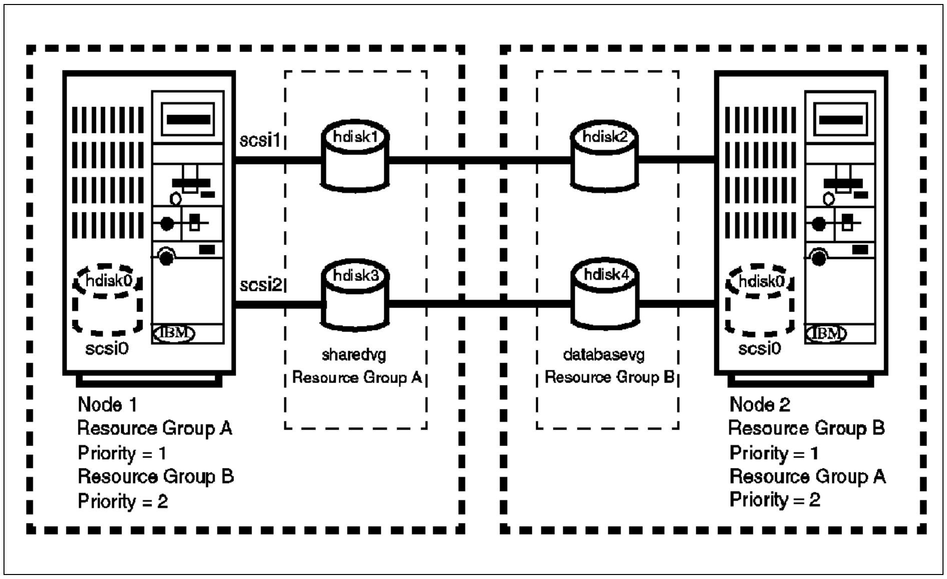


Figure 3. Mutual Takeover Configuration

In this configuration, there are two cascading resource groups: A and B. Resource group A consists of two disks, hdisk1 and hdisk3, and one volume group, sharedvg. Resource group B consists of two disks, hdisk2 and hdisk4, and one volume group, databasevg. Node 1 has priorities of 1 and 2 for resource groups A and B respectively, while Node 2 has priorities of 1 and 2 for resource groups B and A respectively.

During normal operations, nodes 1 and 2 have control of resource groups A and B respectively, and both provide critical services to end users. If either node 1 or node 2 fails, or has to leave the cluster for a scheduled outage, the surviving node acquires the failed node’s resource groups and continues to provide the failed node’s critical services.

When a failed node reintegrates into the cluster, it takes back the resource group for which it has the highest priority. Therefore, even in this configuration, there is a break in service during reintegration. Of course, if you look at it from the point of view of performance, this is the best thing to do, since you have one node doing the work of two when any one of the nodes is down.

Third-Party Takeover Configuration

Figure 4 illustrates a three node cluster in a third-party takeover configuration.

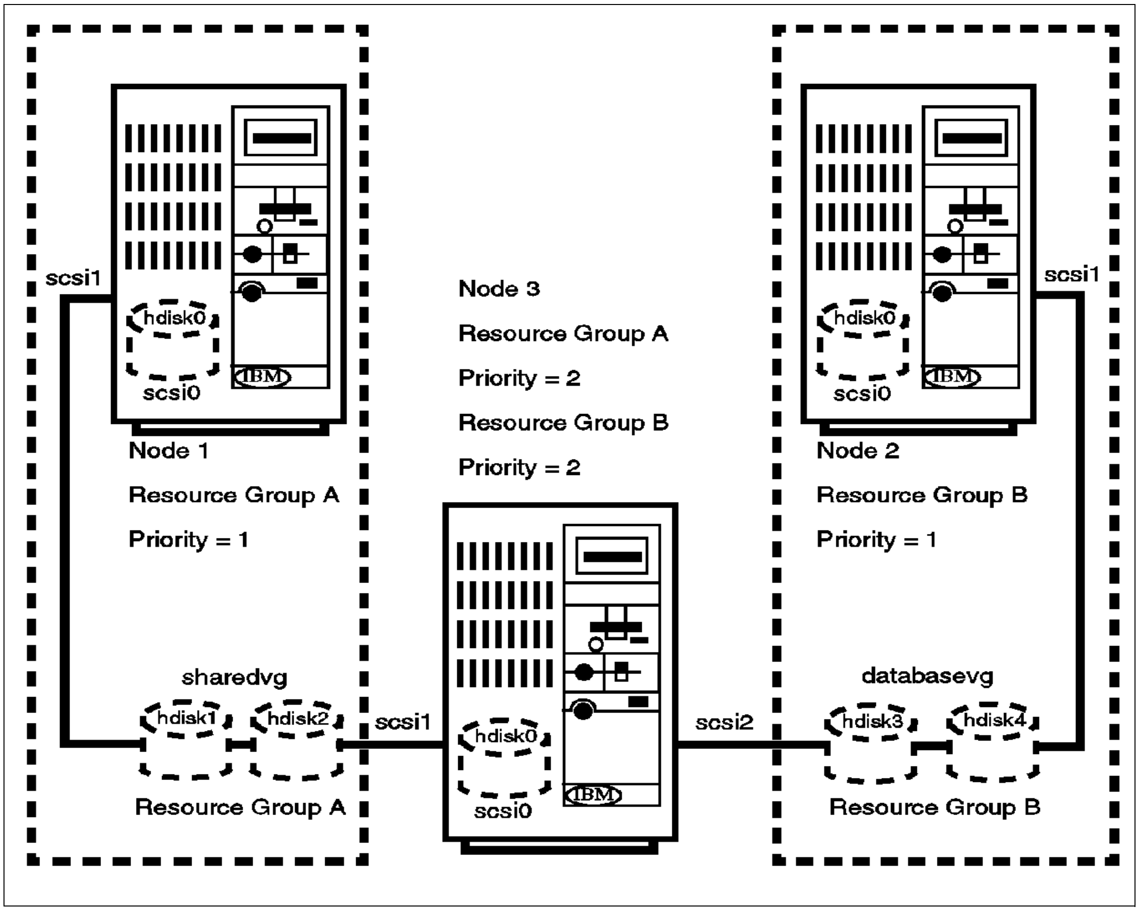


Figure 4. Third-Party Takeover Configuration

This configuration can avoid the performance degradation that results from a failover in the mutual takeover configuration.

Here the resource groups are the same as the ones in the mutual takeover configuration. Also, similar to the previous configuration, nodes 1 and 2 each have priorities of 1 for one of the resource groups, A or B. The only thing different in this configuration is that there is a third node which has a priority of 2 for both the resource groups.

During normal operations, node 3 is either idle or is providing non-critical services. In the case of either node 1 or node 2 failing, node 3 takes over the failed node's resource groups and starts providing its services. When a failed node rejoins the cluster, it reacquires the resource group for which it has the highest priority.

So, in this configuration, you are protected against the failure of two nodes and there is no performance degradation after the failure of one node.

2.4.2.2 Concurrent Disk Access Configurations

A concurrent disk access configuration usually has all its disk storage defined as part of one concurrent resource group. The nodes associated with a concurrent resource group have no priorities assigned to them.

If a 7135 RAIDiant Array Subsystem is used for storage, you can have a maximum of four nodes concurrently accessing a set of storage resources. If you are using the 7133 SSA Disk Subsystem, you can have up to eight nodes concurrently accessing it.This is because of the physical characteristics of SCSI versus SSA.

In the case of a node failure, a concurrent resource group is not explicitly taken over by any other node, since it is already active on the other nodes. However, in order to somewhat mask a node failure from the end users, you should also have cascading resource groups, each containing the service IP address for each node in the cluster. When a node fails, its service IP address will be taken over by another node and users can continue to access critical services at the same IP address that they were using before the node failed.

2.4.3 IP Address Takeover

The goal of IP Address Takeover is to make the server's service address highly available and to give clients the possibility of always connecting to the same IP address. In order to achieve this, you must do the following:

- Decide which types of networks and point-to-point connections to use in the cluster (see 2.2, “Cluster Networks” on page 11 for supported network types)

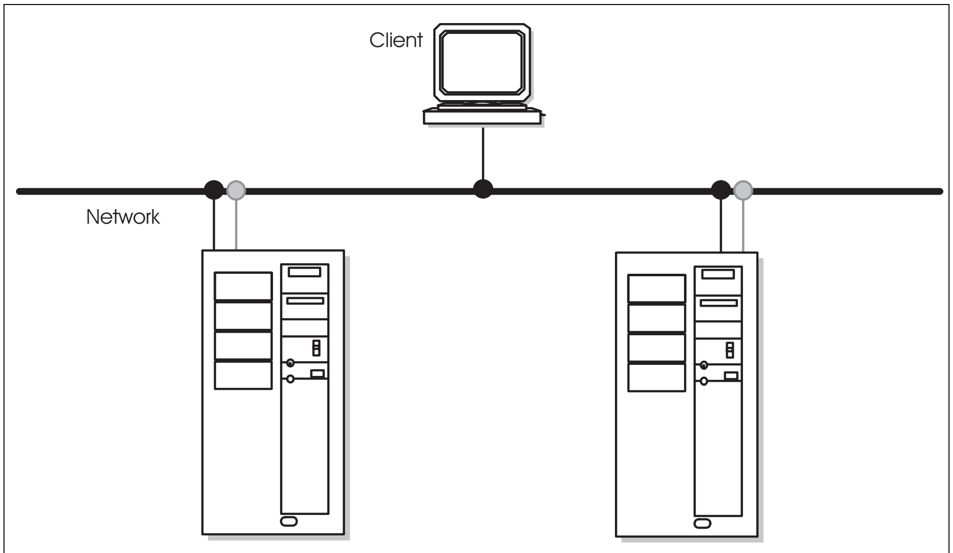
- Design the network topology
- Define a network mask for your site
- Define IP addresses (adapter identifiers) for each node’s service and standby adapters.
- Define a boot address for each service adapter that can be taken over, if you are using IP address takeover or rotating resources.
- Define an alternate hardware address for each service adapter that can have its IP address taken over, if you are using hardware address swapping.

2.4.3.1 Network Topology

The following sections cover topics of network topology.

Single Network

In a single-network setup, each node in the cluster is connected to only one network and has only one service adapter available to clients. In this setup, a service adapter on any of the nodes may fail, and a standby adapter will acquire its IP address. The network itself, however, is a single point of failure. The following figure shows a single-network configuration:



In the single-network setup, each node is connected to one network. Each node has one service adapter and can have none, one, or more standby adapters per public network.

Figure 5. Single-Network Setup

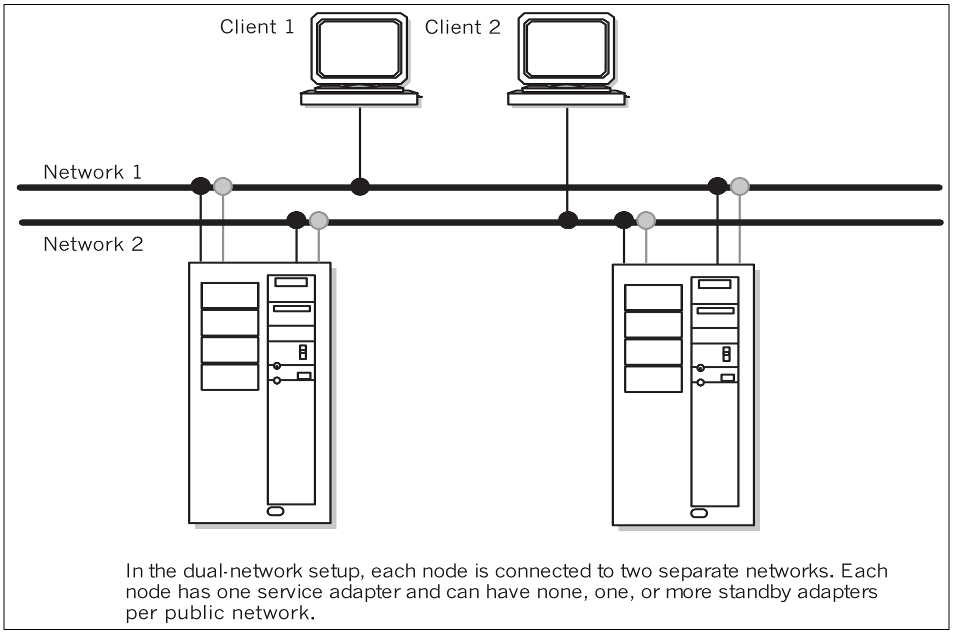
Dual Network

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

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

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

The following figure shows a dual-network setup:



In the dual-network setup, each node is connected to two separate networks. Each node has one service adapter and can have none, one, or more standby adapters per public network.

Figure 6. Dual-Network Setup

Point-to-Point Connection

A point-to-point connection links two (neighboring) cluster nodes directly. SOCC, SLIP, and ATM are point-to-point connection types. In HACMP clusters of four or more nodes, however, use an SOCC line *only* as a private network between neighboring nodes because it cannot guarantee cluster communications with nodes other than its neighbors.

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

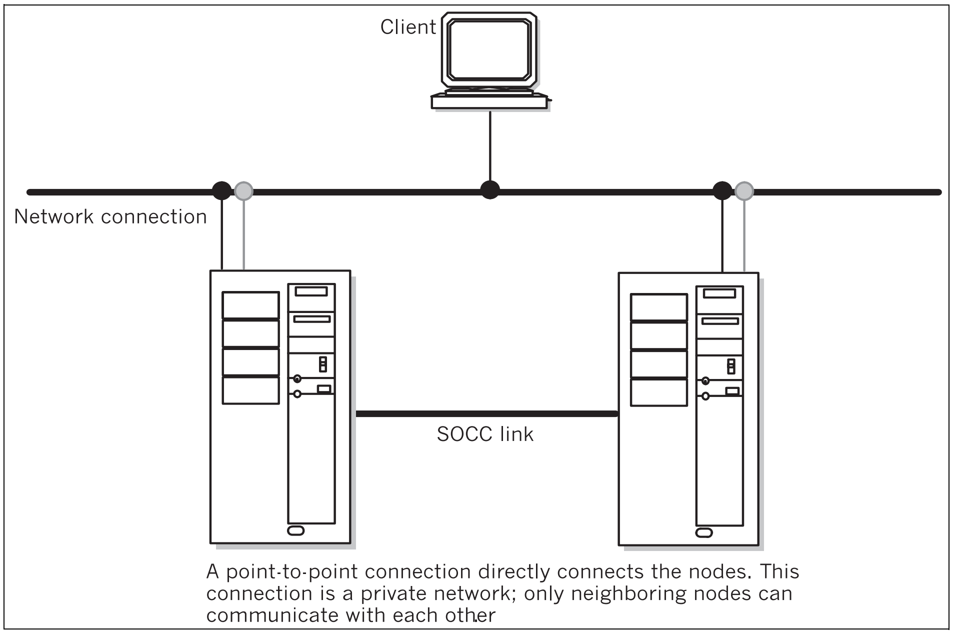

Client
Network connection
SOCC link
A point-to-point connection directly connects the nodes. This connection is a private network; only neighboring nodes can communicate with each other

Figure 7. A Point-to-Point Connection

2.4.3.2 Networks

Networks in an HACMP cluster are identified by name and attribute.

Network Name

The network name is a symbolic value that identifies a network in an HACMP for AIX environment. Cluster processes use this information to determine which adapters are connected to the same physical network. In most cases, the network name is arbitrary, and it must be used consistently. If several adapters share the same physical network, make sure that you use the same network name when defining these adapters.

Network Attribute

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

Public A public network connects from two to 32 nodes and allows clients to monitor or access cluster nodes. Ethernet, Token-Ring, FDDI, and

SLIP are considered public networks. Note that a SLIP line, however, does not provide client access.

Private A private network provides communication between nodes only; it typically does not allow client access. An SOCC line or an ATM network are also private networks; however, an ATM network does allow client connections and may contain standby adapters. If an SP node is used as a client, the SP Switch network, although private, can allow client access.

Serial This network attribute is used for non TCP/IP networks (see 2.2.2, “Non-TCPIP Networks” on page 14.)

2.4.3.3 Network Adapters

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

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

Adapter Label

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

Adapter Function

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

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

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

In an HACMP for AIX environment on the RS/6000 SP, the SP Ethernet adapters can be configured as service adapters but *should not* be configured for IP address takeover. For the SP switch network, service addresses used for IP address takeover are **ifconfig alias** addresses used on the css0 network.

Standby Adapter A standby adapter backs up a service adapter. If a service adapter fails, the Cluster Manager swaps the standby adapter's address with the service adapter's address. Using a standby adapter eliminates a network adapter as a single point of failure. A node can have no standby adapter, or it can have from one to seven standby adapters for each network to which it connects. Your software configuration and hardware slot constraints determine the actual number of standby adapters that a node can support.

The standby adapter is configured on a different subnet from any service adapters on the same system, and its use should be reserved for HACMP only.

In an HACMP for AIX environment on the RS/6000 SP, for an IP address takeover configuration using the SP switch, standby adapters are not required.

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

In an HACMP for AIX environment on the RS/6000 SP, boot addresses used in the IP address for the switch network takeover are **ifconfig alias** addresses used on that css0 network.

When the HACMP for AIX software 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, thus 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 adapter from the boot address to its reacquired service address.

Consider the following scenario: Suppose that Node A fails. Node B acquires Node A’s service address and services client requests directed to that address. Later, when Node A is restarted, it comes up on its boot address and attempts to reintegrate into the cluster on its service address by requesting that Node B release Node A’s service address. When Node B releases the requested address, Node A reclaims the address and reintegrates it into the cluster. Reintegration, however, fails if Node A has not been configured to boot using its boot address.

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

2.4.3.4 Defining Hardware Addresses

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

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

If you do not use Hardware Address Takeover, the ARP cache of clients can be updated by adding the clients' IP addresses to the `PING_CLIENT_LIST` variable in the /usr/sbin/cluster/etc/clinfo.rc file.

2.4.4 NFS Exports and NFS Mounts

There are two items concerning NFS when doing the configuration of a Resource Group:

Filesystems to Export File systems listed here will be NFS exported, so they can be mounted by NFS client systems or other nodes in the cluster.

Filesystems to NFS mount Filling in this field sets up what we call an *NFS cross mount*. Any file system defined in this field will be NFS mounted by all the participating nodes, other than the node that is currently holding the resource group. If the node holding the resource group fails, the next node to take over breaks its NFS mount for this file system, and mounts the file system itself as part of its takeover processing.

2.5 Application Planning

The central purpose for combining nodes in a cluster is to provide a highly available environment for mission-critical applications. These applications must remain available at all times in many organizations. For example, an HACMP cluster could run a database server program that services client applications. The clients send queries to the server program that responds to their requests by accessing a database that is stored on a shared external disk.

Planning for these applications requires that you be aware of their location within the cluster, and that you provide a solution that enables them to be handled correctly, in case a node should fail. 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 specially 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 creating application server resources, see the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278.

2.5.1 Performance Requirements

In order to plan your application’s needs, you must have a thorough understanding of it. One part of that is to have The Application Planning Worksheets, found in Appendix A of the *HACMP for AIX Planning Guide*, SC23-4277, filled out.

Your applications have to be served correctly in an HACMP cluster environment. Therefore, you need to know not only how they run on a single uni- or multiprocessor machine, but also which resources are required by them. How much disk space is required, what is the usual and critical load the application puts on a server, and how users access the application are some critical factors that will influence your decisions on how to plan the cluster.

Within an HACMP environment there are always a number of possible states in which the cluster could be. Under normal conditions, the load is serviced by a cluster node that was designed for this application’s needs. In case of a failover, another node has to handle its own work plus the application it is going to take over from a failing node. You can even plan one cluster node to be the takeover node for multiple nodes; so, when any one of its primary nodes fail, it has to take over its application and its load. Therefore, the performance requirements of any cluster application have to be understood in order to have the computing power available for mission-critical applications in all possible cluster states.

2.5.2 Application Startup and Shutdown Routines

Highly available applications do not only have to come up at boot time, or when someone is starting them up, but also when a critical resource fails and has to be taken over by another cluster node. In this case, there have to be robust scripts to both start up and shut down the application on the cluster nodes. The startup script especially must be able to recover the application from an abnormal termination, such as a power failure. You should verify that it runs properly in a uniprocessor environment before including the HACMP for AIX software.

Note

Application start and stop scripts have to be available on the primary as well as the takeover node. They are not transferred during synchronization; so, the administrator of a cluster has to ensure that they are found in the same path location, with the same permissions and in the same state, i.e. changes have to be transferred manually.

2.5.3 Licensing Methods

Some vendors require a unique license for each processor that runs an application, which means that you must license-protect the application by incorporating processor-specific information into the application when it is installed. As a result, it is possible that even though the HACMP for AIX software processes a node failure correctly, it is unable to restart the application on the failover node because of a restriction on the number of licenses available within the cluster for that application. To avoid this problem, make sure that you have a license for each system unit in the cluster that may potentially run an application.

This can be done by “floating licenses”, where a license server is asked to grant the permission to run an application on request, as well as “node-locked licenses”, where each processor possibly running an application must have the licensing files installed and configured.

2.5.4 Coexistence with other Applications

In case of a failover, a node might have to handle several applications concurrently. This means the applications data or resources *must not* conflict with each other. Again, the Application Worksheets can help in deciding whether certain resources might conflict with others.

2.5.5 Critical/Non-Critical Prioritizations

Building a highly available environment for mission-critical applications also forces the need to differentiate between the priorities of a number of applications. Should a server node fail, it might be appropriate to shut down another application, which is not as highly prioritized, in favor of the takeover of the server node’s application. The applications running in a cluster have to be clearly ordered and prioritized in order to decide what to do under these circumstances.

2.6 Customization Planning

The Cluster Manager’s ability to recognize a specific series of events and subevents permits a very flexible customization scheme. The HACMP for AIX software provides an event customization facility that allows you to tailor cluster event processing to your site.

2.6.1 Event Customization

As part of the planning process, you need to decide whether to customize event processing. If the actions taken by the default scripts are sufficient for your purposes, you do not need to do anything further to configure events during the installation process.

If you decide to tailor event processing to your environment, it is strongly recommended that you use the HACMP for AIX event customization facility described in this chapter. If you tailor event processing, you must register user-defined scripts with HACMP during the installation process. The *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278 describes how to configure event processing for a cluster.

You cannot define additional cluster events.

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

The event customization facility includes the following features:

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

2.6.1.1 Special Application Requirements

Some applications may have some special requirements that have to be checked and ensured before or after a cluster event happens. In case of a failover you can customize events through the definition of pre- and post-events, to act according to your application’s needs. For example, an application might want to reset a counter or unlock a user before it can be started correctly on the failover node.

2.6.1.2 Event Notification

You can specify a `notify` command that sends mail to indicate that an event is about to happen (or has just occurred), and that an event script succeeded or failed. For example, a site may want to use a **network_down** notification event to inform system administrators that traffic may have to be rerouted. Afterwards, you can use a **network_up** notification event to inform system administrators that traffic can again be serviced through the restored network.

2.6.1.3 Predictive Event Error Correction

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

For example, a recovery command can include the retry of unmounting a file system after logging a user off and making sure no one was currently accessing the file system.

If a condition that affects the processing of a given event on a cluster is identified, such as a timing issue, you can insert a recovery command with a retry count high enough to be sure to cover for the problem.

2.6.2 Error Notification

The AIX Error Notification facility detects errors that are logged to the AIX error log, such as network and disk adapter failures, and triggers a predefined response to the failure. It can even act on application failures, as long as they are logged in the error log.

To implement error notification, you have to add an object to the Error Notification object class in the ODM. This object clearly identifies what sort of errors you are going to react to, and how.

By specifying the following in a file:

```
errnotify:
            en_name = "Failuresample"
            en_persistenceflg = 0
            en_class = "H"
            en_type = "PERM"
            en_rclass = "disk"
            en_method = "errpt -a -l $1 | mail -s 'Disk Error' root"
```

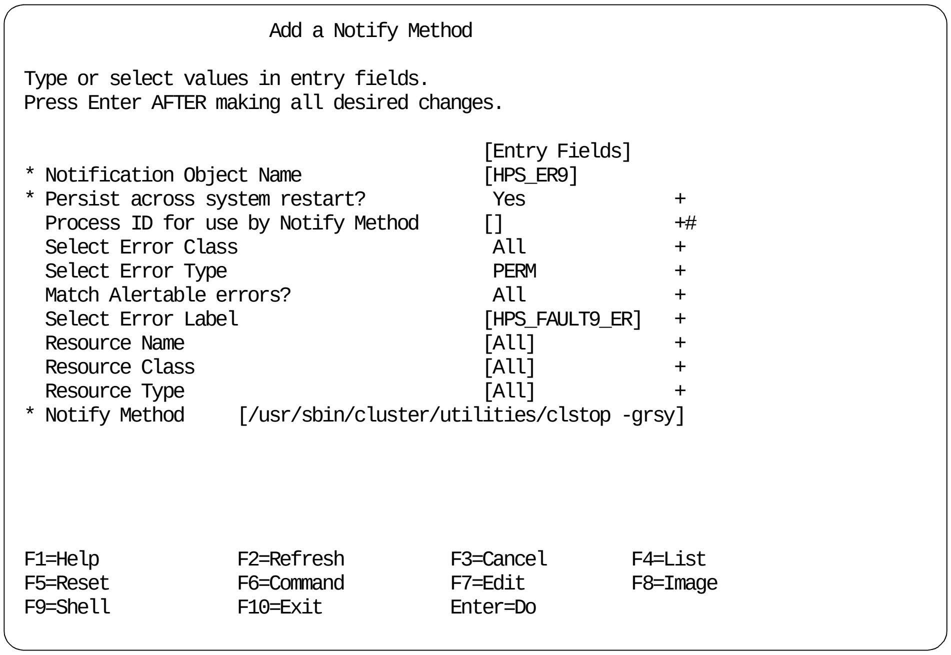
and adding this to the `errnotify` class through the `odmadd <filename>` command, the specified `en_method` is executed every time the error notification daemon finds a matching entry in the error report. In the example above, the root user will get e-mail identifying the exact error report entry.

2.6.2.1 Single Point-of-Failure Hardware Component Recovery

As described in 2.2.1.2, “Special Network Considerations” on page 12, the HPS Switch network is one resource that has to be considered as a single point of failure. Since a node can support only one switch adapter, its failure will disable the switch network for this node. It is strongly recommended to promote a failure like this into a node failure, if the switch network is critical to your operations.

Critical failures of the switch adapter would cause an entry in the AIX error log. Error labels like `HPS_FAULT9_ER` or `HPS_FAULT3_ER` are considered critical, and can be specified to AIX Error Notification in order to be able to act upon them.

With HACMP, there is a SMIT screen to make it easier to set up an error notification object. This is much easier than the traditional AIX way of adding a template file to the ODM class. Under `smit hacmp > RAS Support > Error Notification > Add a Notify Method`, you will find the menu allowing you to add these objects to the ODM. An example of the SMIT panel is shown below:



```
                        Add a Notify Method

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                           [Entry Fields]
* Notification Object Name                 [HPS_ER9]
* Persist across system restart?            Yes             +
  Process ID for use by Notify Method      []               +#
  Select Error Class                        All             +
  Select Error Type                         PERM            +
  Match Alertable errors?                   All             +
  Select Error Label                       [HPS_FAULT9_ER]  +
  Resource Name                            [All]            +
  Resource Class                           [All]            +
  Resource Type                            [All]            +
* Notify Method     [/usr/sbin/cluster/utilities/clstop -grsy]




F1=Help             F2=Refresh          F3=Cancel        F4=List
F5=Reset            F6=Command          F7=Edit          F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 8. Sample Screen for Add a Notification Method

The above example screen will add a Notification Method to the ODM, so that upon appearance of the HPS_FAULT9_ER entry in the error log, the error notification daemon will trigger the execution of the `/usr/sbin/cluster/utilities/clstop -grsy` command, which shuts HACMP down gracefully with takeover. In this way, the switch failure is acted upon as a node failure.

2.6.2.2 Notification

The method that is triggered upon the appearance of a specified error log entry will be run by the error notification daemon with the command `sh -c <en_method>`. Because this a regular shell, any shell script can act as a method.

So, if you want a specific notification, such as e-mail from this event, you can define a script that sends e-mail and then issues the appropriate commands.

Note

Because the Notification Method is an object in the node's ODM, it has to be added to each and every node potentially facing a situation where it would be wise to act upon the appearance of an error log entry.
This is NOT handled by the HACMP synchronization facility. You have to take care of this manually.

Alternatively, you can always customize any cluster event to enable a Notify Command whenever this event is triggered through the SMIT screen for customizing events.

2.6.2.3 Application Failure

Even application failures can cause an event to happen, if you have configured this correctly. To do so, you have to find some method to decide whether an application has failed. This can be as easy as looking for a specific process, or much more complex, depending on the application. If you issue an Operator Message through the

```
errlogger <message>
```

command, you can act on that as you would on an error notification as described in 2.6.2.1, “Single Point-of-Failure Hardware Component Recovery” on page 46.

2.7 User ID Planning

The following sections describe various aspects of User ID Planning.

2.7.1 Cluster User and Group IDs

One of the basic tasks any system administrator must perform is setting up user accounts and groups. All users require accounts to gain access to the system. Every user account must belong to a group. Groups provide an additional level of security and allow system administrators to manipulate a group of users as a single entity.

For users of an HACMP for AIX cluster, system administrators must create duplicate accounts on each cluster node. The user account information stored in the /etc/passwd file, and in other files stored in the /etc/security directory, should be consistent on all cluster nodes. For example, if a cluster node fails, users should be able to log on to the surviving nodes without experiencing problems caused by mismatches in the user or group IDs.

System administrators typically keep user accounts synchronized across cluster nodes by copying the key system account and security files to all cluster nodes whenever a new account is created or an existing account is changed.Typically `rdist` or `rcp` is used, for that. On RS/6000 SP systems `pcp` or `supper` are widely used. For C-SPOC clusters, the C-SPOC utility simplifies the cluster-wide synchronization of user accounts by propagating the new account or changes to an existing account across all cluster nodes automatically.

The following are some common user and group management tasks, and are briefly explained in 8.8, “User Management” on page 178:

- Listing all user accounts on all cluster nodes
- Adding users to all cluster nodes
- Changing characteristics of a user account on all cluster nodes
- Removing a user account from all cluster nodes.
- Listing all groups on all cluster nodes
- Adding groups to all cluster nodes
- Changing characteristics of a group on all cluster nodes
- Removing a group from all cluster nodes

2.7.2 Cluster Passwords

While user and group management is very much facilitated with C-SPOC, the password information still has to be distributed by some other means. If the system is not configured to use NIS or DCE, the system administrator still has to distribute the password information, meaning that found in the /etc/security/password file, to all cluster nodes.

As before, this can be done through `rdist` or `rcp`. On RS/6000 SP systems, there are tools like `pcp` or `supper` to distribute information or better files.

2.7.3 User Home Directory Planning

As for user IDs, the system administrator has to ensure that users have their home directories available and in the same position at all times. That is, they don’t care whether a takeover has taken place or everything is normal. They simply want to access their files, wherever they may reside physically, under the same directory path with the same permissions, as they would on a single machine.

There are different approaches to that. You could either put them on a shared volume and handle them within a resource group, or you could use NFS mounts.

2.7.3.1 Home Directories on Shared Volumes

Within an HACMP cluster, this approach is quite obvious, however, it restricts you to only one machine where a home directory can be active at any given time. If you have only one application that the user needs to access, or all of the applications are running on one machine, where the second node serves as a standby machine only, this would be sufficient.

2.7.3.2 NFS-Mounted Home Directories

The NFS mounted home directory approach is much more flexible. Because the directory can be mounted on several machines at the same time, a user can work with it in several applications on several nodes at the same time.

However, if one cluster node provides NFS service of home directories to other nodes, in case of a failure of the NFS server node, the access to the home directories is barred. Placing them onto a machine outside the cluster doesn’t help either, since this again introduces a single point of failure, and machines outside the cluster are not any less likely to fail than machines within.

2.7.3.3 NFS-Mounted Home Directories on Shared Volumes

So, a combined approach is used in most cases. In order to make home directories a highly available resource, they have to be part of a resource group and placed on a shared volume. That way, all cluster nodes can access them in case they need to.

To make the home directories accessible on nodes that currently do not own the resource where they are physically residing, they have to be NFS exported from the resource group and imported on all the other nodes in case any application is running there, needing access to the users files.

In order to make the directory available to users again, when a failover happens, the takeover node that previously had the directory NFS mounted from the failed node has to break locks on NFS files, if there are any. Next, it must unmount the NFS directory, acquire the shared volume (varyon the shared volume group) and mount the shared file system. Only after that can users access the application on the takeover node again.

Chapter 3. Cluster Hardware and Software Preparation

This chapter covers the steps that are required to prepare the RS/6000 hardware and AIX software for the installation of HACMP and the configuration of the cluster. This includes configuring adapters for TCP/IP, setting up shared volume groups, and mirroring and editing AIX configuration files.

3.1 Cluster Node Setup

The following sections describe important details of cluster node setup.

3.1.1 Adapter Slot Placement

For information regarding proper adapter placement, see the following documentation:

- *PCI Adapter Placement Reference Guide,* SA38-0538
- *Adapters, Devices, and Cable Information for Micro Channel Bus Systems*, SA38-0533
- *Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516

3.1.2 Rootvg Mirroring

Of all the components used to build a computer system, physical disk devices are usually the most susceptible to failure. Because of this, disk mirroring is a frequently used technique for increasing system availability.

File system mirroring and disk mirroring are easily configured using the AIX Logical Volume Manager. However, conventional file system and disk mirroring offer no protection against operating system failure or against a failure of the disk from which the operating system normally boots.

Operating system failure does not always occur instantaneously, as demonstrated by a system that gradually loses access to operating system services. This happens as code and data that were previously being accessed from memory gradually disappear in response to normal paging.

Normally, in an HACMP environment, it is not necessary to think about mirroring the root volume group, because the node failure facilities of HACMP can cover for the loss of any of the rootvg physical volumes. However, it is possible that a customer with business-critical applications will justify mirroring rootvg in order to avoid the impact of the failover time involved in a node failure. In terms of maximizing availability, this technique is just as valid for increasing the availability of a cluster as it is for increasing single-system availability.

The following procedure contains information that will enable you to mirror the root volume group (rootvg), using the advanced functions of the Logical Volume Manager (LVM). It contains the steps required to:

- Mirror all the file systems in rootvg.
- Create an additional boot logical volume (blv).
- Modify the bootlist to contain all boot devices.

You may mirror logical volumes in the rootvg in the same way as any AIX logical volume may be mirrored, either once (two copies), or twice (three copies). The following procedure is designed for mirroring rootvg to a second disk only. Upon completion of these steps, your system will remain available if one of the disks in rootvg fails, and will even automatically boot from an alternate disk drive, if necessary.

If the dump device is mirrored, you may not be able to capture the dump image from a crash or the dump image may be corrupted. The design of LVM prevents mirrored writes of the dump device. Only one of the mirrors will receive the dump image. Depending on the boot sequence and disk availability after a crash, the dump will be in one of the following three states:

1. Not available
2. Available and not corrupted
3. Available and corrupted

State (1) will always be a possibility. If the user prefers to prevent the risk of encountering State (3), then the user must create a non-mirrored logical volume (that is not hd6) and set the dump device to this non-mirrored logical volume.

In AIX 4.2.1, two new LVM commands were introduced: `mirrorvg` and `unmirrorvg`. These two commands where introduced to simplify mirroring or unmirroring of the entire contents of a volume group. The commands will detect if the entity to be mirrored or unmirrored is rootvg, and will give slightly different completion messages based on the type of volume group.The `mirrorvg` command does the equivalent of Procedure steps (2), (3), and (4).

The `mirrorvg` command takes dump devices and paging devices into account. If the dump devices are also the paging device, the logical volume will be mirrored. If the dump devices are NOT the paging device, that dump logical volume will not be mirrored.

3.1.2.1 Procedure

The following steps assume the user has rootvg contained on hdisk0 and is attempting to mirror the rootvg to a new disk: hdisk1.

1. Extend rootvg to hdisk1 by executing the following:

   ```
   extendvg rootvg hdisk1
   ```

2. Disable QUORUM, by executing the following:

   ```
   chvg -Qn rootvg
   ```

3. Mirror the logical volumes that make up the AIX operating system by executing the following:

   ```
   mklvcopy hd1 2 hdisk1 # /home file system
   mklvcopy hd2 2 hdisk1 # /usr file system
   mklvcopy hd3 2 hdisk1 # /tmp file system
   mklvcopy hd4 2 hdisk1 # / (root) file system
   mklvcopy hd5 2 hdisk1 # blv, boot logical volume
   mklvcopy hd6 2 hdisk1 # paging space
   mklvcopy hd8 2 hdisk1 # file system log
   mklvcopy hd9var 2 hdisk1 # /var file system
   ```

   If you have other paging devices, rootvg and non-rootvg, it is recommended that you also mirror those logical volumes in addition to hd6.

   If hd5 consists of more than one logical partition, then, after mirroring hd5 you must verify that the mirrored copy of hd5 resides on contiguous physical partitions. This can be verified with the following command:

   ```
   lslv -m hd5
   ```

   If the mirrored hd5 partitions are not contiguous, you must delete the mirror copy of hd5 (on hdisk1) and rerun the `mklvcopy` for hd5, using the

"-m" option. You should consult documentation on the usage of the "-m" option for mklvcopy.

4. Synchronize the newly created mirrors with the following command:

```
syncvg -v rootvg
```

5. Bosboot to initialize all boot records and devices by executing the following command:

```
bosboot -a -d /dev/hdisk?
```

   where hdisk? is the first hdisk listed under the "PV" heading after the command lslv -l hd5 has executed.

6. Initialize the boot list by executing the following:

```
bootlist -m normal hdisk0 hdisk1
```

Note

Even though this command identifies the list of possible boot disks, it does not guarantee that the system will boot from the alternate disk in all cases involving failures of the first disk. In such situations, it may be necessary for the user to boot from the installation/maintenance media. Select maintenance, reissue the bootlist command leaving out the failing disk, and then reboot. On some models, firmware provides a utility for selecting the boot device at boot time. This may also be used to force the system to boot from the alternate disk.

7. Shutdown and reboot the system by executing the following command:

```
shutdown -Fr
```

   This is so that the "Quorum OFF" functionality takes effect.

3.1.2.2 Necessary APAR Fixes

Table 11. Necessary APAR Fixes

AIX Version	APARs needed
4.1	IX56564 IX61184 IX60521
4.2	IX62417 IX68483 IX70884 IX72058
4.3	IX72550

To determine if either fix is installed on a machine, execute the following:

```
instfix -i -k <apar number>
```

3.1.3 AIX Prerequisite LPPs

In order to install HACMP and HACMP/ES the AIX setup must be in a proper state. The following table gives you the prerequisite AIX levels for the different HACMP versions:

Table 12. AIX Prerequisite LPPs

HACMP Version	Prerequisite AIX and PSSP Version
HACMP 4.1 for AIX	AIX 4.1.5 PSSP 2.2, if installed on an SP
HACMP 4.2 for AIX	AIX 4.1.5 PSSP 2.2, if installed on an SP
HACMP 4.3 for AIX	AIX 4.3.2 PSSP 2.2, if installed on an SP
HACMP/ES 4.2 for AIX	AIX 4.2.1 PSSP 2.2, if installed on an SP
HACMP/ES 4.3 for AIX	AIX 4.3.2 PSSP 3.1, if installed on an SP

The Prerequisites for the HACMP component HAView 4.2 are

- xlC.rte 3.1.3.0
- nv6000.base.obj 4.1.0.0

- nv6000.database.obj 4.1.0.0
- nv6000.Features.obj 4.1.2.0
- nv6000.client.obj 4.1.0.0

and for HAView 4.3

- xlC.rte 3.1.4.0
- nv6000.base.obj 4.1.2.0
- nv6000.database.obj 4.1.2.0
- nv6000.Features.obj 4.1.2.0
- nv6000.client.obj 4.1.2.0

3.1.4 AIX Parameter Settings

This section discusses several general tasks necessary to ensure that your HACMP for AIX cluster environment works as planned. Consider or check the following issues to ensure that AIX works as expected in an HACMP cluster.

- I/O pacing
- User and group IDs (see Chapter 2.7, “User ID Planning” on page 48)
- Network option settings
- /etc/hosts file and nameserver edits
- /.rhosts file edits

3.1.4.1 I/O Pacing

AIX users have occasionally seen poor interactive performance from some applications when another application on the system is doing heavy input/output. Under certain conditions, I/O can take several seconds to complete. While the heavy I/O is occurring, an interactive process can be severely affected if its I/O is blocked, or, if it needs resources held by a blocked process.

Under these conditions, the HACMP for AIX software may be unable to send keepalive packets from the affected node. The Cluster Managers on other cluster nodes interpret the lack of keepalives as node failure, and the I/O-bound node is “failed” by the other nodes. When the I/O finishes, the node resumes sending keepalives. Its packets, however, are now out of sync with the other nodes, which then kill the I/O-bound node with a RESET packet.

You can use I/O pacing to tune the system so that system resources are distributed more equitably during high disk I/O. You do this by setting high-

and low-water marks. If a process tries to write to a file at the high-water mark, it must wait until enough I/O operations have finished to make the low-water mark.

Use the `smit chgsys` fastpath to set high- and low-water marks on the Change/Show Characteristics of the Operating System screen.

By default, AIX is installed with high- and low-water marks set to zero, which disables I/O pacing.

While enabling I/O pacing may have a slight performance effect on very I/O-intensive processes, it is required for an HACMP cluster to behave correctly during large disk writes. If you anticipate heavy I/O on your HACMP cluster, you should enable I/O pacing.

Although the most efficient high- and low-water marks vary from system to system, an initial high-water mark of 33 and a low-water mark of 24 provides a good starting point. These settings only slightly reduce write times and consistently generate correct fallover behavior from the HACMP for AIX software.

See the *AIX Performance Monitoring & Tuning Guide*, SC23-2365, for more information on I/O pacing.

3.1.4.2 Checking Network Option Settings

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

```
thewall = 5120
```

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

```
no -o thewall=5120
```

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

To list the values of other network options (not configurable) that are currently set on a node, enter:

```
no -a
```

3.1.4.3 Editing the /etc/hosts File and Nameserver Configuration

Make sure all nodes can resolve all cluster addresses. See the chapter on planning TCP/IP networks (the section Using HACMP with NIS and DNS) in the *HACMP for AIX, Version 4.3: Planning Guide, SC23-4277* for more information on name serving and HACMP.

Edit the /etc/hosts file (and the /etc/resolv.conf file, if using the nameserver configuration) on each node in the cluster to make sure the IP addresses of all clustered interfaces are listed.

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

```
100.100.50.200 crab_boot
```

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

```
127.0.0.1    loopback localhost
```

3.1.4.4 cron and NIS Considerations

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

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

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

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

3.1.4.5 Editing the /.rhosts File

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

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

The /.rhosts file is not required on SP systems running the HACMP Enhanced Security. This feature removes the requirement of TCP/IP access control lists (for example, the /.rhosts file) on remote nodes during HACMP configuration.

3.2 Network Connection and Testing

The following sections describe important aspects of network connection and testing.

3.2.1 TCP/IP Networks

Since there are several types of TCP/IP Networks available within HACMP, there are several different characteristics and some restrictions on them. Characteristics like maximum distance between nodes have to be considered. You don’t want to put two cluster nodes running a mission-critical application in the same room for example.

3.2.1.1 Cabling Considerations

Characteristics of the different types of cable, their maximum length, and the like are beyond the scope of this book. However, for actual planning of your clusters, you have to check whether your network cabling allows you to put two cluster nodes away from each other, or even in different buildings.

There’s one additional point with cabling, that should be taken care of. Cabling of networks often involves hubs or switches. If not carefully planned, this sometimes introduces another single point of failure into your cluster. To eliminate this you should have at least two hubs.

As shown in Figure 9, failure of a hub would not result in one machine being disconnected from the network. In that case, a hub failure would cause either both service adapters to fail, which would cause a *swap_adapter* event, and the standby adapters would take over the network, or both standby adapters would fail, which would cause *fail_standby* events. Configuring a notify method for these events can alert the network administrator to check and fix the broken hub.

.

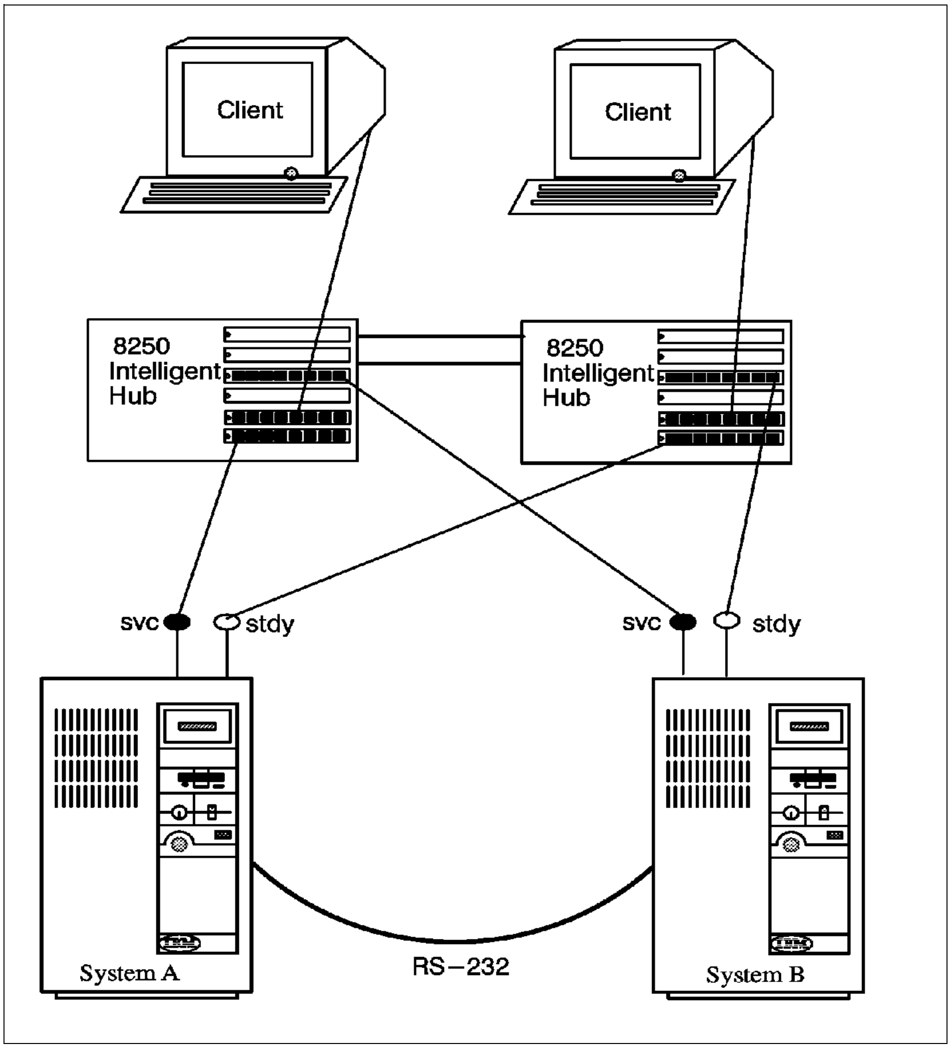


Figure 9. Connecting Networks to a Hub

3.2.1.2 IP Addresses and Subnets

The design of the HACMP for AIX software specifies that:

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

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

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

Note

The netmask for all adapters in an HACMP network must be the same even though the service and standby adapters are on different logical subnets. See the *HACMP for AIX, Version 4.3: Concepts and Facilities, SC23-4276* guide for more information about using the same netmask for all adapters.

See Chapter 2.4.3, “IP Address Takeover” on page 34 for more detailed information.

3.2.1.3 Testing

After setting up all adapters with AIX, you can do several things to check whether TCP/IP is working correctly. Note, that without HACMP being started, the service adapters defined to HACMP will remain on their boot address. After startup these adapters change to their service addresses.

Use the following AIX commands to investigate the TCP/IP subsystem:

- Use the `netstat` command to make sure that the adapters are initialized and that a communication path exists between the local node and the target node.
- Use the `ping` command to check the point-to-point connectivity between nodes.
- Use the `ifconfig` command on all adapters to detect bad IP addresses, incorrect subnet masks, and improper broadcast addresses.

- Scan the /tmp/hacmp.out file to confirm that the /etc/rc.net script has run successfully. Look for a zero exit status.
- If IP address takeover is enabled, confirm that the /etc/rc.net script has run and that the service adapter is on its service address and not on its boot address.
- Use the `lssrc -g tcpip` command to make sure that the inetd daemon is running.
- Use the `lssrc -g portmap` command to make sure that the portmapper daemon is running.
- Use the `arp` command to make sure that the cluster nodes are not using the same IP or hardware address.

3.2.2 Non TCP/IP Networks

Currently three types of non-TCP/IP networks are supported:

- Serial (RS232)
- Target-mode SCSI
- Target-mode SSA

While we use the word serial here to refer to RS232 only, in HACMP definitions, a “serial” network means a non-TCP/IP network of any kind. Therefore, when we are talking about HACMP network definitions, a serial network could also be a target-mode SCSI or target-mode SSA network.

The following describes some cabling issues on each type of non-TCP/IP network, how they are to be configured, and how you can test if they are operational.

3.2.2.1 Cabling Considerations

RS232 Cabling a serial connection requires a null-modem cable. As often cluster nodes are further apart than 60 m (181 ft.), sometimes modem eliminators or converters to fiber channel are used.

TMSCSI If your cluster uses SCSI disks as shared devices, you can use that line for TMSCSI as well. TMSCSI requires Differential SCSI adapters (see Chapter 2.3.2.3, “Supported SCSI Adapters” on page 26). Because the SCSI bus has to be terminated on both ends, and not anywhere else in between, resistors on the adapters should be removed, and cabling should be done as shown in Figure 11 on page 77, that is, with Y-cables that are terminated at one end connected to the adapters where the other end connects to the shared disk device.

TMSSA Target-mode SSA is only supported with the SSA Multi-Initiator RAID Adapters (Feature #6215 and #6219), Microcode Level 1801 or later. You need at least HACMP Version 4.2.2 with APAR IX75718.

3.2.2.2 Configuring RS232

Use the `smit tty` fastpath to create a tty device on the nodes. On the resulting panel, you can add an RS232 tty by selecting a native serial port, or a port on an asynchronous adapter. Make sure that the Enable Login field is set to `disable`. You do not want a getty process being spawned on this interface.

3.2.2.3 Configuring Target Mode SCSI

To configure a target-mode SCSI network on the Differential SCSI adapters, you have to enable the SCSI adapter's feature **TARGET MODE** by setting the **enabled** characteristics to **yes**. Since disks on the SCSI bus are normally configured at boot time, and the characteristics of the parent device cannot be changed as long as there are child devices present and active, you have to set all the disks on that bus to `Defined` with the

```
rmdev -l hdiskx
```

command, before you can enable that feature. Alternatively you can make these changes to the database (ODM) only, and they will be activated at the time of the next reboot.

If you choose not to reboot, instead setting all the child devices to `Defined`, you have to run cfgmgr, to get the tmscsi device created, as well as all the child devices of the adapter back to the available state.

Note

The target mode device created is a logical new device on the bus. Because it is created by scanning the bus for possible initiator devices, a `tmscsix` device is created on a node for each SCSI adapter on the same bus that has the target mode flag enabled, therefore representing this adapter's unique SCSI ID. In that way, the initiator can address packets to exactly one target device.

This procedure has to be done for all the cluster nodes that are going to use a serial network of type tmscsi as defined in your planning sheets.

3.2.2.4 Configuring Target Mode SSA

The node number on each system needs to be changed from the default of zero to a number. All systems on the SSA loop must have a unique node number.

To change the node number use the following command:

```
chdev -l ssar -a node_number=#
```

To show the system's node number use the following command:

```
lsattr -El ssar
```

Having the node numbers set to non-zero values enables the target mode devices to be configured. Run the `cfgmgr` command to configure the `tmssa#` devices on each system. Check that the tmssa devices are available on each system using the following command:

```
lsdev -C | grep tmssa
```

The Target Mode SCSI or SSA serial network can now be configured into an HACMP cluster.

3.2.2.5 Testing RS232 and Target Mode Networks

Testing of the serial networks functionality is similar. Basically you just write to one side's device and read from the other.

Serial (RS323): After configuring the serial adapter and cabling it correctly, you can check the functionality of the connection by entering the command

```
cat < /dev/ttyx
```

on one node for reading from that device and

```
cat /etc/environment > /dev/ttyy
```

on the corresponding node for writing. You should see the first command hanging until the second command is issued, and then showing the output of it.

Target Mode SSA: After configuration of Target Mode SSA, you can check the functionality of the connection by entering the command:

```
cat < /dev/tmssax.tm
```

on one node for reading from that device and:

```
cat /etc/environment > /dev/tmssay.im
```

on the corresponding node for writing. x and y correspond to the appropriate opposite nodenumber. You should see the first command hanging until the second command is issued, and then showing its output.

Target Mode SCSI: After configuration of Target Mode SCSI, you can check the functionality of the connection by entering the command:

```
cat < /dev/tmscsix.tm
```

on one node for reading from that device and:

```
cat /etc/environment > /dev/tmscsiy.im
```

on the corresponding node for writing. You should see the first command hanging until the second command is issued, and then showing the output of that second command.

3.3 Cluster Disk Setup

The following sections relate important information about cluster disk setup.

3.3.1 SSA

The following sections describe cabling, AIX configuration, microcode loading, and configuring a RAID on SSA disks.

3.3.1.1 Cabling

The following rules must be followed when connecting a 7133 SSA Subsystem:

- Each SSA loop must be connected to a valid pair of connectors on the SSA adapter card (A1 and A2 to form one loop, or B1 and B2 to form one loop).
- Only one pair of connectors of an SSA adapter can be connected in a particular SSA loop (A1 or A2, with B1 or B2 cannot be in the same SSA loop).
- A maximum of 48 disks can be connected in an SSA loop.
- A maximum of three dummy disk drive modules can be connected next to each other.
- The maximum length of an SSA cable is 25 m. With Fiber-Optic Extenders, the connection length can be up to 2.4 km.

For more information regarding adapters and cabling rules see 2.3.1, “SSA Disks” on page 16 or the following documents:

- 7133 SSA Disk Subsystems: Service Guide, SY33-0185-02
- 7133 SSA Disk Subsystem: Operator Guide, GA33-3259-01
- 7133 Models 010 and 020 SSA Disk Subsystems: Installation Guide, GA33-3260-02
- 7133 Models 500 and 600 SSA Disk Subsystems: Installation Guide, GA33-3263-02
- 7133 SSA Disk Subsystems for Open Attachment: Service Guide, SY33-0191-00
- 7133 SSA Disk Subsystems for Open Attachment: Installation and User's Guide, SA33-3273-00

3.3.1.2 AIX Configuration

During boot time, the configuration manager of AIX configures all the device drivers needed to have the SSA disks available for usage. The configuration manager can’t do this configuration if the SSA Subsystem is not properly connected or if the SSA Software is not installed. If the SSA Software is not already installed, the configuration manager will tell you the missing filesets. You can either install the missing filesets with `smit`, or call the configuration manager with the -i flag.

The configuration manager configures the following devices:

- SSA Adapter Router
- SSA Adapter
- SSA Disks

Adapter Router

The adapter Router (ssar) is only a conceptual configuration aid and is always in a “Defined” state. It cannot be made “Available.” You can list the ssar with the following command:

```
#lsdev -C | grep ssar
ssar       Defined                 SSA Adapter Router
```

Adapter Definitions

By issuing the following command, you can check the correct adapter configuration. In order to work correctly, the adapter must be in the “Available” state:

```
#lsdev -C | grep ssa
ssa0       Available 00-07          SSA Enhanced Adapter
ssar       Defined                  SSA Adapter Router
```

The third column in the adapter device line shows the location of the adapter.

Disk Definitions

SSA disk drives are represented in AIX as SSA logical disks (hdisk0, hdisk1,...,hdiskN) and SSA physical disks (pdisk0, pdisk1,...,pdiskN). SSA RAID arrays are represented as SSA logical disks (hdisk0, hdisk1,...,hdiskN). SSA logical disks represent the logical properties of the disk drive or array, and can have volume groups and file systems mounted on them. SSA physical disks represent the physical properties of the disk drive. By default, one pdisk is always configured for each physical disk drive. One hdisk is configured for each disk drive that is connected to the using system, or for each array. By default, all disk drives are configured as system (AIX) disk drives. The array management software can be used to change the disks from hdisks to array candidate disks or hot spares.

SSA logical disks:

- Are configured as hdisk0, hdisk1,...,hdiskN.
- Support a character special file (/dev/rhdisk0, /dev/rhdisk1,...,/dev/rhdiskN).
- Support a block special file (/dev/hdisk0, /dev/hdisk1,...,/dev/hdiskN).
- Support the I/O Control (IOCTL) subroutine call for non service and diagnostic functions only.
- Accept the read and write subroutine calls to the special files.
- Can be members of volume groups and have file systems mounted on them.

In order to list the logical disk definitions, use the following command:

```
#lsdev -Cc disk| grep SSA
hdisk3   Available 00-07-L          SSA Logical Disk Drive
hdisk4   Available 00-07-L          SSA Logical Disk Drive
hdisk5   Available 00-07-L          SSA Logical Disk Drive
hdisk6   Available 00-07-L          SSA Logical Disk Drive
hdisk7   Available 00-07-L          SSA Logical Disk Drive
hdisk8   Available 00-07-L          SSA Logical Disk Drive
```

SSA physical disks:

- Are configured as pdisk0, pdisk1,...,pdiskN.
- Have errors logged against them in the system error log.
- Support a character special file (/dev/pdisk0, /dev/pdisk1,...,/dev/p.diskN).
- Support the IOCTLl subroutine for servicing and diagnostic functions.
- Do not accept read or write subroutine calls for the character special file.

In order to list the physical disk definitions use the following command:

```
#lsdev -Cc pdisk| grep SSA
pdisk0 Available 00-07-P 1GB SSA C Physical Disk Drive
pdisk1 Available 00-07-P 1GB SSA C Physical Disk Drive
pdisk2 Available 00-07-P 1GB SSA C Physical Disk Drive
pdisk3 Available 00-07-P 1GB SSA C Physical Disk Drive
pdisk4 Available 00-07-P 1GB SSA C Physical Disk Drive
pdisk5 Available 00-07-P 1GB SSA C Physical Disk Drive
```

Diagnostics

A good tool to get rid of SSA problems are the SSA service aids in the AIX diagnostic program `diag`. The SSA diagnostic routines are fully documented in *A Practical Guide to SSA for AIX*, SG24-4599. The following is a brief overview:

The SSA service aids are accessed from the main menu of the `diag` program. Select **Task Selection -> SSA Service Aids**. This will give you the following options:

Set Service Mode — This option enables you to determine the location of a specific SSA disk drive within a loop and to remove the drive from the configuration, if required.

Link Verification — This option enables you to determine the operational status of a link

Configuration Verification This option enables you to display the relationships between physical (pdisk) and logical (hdisk) disks.

Format Disk This option enables you to format SSA disk drives.

Certify Disk This option enables you to test whether data on an SSA disk drive can be read correctly.

Display/Download... This option enables you to display the microcode level of the SSA disk drives and to download new microcode to individual or all SSA disk drives connected to the system.

> **Note**
>
> When an SSA loop is attached to multiple host systems, do not invoke the diagnostic routines from more than one host simultaneously, to avoid unpredictable results that may result in data corruption.

3.3.1.3 Microcode Loading

To ensure that everything works correctly, install the latest filesets, fixes and microcode for your SSA disk subsystem. The latest information and downloadable files can be found under `http://www.hursley.ibm.com/~ssa`.

Upgrade Instructions

Follow these steps to perform an upgrade:

1. Login as root
2. Download the appropriate microcode file for your AIX version from the web-site mentioned above
3. Save the file upgrade.tar in your /tmp directory
4. Type tar -xvf upgrade.tar
5. Run `smitty install`
6. Select **install & update software**
7. Select **install & update from ALL available software**
8. Use the directory /usr/sys/inst.images as the install device
9. Select all filesets in this directory for install
10. Execute the command
11. Exit Smit

Note

You must ensure that:

- You do not attempt to perform this adapter microcode download concurrently on systems that are in the same SSA loop. This may cause a portion of the loop to be isolated and could prevent access to these disks from elsewhere in the loop.
- You do not run advanced diagnostics while downloads are in progress. Advanced diagnostics causes the SSA adapter to be reset temporarily, thereby introducing a break in the loop, portions of the loop may become temporarily isolated and inaccessible.
- You have complete SSA loops. Check this by using diagnostics in System Verification mode. If you have incomplete loops (such as strings) action must be taken to resolve this before you can continue.
- All of your loops are valid, in this case with one or two adapters in each loop. This is also done by using Diagnostics in System Verification mode.

12. Run `cfgmgr` to install the microcode to adapters.
13. To complete the device driver upgrade, you must now reboot your system.
14. To confirm that the upgrade was a success, type `lscfg -vl ssaX` where X is 0,1... for all SSA adapters. Check the ROS Level line to see that each adapter has the appropriate microcode level (for the correct microcode level, see the above mentioned web-site).
15. Run `lslpp -l|grep SSA` and check that the fileset levels you have match, or are above the levels shown in the list on the above mentioned web-site. If any of the SSA filesets are at a lower level than those shown in the above link, please repeat the whole upgrade procedure again. If, after repeating the procedure, the code levels do not match the latest ones, place a call with your local IBM Service Center.
16. If the adapters are in SSA loops which contain other adapters in other systems, please repeat this procedure on all systems as soon as possible.
17. In order to install the disk microcode, run `ssadload -u` from each system in turn.

    Note

    Allow ssadload to complete on one system before running it on another.

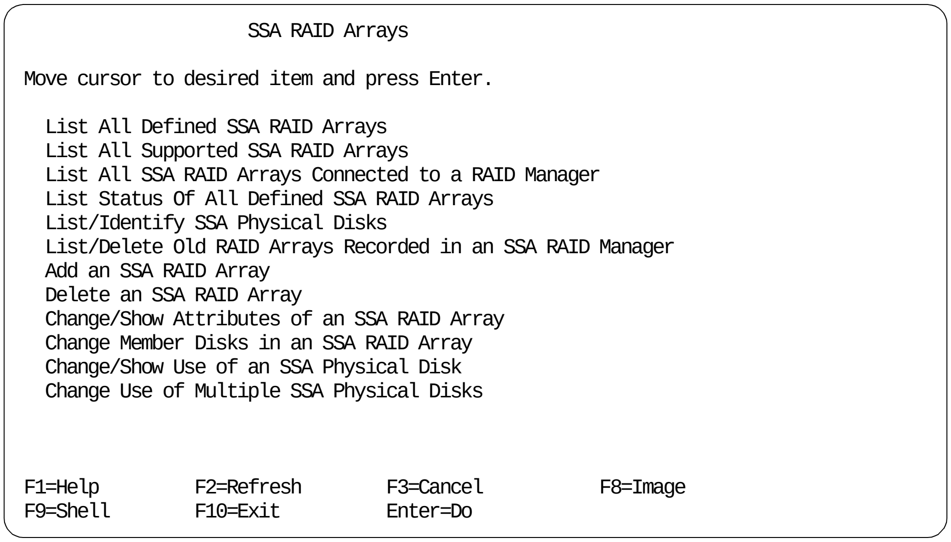
18. To confirm that the upgrade was a success, type `lscfg -vl pdiskX` where X is 0,1... for all SSA disks. Check the ROS Level line to see that each disk has the appropriate microcode level (for the correct microcode level see the above mentioned web-site).

3.3.1.4 Configuring a RAID on SSA Disks

Disk arrays are groups of disk drives that act like one disk as far as the operating system is concerned, and which provide better availability or performance characteristics than the individual drives operating alone. Depending on the particular type of array that is used, it is possible to optimize availability or performance, or to select a compromise between both.

The SSA Enhanced Raid adapters only support RAID level 5 (RAID5). RAID0 (Striping) and RAID1 (Mirroring) is not directly supported by the SSA Enhanced Raid adapters, but with the Logical Volume Manager (LVM), RAID0 and RAID1 can be configured on non-RAID disks.

In order to create a RAID5 on SSA Disks, use the command `smitty ssaraid`. This will show you the following menu:



```
                       SSA RAID Arrays

Move cursor to desired item and press Enter.

