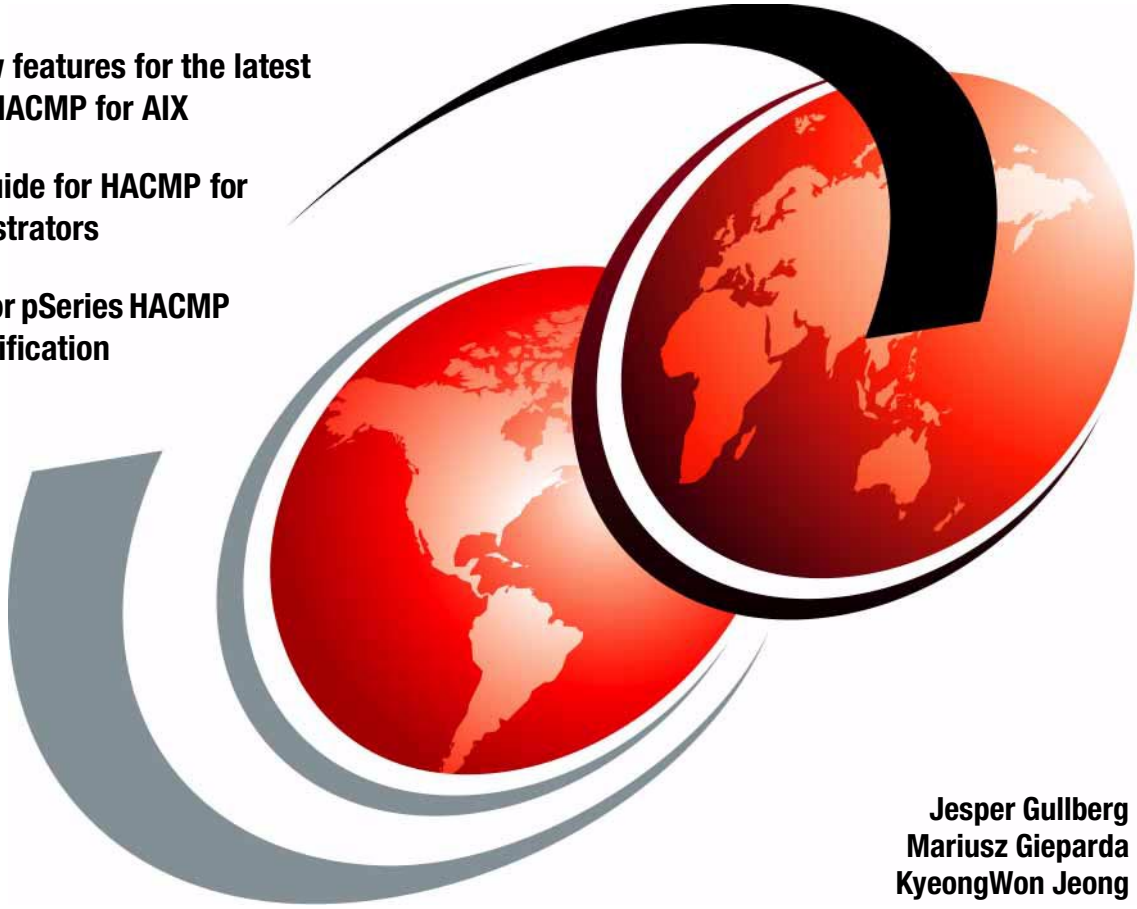
IBM

IBM Certified™
Systems Expert

IBM @server Certification Study Guide - pSeries HACMP for AIX

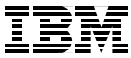
- Update new features for the latest version of HACMP for AIX
- Valuable guide for HACMP for AIX administrators
- Get ready for pSeries HACMP for AIX certification



Jesper Gullberg
Mariusz Gieparda
KyeongWon Jeong

Redbooks

ibm.com/redbooks



International Technical Support Organization

IBM @server Certification Study Guide - pSeries HACMP for AIX

November 2001

SG24-6187-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 313.

First Edition (November 2001)

This edition applies to IBM HACMP and HACMP/Enhanced Scalability (HACMP/ES) for AIX Version 4.4.1, Program Number 5765-E54, for use with the AIX Operating System, and is based on information available in October 2001.

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

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Figures

Tables

Preface

The AIX and IBM @server pSeries (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 IBM @server pSeries (RS/6000). A complete set of professional certifications is available. It includes:

- IBM Certified AIX User
- IBM Certified Specialist - AIX System Administration
- IBM Certified Specialist - AIX System Support
- IBM Certified Specialist - Business Intelligence for RS/6000
- IBM Certified Specialist - Domino for RS/6000
- IBM @server Certified Specialist - pSeries AIX System Administration
- IBM @server Certified Specialist - pSeries AIX System Support
- IBM @server Certified Specialist - pSeries Solution Sales
- IBM Certified Specialist - RS/6000 Solution Sales
- IBM Certified Specialist - RS/6000 SP and PSSP V3
- RS/6000 SP - Sales Qualification
- IBM Certified Specialist - Web Server for RS/6000
- IBM @server Certified Systems Expert - pSeries HACMP for AIX
- 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 @server Certified Systems Expert - pSeries HACMP for AIX.

The pSeries HACMP for AIX certification validates the skills required to successfully plan, install, configure, and support an HACMP for AIX 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 a 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.

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 pSeries HACMP for AIX certification exam, or just want to validate your HACMP skills, this redbook is for you.

For additional information about certification and instructions on How to Register for an exam, contact IBM Learning Services 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.

KyeongWon Jeong is a Consulting IT Specialist at the International Technical Support Organization, Austin Center. He writes extensively on AIX and education materials and teaches IBM classes worldwide on all areas of AIX. Before joining the ITSO three years ago, he worked in IBM Global Learning Services in Korea as a Senior Education Specialist and was a class manager of all AIX classes for customers and interns. He has many years of teaching and development experience. He is an IBM Certified Advanced Technical Expert - RS/6000 AIX.

Jesper Gullberg is a Systems Engineer and works for Pulsen Systems, which is an IBM Business Partner in Sweden. He has three years of experience in RS/6000, AIX, HACMP, and communications. His areas of expertise include TCP/IP, DB2, WAS, and IBM LDAP. He is an IBM Certified Advanced Technical Expert - RS/6000 AIX.

Mariusz Gieparda is a System Analyst and works for ComputerLand S.A., an IBM Business Partner in Poland. He has three years of experience in RS/6000, AIX, HACMP, and ten years of experience in networking and communications. His areas of expertise include Windows NT/2000, UNIX, TCP/IP, internetworking between different operating systems and network devices, and system and network security, including firewall environments. He is an IBM Certified Advanced Technical Expert - RS/6000 AIX and also a Microsoft Certified System Engineer.

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

International Technical Support Organization, Austin Center
Scott Vetter and Wade Wallace

International Technical Support Organization, Poughkeepsie Center
Ella Buslovich

IBM Austin
Darin Hartman and Jorge Ruiz

IBM Atlanta
Shannan L DeBrule

IBM Dallas
Christopher D Deville

Pulsen Systems, Sweden
Bjorn Roden

We would also like to thank the authors of the original publication of this redbook:

David Thiessen, Achim Rehor, Reinhard Zettler

Special notice

This publication is intended to help system administrators, system engineers, and other system professionals who want to pass the pSeries HACMP for AIX certification exam. The information in this publication is not intended as the specification of any of the following programming interfaces that are provided by 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.

IBM trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

AIX®
Approach®
DB2®
e (logo)® 
FlashCopy™
IBM ®
Micro Channel®
OS/390®
POWERparallel®
Redbooks™
RISC System/6000®
S/390®
SP™
AIX 5L™
C/2™
Domino™
Enterprise Storage Server™
HACMP/6000™
IBM.COM™
Notes®
Perform™
pSeries™
Redbooks Logo 
RS/6000®
Seascape®
Versatile Storage Server™

Comments welcome

Your comments are important to us!

We want our IBM Redbooks to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  `ibm.com/redbooks`

- Send your comments in an Internet note to:

  `redbook@us.ibm.com`

- Mail your comments to the address on page ii.

1

Certification overview

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

1.1 IBM @server Certified Systems Expert - pSeries HACMP for AIX

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

1.1.1 Certification requirements (two tests)

To attain the IBM @server Certified Systems Expert - pSeries HACMP for AIX certification, candidates must pass two tests.

Note: One test is the prerequisite in either AIX System Administration or AIX System Support. The second test is the pSeries HACMP for AIX test.

Prior to attaining the IBM @server Certified Systems Expert - pSeries HACMP for AIX certification, candidates must be certified as either an IBM Certified Specialist - AIX System Administration or IBM Certified Specialist - AIX System Support. 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 187: pSeries HACMP for AIX

- Role description

  Certifies a proficiency with implementation skills required to plan, install, and configure High Availability Cluster Multi-Processing (HACMP) for AIX systems, and also the ability to perform administrative and diagnostic activities needed to support Highly Available Clusters.

- Recommended prerequisites

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

- Registration for the certification exam

  For information about how to register for the certification exam, please contact IBM Learning Services or 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 - Pre-installation

  The following items should be considered as part of the pre-installation plan:

  - Conduct a planning session
    - Coordinate/organize 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
    - Qualify applications for HACMP implementation
  - 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
    - Design and implement global network recovery
    - Develop a disk management modification plan
    - Configure single adapter solutions
    - Produce a test plan
- Section 2 - HACMP Implementation

  The following items should be considered for proper implementation:

  - Implement customer dependent options
    - Migrate HA

    - Identify and configure SP specifics with HACMP effect
    - Implement HACMP into existing production environment
- Install/configure nodes, networking, disk, and applications
    - Configure HACMP solutions
    - Install HACMP code
    - Configure an IP address takeover (IPAT)
    - Configure MAC address takeover
    - Configure non-IP heartbeat paths
    - Configure a network adapter
    - Customize/tailor AIX
    - Set up a shared disk (SSA)
    - Set up a shared disk (SCSI)
    - Convert volume groups to concurrent access
    - Verify a cluster configuration
    - Configure HACMP nodes for enterprise management
    - Create an application server
    - Create and populate resource group
- 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 failed takeovers (node, adapter, application, and disk)
    - Troubleshoot a failed IPAT failover
    - Troubleshoot failed shared volume groups
    - Troubleshoot a failed network configuration
    - Troubleshoot failed shared disk tests
    - Troubleshoot failed pre/post event scripts

- Troubleshoot HACMP client recovery
- 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, and 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, some of the following courses are being developed:

► HACMP System Administration 1: Planning and Implementation
► HACMP System Administration 2: Maintenance and Migration
► HACMP System Administration 3: Problem Determination
► HACMP Certification Preparation Workshop
► AIX Version 4 HACMP System Administration
► Implementing HA on the RS/6000 SP
► HAGEO Implementation

The current list of courses can vary according to your geographic location. For a current list of courses, contact IBM Learning Services or visit the following Web site:

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

Table 1-1 through Table 1-5 on page 10 outlines information about the above courses.

Table 1-1 HACMP System Administration 1: Planning and Implementation

Course number	Q1554 (NA); AU54 (Worldwide)
Course duration	Five days
Course abstract	Learn how to plan, design, install, and configure IBM High Availability Cluster Multiprocessing (HACMP) for AIX and HACMP Enhanced Scalability (HACMP/ES) clustering products on the IBM @server pSeries and RS/6000 platforms. Learn how to configure cascading, mutual takeover, rotating, and concurrent modes of operation of the HACMP and HACMP/ES products. Learn how to make configuration changes to an already existing cluster. Also, learn about event processing, change management, and problem determination. The following is the objectives of the course: ► Define high availability ► Design and plan a highly available cluster ► Install and configure HACMP and HACMP/ES in the following modes of operation: cascading and mutual takeover, cascading without fallback, rotating, and concurrent access ► Perform basic system administration for HACMP ► Make configuration and customization changes for HACMP ► Perform basic problem determination and recovery actions The following topics will be covered in the course: ► High availability concepts ► HACMP features and components ► HACMP modes of operation ► Cluster planning and design ► Installation of HACMP ► Configuration of TCP/IP, Logical Volume Manager (LVM), and HACMP ► Testing cluster failover ► Basic cluster configuration changes

Table 1-2 HACMP System Administration 2: Maintenance and Migration

Course number	Q1557 (NA); AU57 (Worldwide)
Course duration	Five days
Course abstract	Learn how to determine, monitor, and change the configuration of an existing High Availability Cluster Multiprocessing (HACMP) or HACMP Enhanced Scalability (HACMP/ES) cluster with little or no impact to an active cluster environment. Also, learn about event processing and change management. Learn how to maintain, and migrate IBM HACMP for AIX and HACMP/ES products on the IBM @server pSeries and RS/6000 platforms with little or no impact to an active cluster. All theory components apply equally to both HACMP and HACMP/ES technologies. The following is the objectives of the course: ► Determine the configuration of a cluster ► Determine the status of the cluster ► Monitor an active cluster with standard AIX commands, utilities provided with HACMP, Enhanced Simple Network Management Protocol (SNMP) based tools (NetView or Tivoli) ► Migrate from HACMP to HACMP and HACMP/ES ► Perform system administration for HACMP ► Make configuration and customization changes for HACMP The following topics will be covered in the course: ► How to monitor a cluster ► How to do an HACMP migration ► Configuring HACMP to work with other network services ► Maintenance of HACMP in a production environment ► Making HACMP configuration changes in a production environment ► Other shared devices, including printers, in an HACMP environment

Table 1-3 HACMP System Administration 3: Problem Determination

Course number	Q1559 (NA); AU59 (Worldwide)
Course duration	Five days
Course abstract	Learn how to perform problem determination and recovery on IBM High Availability Cluster Multiprocessing (HACMP) for AIX and HACMP Enhanced Scalability (HACMP/ES) clustering products on the IBM @server pSeries and RS/6000 platforms. Learn how to correct common administration errors in a running HACMP cluster environment. All theory components apply equally to both HACMP and HACMP/ES technologies. The following are the objectives of the course: ► Diagnose problems associated with an HACMP cluster ► Recover from cluster failures ► Identify and correct network configuration errors ► Identify and correct cluster configuration errors ► Handle HACMP event failures ► Escalate a small failure to a node failure The following topics will be covered in the course: ► Tools for problem identification and determination ► Common reasons for cluster failure ► Handling network configuration errors ► Handling LVM configuration errors ► Fixing configuration errors on a running cluster ► Enabling error notification ► Application monitoring

Table 1-4 HACMP Certification Preparation Workshop

Course number	Q1829 (NA); AW29 (Worldwide)
Course duration	Two days
Course Abstract	This course will refresh your knowledge and increase your confidence for the exam. The exam is given at the end of the second day and is included in the price of the course. Completion of the prerequisites helps the instructor customize the course to fit your needs. The precise content of this course is dictated by your requirements. In order to establish these requirements, you take a mock test at the beginning of the workshop. The following is the objectives of the course: ► Understand what is necessary to succeed on the IBM @server Certified Systems Expert - pSeries HACMP for AIX examination ► Identify items covered on the IBM @server Certified Specialist - pSeries HACMP for AIX examination ► Clarify and enhance knowledge from other HACMP courses or hands-on experience

Table 1-5 AIX Version 4 HACMP System Administration

Course number	Q1150 (NA); 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 course involves a significant number of hands-on exercises to reinforce the concepts. Students are expected to have completed the course AU54 (Q1554) HACMP Installation and Implementation before attending this course. The student learns the administrative, maintenance, and configuration tasks involved in running an existing High-Availability Cluster Multiprocessing (HACMP) cluster. The following is the objectives of the course: ► Maintain an accurate configuration of the cluster ► Monitor the cluster for any failures that occur ► Replace failed hardware in a cluster ► Recover from HACMP function failures ► Perform routine maintenance of user IDs, shared volume groups, and Program Temporary Fixes (PTF) ► Change cluster configurations with minimal impact to users The following topics will be covered in the course: ► Determining cluster configuration ► Determining cluster status ► Integrating HACMP with existing network services ► Failure recovery ► Routine maintenance tasks ► Cluster configuration changes ► Keeping your cluster healthy

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, and 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 application loads and characteristics into account, 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 High availability

A high availability solution will ensure that any failure of any component of the solution, be it hardware, software, or system management, will not cause the application and its data to be inaccessible to the user community. This is achieved through the elimination or masking of both planned and unplanned downtime. A highly available solution should eliminate single point of failure (SPOF) through appropriate design, planning, selection of hardware, configuration of software and carefully controlled change management discipline.

High availability is:

- The masking or elimination of both planned and unplanned downtime.
- The elimination of single points of failure (SPOFs)
- Fault resilience, but not fault tolerance

High availability systems are an excellent solution for applications that can withstand a short interruption should a failure occur, but which must be restored quickly.

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

System downtime is either planned or unplanned, with planned downtime accounting for the vast majority of the total. HACMP allows you to minimize or eliminate this planned downtime from your operation, by allowing you to maintain a service to your customers while performing hardware upgrades, software upgrades, or other maintenance activity at the same time. Services may be moved from one cluster node to another at will, allowing the original node to undergo maintenance without affecting the availability of the service. When the maintenance activity is completed, the service may be moved back to the node which was originally running it.

Unplanned downtime has one of two causes: periods caused by hardware failures, and those periods caused by software failures. Hardware has been getting more and more reliable over time and will continue to do so, but hardware remains a cause of failures. Together with facilities provided by AIX, HACMP can

protect your operation from a hardware failure, by automatically moving the services provided by the failing node to other nodes within the cluster. To see the difference between HACMP versions, see Chapter 10, “HACMP classic versus HACMP/ES” on page 303.

Note: To achieve the system high availability, the HACMP software must be installed and configured on every node within the cluster.

2.2 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.

When a piece of hardware is listed as qualified for use with HACMP, it is qualified for use with all the feature codes of HACMP, including HANFS. The only exception is that disk support for the Concurrent Resource Manager is different than for the remainder of HACMP and these differences are shown in the disk enclosure tables. The designated hardware should only be used on an appropriate IBM @server pSeries, RS/6000 Platform, or 9076 Scalable POWERParallel Platform (SP).

The following sections will deal with the various:

- Operating system levels
- CPU options
- Disk storages for CRM
- Cluster node considerations

available to you when you are planning your HACMP cluster.

2.2.1 Operating system level

Before installation of HACMP, make sure to have the proper version of the operating system. Here is a list of required operating system levels for HACMP versions (see Table 2-1) and Parallel System Support Program versions (see Table 2-2).

Table 2-1 Required OS level for HACMP

AIX OS level	HACMP Version 4.3.1	HACMP Version 4.4.0	HACMP Version 4.4.1
4.3.2	yes	no	no
4.3.3	yes	yes	yes
5.1	no	yes*	yes

*** Note:**
The following restrictions apply to HACMP for AIX Version 4.4.0 support of IBM AIX 5L for Power Version 5.1. IBM intends to remove these restrictions through further APARs on HACMP.

- Enhanced Journaled File Systems are not supported on shared volume groups.
- Fencing is not supported for concurrent mode volume groups created on 9333 disks.
- HACMP can only run on 32-bit AIX kernels. Even if the hardware is capable of supporting 64-bit kernels, it must be booted using the the **`bosboot`** command with a 32-bit kernel.
- The VSM-based xhacmpm configuration utility is not supported.

Table 2-2 PSSP versions for SP installation

HACMP version	Prerequisite PSSP version
HACMP Version 4.3.1 for AIX	PSSP Version 3.1
HACMP Version 4.4.0 for AIX	PSSP Version 3.2
HACMP Version 4.4.1 for AIX	PSSP Version 3.2

2.2.2 CPU options

HACMP is designed to execute with RS/6000 uniprocessors, SMP servers, and SP systems in a no single point of failure server configuration. HACMP supports the IBM @server pSeries and the RS/6000 models that are designed for server application and meet the minimum requirements for internal memory, internal disk, and I/O slots. See Table 2-3 and Table 2-4 for more information.

Table 2-3 HACMP supported RS/6000 models

7043-140	7043-150	7043-240	7043-260	7024-E20
7024-E30	7025-F30	7025-F40	7025-F50	7026-H10
7026-H50	7026-H70	7012-G30	7012-G40	7013-J50
7017-S70	7017-S7A	7017-S80	7011-25S	7011-250
7011-25T	7009-C10	7009-C20	7012-370	7012-380
7012-390	7030-397	7013-J30	7013-J40	7012-39H
7013-570	7013-57F	7013-580	7013-58F	7013-58H
7013-590	7013-59H	7013-591	7013-595	7015-98B
7015-98E	7015-R30	7015-R40	7015-98F	7015-990
7015-99E	7015-99F	7015-99J	7015-99K	7015-R10
7015-R20	7015-R21	7015-R24	7015-R50	7015-R5U
7015-R3U	7015-R4U	7017-S70	7017-S7A	7017-S80
7017-S85	7044-170	7044-270	7025-F80	7025-6F1
7026-H80	7026-6H1	7026-M80	7026-B80	

Table 2-4 HACMP supported SP platforms

204	205	206	207	208
209	20A	2A4	2A5	2A7
2A8	2A9	2AA	304	305
306	307	308	309	30A
3A4	3A5	3A7	3A8	3A9
3AA	3B4	3B5	3B7	3B8
3B9	3BA	404	405	406
407	408	409	4AA	500

550	50H	55H		

Note: 604 High Nodes, 604E High Nodes, the Power2 Super Chip (P2SC) nodes and the 375 MHz Power3 SMP Nodes are also supported.

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

2.2.3 Disk storages

HACMP executing in a concurrent access configuration requires one of the following devices (see Table 2-5).

Table 2-5 HACMP requirements for concurrent access configuration

IBM 7131 SSA Multi-Storage Tower Model 405 (supports up to eight nodes, no CD-ROMs or tapescan be installed)
IBM 7133 SSA Disk Subsystem Models 020, 600, D40, and T40 (supports up to eight nodes)
IBM 7135 RAIDiant Array Models 110 and 210 (supports up to four nodes, dual controllers recommended)
IBM 7137 Disk Array Subsystem Models 413, 414, 415, 415, 513, 514 or 515 (supports up to four nodes)
IBM 2102 Fibre Channel Storage Server Model F10
IBM 2105 Versatile Storage Server (VSS) Models B09 and 100 (supports up to four nodes)
IBM 2105 Enterprise Storage Server (ESS) Models E10, E20, F10, and F20 (supports up to eight nodes)

Note:

- Models E10 and E20 for HACMP Version 4.4 requires APARs IY11563 and IY11565.
- Models F10 and F20 for HACMP Version 4.4 requires APARs IY11480, IY11563, and IY11565.
- Native Fibre Attachment support for ESS Models E10, E20, F10, and F20 added for Version 4.4 upon availability of APARs IY12021, IY12022, IY12056, and IY12057.

2.2.4 Cluster node considerations

Your major goal throughout the planning process is to eliminate single points of failure. A single point of failure exists when a critical cluster function is provided by a single component. If that component fails, the cluster has no other way of providing that function, and the service depending on that component becomes unavailable.

Realize that, while your goal is to eliminate all single points of failure, you may have to make some compromises. There is usually a cost associated with eliminating a single point of failure. For example, purchasing an additional hardware device to serve as backup for the primary device increases cost. The cost of eliminating a single point of failure should be compared against the cost of losing services should that component fail. Again, the purpose of the HACMP for AIX software is to provide a cost-effective, highly available computing platform that can grow to meet future processing demands.

Note: HACMP for AIX is designed to recover from a single hardware or software failure. It may not be able to handle multiple failures, depending on the sequence of failures. For example, the default event scripts cannot do an adapter swap after an IP address takeover (IPAT) has occurred if only one standby adapter exists for that network.

To be highly available, all cluster resources associated with a critical application should have no single points of failure. As you design an HACMP cluster, your goal is to identify and address all potential single points of failure

How to eliminate the single point of failure

Table 2-6 summarizes potential single points of failure within an HACMP cluster and describes how to eliminate them.

Table 2-6 Eliminating cluster objects as single points of failure

Cluster object	Eliminated a single point of failure by...
Node	Using multiple nodes
Power source	Using multiple circuits or uninterruptable power supplies
Network adapter	Using redundant network adapters
Network	Using multiple networks to connect nodes
TCP/IP subsystem	Using serial networks to connect adjoining nodes and clients
Disk adapter	Using redundant disk adapters
Controller	Using redundant disk controllers
Disk	Using redundant hardware and disk mirroring
Application	Assigning a node for application takeover

See the *Concepts and Facilities Guide,* SC23-4276 for an in-depth discussion on eliminating cluster objects as single points of failure.

Note: In an HACMP for AIX environment on an SP machine, the SP Switch adapter is a single point of failure and should be promoted to node failure. See the appendix on using the SP machine in the *Installation Guide*, SC23-4278 for complete information on the SP Switch adapter functionality.

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.

2.3 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.

2.3.1 TCP/IP networks

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

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 “Supported Non-TCP/IP network types” on page 23.

- 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.

Integrating HACMP with network services

HACMP requires IP address-to-name resolution. The three most commonly used methods include:

- Domain Name Service
- Network Information Service
- Flat file name resolution (/etc/hosts)

By default, a name request will look first for the DNS (/etc/resolv.conf), second for NIS, and last for /etc/hosts to resolve the name. Since DNS and NIS both require certain hosts as designated servers, it is necessary to maintain the /etc/hosts file in case the DNS or NIS name server is unavailable, and to identify hosts that are not known to the name server. It is required to have all cluster adapter IP labels in all cluster nodes' /etc/hosts tables.