  List All Defined SSA RAID Arrays
  List All Supported SSA RAID Arrays
  List All SSA RAID Arrays Connected to a RAID Manager
  List Status Of All Defined SSA RAID Arrays
  List/Identify SSA Physical Disks
  List/Delete Old RAID Arrays Recorded in an SSA RAID Manager
  Add an SSA RAID Array
  Delete an SSA RAID Array
  Change/Show Attributes of an SSA RAID Array
  Change Member Disks in an SSA RAID Array
  Change/Show Use of an SSA Physical Disk
  Change Use of Multiple SSA Physical Disks



F1=Help         F2=Refresh        F3=Cancel           F8=Image
F9=Shell        F10=Exit          Enter=Do
```

Select **Add an SSA RAID Array** to do the definitions.

3.3.2 SCSI

The following sections contain important information about SCSI: cabling, connecting RAID subsystems, and adapter SCSI ID and termination change.

3.3.2.1 Cabling

The following sections describe important information about cabling.

SCSI Adapters

A overview of SCSI adapters that can be used on a shared SCSI bus is given in 2.3.2.3, “Supported SCSI Adapters” on page 26. For the necessary adapter changes, see 3.3.2.3, “Adapter SCSI ID and Termination change” on page 77.

RAID Enclosures

The 7135 RAIDiant Array can hold a maximum of 30 single-ended disks in two units (one base and one expansion). It has one controller by default, and another controller can be added for improved performance and availability. Each controller takes up one SCSI ID. The disks sit on internal single-ended buses and hence do not take up IDs on the external bus. In an HACMP cluster, each 7135 should have two controllers, each of which is connected to a separate shared SCSI bus. This configuration protects you against any failure (SCSI adapter, cables, or RAID controller) on either SCSI bus.

Because of cable length restrictions, a maximum of two 7135s on a shared SCSI bus are supported by HACMP.

3.3.2.2 Connecting RAID Subsystems

In this section, we will list the different components required to connect RAID subsystems on a shared bus. We will also show you how to connect these components together.

The 7135-110 RAIDiant Array can be connected to multiple systems on either an 8-bit or a 16-bit SCSI-2 differential bus. The Model 210 can only be connected to a 16-bit SCSI-2 Fast/Wide differential bus, using the Enhanced SCSI-2 Differential Fast/Wide Adapter/A.

To connect a set of 7135-110s to SCSI-2 Differential Controllers on a shared 8-bit SCSI bus, you need the following:

- SCSI-2 Differential Y-Cable

  FC: 2422 (0.765m), PN: 52G7348

- SCSI-2 Differential System-to-System Cable

  FC: 2423 (2.5m), PN: 52G7349

  This cable is used only if there are more than two nodes attached to the same shared bus.

- Differential SCSI Cable (RAID Cable)

  FC: 2901 or 9201 (0.6m), PN: 67G1259 - OR -

FC: 2902 or 9202 (2.4m), PN: 67G1260 - OR -

FC: 2905 or 9205 (4.5m), PN: 67G1261 - OR -

FC: 2912 or 9212 (12m), PN: 67G1262 - OR -

FC: 2914 or 9214 (14m), PN: 67G1263 - OR -

FC: 2918 or 9218 (18m), PN: 67G1264

- Terminator (T)

  Included in FC 2422 (Y-Cable), PN: 52G7350

- Cable Interposer (I)

  FC: 2919, PN: 61G8323

  One of these is required for each connection between an SCSI-2 Differential Y-Cable and a Differential SCSI Cable going to the 7135 unit, as shown in Figure 10.

Figure 10 shows four RS/6000s, each represented by two SCSI-2 Differential Controllers, connected on two 8-bit buses to two 7135-110s, each with two controllers.

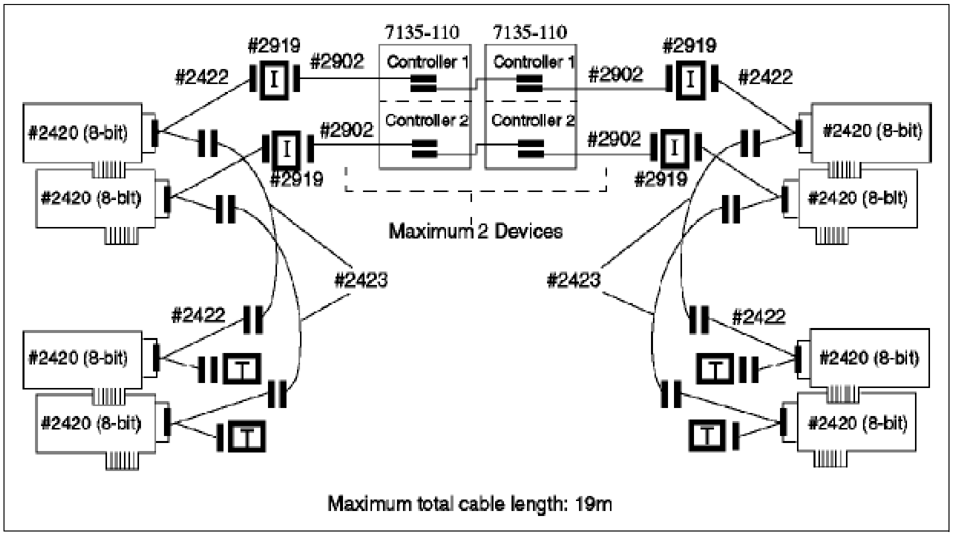


Figure 10. 7135-110 RAIDiant Arrays Connected on Two Shared 8-Bit SCSI Buses

To connect a set of 7135s to SCSI-2 Differential Fast/Wide Adapter/As or Enhanced SCSI-2 Differential Fast/Wide Adapter/As on a shared 16-bit SCSI bus, you need the following:

- 16-Bit SCSI-2 Differential Y-Cable

FC: 2426 (0.94m), PN: 52G4234

- 16-Bit SCSI-2 Differential System-to-System Cable

  FC: 2424 (0.6m), PN: 52G4291 - OR -

  FC: 2425 (2.5m), PN: 52G4233

  This cable is used only if there are more than two nodes attached to the same shared bus.

- 16-Bit Differential SCSI Cable (RAID Cable)

  FC: 2901 or 9201 (0.6m), PN: 67G1259 - OR -

  FC: 2902 or 9202 (2.4m), PN: 67G1260 - OR -

  FC: 2905 or 9205 (4.5m), PN: 67G1261 - OR -

  FC: 2912 or 9212 (12m), PN: 67G1262 - OR -

  FC: 2914 or 9214 (14m), PN: 67G1263 - OR -

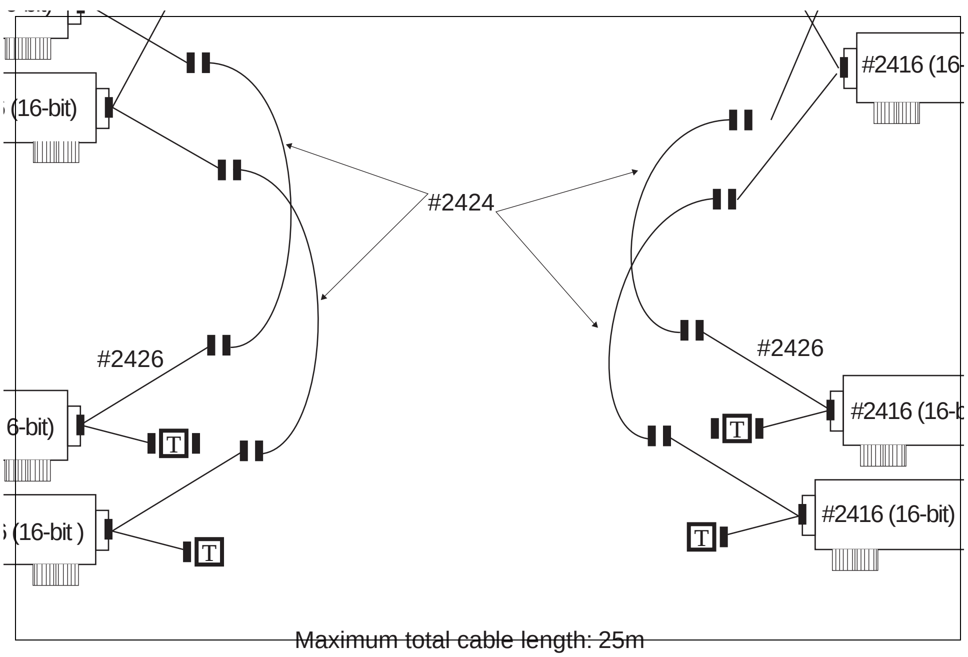
  FC: 2918 or 9218 (18m), PN: 67G1264

- 16-Bit Terminator (T)

  Included in FC 2426 (Y-Cable), PN: 61G8324

Figure 11 shows four RS/6000s, each represented by two SCSI-2 Differential Fast/Wide Adapter/As connected on two 16-bit buses to two 7135-110s, each with two controllers.

The 7135-210 requires the Enhanced SCSI-2 Differential Fast/Wide Adapter/A adapter for connection. Other than that, the cabling is exactly the same as shown in Figure 11, if you just substitute the Enhanced SCSI-2 Differential Fast/Wide Adapter/A (FC: 2412) for the SCSI-2 Differential Fast/Wide Adapter/A (FC: 2416) in the picture.


(16-bit)
#2416 (16-
#2424
#2426
#2426
6-bit)
#2416 (16-b
T
T
T
T
(16-bit)
#2416 (16-bit)
Maximum total cable length: 25m

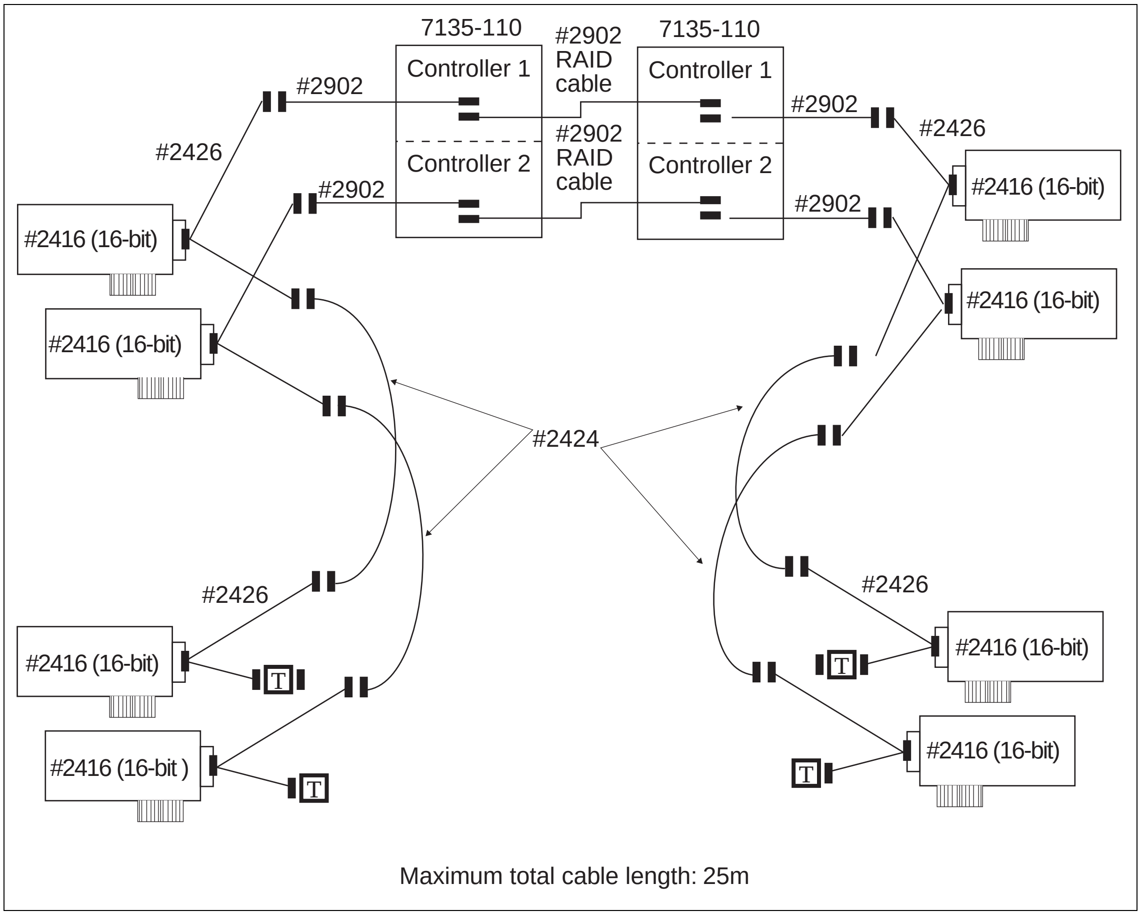


Figure 11. 7135-110 RAIDiant Arrays Connected on Two Shared 16-Bit SCSI Buses

3.3.2.3 Adapter SCSI ID and Termination change

The SCSI-2 Differential Controller is used to connect to 8-bit disk devices on a shared bus. The SCSI-2 Differential Fast/Wide Adapter/A or Enhanced SCSI-2 Differential Fast/Wide Adapter/A is usually used to connect to 16-bit devices but can also be used with 8-bit devices.

In a dual head-of-chain configuration of shared disks, there should be no termination anywhere on the bus except at the extremities. Therefore, you should remove the termination resistor blocks from the SCSI-2 Differential Controller and the SCSI-2 Differential Fast/Wide Adapter/A or Enhanced SCSI-2 Differential Fast/Wide Adapter/A. The positions of these blocks (U8 and U26 on the SCSI-2 Differential Controller, and RN1, RN2 and RN3 on the

SCSI-2 Differential Fast/Wide Adapter/A and Enhanced SCSI-2 Differential Fast/Wide Adapter/A) are shown in Figure 12 and Figure 13 respectively.

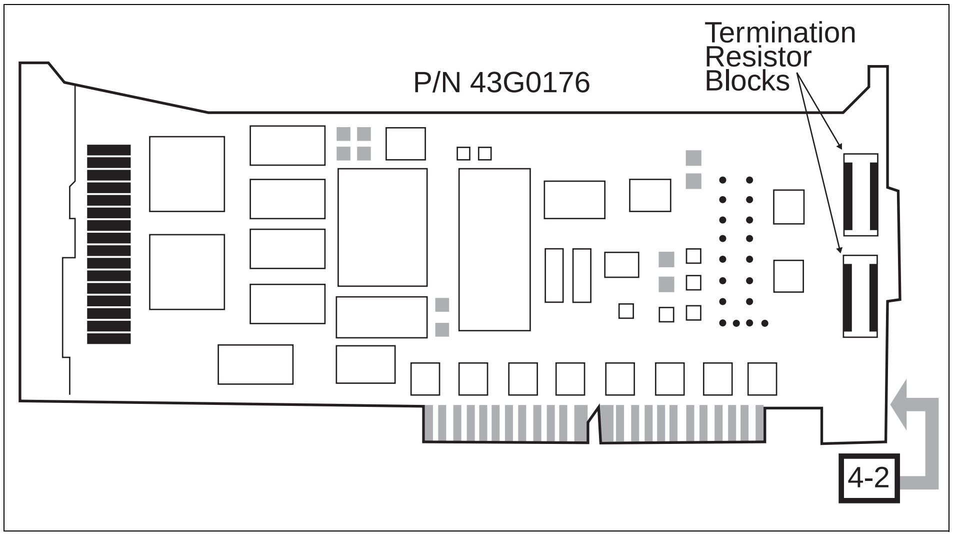


Figure 12. Termination on the SCSI-2 Differential Controller

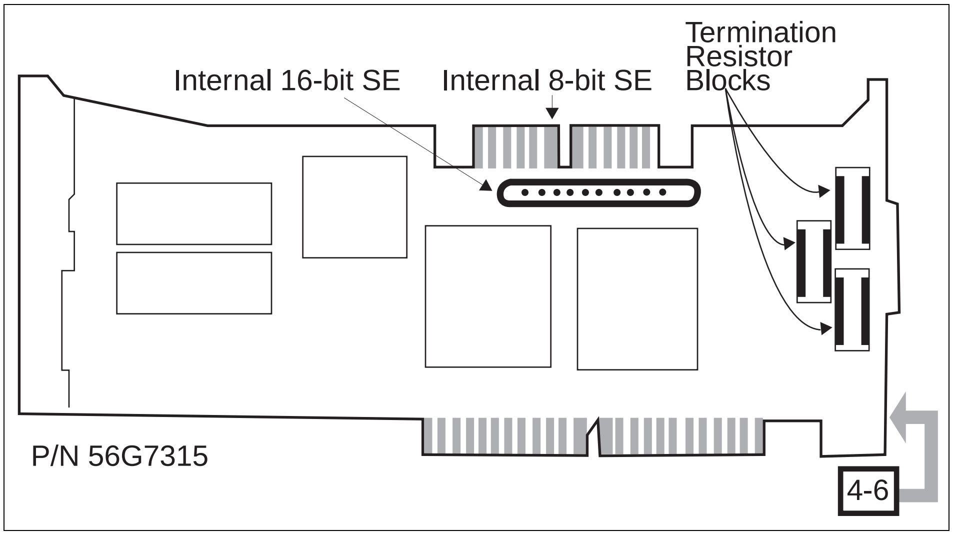
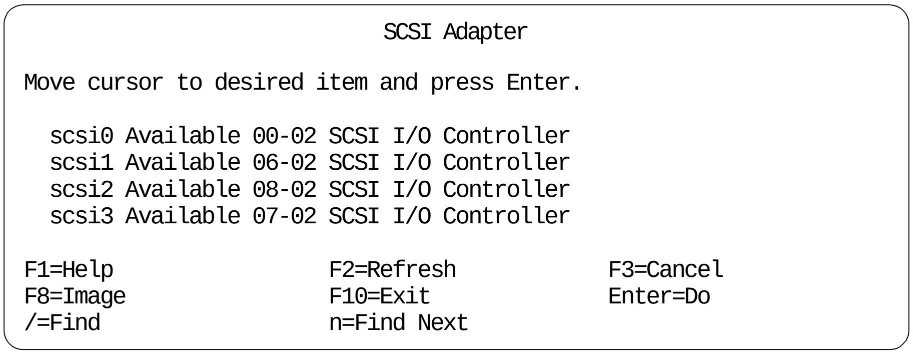


Figure 13. Termination on the SCSI-2 Differential Fast/Wide Adapters

The ID of an SCSI adapter, by default, is 7. Since each device on an SCSI bus must have a unique ID, the ID of at least one of the adapters on a shared SCSI bus has to be changed.

The procedure to change the ID of an SCSI-2 Differential Controller is:

1. At the command prompt, enter `smit chgscsi`.
2. Select the adapter whose ID you want to change from the list presented to you.



```
                              SCSI Adapter

Move cursor to desired item and press Enter.

  scsi0 Available 00-02 SCSI I/O Controller
  scsi1 Available 06-02 SCSI I/O Controller
  scsi2 Available 08-02 SCSI I/O Controller
  scsi3 Available 07-02 SCSI I/O Controller

F1=Help                 F2=Refresh            F3=Cancel
F8=Image                F10=Exit              Enter=Do
/=Find                  n=Find Next
```

3. Enter the new ID (any integer from 0 to 7) for this adapter in the Adapter card SCSI ID field. Since the device with the highest SCSI ID on a bus gets control of the bus, set the adapter’s ID to the highest available ID. Set the Apply change to DATABASE only field to **yes**.

```
 Change / Show Characteristics of a SCSI Adapter

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                   [Entry Fields]
  SCSI Adapter                            scsi0
  Description                             SCSI I/O Controller
  Status                                  Available
  Location                                00-08
  Adapter card SCSI ID                    [6]            +#
  BATTERY backed adapter                   no            +
  DMA bus memory LENGTH                   [0x202000]     +
  Enable TARGET MODE interface            no             +
  Target Mode interface enabled           no
  PERCENTAGE of bus memory DMA area
  for target mode                         [50]           +#
  Name of adapter code download file      /etc/microcode/8d>
  Apply change to DATABASE only           yes            +

F1=Help        F2=Refresh       F3=Cancel       F4=List
F5=Reset       F6=Command       F7=Edit         F8=Image
F9=Shell       F10=Exit         Enter=Do
```

4. Reboot the machine to bring the change into effect.

The same task can be executed from the command line by entering:

```
# chdev -l scsi1 -a id=6 -P
```

Also with this method, a reboot is required to bring the change into effect.

The procedure to change the ID of an SCSI-2 Differential Fast/Wide Adapter/A or Enhanced SCSI-2 Differential Fast/Wide Adapter/A is almost the same as the one described above. Here, the adapter that you choose from the list you get after executing the `smit chgsys` command should be an `ascsi` device. Also, as shown below, you need to change the external SCSI ID only.

```
   Change/Show Characteristics of a SCSI Adapter

SCSI adapter                       ascsi1
Description                        Wide SCSI I/O Control>
Status                             Available
Location                           00-06
Internal SCSI ID                   7                    +#
External SCSI ID                   [6]                  +#
WIDE bus enabled                   yes                  +
...
Apply change to DATABASE only      yes
```

The command line version of this is:

```
# chdev -l ascsi1 -a id=6 -P
```

As in the case of the SCSI-2 Differential Controller, a system reboot is required to bring the change into effect.

The maximum length of the bus, including any internal cabling in disk subsystems, is limited to 19 meters for buses connected to the SCSI-2 Differential Controller, and 25 meters for those connected to the SCSI-2 Differential Fast/Wide Adapter/A or Enhanced SCSI-2 Differential Fast/Wide Adapter/A.

3.4 Shared LVM Component Configuration

This section describes how to define the LVM components shared by cluster nodes in an HACMP for AIX cluster environment.

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

Non-concurrent access environments typically use journaled file systems to manage data, while concurrent access environments use raw logical volumes. This chapter provides different instructions for defining shared LVM components in non-concurrent access and concurrent access environments.

3.4.1 Creating Shared VGs

The following sections contain information about creating non-concurrent VGs and VGs for concurrent access.

3.4.1.1 Creating Non-Concurrent VGs

This section covers how to create a shared volume group on the source node using the SMIT interface. Use the `smit mkvg` fastpath to create a shared volume group. Use the default field values unless your site has other requirements, or unless you are specifically instructed otherwise here.

Table 13. smit mkvg Options (Non-Concurrent)

Options	Description
VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
Activate volume group AUTOMATICALLY at system restart?	Set to **no** so that the volume group can be activated as appropriate by the cluster event scripts.
ACTIVATE volume group after it is created?	Set to **yes.**
Volume Group MAJOR NUMBER	If you are not using NFS, you can use the default (which is the next available number in the valid range). If you are using NFS, you must make sure to use the same major number on all nodes. Use the `lvlstmajor` command on each node to determine a free major number common to all nodes.

3.4.1.2 Creating VGs for Concurrent Access

The procedure used to create a concurrent access volume group varies depending on which type of device you are using: serial disk subsystem (7133) or RAID disk subsystem (7135).

Note

If you are creating (or plan to create) concurrent volume groups on SSA devices, be sure to assign unique non-zero node numbers through the SSAR on each cluster node. If you plan to specify SSA disk fencing in your concurrent resource group, the node numbers are assigned when you synchronize resources. If you do not specify SSA disk fencing, assign node numbers using the following command: `chdev -l ssar -a node_number=x`, where x is the number to assign to that node. You must reboot the system to effect the change.

Creating a Concurrent Access Volume Group on Serial Disk Subsystems

To use a concurrent access volume group, defined on a serial disk subsystem such as an IBM 7133 disk subsystem, you must create it as a *concurrent-capable* volume group. A concurrent-capable volume group can be activated (varied on) in either non-concurrent mode or concurrent access mode. To define logical volumes on a concurrent-capable volume group, it must be varied on in non-concurrent mode.

You can use `smit mkvg` with the following options to build the volume group:

Table 14. smit mkvg Options (Concurrent, Non-RAID)

Options	Description
VOLUME GROUP name	Specify name of volume group.
Physical partition SIZE in megabytes	Accept the default.
PHYSICAL VOLUME NAMES	Specify the names of the physical volumes you want included in the volume group.
Activate volume group AUTOMATICALLY at system restart?	Set this field to **no** so that the volume group can be activated as appropriate by the cluster event scripts.
ACTIVATE volume group after it is created?	Set this field to **no.**
Volume Group MAJOR NUMBER	Accept the default.
Create VG concurrent capable?	Set this field to **yes** so that the volume group can be activated in concurrent access mode by the HACMP for AIX event scripts.
Auto-varyon concurrent mode?	Set this field to **no** so that the volume group can be activated as appropriate by the cluster event scripts.

Creating a Concurrent Access Volume Group on RAID Disk Subsystems

To create a concurrent access volume group on a RAID disk subsystem, such as an IBM 7135 disk subsystem, follow the same procedure as you would to create a non-concurrent access volume group. A concurrent access volume group can be activated (varied on) in either non-concurrent mode or concurrent access mode. To define logical volumes on a concurrent access volume group, it must be varied on in non-concurrent mode.

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

Table 15. smit mkvg Options (Concurrent, RAID)

Options	**Description**
VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
Activate volume group AUTOMATICALLY at system restart?	Set to **no** so that the volume group can be activated as appropriate by the cluster event scripts.
ACTIVATE volume group after it is created?	Set to **yes**.
Volume Group MAJOR NUMBER	While it is only really required when you are using NFS, it is always good practice in an HACMP cluster to have a shared volume group have the same major number on all the nodes that serve it. Use the `lvlstmajor` command on each node to determine a free major number common to all nodes.
Create VG concurrent capable?	Set this field to **no**.

3.4.2 Creating Shared LVs and File Systems

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

Table 16. smit crjfs Options

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

Renaming a jfslog and Logical Volumes on the Source Node

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

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

1. Use the `lsvg -l volume_group_name` command to determine the name of the logical volume and the log logical volume (jfslog) associated with the shared volume groups. In the resulting display, look for the logical volume name that has type jfs. This is the logical volume. Then look for the logical volume name that has type jfslog. This is the log logical volume.
2. Use the `smit chlv` fastpath to rename the logical volume and the log logical volume.
3. After renaming the jfslog or a logical volume, check the /etc/filesystems file to make sure the dev and log attributes reflect the change. Check the log attribute for each file system in the volume group, and make sure that it has the new jfslog name. Check the dev attribute for the logical volume that you renamed, and make sure that it has the new logical volume name.

Adding Copies to Logical Volume on the Source Node

> **Note**
>
> These steps do not apply to RAID devices, which provide their own mirroring of logical volumes.

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

   The copies should reside on separate disks that are controlled by different disk adapters and are located in separate drawers or units, if possible. See Chapter 3.4.3, “Mirroring Strategies” on page 86 for more information.
2. Verify the number of logical volume copies by entering: `lsvg -l volume_group_name.` In the resulting display, locate the line for the logical volume for which you just added copies. Notice that the number in the physical partitions column is *x* times the number in the logical partitions column, where *x* is the number of copies.
3. To verify the placement of logical volume copies, enter: `lspv -l hdiskx`, where `hdiskx` is the name of each disk to which you assigned copies.

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

Testing a File System

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

1. Enter: `fsck /filesystem_name`
2. Verify that you can mount the file system by entering:
   `mount /filesystem_name`
3. Verify that you can unmount the file system by entering:
   `umount /filesystem_name`

3.4.3 Mirroring Strategies

Shared logical volumes residing on non-RAID disk devices should be mirrored in AIX to eliminate the disk as a single point of failure. Shared volume groups residing on a RAID device should not be AIX mirrored; the disk array provides its own data redundancy.

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

3.4.4 Importing to Other Nodes

The following sections cover: varying off a volume group on the source node, importing it onto the destination node, changing its startup status, and varying it off on the destination nodes.

3.4.4.1 Varying Off a Volume Group on the Source Node

After completing the previous tasks, use the `varyoffvg` command to deactivate the shared volume group. You vary off the volume group so that it can be properly imported onto a destination node and activated as appropriate by the cluster event scripts. Enter the following command: `varyoffvg volume_group_name.` Make sure that all the file systems of the volume group have been unmounted, otherwise the `varyoffvg` command will not work.

3.4.4.2 Importing a Volume Group onto the Destination Node

This section covers how to import a volume group onto destination nodes using the SMIT interface. You can also use the TaskGuide utility for this task.

The TaskGuide uses a graphical interface to guide you through the steps of adding nodes to an existing volume group. For more information on the TaskGuide, see 3.4.6, “Alternate Method - TaskGuide” on page 90.

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

You can use the `smit importvg` fastpath to import the volume group.

Table 17. smit importvg Options

Options	Description
VOLUME GROUP name	Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the source node.
PHYSICAL VOLUME name	Enter the name of a physical volume that resides in the volume group. Note that a disk *may have* a different logical name on different nodes. Make sure that you use the disk name as it is defined on the destination node.
ACTIVATE volume group after it is imported?	Set the field to **yes**.
Volume Group MAJOR NUMBER	If you are not using NFS, you may use the default (which is the next available number in the valid range). If you are using NFS, you must make sure to use the same major number on all nodes. Use the `lvlstmajor` command on each node to determine a free major number common to all nodes.

3.4.4.3 Changing a Volume Group’s Startup Status

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

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

Table 18. smit crjfs Options

Options	Description
Activate volume group automatically at system restart?	Set this field to **no**.

Options	Description
A QUORUM of disks required to keep the volume group online?	This field is site-dependent. See 3.4.5, “Quorum” on page 88 for a discussion of quorum in an HACMP cluster.

3.4.4.4 Varying Off the Volume Group on the Destination Nodes

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

3.4.5 Quorum

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

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

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

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

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

3.4.5.1 Quorum at Vary On

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

3.4.5.2 Quorum after Vary On

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

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

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

3.4.5.3 Disabling and Enabling Quorum

Quorum checking is enabled by default. Quorum checking can be disabled using the `chvg -Qn vgname` command, or by using the `smit chvg` fastpath.

Quorum Enabled

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

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

Quorum Disabled

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

Forcing a Varyon

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

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

Quorum in Non-Concurrent Access Configurations

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

Quorum in Concurrent Access Configurations

Quorum must be enabled for an HACMP for AIX concurrent access configuration. Disabling quorum could result in data corruption. Any concurrent access configuration where multiple failures could result in no common shared disk between cluster nodes has the potential for data corruption or inconsistency.

3.4.6 Alternate Method - TaskGuide

The TaskGuide is a 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.

3.4.6.1 TaskGuide Requirements

Before starting the TaskGuide, make sure:

- You have a configured HACMP cluster in place.
- You are on a graphics capable terminal.

3.4.6.2 Starting the TaskGuide

You can start the TaskGuide from the command line by typing: `/usr/sbin/cluster/tguides/bin/cl_ccvg` or you can use the SMIT interface as follows:

1. Type `smit hacmp`.
2. From the SMIT main menu, choose **Cluster System Management -> Cluster Logical Volume Manager ->Taskguide for Creating a Shared Volume Group.** After a pause, the TaskGuide Welcome panel appears.
3. Proceed through the panels to create or share a volume group.

Chapter 4. HACMP Installation and Cluster Definition

This chapter describes issues concerning the actual installation of HACMP Version 4.3 and the definition of a cluster and its resources. It concentrates on the HACMP part of the installation, so, we will assume AIX is already at the 4.3.2 level. Please refer to the *AIX Version 4.3: Migration Guide*, SG24-5116, for details on installation or migration to that level.

This chapter is meant to give an overview of the steps to be taken, and not to be a complete handbook for performing these tasks. When actually performing the HACMP install, the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278 should be consulted.

4.1 Installing HACMP

Before installing, you need to ensure that all the prerequisites are met. Chapter 8 of the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278, gives a detailed list. The AIX Level of the server nodes has to be at AIX 4.3.2, for example, and the required free space in `/usr` must be confirmed. For parts of the product, like HAView, there are prerequisites for other lpps, nv6000 in this case, that have to be ensured.

You can install either from the installation media, from an installation server through Network Installation Management (NIM), or from a hard disk to which the software has been copied.

You will either be installing the HACMP for AIX software for the first time, or upgrading from an earlier version. Both of those situations are discussed in the following sections.

4.1.1 First Time Installs

There are a number of filesets involved in an HACMP Installation. Here is a short overview of them, and what their purpose is.

- cluster.base

  This is the basic component that has to be installed on all server nodes in the cluster, and it contains the following:

  ```
  cluster.base.client.lib          HACMP Base Client Libraries
  cluster.base.client.rte          HACMP Base Client Runtime
  cluster.base.client.utils        HACMP Base Client Utilities
  cluster.base.server.diag         HACMP Base Server Diags
  cluster.base.server.events       HACMP Base Server Events
  cluster.base.server.rte          HACMP Base Server Runtime
  ```

```
cluster.base.server.utils         HACMP Base Server Utilities
```

- cluster.cspoc

  This component includes all of the commands and environment for the C-SPOC utility, the Cluster-Single Point Of Control feature. These routines are responsible for centralized administration of the cluster. There is no restriction on the node from which you run the C-SPOC utility commands, so it should also be installed on all the server nodes. It consists of the following:

```
cluster.cspoc.rte                         HACMP CSPOC Runtime Commands
cluster.cspoc.cmds                        HACMP CSPOC commands
cluster.cspoc.dsh                         HACMP CSPOC dsh and perl
```

- cluster.adt

  This component contains demo clients and their include files, for example, for building a clinfo client on a non-AIX machine. Since these are sample files and demos, you might want to install this on a dedicated machine only. This machine can further be used for development of server or client code:

```
cluster.adt.client.demos          HACMP Client Demos
cluster.adt.client.samples.demos  HACMP Client Demos Samples
cluster.adt.client.samples.clinfo HACMP Client clinfo Samples
cluster.adt.client.samples.clstat HACMP Client clstat Samples
cluster.adt.client.include        HACMP Client includes
cluster.adt.client.samples.libcl  HACMP Client libcl Samples
cluster.adt.server.samples.images HACMP Sample Images
cluster.adt.server.demos          HACMP Server Demos
cluster.adt.server.samples.demos  HACMP Server Sample Demos
```

- cluster.man.en_US.data

  This component contains the man pages in US English. You may like to exchange this with your own language:

```
cluster.man.en_US.cspoc.data      HACMP CSPOC Man pages
cluster.man.en_US.client.data     HACMP Client Man pages
cluster.man.en_US.server.data     HACMP Server Man pages
cluster.man.en_US.haview.data     HACMP HAView Man pages
```

- cluster.msg.en_US

  These filesets contain the messages in US English. In contrast to the man pages, the en_US version must be installed. You might add your language’s messages if you want:

```
cluster.msg.en_US.cspoc           HACMP CSPOC Messages
cluster.msg.en_US.client          HACMP Client Messages
cluster.man.en_US.haview.data     HACMP HAView Messages
```

- cluster.vsm

  The Visual Systems Management Fileset contains Icons and bitmaps for the graphical Management of HACMP Resources, as well as the xhacmpm command:

  ```
  cluster.vsm                        HACMP X11 Dependent
  ```

- cluster.haview

  This fileset contains the files for including HACMP cluster views into a TME 10 Netview Environment. It is installed on a Netview network management machine, and not on a cluster node:

  ```
  cluster.haview                     HACMP HAView
  ```

- cluster.man.en_US.haview.data

  This fileset contains man pages and data for the HAView component:

  ```
  cluster.man.en_US.haview.data      HACMP HAView Manpages
  ```

- cluster.msg.en_US.haview

  This fileset contains the US English messages for the HAView component:

  ```
  cluster.msg.en_US.haview           HACMP HAView Messages
  ```

> **Note**
>
> TME 10 NetView for AIX must be installed on any system where you will install HAView. If NetView is installed using a client/server configuration, HAView should be installed on the NetView client; otherwise, install it on the NetView server node. Also, be aware that the NetView client should not be configured as a cluster node to avoid NetView's failure after a failover.

- cluster.taskguides

  This is the fileset that contains the taskguide for easy creation of shared volume groups:

  ```
  cluster.taskguides.shrvolgrp       HAES Shr Vol Grp Task Guides
  ```

- cluster.clvm

  This fileset contains the Concurrent Resource Manager (CRM) option:

  ```
  cluster.clvm                       HACMP for AIX Concurrent Access
  ```

- cluster.hc

This fileset contains the Application Heart Beat Daemon, Oracle Parallel Server is an application that makes use of it:

```
cluster.hc.rte                    Application Heart Beat Daemon
```

The installation of CRM requires the following software:

```
bos.rte.lvm.usr.4.3.2.0           AIX Run-time Executable
```

Install Server Nodes

From whatever medium you are going to use, install the needed filesets on each node. Refer to Chapter 8 of the *HACMP for AIX, Version 4.3: Installation Guide,* SC23-4278 for details.

Rebooting Servers

The final step in installing the HACMP for AIX software is to reboot each server in your HACMP for AIX environment.

4.1.2 Upgrading From a Previous Version

If you are upgrading your cluster nodes from a previous version, there are some things you have to take care of in order to get your existing cluster back the way you want it after the upgrade is through.

- Ensure that all the prerequisites are met. For details, look into Chapter 8 of the *HACMP for AIX Version 4.3: Installation Guide,* SC23-4278.
- Archive any localized script and configuration files to prevent losing them during an upgrade.
- Its always a good idea to have a mksysb of a working system, so take one of the cluster nodes to be upgraded.
- Commit your current HACMP Version 4.* software (if it is applied but not committed) so that the HACMP 4.3 software can be installed over the existing version.
- Save the current configuration using the cluster snapshot utility, and save any customized event scripts in a directory of your own.
  (To review cluster snapshot instructions, see the chapter on saving and restoring cluster configurations in the *HACMP for AIX, Version 4.3: Administration Guide,* SC23-4279.)

As the version from where you start may differ, there is more than one way to get to the required level.

If your site is currently running an earlier version of the HACMP for AIX software in its cluster environment, except for Version 4.2.2 already running on AIX 4.3, the following procedures describe how to upgrade your existing

HACMP software to HACMP for AIX, Version 4.3. The comments on upgrading the Operating System are not included.

If you are already running AIX 4.3, see the special note at the end of this section.

> **Note**
>
> Although your objective in performing a migration installation is to keep the cluster operational and to preserve essential configuration information, do not run your cluster with mixed versions of the HACMP for AIX software for an extended period of time.

4.1.2.1 Upgrading from Version 4.1.0 through 4.2.2 to Version 4.3

The following procedure applies to upgrading a two-node or multi-node cluster running HACMP Version 4.1.0 through 4.2.2 to Version 4.3 when the installed AIX version is earlier than 4.3.

To perform a rolling AIX migration installation and HACMP upgrade from Version 4.1.0 through Version 4.2.2 to Version 4.3, complete the following steps:

Upgrade AIX on One Node

The following steps describe how to upgrade AIX on one node:

1. If you wish to save your cluster configuration, see the chapter “Saving and Restoring Cluster Configurations” in the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279.
2. Shut down the first node (gracefully with takeover) using the `smit clstop` fastpath. For this example, shut down Node A. Node B will take over Node A's resources and make them available to clients.
   See the chapter “Starting and Stopping Cluster Services” in the *HACMP for AIX, Version 4.3: Administration Guide,* SC23-4279 for more information about stopping cluster services.
3. Perform a **Migration Installation** as described in your *AIX Installation Guide,* SBOF-1803 on Node A.
   The Migration Installation option preserves the current version of the HACMP for AIX software and upgrades the existing base operating system to AIX 4.3.2. Product (application) files and configuration data are also saved.
4. Check the Migration Installation. Verify that all the disks are available. Run `lppchk -v` and `oslevel` to ensure that the system is in a stable state.

Install HACMP 4.3 for AIX on Node A

5. After upgrading AIX and verifying that the disks are correctly configured, install the HACMP 4.3 for AIX software on Node A. For a short description of the filesets, please refer to *4.1.1, “First Time Installs” on page 93* or to Chapter 8 of the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278.

6. The installation process automatically runs the `cl_convert` program. It removes the current HACMP objects from `/etc/objrepos` and saves them to HACMP.old. It creates new HACMP ODM object classes for Version 4.3 in `/etc/objrepos`.

> **Note**
>
> If upgrading from HACMP 4.1 for AIX, you must run the `cl_convert` utility manually. Enter the following command:
>
> `/usr/sbin/cluster/conversion/cl_convert -v 4.1`

7. Start the Version 4.3 software on Node A using the `smit clstart` fastpath. After HACMP is running, start the previous version HACMP software on Node B, if it is not still running. Check to ensure that the nodes successfully join the cluster.

> **Important**
>
> If the node running Version 4.3 fails while the cluster is in this state, the surviving node running the previous version may not successfully mount the file systems that were not properly unmounted due to Node A’s failure.

8. Repeat Steps 2 through 7 on Node B on remaining cluster nodes, one at a time.

> **Important**
>
> In a multi-node cluster, do not synchronize the node configuration or the cluster topology until the last node has been upgraded.

9. When the last node has been upgraded to both AIX 4.3.2 and HACMP 4.3, the cluster install/upgrade process is complete.

Check Upgraded Configuration

10. If using tty devices, check that the tty device is configured as a serial network using the `smit chgtty` fastpath.

11. In order to verify and synchronize the configuration (if desired), you must have /.rhosts files on cluster nodes. If they do not exist, create the /.rhosts

file on Node A using the following command:

```
/usr/sbin/cluster/utilities/cllsif -x >> /.rhosts
```

This command will append information to the /.rhosts file instead of overwriting it. Then, you can ftp this file to the other nodes as necessary.

12. Verify the cluster topology on all nodes using the `clverify` utility.
13. Check that custom event scripts are properly installed.
14. Synchronize the node configuration and the cluster topology from Node A to all nodes (this step is optional).
15. It is recommended that you test the upgraded cluster to ensure proper behavior.

Client-only Migration

If you are migrating from an HACMP 4.1 for AIX through HACMP 4.2 for AIX server node to a client-only node running Version 4.3, first remove the existing server portion of HACMP. If, after upgrading AIX, you install the cluster.base.client.* filesets on a node running an earlier version of HACMP for AIX without de-installing the server, the results are unpredictable.

To determine if there is a mismatch between the HACMP client and server software installed on a node, issue the following command to list the installed software:

```
lslpp -L "cluster*"
```

Examine the list and make sure that all cluster filesets are at 4.3.0.

If you determine that there is a mismatch between the client and server, de-install the server and then repeat the installation of the client software.

In case the node was just a client before, you only have to install the cluster.base.client.* filesets after it has been migrated to AIX Version 4.3.2. Again, please check whether the installation succeeded by issuing the command: `lslpp -L "cluster*"`

4.1.2.2 Upgrading from Version 4.2.2 on AIX 4.3.2 to HACMP Version 4.3

The following steps describe upgrading from Version 4.2.2 on AIX 4.3.2 to HACMP Version 4.3:

1. If your cluster is currently running Version 4.2.2 of the HACMP software on AIX Version 4.3.2, you should upgrade your cluster configuration to HACMP 4.3 for AIX.

2. If you wish to save your cluster configuration, see the chapter Saving and Restoring Cluster Configurations in the *HACMP for AIX, Version 4.3: Administration Guide,* SC23-4279.
3. Commit your current HACMP for AIX software on all nodes.
4. Shut down one node (gracefully with takeover) using the `smit clstop` fastpath. For this example, shut down Node A. Node B will take over Node A’s resources and make them available to clients.
   See the chapter “Starting and Stopping Cluster Services” in the *HACMP for AIX, Version 4.3: Administration Guide,* SC23-4279, for more information on stopping cluster services.
5. Install HACMP for AIX Version 4.3. See Chapter 8 of the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278, starting with the section “Installation Choices”, for instructions.
   The `cl_convert` utility automatically updates the HACMP ODM object classes to the 4.3 version.

> **Note**
>
> If IP address swapping is being used on this node, that is, a boot address is defined for this node, check to ensure that the HACMP changes to `/etc/inittab` and `/etc/rc.net` exist as specified in Appendix A of the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278, before rebooting the node.

6. Reboot Node A.
7. Start the HACMP for AIX software on Node A using the `smit clstart` fastpath and verify that Node A successfully joins the cluster.
8. Repeat Steps 3 through 7 on remaining cluster nodes, one at a time.
9. After all nodes have been upgraded to HACMP Version 4.3, synchronize the node configuration and the cluster topology from Node A to all nodes.
10. Verify the cluster topology on all nodes using the `clverify` utility.
11. Complete a test phase on the cluster before putting it into production.

4.2 Defining Cluster Topology

The cluster topology is comprised of the following components:

- The cluster definition
- The cluster nodes
- The network adapters

- The network modules

You define the cluster topology by entering information about each component into HACMP-specific ODM classes. You enter the HACMP ODM data by using the HACMP SMIT interface or the VSM utility `xhacmpm`. The `xhacmpm` utility is an X Windows tool for creating cluster configurations using icons to represent cluster components. For more information about the `xhacmpm` utility, see the administrative facilities chapter of the *HACMP for AIX, Version 4.3: Concepts and Facilities,* SC23-4276.

Note

The SP Switch network module can support multiple clusters; therefore, its settings should remain at their default values to avoid affecting HACMP event scripts. If you must change these settings, see the chapter on changing the cluster topology in the *HACMP for AIX, Version 4.3: Administration Guide,* SC23-4279 for more information.

4.2.1 Defining the Cluster

The cluster ID and name identifies a cluster in an HACMP environment. The cluster ID and name must be unique for each cluster defined.

Cluster IDs have to be a positive integer in the range from 1 through 99999, and the cluster name is a text string of up to 31 alphanumeric characters, including underscores. It doesn’t necessarily need to match the hostname.

The HACMP software uses this information to create the cluster entries for the ODM.

4.2.2 Defining Nodes

After defining the cluster name and ID, cluster nodes have to be defined. As above, this is usually done through `smit hacmp`. Each of the cluster nodes needs a unique name, so the cluster manager can address them.

Again, a node name is a text string of up to 31 alphanumeric characters that can contain underscores.

You can add more than one node at a time by separating them with whitespace characters.

Note

The node names are logically sorted in their ascii order within HACMP in order to decide which nodes are considered to be neighbors for heartbeat purposes.

In order to build a logical ring, a node always talks to its up- and downstream neighbor in their node name’s ascii order. The uppermost and the lowest node are also considered neighbors.

Adding or Changing a Node Name after the Initial Configuration

If you want to add or change a node name after the initial configuration, use the Change/Show Cluster Node Name screen. See the chapter on changing the cluster topology of the *HACMP for AIX, Version 4.3:Administration Guide, SC23-4279* for more information.

4.2.3 Defining Adapters

To define the adapters after defining the node names, first consult your planning worksheets for both TCP/IP and serial networks.

There are a number of attributes associated with an Adapter in the HACMP configuration which need to be specified:

Adapter IP Label Enter the IP label (the name) of the adapter you have chosen as the service address for this adapter. Adapter labels can be any ASCII text string consisting of alphabetical and numeric characters, underscores, and hyphens, up to 31 characters.

If IP address takeover is defined for that adapter, a boot adapter (address) label has to be defined for it. Use a consistent naming convention for boot adapter labels. (You will choose the **Add an Adapter** option again to define the boot adapter when you finish defining the service adapter.)
You can use hyphens in adapter labels. However, currently it might not be a good idea since the `/usr/sbin/cluster/diag/clverify` utility flags adapter labels that contain hyphens each time it runs.

Network Type Indicate the type of network to which this adapter is connected. Pre-installed network modules are listed on the pop-up pick list.

Network Name Enter an ASCII text string that identifies the network. The network name can include alphabetic and numeric characters and underscores. Use no more than 31 characters. The network name is arbitrary, but must be used consistently for adapters on the same physical network.

If several adapters share the same physical network, make sure you use the same network name for each of these adapters.

Network Attribute Indicate whether the network is **public**, **private**, or **serial**. Press Tab to toggle the values. In the context of HACMP, serial networks means “non-TCP/IP”; public and private networks are TCP/IP networks. Ethernet, Token-Ring, FDDI, and SLIP are public networks. SOCC, ATM, and an SP Switch are private networks. RS232 lines, target mode SSA loops, and target mode SCSI-2 buses are serial networks.

Adapter Function Indicate whether the adapter's function is *service*, *standby*, or *boot*. Press Tab to toggle the values. A node has a single service adapter for each public or private network. A serial network has only a single service adapter.

A node can have none, one, or more standby adapters for each public network. Serial and private networks do not have standby adapters, with the exception of ATM networks. ATM networks must be defined as private, and therefore standby adapters are supported.

In an HACMP environment on the RS/6000 SP, the ethernet adapters can be configured as service adapters but *should not* be configured for IP address takeover. Regarding the SP Switch, network, boot, and service addresses used for IP address takeover are ifconfig alias addresses used on the css0 network. See the appendix entitled HACMP for the RS/6000 SP in the *HACMP Installation Guide* Version 4.3 for more information on adapter functions in an SP Switch environment.

In an ATM network, the adapter function should be listed as svc_s to indicate that the interface is used by HACMP servers. Keep in mind that the netmask for all adapters in an HACMP network must be the same.

Adapter Identifier Enter the IP address in dotted decimal format or a device file name. IP address information is required for non-serial network adapters only if the node's address cannot be obtained from the domain name server or the local /etc/hosts file (using the adapter IP label given). You must enter device filenames for serial network adapters. RS232 serial adapters must have the device filename /dev/ttyN. Target mode SCSI serial adapters must have the device file name /dev/tmscsiN. Target mode SSA adapters must have the device file name /dev/tmssaN.im or /dev/tmssaN.tm**.**

Adapter Hardware Address(optional) Enter a hardware address for the adapter. The hardware address must be unique within the physical network. Enter a value in this field only if: You are currently defining a service adapter, and the adapter has a boot address, and you want to use hardware address swapping. See the chapter on planning TCP/IP networks in the *HACMP for AIX, Version 4.3:Planning Guide*, SC23-4277 for more information on hardware address swapping. This facility is supported only for Ethernet, Token Ring, and FDDI adapters. It does not work with the SP Switch.

Node Name Define a node name for all adapters except for those service adapters whose addresses may be shared by nodes participating in the resource chain for a rotating resource configuration. These adapters are rotating resources. The event scripts use the user-defined configuration to associate these service addresses with the proper node. In all other cases, addresses are associated with a particular node (service, boot, and standby)

Note

Although it is possible to have only one physical network adapter (no standby adapters), this constitutes a potential single point of failure condition and is not recommended for an HACMP configuration.

Note

When IPAT is configured, the run level of the IP-related entries (e. g. rctcpip, rcnfs...) of the `/etc/inittab` are changed to “a”. This has the result that these services are not started at boot time, but with HACMP.

Adding or Changing Adapters after the Initial Configuration

If you want to change the information about an adapter after the initial configuration, use the Change/Show an Adapter screen. See the chapter on changing the cluster topology in the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279, for more information.

4.2.4 Configuring Network Modules

Each supported cluster network in a configured HACMP cluster has a corresponding cluster network module. Each network module monitors all I/O to its cluster network.

The Network Modules are pre-loaded when you install the HACMP software. You do not need to enter information in the Network Module SMIT screens unless you want to change some field associated with a network module, such as the failure detection rate.

Each network module maintains a connection to other network modules in the cluster. The Cluster Managers on cluster nodes send messages to each other through these connections. Each network module is responsible for maintaining a working set of service adapters and for verifying connectivity to cluster peers. The network module is also responsible for reporting when a given link actually fails. It does this by sending and receiving periodic heartbeat messages to or from other network modules in the cluster, and reporting back to the Cluster Manager when it misses a threshold number of heartbeats.

Currently, network modules support communication over the following types of networks:

- Serial (RS232)
- Target-mode SCSI
- Target-mode SSA
- IP (Generic IP)
- Ethernet
- Token-Ring
- FDDI
- SOCC

- SLIP
- SP Switch
- ATM

It is highly unlikely that you will add or remove a network module. For information about changing a characteristic of a Network Module, such as the failure detection rate, see the chapter on changing the cluster topology in the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279. Changing the network module allows the user to influence the rate of heartbeats being sent and received by a Network Module, thereby changing the sensitivity of the detection of a network failure.

In HACMP/ES, topology services and group services are used instead of Network Interface Modules (NIMs) in order to keep track of the status of nodes, adapters or resources.

In HACMP/ES, the tuning of network sensitivity is a little different. Customizable attributes are the interval between heartbeats in seconds and the *Fibrillate Count,* which is the acceptable number of missed heartbeats before some event is triggered. You will find the **Change / Show Topology and Group Services Configuration** in the **Cluster Topology** screen, just like the NIM tuning options.

4.2.5 Synchronizing the Cluster Definition Across Nodes

Synchronization of the cluster topology ensures that the ODM data on all cluster nodes is in sync. The HACMP ODM entries must be the same on each node in the cluster. If the definitions are not synchronized across nodes, the HACMP for AIX software generates a run-time error at cluster startup.

Even if you have a cluster defined with only one node, you must still synchronize the cluster.

The processing performed in synchronization varies depending on whether the cluster manager is active on the local node. If the cluster manager is not active on the local node when you select this option, the ODM data in the system default configuration directory (DCD) on the local node is copied to the ODMs stored in the DCDs on all cluster nodes. The cluster manager is typically not running when you synchronize the initial cluster configuration.

If the cluster manager is active on the local node, the ODM data stored in the DCDs on all cluster nodes are synchronized. In addition, the configuration data stored in the active configuration directory (ACD) on each cluster node is overwritten with the new configuration data, which becomes the new active configuration. If the cluster manager is active on some other cluster nodes but not on the local node, the synchronization operation is aborted.

Before attempting to synchronize a cluster configuration, ensure that all nodes are powered on, that the HACMP software is installed, and that the /etc/hosts and /.rhosts files on all nodes include all HACMP boot and service IP labels.

The /.rhosts file may not be required if you are running HACMP on the SP system.The SP system uses kerberos as its security infrastructure. If you are running HACMP on a node with kerberos enabled (usually an SP node, but could also be a standalone RS/6000 that has been configured with kerberos), you can set a parameter in HACMP to use “Enhanced Security”. This feature removes the requirement of TCP/IP access control lists (for example, the /.rhosts file) on remote nodes during HACMP configuration. Instead, it uses a kerberized version of remote commands to accomplish the synchronization.

It should be noted that Kerberos support is not included in standard AIX 4.3.2. However, Kerberos is available in the public domain, and it is possible to get it and configure it on a non-SP RS/6000 node. This is not very common though, so you will almost always see HACMP Enhanced Security used on the SP system.

When you synchronize the cluster topology, there are two options that control the behavior of this process as follows:

Table 19. *Options for Synchronization of the Cluster Topology*

Ignore Cluster Verification Errors	If you specify **yes**, the result of the cluster verification performed as part of synchronization is ignored and the configuration is synchronized even if verification fails. If you specify **no**, the changes are not synchronized if verification fails. View the error messages in the system error log to determine the configuration problem. For information about the `/usr/sbin/cluster/diag/clverify` utility, see the chapter on verifying a cluster configuration in the *HACMP for AIX, Version 4.3: Administration Guide, SC23-4279*.
Emulate or Actual	If you set this field to **Emulate**, the synchronization will be an emulation and will not affect the Cluster Manager. If you set this field to **Actual**, the synchronization will actually occur, and any subsequent changes will be made to the Cluster Manager. **Emulate** is the default value.

The cluster topology definition (including all node, adapter, and network module information) is copied to the other nodes in the cluster.

4.3 Defining Resources

The HACMP for AIX software provides a highly available environment by identifying a set of cluster-wide resources essential to uninterrupted processing, and then by defining relationships among nodes that ensure these resources are available to client processes. Resources include the following hardware and software:

- Disks
- Volume groups
- File systems
- Network addresses
- Application servers

In the HACMP for AIX software, you define each resource as part of a resource group. This allows you to combine related resources into a single logical entity for easier configuration and management. You then configure each resource group to have a particular kind of relationship with a set of nodes. Depending on this relationship, resources can be defined as one of three types: cascading, concurrent access, or rotating. See 2.4.1, “Resource Group Options” on page 28 for details.

After configuring the cluster topology, you must configure resources and set up the cluster node. This involves:

- Configuring resource groups and node relationships to behave as desired
- Adding individual resources to each resource group
- Setting up run-time parameters for each node
- Synchronizing cluster nodes

4.3.1 Configuring Resource Groups

Resource Groups are initialized by telling the HACMP ODM their names, the participating nodes, and their relationship. The order in which the participating nodes are defined is taken as the priority of the resource chain, that is, priority is decreasing from left to right.

The relationship can be one of Cascading, Rotating or Concurrent. See 2.4.1, “Resource Group Options” on page 28 for details.

4.3.1.1 Configuring Resources for Resource Groups

Once you have defined resource groups, you further configure them by assigning cluster resources to one resource group or another. You can configure resource groups even if a node is powered down. However, SMIT cannot list possible shared resources for the node (making configuration errors likely).

Note

You cannot configure a resource group until you have completed the information on the **Add a Resource Group** screen.

Note

If you configure a cascading resource group with an NFS mount point, you must also configure the resource to use IP Address Takeover. If you do not do this, takeover results are unpredictable.You should also set the field value **Filesystems Mounted Before IP Configured** to **true** so that the takeover process proceeds correctly.

Note

When setting up a cascading resource with an IP Address takeover configuration, each cluster node should be configured in no more than (N+1) resource groups on a particular network. Here, N is the number of standby adapters on a particular node and network.

The following describes the different possibilities for resources that might be added to a resource group:

Table 20. Options Configuring Resources for a Resource Group

Service IP Label	If IP address takeover is being used, list the IP label to be moved when this resource group is taken over. Press F4 to see a list of valid IP labels. These include addresses which rotate or may be taken over.
HTY Service IP Label	NTX adapters are not supported by HACMP for AIX 4.3
File Systems	Identify the file systems to include in this resource group. Press F4 to see a list of the file systems. When you enter a file system in this field, the HACMP for AIX software determines the correct values for the Volume Groups and Raw Disk PVIDs fields.

Service IP Label	If IP address takeover is being used, list the IP label to be moved when this resource group is taken over. Press F4 to see a list of valid IP labels. These include addresses which rotate or may be taken over.
File Systems Consistency Check	Identify the method for checking consistency of file systems, `fsck` (default) or `logredo` (for fast recovery).
File Systems Recovery Method	Identify the recovery method for the file systems, **parallel** (for fast recovery) or **sequential** (default). Do *not* set this field to **parallel** if you have shared, nested file systems. These must be recovered sequentially. (Note that the cluster verification utility, **clverify**, does not report file system and fast recovery inconsistencies.)
File Systems to Export	Identify the file systems to be exported to include in this resource group. These should be a subset of the file systems listed above. Press F4 for a list.
File Systems to NFS Mount	Identify the subset of file systems to NFS mount. All nodes in the resource chain that do not currently hold the resource will attempt to NFS mount these file systems while the owner node is active in the cluster.

These settings also have to be synchronized throughout the cluster. Therefore **Synchronize Cluster Resources** has to be chosen from the corresponding SMIT Menu.

If the Cluster Manager is running on the local node, synchronizing cluster resources triggers a dynamic reconfiguration event (DARE, see 8.5.3, “DARE Resource Migration Utility” on page 169).

4.3.1.2 Configuring Run-Time Parameters

There are two types of Run-Time Parameters for a node that can be chosen. One of them is the debug level, which can be switched from **high** to **low,** meaning all cluster manager actions are logged, or only errors are logged, respectively. The other is the differentiation whether the node uses NIS or DNS nameservice or not, to enable the cluster manager to turn that off in case it would interfere with its actions. Both of these parameters can be changed while the cluster is running.

4.3.1.3 Defining Application Servers

Application servers are another resource that can be configured into a Resource Group. They consist of a (hopefully meaningful) name, in order to enable the cluster manager to identify the application server uniquely, as well as the path locations for start and stop scripts for the application. These scripts have to be in the same location on every service node.

Just as for pre- and post-events, these scripts can be adapted to specific nodes. They don’t need to be equal in content. The system administrator has to ensure, however, that they are in the same location, use the same name, and are executable for the root user.

4.3.1.4 Synchronizing Cluster Resources

After defining these resources and their relationship with the resource group, the act of synchronizing cluster resources sends the information contained on the current node to all defined cluster nodes.

> **Note**
>
> All configured nodes must be on their boot addresses when a cluster has been configured and the nodes are synchronized for the first time. Any node not on its boot address will not have its /etc/rc.net file updated with the HACMP entry; this causes problems for the reintegration of this node into the cluster.

If a node attempts to join the cluster when its configuration is out-of-sync with other active cluster nodes, it will be denied. You must ensure that other nodes are synchronized to the joining member.

4.4 Initial Testing

After installing and configuring your cluster, it is recommended that you do some initial testing in order to verify that the cluster is acting as it should.

4.4.1 Clverify

Running `/usr/sbin/cluster/diag/clverify` is probably a good start to the testing. It allows you to check the software and the cluster.

Software checking is reduced to lpp checking, which is basically checking whether HACMP-specific modifications to AIX files are correct. For correctness of the installation itself, use the `lppcheck -v` command.

Cluster verification is divided into topology and configuration checking. These two parts do basically the same as `smit clverify`, i.e. verifying that the clusters topology as well as the resource configurations are in sync on the cluster nodes.

4.4.2 Initial Startup

At this point in time, the cluster is not yet started. So the cluster manager has to be started first. To check whether the cluster manager is up, you can either look for the process with the `ps` command:

```
ps -ef | grep clstr
```

or look for the status of the cluster group subsystems:

```
lssrc -g cluster
```

or look for the status of the network interfaces. If you have IP Address Takeover (IPAT) configured you should see that the network interface is on its boot address with the `netstat -i` command.

Then start HACMP through `smit clstart.` In the panel that appears, choose the following parameters and press Enter:

1. start **now**
2. broadcast message **true**
3. start cluster lock services **false**
4. start cluster information daemon **true**

Reissue either the `ps` command (see above) or look for the interface state with the `netstat -i` command. Now, you should see that the boot interface is gone in favor of the service-interface.

You also would like to check whether a takeover will work, so, you have to bring up HACMP on all cluster nodes through `smitty clstart` and check whether the cluster gets into a stable state. Use `clstat` for this purpose.

4.4.3 Takeover and Reintegration

When the cluster is up and running, stop one of the node’s cluster managers with `smitty clstop` and choose **graceful with takeover**. One possibility to check whether the takeover went through smoothly is to look at the /tmp/hacmp.out file during the takeover, preferably on the takeover node. You can use the `tail -f /tmp/hacmp.out` command for this.

After the cluster has become stable, you might check the `netstat -i` output again to verify that the takeover node has acquired the IP address of the “failed” node.

For cascading resource groups the failed node is going to reaquire its resources, once it is up and running again. So, you have to restart HACMP on it through `smitty clstart` and check again for the logfile, as well as the clusters status.

Further and more intensive debugging issues are covered in Chapter 7, “Cluster Troubleshooting” on page 143.

4.5 Cluster Snapshot

Now that the actual installation is finished, the cluster is well documented in the planning sheets, all information from there has been implemented in the HACMP ODM, and the cluster is verified and synchronized; provided the initial testing didn’t bring up any curiosities, you should save this working configuration in a cluster snapshot.

The cluster snapshot utility allows you to save, in a file, a record of all the data that defines a particular cluster configuration. This facility gives you the ability to recreate a particular cluster configuration, a process called applying a snapshot, provided the cluster is configured with the requisite hardware and software to support the configuration.

You can perform many of the cluster snapshot utility operations, such as saving a configuration and applying a saved configuration, using the HACMP for AIX VSM application (xhacmpm). For more information, see the administrative facilities chapter in the *HACMP for AIX, Version 4.3: Concepts and Facilities*, SC23-4276, and the online help information available with the application.

In addition, a snapshot can provide useful information for troubleshooting cluster problems. Because the snapshots are simple ASCII files that can be sent via e-mail, they can make remote problem determination easier.

You can also add your own custom snapshot methods to store additional user-specified cluster and system information in your snapshots. The output from these user-defined custom methods is reported along with the conventional snapshot information.

Note

You cannot use the cluster snapshot facility in a cluster with nodes concurrently running different versions of HACMP for AIX.

Essentially, a snapshot saves all the ODM classes HACMP has generated during its configuration. It does not save user customized scripts, such as start or stop scripts for an application server. However, the location and names of these scripts are in an HACMP ODM class, and are therefore saved. It is very helpful to put all the customized data in one defined place, in order to make saving these customizations easier. You can then use a custom snapshot method to save this data as well, by including a user-defined script in the custom snapshot.

4.5.1 Applying a Cluster Snapshot

Applying a cluster snapshot overwrites the data in the existing HACMP for AIX ODM classes on all nodes in the cluster with the new ODM data contained in the snapshot. You can apply a cluster snapshot from any cluster node. However, you have to differentiate between two possible states the cluster could be in when applying the snapshot.

If cluster services are inactive on all cluster nodes, applying the snapshot changes the ODM data stored in the system default configuration directory (DCD). If cluster services are active on the local node, applying a snapshot triggers a cluster-wide dynamic reconfiguration event. In dynamic reconfiguration, in addition to synchronizing the ODM data stored in the DCDs on each node, HACMP for AIX replaces the current configuration data stored in the active configuration directory (ACD) with the changed configuration data in the DCD. The snapshot becomes the currently active configuration.

Note

A cluster snapshot used for dynamic reconfiguration may contain changes to either the cluster topology OR to cluster resources, but not both. You cannot change both the cluster topology and cluster resources in a single dynamic reconfiguration event.

Note

Applying a cluster snapshot may affect both AIX and HACMP for AIX ODM objects and system files as well as user-defined files.

More detailed Information about Cluster Snapshot can be found in the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279, Chapter 11 as well as in the *HACMP for AIX, Version 4.3: Troubleshooting Guide*, SC23-4280.

Chapter 5. Cluster Customization

Within an HACMP for AIX cluster, there are several things that are customizable. The following paragraphs explain the customizing features for events, error notification, network modules and topology services.

5.1 Event Customization

An HACMP for AIX cluster environment acts upon a state change with a set of predefined cluster events (see 5.1.1, “Predefined Cluster Events” on page 117). Whenever a state change is detected by the cluster manager, it decides which event will be started. It then executes the script for that event in a shell, as well as the subevents associated with it. These predefined events can be found under `/usr/sbin/cluster/events`.

The HACMP for AIX software provides an event customization facility that allows you to tailor event processing to your site. This facility can be used to include the following types of customization:

- Adding, changing, and removing custom cluster events
- Pre- and post-event processing
- Event notification
- Event recovery and retry

5.1.1 Predefined Cluster Events

HACMP has the following predefined cluster events:

5.1.1.1 Node Events

This is the sequence of node_up events:

node_up This event occurs when a node joins the cluster. Depending on whether the node is local or remote, this event initiates either a node_up_local or node_up_remote event.

node_up_local This script acquires the service address (or shared address), gets all its owned (or shared) resources, and takes the resources. This includes making disks available, varying on volume groups, mounting file systems, exporting file systems, NFS-mounting file systems, and varying on concurrent access volumes groups.

acquire_service_addr (If configured for IP address takeover.) Configures boot addresses to the corresponding service address, and starts TCP/IP servers and network daemons by running the `telinit -a` command.

acquire_takeover_addr The script checks to see if a configured standby address exists, then swaps the standby address with the takeover address.

get_disk_vg_fs Acquires disk, volume group, and file system resources.

node_up_remote Causes the local node to release all resources taken from the remote node and to place any concurrent volume groups in concurrent mode. Some of the scripts called by node_up_remote include the following:

release_takeover_addr (If configured for IP address takeover.) Identifies a takeover address to be released because a standby adapter on the local node is masquerading as the service address of the remote node. Reconfigures the local standby adapter to its original address (and hardware address, if necessary).

stop_server Stops application servers belonging to the reintegrating node.

release_vg_fs Releases volume groups and file systems belonging to a resource group that the remote node will be taking over.

cl_deactivate_nfs Unmounts NFS file systems.

node_up_complete This event occurs only after a node_up event has successfully completed. Depending on whether the node is local or remote, this event initiates either a node_up_local_complete or node_up_remote_complete event.

node_up_local_complete Calls the start_server script to start application servers. This event occurs only after a node_up_local event has successfully completed.

node_up_remote_complete Allows the local node to do an NFS mount only after the remote node is completely up. This event occurs only after a node_up_remote event has successfully completed.

Sequence of node_down Events

node_down This event occurs when a node intentionally leaves the cluster or fails. Depending on whether the exiting node is local or remote, this event initiates either the node_down_local or node_down_remote event, which in turn initiates a series of subevents.

node_down_local Processes the following events:

stop_server Stops application servers.

release_takeover_addr (If configured for IP address takeover.) Identifies a takeover address to be released because a standby adapter on the local node is masquerading as the service address of the remote node. Reconfigures the local standby with its original IP address (and hardware address, if necessary).

release_vg_fs Releases volume groups and file systems that are part of a resource group the local node is serving.

release_service_addr (If configured for IP address takeover.) Detaches the service address and reconfigures the service adapter to its boot address.

node_down_remote Processes the following events:

acquire_takeover_addr (If configured for IP address takeover.) Checks for a configured standby address currently seen as up by the Cluster Manager, and then does a standby_address to takeover_address swap (and hardware address, if necessary.

get_disk_vg_fs Acquires disk, volume group, and file system resources as part of a takeover.

node_down_complete This event occurs only after a node_down event has successfully completed. Depending on whether the node is local or remote, this event initiates either a node_down_local_complete or node_down_remote_complete event.

node_down_local_complete Instructs the Cluster Manager to exit when the local node has left the cluster. This event occurs only after a node_down_local event has successfully completed.

node_down_remote_complete Starts takeover application servers. This event runs only after a node_down_remote event has successfully completed.

start_server Starts application servers.

5.1.1.2 Network Events

network_down This event occurs when the Cluster Manager determines a network has failed. A network_down event can take one of two forms:

Local network_down, where only a particular node has lost contact with a network.

Global network_down, where all of the nodes connected to a network have lost contact with a network. It is assumed in this case that a network-related failure has occurred rather than a node-related failure.

The network_down event mails a notification to the system administrator, but takes no further action since appropriate actions depend on the local network configuration.

network_down_complete This event occurs only after a network_down event has successfully completed. The default network_down_complete event processing takes no actions since appropriate actions depend on the local network configuration.

network_up This event occurs when the Cluster Manager determines a network has become available for use. The default network_up event processing takes no actions since appropriate actions depend on the local network configuration.

network_up_complete This event occurs only after a network_up event has successfully completed. The default network_up_complete event processing takes no actions since appropriate actions depend on the local network configuration.

5.1.1.3 Network Adapter Events

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

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

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

join_standby This event occurs if a standby adapter becomes available. The join_standby event displays a console message indicating that a standby adapter has become available.

5.1.1.4 Cluster Status Events

config_too_long This event occurs when a node has been in reconfiguration for more than six minutes. The event periodically displays a console message.

reconfig_topology_start This event marks the beginning of a dynamic reconfiguration of the cluster topology.

reconfig_topology_complete This event indicates that a cluster topology dynamic reconfiguration has completed.

reconfig_resource_acquire This event indicates that cluster resources that are affected by dynamic reconfiguration are being acquired by appropriate nodes.

reconfig_resource_release This event indicates that cluster resources affected by dynamic reconfiguration are being released by appropriate nodes.

reconfig_resource_complete This event indicates that a cluster resource dynamic reconfiguration has completed.

5.1.2 Pre- and Post-Event Processing

To tailor event processing to your environment, specify commands or user-defined scripts that should execute before and/or after a specific event is generated by the Cluster Manager. You specify them by selecting the HACMP event to be customized on the **smit hacmp -> Cluster Configuration -> Resources -> Cluster Events -> Change/Show Cluster Events** screen, and then, choosing the one to be tailored. Now, you can enter the location of your pre- or post-event to be executed before or after the chosen event has been processed.

For preprocessing, 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.

5.1.3 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.

This is done on the very same SMIT screen as in 5.1.2, “Pre- and Post-Event Processing” on page 122 in the `Notify` Command field.

For example, a site may want to use a `network_down` notification event to inform system administrators that traffic may have to be rerouted. Afterwards, you can use a `network_up` notification event to tell system administrators that traffic can again be serviced through the restored network.

Event notification in an HACMP cluster can also be done using pre- and post-event scripts, just by adding the script you want to execute for notification into the `pre-` and/or `post-event` command script.

5.1.4 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.

For example, a file system cannot be unmounted, because of a process running on it. Then, you might want to kill that process first, before unmounting the file system, in order to get the event script done. Now, since the event script didn't succeed in its first run, the *Retry* feature enables HACMP for AIX to retry it until it finally succeeds, or the retry count is reached.

5.1.5 Notes on Customizing Event Processing

You must declare a shell (for example `#!/bin/sh`) at the beginning of each script executed by the notify, recovery, and pre- or post-event processing commands.

Notify, recovery, and pre- and post-event processing do not occur when the force option of the `node_down` event is specified.

Synchronizing the cluster configuration does not propagate the actual new or changed scripts; you must add these to each node manually. Also, it is allowed to have different contents in these scripts on different nodes in order to be able to act upon different environments. However, the name of these scripts, their location in the file system, and their permission bits have to be identical.

5.1.6 Event Emulator

To test the effect of running an event on your cluster, HACMP for AIX provides a utility to run an emulation of an event. This emulation lets you predict a cluster's reaction to an event as though the event actually occurred. The emulation runs on all active nodes in your cluster, and the output is stored in an output file. You can select the path and name of this output file using the `EMU_OUTPUT` environment variable, or, use the default /tmp/emuhacmp.out file on the node that invoked the Event Emulator.

For more information on event emulation, see these chapters: “Administrative Facilities” in the *HACMP for AIX, Version 4.3: Concepts and Facilities,* SC23-4276, and “Monitoring an HACMP Cluster” in the *HACMP for AIX, Version 4.3: Administration Guide,* SC23-4279.

5.2 Error Notification

The AIX Error Notification facility detects errors matching predefined selection criteria and responds in a programmed way. The facility provides a wide range of criteria that you can use to define an error condition. These errors are called *notification objects.*

Each time an error is logged in the system error log, the error notification daemon determines if the error log entry matches the selection criteria. If it does, an executable is run. This executable, called a *notify method*, can range from a simple command to a complex program. For example, the notify method might be a mail message to the system administrator or a command to shut down the cluster.

Using the Error Notification facility adds an additional layer of high availability to the HACMP for AIX software. Although the combination of the HACMP for AIX software and the inherent high availability features built into the AIX operating system keeps single points of failure to a minimum, failures still exist that, although detected, are not handled in a useful way.

Take the example of a cluster where an owner node and a takeover node share an SCSI disk. The owner node is using the disk. If the SCSI adapter on the owner node fails, an error may be logged, but neither the HACMP for AIX software nor the AIX Logical Volume Manager responds to the error. If the error has been defined to the Error Notification facility, however, an executable that shuts down the node with the failed adapter could be run, allowing the surviving node to take over the disk.

5.3 Network Modules/Topology Services and Group Services

The HACMP for AIX SMIT interface allows you to add, remove, or change an HACMP for AIX network module. You rarely need to add or remove any of those, however, you may want to change the failure detection rate of a network module.

There are three values to choose from: *Fast, Normal* and *Slow.* The normal heartbeat rate is usually optimal. Speeding up or slowing down failure detection is an area where you can adjust cluster failover behavior.

If you decide to change the failure detection rate of a network module, keep the following considerations in mind:

- Failure detection is dependent on the fastest network linking two nodes.
- Faster heartbeat rates may lead to false failure detections, particularly on busy networks. For example, bursts of high network traffic may delay heartbeats and this may result in nodes being falsely ejected from the cluster. Faster heartbeat rates also place a greater load on networks.
- If your networks are very busy and you experience false failure detections, you can try changing the failure detection speed on the network modules to slow to avoid this problem.

The failure rate of networks varies, depending on their characteristics. For example, for an Ethernet, the normal failure detection rate is two keepalives per second; fast is about four per second; slow is about one per second. For an HPS network, because no network traffic is allowed when a node joins the cluster, normal failure detection is 30 seconds; fast is 10 seconds; slow is 60 seconds.

The **Change / Show Topology and Group Services Configuration** screen includes the settings for the length of the Topology and Group services logs. The default settings are highly recommended. The screen also contains entries for heartbeat settings, but these are not operable (see *HACMP/ES Installation and Administration Guide*, SC23-4284, Chapter 18). The heartbeat rate is now set for each network module in the corresponding screen (see above).

To learn more about Topology and Group Services, see Chapter 32 of the *HACMP/ES Installation and Administration Guide*, SC23-4284.

5.4 NFS considerations

For NFS to work correctly in an HACMP cluster environment, you have to take care of some special NFS characteristics.

The HACMP scripts have only minimal NFS support. You may need to modify them to handle your particular configuration. The following sections contain some suggestions for handling a variety of issues.

5.4.1 Creating Shared Volume Groups

When creating shared volume groups, normally, you can leave the Major Number field blank and let the system provide a default for you. However, unless all nodes in your cluster are identically configured, you will have problems using NFS in an HACMP environment. The reason is that the system uses the major number as part of the file handle to uniquely identify a Network File System.

In the event of node failure, NFS clients attached to an HACMP cluster operate exactly the way they do when a standard NFS server fails and reboots. If the major numbers are not the same, when another cluster node takes over the file system and re-exports it, the client application will not recover, since the file system exported by the node will appear to be different from the one exported by the failed node.

To prevent problems with NFS file systems in an HACMP cluster, make sure that each shared volume group has the same major number on all nodes. The `lvlstmajor` command lists the free major numbers on a node. Use this command on each node to find a major number that is free on all cluster nodes, then, record that number in the Major Number field on the *Shared Volume Group/File System (Non-Concurrent Access)* worksheet in Appendix A, Planning Worksheets, of the *HACMP for AIX, Version 4.3: Planning Guide, SC23-4277* for a non-concurrent access configuration.

Alternatively, if you use the Task Guide to create your shared volume groups, it will make sure that the major number is the same on all nodes that will share it.

5.4.2 Exporting NFS File Systems

The default scripts provided with HACMP do not use the /etc/exports file. Instead, the default scripts provided call a `cl_export_fs` utility that uses the `exportfs` command with the `-i` flag and specifies the file system names stored in the HACMP ODM object class.

Therefore export options specified in the /etc/exports file are ignored. However, export options may be specified by modifying the `cl_export_fs` utility. Alternately, the /etc/exports file can be used as is typical in an NFS environment by simply removing the `-i` flag from the `exportfs` command in the cl_export_fs utility.

5.4.3 NFS Mounting

For HACMP and NFS to work together properly, you must be aware of the following mount issues:

5.4.3.1 Creating NFS Mount Points on Clients

A mount point is required in order to mount a file system with NFS. Mount points are required for NFS clients, not servers; however, you should be aware that a server can also be a client.

5.4.4 Cascading Takeover with Cross Mounted NFS File Systems

This section describes how to set up cascading resource groups with cross mounted NFS file systems.

5.4.4.1 Server-to-Server NFS Cross Mounting

HACMP allows you to configure a cluster so that servers can NFS-mount each other’s file systems. The following figure shows an example:

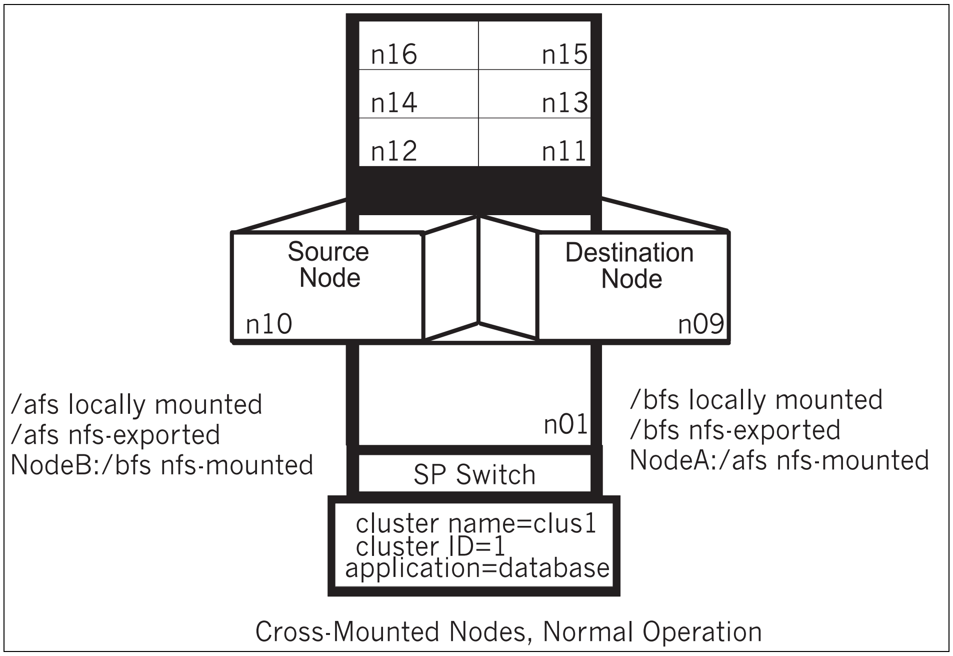


Figure 14. NFS Cross Mounts

When Node A fails, Node B uses the `cl_nfskill` utility to close open files in Node A:/afs, unmounts it, mounts it locally, and re-exports it to waiting clients.

After takeover, Node B has:

```
/bfs locally mounted
/bfs nfs-exported
/afs locally mounted
/afs nfs-exported
```

Ensure that the shared volume groups have the same major number on the server nodes. This allows the clients to re-establish the NFS-mount transparently after the takeover.

Caveats about Node Names and NFS

In the configuration described above, the node name is used as the NFS hostname for the mount. This can fail if the node name is not a legitimate TCP/IP adapter label.

To avoid this problem do one of the following:

- Ensure that node name and the service adapter label are the same on each node in the cluster
  or
- Alias the node name to the service adapter label in the /etc/hosts file.

5.4.5 Cross Mounted NFS File Systems and the Network Lock Manager

If an NFS client application uses the Network Lock Manager, there are additional considerations to ensure a successful failover. Consider the following scenario: Node A has a file system mounted locally and exported for use by clients. Node B is an NFS client and mounts the exported file system for local use by an application that issues lock requests using the `flock()` system call. Node A fails. Node B then attempts to unmount the NFS mounted file system, mount it as a local file system, and export it for client use. However, the unmount fails because of outstanding lock requests against the file system.

Adding the following lines to the cl_deactivate_nfs script will clear outstanding locks against the failed node and will allow the file system to be unmounted. However, it will result in the loss of all locks. Consider your configuration carefully. If you have non-cluster related NFS file systems where losing locks would be unacceptable, you may need to take appropriate steps before using this addition to the cl_deactivate_nfs script.

Add the code below between the following two lines (three places):

```
######### Add for NFS Lock Removal (start) #########
######### Add for NFS Lock Removal (finish) #########
################################################################################
#
#  Name:  cl_deactivate_nfs
#
#  Given a list of nfs-mounted filesystems, we try and unmount -f
#  any that are currently mounted.
#
#  Arguments: list of filesystems.
#
################################################################################
PROGNAME="$0"
MOUNTED="false"
######### Add for NFS Lock Removal (start) #########
STOPPED="false"
######### Add for NFS Lock Removal (finish) #########
SLEEP="2"
if [ "$VERBOSE_LOGGING" = "high" ]
then
   set -x
fi
set -u
if [ $# -ne 0 ]
then
  FILELIST=‘for i in $*; do /bin/echo $i; done | /bin/sort -r‘
  for fs in $FILELIST
  do
      # Is the filesystem mounted?
      # -s says only return status, -x says exact match
      # we use awk instead of cut because mount outputs
      # lots of leading blanks that confuse cut
    /etc/mount | awk ’{ print $2 }’ | fgrep -s -x "$fs"
    if [ $? -eq 0 ]
    then
      # At least one filesystem is mounted
      MOUNTED="true"
       # This filesystem is mounted
######### Add for NFS Lock Removal (start) #########
       # Determine the host which is making the filesystem
        # available
       # This will clear any outstanding locks against the
        # failed node, not preserve their state, and is thus
        # considered a forceful move.
      host=‘/etc/mount|grep nfs|grep "$fs"|awk ’{ print $1 }’‘
      if [ -n "$host" ]
      then
       if [ "$STOPPED" = "false" ]
       then
         stopsrc -s rpc.lockd
         stopsrc -s rpc.statd
         STOPPED="true"
       fi
       /bin/rm -f /etc/sm.bak/$host
       /bin/rm -f /etc/sm/$host
       /bin/rm -f /etc/state
      fi
######### Add for NFS Lock Removal (finish) #########
        # Send a SIGKILL to all processes having open file
        # descriptors within this logical volume to allow
        # the unmount to succeed..
      cl_nfskill -k -u $fs
    fi
  done
else
  cl_echo 27 ’$PROGNAME:  Bad number of arguments ’ $PROGNAME
  exit 2
fi
# Make sure all processes have time to die
# Only wait if at least one filesystem is mounted
if [ "$MOUNTED" = "true" ]
then
  sleep $SLEEP
fi
FILELIST=‘for i in $*; do /bin/echo $i; done | /bin/sort -r‘
for fs in $FILELIST
do
    # Is the filesystem mounted?
    # -s says only return status, -x says exact match
    # we use awk instead of cut because mount outputs
    # lots of leading blanks that confuse cut
  /etc/mount | awk ’{ print $2 }’ | fgrep -s -x "$fs"
  if [ $? -eq 0 ]
  then
      # At least one filesystem is mounted
    until /etc/umount -f $fs
    do
      sleep 2
    done
  fi
done
######### Add for NFS Lock Removal (start) #########
if [ "$STOPPED" = "true"  ]
then
  startsrc -s rpc.statd
  startsrc -s rpc.lockd
fi
######### Add for NFS Lock Removal (finish) #########
exit 0
```

Chapter 6. Cluster Testing

Before you start to test the HACMP configuration, you need to guarantee that your cluster nodes are in a stable state. Check the state of the:

- Devices
- System parameters
- Processes
- Network adapters
- LVM
- Cluster
- Other items such as SP Switch, printers, and SNA configuration

6.1 Node Verification

Here is a series of suggested actions to test the state of a node before including HACMP in the testing.

6.1.1 Device State

- Run `diag -a` in order to clean up the VPD.
- Look in the errorlog for unusual errors by issuing the command `errpt | more` or `errpt -a | more`.
- Check that all devices are in the available state (`lsdev -C | more`).
- Check that the SCSI addresses of adapters on shared buses are unique (`lsattr -E -l ascsi0`).
- If you are using target mode SCSI networks, check the connection by issuing `cat < /dev/tmscsi#.tm` on the first node and `cat /etc/hosts > /dev/tmscsi#.im` (enter twice!) on the second node where `#` is the appropriate tmscsi device number. Repeat the test in the other direction. Note that cluster services must be stopped on both nodes to perform this test.
- To check a serial line between two nodes type `stty < /dev/tty#` on both nodes where `#` is the appropriate tty device number for the RS232 heartbeat connection. Note that cluster services must be stopped on both nodes to perform this test.

6.1.2 System Parameters

- Type date on all nodes to check that all the nodes in the cluster are running with their clocks on the same time.
- Ensure that the number of user licenses has been correctly set (lslicense).
- Check high water mark and other system settings (smitty chgsys).
- Type sysdumpdev -l and sysdumpdev -e to ensure that the dump space is correctly set and that the primary dump device (lslv hd7) is large enough to accomodate a dump.
- Check that applications to be controlled by HACMP are not started here, and that extraneous processes which might interfere with HACMP and/or dominate system resources are not started (more /etc/inittab).
- Check list of cron jobs (crontab -l).

6.1.3 Process State

- Check the paging space usage by issuing lsps -a.
- Look for all expected processes with ps -ef | more.
- Check that the run queue is < 5 and that the CPU usage is at an acceptable level (vmstat 2 5).

6.1.4 Network State

- Type for example ifconfig lo0, ifconfig en0 and ifconfig en1 to check the network adapter configuration, if you are using ethernet adapters. For other types of adapters, use the appropriate device name.
- To check the configuration of an SP Switch adapter, type: /usr/lpp/ssp/css/ifconfig css0.
- Use netstat -i or netstat -in to show the network configuration of the node.
- To check the alternate MAC address, issue netstat -v ent0 | more.
- Look at mbufs sizing relative to requests for memory denied (netstat -m | more).
- Type netstat -r or netstat -rAn to ensure that there are valid routes to the other cluster node interfaces and to clients.
- Run no -a | more and look at the setting of ipforwarding and ipsendredirects.

- Check that all interfaces communicate (`ping <ip-address>` or `ping -R <ip-address>)`.
- List the arp table entries with `arp -a`.
- Check the status of the TCP/IP daemons (`lssrc -g tcpip`).
- Ensure that there are no bad entries in the /etc/hosts file, especially at the bottom of the file.
- Verify that, if DNS is in use, the DNS servers are correctly defined (`more /etc/resolv.conf`).
- Check the status of NIS by typing `ps -ef | grep ypbind` and `lssrc -g yp`.
- The command `exportfs` shows non-HACMP controlled NFS exports.
- Run `snmpinfo -m dump -o /usr/sbin/cluster/hacmp.defs address` show snmp information for Cluster network addresses (including the serial interfaces).

6.1.5 LVM State

- Ensure that the correct VG's are defined, that quorum and auto-varyon are correctly defined, and that the shared VG's are in the correct state (`lsvg` and `lsvg -o`).
- Check that there are no stale partitions (`lsvg -l`).
- Check that all appropriate file systems have been mounted and that none of the rootvg file systems are full (`df -k`).
- Check that PVid's have been assigned where necessary and that there are no ghost disks(`lspv`).
- Verify that all entries in the /etc/filesystems file are correct and that there are no erroneous entries (`more /etc/filesystems` and `lsfs`).

6.1.6 Cluster State

- Check the status of the cluster daemons by issuing `lssrc -g cluster` and `lssrc -g lock`.
- Run `/usr/sbin/cluster/clstat` to check the status of the cluster and the status of the network interfaces.
- Check the cluster logfiles with `tail -f /tmp/hacmp.out`, `more /usr/sbin/cluster/history/cluster.mmdd` (mmdd = current date), `tail -f /var/adm/cluster.log` and `more /tmp/cm.log`.
- Check that the nodename is correct (`odmget HACMPcluster`).

- Verify the cluster configuration by running `/usr/sbin/cluster/diag/clconfig -v '-tr'`.
- To show cluster configuration, run: `/usr/sbin/cluster/utilities/cllscf`.
- To show the clstrmgr version, type: `snmpinfo -m dump -o /usr/sbin/cluster/hacmp.defs clstrmgr`.

6.2 Simulate Errors

The following paragraphs will give you hints on how you can simulate different hardware and software errors in order to verify your HACMP configuration. As an example, we will use a cluster consisting of two nodes and a cascading resource group definition.

The term “NodeF” is used for the node to be failed, and the term “NodeT” for the takeover node of NodeF's resource group.

When executing the test plan, it is helpful to monitor cluster activities during failover with the following commands. Note that the /tmp/hacmp.out file is the most useful to monitor, especially if the Debug Level of the HACMP Run Time Parameters for the nodes has been set to “high”, and if the Application Server Scripts include the `set -x` flag and periodic `echo` commands.

6.2.1 Adapter Failure

The following sections cover adapter failure.

6.2.1.1 Ethernet or Token Ring Interface Failure

In case of an Ethernet or Token Ring interface failure, perform the following steps:

- Check that all the nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).
- Monitor the cluster log files on NodeT.
- Use ifconfig to shut off the appropriate service interface (but not the Administrative SP Ethernet) on NodeF (for example, `ifconfig en0 down`). This will cause the service IP address to failover to the standby adapter on NodeF.
- Verify that the swap adapter has occurred (including MAC Addressfailover) and that HACMP has turned the original service interface back on as the standby interface.

- Use ifconfig to swap the service address back to the original service interface back (`ifconfig en1 down`). This will cause the service IP address to failover back to the service adapter on NodeF.

6.2.1.2 Ethernet or Token Ring Adapter or Cable Failure

Perform the following steps in the event of an Ethernet or Token Ring adapter or cable failure:

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).
- Monitor the cluster log files on NodeT.
- Disconnect the network cable from the appropriate service interface (but not the Administrative SP Ethernet) on NodeF. This will cause the service IP and MAC addresses to failover to the standby adapter on NodeF.
- Verify that the swap adapter has occurred.
- Reconnect the network cable to the service interface. This will cause the original service interface to become the standby interface.
- Initiate a swap adapter back to the original service interface by disconnecting the network cable from the new service interface (originally the standby interface). This will cause the service IP and MAC addresses to failover back to the service adapter on NodeF.
- Verify that the swap adapter has occurred.
- Reconnect the cable to the original standby interface.
- Verify that the original standby interface is operating with the standby IP address.

6.2.1.3 Switch Adapter Failure

Perform the following steps in case of switch adapter failure:

Note

Do not disconnect live switch cables to simulate a switch failure!

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Assign NodeF to be Eprimary.
- Optional: Prune the error log on NodeF (errclear 0).
- Monitor the cluster log files on NodeT.

- Generate the switch error in the error log which is being monitored by HACMP Error Notification (for configuration see 2.6.2.1, “Single Point-of-Failure Hardware Component Recovery” on page 46), or, if the network_down event has been customized, bring down css0 (`ifconfig css0 down`) or fence out NodeF from the Control Workstation (`Efence NodeF`).
- If the first failure simulation method is used, the switch failure will be detected in the error log (`errpt -a | more`) on NodeF and cause a node failover to NodeT. The other two methods will cause HACMP to detect a network_down event, with the same result. (Note that if there is another node in the cluster with a lower alphanumeric node name than NodeT, then that node will become Eprimary. HACMP does not take care of the Eprimary if a new SP-Switch is used).
- Verify that failover has occurred (`netstat -i` and `ping` for networks, `lsvg -o` and `vi` of a test file for volume groups, `ps -U <appuid>` for application processes, and `Eprimary` for Eprimary).
- Start HACMP on NodeF (`smit clstart`). NodeT will release NodeF's cascading Resource Groups and NodeF will take them back over, but NodeT (or a lower alphanumeric node) will remain Eprimary.
- Verify that re-integration has occurred (`netstat -i` and `ping` for networks, `lsvg -o` and `vi` of a test file for volume groups, `ps -U <appuid>` for application processes, and `Eprimary` for Eprimary).

6.2.1.4 Failure of a 7133 Adapter

Perform the following steps in the event of a 7133 Adapter failure:

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).
- Monitor cluster logfiles on NodeT if HACMP has been customized to monitor 7133 disk failures.
- Pull all the cable from the SSA adapter.
- The failure of the 7133 adapter should be detected in the error log (`errpt -a | more`) on NodeF or should be noted in the appropriate diagnostics tool, and the logical volume copies on the disks in drawer 1 will be marked stale (`lsvg -l NodeFvg`).
- Verify that all sharedvg file systems and paging spaces are accessible (`df -k` and `lsps -a`).
- Re-attach the cables.

- Verify that all sharedvg file systems and paging spaces are accessible (`df -k` and `lsps -a`).

6.2.2 Node Failure / Reintegration

The following sections deal with issues of node failure and reintegration.

6.2.2.1 AIX Crash

Perform the following steps in the event of an AIX crash:

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).
- If NodeF is an SMP, you may want to set the fast reboot switch (`mpcfg -cf 11 1`).
- Monitor cluster logfiles on NodeT.
- Crash NodeF by entering `cat /etc/hosts > /dev/kmem`. (The LED on NodeF will display 888.)
- The OS failure on NodeF will cause a node failover to NodeT.
- Verify that failover has occurred (`netstat -i` and `ping` for networks, `lsvg -o` and `vi` of a test file for volume groups, and `ps -U <appuid>` for application processes).
- Power cycle NodeF. If HACMP is not configured to start from /etc/inittab, (on restart) start HACMP on NodeF (`smit clstart`). NodeF will take back its cascading Resource Groups.
- Verify that re-integration has occurred (`netstat -i` and `ping` for networks, `lsvg -o` and `vi` of a test file for volume groups, and `ps -U <appuid>` for application processes).

6.2.2.2 CPU Failure

Perform the following steps in the event of CPU failure:

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).
- If NodeF is an SMP, you may want to set the fast reboot switch (`mpcfg -cf 11 1`).
- Monitor cluster logfiles on NodeT.
- Power off NodeF. This will cause a node failover to NodeT.

- Verify that failover has occurred (`netstat -i` and `ping` for networks, `lsvg -o` and `vi` of a test file for volume groups, and `ps -U <appuid>` for application processes).
- Power cycle NodeF. If HACMP is not configured to start from /etc/inittab (on restart), start HACMP on NodeF (`smit clstart`). NodeF will take back its cascading Resource Groups.
- Verify that re-integration has occurred (`netstat -i` and `ping` for networks, `lsvg -o` and `vi` of a test file for volume groups, and `ps -U <appuid>` for application processes).

6.2.2.3 TCP/IP Subsystem Failure

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (errclear 0).
- Monitor the cluster log files on NodeT.
- On NodeF, stop the TCP/IP subsystem (`sh /etc/tcp.clean`) or crash the subsystem by increasing the size of the sb_max and thewall parameters to large values (`no -o sb_max=10000; no -o thewall=10000`) and ping NodeT. Note that you should record the values for sb_max and thewall prior to modifying them, and, as an extra check, you may want to add the original values to the end of /etc/rc.net.
- The TCP/IP subsystem failure on NodeF will cause a network failure of all the TCP/IP networks on NodeF. Unless there has been some customization done to promote this type of failure to a node failure, only the network failure will occur. The presence of a non-TCP/IP network (RS232, target mode SCSI or target mode SSA) should prevent the cluster from triggering a node down in this situation.
- Verify that the network_down event has been run by checking the /tmp/hacmp.out file on either node. By default, the network_down script does nothing, but it can be customized to do whatever is appropriate for that situation in your environment.
- On NodeF, issue the command `startsrc -g tcpip`. This should restart the TCP/IP daemons, and should cause a network_up event to be triggered in the cluster for each of your TCP/IP networks.

6.2.3 Network Failure

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).

- Monitor the cluster log files on NodeT.
- Disconnect the network cable from the appropriate service and all the standby interfaces at the same time (but not the Administrative SP Ethernet) on NodeF. This will cause HACMP to detect a network_down event.
- HACMP triggers events dependent on your configuration of the network_down event. By default, no action is triggered by the network_down event.
- Verify that the expected action has occurred.

6.2.4 Disk Failure

The following sections deal with issues of disk failure.

6.2.4.1 Mirrored rootvg Disk (hdisk0) Failure

Perform the following steps in case of mirrored rootvg disk (hdisk0) failure:

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Verify that the bootlist contains hdisk0 and hdisk1, if for example, hdisk1 is the mirror of hdisk0 (`bootlist -m normal -o`).
- Optional: Prune the error log on NodeF (`errclear 0`).
- Monitor cluster logfiles on NodeT if HACMP has been customized to monitor SCSI disk failures.
- Slide back cover/casing on NodeF to get access to hdisk0 (this may first require turning the key to service mode). Pull the power cable (several colored wires with a white plastic connector) from the rear of hdisk0 (the lower internal disk is hdisk0, and the upper internal disk is hdisk1 on most systems). If you have a hot-pluggable disk, just pull the disk out of the frame.
- The failure of hdisk0 should be detected in the error log (`errpt -a | more`) on NodeF.
- Verify that all rootvg file systems and paging spaces are accessible (`df; lsps -a`).
- Shutdown (`smit clstop`; `shutdown -F`) and power off NodeF.
- Turn key to normal mode, power on NodeF, and verify that the system boots correctly. Log in and verify that all the rootvg file systems have been mounted (`df`).
- Shutdown (`shutdown -F`) and power off NodeF.

- Reconnect hdisk0, close the casing, and turn the key to normal mode.
- Power on NodeF then verify that the rootvg logical volumes are no longer stale (`lsvg -l rootvg`).

6.2.4.2 7135 Disk Failure

Perform the following steps in the event of a disk failure:

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).
- Monitor cluster logfiles on NodeT if HACMP has been customized to monitor 7135 disk failures.
- Mark a shared disk failed through smit (`smit raidiant`; **RAIDiant Disk Array Manager -> Change/Show Drive Status -> select the appropriate hdisk -> select the appropriate physical disk -> F4 to select a Drive Status of 83 Fail Drive)**, or if the disk is hot pluggable, remove the disk.
- The amber light on the front of the 7135 comes on, and can also be seen in SMIT (`smit raidiant`; **RAIDiant Disk Array Manager -> List all SCSI RAID Arrays**).
- Verify that all sharedvg file systems and paging spaces are accessible (`df` and `lsps -a`).
- If using RAID5 with Hot Spare, verify that reconstruction has completed to the Hot Spare, then un-mark or plug the failed disk back in. If using RAID1, sync the volume group (`syncvg NodeFvg`).
- If using RAID5 without Hot Spare, mark the failed disk Optimal (`smit raidiant`; **RAIDiant Disk Array Manager -> Change/Show Drive Status; select the appropriate hdisk -> select the appropriate physical disk -> F4 to select a Drive Status of 84 Replace and Reconstruct Drive**).
- Verify that the reconstruction has completed (`smit raidiant`;**RAIDiant Disk Array Manager -> List all SCSI RAID Arrays**).
- Verify that all sharedvg file systems and paging spaces are accessible (df and lsps -a) and that the partitions are not stale (`lsvg -l sharedvg`). Also verify that the yellow light has turned off on the 7135.

6.2.4.3 Mirrored 7133 Disk Failure

- Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
- Optional: Prune the error log on NodeF (`errclear 0`).

- Monitor cluster logfiles on NodeT if HACMP has been customized to monitor 7133 disk failures.
- Since the 7133 disk is hot pluggable, remove a disk from drawer 1 associated with NodeF's shared volume group.
- The failure of the 7133 disk will be detected in the error log (`errpt -a | more`) on NodeF, and the logical volumes with copies on that disk will be marked stale (`lsvg -l NodeFvg`).
- Verify that all NodeFvg file systems and paging spaces are accessible (`df -k and lsps -a`).
- Plug the failed disk back in, then sync the volume group (`syncvg NodeFvg`).
- Verify that all NodeFvg file systems and paging spaces are accessible (`df -k` and `lsps -a`) and that the partitions are not stale (`lsvg -l NodeFvg`).

6.2.5 Application Failure

By default, HACMP does not recognize application failures. With some additional configuration, it is possible to “teach” HACMP application failures and trigger events (for more information see 2.6.2.3, “Application Failure” on page 47).

So, the way of testing application failures is strongly dependent on your configuration.

Before you start to do the configuration and the testing application failure notification, analyse your application for possible failures. Then try to reproduce them.

Chapter 7. Cluster Troubleshooting

Typically, a functioning HACMP cluster requires minimal intervention. If a problem occurs, however, diagnostic and recovery skills are essential. Thus, troubleshooting requires that you identify the problem quickly and apply your understanding of the HACMP for AIX software to restore the cluster to full operation.

In general, troubleshooting an HACMP cluster involves:

- Becoming aware that a problem exists
- Determining the source of the problem
- Correcting the problem

Becoming aware of a problem is often through system messages on the console, end-users complaining about slow or unavailable services or through some sort of monitoring of your cluster. When an HACMP for AIX script or daemon generates a message, the message is written to the system console and to one or more cluster log files. Messages written to the system console may scroll off screen before you notice them. The following paragraphs provide an overview of the log files, which are to be consulted for cluster troubleshooting, as well as some information on specific cluster states you may find there.

7.1 Cluster Log Files

HACMP for AIX scripts, daemons, and utilities write messages to the following log files:

Table 21. HACMP Log Files

Log File Name	Description
`/var/adm/cluster.log`	Contains time-stamped, formatted messages generated by HACMP for AIX scripts and daemons. In this log file, there is one line written for the start of each event, and one line written for the completion.
`/tmp/hacmp.out`	Contains time-stamped, formatted messages generated by the HACMP for AIX scripts. In verbose mode, this log file contains a line-by-line record of each command executed in the scripts, including the values of the arguments passed to the commands. By default, the HACMP for AIX software writes verbose information to this log file; however, you can change this default. Verbose mode is recommended.

Log File Name	Description
system error log	Contains time-stamped, formatted messages from all AIX subsystems, including the HACMP for AIX scripts and daemons.
/usr/sbin/cluster/ history/cluster.mmdd	Contains time-stamped, formatted messages generated by the HACMP for AIX scripts. The system creates a new cluster history log file every day that has a cluster event occurring. It identifies each day's file by the filename extension, where *mm* indicates the month and *dd* indicates the day.
/tmp/cm.log	Contains time-stamped, formatted messages generated by HACMP for AIX clstrmgr activity. Information in this file is used by IBM Support personnel when the clstrmgr is in debug mode. Note that this file is overwritten every time cluster services are started; so, you should be careful to make a copy of it before restarting cluster services on a failed node.
/tmp/cspoc.log	Contains time-stamped, formatted messages generated by HACMP for AIX C-SPOC commands. Because the C-SPOC utility lets you start or stop the cluster from a single cluster node, the /tmp/cspoc.log is stored on the node that initiates a C-SPOC command.
/tmp/dms_logs.out	Stores log messages every time HACMP for AIX triggers the deadman switch.
/tmp/emuhacmp.out	Contains time-stamped, formatted messages generated by the HACMP for AIX Event Emulator. The messages are collected from output files on each node of the cluster, and cataloged together into the /tmp/emuhacmp.out log file.In verbose mode (recommended), this log file contains a line-by-line record of every event emulated. Customized scripts within the event are displayed, but commands within those scripts are not executed.

For a more detailed description of the cluster log files consult Chapter 2 of the *HACMP for AIX, Version 4.3: Troubleshooting Guide*, SC23-4280.

7.2 config_too_long

If the cluster manager recognizes a state change in the cluster, it acts upon it by executing an event script. However, some circumstances, like errors within the script or special conditions of the cluster, might cause the event script to hang. After a certain amount of time, by default 360 seconds, the cluster manager will issue a config_too_long message into the /tmp/hacmp.out file.

The message issued looks like this:

```
The cluster has been in reconfiguration too long;Something may be wrong.
```

In most cases, this is because an event script has failed. You can find out more by analyzing the /tmp/hacmp.out file.The error messages in the /var/adm/cluster.log file may also be helpful. You can then fix the problem identified in the log file and execute the `clruncmd` command on the command line, or by using the `SMIT Cluster Recovery Aids` screen. The `clruncmd` command signals the Cluster Manager to resume cluster processing.

Note, however, that sometimes scripts simply take too long, so the message showing up isn't always an error, but sometimes a warning. If the message is issued, that doesn't necessarily mean that the script failed or never finished. A script running for more than 360 seconds can still be working on something and eventually get the job done. Therefore, it is essential to look at the /tmp/hacmp.out file to find out what is actually happening.

7.3 Deadman Switch

The term "deadman switch" describes the AIX kernel extension that causes a system panic and dump under certain cluster conditions if it is not reset. The deadman switch halts a node when it enters a hung state that extends beyond a certain time limit. This enables another node in the cluster to acquire the hung node's resources in an orderly fashion, avoiding possible contention problems.

If this is happening, and it isn't obvious why the cluster manager was kept from resetting this timer counter, for example because some application ran at a higher priority as the `clstrmgr` process, customizations related to performance problems should be performed in the following order:

1. Tune the system using I/O pacing.
2. Increase the syncd frequency.
3. If needed, increase the amount of memory available for the communications subsystem.
4. Change the Failure Detection Rate.

Each of these options is described in the following sections.

7.3.1 Tuning the System Using I/O Pacing

Use I/O pacing to tune the system so that system resources are distributed more equitably during large disk writes. Enabling I/O pacing is required for an HACMP cluster to behave correctly during large disk writes, and it is strongly recommended if you anticipate large blocks of disk writes on your HACMP cluster.

You can enable I/O pacing using the `smit chgsys` fastpath to set high- and low-water marks. These marks are by default set to zero (disabling I/O pacing) when AIX is installed. While the most efficient high- and low-water marks vary from system to system, an initial high-water mark of 33 and a low-water mark of 24 provide a good starting point. These settings only slightly reduce write times, and consistently generate correct failover behavior from HACMP for AIX. If a process tries to write to a file at the high-water mark, it must wait until enough I/O operations have finished to make the low-water mark. See the *AIX Performance Monitoring & Tuning Guide*, SC23-2365 for more information on I/O pacing.

7.3.2 Extending the syncd Frequency

Edit the /sbin/rc.boot file to increase the `syncd` frequency from its default value of 60 seconds to either 30, 20, or 10 seconds. Increasing the frequency forces more frequent I/O flushes and reduces the likelihood of triggering the deadman switch due to heavy I/O traffic.

7.3.3 Increase Amount of Memory for Communications Subsystem

If the output of `netstat -m` reports that requests for mbufs are being denied, or if errors indicating `LOW_MBUFS` are being logged to the AIX error report, increase the value associated with “`thewall`” network option. The default value is 25% of the real memory. This can be increased to as much as 50% of the real memory.

To change this value, add a line similar to the following at the end of the /etc/rc.net file:

```
no -o thewall=xxxxx
```

where *xxxxx* is the value you want to be available for use by the communications subsystem. For example,