To ensure the most rapid name resolution of cluster nodes, it is recommended that you change the default order for name serving so that /etc/hosts is used first (at least for cluster nodes). To do this, edit the /etc/netsvc.conf file so that this line appears as shown here:

```
Hosts=local,nis,bind
```

Putting the local option first tells the system to use /etc/hosts first, then NIS. You can also change the order for name resolution by changing the environment variable NSORDER so it looks like this:

```
NSORDER=local,bind,nis
```

Note: If you are using NIS, we recommend that you have the NIS master server outside the cluster, and have the cluster nodes run as NIS slave servers. At a minimum, every HACMP node must be able to access NIS master or slave servers on a local subnet, and not via a router.

If you are using DNS or NIS, this fact must be configured in the run-time parameters of all nodes in the cluster. The Cluster Manager then knows to stop and restart the named daemon when it is changing adapter IP addresses (in an adapter_swap event).

For more information on how to test the name resolution configuration, please refer to Section 6.1.4, “Network state” on page 163.

Note: You cannot use DHCP to allocate IP addresses to HACMP cluster nodes. Clients may use this method, but cluster nodes cannot.

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. For example:

- *Ethernet* supports 10, 100, and 1000 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 does not 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.
- *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 300 MBps. 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 if you would like to use the IPAT feature, 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.

2.3.2 Non-TCP/IP 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.

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.

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. In case the number of native serial ports does not 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.

Note:
On the SP thin or wide nodes there are no serial ports available. Therefore, any HACMP/ES configurations that require a tty network need to make use of a serial adapter card (8-port async EIA-232 adapter, FC/2930), available on the SP as an RPQ.

The 7013-S70, 7015-S70, and 7017-S70 do not support the use of native serial ports in an HACMP/ES RS232 serial network. Configuration of an RS232 serial network in an S70 system requires a PCI multi-port Async card.

Configuration of RS232 port should be as follows:

- BAUD rate: 9600 bps
- PARITY: none
- BITS per character: 8
- Number of STOP BITS: 1
- Enable LOGIN: disable

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-1 Single-Ended and SCSI-2 Single-Ended do not support serial networks in an HACMP cluster. The recommendation is to not use more than four 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 (feature code #6215, #6225, #6230, or #6219), since these are the only current adapters that support the Multi-Initiator Feature.

2.4 Cluster disks

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

2.4.1 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 2-1.

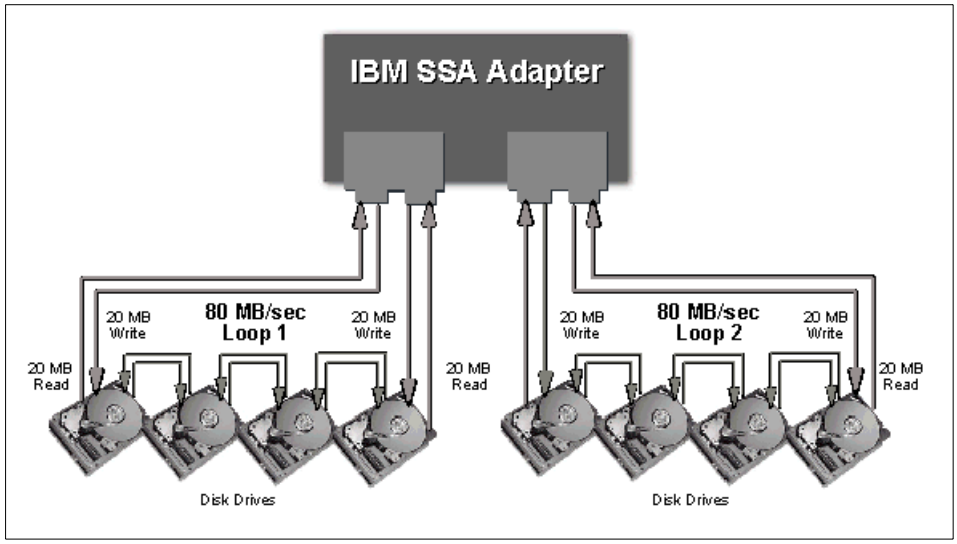


Figure 2-1 Basic SSA configuration

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 160 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 RS/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. Models 010 and 500 are not supported from HACMP Version 4.4.0. More recently, in November 1998, the models D40 and T40 were announced.

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

SSA adapters

Here is a short comparison of SSA adapters. Table 2-7 lists the different SSA adapters and presents an overview of their characteristics.

Table 2-7 SSA adapters

Feature code	Bus	Adapter description	Number of adapters per loop	Hardware RAID types
6214	MCA	Classic	2	n/a
6215	PCI	Enhanced RAID-5	8*	5
6216	MCA	Enhanced	8	n/a
6219	MCA	Enhanced RAID-5	8*	5
6225	PCI	Advanced SerialRAID Adapter	8*	0, 5
6230	PCI	Advanced SerialRAID Plus Adapter	8*	0, 0+1, 5

*** Note:** See “Rules for SSA loops” on page 26 for more information. You cannot have more than four adapters in a single system.

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 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 SSA loops that include an Advanced Serial RAID Adapter, the following rules apply:

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

- All members of an array must be on the same SSA loop.
- A maximum of 48 devices can be connected in a particular SSA loop.
- If an SSA adapter that is in a two-way configuration is connected to two SSA loops, and a second adapter is connected to each loop, both loops must be connected to the same second adapter.
- Each SSA loop can be connected to no more than two adapters on any one in-use system.
- The number of adapters that are supported in an SSA loop is determined by whether any disk drivers are configured for RAID or fast-write operations, and by type of adapter.

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.

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 possible enhancements 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. A 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 Table 2-8.

Table 2-8 The advantages and disadvantages of the different RAID levels

RAID level	Availability mechanism	Capacity	Performance	Cost
0	None	100 percent	High	Medium
1	Mirroring	50 percent	Medium/high	High
3	Parity	80 percent	Medium	Medium
5	Parity	80 percent	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 is 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.

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, 10 km 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.

2.4.2 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.

Capacities

The capacities of the disks and the the disk subsystems are as follows:

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 (maximum of 67.5 GB) in the base configuration and 30 disks (maximum of 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.

Amount in a string

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

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

High availability features

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.

2.4.3 Fibre Channel adapters

Fibre Channel is a 100 MB/s or 1 GB/s, full-duplex, serial communication technology used to interconnect input/output (I/O) devices and host systems that can be separated by tens of kilometers. It incorporates the best features of traditional I/O interfaces, such as throughput and reliability, found in SCSI and PCI, with the best features of networking interfaces, such as connectivity and scalability, found in Ethernet and Token Ring. It provides a new transport mechanism for the delivery of existing commands, and provides an architecture that achieves high performance by allowing a significant amount of processing to be performed in hardware. It can operate with legacy protocols and drivers like SCSI and IP, enabling it to be introduced easily into existing infrastructures.

Fibre Channel transfers information between the sources and the users of the information. This information can include commands, controls, files, graphics, video, and sound. Fibre Channel connections are established between Fibre Channel ports residing in I/O devices, host systems, and the network interconnecting them. The network consists of elements like switches, hubs, bridges, and repeaters, that are used to interconnect the Fibre Channel ports.

The PCI adapter is a 32/64 bit adapter card and is installed as a SCSI controller. AIX and HACMP supports the Gigabit Fibre Channel Adapters like FC6227 and FC6228 (for 64-bit PCI Bus). With Fibre Channel Adapters you can connect, for example, storages like IBM Enterprise Storage Server (ESS).

Fibre Channel topologies

There are three Fibre Channel topologies defined in the Fibre Channel architecture. These are Point-to-Point, Switched Fabric, and Arbitrated Loop.

Point-to-point

This is the simplest Fibre Channel topology. See Figure 2-2 on page 35.

Figure 2-2 Point-to-point topology

It allows exactly two Fibre Channel end points (for example, I/O device or host system) to be directly connected via a Fibre Channel cable. This topology can satisfy only the very basic configuration requirements. It does not provide sufficient connectivity to support complex configurations, but it does support the maximum bandwidth capability of Fibre Channel. ESS fully supports the point-to-point topology. It supports this topology with what has come to be called its point-to-point protocol. This protocol is also used by ESS when directly attached to a fabric in a switched fabric topology.

Switched Fabric

In this topology two or more Fibre Channel end points are interconnected through one or more switches. See Figure 2-3.

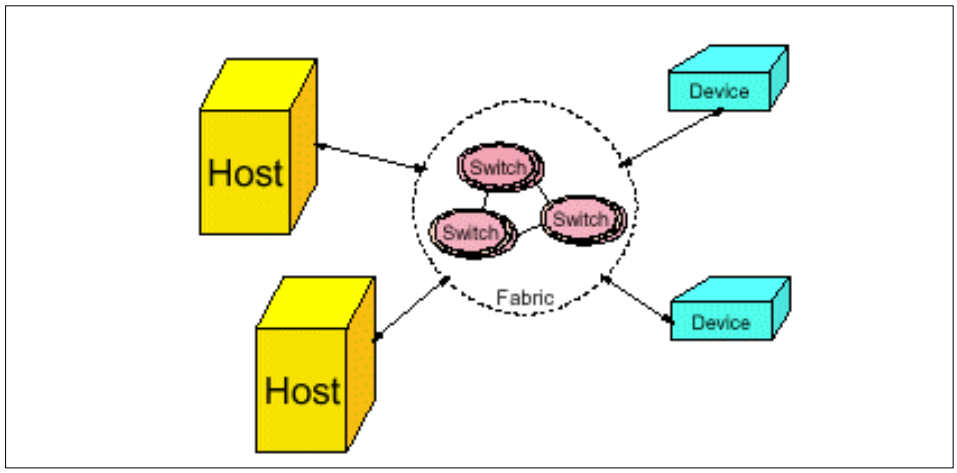


Figure 2-3 Switched Fabric topology

Each switch can support up to 256 ports (per the Fibre Channel architecture), but most switches today support only 8, 16, 32, or 64 ports. When multiple switches are interconnected, they are said to be cascaded. The set of cascaded switches is commonly called a fabric. The term fabric is used to describe a routing structure that receives addressed information and routes it to its appropriate destination. A fabric may consist of one or more switches, functioning as a single routing mechanism.

As mentioned earlier, the architected maximum number of end points that may be connected in a fabric is somewhat less than 8 million when maximum use is made of loop configurations, and somewhat less than 16 million when no loops are used. Most fabrics in existence today, though, support only 10s to 100s of ports. The Fibre Channel end points have no awareness of the internal structure of the fabric. Fibre Channel architecture does not support the cascading of independent fabrics where each fabric contains its own separate routing mechanism. The addressing structure within Fibre Channel architecture will only support a single fabric router. Switches (or switched fabrics) also include a function commonly called Zoning.

This function allows the user to partition the switch ports into port groups. The ports within a port group, or zone, can only communicate with other ports in the same port group (zone). By using zoning, the I/O from one group of hosts and devices can be completely separated from that of any other group, thus preventing the possibility of any interference between the groups.

One example where zoning might be used is in an environment where both a production system and a test system coexist. The customer will want to make sure that his test system I/O does not impact his production system I/O. He can accomplish this very simply by setting up two zones within the fabric, one for the hosts and devices in the production system, and the other for the hosts and devices in the test system. He could, of course, accomplish the same result by implementing two separate switch fabrics, but that would be more expensive and also require more complicated configuration management.

This zoning just described is known as hard zoning, since it is enforced by the switch. There is another kind of zoning, called soft zoning, that can be enabled as well. It is not enforced by the switch, but rather operates on the honor system. The way this zoning works is that the user assigns nodes to a zone according to the node's World Wide Name - either the World Wide Port Name (WWPN) or the World Wide Node Name (WWNN). This information is captured by the name server, which is a function embedded within the switch. Then, whenever a port communicates with the name server to find out to which nodes it is allowed to connect, the name server will respond only with the nodes that are within that port's zone. Since the standard Fibre Channel device drivers do communicate with the name server in this manner, this type of zoning is adequate for most situations. However, it is possible that a device driver could be designed that would attempt to access nodes not in its list of allowed connections. If this occurred, the switch would neither prevent nor detect the violation.

ESS fully supports the switched fabric topology. As stated previously, it supports this topology with its point-to-point protocol. Through the use of switched fabrics, ESS can attach to multiple Fibre Channel hosts via a single Fibre Channel adapter.

ESS supports connections to a maximum of 128 hosts per Fibre Channel port, and a maximum of 512 hosts across all Fibre Channel and SCSI ports configured on one machine. In switched configurations, because of possible interactions among the Fibre Channel adapters from one or more hosts, it is recommended that zones be established that contain the desired number of ESS ports and only a single host port. These zones can be created as either soft zones or hard zones. The ESS ports can be members of multiple zones, so that zones can be established with each host port to allow the hosts to have shared access to the ESS ports. Creating the zones in this manner will prevent any unwanted communication between the host's ports. In general, fabrics allow all of its ports to use the full bandwidth of the Fibre Channel architecture simultaneously. This provides for a very high performance interconnection topology.

Arbitrated loop

In this topology, two or more (up to a maximum of 126) Fibre Channel end points are interconnected via a looped interface (Figure 2-4 on page 38).

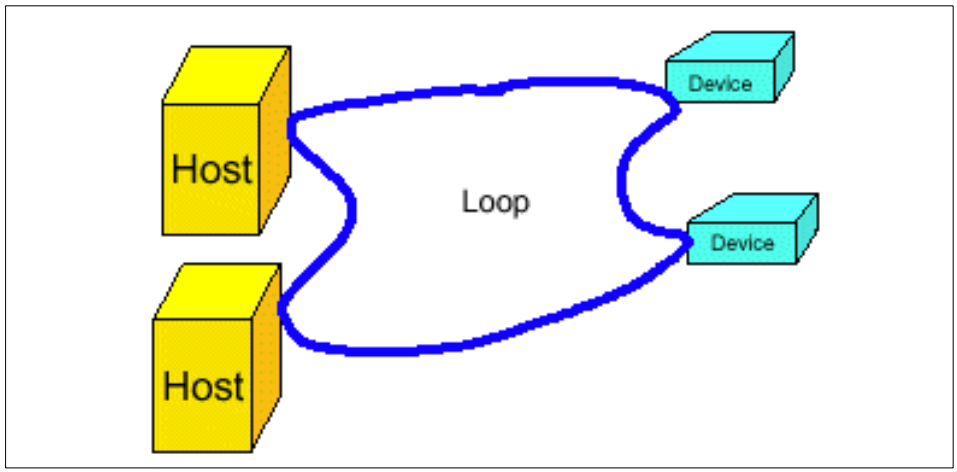


Figure 2-4 Arbitrated loop topology

Information is routed around the loop and repeated by intermediate ports until it arrives at its destination. The Fibre Channel ports that support this topology must contain these routing and repeating functions in addition to all the functions required by the Point-to-Point ports. This topology is called Fibre Channel - Arbitrated Loop, and is often referred to by its acronym FC-AL. All end points share the FC-AL interface and therefore also share the bandwidth of the interface. Only one active connection may be ongoing at a time. A hub incorporates the structure of FC-AL in a package that provides ports physically similar to those of a switch. This allows more centralized cabling and, due to added functionality provided by hubs, increases the reliability and availability of the Arbitrated Loop topology. It does not, however, improve the performance of the loop, since it is still just a single loop that has been repackaged.

Today's hubs generally have 8, 16, or 32 ports. Additional ports can be configured by cascading hubs together, but the resulting configuration is still a single loop. One important point to remember concerning loops, whether or not a hub is utilized, is that the loop goes through an architected Loop Initialization Process (LIP) each time the loop is broken and reconnected or whenever a host or device is added to or removed from the loop. This includes whenever hosts or devices that are attached to the loop are powered on or off. This LIP disrupts any I/O operations currently in progress. Thus, if you have multiple hosts on a loop, then whenever any host is booted up it will cause a LIP and therefore disrupt any ongoing I/O. For this reason it is normally recommended to only have a single host on any loop. Devices have the same effect, but they are generally not booted very often. It is therefore quite common to have multiple devices on a loop.

ESS currently does not support the arbitrated loop topology (when there are more than two Fibre Channel components in the loop). It supports attachment to a hub only as a distance extender where it is attached to a single host. No other devices or hosts may be attached to the hub.

A common variation on the loop topology is the point-to-point loop (Figure 2-5).

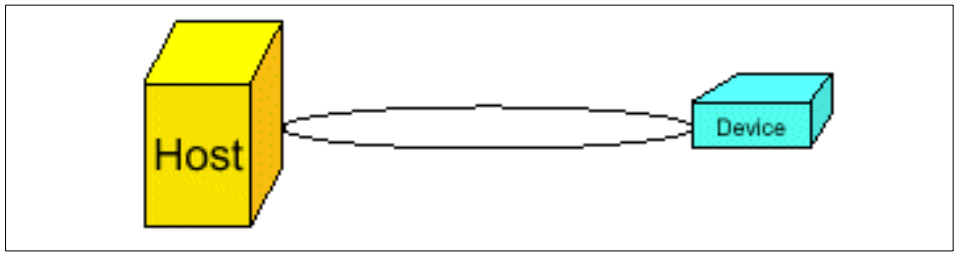


Figure 2-5 Point-to-point loop topology

This topology is very similar to the point-to-point topology. There is only one I/O initiator (host) and one I/O target (device) on the loop, just like in the point-to-point case. The difference, though, is in the protocol used to communicate between the initiator and the target. In the point-to-point topology described previously, the point-to-point protocol was used. In this topology, the loop protocol is used. These protocols are significantly different from each other. The point-to-point protocol is architected for the point-to-point and fabric topologies. The loop protocol is architected for a point-to-point topology with exactly two Fibre Channel end points connected together, and also for a loop topology with anywhere from 3 to 126 Fibre Channel end points connected together. Therefore, when just two Fibre Channel end points are connected together, either the point-to-point protocol or the loop protocol can be utilized, but both end points must use the same protocol. Most host Fibre Channel adapters will default to using the loop protocol whenever they are not directly connected to a fabric. In these cases the ESS adapter must also be configured for the loop protocol. And, of course, when directly connected to a fabric, the adapters will use the point-to-point protocol.

For more information about Fibre Channel technology and ESS Fibre Channel Attachment please refer to *Enterprise Storage Server Fibre Channel Attachment Version 6.0*, found at:

`http://www.storage.ibm.com/hardsoft/products/ess/whitepaper.htm`

ESS technology

The IBM Enterprise Storage Server is a second-generation Seascape disk storage system that provides industry-leading availability, performance, manageability, and scalability.

The IBM Enterprise Storage Server does more than simply enable shared storage across enterprise platforms; it can improve the performance, availability, scalability, and manageability of enterprise-wide storage resources through a variety of powerful functions:

- FlashCopy provides fast data duplication capability. This option helps eliminate the need to stop applications for extended periods of time in order to perform backups and restores.
- Peer-to-Peer Remote Copy maintains a synchronous copy (always up-to-date with the primary copy) of data in a remote location. This backup copy of data can be used to quickly recover from a failure in the primary system without losing any transactions; an optional capability that can literally keep your e-business applications running.
- Extended Remote Copy (XRC) provides a copy of OS/390 data at a remote location (which can be connected using telecommunications lines at unlimited distances) to be used in case the primary storage system fails. The Enterprise Storage Server enhances XRC with full support for unplanned outages. In the event of a telecommunications link failure, this optional function enables the secondary remote copy to be re synchronized quickly without requiring duplication of all data from the primary location for full disaster recovery protection.
- Custom volumes enable volumes of various sizes to be defined for S/390 servers, enabling administrators to configure systems for optimal performance.
- Storage partitioning uses storage devices more efficiently by providing each server access to its own pool of storage capacity. Storage pools can be shared among multiple servers.

2.5 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.

2.5.1 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`.

A parameter called *Cascading Without Fallback* (CWOF) is an attribute of cascading resource groups that defines its fallback behavior. (this is a new feature in HACMP Version 4.4.1). When this flag is set to `TRUE`, a cascading resource group will not fallback to a higher priority node as it joins or reintegrates into the cluster. A cascading group with CWOF set to `FALSE` will exhibit fallback behavior.

You can use cascading configuration in case you have servers (nodes) in a different configuration, for example, the first node has better performance and many more resources (memory, processors and so on) than the second node. In this case, you can use the slowest, less powerful node as a temporary server for your resources and applications if first node fail. If the first node comes up, resources automatically will be moved to the first node (with highest 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.

As an example of using concurrent resource group is a database, where you can spread their workload across the cluster.

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.

2.5.2 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.

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 2-6 on page 44 illustrates a two node cluster in a hot-standby configuration.

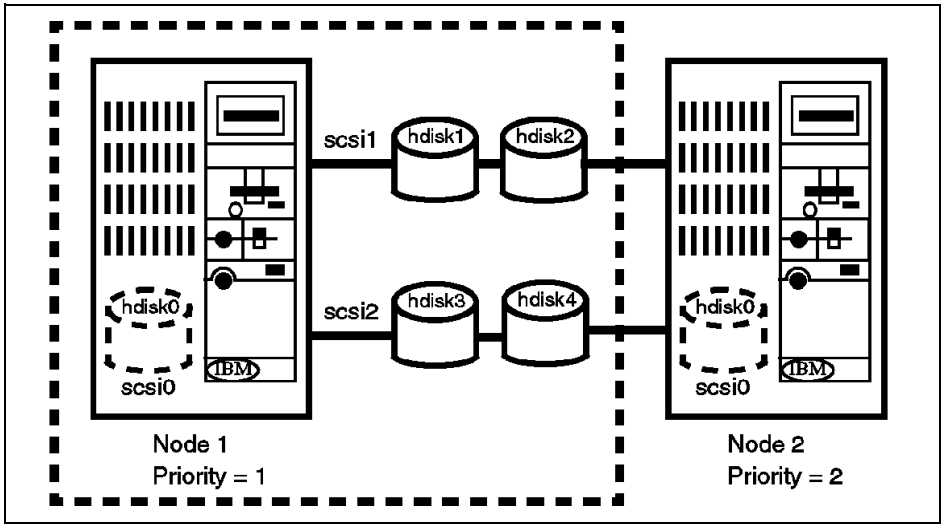


Figure 2-6 Hot-standby configuration

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 2-7 illustrates a two node cluster in a mutual takeover configuration.

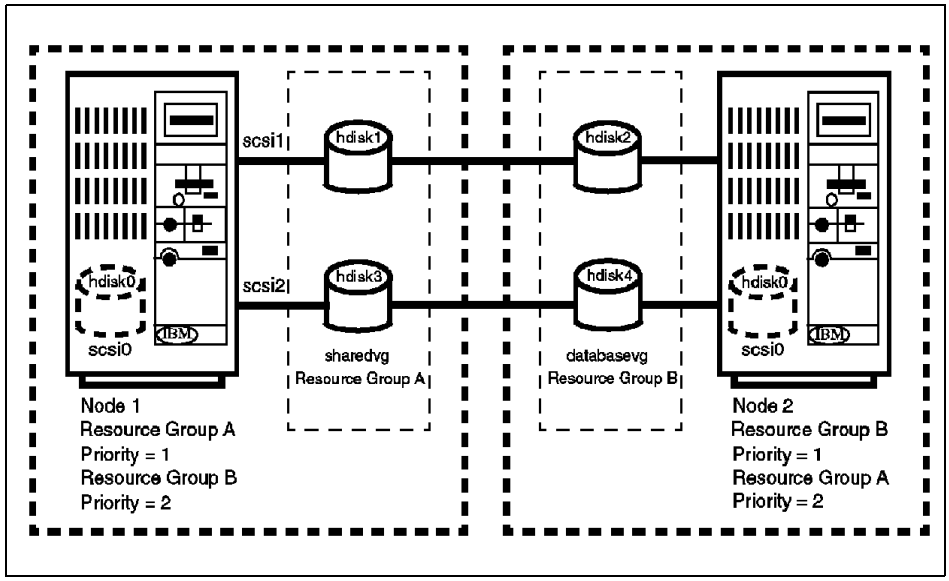


Figure 2-7 Mutual takeover configuration

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 2-8 on page 47 illustrates a three node cluster in a third-party takeover configuration.

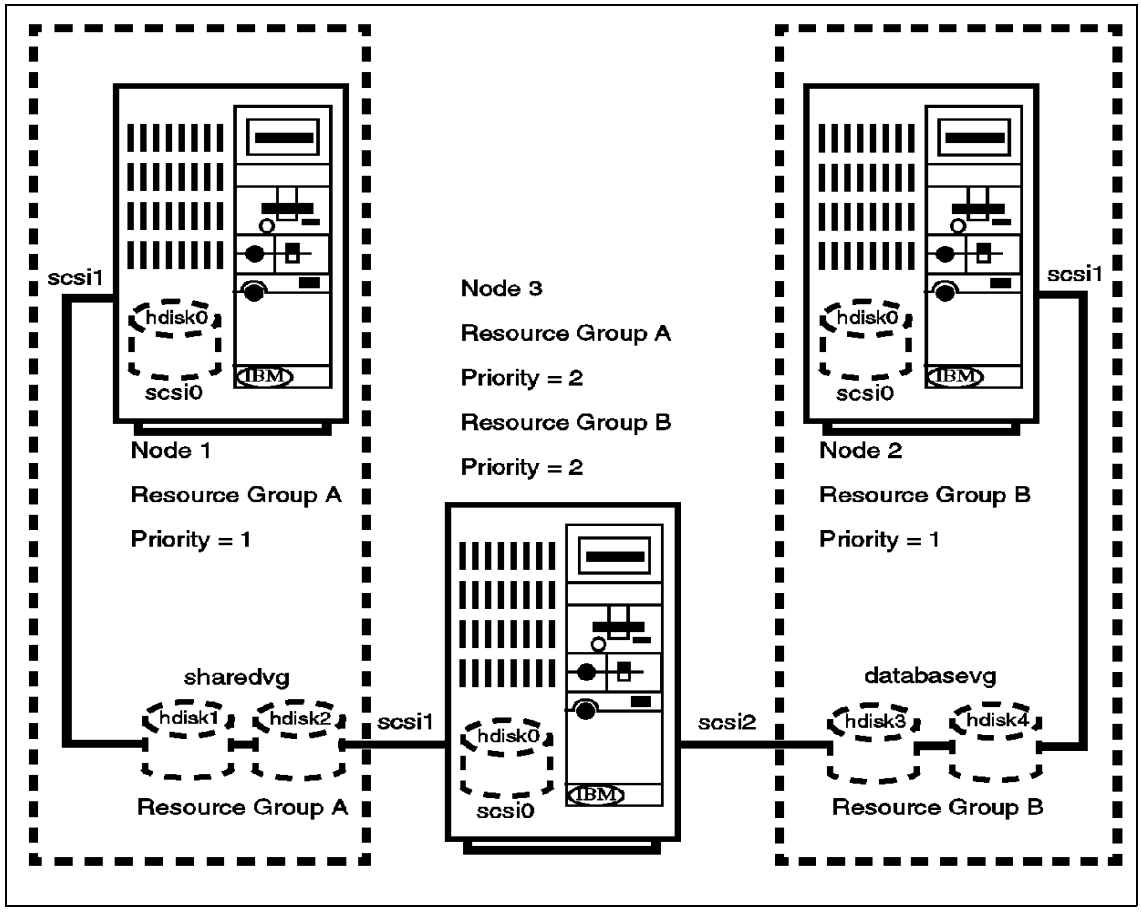


Figure 2-8 Third-party takeover configuration

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.

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.

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

2.5.3 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 Section 2.3, "Cluster networks" on page 19 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.

Note: It is a good idea to make a copy of /etc/inittab before you start configuring IPAT. See the last note in Section 4.2.4, "Defining adapters" on page 124.

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. Figure 2-9 on page 50 shows a single-network configuration.

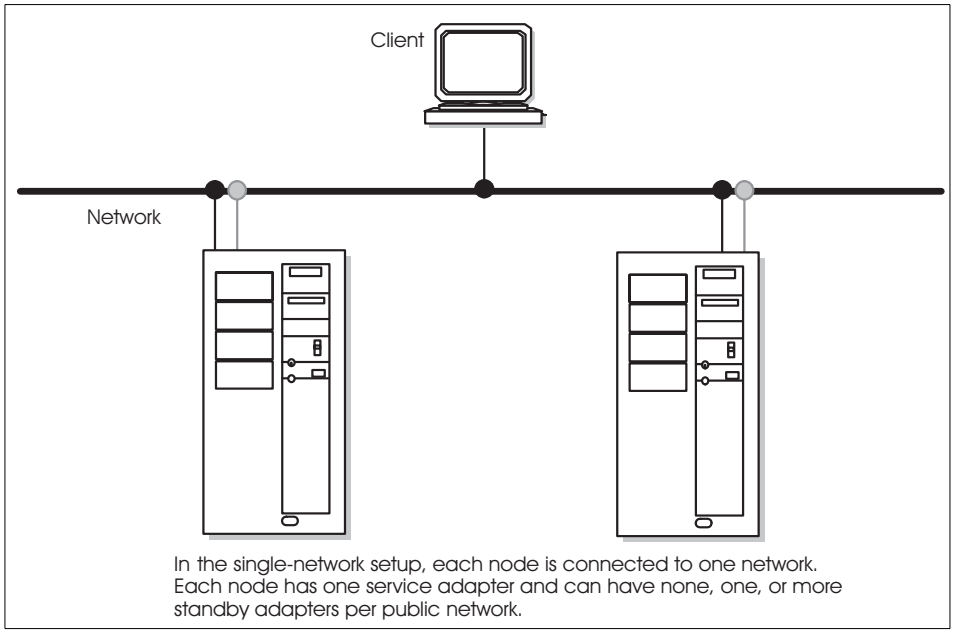


Figure 2-9 Single-network setup

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.

Figure 2-10 on page 51 shows a dual-network setup.

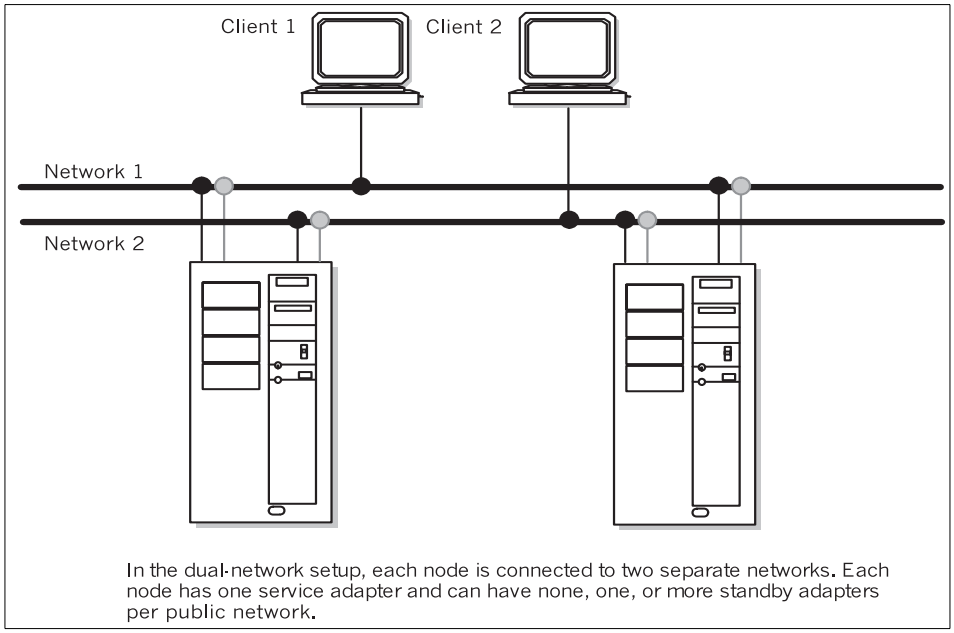


Figure 2-10 Dual-network setup

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.

Figure 2-11 on page 52 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.

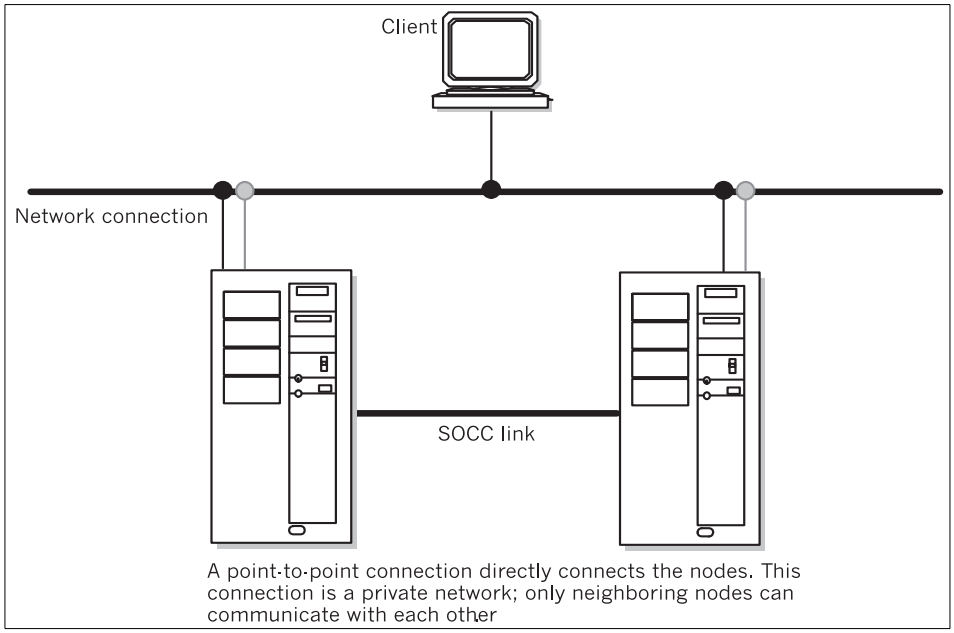


Figure 2-11 Point-to-point connection

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 the 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 Section 2.3.2, “Non-TCP/IP networks” on page 23).

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 host name, 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 re-configured 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 name server configuration.

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.

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.

2.5.4 NFS exports and NFS mounts

There are three items concerning NFS when doing the configuration of a resource group:

- Filesystems/Directories to Export

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

- Filesystems/Directories to NFS mount

  Filling in this field sets up what we call an *NFS cross mount*. Any file system or directory 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 or directory, and mounts the file system itself as part of its takeover processing.

- Network for NFS mount

  Choose a previously defined IP network where you want to NFS mount the file systems. This field is relevant only if you have filled in the previous field. The Service IP Label field should contain a service label that is on the network you choose. You can specify more than one service label in the Service IP Label field. It is highly recommended that at least one entry be an IP label on the network chosen here. If the network you have specified is unavailable when the node is attempting to NFS mount, it will seek other defined, available IP networks in the cluster on which to establish the NFS mount (this field is optional).

2.6 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 failover occurs. For more information about creating application server resources, see the *HACMP for AIX, Version 4.4: Installation Guide*, SC23-4278.

2.6.1 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.

2.6.2 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, for example, changes have to be transferred manually.

2.6.3 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.

2.6.4 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.

2.6.5 Critical/non-critical prioritization

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.

2.7 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.

2.7.1 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.

Note: HACMP without customization will cover only three events: node failure, network failure, and adapter failure.

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 4.4: 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

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.

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.

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.

2.7.2 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.

Single point of failure hardware component recovery

As described in “Special network considerations” on page 21, 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 (see Example 2-1).

Example 2-1 Sample screen for Add a Notification Method

```
                        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
```

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.

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.

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 "Single point of failure hardware component recovery" on page 61.

2.8 User ID planning

The following sections describe various aspects of user ID planning.

2.8.1 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 and passwords 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 Section 8.8, “User management” on page 277:

- 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
- Changing users passwords in the cluster

2.8.2 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. In HACMP Version 4.4.1, password management is also possible using C-SPOC. If the system is not configured to use NIS or DCE, the system administrator still has to distribute the password information found in the /etc/security/password file to all cluster nodes.

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

For more information about password management, see Section 8.8.3, “C-SPOC password enhancement” on page 278.

2.8.3 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 do not 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.

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.

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 to home directories of 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 does not 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.

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.

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

There are several important pieces of information about adapter slots.

PCI slots

Each PCI bus has a limit on the number of adapters it can support. Typically, this limit can range from two adapters to six adapters per bus. To overcome this limit, the system design can implement multiple PCI buses. You can use two different methods to add PCI buses to your system. These two methods are:

- Adding secondary PCI buses off the primary PCI bus.
- Implementing multiple primary buses.

Secondary PCI bus

If you want to increase the number of PCI slots when designing a system, add a secondary PCI bus. A PCI-to-PCI bridge chip can connect a secondary bus to a primary bus. Figure 3-1 shows how to use a primary PCI bus to increase the total number of PCI slots.

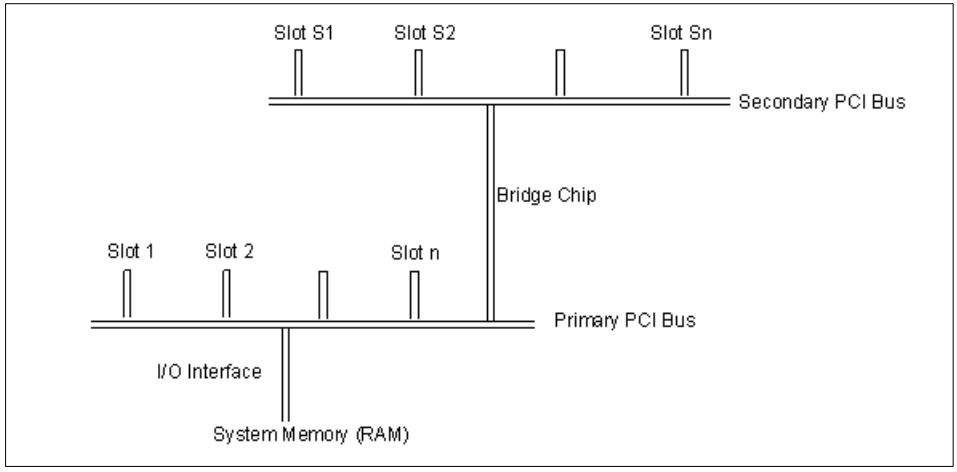


Figure 3-1 Secondary PCI bus

Because the slots on the secondary PCI bus must pass through the bridge chip, certain adapters on a secondary PCI bus may experience lower performance.

Multiple primary PCI buses

To add more PCI slots in a different way, design the system with two or more primary PCI buses. This design requires a more sophisticated I/O interface with the system memory. Figure 3-2 shows another method of increasing the number of PCI slots.

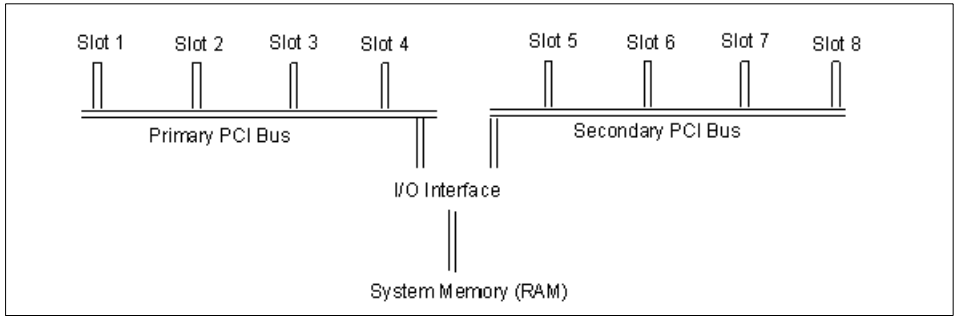


Figure 3-2 Multiple primary PCI buses

This design improves I/O performance over the secondary bus method because the I/O interface has created multiple parallel paths into the system memory.

Integrated adapters

The main processor board now integrates a number of devices, but they physically connect to one of the PCI buses. For this reason, some of the buses may only have two or three slots available to install adapters. Integrated PCI adapters include SCSI adapters and Ethernet adapters.

32-bit versus 64-bit PCI slots

Choosing between 32-bit and 64-bit slots influences slot placement and affects performance. Higher-speed adapters use 64-bit slots because they can transfer 64 bits of data for each data transfer phase. 32-bit adapters can typically function in 64-bit PCI slots; however, 32-bit adapters still operate in 32-bit mode and offer no performance advantage in a 64-bit slot. Likewise, most 64-bit adapters can operate in 32-bit PCI slots but the 64-bit adapter operates in 32-bit mode and reduces performance potential.

33 MHz versus 50 MHz 64-Bit PCI slots

Some systems (for example, 7025 Model F50 and 7026 Model H50) offer 50 MHz capability on 64-bit slots. Adapters capable of functioning at 50 MHz may take advantage of this. If you plug a 33 MHz adapter into a 50 MHz 64-bit slot, the slot switches to 33 MHz and also switches the remaining slots on this PCI bus to 33 MHz.

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

- *PCI Adapter Placement Reference Guide,* SA38-0538
- *RS/6000 Adapters, Devices and Cable Information for Micro Channel Bus Systems*, SA38-0533
- *RS/6000 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 Version 4.2.1, two new LVM commands were introduced: **mirrorvg** and **unmirrorvg**. These two commands were 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.

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
# mklvcopy hd10opt 2 hdisk1 # /opt 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. Use **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
   ```

3.1.3 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 Section 2.8, “User ID planning” on page 62)
- Network option settings
- /etc/hosts file and name server edits
- /.rhosts file edits

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.

To see how to tune I/O pacing, please refer to Section 7.3.1, “Tuning the system using I/O pacing” on page 199 for more information.

Checking network option settings

To ensure that HACMP for AIX requests for memory are handled correctly, you can check (on every cluster node) the thewall network option. The default values for this option depends on version AIX you are using and total amount of memory in your server.

To check a value of the thewall variable, you can use **no -a** command:

```
# no -a | grep thewall
```

Please refer to Section 7.3.3, “Increase amount of memory for communications subsystem” on page 201 for information on how to change this value.

Note: In AIX Version 4.3, the default value is 1/8 of real memory or 131072 (128 MB), which ever is smaller. In AIX Version 4.3.1, the default value is 1/2 of real memory or 131072 (128 MB), which ever is smaller. In AIX Version 4.3.2 and later, the default value depends on whether you are running on a Common Hardware Reference Platform (CHRP) machine or not. For non-CHRP machines, the default value is 1/2 of real memory or 262144 (256 MB), whichever is smaller. For CHRP machines, the default value is 1/2 of real memory or 1048576 (1 GB). A thewall is a run-time attribute.

Editing the /etc/hosts file and name server configuration

Make sure all nodes can resolve all cluster addresses. See Chapter 3 “Planning TCP/IP Networks” in the *HACMP for AIX 4.4.1: 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 name server 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:

```
192.168.100.35 austinboot
```

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

```
127.0.0.1   loopback localhost
```

cron and NIS considerations

If your 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 additional entry after the rcnfs position (Start NFS Daemons) to the /etc/inittab file with run level -a, calling the script in Example 3-1. Make this script executable and set permission to 600 for the root user. 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.

Example 3-1 Additional entry in /etc/inittab

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

Editing the /.rhosts file

Make sure that each node's service adapters and boot addresses are listed in the /.rhosts file on 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, such as maximum distance between nodes, have to be considered. You do not want to put two cluster nodes running a mission-critical application in the same room, for example.

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 is one additional point with cabling that should be considered. 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 problem, you should have at least two hubs.

As shown in Figure 3-3 on page 77, 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.

.

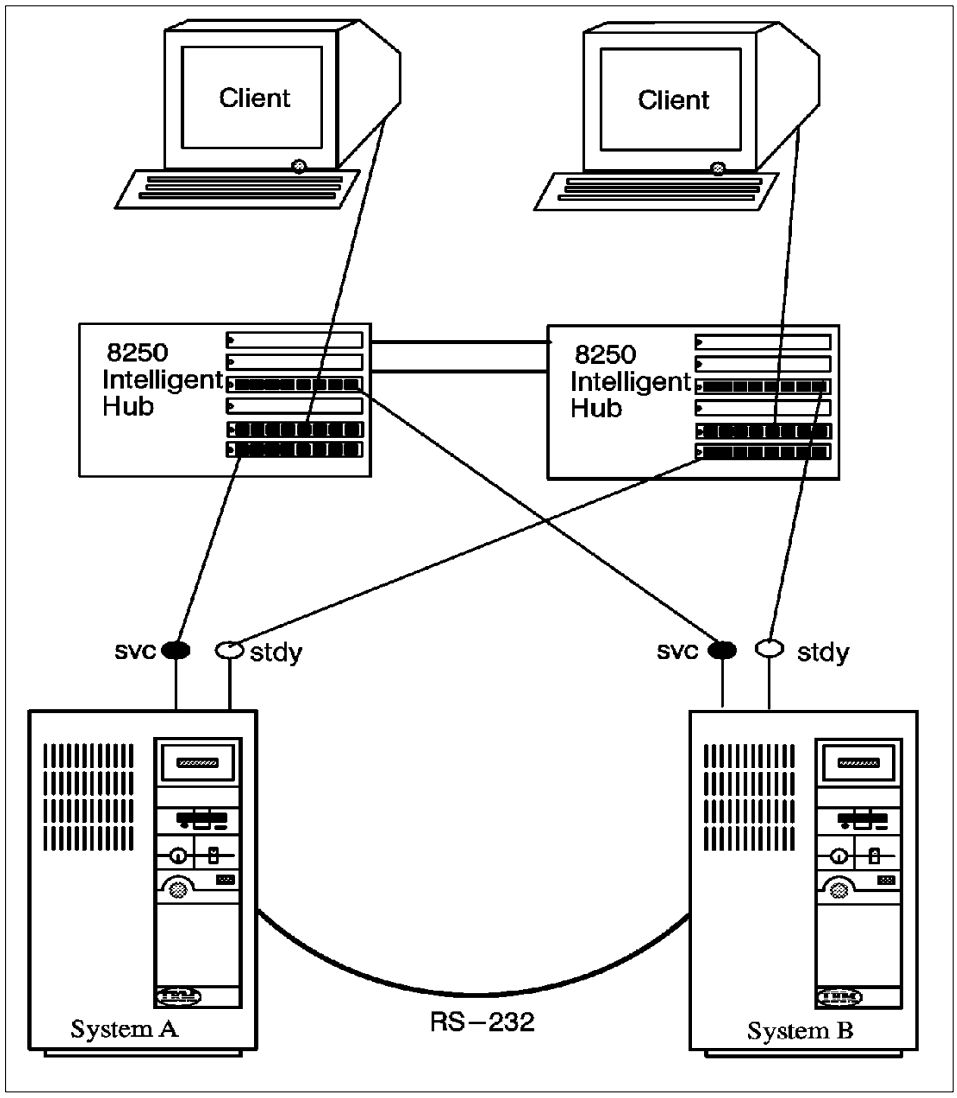


Figure 3-3 Connecting networks to a hub

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. All standby adapters *must* be together on the same subnet and 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 the identification of 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 4.4.1: Concepts and Facilities*, SC23-4276 guide for more information about using the same netmask for all adapters.

See Section 2.5.3, "IP address takeover" on page 48 for more detailed information.

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 network interfaces to detect bad IP addresses, incorrect subnet masks, and improper broadcast addresses.
- 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.

Cabling considerations

RS232 Cabling a serial connection requires a null-modem cable. As cluster nodes are often further apart than 60 m (181 feet), 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 “Supported SCSI adapters” on page 33). 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 3-5 on page 92, 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 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, #6225, #6230, #6235 or #6219), since these are the only current adapters that support the Multi-Initiator Feature.

Configuring RS232

Use the **smit tty** fast path 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.

Configuring target-mode SCSI

To configure a target-mode SCSI network on the Differential SCSI adapters, you have to enable the SCSI adapter's TARGET MODE feature 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 following command, before you can enable that feature:

```
# rmdev -l hdiskx
```

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.

Configuring target-mode SSA

Before configuring target-mode SSA, make sure you have fileset devices.ssa.tm.rte installed. To check if this fileset is installed use, the following command (See Example 3-2 for the output):

```
# lslpp -l | grep devices.ssa.tm.rte
```

Example 3-2 Output from lslpp command

```
# lslpp -l | grep devices.ssa.tm.rte
  devices.ssa.tm.rte         5.1.0.0  COMMITTED  Target Mode SSA Support
```

```
devices.ssa.tm.rte          5.1.0.0  COMMITTED  Target Mode SSA Support
```

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.

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. For more informations about testing, see Section 6.1.1, "Device state" on page 162.

- Serial (RS232)

  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 configuring a 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 node number. You should see the first command hanging until the second command is issued, and then showing its output.

- Target-mode SCSI

  After configuring a 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.

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 Section 2.4.1, “SSA disks” on page 25 or the following documents:

- *7133 SSA Disk Subsystems: Service Guide*, SY33-0185

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

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 cannot 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 what filesets are missing. 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 10-70         IBM SSA Enhanced RAID Adapter
ssar       Defined                 SSA Adapter Router
tmssar     Available               Target Mode SSA Router
tmssa1     Available               Target Mode SSA Device
```

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
```

Example 3-3 show the output of the command.

Example 3-3 Result of lsdev command for logical disk definitions

```
# lsdev -Cc disk | grep SSA
hdisk5  Available 10-70-L       SSA Logical Disk Drive
hdisk6  Available 10-70-L       SSA Logical Disk Drive
hdisk7  Available 10-70-L       SSA Logical Disk Drive
hdisk8  Available 10-70-L       SSA Logical Disk Drive
hdisk9  Available 10-70-L       SSA Logical Disk Drive
hdisk10 Available 10-70-L       SSA Logical Disk Drive
hdisk11 Available 10-70-L       SSA Logical Disk Drive
hdisk12 Available 10-70-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 IOCTL 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
```

Example 3-4 shows the output of the command.

Example 3-4 Result of lsdev command for physical disk definitions

```
# lsdev -Cc pdisk|grep SSA
pdisk0  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk1  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk2  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk3  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk4  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk5  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk6  Available 10-70-P 2GB SSA C Physical Disk Drive
pdisk7  Available 10-70-P 2GB 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 Serial Storage Architecture 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.

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 the following URL:

`http://www.storage.ibm.com/hardsoft/products/ssa/rs6k/index.html`

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 upgrade.tar file to 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 /usr/sys/inst.images directory 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), steps 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 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 that 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).

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 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 **smitty ssaraid** command. This will show you the menu shown in Example 3-5.

Example 3-5 Smit ssaraid 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
  List Status of Hot Spare Pools
  List Status of Hot Spare Protection for an SSA RAID Array
  List Components in a Hot Spare Pool
  Add a Hot Spare Pool
  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
  Change/Show/Delete a Hot Spare Pool
  Array Copy Services



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

First, you have to define Hot Spare disks, if any, then you need to change disks status to Array Candidate Disk using Change/Show Use of an SSA Physical Disk SMIT menu.