```
no -o thewall=65536
```

7.3.4 Changing the Failure Detection Rate

Use the `SMIT Change/Show a Cluster Network Module` screen to change the failure detection rate for your network module *only* if enabling I/O pacing or extending the syncd frequency did not resolve deadman problems in your cluster. By changing the failure detection rate to “Slow”, you can extend the time required before the deadman switch is invoked on a hung node and before a takeover node detects a node failure and acquires a hung node’s resources. See the *HACMP for AIX, Version 4.3: Administration Guide, SC23-4279* for more information and instructions on changing the Failure Detection Rate.

Note

I/O pacing must be enabled before completing these procedures; it regulates the number of I/O data transfers. Also, keep in mind that the `Slow` setting for the Failure Detection Rate is network specific.

7.4 Node Isolation and Partitioned Clusters

Node isolation occurs when all networks connecting nodes fail but the nodes remain up and running. One or more nodes can then be completely isolated from the others. A cluster in which this has happened is called a *partitioned cluster*. A partitioned cluster has two groups of nodes (one or more in each), neither of which cannot communicate with the other. Let’s consider a two node cluster where all networks have failed between the two nodes, but each node remains up and running.

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

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

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

When a cluster becomes partitioned, and the network problem is cleared after the point when takeover processing has begun so that keepalive packets start flowing between the partitioned nodes again, something must be done to restore order in the cluster. This order is restored by the DGSP Message.

7.5 The DGSP Message

A DGSP message (short for Diagnostic Group Shutdown Partition) is sent when a node loses communication with the cluster and then tries to re-establish communication.

For example, if a cluster node becomes unable to communicate with other nodes, yet it continues to work through its process table, the other nodes conclude that the “missing” node has failed because they no longer are receiving keepalive messages from it. The remaining nodes then process the necessary events to acquire the disks, IP addresses, and other resources from the “missing” node. This attempt to take over resources results in the dual-attached disks receiving resets to release them from the “missing” node and the start of IP address takeover scripts.

As the disks are being acquired by the takeover node (or after the disks have been acquired and applications are running), the “missing” node completes its process table (or clears an application problem) and attempts to resend keepalive messages and rejoin the cluster. Since the disks and IP addresses are in the process of being successfully taken over, it becomes possible to have a duplicate IP address on the network and the disks may start to experience extraneous traffic on the data bus.

Because the reason for the “missing” node remains undetermined, you can assume that the problem may repeat itself later, causing additional down time of not only the node but also the cluster and its applications. Thus, to ensure the highest cluster availability, a DGSP message is sent to all nodes in one of the partitions. Any node receiving a DGSP message halts immediately, in order to not cause any damage on disks or confusion on the networks.

In a partitioned cluster situation, the smaller partition (lesser number of nodes) is shut down, with each of its nodes getting a DGSP message. If the partitions are of equal size, the one with the node name beginning in the lowest name in the alphabet gets shut down. For example, in a cluster where one partition has NodeA and the other has NodeB, NodeB will be shut down.

7.6 User ID Problems

Within an HACMP cluster, you always have more than one node potentially offering the same service to a specific user or a specific user id.

As the node providing the service can change, the system administrator has to ensure that the same user and group is known to all nodes potentially running an application. So, in case one node is failing, and the application is taken over by the standby node, a user can go on working since the takeover node knows that user under exactly the same user and group id.

Since user access within an NFS mounted file system is granted based on user IDs, the same applies to NFS mounted file systems.

For more information on managing user and group accounts within a cluster, refer to Chapter 2.7, “User ID Planning” on page 48, or to Chapter 12, “Managing User and Groups in a Cluster” of the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279.

7.7 Troubleshooting Strategy

In order to quickly find a solution to a problem in the cluster, some sort of strategy is helpful for pinpointing the problem. The following guidelines should make the troubleshooting process more productive:

- Save the log files associated with the problem before they become unavailable. Make sure you save the /tmp/hacmp.out and /tmp/cm.log files before you do anything else to try to figure out the cause of the problem.
- Attempt to duplicate the problem. Do not rely too heavily on the user’s problem report. The user has only seen the problem from the application level. If necessary, obtain the user’s data files to recreate the problem.
- Approach the problem methodically. Allow the information gathered from each test to guide your next test. Do not jump back and forth between tests based on hunches.
- Keep an open mind. Do not assume too much about the source of the problem. Test each possibility and base your conclusions on the evidence of the tests.
- Isolate the problem. When tracking down a problem within an HACMP cluster, isolate each component of the system that can fail and determine whether it is working. Work from top to bottom, following the progression described in the following section.

- Go from the simple to the complex. Make the simple tests first. Do not try anything complex and complicated until you have ruled out the simple and obvious.
- Do not make more than one change at a time. If you do, and one of the changes corrects the problem, you have no way of knowing which change actually fixed the problem. Make one change, test the change, and then, if necessary, make the next change.
- Do not neglect the obvious. Small things can cause big problems. Check plugs, connectors, cables, and so on.
- Keep a record of the tests you have completed. Record your tests and results, and keep an historical record of the problem, in case it reappears.

Chapter 8. Cluster Management and Administration

This chapter covers all aspects of monitoring and managing an existing HACMP cluster. This includes a description of the different monitoring methods and tools available, how to start and stop the cluster, changing cluster or resource configurations, applying software fixes, user management, and other things.

8.1 Monitoring the Cluster

By design, HACMP for AIX compensates for various failures that occur within a cluster. For example, HACMP for AIX compensates for a network adapter failure by swapping in a standby adapter. As a result, it is possible that a component in the cluster could have failed and that you would be unaware of the fact. The danger here is that, while HACMP for AIX can survive one or possibly several failures, a failure that escapes your notice threatens a cluster's ability to maintain a highly available environment.

HACMP for AIX provides the following tools for monitoring an HACMP cluster:

- The `/usr/sbin/cluster/clstat` utility, which reports the status of key cluster components—the cluster itself, the nodes in the cluster, and the network adapters connected to the nodes.
- The `HAView` utility, which monitors HACMP clusters through the NetView for AIX graphical network management interface. It lets users monitor multiple HACMP clusters and cluster components across a network from a single node.
- The `SMIT Show Cluster Services` screen, which shows the status of the HACMP for AIX daemons
- The following log files: the /var/adm/cluster.log file, which tracks cluster events, the /tmp/hacmp.out file, which records the output generated by configuration scripts as they execute, the /usr/sbin/cluster/history/cluster.mmdd log file, which logs the daily cluster history, and the /tmp/cspoc.log file, which logs the status of C-SPOC commands executed on cluster nodes.

When you monitor a cluster, use the `clstat` utility to examine the cluster and its components. Also, constantly monitor the /tmp/hacmp.out file. Use the *SMIT Show Cluster Services* screen to make sure that the necessary HACMP for AIX daemons are running on each node. Finally, if necessary, examine the other cluster log files to get a more in-depth view of the cluster status.

Consult the *HACMP for AIX, Version 4.3: Troubleshooting Guide*, SC23-4280, for help if you detect a problem with an HACMP cluster.

8.1.1 The clstat Command

HACMP for AIX provides the `/usr/sbin/cluster/clstat` command for monitoring a cluster and its components. The `clstat` utility is a clinfo client program that uses the Clinfo API to retrieve information about the cluster. Clinfo must be running on a node for this utility to work properly.

The `/usr/sbin/cluster/clstat` utility runs on both ASCII and X Window Display clients in either single-cluster or multi-cluster mode. Multi-cluster mode requires that you use the `-i` flag when invoking the `clstat` utility. The client display automatically corresponds to the capability of the system. For example, if you run `clstat` on an X Window client, a graphical display for the utility appears. However, you can run an ASCII display on an X-capable machine by specifying the `-a` flag. In order to set up a connection to the cluster nodes, the /usr/sbin/cluster/etc/clhosts file must be configured on the client.

The `clstat` utility reports whether the cluster is up, down, or unstable. It also reports whether a node is up, down, joining, leaving, or reconfiguring, and the number of nodes in the cluster. For each node, the utility displays the IP label and address of each network interface attached to the node, and whether that interface is up or down. See the `clstat` man page for additional information about this utility.

8.1.2 Monitoring Clusters using HAView

HAView is a cluster monitoring utility that allows you to monitor HACMP clusters using NetView for AIX. Using NetView, you can monitor clusters and cluster components across a network from a single management station.

HAView creates and modifies NetView objects that represent clusters and cluster components. It also creates submaps that present information about the state of all nodes, networks, and network interfaces associated with a particular cluster. This cluster status and configuration information is accessible through NetView’s menu bar.

HAView monitors cluster status using the Simple Network Management Protocol (SNMP). It combines periodic polling and event notification through traps to retrieve cluster topology and state changes from the HACMP management agent, that is, the Cluster SMUX peer daemon (`clsmuxpd`).

More details on how to configure HAView and on how to monitor your cluster with HAView can be found in Chapter 3, “Monitoring an HACMP cluster” in *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279.

8.1.3 Cluster Log Files

HACMP for AIX writes the messages it generates to the system console and to several log files. Because each log file contains a different subset of the types of messages generated by HACMP for AIX, you can get different views of cluster status by viewing different log files. HACMP for AIX writes messages into the log files described below. See Chapter 2, “Examining Cluster Log Files”, in the *HACMP for AIX, Version 4.3: Troubleshooting Guide,* SC23-4280 for more information about these files.

8.1.3.1 /var/adm/cluster.log

The /var/adm/cluster.log file is the main HACMP for AIX log file. HACMP error messages and messages about HACMP for AIX-related events are appended to this log with the time and date when they occurred.

8.1.3.2 /tmp/hacmp.out

The /tmp/hacmp.out file records the output generated by the configuration and startup scripts as they execute. This information supplements and expands upon the information in the /var/adm/cluster.log file. To receive verbose output, the Debug Level run-time parameter should be set to *high*, which is the default.

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

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

While it is more likely that you will use these files during troubleshooting, you should occasionally look at them to get a more detailed picture of the activity within a cluster.

8.1.3.4 System Error Log

The system error log file contains timestamped, formatted messages from all AIX subsystems, including HACMP for AIX scripts and daemons. Cluster events are logged as operator messages (error id: AA8AB241) in the system error log.

8.1.3.5 /tmp/cm.log

Contains timestamped, formatted messages generated by HACMP for AIX `clstrmgr` activity. This file is typically used by IBM support personnel.

8.1.3.6 /tmp/cspoc.log

Contains timestamped, formatted messages generated by HACMP for AIX C-SPOC commands. The /tmp/cspoc.log file resides on the node that invokes the C-SPOC command.

8.1.3.7 /tmp/emuhacmp.out

The /tmp/emuhacmp.out file records the output generated by the event emulator scripts as they execute. The /tmp/emuhacmp.out file resides on the node from which the event emulator is invoked. You can use the environment variable EMUL_OUTPUT to specify another name and location for this file, but the format and information remains the same.

With HACMP/ES, because of its RSCT technology, there are 3 more logfiles you may want to watch. These are:

8.1.3.8 /var/ha/log/grpsvcs.<filename>

Contains timestamped messages in ASCII format. These track the execution of internal activities of the grpsvcs daemon. IBM support personnel use this information for troubleshooting. The file gets trimmed regularly. Therefore, please save it promptly if there is a chance you may need it.

8.1.3.9 /var/ha/log/topsvcs.<filename>

Contains timestamped messages in ASCII format. These track the execution of internal activities of the topsvcs daemon. IBM support personnel use this information for troubleshooting. The file gets trimmed regularly. Therefore, please save it promptly if there is a chance you may need it.

8.1.3.10 /var/ha/log/grpglsm

The /var/ha/log/grpglsm file tracks the execution of internal activities of the grpglsm daemon. IBM support personnel use this information for troubleshooting. The file gets trimmed regularly. Therefore please save it promptly if there is a chance you may need it.

8.2 Starting and Stopping HACMP on a Node or a Client

This paragraph explains how to start and stop cluster services on cluster nodes and clients. It also describes how the Cluster-Single Point of Control

(C-SPOC) utility can be used to start and stop cluster services on all nodes in cluster environments.

Starting cluster services refers to the process of starting the HACMP for AIX daemons that enable the coordination required between nodes in a cluster. Starting cluster services on a node also triggers the execution of certain HACMP for AIX scripts that initiate the cluster. Stopping cluster services refers to stopping these same daemons on a node. This action may or may not cause the execution of additional HACMP for AIX scripts, depending on the type of shutdown you perform.

8.2.1 HACMP Daemons

The following lists the required and optional HACMP for AIX daemons.

8.2.1.1 Cluster Manager daemon (clstrmgr)

This daemon maintains the heartbeat protocol between the nodes in the cluster, monitors the status of the nodes and their interfaces, and invokes the appropriate scripts in response to node or network events. All cluster nodes must run the `clstrmgr` daemon.

8.2.1.2 Cluster SMUX Peer daemon (clsmuxpd)

This daemon maintains status information about cluster objects. This daemon works in conjunction with the Simple Network Management Protocol (`snmpd`) daemon. All cluster nodes must run the `clsmuxpd` daemon.

> **Note**
>
> The `clsmuxpd` daemon cannot be started unless the `snmpd` daemon is running.

8.2.1.3 Cluster Lock Manager daemon (cllockd)

This daemon provides advisory locking services. The `cllockd` daemon may be required on cluster nodes if those nodes are part of a concurrent access configuration, but this is not necessarily so. Check with your application vendor to see if it is required.

> **Note**
>
> If the `clsmuxpd` daemon or the `cllockd` daemon cannot be started by the Cluster Manager (e. g. the ports are already in use), the Cluster Manager logs an error message and dies.

8.2.1.4 Cluster Information Program daemon (clinfo)

This daemon provides status information about the cluster to cluster nodes and clients and invokes the `/usr/sbin/cluster/etc/clinfo.rc` script in response to a cluster event. The `clinfo` daemon is optional on cluster nodes and clients. However, it is a prerequisite for running the clstat utility.

With RSCT (RISC System Cluster Technology) on HACMP/ES Version 4.3, there are several more daemons.

8.2.1.5 Cluster Topology Services daemon (topsvcsd)

This daemon monitors the status of network adapters in the cluster. All HACMP/ES cluster nodes must run the `topsvcsd` daemon

8.2.1.6 Cluster Event Management daemon (emsvcsd)

This daemon matches information about the state of system resources with information about resource conditions of interest to client programs (applications, subsystems, and other programs).The `emsvcsd` daemon runs on each node of a domain.

8.2.1.7 Cluster Group Services daemon (grpsvcsd)

This daemon manages all of the distributed protocols required for cluster operation. All HACMP/ES cluster nodes must run the `grpsvcsd` daemon.

8.2.1.8 Cluster Globalized Server Daemon daemon (grpglsmd)

This daemon operates as a `grpsvcs` client; its function is to make switch adapter membership global across all cluster nodes. All HACMP/ES cluster nodes must run the `grpglsmd` daemon.

8.2.2 Starting Cluster Services on a Node

You start cluster services on a node by executing the HACMP `/usr/sbin/cluster/etc/rc.cluster` script. Use the *Start Cluster Services SMIT* screen to build and execute this command. The `rc.cluster` script initializes the environment required for HACMP by setting environment variables and then calls the `/usr/sbin/cluster/utilities/clstart` script to start the HACMP daemons. The clstart script is the HACMP script that starts all the cluster services. It does this by calling the SRC `startsrc` command to start the specified subsystem or group.

Using the C-SPOC utility, you can start cluster services on any node (or on all nodes) in a cluster by executing the C-SPOC `/usr/sbin/cluster/utilities/cl_rc.cluster` command on a single cluster node. The C-SPOC `cl_rc.cluster` command calls the `rc.cluster` command to start cluster services on the nodes specified from the one node. The nodes are started in sequential order - not in parallel. The output of the command run on the remote node is returned to the originating node. Because the command is executed remotely, there can be a delay before the command output is returned.

8.2.2.1 Automatically Restarting Cluster Services

You can optionally have cluster services start whenever the system is rebooted. If you specify the `-R` flag to the `rc.cluster` command, or specify **restart or both** in the *Start Cluster Services SMIT* screen, the `rc.cluster` script adds the following line to the /etc/inittab file.

```
hacmp:2:wait:/usr/sbin/cluster/etc/rc.cluster -boot> /dev/console 2>&1
# Bring up Cluster
```

At system boot, this entry causes AIX to execute the `/usr/sbin/cluster/etc/rc.cluster` script to start HACMP Cluster Services.

> **Note**
>
> Be aware that if the cluster services are set to restart automatically at boot time, you may face problems with node integration after a power failure and restoration, or you may want to test a node after doing maintenance work before having it rejoin the cluster.

8.2.2.2 Starting Cluster Services with IP Address Takeover Enabled

If IP address takeover is enabled, the `/usr/sbin/cluster/etc/rc.cluster` script calls the `/etc/rc.net` script to configure and start the TCP/IP interfaces and to set the required network options.

8.2.3 Stopping Cluster Services on a Node

You stop cluster services on a node by executing the HACMP `/usr/sbin/cluster/etc/clstop` script. Use the *HACMP for AIX Stop Cluster Services SMIT* screen to build and execute this command. The `clstop` script stops an HACMP daemon or daemons. The `clstop` script starts all the cluster services or individual cluster services by calling the SRC command `stopsrc`.

Using the C-SPOC utility, you can stop cluster services on a single node or on all nodes in a cluster by executing the C-SPOC `/usr/sbin/cluster/utilities/cl_clstop` command on a single node. The C-SPOC `cl_clstop` command performs some cluster-wide verification and then calls the `clstop` command to stop cluster services on the specified nodes. The nodes are stopped in sequential order—not in parallel. The output of the command that is run on the remote node is returned to the originating node. Because the command is executed remotely, there can be a delay before the command output is returned.

8.2.3.1 When to Stop Cluster services

You typically stop cluster services in the following situations:

- Before making any hardware or software changes or other scheduled node shutdowns or reboots. Failing to do so may cause unintended cluster events to be triggered on other nodes.
- Before certain reconfiguration activity. Some changes to the cluster information stored in the ODM require stopping and restarting the cluster services on *all* nodes for the changes to become active. For example, if you wish to change the name of the cluster, the name of a node, or the name of an adapter, you must stop and restart the cluster.

8.2.3.2 Types of Cluster Stops

When you stop cluster services, you must also decide how to handle the resources that were owned by the node you are removing from the cluster. You have the following options:

Graceful In a graceful stop, the HACMP software shuts down its applications and releases its resources. The other nodes do not take over the resources of the stopped node.

Graceful with Takeover In a graceful with takeover stop, the HACMP software shuts down its applications and releases its resources. The surviving nodes take over these resources. This is also called *intentional failover.*

Forced In a forced stop, the HACMP daemons only are stopped, without releasing any resources. For example, the stopped node stays on its service address if IP Address Takeover has been enabled. It does not stop its applications, unmount its file systems or varyoff its shared volume groups. The other nodes do not take over the resources of the stopped node. Please note that the forced option is currently not supported at the Version 4.3 level in HACMP/ES, only in HACMP Classic.

8.2.3.3 Abnormal Termination of a Cluster Daemon

If the SRC detects that any HACMP daemon has exited abnormally (without being shut down using the `clstop` command), it executes the `/usr/sbin/cluster/utilities/clexit.rc` script to halt the system. This prevents unpredictable behavior from corrupting the data on the shared disks. See the `clexit.rc` man page for additional information.

Important Note

Never use the `kill -9` command on the `clstrmgr` daemon. Using the `kill` command causes the `clstrmgr` daemon to exit abnormally. This causes the SRC to run the /usr/sbin/cluster/utilities/clexit.rc script which halts the system immediately, causing the surviving nodes to initiate failover.

8.2.4 Starting and Stopping Cluster Services on Clients

Use the `/usr/sbin/cluster/etc/rc.cluster` script or the `startsrc` command to start `clinfo` on a client, as shown below:

```
/usr/sbin/cluster/etc/rc.cluster
```

You can also use the standard AIX `startsrc` command:

```
startsrc -s clinfo
```

Use the standard AIX `stopsrc` command to stop `clinfo` on a client machine:

```
stopsrc -s clinfo
```

8.2.4.1 Maintaining Cluster Information Services on Clients

In order for the clinfo daemon to get the information it needs, you must edit the /usr/sbin/cluster/etc/clhosts file. As installed, the clhosts file on an HACMP client node contains no hostnames or addresses. HACMP server addresses must be provided by the user at installation time. This file should contain all boot and service names or addresses of HACMP servers from any cluster accessible through logical connections with this client node. Upon startup, `clinfo` uses these names or addresses to attempt communication with a `clsmuxpd` process executing on an HACMP server.

An example list of hostnames/addresses in a clhosts file follows:

```
n0_cl83   #  n0 service
n2_cl83   #  n2 service
n3_cl83   #  n3 service
```

For more detailed information on the clinfo command refer to Chapter 2, “Starting and Stopping Cluster Services”, *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279.

8.3 Replacing Failed Components

From time to time, it will be necessary to perform hardware maintenance or upgrades on cluster components. Some replacements or upgrades can be performed while the cluster is operative, while others require planned downtime. Make sure you plan all the necessary actions carefully. This will spare you a lot of trouble.

8.3.1 Nodes

When maintaining or upgrading a node, cluster services must usually be stopped on the node. This means down time for the applications usually running on this node, at least during the takeover to other nodes.

Consider the following points when replacing the whole or components of a node:

- Make sure you have at least the same amount of RAM in the replacement system.
- If your applications have been optimized for a particular processor or architecture, ensure that the new node is the same type of system. Uniprocessor applications may run slower on SMP systems.
- Slot capacity of the new node must be the same or better.
- Check the appropriate documentation for a proper adapter placement in your new node.
- The license of your application may be dependent on the CPU ID. You may need to apply for a new license before trying to bring the new node into service.

8.3.2 Adapters

In order to replace or add an adapter, the node must be powered off. This means down time for the applications usually running on this node, at least during the takeover to other nodes.

Consider the following points when replacing or adding adapters in a node:

- Make sure that the adapter is your problem and not faulty cabling. Bad cables are much more common than defective adapters. Most network and SSA cables can be changed online. Do some testing, for example, exchange the cables or try to connect to another port in your hub to see if the hub is your problem.

- The new adapter must be of the same type or a compatible type as the replaced adapter.
- When replacing or adding an SCSI adapter, remove the resistors for shared buses. Furthermore, set the SCSI ID of the adapter to a value different than 7.

8.3.3 Disks

Disk failures are handled differently according to the capabilities of the disk type and the HACMP version you are running. Whether your data is still available after a disk crash, and whether you will need down time to exchange it, will depend on the following questions:

- Is all the data on the failed disk mirrored to another disk, or is the failed disk part of a RAID array?
- Will the volume group stay online (Quorum)?
- Is the type of disk you are using hot-swappable?

8.3.3.1 SSA/SCSI Disk Replacement (RAID)

RAID arrays are typically designed for concurrent maintenance. No command line intervention should be necessary to replace a failed disk in a RAID array.

Do the following steps in order to replace a disk that is a member of a RAID array:

1. Remove the disk logically from the RAID array (for example with the appropriate SMIT menu). Removing a disk from a RAID array is known as reducing the RAID array. No more than one disk can be removed from an array at one time.
2. Remove the failed disk and plug in the substitute disk.
3. Add the replacement disk logically to the RAID array. All information from the original disk will be regenerated on the substitute disk. Once data regeneration has completed on the new disk, the array will return to its normal optimal mode of operation.

8.3.3.2 Disk Replacement (Non-RAID) before HACMP version 4.3

If LVM mirroring is used, some careful manual steps must be followed to replace a failed SCSI or SSA disk:

1. Identify which disk has failed, using `errpt`, `lspv`, `lsvg`, `diags`.
2. Remove all LV copies from the failed disk (`rmlvcopy`).
3. Remove the disk from the VG (`reducevg`).

4. Logically remove the disk from the system (rmdev -l hdiskX -d; rmdev -l pdiskY -d if a SSA disk) on all nodes.
5. Physically remove the failed disk and replace it with a new disk.
6. Add the disk to the ODM (mkdev or cfgmgr) on all nodes.
7. Add the disk to the shared volume group (extendvg).
8. Increase the number of LV copies to span across the new disk (mklvcopy).
9. Synchronize the volume group (syncvg)

> **Note**
>
> Steps 10 and 11 are only necessary in HACMP versions prior to 4.2. With HACMP 4.2 and later Lazy Update will export/import the volume group on the backup node in case of a takeover. However, it is necessary to update the PVID of the replaced disk on the backup nodes manually.

10. Stop all the application(s) using the shared volume group, varyoff the shared volume group and export/import it on the backup node(s). Furthermore set the characteristics of the shared volume group (autovaryon and quorum) on the backup node(s), then vary it off again.
11. Varyon the shared volume group on it's "normal" node and start the application(s).

8.3.3.3 Disk Replacement (Non-RAID) with HACMP version 4.3

With the HACMP 4.3 enhancements to the C-SPOC LVM utilities, the disk replacement does not cause system down time, as long as the failed disk was part of a RAID array, or if all the LVs on it are mirrored to other disks, and the failed disk is hot-swappable.

1. Identify which disk has failed using errpt, lspv, lsvg, diag.
2. Remove all LV copies from the failed disk (smit cl_lvsc).
3. Remove the disk from the VG (smit cl_vgsc).
4. Logically remove the disk from the system (rmdev -l hdiskX -d, rmdev -l pdiskY -d if SSA disk)
5. Physically remove the failed disk and replace it with a new disk.
6. Add the new disk to the ODM (mkdev or cfgmgr).
7. Add the new disk to the sharedvg (smit cl_vgsc).
8. Increase the number of LV copies to span across the new disk (smit cl_lvsc).
9. Sync the volume group (smit cl_syncvg).

8.4 Changing Shared LVM Components

Changes to VG constructs are probably the most frequent kind of changes to be performed in a cluster. As a system administrator of an HACMP for AIX cluster, you may be called upon to perform any of the following LVM-related tasks:

- Creating a new shared volume group
- Extending, reducing, changing, or removing an existing volume group
- Importing, mirroring, unmirroring, or synchronizing mirrors of a volume group
- Creating a new shared logical volume
- Extending, reducing, changing, copying, or removing an existing logical volume (or a copy)
- Creating a new shared file system
- Extending, changing, or removing an existing file system

The varyon of a shared volume group will only succeed if the information stored in the VGDA on the disks of the shared volume group and the information stored in the ODM are equal. After changes in the volume group (e. g. increasing the size of a file system), the information about the volume group in ODM and in the VGDA on the disks are still equal, but it will be different from the information in the ODM of a node that did not have the volume group varied on at the time of the change. In order to keep a takeover from failing, the volume group information must be synchronized. There are four distinct ways to keep all the volume group ODMs synchronized:

- Manual Update
- Lazy Update
- C-SPOC
- TaskGuide

Chapters 4 and 5 of the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279, describe in detail how to change shared LVM components.

8.4.1 Manual Update

Sometimes, manual updates of shared LVM components are inevitable because you cannot do some of the tasks mentioned above with any of the tools. For example, neither with C-SPOC nor with TaskGuide or Lazy Update is it possible to remove a VG on all of the cluster nodes.

When changing shared LVM components manually, you will usually need to run through the following procedure:

1. Stop HACMP on the node owning the shared volume group (sometimes a stop of the applications using the shared volume group may be sufficient).
2. Make the necessary changes to the shared LVM components.
3. Unmount all the file systems of the shared volume group.
4. Varyoff the shared volume group.
5. Export the old volume group definitions on the next node.
6. Import the volume group from one of its disks on the next node. Make sure you use the same VG major number.
7. Change the volume group to not auto-varyon at system boot time.
8. Mount all the file systems of the shared volume group.
9. Test the file systems.
10. Unmount the file systems of the shared volume group.
11. Varyoff the shared volume group.
12. Repeat steps 6 through 11 for all the other nodes with an old ODM of the shared volume group.
13. Start HACMP again on the node usually owning the shared volume group.

8.4.2 Lazy Update

For LVM components under the control of HACMP for AIX, you do not have to explicitly export and import to bring the other cluster nodes up-to-date. Instead, HACMP for AIX can perform the export and import when it activates the volume group during a failover. In a cluster, HACMP controls when volume groups are activated. HACMP for AIX implements a function, called Lazy Update, by keeping a copy of the timestamp from the volume group’s VGDA. AIX updates this timestamp whenever the LVM component is modified. When another cluster node attempts to vary on the volume group, HACMP for AIX compares its copy of the timestamp (kept in the /usr/sbin/cluster/etc/vg file) with the timestamp in the VGDA on the disk. If the values are different, the HACMP for AIX software exports and re-imports the volume group before activating it. If the timestamps are the same, HACMP for AIX activates the volume group without exporting and re-importing.

The time needed for takeover expands by a few minutes if a Lazy Update occurs. A Lazy Update is always performed the first time a takeover occurs in order to create the timestamp file on the takeover node.

Lazy Update has some limitations, which you need to consider when you rely on Lazy Update in general:

- If the first disk in a sharedvg has been replaced, the `importvg` command will fail as Lazy Update expects to be able to match the hdisk number for the first disk to a valid PVID in the ODM.
- Multi-LUN support on the SCSI RAID cabinets can be very confusing to Lazy Update as each LUN appears as a new hdisk known to only one node in the cluster (remember that Lazy Update works on LVM constructs).

8.4.3 C-SPOC

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 (see 8.8, “User Management” on page 178).
- Maintain shared Logical Volume Manager (LVM) components.
- Control HACMP services on a cluster-wide basis (see Chapter 8.2, “Starting and Stopping HACMP on a Node or a Client” on page 154).

Without C-SPOC functionality, the system administrator must spend time executing administrative tasks individually 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. For example, to add a user, you usually must perform this task on each cluster node. 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.

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, Version 4.3: Administration Guide, SC23-4279* for detailed information on using C-SPOC SMIT options.

With C-SPOC, you can perform the following tasks:

- Shared volume groups
    - List all volume groups in the cluster.
    - Import a volume group (with HACMP 4.3 only).
    - Extend a volume group (with HACMP 4.3 only).
    - Reduce a volume group (with HACMP 4.3 only).
    - Mirror a volume group (with HACMP 4.3 only).
    - Unmirror a volume group (with HACMP 4.3 only).
    - Synchronize volume group mirrors (with HACMP 4.3 only).
- Shared logical volumes
    - List all logical volumes by volume group.
    - Add a logical volume to a volume group (with HACMP 4.3 only).
    - Make a copy of a logical volume.
    - Remove a copy of a logical volume.
    - Show the characteristics of a logical volume.
    - Set the characteristics of a logical volume (name, size); this is only possible in non-concurrent mode and with HACMP 4.3.
    - Remove a logical volume.
- Shared file systems (only applicable for non-concurrent VGs)
    - List all shared file systems.
    - Change/View the characteristics of a shared file system.
    - Remove a shared file system.

C-SPOC has the following limitations:

- C-SPOC does not offer the option for creating volume groups. Use the TaskGuide or standard AIX commands.
- The new Volume Group must be imported manually to other nodes in the resource group (TaskGuide does it automatically).
- The Volume Group must be defined in a resource group, and cluster resources must be synchronized, prior to using C-SPOC to manage it.
- C-SPOC does not offer an option for creating file systems. Use standard AIX commands or SMIT menus to create file systems, and use C-SPOC to update the VG information on the other nodes.
- C-SPOC cannot be used for concurrent shared LVM components prior to HACMP 4.3 for AIX.

To use the SMIT shortcuts to C-SPOC, type `smit cl_lvm` or `smit cl_conlvm` for concurrent volume groups. Concurrent volume groups must be varied on in concurrent mode to perform tasks.

8.4.4 TaskGuide

The TaskGuide is a 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.

8.4.4.1 TaskGuide Requirements

TaskGuide is only available since HACMP for AIX version 4.3. Before you start the TaskGuide, make sure that:

- You have a configured HACMP cluster in place
- You are on a graphics capable terminal

8.4.4.2 Starting the TaskGuide

You can start the TaskGuide from the command line by typing: `/usr/sbin/cluster/tguides/bin/cl_ccvg`, or you can use the SMIT interface as follows:

1. Type `smit hacmp`
2. From the SMIT main menu, choose **Cluster System Management -> Cluster Logical Volume Manager ->Taskguide for Creating a Shared Volume Group.** After a pause, the TaskGuide Welcome panel appears.
3. Proceed through the panels to create or share a volume group.

8.5 Changing Cluster Resources

In HACMP for AIX, you define each resource as part of a resource group. This allows you to combine related resources into a single logical entity for easier configuration and management. You then configure each resource group to have a particular kind of relationship with a set of nodes. Depending on this relationship, resources can be defined as one of three types: cascading, rotating, or concurrent access. You also assign a priority to each participating node in a cascading resource group chain.

To change the nodes associated with a given resource group, or to change the priorities assigned to the nodes in a resource group chain, you must redefine the resource group. You must also redefine the resource group if you add or change a resource assigned to the group. This section describes how to add, change, and delete a resource group.

8.5.1 Add/Change/Remove Cluster Resources

You can add, change and remove a resource group in an active cluster. You do not need to stop and then restart cluster services for the resource group to become part of the current cluster configuration.

Use the following SMIT shortcuts:

To add a resource group, use `smit cm_add_grp`.

To remove a resource group, use `smit cm_add_res`.

To change a resource group, use `smit cm_add_res`.

Whenever you modify the configuration of cluster resources in the ODM on one node, you must synchronize the change across all cluster nodes.

8.5.2 Synchronize Cluster Resources

You perform a synchronization by choosing the **Synchronize Cluster Resources** option from the Cluster Resources SMIT screen.

> **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.

The processing performed in synchronization varies depending on whether the Cluster Manager is active on the local node:

- If the cluster manager is not active on the local node when you select this option, the ODM data in the DCD (Default Configuration Directory–for more information, see Chapter 3 in the *HACMP for AIX, Version 4.3: Concepts and Facilities*, SC23-4276) on the local node is copied to the ODMs stored in the DCDs on all cluster nodes.

- If the Cluster Manager is active on the local node, synchronization triggers a cluster-wide, dynamic reconfiguration event. In dynamic reconfiguration, the configuration data stored in the DCD is updated on each cluster node, and, in addition, the new ODM data replaces the ODM data stored in the ACD (Active Configuration Directory) on each cluster node. The cluster daemons are refreshed and the new configuration becomes the active configuration. In the HACMP for AIX log file, reconfig_resource_release, reconfig_resource_acquire, and reconfig_resource_complete events mark the progress of the dynamic reconfiguration.
- If the Cluster Manager is active on some cluster nodes but not on the local node, the synchronization is aborted.

8.5.3 DARE Resource Migration Utility

The HACMP for AIX software provides a Dynamic Reconfiguration (DARE) Resource Migration utility that allows for improved cluster management by allowing a system administrator to alter the placement of resource groups (along with their resources—IP addresses, applications, and disks) to specific cluster nodes using the `cldare` command. The command lets you move the ownership of a series of resource groups to a specific node in that resource group’s node list, as long as the requested arrangement is not incompatible with the current resource group configuration. It also lets you disable resource groups, preventing them from being acquired during a failover or reintegration.

Dynamic resource group movement essentially lets a system administrator better use hardware resources within the cluster, forcing resource traffic onto one or more high-powered or better-connected nodes without having to shut down HACMP on the node from which the resource group is moved. Dynamic resource group movement also lets you perform selective maintenance without rebooting the cluster or disturbing operational nodes.

Using the DARE Resource Migration utility does not affect other resource groups that might currently be owned by that node. The node that currently owns the resource group will release it as it would during a “graceful shutdown with takeover”, and the node to which the resource group is being moved will acquire the resource group as it would during a node failover.

The following section covers the types and location keywords used in DARE resource migrations, and also how to use the `cldare` command and the -M flag to perform the migration.

8.5.3.1 Resource Migration Types

Before performing a resource migration, decide if you will declare the migration `sticky` or `non-sticky`.

Sticky Resource Migration

A sticky migration permanently attaches a resource group to a specified node. The resource group attempts to remain on the specified node during a node failover or reintegration.

Since stickiness is a behavioral property of a resource group, assigning a node as a sticky location makes the specified resource group a sticky resource. Older sticky locations are superseded only by new sticky migration requests for the same resource group, or they are removed entirely during non-sticky migration requests for the same resource group. If it is not possible to place a resource group on its sticky location (because that node is down), the normal resource policy is invoked, allowing the resource to migrate according to the takeover priority specified in the resource group's node list.

For both cascading and rotating resource groups, a normal resource policy means that other cluster nodes in the group's node list are consulted at the time the sticky location fails to find the highest-priority node active. After finding the active node, cascading resource groups will continually migrate to the highest-priority node in the group's node list (ultimately residing at the sticky location). Rotating resource groups stay put until the sticky location returns to the cluster.

You can attach the optional keyword `sticky` to any migration you perform, regardless of the resource group configuration (rotating or cascading). However, with very few exceptions, you always use the sticky location for cascading configurations, and do not use it for rotating configurations.

Non-Sticky Resource Migration

Resource groups on nodes not designated sticky are by default transient, non-sticky resources. These resources are temporarily placed on the specified node with the highest priority in the node list until the next failover or reintegration occurs. Non-sticky resources are best suited for use with rotating resource group configurations because of this transient behavior.

Because the normal behavior of cascading resources is to bound back to the highest available node in their node list, non-sticky migrations are usually not the best choice. The one instance in which a non-sticky migration of a cascading resource might make sense is if this resource has the

INACTIVE_TAKEOVER flag set to false and has not yet started because its primary node is down.

In general, however, only rotating resource groups should be migrated in a non-sticky manner. Such migrations are one-time events and occur similar to normal rotating resource group flavors. After migration, the resource group immediately resumes a normal rotating resource group failover policy, but from the new location.

Note

The `cldare` command attempts to perform all requested migrations simultaneously. If, for some reason, the command cannot simultaneously cause all specified resources to be released and cannot simultaneously reacquire them at the new locations, it fails, and no migrations occur.

8.5.3.2 Locations

You can specify the location for a resource group by entering a node name or a keyword.

Node Name

In most cases, you enter a node name in the location field to specify which node will contain sticky or non-sticky resource groups. Node names can be arbitrary and apply to both rotating and cascading resource group configurations.

The DARE Resource Migration utility also provides the following special keywords you can use in the location field to determine the placement of migrated resource groups: `default` and `stop`. The `default` and `stop` locations are special locations that determine resource group behavior and whether the resources can be reacquired.

Default Location

If you use the `default` keyword as the location specifier, the DARE Resource Migration utility removes all previous stickiness for the resource group and returns the resource group to its default failover behavior where node priorities apply (for either cascading or rotating resources). The use of a default destination for a cascading resource group returns it to its normal behavior (the resource group will migrate to the highest priority node currently up). Using a default destination for a rotating resource group releases the group from wherever it resides and lets the highest priority node with a boot address reacquire the resource.

If you do not include a location specifier in the location field, the DARE Resource Migration utility performs a default migration, again making the resources available for reacquisition.

Note

A default migration can be used to start a cascading resource group that has INACTIVE_TAKEOVER set to **false** and that has not yet started because its primary node is down.

Stop Location

The second special location keyword, `stop`, causes a resource group to be made inactive, preventing it from being reacquired, though it remains in the resource configuration. Its resources remain unavailable for reacquisition even after a failover or reintegration.

8.5.3.3 Using the cldare Command to Migrate Resources

The `cldare` command can be used to perform dynamic resource group migrations to other cluster nodes in conjunction with other `cldare` resource functionality. It lets you specify multiple resource groups and nodes on the command line, as long as the final resource group configuration is consistent. After some error checking, the resources are released and reacquired by the specified cluster nodes. Resource migration first releases all specified resources (wherever they reside in the cluster); then it reacquires these resources on the newly specified nodes.

You can also use this command to swap resources on nodes in the resource group’s node list, but you cannot mix keywords—`default, stop`, and `node`—when using the `cldare` command.

To migrate resource groups (and their resources) using the `cldare` command, enter the following command:

```
cldare -M <resgroup name>:[location|[default|stop]][:sticky] ...
```

where `-M` specifies migration, and where resource group names must be valid names of resource groups in the cluster. You can specify a node name (or special location) or the keyword `stop` or `default` after the first colon. The node name must represent a cluster node that is up and in the resource group’s node list. You can specify a migration type after the second colon. Repeat this syntax on the command line for each resource group you want to migrate. Do not include spaces between arguments.

Note that you cannot add nodes to the resource group list with the DARE Resource Migration utility. This task is performed through SMIT.

Stopping Resource Groups

If the location field of a migration contains the keyword `stop` instead of an actual nodename, the DARE Resource Migration utility attempts to stop the resource group, which includes taking down any service label, unmounting file systems, and so on. You should typically supplement the keyword `stop` with the migration type `sticky` to indicate that the resource stays down, even if you reboot the cluster.

As with sticky locations, sticky stop requests are superseded by new sticky migration requests for the same resource group, or they are removed by `default`, non-sticky migration requests for the same resource group. Thus, a stopped resource will be restarted at the time of the next migration request.

> **Note**
>
> Be careful when using a non-sticky stop request, since the resource group will likely be restarted at the next major cluster event. As a result, all non-sticky requests produce warning messages. A non-sticky stop could be used to halt a cascading resource group that has INACTIVE_TAKEOVER set to false during periods in which its primary node is down.

8.5.3.4 Using the clfindres Command

To help you locate resources placed on a specific node, the DARE Resource Migration utility includes a command, `clfindres`, that makes a best-guess estimate (within the domain of current HACMP configuration policies) of the state and location of specified resource groups. It also indicates whether a resource group has a sticky location, and it identifies that location.

See Appendix A of the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279, for the syntax and typical output of the `clfindres` command.

8.5.3.5 Removing Sticky Markers When the Cluster is Down

Sticky location markers are stored in the HACMPresource class in the HACMP ODM and are a persistent cluster attribute. While the cluster is up, you can only remove these locations by performing a subsequent non-sticky migration on the same resource group, using the `default` special location keyword or specifying no location.

Be aware that persistent sticky location markers are saved and restored in cluster snapshots. You can use the `clfindres` command to find out if sticky markers are present in a resource group.

If you want to remove sticky location markers while the cluster is down, the `default` keyword is not a valid method, since it implies activating the resource. Instead, when the cluster is down, you use a transient `stop` request, as in this example:

```
cldare -v -M <resgroup name>:stop
```

(The optional -v flag indicates that verification is skipped.)

8.6 Applying Software Maintenance to an HACMP Cluster

You can install software maintenance, called Program Temporary Fixes (PTFs), to your HACMP cluster while running HACMP for AIX cluster services on cluster nodes; however, you must stop cluster services on the node on which you are applying a PTF. As with everything else in a cluster, applying software fixes should be done in a controlled fashion.

With the method described below, you might even be able to keep your mission-critical application up and running during the update process, provided that the takeover node is designed to carry its own load and the takeover load as well.

The normal method of applying AIX fixes is to do the following:

1. Use the `smit clstop` fastpath to stop cluster services on the node on which the PTF is to be applied. If you would like the resources provided by this node to remain available to users, stop cluster with takeover so that the takeover node will continue to provide these resources to users.
2. Apply the software maintenance to this node using the procedure described in the documentation distributed with the PTF.
3. Run the `/usr/sbin/cluster/diag/clverify` utility to ensure that no errors exist after installing the PTF. Test the fix as thoroughly as possible.
4. Reboot the node to reload any HACMP for AIX kernel extensions that may have changed as a result of the PTF being applied.
   If an update to the cluster.base.client.lib file set has been applied and you are using Cluster Lock Manager or Clinfo API functions, you may need to relink your applications.

5. Restart the HACMP for AIX software on the node using the `smit clstart` fastpath and verify that the node successfully joined the cluster.
6. Repeat Steps 1 through 5 on the remaining cluster nodes.

Figure 15 below shows the procedure:

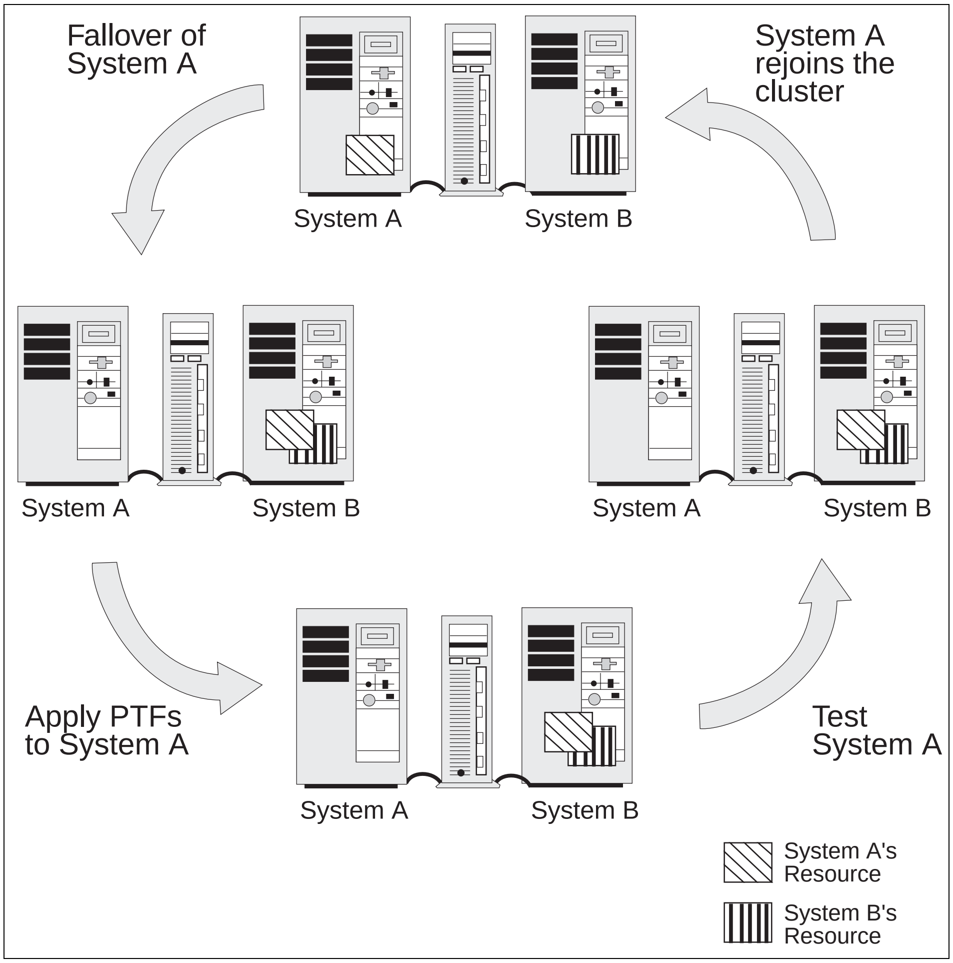


Figure 15. Applying a PTF to a Cluster Node

Along with the normal rules for applying updates, the following general points should be observed for HACMP clusters:

- Cluster nodes should be kept at the same AIX maintenance levels wherever possible. This will, of course, not be true while the update is being applied, but should be true at all other times.

- Cluster nodes should be running the same HACMP maintenance levels. There might be incompatibilities between various maintenance levels of HACMP, so you must ensure that consistent levels are maintained across all cluster nodes. The cluster must be taken down to update the maintenance levels.

8.7 Backup Strategies

HACMP software masks hardware failures in clustered RISC System/6000 environments by quickly switching over to backup machines or other redundant components. However, installing HACMP is not a substitute for having a formal backup and recovery procedure.

In general, a backup of user and system data is kept in case data is accidentally removed or in case of a disk failure. A formal backup process is really an insurance policy. You invest in the technology and time to back up systems so that, in the event of a problem, you can quickly rebuild the system.

Since system and application backups are preferably done during periods of no usage (for instance, in the middle of the night), many installations implement an automated backup procedure using the AIX cron facility. While this is a very good procedure, the HACMP cluster environment presents some special challenges. The problem is, you never know which machine has your application data online, so you need to ensure that exactly the node that has a resource online will initiate the backup of data.

It isn’t actually important which of the several backup commands you are using, what is important is the strategy. For the features and/or restrictions of backup commands like `tar, cpio, dd` or `backup`, refer to the *AIX Commands Reference Version 4.3*, SBOF-1877.

8.7.1 Split-Mirror Backups

No file system can be safely backed up while update activity is occurring. If you are going to have any assurance as to which updates are on the backup and which updates are not, you need to be able to demark exactly where the backup was made. Therefore, it may be difficult to do a good backup on systems that have applications or data that must be online continuously or offline for only a very short time. In some installations, the time required to do a full backup to an archival device, or even to another, might be longer than the availability requirements of the application will allow it to be offline. The mirroring capability of the AIX Logical Volume Manager (LVM) can be used to address this issue.

8.7.1.1 How to do a split-mirror backup

This same procedure can be used with just one mirrored copy of a logical volume. If you remove a mirrored copy of a logical volume (and file system), and then create a new logical volume (and file system) using the allocation map from that mirrored copy, your new logical volume and file system will contain the same data as was in the original logical volume.

Now, you can mount this new file system (read-only is recommended), back it up, and you are really backing up a mirrored copy of the data in the original file system, as it was when we removed the mirror copy. Since this file system, created from the mirror copy, is mounted read-only, no inconsistency in the file system from the point at which you removed the mirror originally is created during the backup. After that, you can delete the new file system to release the physical partitions back to the free pool. Finally, you can add and synchronize a mirror back onto the original file system, and you are back to a mirrored mode of operation, with fully updated data.

The `splitlvcopy` command of AIX does much of the work required to implement this solution.

We can summarize the steps to do a split-mirror backup of a file system as follows:

1. Use the `lsvg -l VGNAME` command to take note of the logical volumet name that contains the file system you want to back up.
2. Stop any application using the file system and unmount the file system.
3. Use the `splitlvcopy` command to break off one mirror of the logical volume, and create a new logical volume with its contents. For example, if the existing logical volume is named fslv, the command would be `splitlvcopy -y newlv fslv`.
4. It is important to note that there is now one less mirror copy available to the user in fslv.
5. Remount the file system and restart the application that was using it.
6. You can see that the application and data are offline for only a very short time.
7. Create a file system on your new logical volume and mount it read-only. This is to ensure that no update activity will occur in the new logical volume, and the consistency of the file system is guaranteed.
8. Perform the backup on the new file system by any means desired, such as `backup`, `tar`, `cpio`, and `pax`.

9. After the backup is complete and verified, unmount and delete the new file system and the logical volume you used for it.

10. Use the `mklvcopy` command to add back the logical volume copy you previously split off to the fslv logical volume.

11. Resynchronize the logical volume.

Once the mirror copy has been recreated on the logical volume, the `syncvg` command will resynchronize all physical partitions in the new copy, including any updates that have occurred on the original copy during the backup process.

It is always a good idea to check a backup for validity.

8.7.2 Using Events to Schedule a Backup

As described above, a `crontab` entry is often used for scheduling nightly backups during off-peak hours of the application. Now as you have several cluster nodes, each of them would need a `crontab` entry, in order to get its own data backed up. This `crontab` entry can determine whether only the “normal” data is backed up, i.e. the data this cluster node cares about during “normal” operations, or, in case of another’s node failure and a subsequent takeover of this node’s resources, backing up both of the cluster nodes’ data.

Whenever one node takes over the reources of another node, the *node_down_remote* event has happened. You can use a post-event to the *node_down_remote* event to change the `crontab` entry from backing up only the local node’s data into backing up both nodes’ data.

Furthermore, if the second node eventually comes up again and takes its resources back, you will see a *node_up_remote* event in your logs. Thus, you can configure a post-event to the *node_up_remote* event to change the crontab entry back to the “normal” setting.

If you want to do a split-mirror backup, the crontab entry has to invoke a script, implementing the steps described above.

A more detailed description of this procedure can be found in the redbook *HACMP/6000 Customization Examples*, SG24-4498, Chapter 6.

8.8 User Management

As 2.7, “User ID Planning” on page 48 described, on an HACMP cluster, the administrator has to take care of user and group IDs throughout the cluster. If they don't match, the user won't get anything done after a failover happened. So, the administrator has to keep definitions equal throughout the cluster.

Fortunately, the C-SPOC utility, as of HACMP Version 4.3 and later, does this for you. When you create a cluster group or user using C-SPOC, it makes sure that it has the same group id or user id throughout the cluster.

8.8.1 Listing Users On All Cluster Nodes

To obtain information about all user accounts on cluster nodes (or about a particular user account), you can either use the AIX `lsuser` command in `rsh` to one cluster node after another, or use the C-SPOC `cl_lsuser` command, or the C-SPOC SMIT List all the Users on the Cluster screen. The `cl_lsuser` command executes the AIX `lsuser` command on each node. To obtain a listing of all user accounts in the cluster, you must specify the `ALL` argument.

If you specify a user name that does not exist on one of the cluster nodes, the `cl_lsuser` command outputs a warning message but continues execution of the command on other cluster nodes.

Note

If you have a Network Information Service (NIS) database installed on any cluster node, some user information may not appear when you use the `cl_lsuser` command.

8.8.2 Adding User Accounts on all Cluster Nodes

Adding a user to the cluster involves three steps:

1. Add an entry for the new user to the /etc/passwd file and other system security files.
2. Create a home directory for the new user.
3. Add the user to a group file.

On AIX systems, you use the `mkuser` command to perform these tasks. This command adds entries for the new user to various system security files, including /etc/passwd and /etc/security/passwd, adds the new user to a group, and creates a home directory for the new user. Every user account has a number of attributes associated with it. When you create a user, the `mkuser` command fills in values for these attributes from the system default /usr/lib/security/mkuser.default file. You can override these default values by specifying an attribute and a value on the `mkuser` command line.

To add a user on one or more nodes in a cluster, you can either use the AIX mkuser command in a rsh to one clusternode after the other, or use the C-SPOC cl_mkuser command or the Add a User to the Cluster SMIT screen. The cl_mkuser command calls the AIX mkuser command to create the user account on each cluster node you specify. The cl_mkuser command creates a home directory for the new account on each cluster node.

8.8.3 Changing Attributes of Users in a Cluster

On AIX systems, you can change any of the attributes associated with an existing user account by using the chuser command. Using the chuser command, you specify the name of the user account you want to change and then specify the attributes with their new values. If you use the SMIT Change User Attributes screen, the complete list of user attributes is displayed and you can supply new values for any attributes. The chuser command modifies the user information stored in the /etc/passwd file and the files in the /etc/security directory.

To change the attributes of a user account on one or more cluster nodes, you can either use the AIX chuser command in rsh to one cluster node after the other, or use the C-SPOC cl_chuser command or the C-SPOC Change User Attributes SMIT screen. The cl_chuser command executes the AIX chuser command on each cluster node.

Note

Do not use the cl_chuser command if you have an NIS (Network Information Service) database installed on any node in your cluster.

Both cluster nodes must be active and a user with the specified name must exist on both the nodes for the change operation to proceed. Optionally, you can specify that the cl_chuser command continue processing if the specified user name exists on any of the cluster nodes. See the cl_chuser command man page for more information.

8.8.4 Removing Users from a Cluster

On AIX systems, you remove a user account by using the rmuser command or the SMIT Remove a User From the System screen. Using the rmuser command you specify the name of the user account you want to remove and specify whether you want the user password and other authentication information removed from the /etc/security/passwd file.

To remove a user account from one or more cluster nodes, you can either use the AIX `rmuser` command on one cluster node after the other, or use the C-SPOC `cl_rmuser` command or the C-SPOC Remove a User from the Cluster SMIT screen. The `cl_rmuser` command executes the AIX `rmuser` command on all cluster nodes.

> **Note**
>
> The system removes the user account but does not remove the home directory or any files owned by the user. These files are only accessible to users with root authority or by the group in which the user was a member.

8.8.5 Managing Group Accounts

In order to manage a number of similar users as a single entity, AIX provides the administrator with the group concept. Members of one group share the same permissions, the same attributes and limits, and so on.

Commands for managing group accounts are just like the user managing commands very much alike to the native AIX commands. The restrictions on NIS are just the same as for users, and therefore are not explained here in detail.

For more detailed information, please refer to Chapter 12 of the *HACMP for AIX, Version 4.3: Administration Guide*, SC23-4279.

8.8.6 C-SPOC Log

Because these commands are running and executing while distributed amongst the cluster, it could happen that something doesn’t work exactly like it should. The C-SPOC utility, therefore, maintains a log on the initiating node. It can be found under `/tmp/cspoc.log`.

Note that the initiating node doesn’t have to be the same in all cases, so the log file might be present on different cluster nodes, and doesn’t contain the same data.

Chapter 9. Special RS/6000 SP Topics

This chapter will introduce you to some special topics that only apply if you are running HACMP on the SP system.

9.1 High Availability Control Workstation (HACWS)

If you are thinking about what could happen to your SP whenever the Control Workstation might fail, you will probably think about installing HACWS for that. These paragraphs will not explain HACWS in full detail, but will concentrate on the most important issues for installation and configuration. For more details, refer to Chapter 3, “Installing and Configuring the High Availability Workstation”, in the *IBM Parallel System Support Programs for AIX Installation and Migration Guide*, GA22-7347, or to Chapter 4,”Planning for a High Availability Workstation”, in the *IBM RS/6000 SP Planning Volume 2, Control Workstation and Software Environment*, GA22-7281.

Some services of the control workstation (or cws for short) are vital, so a failure would impact your ability to manage an SP system. Also, the failure of the control workstation could cause the switch network to fail. HACWS covers the following cases with a fully functional environment:

- Continues running your SP system after a cws failure
- Shuts down the cws for deferred hardware and software maintenance without having a system outage
- Maintains the SP system function and reliability when the cws fails
- Fails over the cws to a backup

9.1.1 Hardware Requirements

To build a cluster consisting of two control workstations, you have to think about shared resources. The spdata file system holding the SDR data and other vital data has to be accessible from both control workstations, so, it has to be put onto a shared disk.

The cws connects to the frames of an RS/6000 SP with RS232 lines as its supervisor network. If the RS/6000 SP consists of multiple frames, you will probably have an 8-port adapter installed in the Control Workstation in order to provide the needed number of ttys.

To connect a backup cws to the frames, you need exactly the same tty port configuration as on the primary cws, that is, when frame 3 connects to tty3 on the primary cws, it has to connect to tty3 on the backup cws as well. Also, you need to have the frame supervisors support dual tty lines in order to get both control workstations connected at the same time. Contact your IBM representative for the neccessary hardware (see Figure 16 on page 184).

Both the tty network and the RS/6000 SP internal ethernet are extended to the backup cws. In contrast to standard HACMP, you don't need to have a second ethernet adapter on the backup cws. In case you have only one, the HACWS software will work with ip aliasing addresses on one adapter.

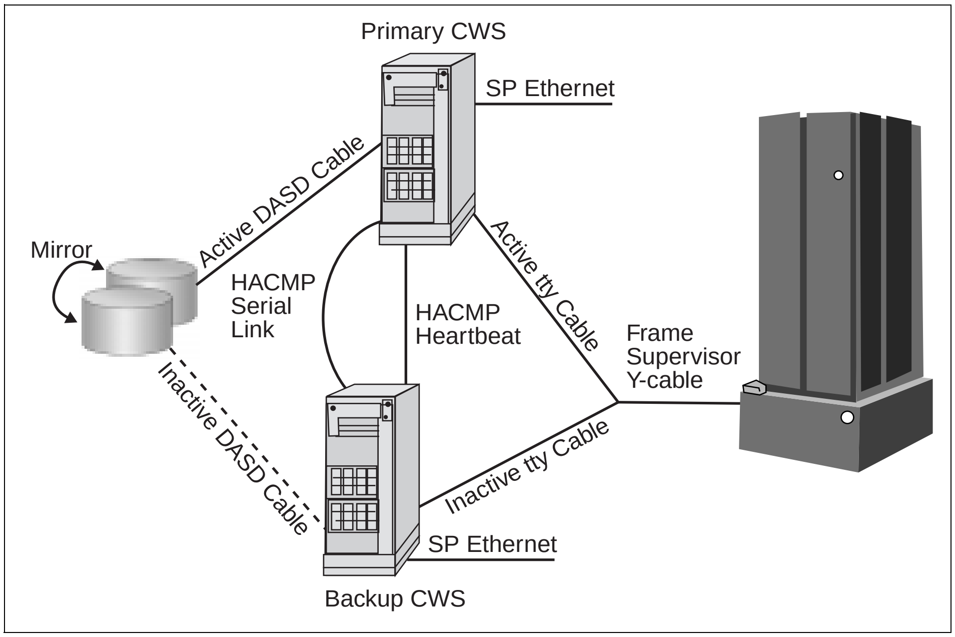


Figure 16. A Simple HACWS Environment

9.1.2 Software Requirements

Both of the control workstations must have the same software installed, that is, they must be on the same AIX level, use the same PSSP software level and have to have HACMP on the same level as well. For example, if you want to use HACMP 4.3 for AIX, you have to use PSSP 3.1 and therefore AIX Version 4.3.2 on the primary cws and on the backup cws.

9.1.3 Configuring the Backup CWS

The primary cws is configured exactly as usual, as far as the AIX and PSSP software is concerned, as if there were no HACWS at all.

The backup cws has to be installed with the same level of AIX and PSSP. Depending on the kerberos configuration of the primary cws, the backup cws has to be configured either as a *secondary authentication server* for the authentication realm of your RS/6000 SP when the primary cws is an authentication server itself, or as an *authentication client* when the primary cws is an authentication client of some other server. To do so will enable a correct kerberos environment on the backup cws; so, remote commands will succeed through kerberos authentication as on the primary cws.

After the initial AIX and PSSP setup is done, the HACWS software has to be installed.

9.1.4 Install High Availability Software

On both control workstations, the HACMP software has to be installed now, according to the instructions in the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278. Verification, as described in Chapter 10 of the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278, should be performed. For HACWS control workstations, the ssp.hacws fileset has to be installed as well.

9.1.5 HACWS Configuration

Since the cws might have some daemons active that could interfere with the definition and configuration of networks, you have to stop them, in order to get the configuration done, with the command:

```
/usr/sbin/hacws/spcw_apps -d
```

This will stop the subsystems `spmgr splogd hardmon sysctld supfilesrv sp_configd` if they have been active with the corresponding SRC command.

Now configure the serial network. You can either use target mode SCSI, target mode SSA or the raw RS-232 serial line, or any combination.

Both machines, primary and backup, need to be configured to boot up on their boot address in order to not confuse a working cws at boot time of the backup cws.

If not previously done, you have to migrate the /spdata file system to an external volume group, to make it accessible from both sides.

After the /spdata file system is set up so that a `varyonvg` of its vg will work on either cws, you have to complete the Administration Tasks, like on an ordinary HACMP cluster, as it is described in Chapter 7 of the *HACMP for AIX, Version 4.3: Installation Guide*, SC23-4278.

Now the cluster environment has to be configured. Define a cluster ID and name for your HACWS cluster and define the two nodes to HACMP.

Adapters have to be added to your cluster definition as described before. You will have to add a boot adapter and a service adapter for both primary and backup cws. Now that the cluster topology is defined to HACMP, you can configure Network Modules as in the *HACMP for AIX Installation Guide*, SC23-4278, and synchronize the definitions across the two cluster nodes.

With the ssp.hacws filesets comes a predefined start- and stop-script, which has to be defined to HACMP as part of the application server definition, which in turn has to be included in a resource group.

Recommended settings for this resource group are:

Resource Group Name	[hacws_group1]
Node Relationship	[rotating]
Participating Node Names	[“nodename of primary cws” “nodename of backup cws”]
Service IP label	at least the hostname of the primary cws
File System	the name of the file system, most probably /spdata
Volume Groups	the name of the shared volume group containing /spdata
Application Servers	the name you gave the application server before

9.1.6 Setup and Test HACWS

Both the primary and backup cws have to be addressable by their hostname, in order to finish the configuration and check that everything is in order. So, check if the primary cws can address the backup cws by its hostname and vice versa. If not, use the `ifconfig` command to temporarily set the interface to the hostname on each cws. Do *NOT* use `smit chinet` for this, since this would be a permanent change.

Run the command:

```
/usr/sbin/hacws/install_hacws -p primary_hostname -b backup_hostname -s
```

on the primary cws to set up HACWS with the 2 node names.

After that, identify the HACWS event scripts to HACMP by executing the

```
/usr/sbin/hacws/spcw_addevents
```

command, and verify the configuration with the

```
/usr/sbin/hacws/hacws_verify
```

command. You should also check the cabling from the backup cws with the

```
/usr/sbin/hacws/spcw_verify_cabling
```

command. Then reboot the primary and the backup cws, one after the other, and start cluster services on the primary cws with `smit clstart`. After cluster services is up and running, check that control workstation services, such as `SDRGetObjects`, are working as expected. If everything is fine, start up cluster services on the backup cws as well. Check for the completion of the cluster services startup with the following command:

```
grep "SPCW_APPS COMPLETE" /tmp/hacmp.out
```

Now you can cause a failover by stopping cluster services on the primary cws and see whether cws services are still available afterwards.

9.2 Kerberos Security

To understand security, we have to clarify some definitions first.

Identification is the process by which an entity tells another who it is.

Authentication is the process by which the other entity verifies this identity.

Authorization is the process performed by an entity to check if an agent, whose identity has previously been authenticated, has or does not have the necessary privileges to carry out some action.

Additionally, if information is transferred over an insecure network, as any TCP/IP network basically is, there is always a chance that someone is listening, so some sort of encryption is required.

These issues are solved with kerberos.

> **Kerberos**
>
> Also spelled Cerberus - The watchdog of Hades, whose duty was to guard the entrance (against whom or what does not clearly appear); it is known to have had three heads.
>
> - Ambrose Bierce, *The Enlarged Devil's Dictionary*

The following is simply a shortened description on how kerberos works. For more details, the redbook *Inside the RS/6000 SP*, SG24-5145, covers the subject in much more detail.

When dealing with authentication and Kerberos, three entities are involved: the *client*, who is requesting service from a *server*; the second entity, and the *Key Distribution Center* or *Kerberos server*, which is a machine that manages the database, where all the authentication data is kept and maintained.

Kerberos is a third-party system used to authenticate users or services that are known to Kerberos as *principals*. The very first action to take regarding Kerberos and principals is to register the latter to the former. When this is done, Kerberos asks for a principal's password, which is converted to a principal (user or service) 56-bit key using the DES (Data Encryption Standard) algorithm. This key is stored in the Kerberos server database.

When a client needs the services of a server, the client must prove its identity to the server so that the server knows to whom it is talking.

Tickets are the means the Kerberos server gives to clients to authenticate themselves to the service providers and get work done on their behalf on the services servers. Tickets have a finite life, known as the ticket life span.

In Kerberos terms, to make a Kerberos authenticated service provider work on behalf of a client is a three-step process:

- Get a ticket-granting ticket.
- Get a service ticket.
- Get the work done on the service provider.

The main role of the ticket-granting ticket service is to avoid unnecessary password traffic over the network; so, the user should issue his password only once per session. What this ticket-granting ticket service does is to give the client systems a ticket that has a certain time span, whose purpose is to allow the clients to get service tickets to be used with other servers without the need to give them the password every time they request services.

So, given a user has a ticket-granting ticket, if a user requests a kerberized service, he has to get a service ticket for it. In order to get one, the kerberized command sends an encrypted message, containing the requested service name, the machine's name, and a time-stamp to the Kerberos server. The Kerberos server decrypts the message, checks whether everything is in order, and if so, sends back a service ticket encrypted with the service's private key, so that only the requested service can decrypt it. The client sends his request along with the just received ticket to the service provider, who in turn decrypts and checks authorization, and then, if it is in order, provides the requested service to the client.

9.2.1 Configuring Kerberos Security with HACMP Version 4.3

With HACMP Version 4.3 there is a handy script to do the kerberos setup for you, called `cl_setup_kerberos`. It sets up all the IP labels defined to the HACMP cluster together with the needed kerberos principals, so that remote kerberized commands will work.

On an SP the `setup_authent` command does the SP-related kerberos setup, which is based on the IP labels found in the SDR. Since the SDR does not allow multiple IP labels to be defined on the same interface, whereas HACMP needs to have multiple IP labels on one interface during IPAT, the kerberos setup for HACMP has to be redone, every time the `setup_authent` command is run explicitly or implicitly through the `setup_server` command.

You can either do that manually, or use the `cl_setup_kerberos` tool. To manually add the kerberos principals, use the `kadmin` command. Necessary principals for kerberized operation in enhanced security mode are the (remote) rcmd principals and the godm principals. As always, a kerberos principal consists of a name, godm for example, an IP label, like hadave1_stby and a realm, so that the principal in its full length would look like godm.hadave1_stby@ITSO.AUSTIN.IBM.COM.

Now after adding all the needed principals to the kerberos database, you must also add them to the /etc/krb-srvtab file on the nodes. To do that, you will have to extract them from the database and copy them out to the nodes, replacing their kerberos file.

Now you can extend root's .klogin file and /etc/krb.realms file to reflect the new principals, and copy these files out to the node as well.

After setting the cluster's security settings to enhanced for all these nodes, you can verify that it is working as expected, for example, by running clverify, which goes out to the nodes and checks the consistency of files.

9.3 VSDs - RVSDs

VSDs (Virtual Shared Disks) and RVSDs (Recoverable Virtual Shared Disks) are SP-specific facilities that you are likely to use in an HACMP environment.

9.3.1 Virtual Shared Disk (VSDs)

Virtual Shared Disk (VSD) allows data in logical volumes on disks physically connected to one node to be transparently accessed by other nodes. Importantly, VSD supports only raw logical volumes, not file systems. The VSD facility is included in the ssp.csd.vsd fileset of PSSP.

IBM developed VSD to enable Oracle's parallel database on the SP. Oracle's database architecture is strongly centralized. Any processing element, or node, must be able to "see" the entire database. In the case of the parallel implementation of Oracle, all nodes must have access to all disks of the database, regardless of where those disks are physically attached.

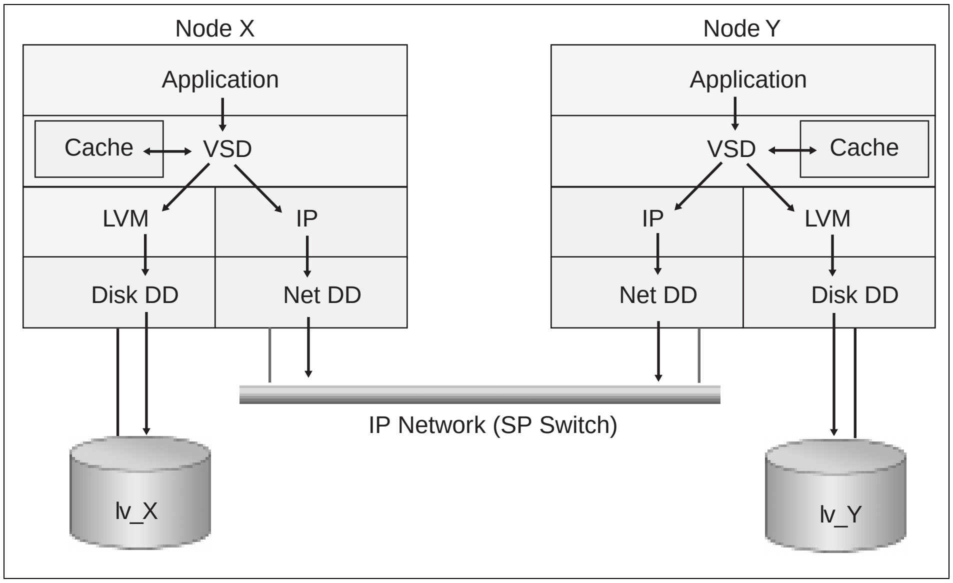


Figure 17. VSD Architecture

With reference to Figure 17 above, imagine two nodes, Node X and Node Y, running the same application. The nodes are connected by the switch and have locally-attached disks. On Node X's disk resides a volume group containing the raw logical volume lv_X. Similarly, Node Y has lv_Y. For the sake of illustration, let us suppose that lv_X and lv_Y together constitute an Oracle Parallel Server database to which the application on each node makes I/O requests.

The application on Node X requests a piece of data in the database. After the node's Virtual Memory Manager (VMM) determines that the data is not in memory, it talks not to the regular Logical Volume Manager (LVM), but rather to the VSD device driver. The VSD device driver is loaded as a kernel extension. Thus VSDs configured in the SP are known to the appropriate nodes at the kernel level.

The VSD device driver can fetch the data from one of three places:

1. From the VSD cache, if the data is still there from previous requests. VSD cache is shared by all VSDs configured on a node. Data is stored in 4KB blocks, a size optimized for Oracle Parallel Server. If your I/O patterns involve I/O operations larger than 4KB, we recommend disabling VSD cache, because its management becomes counterproductive.
2. From lv_X, in which case the VSD device driver exploits Node X's normal LVM and Disk Device Driver (Disk DD) pathway to fetch the data.
3. From lv_Y, in which case the VSD device driver issues the request through the IP and Network Device Driver (Net DD) pathway to access Node Y. For performance, VSD uses its own stripped-down IP protocol. Once the request is passed up through Node Y's Net DD and IP layers, Node Y's VSD device driver accesses the data either from VSD cache or from lv_Y.

The VSD server node uses the *buddy buffer* to temporarily store data for I/O operations originating at a client node, and to handle requests that are greater than the IP message size. In contrast to the data in the cache buffer, the data in a buddy buffer is purged immediately after the I/O operation completes. Buddy buffers are used only when a shortage in the switch buffer pool occurs, or, on certain networks with small IP message sizes (for example, Ethernet). The maximum and minimum size for the buddy buffer must be defined when the VSD is created. For best performance, you must ensure that your buddy buffer limits accommodate your I/O transaction sizes to minimize the packetizing workload of the VSD protocol. Buddy buffers are discussed in detail in *IBM Parallel System Support Programs for AIX Managing Shared Disks*, SA22-7279.

The VSDs in this scenario are mapped to the raw logical volumes lv_X and lv_Y. Node X is a client of Node Y's VSD, and vice versa. Node X is also a direct client of its own VSD (lv_X), and Node Y is a direct client of VSD lv_Y. VSD configuration is flexible. An interesting property of the architecture is that a node can be a client of any other node's VSD(s), with no dependency on that client node owning a VSD itself. You could set up three nodes with powerful I/O capacity to be VSD servers, and ten application nodes, with no disk other than for AIX, PSSP, and the application executables, as clients of the VSDs on these server nodes.

VSDs are defined in the SDR and managed by either SP SMIT panels or the VSD Perspective. VSDs can be in one of five states as shown in Figure 18 on page 192.

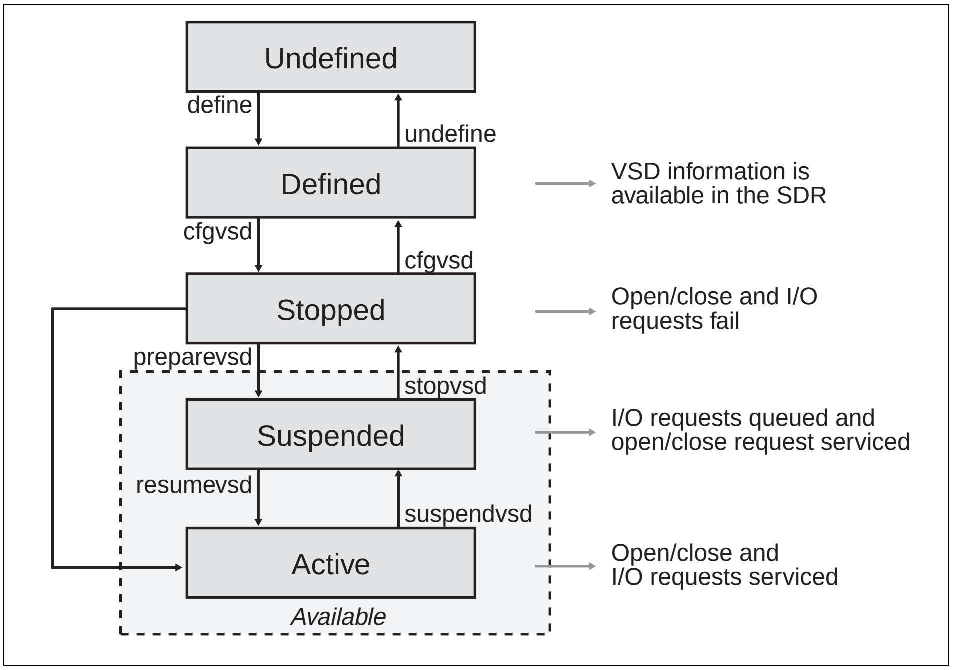

Undefined
define
undefine
Defined
VSD information is available in the SDR
cfgvsd
cfgvsd
Stopped
Open/close and I/O requests fail
preparevsd
stopvsd
Suspended
I/O requests queued and open/close request serviced
resumevsd
suspendvsd
Active
Open/close and I/O requests serviced
Available

Figure 18. VSD State Transitions

This figure shows the possible states of a VSD and the commands used to move between states. VSD configuration changes, or manual recovery of a failed VSD, require you to move the VSD between various states.

The distributed data access aspect of VSD scales well. The SP Switch itself provides a very high-bandwidth, scalable interconnect between VSD clients and servers, while the VSD layers of code are efficient. The performance impact of servicing a local I/O request through VSD relative to the normal VMM/LVM pathway is very small. IBM supports any IP network for VSD, but we recommend the switch for performance.

VSD provides distributed data access, but not a locking mechanism to preserve data integrity. A separate product such as Oracle Parallel Server must provide the global locking mechanism.

9.3.2 Recoverable Virtual Shared Disk

Recoverable Virtual Shared Disk (RVSD) adds availability to VSD. RVSD allows you to twin-tail disks, that is, physically connect the same group of disks to two or more nodes, and provide transparent failover of VSDs among the nodes. RVSD is a separately-priced IBM LPP.

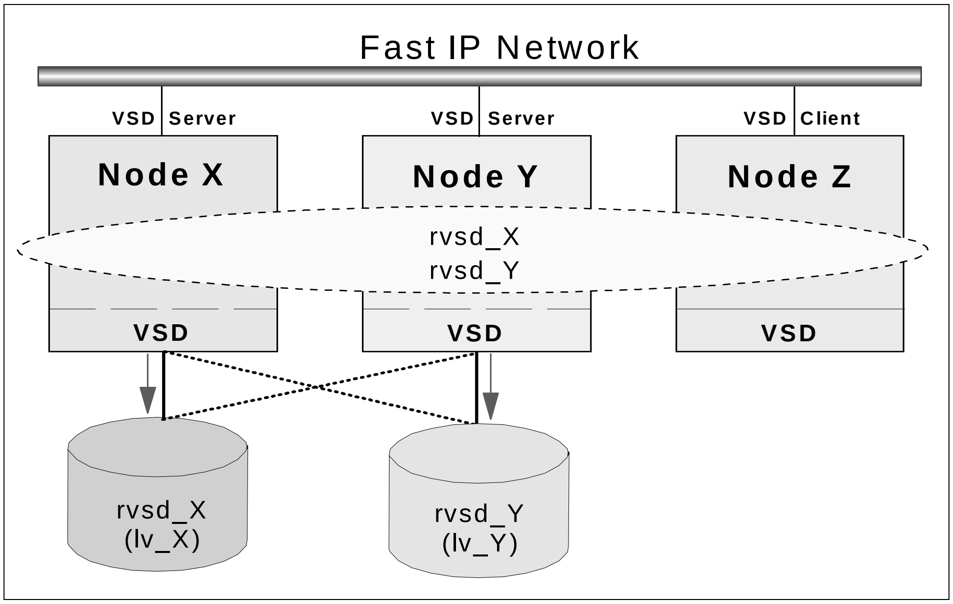


Figure 19. RVSD Function

With reference to Figure 19 above, Nodes X, Y, and Z form a group of nodes using VSD. RVSD is installed on Nodes X and Y to protect VSDs rvsd_X and rvsd_Y. Nodes X and Y physically connect to each other’s disk subsystems where the VSDs reside. Node X is the primary server for rvsd_X and the secondary server for rvsd_Y, and vice versa for Node Y. Should Node X fail, RVSD will automatically fail over rvsd_X to Node Y. Node Y will take ownership of the disks, varyon the volume group containing rvsd_X and make the VSD available. Node Y then serves both rvsd_X and rvsd_Y. Any I/O operation that was in progress, as well as new I/O operations against rvsd_X, are suspended until failover is complete. When Node X is repaired and rebooted, RVSD switches the rvsd_X back to its primary, Node X.

The RVSD subsystems are shown in Figure 20 on page 194. The rvsd daemon controls recovery. It invokes the recovery scripts whenever there is a change in the group membership, which it is recognizing through the use of Group Services, which in turn relies on information from Topology Services. When a failure occurs, the rvsd daemon notifies all surviving providers in the RVSD node group, so they can begin recovery. Communication adapter failures are treated the same as node failures.

The hc daemon is also called the Connection Manager. It supports the development of recoverable applications. The hc daemon maintains a membership list of the nodes that are currently running hc daemons and an incarnation number that is changed every time the membership list changes. The hc daemon shadows the rvsd daemon; recording the same changes in state and management of VSD that rvsd records. The difference is that hc only records these changes after rvsd processes them, to assure that RVSD recovery activities begin and complete before the recovery of hc client applications takes place. This serialization helps ensure data integrity.

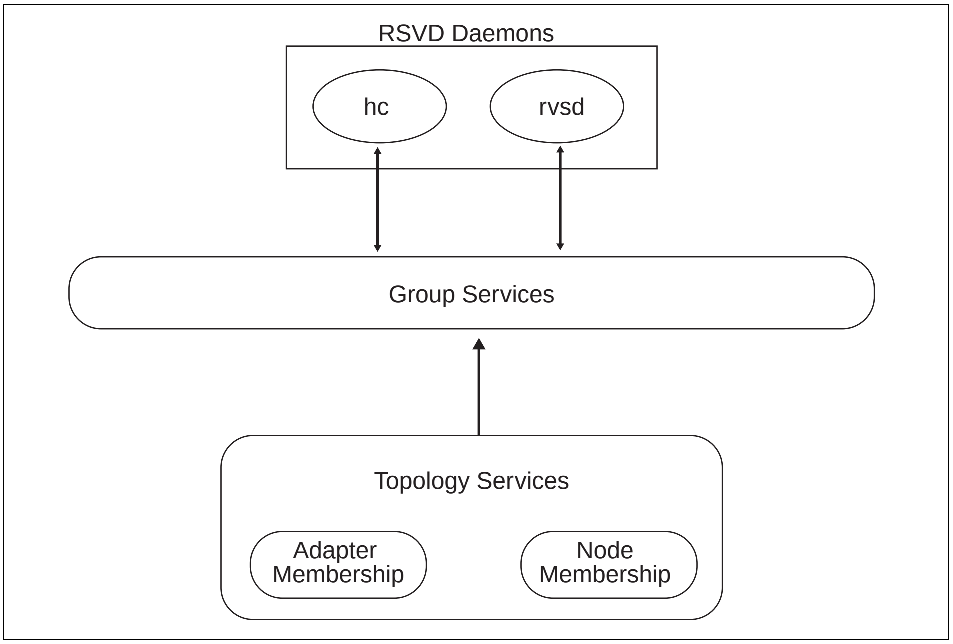


Figure 20. RVSD Subsystem and HA Infrastructure

9.4 SP Switch as an HACMP Network

One of the fascinating things with an RS/6000 SP is the switch network. It has developed over time; so, currently there are two types of switches at customer sites. The “older” HPS or HiPS switch (High Performance Switch), also known as the TB2 switch, and the “newer” SP Switch, also known as the TB3 switch.

The HPS switch is no longer supported with PSSP Version 3.1, and the same applies to HACMP/ES Version 4.3.

The two different types of switches differ in their availability design from the hardware point of view significantly. For example, any fault service action on the HPS switch caused a total network disruption for a small fraction of time. For example, running an Estart to get new nodes up on the switch affected running nodes.

The SP switch however, was designed to do all actions regarding the switch fabric on the link level, so only the selected node is affected. All the others continue working without even noticing that something has happened on the switch network.

9.4.1 Switch Basics Within HACMP

Although it has already been mentioned in other places, the following is a short summary of basics you have to remember when you configure a switch as a network to HACMP.

- As the switch network is a point-to-point network, you must configure it to HACMP as a *private* network.
- For IPAT to work on the switch network, you must enable ARP on the switch network. However, hardware address takeover is not supported with the switch.
- If you configure IPAT, the service and boot addresses are ifconfig alias addresses on the `css0` adapter. Since there is currently no support for more than one switch adapter in a node, this is the way HACMP covers the normally-needed second adapter for redundancy.
- The base address for the switch adapter, i.e. the switch address known to the SDR, should not be configured into an HACMP network. This would lead to confusion for the PSSP switch management software.
- The netmask associated with the css0 base IP address is used as the netmask for all HACMP SP Switch network addresses.

9.4.2 Eprimary Management

The SP switch has an internal primary backup concept, where the primary node, known as the Eprimary, is backed up automatically by a backup node. So, in case any serious failure happens on the primary, it will resign from work, and the backup node will take over the switch network handling, keeping track of routes, working on events, and so on.

HACMP/ES used to have an Eprimary management function with versions below 4.3; so, if you upgrade to Version 4.3 and also upgrade your switch to the SP switch, and you had configured Eprimary management previously within the HACMP definitions, you have to unmanage it.

To check whether the Eprimary is set to be managed, issue the following command:

```
odmget -q'name=EPRIMARY' HACMPsp2
```

If the switch is set to MANAGE, before changing to the new switch, run the script:

```
/usr/es/sbin/cluster/events/utils/cl_HPS_Eprimary unmanage
```

As the SP switch has its availability concept built-in, there is no need to do it outside the PSSP software, so, HACMP doesn’t have to take care of it any more.

9.4.3 Switch Failures

As mentioned before, a node in the SP is still restricted to have a maximum of one switch adapter installed. Therefore, even with the software being able to assign a new primary node within the SP and outside of HACMP, the switch adapter is still a single point of failure.

If the switch adapter in a node resigns from work due to a software or hardware problem, the switch network is down for that node.

If any application running on that node relies on the switch network, this means that the application has virtually died on that node. Therefore, it might be advisable to promote the switch network failure into a node failure, as described in 2.6.2.1, “Single Point-of-Failure Hardware Component Recovery” on page 46. HACMP would be able to recognize the network failure when you configure the switch network as an HACMP network, and thus would react with a *network_down* event, which in turn would shut down the node from HACMP, causing a takeover.

In case this node was the Eprimary node on the switch network, and it is an SP switch, then the RS/6000 SP software would have chosen a new Eprimary independently from the HACMP software as well.

Chapter 10. HACMP Classic vs. HACMP/ES vs. HANFS

So, why would you prefer to install one version of HACMP instead of another? This chapter summarizes the differences between them, to give you an idea in which situation one or the other best matches your needs. The certification test itself does not refer to these different HACMP flavors, but it is useful to know the differences anyway.

The following paragraphs are based on the assumption that you are using Version 4.3. For an overview of previous Versions and their corresponding AIX levels, as well as the supported hardware to run on, see Table 3 on page 8.

10.1 HACMP for AIX Classic

High Availability Cluster Multi-Processing for AIX (HACMP for AIX) Version 4.3 comes in two flavors. One of them directly derives from previous versions, and therefore is called Classic, and the other, which utilizes another technology for heartbeating, is called HACMP Extended Scalability (HACMP/ES) see below for details.

Basically, these two versions differ only in the way the cluster manager keeps track of the status of nodes, adapters and networks. In the Classic Version, this is done through the use of Network Interface Modules.

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 neighboring 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.

10.2 HACMP for AIX / Enhanced Scalability

HACMP/ES no longer used NIMs, but utilizes a technolgy that was originally developed on the RS/6000 SPs. Since PSSP Version 2.2. RS/6000 SP Systems come with the Phoenix technology for managing availability of the nodes. This technology was already designed as a basic instrument for handling membership and event management by using heartbeats. On the SP, the original High Availability infrastructure was built on this technology, and HACMP/ES Version 4.3. is now another instance relying on it. As of AIX 4.3.2 and PSSP 3.1, the High Availability infrastructure, which previously was tightly coupled to PSSP, was externalized into a package called RISC System Cluster Technology (RSCT). This package can be installed and run, not only on SP nodes, but also on regular RS/6000 systems. This allows HACMP/ES to also be available on non-SP RS/6000s as of Version 4.3.

10.2.1 IBM RISC System Cluster Technology (RSCT)

The High Availability services previously packaged with the IBM PSSP for AIX Availability Services, also known as the ssp.ha fileset, are now an integral part of the HACMP/ES software. The IBM RS/6000 Cluster Technology (RSCT) services provide greater scalability, notify distributed subsystems of software failure, and coordinate recovery and synchronization among all subsystems in the software stack.

Packaging these services with HACMP/ES makes it possible to run this software on all RS/6000s, not just on SP nodes.

RSCT Services include the following components:

Event Manager A distributed subsystem providing a set of high availability services. It creates events by matching information about the state of system resources with information about resource conditions of interest to client programs. Client programs, in turn, can use event notifications to trigger recovery from system failures.

Group Services A system-wide, fault-tolerant, and highly available facility for coordinating and monitoring changes to the state of an application running on a set of nodes. Group Services helps both in the design and implementation of fault-tolerant applications and in the consistent recovery of multiple applications. It accomplishes these two distinct tasks in an integrated framework.

Topology Service A facility for generating heartbeats over multiple networks and for providing information about adapter membership, node membership, and routing. Adapter and node membership provide indications of adapter and node failures respectively. Reliable Messaging uses the routing information to route messages between nodes around adapter failures.

See Part 4 of *HACMP for AIX, Version 4.3: Enhanced Scalability Installation and Administration Guide*, SC23-4284, for more information on these services.

10.2.2 Enhanced Cluster Security

With HACMP Version 4.3 comes an option to switch security Mode between Standard and Enhanced.

Standard Synchronization is done through the `/.rhosts` remote command facilities. To avoid the compromised security that the presence of this file presents, the administrator is strongly encouraged to remove these files after the synchronization/verification is done.

Enhanced Kerberos authentication is used for remote commands. That means the kerberos daemons can decide whether a remote host is who they claim to be. This is done by granting access on the basis of tickets, which are provided only to those hosts having the correct identification.

10.3 High Availability for Network File System for AIX

The HANFS for AIX software provides a reliable NFS server capability by allowing a backup processor to recover current NFS activity should the primary NFS server fail.

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

HANFS for AIX is based on High Availability Cluster Multi-Processing for AIX, Version 4.3 (HACMP for AIX Classic) product architecture, which ensures that critical resources, configured as part of a cluster, are highly available for processing. The HANFS for AIX software extends HACMP for AIX by taking advantage of AIX extensions to the standard NFS functionality that enable it to handle duplicate requests correctly and restore lock state during NFS server failover and reintegration.

Note

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

10.4 Similarities and Differences

All three products have the basic structure in common. They all use the same concepts and structures. So, a cluster or a network, in the HACMP context, is the same, no matter what product is being used. There is always a Cluster Manager controlling the node, keeping track of the cluster’s status, and triggering events. The differences are in the technologies being used underneath, or in some special cases, the features available.

The technique of keeping track of the status of a cluster by sending and receiving heartbeat messages is the major difference between HACMP Classic and HACMP/ES Version 4.3. HACMP Classic uses the network modules (NIMs) for this purpose. These communicate their results straight through to the HACMP Cluster Manager. HACMP/ES, uses the facilities of RSCT, namely Topology Services, Group Services, and Event Management, for its heartbeating. Since Version 4.3, the restriction to run HACMP/ES on RS/6000 SP systems only has been withdrawn. However, if you run it on an RS/6000 SP, you need to have PSSP Version 3.1 installed. As the HPS Switch is no longer supported with PSSP Version 3.1, you need to upgrade to the SP Switch, in case you haven’t already, or you will have a switchless system.

You can still run HACMP Classic on RS/6000 SP Nodes, just as on standalone RISC System/6000s. It has no references into the PSSP code whatsoever.

HANFS for AIX Version 4.3 is basically, a modified HACMP Classic, enhanced with the capability of the takeover node to recover current NFS activity, should the primary NFS server fail. By means of AIX extensions to standard NFS functionality, HANFS for AIX is enabled to handle duplicate requests correctly or restore the lock state in case of an NFS server failover or reintegration. Remember though, that HANFS is somewhat restricted, in that it only supports two-node clusters and cascading resource groups.

10.5 Decision Criteria

Your decision of what type of high availability software you are going to use can be based on various criteria. Existing hardware is one of them.

If you still use the “old” HPS Switch and don’t want to lose its functionality, you are bound to the use of PSSP 2.4 or lower. Therefore HACMP Classic or HACMP/ES up to Version 4.2.2 only is the choice for you.

For switchless RS/6000 SP systems or SPs with the newer SP Switch, the decision will be based on a more functional level.

Event Management is much more flexible in HACMP/ES, since you can define custom events. These events can act on anything that `haemd` can detect, which is virtually anything measurable on an AIX system. How to customize events is explained in great detail in the redbook *HACMP Enhanced Scalability*, SG24-2081.

If you have an NFS server that you need to make highly available, especially if it is heavily used, and NFS file locking is a major issue, you will need to run HANFS for AIX.

Appendix A. Special Notices

This publication is intended to help System Administrators, System Engineers and other System Professionals to pass the IBM HACMP Certification Exam. The information in this publication is not intended as the specification for any of the following programming interfaces: HACMP, HACMP/ES, HANFS or HACWS. See the PUBLICATIONS section of the IBM Programming Announcement for those products for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594 USA.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The information about non-IBM ("vendor") products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

Any performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

The following document contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples contain the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

AIX
Application System/400
AS/400
AT
BookManager
CT
HACMP/6000
Home Director
IBM ®
Micro Channel
NetView
POWERparallel
POWERserver
RISC System/6000
RS/6000
SP
SP1
System/390
Ultrastar
XT
400

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc.

Java and HotJava are trademarks of Sun Microsystems, Incorporated.

Microsoft, Windows, Windows NT, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

Pentium, MMX, ProShare, LANDesk, and ActionMedia are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

Network File System and NFS are trademarks of SUN Microsystems, Inc.

SUN Microsystems is a trademark of SUN Microsystems, Inc.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

Other company, product, and service names may be trademarks or service marks of others.

Appendix B. Related Publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

B.1 International Technical Support Organization Publications

For information on ordering these ITSO publications see “How to Get ITSO Redbooks” on page 211.

- *A Practical Guide to Serial Storage Architecture for AIX,* SG24-4599
- *HACMP Enhanced Scalability,* SG24-2081
- *HACMP Enhanced Scalability Handbook*, SG24-5328
- *HACMP Enhanced Scalability: User-Defined Events*, SG24-5327
- *HACMP/6000 Customization Examples,* SG24-4498
- *High Availability on the RISC System/6000 Family,* SG24-4551
- *Inside the RS/6000 SP,* SG24-5145
- *Monitoring and Managing IBM SSA Disk Subsystems*, SG24-5251
- *AIX Version 4.3 Migration Guide*, SG24-5116

B.2 Redbooks on CD-ROMs

Redbooks are also available on CD-ROMs. **Order a subscription** and receive updates 2-4 times a year at significant savings.

CD-ROM Title	Subscription Number	Collection Kit Number
System/390 Redbooks Collection	SBOF-7201	SK2T-2177
Networking and Systems Management Redbooks Collection	SBOF-7370	SK2T-6022
Transaction Processing and Data Management Redbook	SBOF-7240	SK2T-8038
Lotus Redbooks Collection	SBOF-6899	SK2T-8039
Tivoli Redbooks Collection	SBOF-6898	SK2T-8044
AS/400 Redbooks Collection	SBOF-7270	SK2T-2849
RS/6000 Redbooks Collection (HTML, BkMgr)	SBOF-7230	SK2T-8040
RS/6000 Redbooks Collection (PostScript)	SBOF-7205	SK2T-8041
RS/6000 Redbooks Collection (PDF Format)	SBOF-8700	SK2T-8043
Application Development Redbooks Collection	SBOF-7290	SK2T-8037

B.3 Other Publications

These publications are also relevant as additional sources of information:

- *IBM RS/6000 SP: Planning, Volume 2, Control Workstation and Software Environment*, GA22-7281
- *IBM PSSP for AIX: Installation and Migration Guide*, GA22-7347
- *IBM PSSP for AIX: Managing Shared Disks*, SA22-7279
- *Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516
- *Adapters, Devices, and Cable Information for Micro Channel Bus Systems*, SA38-0533
- *PCI Adapter Placement Reference*, SA38-0538
- *AIX Commands Reference*, SBOF-1877
- *AIX Performance Monitoring and Tuning Guide*, SC23-2365
- *AIX HACMP for AIX,Version 4.3: Concepts and Facilities*, SC23-4276
- *AIX HACMP for AIX,Version 4.3: Planning Guide*, SC23-4277
- *AIX HACMP for AIX,Version 4.3: Installation Guide*, SC23-4278
- *AIX HACMP for AIX,Version 4.3: Administration Guide*, SC23-4279
- *AIX HACMP for AIX,Version 4.3: Troubleshooting Guide*, SC23-4280
- *AIX HACMP for AIX,Version 4.3: Programming Locking Applications*, SC23-4281
- *AIX HACMP for AIX,Version 4.3: Programming Client Applications*, SC23-4282
- *AIX HACMP for AIX,Version 4.3: HANFS for AIX Installation and Administration Guide*, SC23-4283
- *AIX HACMP for AIX,Version 4.3: Enhanced Scalability Installation and Administration Guide*, SC23-4284
- *AIX HACMP for AIX,Version 4.3: Master Index and Glossary*, SC23-4285

How to Get ITSO Redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, CD-ROMs, workshops, and residencies. A form for ordering books and CD-ROMs is also provided.

This information was current at the time of publication, but is continually subject to change. The latest information may be found at `http://www.redbooks.ibm.com/`.

How IBM Employees Can Get ITSO Redbooks

Employees may request ITSO deliverables (redbooks, BookManager BOOKs, and CD-ROMs) and information about redbooks, workshops, and residencies in the following ways:

- **Redbooks Web Site on the World Wide Web**

  `http://w3.itso.ibm.com/`

- **PUBORDER** – to order hardcopies in the United States
- **Tools Disks**

  To get LIST3820s of redbooks, type one of the following commands:

  ```
  TOOLCAT REDPRINT
  TOOLS SENDTO EHONE4 TOOLS2 REDPRINT GET SG24xxxx PACKAGE
  TOOLS SENDTO CANVM2 TOOLS REDPRINT GET SG24xxxx PACKAGE (Canadian users only)
  ```

  To get BookManager BOOKs of redbooks, type the following command:

  ```
  TOOLCAT REDBOOKS
  ```

  To get lists of redbooks, type the following command:

  ```
  TOOLS SENDTO USDIST MKTTOOLS MKTTOOLS GET ITSOCAT TXT
  ```

  To register for information on workshops, residencies, and redbooks, type the following command:

  ```
  TOOLS SENDTO WTSCPOK TOOLS ZDISK GET ITSOREGI 1998
  ```

- **REDBOOKS Category on INEWS**
- **Online** – send orders to: USIB6FPL at IBMMAIL or DKIBMBSH at IBMMAIL

Redpieces

For information so current it is still in the process of being written, look at "Redpieces" on the Redbooks Web Site (`http://www.redbooks.ibm.com/redpieces.html`). Redpieces are redbooks in progress; not all redbooks become redpieces, and sometimes just a few chapters will be published this way. The intent is to get the information out much more quickly than the formal publishing process allows.

How Customers Can Get ITSO Redbooks

Customers may request ITSO deliverables (redbooks, BookManager BOOKs, and CD-ROMs) and information about redbooks, workshops, and residencies in the following ways:

- **Online Orders** – send orders to:

	IBMMAIL	**Internet**
In United States	usib6fpl at ibmmail	usib6fpl@ibmmail.com
In Canada	caibmbkz at ibmmail	lmannix@vnet.ibm.com
Outside North America	dkibmbsh at ibmmail	bookshop@dk.ibm.com

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	(long distance charges apply)
(+45) 4810-1320 - Danish	(+45) 4810-1020 - German
(+45) 4810-1420 - Dutch	(+45) 4810-1620 - Italian
(+45) 4810-1540 - English	(+45) 4810-1270 - Norwegian
(+45) 4810-1670 - Finnish	(+45) 4810-1120 - Spanish
(+45) 4810-1220 - French	(+45) 4810-1170 - Swedish

- **Mail Orders** – send orders to:

IBM Publications
Publications Customer Support
P.O. Box 29570
Raleigh, NC 27626-0570
USA

IBM Publications
144-4th Avenue, S.W.
Calgary, Alberta T2P 3N5
Canada

IBM Direct Services
Sortemosevej 21
DK-3450 Allerød
Denmark

- **Fax** – send orders to:

United States (toll free)	1-800-445-9269
Canada	1-800-267-4455
Outside North America	(+45) 48 14 2207 (long distance charge)

- **1-800-IBM-4FAX (United States)** or **(+1) 408 256 5422 (Outside USA)** – ask for:

Index # 4421 Abstracts of new redbooks
Index # 4422 IBM redbooks
Index # 4420 Redbooks for last six months

- **On the World Wide Web**

Redbooks Web Site	http://www.redbooks.ibm.com
IBM Direct Publications Catalog	http://www.elink.ibmlink.ibm.com/pbl/pbl

Redpieces

For information so current it is still in the process of being written, look at "Redpieces" on the Redbooks Web Site (`http://www.redbooks.ibm.com/redpieces.html`). Redpieces are redbooks in progress; not all redbooks become redpieces, and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

IBM Redbook Order Form

Please send me the following:

Title	**Order Number**	**Quantity**

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

□ Invoice to customer number

□ Credit card number

Credit card expiration date Card issued to Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

List of Abbreviations

AIX	Advanced Interactive Executive
APA	All Points Addressable
APAR	Authorized Program Analysis Report The description of a problem to be fixed by IBM defect support. This fix is delivered in a PTF (see below).
ARP	Address Resolution Protocol
ASCII	American Standard Code for Information Interchange
AS/400	Application System/400
CDF	Cumulative Distribution Function
CD-ROM	Compact Disk - Read Only Memory
CLM	Cluster Lock Manager
CLVM	Concurrent Logical Volume Manager
CPU	Central Processing Unit
CRM	Concurrent Resource Manager
DE	Differential Ended
DLC	Data Link Control
DMS	Deadman Switch
DNS	Domain Name Service
DSMIT	Distributed System Management Interface Tool
FDDI	Fiber Distributed Data Interface
F/W	Fast and Wide (SCSI)
GB	Gigabyte
GODM	Global Object Data Manager
GUI	Graphical User Interface
HACMP	High Availability Cluster Multi-Processing
HANFS	High Availability Network File System
HCON	Host Connection Program
IBM	International Business Machines Corporation
I/O	Input/Output
IP	Interface Protocol
IPL	Initial Program Load (System Boot)
ITSO	International Technical Support Organization
JFS	Journaled File System
KA	Keepalive Packet
KB	kilobyte
Kb	kilobit
LAN	Local Area Network
LU	Logical Unit (SNA definition)
LUN	Logical Unit (RAID definition)
LVM	Logical Volume Manager
MAC	Medium Access Control
MB	megabyte
MIB	Management Information Base
MTBF	Mean Time Between Failure

NETBIOS	Network Basic Input/Output System
NFS	Network File System
NIM	Network Interface Module (This is the definition of NIM in the HACMP context. NIM in the AIX 4.1 context stands for Network Installation Manager).
NIS	Network Information Service
NVRAM	Non-Volatile Random Access Memory
ODM	Object Data Manager
POST	Power On Self Test
PTF	Program Temporary Fix A fix to a problem described in an APAR (see above).
RAID	Redundant Array of Independent (or Inexpensive) Disks
RISC	Reduced Instruction Set Computer
SCSI	Small Computer Systems Interface
SLIP	Serial Line Interface Protocol
SMIT	System Management Interface Tool
SMP	Symmetric Multi-Processor
SMUX	SNMP (see below) Multiplexor
SNA	Systems Network Architecture
SNMP	Simple Network Management Protocol
SOCC	Serial Optical Channel Converter
SPOF	Single Point of Failure
SPX/IPX	Sequenced Package Exchange/Internetwork Packet Exchange
SRC	System Resource Controller
SSA	Serial Storage Architecture
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Interface Protocol
UDP	User Datagram Protocol
UPS	Uninterruptible Power Supply
VGDA	Volume Group Descriptor Area
VGSA	Volume Group Status Area
WAN	Wide Area Network

Index

ITSO Redbook Evaluation

IBM Certification Study Guide AIX HACMP
SG24-5131-00

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at http://www.redbooks.ibm.com
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?
_ Customer _ Business Partner _ Solution Developer _ IBM employee
_ None of the above

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction __________

Please answer the following questions:

Was this redbook published in time for your needs? Yes___ No___

If no, please explain:

____

____

____

____

What other redbooks would you like to see published?

____

____

____

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

____

____

____

____

IBM Certification Study Guide AIX HACMP **SG24-5131-00**

SG24-5131-00
Printed in the U.S.A.