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.

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 “Supported SCSI adapters” on page 33. For the necessary adapter changes, see “Adapter SCSI ID and termination change” on page 93.

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.

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)

- SCSI-2 Differential System-to-System Cable

FC: 2423 (2.5m)

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) or

  FC: 2902 or 9202 (2.4m) or

  FC: 2905 or 9205 (4.5m) or

  FC: 2912 or 9212 (12m) or

  FC: 2914 or 9214 (14m) or

  FC: 2918 or 9218 (18m) or

- Terminator (T)

  Included in FC 2422 (Y-Cable)

- Cable Interposer (I)

  FC: 2919

  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 3-4 on page 91.

Figure 3-4 on page 91 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.

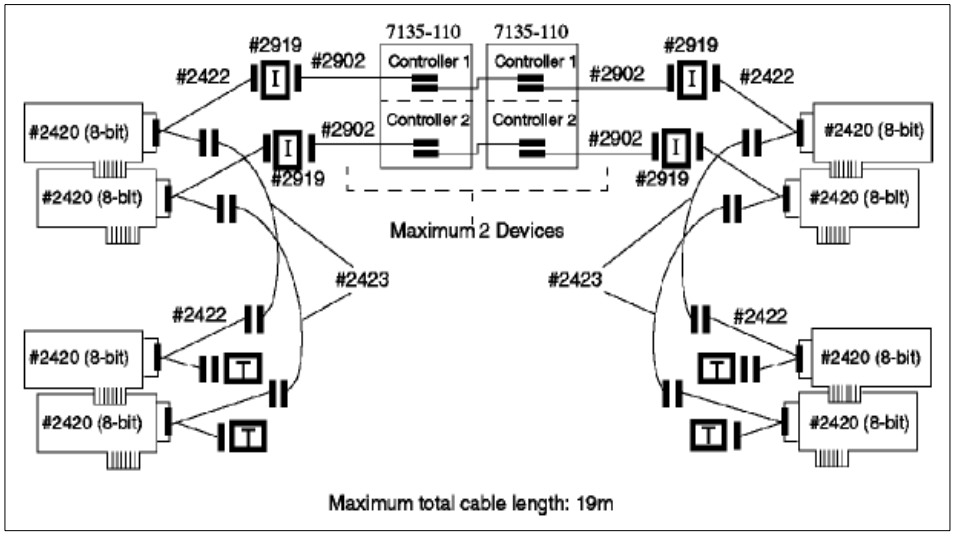


Figure 3-4 7135-110 RAIDiant arrays connected on two shared 8-Bit SCSI

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)

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

  FC: 2424 (0.6m) or

  FC: 2425 (2.5m)

  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) or

  FC: 2902 or 9202 (2.4m) or

  FC: 2905 or 9205 (4.5m) or

  FC: 2912 or 9212 (12m) or

  FC: 2914 or 9214 (14m) or

  FC: 2918 or 9218 (18m) or

- 16-Bit Terminator (T)

Included in FC 2426 (Y-Cable)

Figure 3-5 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 3-5, 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.

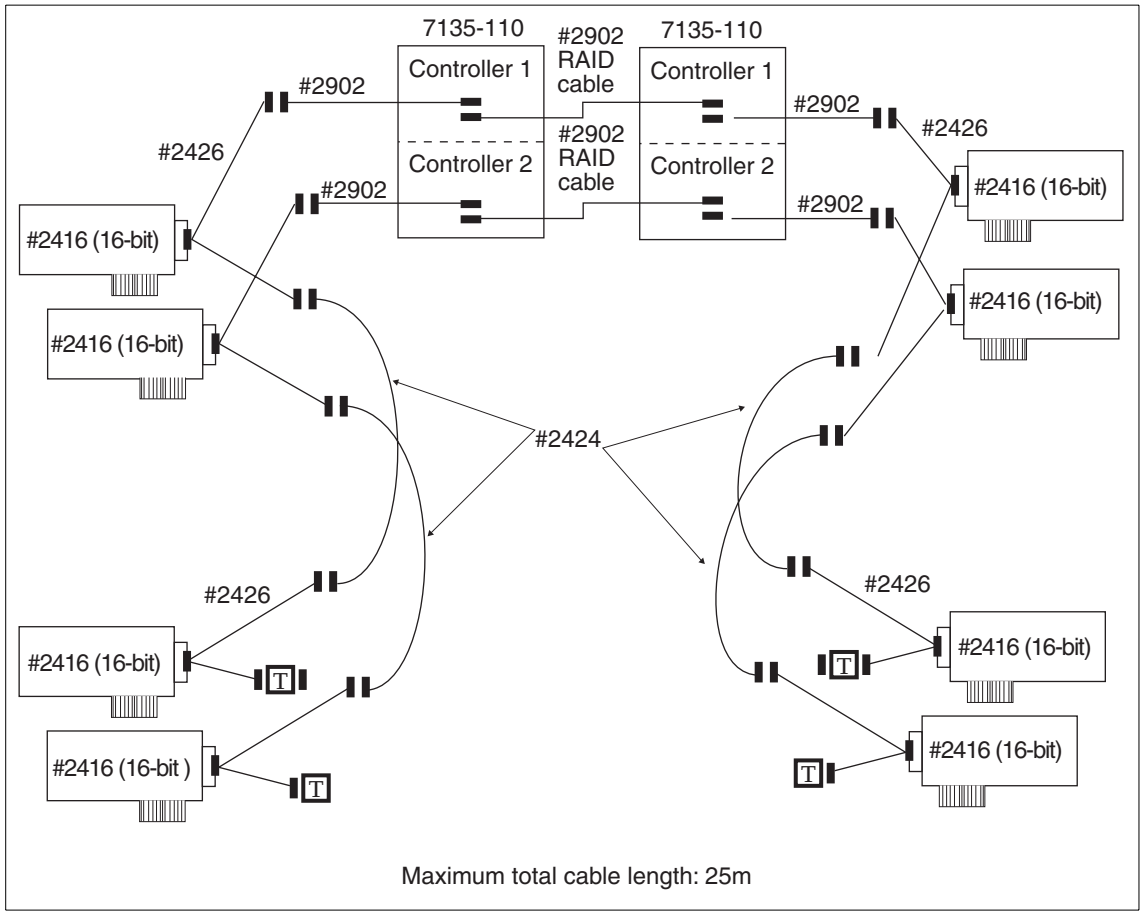


Figure 3-5 7135-110 RAIDiant arrays connected on two shared 16-Bit SCSI

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 3-6 and Figure 3-7 on page 94, respectively.

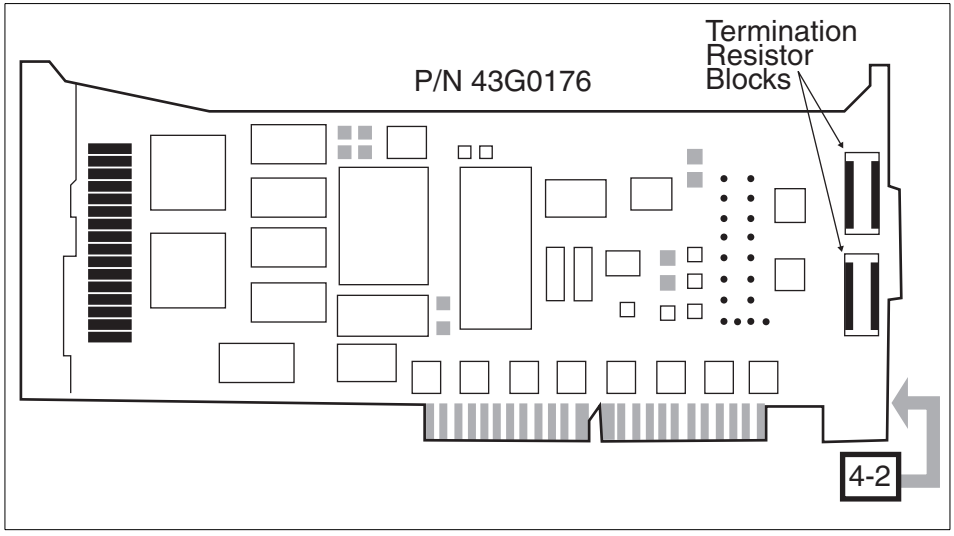


Figure 3-6 Termination on the SCSI-2 Differential Controller

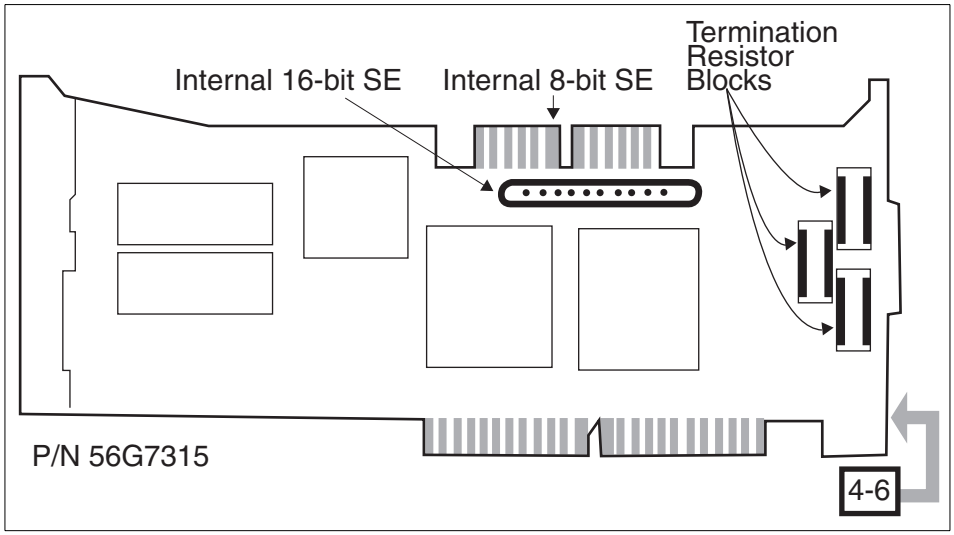


Figure 3-7 Termination on the SCSI-2 Differential Fast/Wide Adapters

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 (see Example 3-6).

Example 3-6 Smit chgscsi screen

```
                              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 (see Example 3-7).

Example 3-7 Characteristics of a SCSI adapter

```
            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 should be an *ascsi* device. Also, you need to change the external SCSI ID only (see Example 3-8).

Example 3-8 Characteristics of a SCSI adapter

```
              Change/Show Characteristics of a SCSI Adapter

SCSI adapter                          ascsi1
Description                           Wide SCSI I/O Control>
Status                                Av
ailable
```

```
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.

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** fast path 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. See Table 3-1 for the **smit mkvg** options.

Table 3-1 smit mkvg options (non-concurrent)

Options	Description
VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
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 to `no` so that the volume group can be activated as appropriate by the cluster event scripts.
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.
Create VG concurrent capable?	Set this field to `no` (leave default)
Auto-varyon concurrent mode?	Accept the default.

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 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 options shown in Table 3-2 to build the volume group.

Table 3-2 smit mkvg options (concurrent, non-RAID)

Options	Description
VOLUME GROUP name	The name of the shared volume group should be unique within the cluster.
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.
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 `yes` so that the volume group can be activated in concurrent access mode by the HACMP for AIX event scripts.

Options	Description
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`** fast path 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 3-3 lists the **`smit mkvg`** options.

Table 3-3 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.
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 3-4 lists the **smit crjfs** options.

Table 3-4 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. Follow these steps to rename the logical volumes:

Follow these steps to rename the logical volumes:

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** fast path 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.

Follow these steps to add copies to the logical volume:

1. Use the **smit mklvcopy** fast path 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 Section 3.4.3, “Mirroring strategies” on page 101 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

Follow these steps to test a file system:

1. To run a consistency check on each file system, enter the command:

   ```
   # 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.

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.

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 Section 3.4.6, “Alternate method - TaskGuide” on page 106.

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** fast path to import the volume group.

Table 3-5 lists the **smit importvg** options.

Table 3-5 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.

Options	Description
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.

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** fast path to change the characteristics of a volume group.

Table 3-6 lists the **smit crjfs** options.

Table 3-6 smit crjfs options

Options	Description
Activate volume group automatically at system restart?	Set this field to no.
A QUORUM of disks required to keep the volume group online?	This field is site-dependent. See Section 3.4.5, “Quorum” on page 103 for a discussion of quorum in an HACMP cluster.

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

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

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

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

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

Quorum at varyon

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

Quorum after varyon

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.

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** fast path.

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 failover. 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.

TaskGuide requirements

Before starting the TaskGuide, make sure:

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

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.

4

HACMP installation and cluster definition

This chapter describes issues concerning the actual installation of HACMP Version 4.4.1 and the definition of a cluster and its resources. It concentrates on the HACMP part of the installation, so we will assume that AIX is already installed.

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 4.4.1: 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 4.4.1: Installation Guide*, SC23-4278, gives a detailed list. The AIX Level of the server nodes has to be at AIX Version 4.3.3.25 or higher, for example, and the required free space in /usr must be confirmed. 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 install

You must install the HACMP for AIX software on each server machine. You can install the software from an installation server, from the installation media, or from a hard disk to which the software has been copied.

There are a number of filesets involved in an HACMP Installation. Here is a short overview of them, and what each one's 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.libHACMP Base Client Libraries
  cluster.base.client.rteHACMP Base Client Runtime
  cluster.base.client.utilsHACMP Base Client Utilities
  cluster.base.server.diagHACMP Base Server Diags
  cluster.base.server.eventsHACMP Base Server Events
  cluster.base.server.rteHACMP Base Server Runtime
  cluster.base.server.utilsHACMP 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.rteHACMP CSPOC Runtime Commands
  cluster.cspoc.cmdsHACMP CSPOC Commands
  cluster.cspoc.dshHACMP CSPOC dsh
  ```

- 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.demosHACMP Client Demos
  cluster.adt.client.samples.demosHACMP Client Demos Samples
  cluster.adt.client.samples.clinfoHACMP Client CLINFO Samples
  cluster.adt.client.samples.clstatHACMP Client Clstat Samples
  cluster.adt.client.includeHACMP Client include Files
  cluster.adt.client.samples.libclHACMP Client LIBCL Samples
  cluster.adt.server.samples.imagesHACMP Sample Images
  cluster.adt.server.demosHACMP Server Demos
  cluster.adt.server.samples.demosHACMP Server Sample Demos
  ```

- cluster.man.en_US

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

  ```
  cluster.man.en_US.cspoc.dataHACMP CSPOC Man pages
  cluster.man.en_US.client.dataHACMP Client Man pages
  cluster.man.en_US.server.dataHACMP Server Man pages
  ```

- cluster.man.en_US.haview

  This component contains the man pages for HAView:

  ```
  cluster.man.en_US.haview.dataHACMP 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.cspocHACMP CSPOC Messages
  cluster.msg.en_US.clientHACMP Client Messages
  cluster.man.en_US.haview.dataHACMP 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 Visual System Management Configuration Utility
  ```

- cluster.hativoli

  The filesets needed when you plan to monitor this cluster with Tivoli:

  ```
  cluster.hativoli.client
  cluster.hativoli.server
  ```

- cluster.haview

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

  ```
  cluster.haviewHACMP HAView
  ```

- cluster.man.en_US.haview.data

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

  ```
  cluster.man.en_US.haview.dataHACMP HAView Manpages
  ```

- cluster.msg.en_US.haview

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

  ```
  cluster.msg.en_US.haviewHACMP HAView Messages
  ```

  Note: Tivoli 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.shrvolgrpHAES Shr Vol Grp Task Guides
  ```

The installable images on the CRM installation media are listed here:

- 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.rteApplication Heart Beat Daemon
  ```

The installation of CRM requires the following software:

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

HAView installation notes

HAView requires Tivoli NetView for AIX. Install NetView before installing HAView.

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

Note: It is recommended that you install the HAView components on a node outside the cluster. Installing HAView outside the cluster minimizes the probability of losing monitoring capabilities during a cluster node failure.

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 4.4.1: Installation Guide*, SC23-4278 for details.

To install the base high availability subsystem on a server node:

1. Insert the installation medium and enter:

   ```
   # smit install_selectable_all
   ```

2. Enter the device name of the installation medium or Install Directory in the INPUT device/directory for software field and press Enter.
3. If you are unsure about the input device name or about Install Directory, press F4 to list available devices. Then select the proper drive or directory and press Enter. The correct value is entered into the INPUT device/directory field as the valid input device.
4. Press Enter to display the Install/Update From All Available Software screen.
5. Enter field values as follows:
   a. SOFTWARE to install

      Enter `cluster*` or `all` to install all server and client images, or press F4 for a software listing. If you press F4, a popup window appears, listing all installable software. Use the arrow keys to locate all software modules associated with the following Version 4.4.1 cluster images: **c**luster.base, cluster.cspoc, cluster.adt, cluster.man.en_US, cluster.vsm, cluster.haview, and cluster.man.en_US.haview.data, and cluster.taskguides. The cluster.base image (which contains the HACMP for AIX run-time executables) and the cluster.cspoc image are required and must be installed on all servers. If you select either cluster.base or cluster.cspoc, the other will be installed automatically.

      Next, press F7 to select either an image or a module. Then press Enter after making all selections. Your selections appear in this field.

Note that selecting cluster.base installs the base high availability subsystem and all associated messages.

b. PREVIEW ONLY?

Change the value to `no`.

c. OVERWRITE same or newer versions?

Leave this field set to `no`. Set it to `yes` if you are reinstalling or reverting to Version 4.4.1 from a newer version of the HACMP for AIX software.

d. AUTOMATICALLY Install requisite software

Set this field to `no` if the prerequisite software for Version 4.4.1 is installed or if the OVERWRITE same or newer versions? field is set to `yes`; otherwise, set this field to `yes` to install required software.

6. Enter values for the other fields as appropriate for your site.
7. Press Enter for confirmation in the SMIT screen.
8. Press Enter again.

Rebooting servers

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

Note: Read the HACMP Version 4.4.1 release_notes file in the /usr/lpp/cluster/doc directory for further instructions.

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 9 of the *HACMP for AIX 4.4.1: Installation Guide*, SC23-4278.
- Archive any localized script and configuration files to prevent losing them during an upgrade.
- Commit your current HACMP for AIX Version 4.* software (if it is applied but not committed) so that the HACMP for AIX 4.4.1 software can be installed over the existing version. To see if your configuration is already committed, enter:

```
# lslpp -h "cluster.*"
```

If the word COMMIT is displayed under the Action header for all the cluster filesets, continue to the next step. If not, run the **smit install_commit** utility before installing the Version 4.4.1 software (see Example 4-1).

Example 4-1 How to commit applied cluster software.

```
          Commit Applied Software Updates (Remove Saved Files)

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

                                                      [Entry Fields]
* SOFTWARE name                                      [cluster.*]              +
  PREVIEW only? (commit operation will NOT occur)     no                      +
  COMMIT requisites?                                  yes                     +
  EXTEND file systems if space needed?                yes                     +
  DETAILED output?                                    no                      +


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

- Make a mksysb backup on each node. This saves a backup of the AIX root volume group.

  Note: If using SCSI disks, and, for some reason, you do restore a mksysb back onto your system, you will need to reset the SCSI IDs on the system.

- Save the current configuration, using the cluster snapshot utility, and save any customized event scripts in a directory of your own. (To see how to save and restore cluster configurations, see the *HACMP for AIX 4.4.1: Administration Guide*, SC23-4279.)

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.

Upgrading from 4.2.2, 4.3.0, or 4.3.1 to 4.4.1 version

The following procedure applies to you upgrading your existing HACMP or HANFS software Version 4.2.2, 4.3.0, or 4.3.1 to HACMP for AIX, Version 4.4.1 in a two-node or multi-node cluster environment.

Note: Earlier software versions are no longer supported and should be removed prior to upgrading to HACMP for AIX Version 4.4.1.

If you plan to upgrade from HACMP Versions 4.2.2 through 4.3.1 to Version 4.4.1, you must perform a migration installation of AIX to upgrade to AIX Version 4.3.3.25 or later on all cluster nodes.

Note: You cannot migrate directly from HACMP software Version 4.2.2 or 4.3.0 to Version 4.4.1. To migrate to version 4.4.1, you must first do an upgrade from Version 4.2.2 or 4.3.0 to Version 4.3.1.

HACMP conversion utilities provide easy conversion between the HACMP versions and these products:

- HACMP Version 4.3.1 to HACMP Version 4.4.1
- HACMP ES Version 4.3.1 to HACMP ES Version 4.4.1
- HACMP Version 4.4.0 to HACMP Version 4.4.1
- HACMP ES Version 4.4.0 to HACMP ES Version 4.4.1
- HACMP Version 4.4.1 to HACMP ES Version 4.4.1
- HANFS Version 4.3.1 to HACMP Version 4.4.1

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 4.4.1: Administration Guide*, SC23-4279.
2. Shut down the first node (gracefully with takeover) using the **`smit clstop`** fast path. For this example, shut down Node A. Node B will take over Node A's resources and make them available to clients.
3. Perform a Migration Installation 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 4.3.3. Product (application) files and configuration data are also saved.
4. Check the Migration Installation. Verify that all the disks are available. Run the **`lppchk -v`** and **`oslevel`** commands to ensure that the system is in a stable state.

Install HACMP for AIX 4.4.1 on node A

1. After upgrading AIX and verifying that the disks are correctly configured, install the HACMP for AIX Version 4.4.1 software on Node A. For a short description of the filesets, please refer to Section 4.1.1, “First time install” on page 108.
2. The HACMP conversion utilities are **cl_convert** and **clconvert_snapshot.** Upgrading HACMP software to the newest version involves converting the ODM from a previous release to that of the current release. When you install HACMP, **cl_convert** is run automatically. However, if installation fails, you must run **cl_convert** from the command line. The **clconvert_snapshot** is not run automatically during installation, and must always be run from the command line.

   The **cl_convert** utility logs conversion progress to the /tmp/clconvert.log file so that you can gauge conversion success. This log file is regenerated each time **cl_convert** or **clconvert_snapshot** is executed
3. Run **clconvert_snapshot** to upgrade cluster snapshots. For more information on **cl_convert** and **clconvert_snapshot**, refer to the *HACMP for AIX, Version 4.4.1: Administration Guide*, SC23-4279-02.

   Note: Root user privilege is required to run a conversion utility. You must know the HACMP version from which you are converting in order to run these utilities.

4. Start the HACMP for AIX Version 4.4.1 software on Node A using the **smit clstart** fast path. After HACMP is running, start the previous version of HACMP software on Node B, if it is not still running. Check to ensure that the nodes successfully join the cluster.

   Note: 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

5. Repeat Steps 2 through 8 on Node B on remaining cluster nodes, one at a time.
6. When the last node has been upgraded to both AIX 4.3.3 or later and HACMP for AIX 4.4.1, the cluster install/upgrade process is complete.

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

Check upgraded configuration

1. If using tty devices, check that the tty device is configured as a serial network using the **smit chgtty** fast path.
2. 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.
3. Verify the cluster topology on all nodes using the **clverify** utility.
4. Check that custom event scripts are properly installed.
5. Synchronize the node configuration and the cluster topology from Node A to all nodes (this step is optional).
6. It is recommended that you test the upgraded cluster to ensure proper behavior.

Client-only migration

If you are migrating from an HACMP for AIX, Version 4.2.2 through 4.3.1 server node to a client-only node running Version 4.4.1, 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 4.4.1. 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.

4.1.3 Migrating from HANFS to HACMP Version 4.4.1

HACMP Version 4.4.1 provides HANFS users a node-by-node migration path from HANFS Version 4.3.1 to HACMP Version 4.4.1. HACMP Version 4.4.1 supports the NFS export behavior of the HANFS cluster.

You can perform a migration from a running HANFS Version 4.3.1 cluster to a running HACMP Version 4.4.1 cluster without bringing the cluster offline, thereby keeping all cluster resources running during the migration process.

Prerequisites

In order to perform HANFS 4.3.1 to HACMP 4.4.1 node-by-node migration, the following prerequisites must apply:

1. Both nodes in the cluster must have HANFS Version 4.3.1 installed and committed. (If you are running an older versions of HANFS, you must upgrade to Version 4.3.1 before migrating to HACMP Version 4.4.1).
2. Both nodes in the cluster must be up and running the HANFS software.
3. The cluster must be in a stable state.

Note: As in any migration, do not attempt to make any changes to the cluster topology or configuration once you have started the migration process. You can make any necessary changes after both nodes have finished the migration process.

Procedure for HANFS to HACMP Version 4.4.1 node-by-node migration

The prerequisite steps are:

- If your version of HANFS is less than 4.3.1, upgrade both nodes to HANFS Version 4.3.1.
- Upgrade AIX to Version 4.3.3.25 on both nodes.

Take the following steps to perform a node-by-node migration from HANFS Version 4.3.1 to HACMP Version 4.4.1:

1. Stop cluster services on one of the nodes running HANFS Version 4.3.1 using the “graceful with takeover” method.
2. Install HACMP Version 4.4.1 on the node. See Section 4.1, “Installing HACMP” on page 108 for more details.

Note: The HACMP installation utility first checks the current version of HANFS. If your HANFS software is an earlier version than 4.3.1, you will see an error message and the installation will be aborted. If you are running Version 4.3.1, you will see a message indicating that a migration has been requested and is about to proceed.

3. After you install the HACMP software, reboot the node.

4. Using the SMIT Start Cluster Services screen, start the HACMP software.
5. Two events take place when you start HACMP:
   - The HACMP Cluster Manager communicates with the HANFS Cluster Manager on the other node.
   - The HACMP software reacquires the resources assigned to the node.
6. Repeat steps 1 through 4 for the other cluster node.

Backout procedure

If the migration process fails (a node crash, for example):

- If the first node fails, you must uninstall all HACMP and/or HANFS software, reinstall the HANFS Version 4.3.1 software, and re-synchronize the cluster from the other HANFS cluster node.
- If the second node fails, you must uninstall all HACMP and/or HANFS software, reinstall the HACMP Version 4.4.1 software, and re-synchronize the cluster from the other (already migrated) HACMP Version 4.4.1 cluster node.

4.1.4 Installing the concurrent resource manager

To install the concurrent access feature on cluster nodes, complete the following procedure:

Note: In HACMP for AIX Version 4.4.1 environments, concurrent access is available using only an IBM 7135-110 or 210 Disk Array, an IBM 7137 Disk Array, IBM 2105-B09 and 100 Versatile Storage Servers, IBM 2105-E10 and E-20 Enterprise Storage Servers, an IBM 7133 SSA disk subsystem, or an IBM 9333 disk subsystem. RAID devices from other manufacturers may not support concurrent access.

Use the **`smit install_selectable_all`** fast path to load the concurrent access install image on a node. See the section Installation Choices for a list of software images to install. Depending on the AIX level installed on your system, not all images are required.

To install the concurrent access software on a server:

1. Insert the installation media and enter:

   ```
   # smit install_selectable_all
   ```

2. Enter the device name of the installation media or Install Directory in the INPUT device/directory for software field and press Enter. If you are unsure about the input device name or about Install Directory, press F4 to list available devices. Then select the proper media or directory and press Enter.

The correct value is entered into the INPUT device/directory field as the valid input device.

3. Press Enter. SMIT refreshes the screen.
4. Enter field values as follows:
   a. SOFTWARE to install

      Change the value in this field to cluster.clvm. Note that the run-time executables for the HACMP for AIX software and associated images are automatically installed when you select this image.

      Note: If using Oracle Parallel Server, you must also install cluster.hc.

   b. PREVIEW ONLY?

      Change the value to `no`.

   c. OVERWRITE same or newer versions?

      Leave this field set to `no`. Set it to `yes` if you are reinstalling or reverting to Version 4.4.1 from a newer version of the HACMP for AIX software.

   d. AUTOMATICALLY Install requisite software

      Set this field to `no` if the prerequisite software for Version 4.4.1 is installed or if the OVERWRITE same or newer versions? field is set to `yes`; otherwise, set this field to `yes` to install required software.
5. Enter values for other fields appropriate for your site.
6. Press Enter when you are satisfied with the entries. SMIT responds:

   ```
   ARE YOU SURE?
   ```
7. Press Enter again.

Note: Read the HACMP Version 4.4.1 release_notes file in the /usr/lpp/cluster/doc directory for further instructions.

4.1.5 Problems during the installation

If you experience problems during an installation, the installation program automatically performs a cleanup process. If for some reason the cleanup is not performed after an unsuccessful installation, do the following:

1. Enter the following command:

   ```
   # smit maintain_software
   ```
2. Select Clean Up After an Interrupted Installation.

3. Review the SMIT output (or examine the /smit.log file) for the interruption's cause.
4. Fix any problems and repeat the installation process.

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.

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 Chapter 6, “Changing the cluster topology”, in the *HACMP for AIX, Version 4.4.1: 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. The good naming convention is setting names in conjunction to the place, where cluster is or/and organization name, for example: C1ITSOAUTX, where C = Cluster, 1 = First cluster, ITSO = Organization, AU = Austin, and TX = Texas.

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 space 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 (see Figure 4-1).

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.

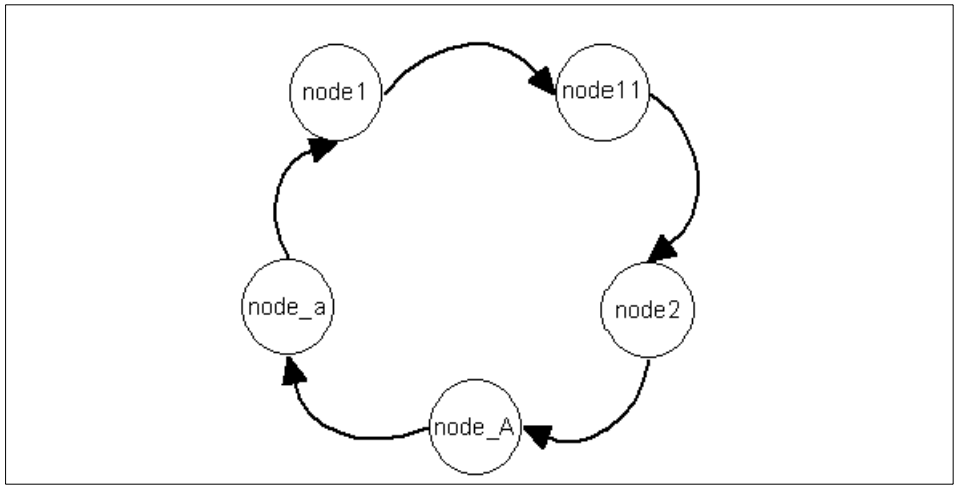


Figure 4-1 Logical ring

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 Chapter 6, "Changing the cluster topology", in the *HACMP for AIX Version 4.4.1: Administration Guide*, SC23-4279 for more information.

4.2.3 Defining networks

Once you have defined all cluster node names, continue setting up your cluster topology by defining the HACMP networks and associated adapters. You can define the cluster topology by defining the networks and adapters as in previous releases, or you can take advantage of the automatic network discovery feature. You can choose to define boot adapters on each node, in the process defining the network, or you can define the networks and then the adapters. You can use automatic discovery at various points in the process no matter which way you choose to start the topology configuration process.

Defining networks with automatic network discovery

Automatic network discovery simplifies the process of defining HACMP cluster topology. Using SMIT, you select the type of network to add (IP-based or non IP-based).Then you can run the automatic discovery process either locally or cluster-wide. The software gathers information about the configured physical networks (sets of subnets) and associated network adapters, and automatically creates the necessary lists for you to select from to complete the HACMP network and adapter definitions.

Cluster-wide network discovery uses C-SPOC to reach all nodes. However, each node must have a boot/service adapter configured before C-SPOC can reach the node. Use local network discovery when you configure the networks for the first time.

You need only assign a name to each network in SMIT, and then add the associated adapters to the configuration. You define the adapters by selecting from a list of known labels instead of typing them. In addition, HACMP will fill in the network name, type, and attribute fields in the SMIT screen, speeding up the process.

The general procedure is:

- On the first node, create your first network. The subnet list that comes up will be only the subnets on the local host, since no other nodes can be reached at this point. You can add additional subnets to this network.
- Then create adapters and add the boot/service adapter labels to this network. Pressing F4 on the SMIT IP Label field gives you a list of IP Labels that belong to any of the subnets of the network. This list consists of two parts: first, the Configured Interfaces, and then the ones in the /etc/hosts file. Select a label and fill in the node field, and the rest is automatically filled out.
- After you add a boot/service adapter to a node, C-SPOC can harvest information from that node.

- If you have networks that are not on the local node, but only on other nodes, you can go back at this point and create the other networks, since the list of subnets can be gathered from the newly reachable nodes also. Then you can add adapters to that network.

Prerequisites and limitations

- Before HACMP can create networks for you, all nodes must have HACMP Version 4.4.1or later installed. The network discovery/creation functionality will not work in earlier versions of HACMP or in mixed-version clusters.
- The /etc/hosts file must contain all adapter labels and associated IP addresses. HACMP uses this file to gather adapter information for creating networks.

Steps for defining IP networks

1. In SMIT, select **Cluster Topology** -> **Configure Networks** -> **Configure IP-Based Network** -> **Discover Current Network Configuration** -> **Local Network Configuration**. When you press Enter, the software builds a list of logical interfaces configured on this node, with additional information, and saves this information in a local file. (This file is updated when you run another discovery operation.)
2. Return to the Configure IP-Based Network screen, and select Add a Network. SMIT displays the Add an IP Network screen.
3. Fill in the fields as follows:

   Network Name — Name the network, using no more than 31 alphanumeric characters and underscores. Do not use reserved names to name the network. (Note that the "network" is actually a collection of subnets found to be on the same physical network.)

   Network Attribute — `Public` or `private`. The default is public. Ethernet, Token-Ring, FDDI, ATM LANE, and SLIP are public networks. SOCC, ATM, and the SP Ethernet are private networks.

   Network Type — Select a network type from the F4 list or enter the type directly. If you do not enter a value, the type will be filled in according to the first subnet chosen on the following field.

   Subnets — Press F4 to see the list of discovered subnets on the local node. Select the subnets for the boot and standby adapters. Choose subnets for the IP Network from the list of available subnets.
4. When you finish adding subnets, press Enter.

5. Press Enter to complete the definition of the network you have just named.
6. Repeat these steps to define all the IP HACMP networks.

Note: To create networks not connected to the first node, you first need to define the adapters for the first network, including those for other cluster nodes. Once a node has an adapter on one network, you can create the other networks connected to that node.

Steps for defining non IP-based networks

1. In SMIT, select **Cluster Topology** -> **Configure Networks** -> **Configure Non IP-Based Networks** -> **Add a Network**.
2. When you press Enter, SMIT displays the Add a Non IP-based Network screen.
3. Fill in the fields on the Add a non IP-based Network screen as follows:

   Network Name Name the network, using no more than 31 alphanumeric characters and underscores. Do not use reserved names to name the network. For a list of reserved names, refer to Chapter 8, “Verifying Cluster Configuration”, in the *HACMP for AIX, Version 4.4.1: Administration Guide*, SC23-4279.

   Network Type Valid types are `RS232`, `tmssa`, and `tmscsi`.
4. Note that the Network Attribute is automatically serial, and does not appear on the SMIT screen. Press Enter to complete the definition of the network you have just named.
5. Repeat these steps to define all the non IP-based HACMP networks.

4.2.4 Defining adapters

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

Note: Note that you can start configuring the cluster topology by defining adapters instead of networks. You can add networks on the Choose Network for New Adapter screen.

Defining adapters for IP-based networks

Once the networks are defined, you can discover the network configuration. Then the Add an Adapter screen is simple to complete. You can select from a list of adapters known to be on the physical network.

1. In SMIT, select **Cluster Topology** -> **Configure Adapters** -> **Configure IP-based Adapters** -> **Discover Current Network Configuration** and press Enter. The information is saved in a file.
2. In SMIT, return to Configure IP-based Adapters and select Add an Adapter.
3. SMIT displays the Choose Network for New Adapter screen. Select the network for which to add adapters. (If you have not previously defined any networks, you can choose to Add an Adapter on a new Network to create the network.)
4. Select an existing IP-based network. The Add an IP-based Adapter screen appears.
5. Fill in the values as follows:

Adapter IP Label	Press F4 and choose the desired adapter from the subnet list.
Network Type	This field is automatically filled in.
Network Name	This field is automatically filled in.
Adapter Function	Indicate whether the adapter's function is `service`, `standby`, or `boot`. Press the Tab key to toggle the values. Each node must have an adapter that performs the service function for each public, private, and serial network in which it participates. A node can also have one or more standby adapters, or none, configured for each public network. Serial and private networks do not normally have standby adapters. If the cluster uses IP address takeover (IPAT) or rotating resources, each node that can have its IP address taken over must have an adapter configured for the boot function. (Remember that a boot “adapter” is not actually a separate adapter, but a second address for the service adapter.) Keep in mind that the netmask for all adapters in an HACMP network must be the same to avoid communication problems between standby adapters after an adapter swap and after the

	adapter is reconfigured with its original standby address.
Adapter Identifier	This field is populated automatically (based on what you entered in the IP Label field) when you hit Enter to put your changes in effect.
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 Chapter 3, “Planning TCP/IP networks”, in the *HACMP for AIX, Version 4.4.1: Planning Guide*, SC23-4277 for more information on hardware address swapping. This facility is supported only for Ethernet, Token Ring, FDDI, and ATM adapters. It does not work with the SP Switch.

Note: Note that the hardware address is 12 digits for Ethernet, Token-Ring, and FDDI, and 14 digits for ATM.

Node Name	Define a node name for all adapters except for those 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)
Netmask (optional)	If you enter a netmask, the subnet will automatically be added to the network.

6. Press Enter after filling in all required fields. HACMP now checks the validity of the adapter configuration.
7. Repeat until each node has all appropriate adapters defined.

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 for AIX configuration. You should configure at least one standby adapter for each public network.

When IPAT is configured, the run level of the IP-related entries (for example, rctcpip, rcnfs, and so on) 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.

Defining adapters for non IP-based networks

Take the following steps to define service adapters on your non IP-based networks:

1. On SMIT, select **Cluster Topology** -> **Configure Adapters** -> **Configure Non IP-based Adapters** -> **Add an Adapter** > **Choose Network for new Adapter**.
2. Select the non IP-based network to which you want to add the adapters. The Add a Non IP-based Adapter screen appears.
3. Fill in the values as follows:

Network Name	This field is automatically filled in (unless you chose Add an Adapter on a New Network in the previous screen).
Network Type	This field is automatically filled in (RS232, tmssa, or tmscsi) when you enter the device name. If you chose Add an Adapter on a New Network in the previous screen, you must toggle this field.
Adapter Label	Enter a name for this adapter.
Device Name	Enter a device file name. RS232 serial adapters must have the device file name /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
Node Name	Define a node name for all serial service adapters.

4. Press Enter after filling in all required fields. HACMP now checks the validity of the adapter configuration. You may receive warnings if a node cannot be reached.
5. Repeat until each node has all appropriate adapters defined.

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 Chapter 6, “Changing the cluster topology”, in the *HACMP for AIX, Version 4.4.1: Administration Guide*, SC23-4279, for more information.

4.2.5 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.

Note: 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 Chapter 6, “Changing the cluster topology”, in the *HACMP for AIX, Version 4.4.1: 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.

4.2.6 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.

Note: 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.

Note: Before attempting to synchronize a cluster configuration, ensure that all nodes are powered on, that the HACMP for AIX software is installed, and that the /etc/hosts and /.rhosts files on all nodes include all HACMP for AIX 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 "Kerberised" version of remote commands to accomplish the synchronization.

To synchronize a cluster definition across nodes, complete the following steps:

1. Enter the following fast path in SMIT:

   ```
   # smit hacmp
   ```

2. From the Cluster Configuration menu, select **Cluster Topology** -> **Synchronize Cluster Topology** and press Enter. SMIT displays the Synchronize Cluster Topology screen.

3. Enter field data as follows:

   Skip Cluster Verification — By default, this field is set to `no` and the cluster topology verification program is run. To save time in the cluster synchronization process, you can toggle this entry field to `yes`. By doing so, cluster verification will be skipped.

   Ignore Cluster Verification Errors — By choosing `yes`, the result of the cluster verification is ignored and the configuration is synchronized even if verification fails. By choosing `no`, the synchronization process terminates; view the error messages in the system error log to determine the configuration problem.

   Emulate or Actual — If you set this field to `Emulate`, the synchronization is an emulation and does not affect the Cluster Manager. If you set this field to `Actual`, the synchronization actually occurs, and

any subsequent changes affect the Cluster Manager. `Actual` is the default value.

4. After you specify values and press Enter, SMIT displays a screen asking if you want to continue with the synchronization. If you want to proceed, press Enter. 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
- SCSI tape drives
- 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 Section 2.5.1, “Resource group options” on page 41 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. See the following fields in the Add Resource Group screen:

Resource Group Name — Enter an ASCII text string that identifies the resource group. The resource group name can include alphabetic or numeric characters and underscores. Use no more than 31 characters. Duplicate entries are not allowed.

Node Relationship — Toggle the entry field between `Cascading`, `Concurrent`, and `Rotating`.

Participating Node Names — Enter the names of the nodes that you want to be members of the resource chain for this resource group. Enter the node names in order from highest to lowest priority (left to right). Leave a space between node names. Priority is ignored for concurrent resource groups.

The relationship can be one of cascading, rotating or concurrent. See Section 2.5.1, “Resource group options” on page 41 for details.

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).

General considerations for configuring resources

- You cannot configure a resource group until you have completed the information on the Add a Resource Group screen.
- If you plan to configure AIX Fast Connect services, you must remove any configured AIX Connections services from the resource group first. You cannot have both in the same resource group.
- 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.
- 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.

- Failure to use Kerberos or /.rhosts with a cascading without fallback resource group will result in resource group failure.
- HACMP limits the number of nodes participating in a cascading without fallback resource group to two.
- If you plan to automatically import volume groups, make sure that the appropriate volume groups have been previously discovered by HACMP. To collect information about the volume group configuration on a local node or cluster-wide, use the Discover Current Volume Group Configuration panel under the Cluster Resources menu in SMIT.
- Tape resources may not be part of concurrent resource groups.

The following describes the different possibilities for resources that might be added to a resource group. See the Configure a Resource Group screen from SMIT menu:

- Resource Group Name

  Reflects the choice you made on the previous screen; the resource group to configure.

- Node Relationship

  Reflects the fallover strategy entered when you created the resource group.

- Participating Node Names

  Reflects the names of the nodes that you entered as members of the resource chain for this resource group. Node names are listed in order from highest to lowest priority (left to right), as you designated them.

- Service IP Label

  If IP address takeover is being used, list the IP labels to be taken over when this resource group is taken over. Press F4 to see a list of valid IP labels. These include addresses which rotate or may be taken over.

- Filesystems (default is All)

  Leave this field blank, if you want ALL file systems in the specified volume groups to be mounted by default when the resource group, containing this volume group, is brought online.

If you leave the Filesystems (default is All) field blank and specify the shared volume groups in the Volume Groups field, then all file systems will be mounted in the resource group. If you leave the Filesystems field blank and do not specify the volume groups in the field, no file systems will be mounted. You may also select individual file systems to include in the resource group. Press F4 to see a list of the file systems. In this case, only the specified file systems will be mounted when the resource group is brought online.

- Filesystems (default is All)

  A valid option ONLY for non-concurrent resource groups. When you select a file system in this field, the HACMP for AIX software determines the correct values for the Volume Groups and Raw Disk PVIDs fields. If you will be configuring Fast Connect file shares, be sure their file systems are configured here.

- Filesystems Consistency Check

  Identifies the method of checking consistency of file systems, fsck (default) or logredo (for fast recovery).

- Filesystems Recovery Method

  Identifies 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.)

- Filesystems/Directories to Export

  Identifies the file systems or directories to be exported. The file systems should be a subset of the file systems listed above. The directories for export should be contained in one of the file systems listed above. Press F4 for a list.

- Filesystems/Directories to NFS Mount

  Identifies the file systems or directories to NFS mount. All nodes in the resource chain will attempt to NFS mount these file systems or directories while the owner node is active in the cluster.

- Network for NFS Mount (This field is optional.)

  Choose a previously defined IP network where you want to NFS mount the file systems. The F4 key lists valid networks. This field is relevant only if you have filled in the previous field. The Service IP Label field should contain a service label which is on the network you choose.

Note: You can specify more than one service label in the Service IP Label field. It is highly recommended that at least one entry be an IP label on the network chosen here.

If the network you have specified is unavailable when the node is attempting to NFS mount, it will seek other defined, available IP networks in the cluster on which to establish the NFS mount.

- Volume Groups

  Identifies the shared volume groups that should be varied on when this resource group is acquired or taken over. Choose the volume groups from the list or enter desired volume groups names in this field.

  If you previously requested that HACMP collect information about the appropriate volume groups, then pressing F4 will either give you a list of all existing volume groups on a local node, or all shared volume groups in the resource group and the volume groups that are currently available for import onto the resource group nodes.

  Specify the shared volume groups in this field if you want to leave the field Filesystems (default is All) blank, and to mount all file systems in the resource group. If you specify more than one volume group in this field, then all file systems in all specified volume groups will be mounted; you cannot choose to mount all file systems in one volume group and not to mount them in another.

  If you have previously selected individual file systems in the Filesystems (default is All) field, the appropriate volume groups are already known to the HACMP for AIX software.

  If you are using raw logical volumes in non-concurrent mode, you only need to specify the volume group in which the raw logical volume resides to include the raw logical volumes in the resource group.

- Concurrent Volume Groups

  Identify the shared volume groups that can be accessed simultaneously by multiple nodes. Choose the volume groups from the list or enter desired volume groups names in this field.

  If you previously requested that HACMP collect information about the appropriate volume groups, then pressing F4 will either give you a list of all existing volume groups on a local node, or all existing concurrent capable volume groups that are currently available in the resource group, and concurrent capable volume groups available to be imported onto the nodes in the resource group.

- Raw Disk PVIDs

  Press F4 for a listing of the PVIDs and associated hdisk device names. If you have previously entered values in the Filesystems or Volume groups fields, the appropriate disks are already known to the HACMP for AIX software.

  If you are using an application that directly accesses raw disks, list the raw disks here.

- AIX Connections Services

  Press F4 to choose from a list of all realm/service pairs that are common to all nodes in the resource group. You can also type in realm/service pairs. Use % as a divider between service name and service type; do not use a colon.

  Note: You cannot configure both AIX Connections and AIX Fast Connect in the same resource group.

- AIX Fast Connect Resources

  Press F4 to choose from a list of Fast Connect resources common to all nodes in the resource group. If you configure Fast Connect file shares, make sure you have defined their file systems in the resource group in the Filesystems field.

  Note that you cannot configure both AIX Connections and AIX Fast Connect in the same resource group. See “General considerations for configuring resources” on page 132 for further notes on this.

- Tape resources

  Enter the tape resources that you want started on the resource group. Press F4 to choose from a list of resources previously defined in the Define Tape Resources screen.

- Application servers

  Indicate the application servers to include in the resource group. Press F4 to see a list of application servers. See the section Configuring Application Servers for information on defining application servers.

- Highly Available Communications Links

  Indicate the communications links to include in the resource group. Press F4 to see a list of communications links.

- Miscellaneous Data

  A string you want to place into the topology, along with the resource group information. It is accessible by the scripts, for example, Database1.

- Automatically Import Volume Groups

  This field specifies whether HACMP should automatically import those volume groups that are defined in the Volume Groups or Concurrent Volume Groups fields. By default, Automatically Import Volume Groups flag is set to `false`.

  If Automatically Import Volume Groups is set to `false`, then selected volume groups will not be imported automatically. In this case, when you add volume groups to the resource group, make sure that the selected volume groups have already been imported to each of the nodes using the **`importvg`** command or C-SPOC.

  If Automatically Import Volume Groups is set to `true`, then when you press Enter, HACMP determines whether the volume group that you entered or selected in the Volume Groups or Concurrent Volume Groups fields needs to be imported to any of the nodes in the resource group, and automatically imports it, if needed.

- Inactive Takeover Activated

  Set this variable to control the initial acquisition of a resource group by a node when the node/resource relationship is cascading. This variable does not apply to rotating or concurrent resource groups.

  If Inactive Takeover is `true`, then the first node in the resource group to join the cluster acquires the resource group, regardless of the node's designated priority. If Inactive Takeover is `false`, the first node to join the cluster acquires only those resource groups for which it has been designated the highest priority node. The default is `false`.

- Cascading without Fallback Enabled

  Set this variable to determine the fallback behavior of a cascading resource group. When the CWOF variable is set to `false`, a cascading resource group will fallback as a node of higher priority joins or reintegrates into the cluster. When CWOF is `true`, a cascading resource group will not fallback as a node of higher priority joins or reintegrates into the cluster. It migrates from its owner node only if the owner node fails. It will not fallback to the owner node when it reintegrates into the cluster. The default for CWOF is `false`.

- 9333 Disk Fencing Activated

  By default, 9333 disk fencing is disabled in a concurrent access environment. To enable 9333 disk fencing, set the field to `true`. Once set, the values in a fence register can typically only be changed by power-cycling the 9333 unit. The fence register is immune to all other "reset" conditions.

Certain occurrences (for example, powering the disk up or down or killing the Cluster Manager) could leave the 9333 disks fenced out from a node (becoming) responsible for managing them. Therefore, the HACMP for AIX software provides a command to clear fence register contents in the same way that power-cycling the disks would. If a node needs access to a disk that is fenced out, you can clear the fence registers for that disk to allow the node access to disk resources. Use the command provided on the Cluster Recovery Aids SMIT screen to do this in extraordinary circumstances only.

- SSA Disk Fencing Activated

  By default, SSA disk fencing is disabled in a concurrent access environment. SSA disk fencing is only available for concurrent access configurations. To enable SSA disk fencing, set this field to `true`. Once set, the values in a fence register cannot be changed by power-cycling the SSA unit. Use the Cluster Recovery Aids SMIT screen to clear the fence registers.

- Filesystems Mounted Before IP Configured

  This field specifies whether, on fallover, HACMP takes over volume groups and mounts file systems before or after taking over the failed node's IP address or addresses.

  The default is `false`, meaning the IP address is taken over first. Similarly, upon reintegration of a node, the IP address is acquired before the file systems.

  Set this field to `true` if the resource group contains file systems to export. This is so that the file systems will be available once NFS requests are received on the service IP address.

After entering field values, you need to synchronize cluster resources. To synchronize the cluster definition, go to the Cluster Resources SMIT screen and select the Synchronize Cluster.

If the Cluster Manager is running on the local node, synchronizing cluster resources triggers a dynamic reconfiguration event (see Section 8.5.3, “DARE resource migration utility” on page 259).

Configuring run-time parameters

There are two types of run-time parameters that can be chosen for a node. One of them is the Debug Level, which can be switched from `high` (default) to `low`**,** meaning all cluster manager actions are logged, or only errors are logged, respectively. The other is the Host uses NIS or Name Server, if the cluster uses Network Information Services (NIS) or name serving, set this field to `true`. The HACMP for AIX software disables these services before entering reconfiguration, and enables them after completing reconfiguration. The default is `false`. Both of these parameters can be changed while the cluster is running.

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 do not 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.

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 for AIX 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. See the Example 4-2.

Example 4-2 Clverify command menu

```
-------------------------------------------------------
 To get help on a specific option, type: help <option>
 To return to previous menu, type: back
 To quit the program, type: quit
```

```
-----------------------------------------------------
Valid Options are:
software
cluster

clverify>
```

Software checking is reduced to lpp checking, which is basically checking whether HACMP-specific modifications to AIX files are correct. For checking the 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 thing as **smit clverify,** that is, verifying that the clusters topology as well as the resource configurations are in sync on the cluster nodes.

If you have configured Kerberos on your system, the **clverify** utility also verifies that:

- All IP labels listed in the configuration have the appropriate service principals in the .klogin file on each node in the cluster.
- All nodes have the proper service principals.
- Kerberos is installed on all nodes in the cluster.
- All nodes have the same security mode setting.

For more information about **clverify** command, please refer to *HACMP for AIX 4.4.1: Installation Guide*, SC23-4278.

For more information about cluster testing, please refer to the Chapter 6, “Cluster testing” on page 161.

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 (or in directory where HACMP log files are redirected) 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 reacquire 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 log file, as well as the cluster's status.

Further and more intensive debugging issues are covered in Chapter 7, "Cluster troubleshooting" on page 193.

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 did not bring up any anomalies, 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 use the cluster snapshot files as a cluster configuration and definition backup.

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 concurrently running different versions of HACMP for AIX.

What information is saved in a cluster snapshot

The primary information saved in a cluster snapshot is the data stored in the HACMP for AIX ODM classes. This is the information used to recreate the cluster configuration when a cluster snapshot is applied. The cluster snapshot utility saves the following HACMP for AIX ODM classes in the cluster snapshot:

HACMPcluster	Cluster configuration information, including cluster ID number, cluster name, names of participating nodes, and information about their IDs.
HACMPnode	Node information, including name and ID number.
HACMPnetwork	Network information, including the name, attribute, and cluster ID.
HACMPnim	Network interface information, including name, description, and path name of the module.
HACMPadapter	Adapter information, including the type, IP label, and function.

HACMPgroup	Resource group information, including name, type, and participating nodes.
HACMPresource	Resource group information, including the values of the Inactive takeover attribute and the disk fencing attribute.
HACMPserver	Information about application servers.
HACMPevent	Event information, including the name, description, and names of pre- and post-processing scripts.
HACMPcommand	Data needed by certain HACMP for AIX commands.
HACMPfence	Settings of the IBM 9333 fence registers. This data, which is device dependent, is saved but not restored. The objects in this class are regenerated as part of applying a snapshot.
HACMPdaemons	HACMP for AIX daemon startup and stop parameters, which may or may not be node specific. This node-specific information is preserved during restoration; however, the object class itself (including the object data) is the same on all nodes. This class also contains the Cluster Lock Manager Resource Allocation parameters.
HACMPsp2	Data HACMP for AIX requires to support the IBM Scalable POWERparallel (SP) system. This information is device specific and is not used during a snapshot restoration.
HACMPcustom	Custom verification, custom event, and custom snapshot method information: method name, full path name to the method, method description, and type definition.

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.

Format of a cluster snapshot

The cluster snapshot utility stores the data it saves in two separate files:

ODM Data File	This file contains all the data stored in the HACMP for AIX ODM object classes for the cluster. This file is given a user-defined base name with the .odm file extension. Because the ODM information must be largely the same

on every cluster node, the cluster snapshot saves the values from only one node.

Cluster State Information File This file contains the output from standard AIX and HACMP for AIX system management commands. This file is given the same user-defined base name with the .info file extension. Output from any custom snapshot method is appended to the .info file.

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.

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 the cluster snapshot can be found in the *HACMP for AIX, Version 4.4.1: Administration Guide*, SC23-4279, as well as in the *HACMP for AIX, Version 4.4.1 Troubleshooting Guide*, SC23-4280.

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 Section 5.1.1, “Predefined cluster events” on page 146). 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:

Node events

The following sections describe the sequence of node_up and node_down events.

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. This script calls the node_up_remote include release_takeover_addr, stop_server, release_vg_fs, and cl_deactivate_nfs events.
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	This script calls the stop_server, release_takeover_addr, release_vg_fs, and release_service_addr 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	This script calls the acquire_takeover_addr, get_disk_vg_fs, and acquire_service_addr 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.

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 because 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, because appropriate actions depend on the local network configuration.

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.

interface_maintenance_mode This event script is called when a maintenance mode is started on an interface.

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.

Application monitor events

server_down This event script is called when an application that is monitored fails. It runs the notify script for that monitor, if one is defined. (HACMP/ES)

server_down_complete This event script is called when an application that is being monitored fails. (HACMP/ES)

server_restart This event script is called when one of the application monitoring needs to restart an application server. (HACMP/ES)

server_restart_complete This event script is called when an application server has been successfully restarted. (HACMP/ES)

Other events

site_down This event script is called when the last node in a site goes down. The script checks to see whether the site name is the local node or not, then calls sub-event scripts accordingly: site_down_local and site_down_remote. (HACMP/ES)

site_down_complete This event script is called after the site_down script successfully completes. The script checks the site name, then calls one of the two sub-event scripts appropriately: site_down_local_complete, and site_down_remote_complete. (HACMP/ES)

site_up This script is called when the first node in a site joins the cluster. The script checks the ID of the node, then calls one of the two sub-event script accordingly: site_up_local and site_up_remote. (HACMP/ES)

site_up_complete This script is called when the site_up script successfully completes. (HACMP/ES)

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 pre-processing, for example, you may want to send a message to specific users informing them to stand by while a certain event occurs. For post-processing, you may want to disable login for a specific group of users if a particular network fails.

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 described in Section 5.1.2, “Pre- and post-event processing” on page 151 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. Because the event script did not 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 the chapters “Administrative Facilities” in *HACMP for AIX, Version 4.4.1: Concepts and Facilities,* SC23-4276, and “Monitoring an HACMP Cluster” in the *HACMP for AIX, Version 4.4.1: 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 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. Now you can chose between predefined values:

- slow
- normal
- fast

or you can customize this value. To see, how to tune this parameters, refer to Section 7.3.4, “Changing the Failure Detection Rate” on page 202.

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.

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, you can normally 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, and 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 4.4.1: Planning Guide*, SC23-4277 for a non-concurrent access configuration.

To check the Major Number of a volume group that already has been created, use the following command:

```
# ls -l /dev/name_of_volume_group
```

In the fifth column, there is a number, which indicates the Major Number (see Example 5-1).

Example 5-1 How to determine the major number

```
# ls -l /dev/test1
crw-rw----   1 root     system    50,  0 Oct 21 21:12 /dev/test1
```

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 the file system, the client application will not recover, because the file system exported by the node will appear to be different from the one exported by the failed node.

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 and directories

The process of NFS-exporting file systems and directories in HACMP for AIX is different from that in AIX.

While in AIX, you list file systems and directories to NFS-export in the /etc/exports file. In HACMP for AIX, you specify file systems and directories to NFS-export by including them in a resource group in the HACMP SMIT NFS Filesystems/Directories to export field.

If you want to specify special options in for NFS-exporting in HACMP, you can create a /usr/sbin/cluster/etc/exports file. This file has the same format as the regular /etc/exports file used in AIX

Note: Use of this alternate exports file is optional. HACMP checks the /usr/sbin/cluster/etc/exports file when NFS-exporting a file system or directory. If there is an entry for the file system or directory in this file, HACMP will use the options listed. If the file system or directory for NFS-export is not listed in the file, or, if the user has not created the /usr/sbin/cluster/etc/exports file, the file system or directory will be NFS-exported with the default option of root access for all cluster nodes.

5.4.3 NFS mounting

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

- To NFS mount, a resource group must be configured with IPAT.
- If you want to use the Reliable NFS Server capability that preserves NFS locks, the IPAT adapter for the resource group must be configured to use Hardware Address Takeover.

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.

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

Creating NFS mount points on clients

A mount point is required in order to mount a file system via NFS. In a cascading resource group, all the nodes in the resource group will NFS mount the file system; thus, you must create a mount point on each node in the resource group. On each of these nodes, create a mount point by executing the following command:

```
# mkdir /mountpoint
```

where mountpoint is the name of the local mountpoint over which the remote file system will be mounted.

Note: A good habit is to *not* create mountpoints in the root (/) directory. Use a subdirectory, such a /nfs/<mountpoint> instead.

Setting up NFS mount point different from local mount point

HACMP handles NFS mounting in cascading resource groups as follows:

The node that currently owns the resource group will mount the file system over the file system's local mount point, and this node will NFS export the file system. All the nodes in the resource group (including the current owner of the group) will NFS mount the file system over a different mount point. Therefore, the owner of the group will have the file system mounted twice - once as a local mount, and once as an NFS mount.

Since IPAT is used in resource groups that have NFS mounted file systems, the nodes will not unmount and remount NFS file systems in the event of a fallover. When the resource group falls over to a new node, the acquiring node will locally mount the file system and NFS-export it. (The NFS mounted file system will be temporarily unavailable to cluster nodes during fallover.) As soon as the new node acquires the IPAT label, access to the NFS file system is restored.

All applications must reference the file system through the NFS mount. If the applications used are dependent upon always referencing the file system by the same mount point name, you can change the mount point for the local file system mount (for example, change it to mount point_local and use the previous local mount point as the new NFS mount point).

In the Change/Show Resources/Attributes for a Resource Group SMIT screen, the Filesystem to NFS Mount field must specify both mount points. Put the NFS mount point, then the local mount point, separating the two with a semicolon, for example “nfspoint;localpoint.” If there are more entries, separate them with a space, for example:

```
nfspoint1;local1 nfspoint2;local2
```

If there are nested mount points, the NFS mount points should be nested in the same manner as the local mount points so that they match up properly. When cross mounting NFS file systems, you must also set the Filesystems mounted before IP configured field of the Resource Group to `true`.

Server-to-server NFS cross mounting

HACMP/ES allows you to configure a cluster so that servers can NFS-mount each other’s file systems. Configuring cascading resource groups allows the cluster manager to decide which node should take over a failed resource, based on priority and node availability. 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.

In the example cluster shown here, you have two resource groups, NodeX_rg and NodeY_rg. These resource groups are defined in SMIT as follows:

Resource Group	NodeX_rg
Participating node names	NodeX NodeY
Filesystems	/xfs (File systems to be locally mounted by node currently owning the resource group)
Filesystems to export	/xfs (File system to NFS-export by node currently owning resource group. File system is subset of file system listed above.)
Filesystems to NFS mount	/mountpointx;/xfs (File systems/directories to be NFS-mounted by all nodes in the resource group. First value is NFS mount point; second value is local mount point)
Resource Group	NodeY_rg
Participating node names	NodeY NodeX

Filesystems	/yfs
Filesystems to export	/yfs
Filesystems to NFS mount	/mountpointy;/yfs

The file system you want the local node (NodeX) in this resource group to locally mount and export is /xfs, on NodeX.You want the remote node (NodeY) in this resource group to NFS-mount /xfs from NodeX. Setting up your cascading resource groups like this ensures the expected default server-to-server NFS behavior described above. On reintegration, /xfs is passed back to NodeX, locally mounted, and exported. NodeY mounts it via NFS again. When the cluster as originally defined is up and running on both nodes, the file systems are mounted, as shown in Figure 5-1.

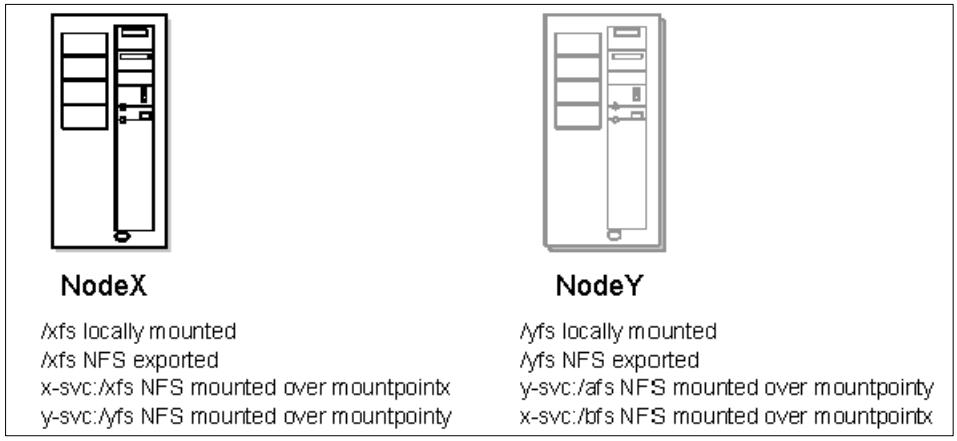


Figure 5-1 NFS crossmount example

When NodeX fails, NodeY uses the cl_nfskill utility to close open files in NodeX:/xfs, unmounts it, mounts it locally, and re-exports it to waiting clients.

After takeover, NodeY has:

- /yfs locally mounted
- /yfs NFS-exported
- /xfs locally mounted
- /xfs NFS-exported
- x-svc:/xfs NFS mounted over /mountpointx
- y-svc:/yfs NFS mounted over /mountpointy

Note: See the man page in /usr/sbin/cluster/events/utils for information about the usage and syntax for the **cl_nfskill** command.

Caveats about node names and NFS

In the configuration described above, the node name is used as the NFS host name 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
- Alias the node name to the service adapter label in the /etc/hosts file.

6

Cluster testing

It is very important to test your HACMP configuration to verify that all efforts have been done to eliminate single point of failure (SPOF) in your HACMP cluster environment. When the HACMP cluster environment that has been set up goes to production all the possible scenarios for failure should have been tested and documented.

Before you test the HACMP cluster configuration, verify that you have made an test scenarios list for all the tests you are about to do for your HACMP configuration. This makes it easier to verify that everything you want to test is covered.

Check the state of the following components:

- Devices
- System parameters
- Processes
- Network adapters
- LVM
- Cluster
- Other items, such as SP Switch, printers, and SNA configuration

6.1 Node verification

Before you simulate errors in your HACMP cluster environment, we suggest that you do some actions to test the state of each node in your cluster. Remember that this is only an example, so you have to customize your own test list depending on your HACMP environment requirement.

6.1.1 Device state

You can use the following ways to verify the state of the device:

- Run **diag -a** in order to clean up the VPD.
- Look in the errorlog for unusual errors by issuing the **errpt | more** or **errpt -a | more** command.
- Check that all devices are in the available state (use **lsdev -C | more**).
- Check that the SCSI addresses of adapters on shared buses are unique (use **lsattr -E -l ascsi0**).
- 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. Example 6-1 shows a serial check.

Example 6-1 Serial line check output

```
#stty < /dev/tty0
speed 9600 baud; -parity hupcl clocal
intr = ^@; quit = ^@; erase = ^@; kill = ^@; eof = ^@; start = ^@
stop = ^@; susp = ^@; dsusp = ^@; reprint = ^@; discrd = ^@; werse = ^@
lnext = ^@
ignbrk -inpck -istrip -ixon -opost
-isig -icanon -echo -echoe -echok
```

- To test a target-mode SSA heartbeat connection between two nodes, invoke one of the nodes and listen to the SSA target mode device by issuing the **cat < /dev/tmssa#.tm** command on one node; the other node will then send the SSA target mode device the **cat /etc/hosts > /dev/tmssa#.im** command. This will show the /etc/hosts file on the listing node. Repeat the test in the other direction. Note that the cluster services must be stopped on both nodes to perform this test.
- If you are using target-mode SCSI heartbeat connection between two nodes the procedure is about the same as for testing target mode SSA. Invoke one of the nodes to listen to the SCSI target-mode device by issuing the **cat < /dev/tmscsi#.tm** (enter twice); the other node sends **cat /etc/hosts > /dev/tmscsi#.im** command to the SCSI target-mode device. This will show the /etc/hosts file on the listing node. Repeat the test in the other direction.

Note that the cluster services must be stopped on both nodes to perform this test.

6.1.2 System parameters

You can use the following ways to verify the 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 (use **lslicense**).
- Check high water mark and other system settings (use **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 accommodate 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

You can use the following ways to verify the state of the process:

- Check the paging space usage by issuing the **lsps -a** command.
- Look for all expected processes with the **ps -ef | more** command.
- Check that the run queue is < 5 and that the CPU usage is at an acceptable level (use **vmstat 2 5**).

For more information about performance tuning, see *AIX 5L Performance Tools Handbook*, SG24-6039.

6.1.4 Network state

You can use the following ways to verify the state of the network:

- Type, for example, **ifconfig lo0**, **ifconfig en0** and **ifconfig en1** to check the network adapter interface 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 the command:

```
# /usr/lpp/ssp/css/ifconfig css0
```

- Use **netstat -i** or **netstat -in** to show the network configuration of the node.
- To check the alternate Ethernet MAC address, issue the **netstat -v ent0 | more** command.
- Look at mbufs sizing relative to requests for memory denied (use **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 (use **ping <ip-address>** or **ping -R <ip-address>**).

  Verify that all boot, service and standby addressers have the same subnetmask and that standby addresses will not be on the same subnet as the boot and service addressers.

  Check that all services and boot addresses on the same subnet can communicate.

  Verify that all standby interfaces on the same subnet can communicate. Test to see if the service and boot interface cannot communicate with the standby interface and vice versa. To test this, do the following steps:

  a. If the HACMP cluster is running, then stop all nodes in the HACMP cluster.
  b. On one node, disconnect the network cables for every standby interface on that node.
  c. From the same node, **ping** all the other nodes boot addressers or, if some nodes do not have a boot address, **ping** their services address. You should be able to make a connection to all those boot and service addresses if they are on the same subnet that your boot or service address is.
  d. Next, **ping** the other nodes standby addresses. You should not be able to communicate to the other nodes standby addresses from your boot or service address. If you are able to communicate with those standby adapters that have the same subnet as this node's standby adapters, then you must check the IP configuration for this cluster.
  e. Start up all the cluster nodes besides the one with disconnected standby adapter cables.
  f. Repeat step 3, but only on those nodes that previously had a boot address.
  g. Reconnect the cables.

h. **ping** all the standby adapters in the cluster, and all those standby adapters that are on the same subnet as your standby adapters should communicate.

i. 9. Repeat this test on every node in the cluster.

- List the ARP table entries with **arp -a**.
- Check the status of the TCP/IP daemons (use **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 (use **more /etc/resolv.conf**).
- Check the status of NIS by typing **ps -ef | grep ypbind** and **lssrc -g yp**.
- If NIS or DNS is used for the name resolution, then verify that the /etc/netsvc.conf file is set so that the /etc/hosts will be used first for name resolution on all the cluster nodes by adding the line `HOSTS=local,bind` into the */*etc/netsvc.conf file. If only the *NSORDER* variable is set, then this will not be permanent after a system reboot.
- If NIS or DNS is used, verify that those nodes have "Host uses NIS or Name Server" set to `true`. To check the settings, type the following command:

  ```
  # /usr/sbin/cluster/utilities/clshowres | more
  ```

  See the output of this command in Example 6-2.

Example 6-2 Fragment clshowres output

```
Node Name                                      austin
Debug Level                                    high
Host uses NIS or Name Server                   false

Node Name                                      boston
Debug Level                                    high
Host uses NIS or Name Server                   false
```

- The command **exportfs** shows non-HACMP controlled NFS exports.
- Run **snmpinfo -m dump -o /usr/sbin/cluster/hacmp.defs address** to show snmp information for Cluster network addresses (including the serial interfaces).

6.1.5 LVM state

You can use the following ways to verify the state of the LVM:

- Ensure that the correct VGs are defined, that quorum and auto-varyon are correctly defined, and that the shared VGs are in the correct state (use **lsvg** and **lsvg -o**).
- Check that there are no stale partitions (use **lsvg -l**).
- Check that all appropriate file systems have been mounted and that none of the rootvg file systems are full (use **df -k**).
- Check that PVIDs have been assigned where necessary and that there are no ghost disks (use **lspv**).
- Verify that all entries in the /etc/filesystems file are correct and that there are no erroneous entries (use **more /etc/filesystems** and **lsfs**).
- If NFS Cross mounting is used, then verify that the Volume Group (VG) number is the same for all nodes in the cluster for a shared VG. To verify what the VG number is for a given VG, type the command **ls -l /dev/vgname**. This has to be done on every node that has the shared VG imported. Example 6-3 shows an output from all VG on one of the nodes in the cluster.

Example 6-3 VG list output

```
#ls -l /dev/*vg*
crw-rw----   1 root     system    10,  0 Sep 25 13:31 /dev/rootvg
crw-r-----   1 root     system    50,  0 Oct 11 11:40 /dev/shared_vg
```

6.1.6 Cluster state

You can use the following ways to verify the state of the cluster:

- 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 log files 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 (use **odmget HACMPcluster**).
- Verify that all the HACMP Configuration is synchronized. To check the state of the HACMP cluster nodes, you have to test the topology and configuration of your HACMP cluster. This is done by using the **clverify** command. Example 6-4 on page 167 shows the cluster topology verification.

Example 6-4 Fragmented output from the cluster topology verification

```
#/usr/sbin/cluster/diag/clverify cluster topology check
Contacting node austin ...
HACMPnode ODM on node austin verified.

Verification to be performed on the following:
        Cluster Topology

Retrieving Cluster Topology...

Verifying Cluster Topology...

Remember to redo automatic error notification if configuration has changed.

Verification has completed normally.
Command completed.
```

Example 6-5 shows the cluster configuration verification done by the **clverify** command.

Example 6-5 Fragmented output from the cluster config verification

```
# /usr/sbin/cluster/diag/clverify cluster config all
Contacting node austin ...
HACMPnode ODM on node austin verified.

Contacting node austin ...
HACMPnetwork ODM on node austin verified.Verification to be performed on the
following:
        Cluster Topology
        Resources

Retrieving Cluster Topology...
Verifying Configured Resources...
Retrieving Resources from Node: austin...
Retrieving Resources from Node: boston...
Performing volume group consistency check.
----------------------------------------
Verifying Resource Group: austin_rg
-----------------------
Verifying Resources on each Node...
Verifying SERVICE_LABEL: serv1
WARNING: Cascading Without Fallback requires the use of remote authorization
via Kerberos or .rhosts files. The absence of remote authorization
will, under certain circumstances result in the failure of a Cascading
Without Fallback resource group.
Verifying Resource Group: boston_rg
ged.
```

```
Verification has completed normally.
Command completed.
```

- To show the clstrmgr version, type the following command:

  ```
  snmpinfo -m dump -o /usr/sbin/cluster/hacmp.defs clstrmgr.
  ```

6.2 Simulate errors

The following section 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 for node boston and a cascading without fallback (CWOF) Resource Group definition for node austin.

The node boston is an failover node for node austin and austin's Resource Group austin_rg.

The node austin is a failover node for boston and boston's Resource Group boston_rg

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`, which is the default, and if the Application Server Scripts include the **`set -x`** command and periodic **`echo`** commands.

6.2.1 Adapter failure

The following sections cover adapter failure.

Ethernet or token ring interface failure

In case of an Ethernet or Token Ring interface failure, perform the following steps:

1. Check that all the nodes in the cluster are up and running.
2. Optional: Prune the error log on austin (use **`errclear 0`**).
3. Monitor the cluster log files on boston.
4. Use the **`ifconfig`** command to shut off the appropriate service interface (but not the Administrative SP Ethernet) on austin (for example, **`ifconfig en0 down`**). This will cause the service IP address to failover to the standby adapter on austin.

5. Verify that the swap adapter has occurred (including MAC Address failover if you configured HWAT) and that HACMP has turned the original service interface back on as the standby interface.
6. Use the **ifconfig** command 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 austin.

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:

1. Check, by way of the verification commands, that all the nodes in the cluster are up and running.
2. Optional: Prune the error log on austin (use **errclear 0**).
3. Monitor the cluster log files on boston.
4. Disconnect the network cable from the appropriate service interface (but not the Administrative SP Ethernet) on austin. This will cause the service IP and MAC addresses to failover to the standby adapter on austin.
5. Verify that the swap adapter has occurred.
6. Reconnect the network cable to the service interface. This will cause the original service interface to become the standby interface.
7. 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 austin.
8. Verify that the swap adapter has occurred.
9. Reconnect the cable to the original standby interface.
10. Verify that the original standby interface is operating with the standby IP address.

Switch adapter failure

This scenario will show how the cascading Resource Group boston_rg on node boston will act in a case of failover to standby node austin and also how the Resource Group boston_rg will fallback to its primary node boston.

Note: Do not disconnect live switch cables to simulate a switch failure!

1. Check, by way of the verification commands, that all the nodes in the cluster are up and running.

2. Assign boston to be Eprimary.
3. Optional: Prune the error log on boston (use **errclear 0**).
4. Monitor the cluster log files on austin.
5. For the status and location of a Resource Group in the cluster, use the **clfindres** command. The output in Example 6-6 shows that Resource Group austin_rg is up on its primary node austin and Resource Group boston_rg is up on its primary node boston.

Example 6-6 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

6. Generate the switch error in the error log that is being monitored by HACMP Error Notification (for configuration, see “Single point of failure hardware component recovery” on page 61), or, if the network_down event has been customized, bring down css0 (use **ifconfig css0 down**) or fence out austin from the Control Workstation (use **Efence austin**).
7. If the first failure simulation method is used, the switch failure will be detected in the error log (**errpt -a | more**) on boston and cause a node failover to austin. 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 austin, then that node will become Eprimary.
8. Verify that failover has occurred (use **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).
9. Use the **clfindres** command to verify that the Resource Group boston_rg now is located on node austin, as in Example 6-7.

Example 6-7 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     austin
```

10. Start HACMP on boston (use **smit clstart**; see Example 6-8 on page 171). austin will release austin's Resource Groups and boston will take them back over, but austin (or a lower alphanumeric node) will remain Eprimary.

Example 6-8 smit clstart output

```
                        Start Cluster Services

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

                                                        [Entry Fields]
* Start now, on system restart or both                  now

  BROADCAST message at startup?                         true
  Startup Cluster Lock Services?                        false
  Startup Cluster Information Daemon?                   false
```

11. Verify that re-integration has occurred (use **`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).
12. Again verify, with the **`clfindres`** command, that the Resource Group boston_rg has fallen back to its primary node boston, as shown in Example 6-9.

Example 6-9 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

Failure of a 7133 adapter

Perform the following steps in the event of a 7133 adapter failure:

1. Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
2. Optional: Prune the error log on austin (use **`errclear 0`**).
3. Monitor cluster log files on boston if HACMP has been customized to monitor 7133 disk failures.
4. Pull all the cable from the SSA adapter.
5. The failure of the 7133 adapter should be detected in the error log (use **`errpt -a | more`**) on austin or should be noted in the appropriate diagnostics tool, and the logical volume copies on the disks in drawer 1 will be marked stale (use **`lsvg -l austin_vg`**).
6. Verify that all sharedvg file systems and paging spaces are accessible (use **`df -k`** and **`lsps -a`**).

7. Re-attach the cables.
8. Verify that all sharedvg file systems and paging spaces are accessible (use **df -k** and **lsps -a**).

6.2.2 Node failure/reintegration

This scenario will show how the CWOF Resource Group austin_rg on node austin will act in a case of failover to standby node boston and also how the Resource Group austin_rg will act different from a cascading Resource Group definition when it is supposed to fallback to its primary node austin, because Resource Group austin_rg has to be manually invoked to its primary node austin.

AIX crash

Perform the following steps in the event of an AIX crash:

1. Check, by way of the verification commands, that all the Nodes in the cluster are up and running.
2. Optional: Prune the error log on austin (use **errclear 0**).
3. Verify the state of the cluster to check that all nodes in the cluster are up by issuing the **/usr/sbin/cluster/clstat** command. Example 6-10 shows that both node austin and boston is up.

Example 6-10 clstat output

```
               clstat - HACMP Cluster Status Monitor
               -------------------------------------------

Cluster: cl_hacmp441    (1)             Fri Oct 12 15:33:21 CDT 2001
               State: UP               Nodes: 2
               SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

4. For the status and location of the Resource Groups in the cluster, use the **clfindres** command. The output in Example 6-11 shows that Resource Group austin_rg is up on its primary node austin and Resource Group boston_vg is located on its primary node boston.

Example 6-11 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

5. If austin is an SMP, you may want to set the fast reboot switch (use **mpcfg -cf 11 1**).
6. Monitor cluster log files on boston.
7. Crash austin by entering **cat /etc/hosts > /dev/kmem.** (The LED on austin will display 888.)
8. The OS failure on austin will cause a node failover to boston.
9. Verify that failover has occurred by monitor /tmp/hacmp.out on node boston and (use **netstat -i** and **ping** for networks, **lsvg -o** and **vi** of a test file for volume groups, and **ps -U <appuid>** for application processes). Example 6-12 shows an fragmented /tmp/hacmp.out output of the event scripts that has been triggered when node austin crashed.

Example 6-12 Fragment output of /tmp/hacmp.out

```
Oct 14 08:04:50 EVENT START: node_down austin
Oct 14 08:04:50 EVENT START: node_down_remote austin
Oct 14 08:04:50 EVENT START: acquire_takeover_addr austin
Oct 14 08:04:53 EVENT COMPLETED: acquire_takeover_addr austin
Oct 14 08:04:53 EVENT START: get_disk_vg_fs
Oct 14 08:04:54 EVENT COMPLETED: get_disk_vg_fs
Oct 14 08:04:54 EVENT COMPLETED: node_down_remote austin
Oct 14 08:04:54 EVENT COMPLETED: node_down austin
Oct 14 08:04:55 EVENT START: node_down_complete austin
Oct 14 08:04:55 EVENT START: node_down_remote_complete austin
Oct 14 08:04:55 EVENT COMPLETED: node_down_remote_complete austin
Oct 14 08:04:55 EVENT COMPLETED: node_down_complete austin
Oct 14 08:04:57 EVENT START: fail_standby boston 192.168.2.20
Oct 14 08:04:57 EVENT COMPLETED: fail_standby boston 192.168.2.20
```

10. The state of cluster can be found by using the **/usr/sbin/cluster/clstat** command. Here in Example 6-13 on page 174 shows that node austin is down.

Example 6-13 clstat output

```
                clstat - HACMP Cluster Status Monitor
               ---------------------------------------

Cluster: cl_hacmp441    (1)             Fri Oct 12 15:23:42 CDT 2001
                State: UP               Nodes: 2
                SubState: STABLE
        Node: austin            State: DOWN
           Interface: boot1 (0)                 Address: 192.168.1.11
                                                State:   DOWN
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

11. Use the **clfindres** command to verify that the Resource Group austin_rg has now been located on node boston (see the output in Example 6-14).

Example 6-14 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     boston
boston_rg   cascading      UP     boston
```

12. Power cycle austin. If HACMP is not configured to start from /etc/inittab, (on restart) start HACMP on austin (use **smit clstart**). austin will not take back its cascading without fallback (CWOF) Resource Groups.

13. Check with the **/usr/sbin/cluster/clstat** command that both nodes austin and bostin are up. See the output in Example 6-15 on page 175 for more details.

Example 6-15 clstat output

```
               clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Fri Oct 12 15:33:21 CDT 2001
               State: UP               Nodes: 2
               SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

14. After HACMP has been started successfully on node austin, we invoke the **clfindres** command again to verify that austin_rg still is located on node boston. See the output in Example 6-16 for more details.

Example 6-16 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     boston
boston_rg   cascading      UP     boston
```

15. To bring back the Resource Group austin_rg to its primary node austin, we have to migrate the Resource Group back to its primary node austin. A Resource Group can be migrated to another node that is part of the Resource Group if that is required. To migrate Resource Group austin_rg back to its primary node austin, use the **/usr/sbin/cluster/utilities/cldare** command or SMIT HACMP menu fast path **smit cl_resgrp_start.select** (as in Example 6-17 on page 176) and then select austin_rg.

Example 6-17 smit cl_resgrp_start.select output

```
+--------------------------------------------------------------------------+
|                        Select a Resource Group                           |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   austin_rg                                                              |
|   boston_rg                                                              |
|                                                                          |
|    F1=Help                F2=Refresh             F3=Cancel               |
|    F8=Image               F10=Exit               Enter=Do                |
|    /=Find                 n=Find Next                                    |
+--------------------------------------------------------------------------+
```

The following output shows the options for bringing a resouce group online. The option Perform Cluster Verification First? can be set to `Yes` or `No`. `No` is the default (see Example 6-18). Option `No` can be used when there has not been any changes to the cluster configuration since the last synchronization.

Example 6-18 smit Bring a Resource Group Online output

```
                        Bring a Resource Group Online

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

                                                      [Entry Fields]
  Resource Group to Bring Online                     austin_rg
  Emulate or Actual?                                 Actual
  Perform Cluster Verification First?                Yes
  Ignore Cluster Verification Errors?                No
```

Example 6-19 shows that the migration of Resource Group austin_rg has completed successfully.

Example 6-19 Fragment output from Resource Group austin_rg migration

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

[TOP]
Executing cldare command: cldare -M austin_rg:default

Performing final check of resource group locations:
```

```
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
-------------------------------------------------------

Requested migrations succeeded.
```

We now invoke the **clfindres** command again to verify that Resource Group austin_rg is located on node austin (see the output from Example 6-20).

Example 6-20 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

16. Verify that re-integration has occurred (use **netstat -i** and **ping** for networks, **lsvg -o** and **vi** of a test file for volume groups, and **ps -U <appuid>** for application processes).

CPU failure

Perform the following steps in the event of CPU failure on node boston. Node bostons Resource Group boston_rg will failover to node austin and automatically fallback to node boston after that node boston has been sucessfully started with HACMP. This failover scenario is the same for Resource Group boston_rg, as it was in the failover scenario in “Switch adapter failure” on page 169.

1. Check, by way of the verification commands, that all the Nodes in the cluster are up and running by issuing the **/usr/sbin/cluster/clstat** command.
2. Optional: Prune the error log on austin (use **errclear 0**).
3. If austin is an SMP, you may want to set the fast reboot switch (use **mpcfg -cf 11 1**).
4. Monitor cluster log files on austin.
5. Power off boston. This will cause a node failover to austin.
6. Verify that failover has occurred (use **netstat -i** and **ping** for networks, **lsvg -o** and **vi** of a test file for volume groups, and **ps -U <appuid>** for application processes).

7. Use the commands **/usr/sbin/cluster/clstat** and **/usr/sbin/cluster/utilities/clfindres** to verify that node boston is down and the Resource Group boston_rg is located on node austin.
8. Power cycle on node austin. If HACMP is not configured to start from /etc/inittab (on restart), start HACMP on austin (use **smit clstart**). austin will take back its cascading Resource Groups.
9. Verify that re-integration has occurred (use **netstat -i** and **ping** for networks, **lsvg -o** and **vi** of a test file for volume groups, and **ps -U <appuid>** for application processes).
10. Again verify with the commands **/usr/sbin/cluster/clstat** and **/usr/sbin/cluster/utilities/clfindres** that node boston now is up and the Resource Group boston_rg also has fallback to its primary node boston.

TCP/IP subsystem failure

Perform the following steps to simulate a TCP/IP failure:

1. Check, by way of the verification commands, that all the Nodes in the cluster are up and running with the **/usr/sbin/cluster/clstat** command.
2. Optional: Prune the error log on austin (use **errclear 0**).
3. Monitor the cluster log files on boston.
4. On austin, stop the TCP/IP subsystem (use **sh /etc/tcp.clean**) or crash the subsystem by increasing the size of the sb_max and thewall parameters to large values (use **no -o sb_max=10000**; **no -o thewall=10000**) and ping boston. 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.
5. The TCP/IP subsystem failure on austin will cause a network failure of all the TCP/IP networks on austin. Unless there has been some customizing 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.
6. Example 6-21 shows the /tmp/hacmp.out, you can see that the network_down_event for node austin has occurred.

Example 6-21 Fragmented /tmp/hacmp.out output.

```
Oct 12 16:09:12 EVENT COMPLETED: network_down_complete austin ether_1
```

7. By invoking **/usr/sbin/cluster/clstat**, you can see, in Example 6-22 on page 179, that the network interface is down for node austin, but the non-TCP/IP network is up.

Example 6-22 clstat output

```
                clstat - HACMP Cluster Status Monitor
               ---------------------------------------

Cluster: cl_hacmp441    (1)             Fri Oct 12 16:17:47 CDT 2001
                State: UP               Nodes: 2
                SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   DOWN
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP

        Node: boston            State: UP
           Interface: boot2 (0)                 Address: 192.168.1.21
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

8. On austin, issue the **startsrc -g tcpip** command. 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. Example 6-23 shows that a network_up_event has happen on node boston.

Example 6-23 Fragmented /tmp/hacmp.out output

```
Oct 12 16:34:35 EVENT COMPLETED: network_up_complete austin ether_1
```

6.2.3 Network failure

Perform the following steps to simulate a network failure:

1. Check, by way of the verification commands, that all the Nodes in the cluster are up and running with the **/usr/sbin/cluster/clstat** command.
2. Optional: Prune the error log on boston (use **errclear 0**).
3. Monitor the cluster log files on austin
4. Disconnect the network cable from the appropriate service and all the standby interfaces at the same time (but not the Administrative SP Ethernet) on node boston. This will cause HACMP to detect a network_down event. Example 6-24 on page 180 shows the networ_down_event for node boston.

Example 6-24 Fragment /tmp/hacmp.out output

```
Oct 12 16:09:12 EVENT COMPLETED: network_down_complete boston ether_1
```

5. HACMP triggers events depending on your configuration of the network_down event. By default, no action is triggered by the network_down event.

6.2.4 Disk failure

The following sections deal with issues of disk failure.

Mirrored rootvg disk (hdisk0) failure

Perform the following steps in case of mirrored rootvg disk (hdisk0) failure:

1. Check, by way of the verification commands, that all the Nodes in the cluster are up and running by issuing the **/usr/sbin/cluster/clstat** command.
2. Verify that the bootlist contains hdisk0 and hdisk1, if, for example, hdisk1 is the mirror of hdisk0 (use **bootlist -m normal -o**). See Example 6-25.

Example 6-25 bootlist output

```
# bootlist -m normal -o
hdisk0
hdisk1
```

3. Optional: Prune the error log on austin (use **errclear 0**).
4. Monitor the cluster log files on boston to see if HACMP has been customized to monitor SCSI disk failures.
5. Slide back the cover/casing on austin to get access to hdisk0 (this may require you to turn 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.
6. The failure of hdisk0 should be detected in the error log (use **errpt -a | more**) on austin.
7. Verify that all rootvg file systems and paging spaces are accessible (use **df; lsps -a**).
8. Shutdown (use **smit clstop; shutdown -F**) and power off austin.
9. Turn the key to normal mode, power on austin, and verify that the system boots correctly. Log in and verify that all the rootvg file systems have been mounted (use **df**).
10. Shutdown (use **shutdown -F**) and power off austin.

11. Reconnect hdisk0, close the casing, and turn the key to normal mode.
12. Power on austin, then verify that the rootvg logical volumes are no longer stale (use **lsvg -l rootvg**).

7135 disk failure

Perform the following steps in the event of a 7135 disk failure:

1. Check, by way of the verification commands, that all the nodes in the cluster are up and running.
2. Optional: Prune the error log on austin (use **errclear 0**).
3. Monitor the cluster log files on boston to see if HACMP has been customized to monitor 7135 disk failures.
4. Mark a shared disk failed through smit (use **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.
5. The amber light on the front of the 7135 comes on, and can also be seen in SMIT (use **smit raidiant**; **RAIDiant Disk Array Manager -> List all SCSI RAID Arrays**).
6. Verify that all sharedvg file systems and paging spaces are accessible (use **df** and **lsps -a**).
7. If using RAID5 with Hot Spare, verify that reconstruction has completed to the Hot Spare, then unmark or plug the failed disk back in. If using RAID1, sync the volume group (use **syncvg austin_vg**).
8. If using RAID5 without Hot Spare, mark the failed disk Optimal (use **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**).
9. Verify that the reconstruction has completed (use **smit raidiant**;**RAIDiant Disk Array Manager -> List all SCSI RAID Arrays**).
10. Verify that all sharedvg file systems and paging spaces are accessible (use **df** and **lsps -a**) and that the partitions are not stale (use **lsvg -l sharedvg**). Also verify that the yellow light has turned off on the 7135.

Mirrored 7133 disk failure

Perform the following steps in the event of a mirrored 7133 disk failure:

1. Check, by way of the verification commands, that all the nodes in the cluster are up and running.
2. Optional: Prune the error log on austin (use **errclear 0**).

3. Monitor the cluster log files on boston to see if HACMP has been customized to monitor 7133 disk failures.
4. Since the 7133 disk is hot pluggable, remove a disk from drawer 1 associated with austin's shared volume group.
5. The failure of the 7133 disk will be detected in the error log (use **errpt -a | more**) on austin, and the logical volumes with copies on that disk will be marked stale (use **lsvg -l NodeFvg**).
6. Verify that all austin_vg file systems and paging spaces are accessible (use **df -k** and **lsps -a)**.
7. Plug the failed disk back in, then sync the volume group (use **syncvg austin_vg**).
8. Verify that all austin_vg file systems and paging spaces are accessible (use **df -k** and **lsps -a**) and that the partitions are not stale (use **lsvg -l austin_vg**).

6.2.5 Application failure

By default, HACMP HAS, also known as HACMP Classic, does not recognize application failures. If monitoring of an application in HACMP HAS is needed, all scripts must be customized and configured for any possible failure scenario that could occur for the application. The scripts can be configured in HACMP HAS to be triggered by pre- or post events and if the application logs output to AIX errorlog, then that output can be used to configure Error Notification in HACMP HAS. Remember that all monitoring for the application is no more reliable than those scripts that has been customized for the applications needs.

In HACMP/ES Version 4.4, application monitoring is available to monitor one or more applications, defined through the SMIT menu, and to specify actions that the system should take upon detection of a process death or application failure.

There are two types of monitoring in ES:

- Process application monitor
- Custom application monitor

The process application monitor relies upon the Event Management (EM) infrastructure provided by RS/6000 Cluster Technology (RSCT), while the custom application monitor uses programs or shell script's that is written by users.

Application monitor program overview

The process application monitor is easier to configure because it uses RSCT built-in monitoring capability and there is no need for any customized shell script, besides the Notify Method script that must be written by a user. However, the limits in the process application monitor is that it cannot monitor any shell script; the processes of the program that is monitored by the process application monitor must be executable binaries, because the shell script are abbreviated in the system table rather than being listed by the full shell script name.

Startup process versus custom application monitoring

When starting up the application monitoring, there are some differences for process and custom application monitor that need to be described.

When a resource group is acquired by a node, the cluster manager (clstrmgr) checks to see if an application server has been configured to monitor process or custom application monitoring inside the resource group. If a monitor is configured, then the clstrmgr starts the run_clappmond deamon, which launches the clappmond daemon.

If the application server is configured to monitor process application monitoring, then the clappmond daemon registers a resource variable to the EM to monitor the processes for the program that has been configured to be monitored (see Figure 6-1).

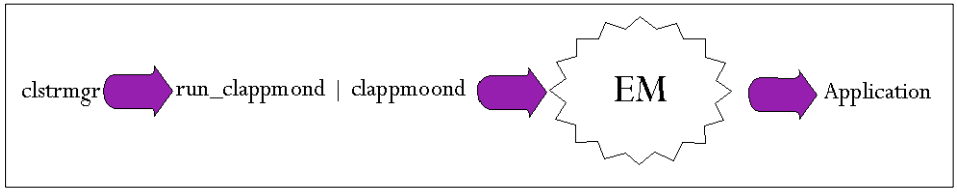


Figure 6-1 Process application monitoring

If custom application monitor is configured to an application server, then the clappmond invokes the user written shell script that has been configured to monitor the given application (see Figure 6-2).

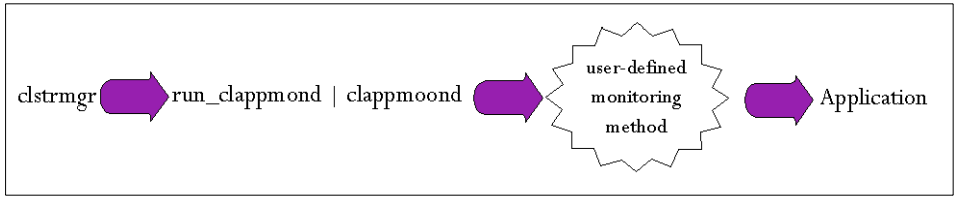


Figure 6-2 Custom application monitoring

There are some considerations to be aware of when planning both process and customer monitoring:

- Application monitoring is only available from HACMP ES Version 4.4.
- Any application to be monitored must be defined inside an application server, and the application server has to be present in a resource group.
- Only one application server per resource group can be monitored. In case you need to monitor multiple applications, one resource group must be configured for each application.
- Any monitored application can only be present in one resource group.
- The monitored application cannot be under the control of the system resource controller (SRC).

6.2.6 Configure the process application monitor parameters

To configure a process application monitor parameters, use the **smit clappserv_to_monitor_by_process.select** fast path (see Example 6-26).

Example 6-26 Add process application monitor output

```
                    Add Process Application Monitor

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

                                                      [Entry Fields]
* Application Server Name                             appsrv
* Processes to Monitor                               []
* Process Owner                                      []
  Instance Count                                     []
* Stabilization Interval                             []
* Restart Count                                      []
  Restart Interval                                   []
* Action on Application Failure                      [notify]
  Notify Method                                      []
  Cleanup Method [/usr/sbin/cluster/events/utils/stop_imaged>
  Restart Method [/usr/sbin/cluster/events/utils/start_image>
```

The fields for the process application monitor are:

Application Name Server

The name for the application server that this process application monitor will be present in.

Process to Monitor
The name of the process; if there is more than one process that will be monitored, use spaces to separate them. To discover the name for the process, use **`ps -el`** rather than **`ps-ef.`**

Process Owner
The name for the user that is the owner for the process.

Instance Count
Specify how many instances of a process are to be monitored; the default is 1. If there is more than one process in the Process to Monitor field, then this value must be 1, because each process is only allowed to be used by one instance.

Stabilization count
Specify the time in seconds that the application needs to start up to be in a stable state. This parameter allows you to delay the monitoring until the application is up and running.

Restart Count
Specify how many times the ES will restart the process on the same cluster node where it failed. If the application has been restarted the number of times that has been specified, then ES executes the Failure Action. The default value is 3 and if = is specified, that means that the ES will not try to restart the application.

Restart Interval
Specify the interval in seconds that the application must remain stable before the Failure Count is reset to 0. If you leave this field empty, ES will assign the value as follows:

```
((Restart Count)*(Stabilization Interval)*1.1)
```

This is the minimum value that you can specify.

Action on Application Failure
If the failed application not have been successfully started within the Restart Count value, this action will be executed by ES. There are two possible action that can be triggered: *notify*, which is the default, or *fallover*.

If notify is selected, then ES run the user customized script that is defined in the Notify Method field.

If fallover is selected, then ES will move the resource group that contain the failed application server to a the next highest priority node that is present in this resource group.

Note: When a fallover action is triggered for a monitored application, then only that resource group that the monitored application's Application Server is present in will fallover to the node with the next highest priority for the resource group; there will not be a node failover of this reason.

Notify Method
Specify the full path for the user customized notify script. This script will be executed each time the monitored application successfully restarts, fails, or moves to a standby node.

Cleanup Method
Specify the full path for the stop script for the monitored application. If the application fails, then this script is invoked every time before the application is restarted to clean up. By default, this is the same as the applications server's stop script that has been defined for the monitored application.

Restart Method
Specify the full path for the monitored application's start script; by default, this is set to the start script that has been defined for the application server for the monitored application. If the Restart Count field is set to 0, then there is no need to specify a start script.

Note: The configuration parameters are saved in a new ODM object class called *HACMPmonitor*. To list information from HACMPmonitor, use the **`odmget HACMPmonitor`** command.

6.2.7 Configure the custom application monitor parameters

The configuration parameters for the custom application monitor is almost the same as for process application monitor and for that reason, we only describe those parameters that differ.To configure the custom application parameter, use **`smit clappserv_to_custom_monitor.select`** (see Example 6-27).

Example 6-27 Custom application monitor output

```
                    Add Custom Application Monitor

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

                                                      [Entry Fields]
* Application Server Name                             appsrv
* Monitor Method                                     []
  Monitor Interval                                   []
```

```
  Hung Monitor Signal                                []
* Stabilization Interval                            []
* Restart Count                                     []
  Restart Interval                                  []
* Action on Application Failure                     [notify]
  Notify Method                                     []
  Cleanup Method [/usr/sbin/cluster/events/utils/stop_imaged>
  Restart Method [/usr/sbin/cluster/events/utils/start_image>
```

The following describes those parameters for custom application monitor that differs from process application monitor:

Monitor Method

Specify the full path for the user customized script that will monitor to see if the defined application is running or not. This script must exit with a return code 0 if the application is up and running, and a exit return code that is not 0 if the application not is running. Arguments cannot be passed to the Monitor Method script on this field.

Monitor interval

Specify, in seconds, the interval that the Monitor Method shell script will be executed. If the execution time for the Monitor Method script is longer than the Monitor Interval, the script is delivered the signal specified in the Hung Monitor Signal state.

Hung Monitor Signal

Specify the number of seconds that must pass before terminating the Monitor Method if it has not been completed within the time that is specified in the Monitor Interval field.

Restart Interval

Specify, in seconds, the time that the application must remain stable before the Restart Count is reset to 0; if this field is empty, then ES will use following formula to calculate a value:

```
((Restart Count)*(Stabilization Interval + Monitor Interval)*1.1)
```

This is the minimum value that can be specified.

- For more information about ES Application Monitoring, see the *HACMP for AIX 4.4.1: Enhanced Scalability & Administration Guide*, SC23-4284

Application failure scenario

In this application failure scenario, we are monitoring the *imserv* application and its process. In case of a failure, application monitoring will restart the imserv application up to two times. The imserv application must be, for this configuration, up and running for at least 44 seconds before it is treated as stable and then it will reset Restart Count to 0. If there is more than two restarts within this time limit, for each server_restart event, application monitoring will not try to restart the imserv application and will instead trigger a server_down event. Each time the imserv application fails, it will also be logged in /tmp/imserv.log by the notify script /usr/local/hacmp/imserv.sh (see Example 6-28).

Example 6-28 imserv.sh output

```
#!/bin/ksh
#
# This shell scipts executes if the application  imserv's process fails.

print "The process imserv died at `/bin/date`" >> /tmp/imserv.log
```

The *appsrv* application server monitors the imserv application and is defined in resource group austin_rg. Example 6-29 shows the configuration for application monitoring of the imserv application.

Example 6-29 Application monitoring output

```
                        Change/Show Process Application Monitor

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

                                       [Entry Fields]
  Application Server Name              appsrv
* Processes to Monitor                 [imserv]
* Process Owner                        [root]
  Instance Count                       [1]
* Stabilization Interval               [20]
* Restart Count                        [2]
  Restart Interval                     [44]
* Action on Application Failure        [notify]
  Notify Method                        [/usr/local/hacmp/imserv.sh]
  Cleanup Method [/usr/sbin/cluster/events/utils/stop_imaged>
  Restart Method [/usr/sbin/cluster/events/utils/start_image>
```

- Verify that all the nodes in the cluster are up and running by issuing the **/usr/sbin/cluster/clstat** command (see the output of the command in Example 6-30 on page 189).

Example 6-30 clstat output

```
               clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Sun Oct 14 19:42:52 CDT 2001
               State: UP                Nodes: 2
               SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                 Address: 192.168.1.10
                                                 State:   UP
           Interface: austin_tty0 (1)            Address: 0.0.0.0
                                                 State:   UP
           Interface: austin_tmssa (2)           Address: 0.0.0.0
                                                 State:   UP

        Node: boston            State: UP
           Interface: boston (0)                 Address: 192.168.1.20
                                                 State:   UP
           Interface: boston_tty0 (1)            Address: 0.0.0.0
                                                 State:   UP
           Interface: boston_tmssa1 (2)          Address: 0.0.0.0
                                                 State:   UP
```

- Verify that the resource group austin_rg is up and running and located on node austin (see Example 6-31).

Example 6-31 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

- Check that the imserv application is up and running on node austin (see Example 6-32).

Example 6-32 ps -ef output

```
# ps -ef | grep imserv
root 31178 24820   0 23:02:43      -  0:00 imserv 192.168.1.10
```

- Terminate the PID for the imserv application's PID by issuing **kill -9 31178**.
- In Example 6-33 on page 190, we can see that the imserv process has been terminated because it is up and running with a new PID; ES will detect that the process for imserv has failed and then restart the application imserv.

Example 6-33 ps -ef output

```
# ps -ef|grep imserv
root 44408 27222   0 23:04:27      -  0:00 imserv 192.168.1.10
```

- Again, we terminate the PID for the application imserv by issuing **kill -9 44408**.
- ES has restarted the application and imserv has a new PID, as can be seen in Example 6-34.

Example 6-34 ps -ef output

```
# ps -ef|grep imserv
root 43842 31908   0 23:04:50      -  0:00 imserv 192.168.1.10
```

- Again, we terminate the imserv application's PID by issuing **kill -9 43842**.
- The output from /tmp/imserv.log (Example 6-35) shows that the application imserv has failed three times.

Example 6-35 /tmp/imserv.log output

```
# pg /tmp/imserv.log
The process imserv died at Sun Oct 14 23:04:26 CDT 2001
The process imserv died at Sun Oct 14 23:04:49 CDT 2001
The process imserv died at Sun Oct 14 23:05:19 CDT 2001
```

- The /tmp/hacmp.out output in Example 6-36 shows that the server_restart has been invoked two times and when the Restart Count was equal to two, then the server_down event was triggered.

Example 6-36 /tmp/hacmp.out output

```
[root /] cat /tmp/hacmp.out |grep EVENT
Oct 14 23:04:26 EVENT START: server_restart austin 21
Oct 14 23:04:26 EVENT COMPLETED: server_restart austin 21
Oct 14 23:04:26 EVENT START: server_restart_complete austin 21
Oct 14 23:04:27 EVENT COMPLETED: server_restart_complete austin 21
Oct 14 23:04:49 EVENT START: server_restart austin 21
Oct 14 23:04:49 EVENT COMPLETED: server_restart austin 21
Oct 14 23:04:50 EVENT START: server_restart_complete austin 21
Oct 14 23:04:50 EVENT COMPLETED: server_restart_complete austin 21
Oct 14 23:05:19 EVENT START: server_down austin 21
Oct 14 23:05:19 EVENT COMPLETED: server_down austin 21
Oct 14 23:05:19 EVENT START: server_down_complete austin 21
Oct 14 23:05:19 EVENT COMPLETED: server_down_complete austin 21
```

If we instead configured the Application Monitoring to trigger a falllover instead of notify as can be seen in Example 6-29 on page 188, then a fallover to node boston would occur. The fallover would only be for the Resource Group austin_rg not a node failover.

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

There are different ways to become aware of the fact that the HACMP cluster is experiencing problems, for example, by monitoring the cluster with HATivoli or by verifying the status of the cluster and the substate, the network state, and the participating nodes in the cluster and their states.

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. Therefore, you must scan the HACMP log files for error, messages, triggered event scripts, and other events that not should not be available in those log files. The output in those log files will indicate if something wrong has happened to the cluster or that the cluster is in a unstable state.

There are various signs that will indicate that an HACMP cluster should be scanned to verify the HACMP cluster state. These signs include:

- System messages on the console
- E-mail to root or to other users that state that a hardware or software error has occurred on a host in the HACMP cluster
- End users complaining about slow or unavailable services

The following section provides an overview of the log files that 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 log files shown in Table 7-1.

Table 7-1 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.
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 file name 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.

Log file name	Description
/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 config_too_long message will continue to be appended into the /tmp/hacmp.out every 30 seconds until action is taken.

The message issued looks like this:

```
config_too_long 360 $event_name $argument
```

Where:

- $event_name is the reconfig event that has failed.
- $argument are the parameters used by the event.

In most cases, this is because an event script has failed and then it often hangs. Use the **ps -ef** command to find the script's PID and then terminate it by issuing **kill -9 PID**.

You can find out more by analyzing the /tmp/hacmp.out file, where a config_too_long message will appear if the time for the script runs longer than the time limit before config_too_long has been set to be triggered by clstrmgr.

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 **/usr/sbin/cluster/utilities/clruncmd** command on the command line, or by using the SMIT Cluster Recovery Aids screen. The **/usr/sbin/cluster/utilities/clruncmd** command signals the Cluster Manager to resume cluster processing.

Note, however, that sometimes scripts simply take too long, so the message showing up is not always an error, but sometimes a warning. If the message is issued, that does not 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.

If it takes longer than six minutes for the script or program to be successfully started, then you should clock the script to determine the actual time the script or program needs to be successfully up and running, then you increase the time to wait before calling the config_too_long by using the following command:

```
chssys -s clstrmgr -a "-u millie_seconds_to_wait"
```

For example:

```
chssys -s clstrmgr -a "-u 6000"
```

This will set the time before calling the config_too_long to 10 minutes instead of default of six minutes.

7.3 Deadman switch

The term *deadman switch* (DMS) describes the AIX kernel extension that causes a system panic and dump under certain cluster conditions if it is not reset.

The reason to use a DMS is to protect the data on the external disks. 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 problems, in particular for the external shared disks.

If this is happening, it is not always obvious why the cluster manager was kept from resetting this timer counter, for example, because some application ran at a higher priority than the clstrmgr process, or the I/O activity was too high because the system ran out of memory and caused a trashing paging activity. 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.

In HACMP Version 4.4, you can change all tuning options from the SMIT HACMP menu “Advanced Performance Tuning Parameters” that is shown in Example 7-1. You can reach the menu with the **smit cm_configure_menu** fast path.

Example 7-1 Cluster Configuration output

```
                        Cluster Configuration

Move cursor to desired item and press Enter.

  Cluster Topology
  Cluster Security
  Cluster Resources
  Cluster Snapshots
  Cluster Verification
  Cluster Custom Modification
  Restore System Default Configuration from Active Configuration
  Advanced Performance Tuning Parameters
```

If you select the Advanced Performance Tuning Parameters it will show all the tuning options that are available from the SMIT HACMP menu, as shown in Example 7-2.

Example 7-2 Advanced Performance Tuning Parameters output

```
                Advanced Performance Tuning Parameters

Move cursor to desired item and press Enter.

  Change/Show I/O pacing
  Change/Show syncd frequency
  Configure Network Modules
```

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 menu or the **chdev** command.To use SMIT, use the **smit chgsys** fast path or the **smit cm_configure_menu** fast path to set high- and low-water marks, as shown in Example 7-3, or type the **chdev -l sys0 -a maxpout=X** command for the high water mark and the **chdev -l sys0 minpout=X** command for the low water mark.

Example 7-3 Change/Show I/O pacing output

```
                        Change/Show I/O pacing

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

                                                      [Entry Fields]
  HIGH water mark for pending write I/Os per file    [0]
  LOW water mark for pending write I/Os per file     [0]
```

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, and if you want to change the recommended values, use the following formula:

```
High water mark = m * 4 +1
Low water mark = n * 4
```

Where *m* and *n* are non-negative integers.

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 5L Performance Tools Handbook*, SC24-6039 for more information on I/O pacing.

Note: If I/O pacing needs to be used in the cluster, it must configured on each node manually, where I/O pacing is needed because this parameter is not stored in the HACMP ODM object classes. It does not matter if I/O pacing is configured through the **smit chgsys** or **smit cm_configure_menu** fast path.

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.

7.3.2 Extending the syncd frequency

The syncd daemon is responsible for flushing all unwritten system buffers to disk. By default, it is started automatically at IPL from /sbin/rc.boot and is invoked by AIX every 60 seconds.

Edit the /sbin/rc.boot file or select Advanced Performance Tuning Parameters through the **smit cm_configure_menu** fast path, as can be seen in Example 7-1 on page 198, to increase the syncd frequency from its default value of 60 seconds to either 30, 20, or 10 seconds (10 seconds is recommended for the most clusters). Increasing the frequency forces more frequent I/O flushes and reduces the likelihood of triggering the deadman switch due to heavy I/O traffic.

To set the new syncd time frequency, see Example 7-4.

Example 7-4 Change/Show syncd frequency output

```
                  Change/Show syncd frequency

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

                                                      [Entry Fields]
  syncd frequency (in seconds)                       [10]
```

Example 7-5, we see that the syncd daemon was stopped and started so the changes are now available.

Example 7-5 syncd output

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Killing the syncd daemon.
```

```
Starting the syncd daemon.
```

Note: If the syncd frequency needs to be increased, this must be done manually on every node in the cluster, because this parameter is not stored in the HACMP ODM object classes. It does not matter if the syncd frequency is configured through the /sbin/rc.boot or the **smit cm_configure_menu** fast path. If the syncd frequency is changed by editing the /sbin/rc.boot, do not forget to run the /sbin/rc.boot to make the new syncd frequency available.

7.3.3 Increase amount of memory for communications subsystem

If you experience network performance problems, setting the **no** option by using **no -o extendednetstats=1** command will give you more information when using the **netstat -am** command. 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 output in Example 7-6 will show that the **no** option thewall does not need to be increased.

Example 7-6 netstat -m output

```
# netstat -m
1092 mbufs in use:
1088 mbuf cluster pages in use
4625 Kbytes allocated to mbufs
0 requests for mbufs denied
0 calls to protocol drain routines
0 sockets not created because sockthresh was reached
```

Note: When the **no** flag extendednetstats is enabled, it will cause performance to degrade on a multiprocessor system and should only be used to aid problem determination. So when there is no need to use the extended information, disable the extendednetstats flag by typing **no -o extendednetstats=0**. If this value was edited in the /etc/rc.net file, it must also be changed, otherwise the value of the exetendednetstats flag will be enabled for that host after an system reboot.

The default value is dependent on your AIX level that you are running and the total amount of physical memory that your host is configured with, so to check the default amount of the real memory on your system, use the **lsattr -El mem0** command (see Example 7-7 on page 202). Also, verify how much the max thewall value can be increased; this is also dependent on the AIX level you are running on your system.

Example 7-7 mem0 output

```
# lsattr -El mem0
size     1024 Total amount of physical memory in Mbytes  False
goodsize 1024 Amount of usable physical memory in Mbytes False
```

To list the current value of the thewall, use the **no -a |grep thewall** command (see Example 7-8 for the output of the command).

Example 7-8 thewall output

```
# no -a |grep thewall
                 thewall = 524208
```

How much the attribute thewall can be increased to use the physical amount of memory is also dependent on the AIX level you are running on your system.

To change this value, use the **no** command; to make the thewall attribute permanent you must 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=624208
```

For more information about tuning the thewall attribute, see the *AIX 5L Performance Tools Handbook*, SG24-6039.

7.3.4 Changing the Failure Detection Rate

In a HACMP cluster, every node is connected to at least one TCP/IP network and should also have one or more Non-TCP/IP network. Every supported network has a corresponding Network Interface Module (NIM), which maintains a connection to the other nodes' NIMs in the cluster.

The technique of keeping track of the status of a cluster by sending and receiving heartbeat messages is the major difference between HACMP HAS and HACMP/ES. HACMP HAS 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 heartbeats

Both the TCP/IP network and non-TCP/IP network is used for sending and receiving keepalive packets, and each NIM, for example, ether, tmssa, or tmscsi, has parameters to tune their rate; this rate is often referred as to the heartbeat rate of the Failure Detection Rate.

There are two parameters in ES 4.4,that determine the heartbeat: frequency and sensitivity.

- Frequency

The time interval between keepalive packets.

- Sensitivity

The number of consecutive keepalive packets that must be missed before the interface is consider to have failed.

The time that is needed by ES Version 4.4 to detect a failure can be calculated by using following formula:

```
Frequency * Sensitivity * 2
```

Tuning the network interface module (NIM)

Example 7-9 shows the **smit cm_config_networks** fast path, which is used to tune a Network Interface Module (NIM).

Example 7-9 smit cm_config_networks output

```
            Move cursor to desired item and press Enter.

  Change a Network Module using Predefined Values
  Change a Network Module using Custom Values
  Show a Network Module
```

When you are about to change the Failure Detection Rate for a NIM, these four values can be selected.

1. Slow
2. Normal
3. Fast
4. Custom

To select slow, normal or fast for tuning a NIM, use the SMIT menu Change a Network Module using Predefined Values, and select the NIM that you want to change. The output from Example 7-10 on page 204 shows all of the supported networks for the heartbeat rate. Here we decide to tune an Ethernet NIM by changing it to slow rate.

Example 7-10 smit Network Module to Change output

```
                              Network Type

 Move cursor to desired item and press Enter.

   IP
   atm
   ether
   fddi
   hps
   token
```

After the ether NIM is selected, you change the Failure Detection Rate to slow, as shown in Example 7-11.

Example 7-11 Change a Network Module ether output

```
             Change a Network Module using Predefined Values

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

                                                        [Entry Fields]
* Network Module Name                                  ether
  Description                                          Ethernet Protocol
  Failure Detection Rate                               Slow

  Note:  Changes made in this panel must be
  propagated to the other nodes by synching
  topology.
```

When the ether NIM has been changed, you will se the new Failure Detection Rate, as shown in Example 7-12.

Example 7-12 smit ether failure rate output

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Adapter failure Detection Time is now 12 * 2 * 2, or 48 seconds
```

Example 7-13 on page 205 shows how to change a tmssa NIM to a slower value by changing the Custom rate from Change a Network Module using Custom Values, as shown in Example 7-9 on page 203.

Example 7-13 smit Change a NIM for tmssa output

```
            Change a Network Module using Custom Values

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

                                                      [Entry Fields]
* Network Module Name                                  tmssa
  New Network Module Name                             []
  Description                                         [TMSSA Serial protocol]
  Parameters                                          []
  Grace Period                                        [60]
  Failure Cycle                                       [8]
  Interval between Heartbeats (seconds)               [3]

  Note: The combination of heartbeat rate
  and failure cycle determines how quickly a failure
   can be
  detected and may be calculated using this formula:
  (heartbeat rate) * (failure cycle) * 2 seconds


  Note:  Changes made in this panel must be
  propagated to the other nodes by synching
  topology.
```

The terminology in Example 7-13 is explained here:

- Grace Period

  The Grace Period is the time limit within which a network fallover must be taken care of.

- Failure Detection Rate

  Select one of the four values: Slow, Normal, Fast or Custom.The failure cycle and heartbeat interval determine how soon a failure can be detected (see Example 7-3 on page 199 for the output).

- Failure Cycle

  The number of missed heartbeat messages that are allowed before an adapter is declared down. If you choose to modify the Custom failure rate, then you can enter a number from 1 to 21474.

- Heartbeat Rate

  The Heartbeat Rate is the interval at which cluster services sends keepalive messages between adapters in the cluster. If you choose to modify the Custom failure rate, then you can enter a number from 1 to 21474.

When the tmssa NIM has been changed, you will see the new Failure Detection Rate, as shown in Example 7-14.

Example 7-14 tmssa custom heartbeat rate output

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Adapter failure Detection Time is now 8 * 3 * 2, or 48 seconds
```

'The default Failure Detection Rate is Normal for each NIM, and does not need to be tuned unless there are problems with the hearbeat messenger default value, such as when your cluster environment starts to experience false fail overs because of loss of heartbeats on a very busy network. When you tune the network interface module, it is recommended that you:

- Change the Failure Detection Rate to a slower value, for example, from normal to slow.
- Set the Failure Detection to Custom and tune the Failure Cycle field by selecting a higher number.
- Adjust the Heartbeat Rate by choosing a higher value.

Note: When any change has been done to the network interface module (NIM) parameters, a synchronization must be done to the cluster. Before you synchronize the cluster topology run **/usr/sbin/cluster/clstat** to verify that the HACMP cluster is down.To synchronize the cluster topology, use the SMIT menu or **/usr/sbin/cluster/diag/clverify cluster topology sync**.

7.4 ES 4.4.0 and later creates entries in AIX error log

A new feature, known as the *topology service*, which creates AIX Error Log entries when certain abnormal conditions occur, has been available through ES 4.4 that relies upon RS/6000 Cluster Technology (RSCT) 1.2. Both PSSP and ES uses RSCT, so some of the error labels are only apply to ES while other error labels only apply to PSSP.

The log files under /var/ha is often very difficult to understand because the information is so detailed in regards to topology services activities, but now when RSCT 1.2 is used to create new entries in the AIX Error Log, it has been much easier to troubleshooting topology services problems. Example 7-15 shows some of the new entries that are created by RSCT for HACMP/ES.

Example 7-15 errpt -t|grep TS_ output

```
# errpt -t|grep TS_
02B7B359 TS_SERVICE_ER       PERM S  Cannot start or refresh. No service entr
0A428C61 TS_NODENUM_ER       PERM S  Cannot start. Cannot get local node numb
0D26EA96 TS_FD_INVAL_ADDR_ST INFO O  Missing or incorrect adapter address
1520F81 TS_CL_PTPOFFSETS_DE UNKN S  Point-to-point network uses too many hea
15EA6E3E TS_CL_NODENUM_ER    PERM S  Cannot refresh/startup: incorrect node n
1BFFE224 TS_LIBERR_EM        PEND S  Topology Services client library encount
1D461OEC TS_SP_DIR_ER        PERM S  Cannot start. Cannot create working dire
20CD2OE5 TS_DEATH_TR         UNKN U  Contact with a neighboring adapter lost
20D516B8 TS_SDR_ER           PERM S  Cannot start. Cannot access or update da
26ADF581 TS_SHMAT_ER         PERM S  Cannot start. Cannot attach to a shared
284EE577 TS_CTIPDUP_ER       PERM S  Cannot start or refresh. IP address dupl
2A188FDE TS_DUPNETNAME_ER    PERM S  Cannot start/refresh: duplicated network
2C5D2DO8 TS_UNS_SIN_TR       UNKN U  Local adapter disabled after unstable si
2F9OO3AC TS_RSOCK_ER         PERM S  Cannot start. Cannot open UDP socket for
32E65B88 TS_MIGRATE_ER       PERM S  Error encountered during migration-refre
EDOC4C29 TS_LATEHB_PE        PERF U  Late in sending heartbeat
EEFO83B9 TS_SYSPAR_ER        PERM S  Cannot start. Cannot get system partitio
F2A1FF2D TS_SPIPDUP_ER       PERM S  Cannot start or refresh. IP address dupl
F39A8E61 TS_REFRESH_ER       PERM U  Error encountered during refresh operati
F49C1DD6 TS_IOCTL_ER         PERM S  Cannot retrieve network interface config
F5618B1A TS_CL_PTPENDPT_DE   UNKN S  Point-to-point network does not have 2 e
F669B4OD TS_ASSERT_EM        PEND S  Topology Services daemon exit abnormally
FCA9F78C TS_LSOCK_ER         PERM S  Cannot start. Cannot open listening sock
```

7.4.1 The topology services subsystem

The topology services daemon is contained in the executable **/usr/sbin/rsct/bin/hatsd** file, and this runs on every node in a HACMP/ES cluster. If HACMP/ES is installed on a SP system, then it will be two daemons running, one for the SP system and one for the HACMP/ES. The name of these two daemons differ and this can be shown by using the **lssrc** command. The *hats* daemon is used in the SP subsystem and the *topsvcs* daemon is used in the HACMP/ES subsystem.

When each daemon starts up, it will first read its configuration file from a file set up by the topsvcs startup script, the same as the subsystem name. This file has all the machines (nodes) listed that are part of the configuration and the IP addresses for each adapter in the configuration for each adapter; therefore, this file is called the *machine list* file. The machine list file is built from the HACMP/ES configuration that is stored in the Global ODM and this is how the topsvcs daemon knows the IP addresses and node number of all potential heartbeats ring members.

The topology services subsystem is directed to form as large a heartbeat ring as possible; to do this, the daemon must alert the other daemons on the other nodes of their presence using a PROCLAIM message.

To verify the operational status of the topology services subsystem, use the **lssrc -ls topsvcs** command on one node in the cluster (see Example 7-16 on page 209). This will only show the status of the networks that are configured on that node the command was executed.

The output of the **lssrc -ls topsvcs** command will show the topsvcs status and two lines for each network for which this node has an adapter, which includes the following:

- The network name.
- The network index.
- The number of defined members/adapters that the configuration reported existing for this network.
- The number of members/adapters that is currently in the heartbeat ring.
- The state of the heartbeat ring, denoted by S, U, or D. The S stands for stable, U stands for unstable, and D stands for disable.
- Adapter ID, which is the address and instance number for the local adapter in this heartbeat ring.
- Group ID, which is the address and instance number of the heartbeat ring. The address of the heartbeat ring is also the address of the group leader.
- HB interval, which is the interval in seconds between heartbeats. This exists both on a per network basis and the default value that can be different. The per network value overrides the default value for that network, if it exists.
- HB sensitivity, which is the number of missed heartbeats that constitute a failed adapter.
- The number of clients connected with a process name and process ID.
- Configuration instance, which is the instance number of the machine list file.

Example 7-16 lssrc -ls topsvcs output

```
# lssrc -ls topsvcs
Subsystem         Group            PID     Status
 topsvcs          topsvcs          21942   active
Network Name   Indx Defd Mbrs St Adapter ID      Group ID
ether1_0       [ 0]    2    2  S 192.168.1.10    192.168.1.20
ether1_0       [ 0] en0          0x43e774b6      0x43e7756b
HB Interval = 1 secs. Sensitivity = 10 missed beats
Packets sent    : UDP 749 ICMP 8 Errors: 0 No mbuf: 0
Packets received: UDP 1030 ICMP 19 Dropped: 0
NIM's PID: 19944
ether1_1       [ 1]    2    2  S 192.168.2.10    192.168.2.20
ether1_1       [ 1] en2          0x43e77460      0x43e77503
HB Interval = 1 secs. Sensitivity = 10 missed beats
Packets sent    : UDP 689 ICMP 0 Errors: 0 No mbuf: 0
Packets received: UDP 998 ICMP 0 Dropped: 0
NIM's PID: 21680
rs232_0        [ 2]    2    2  S 255.255.0.0     255.255.0.1
rs232_0        [ 2] tty0         0x83e7753b      0x83e7750b
HB Interval = 2 secs. Sensitivity = 5 missed beats
tmssa_0        [ 3]    2    2  S 255.255.2.0     255.255.2.1
tmssa_0        [ 3] ssa20        0x83e77547      0x83e77515
HB Interval = 2 secs. Sensitivity = 5 missed beats
  2 locally connected Clients with PIDs:
haemd( 21476) hagsd( 19018)
  Dead Man Switch Enabled:
     reset interval = 1 seconds
     trip  interval = 20 seconds
  Configuration Instance = 5
  Default: HB Interval = 1 secs. Sensitivity = 4 missed beats
  Daemon employs no security
  Segments pinned: Text Data.
  Text segment size: 700 KB. Static data segment size: 317 KB.
  Dynamic data segment size: 3085. Number of outstanding malloc: 229
  User time 0 sec. System time 1 sec.
  Number of page faults: 223. Process swapped out 0 times.
  Number of nodes up: 2. Number of nodes down: 0.
```

We can see the settings in Example 7-16 for the ether, rs232, and a tmssa heartbeat network that has been registered to be monitored by the Topology Services subsystem.

7.4.2 Missing heartbeat creates entries in AIX error log

If you need to change the Failure Detection Rate for a network type, you must calculate the time for the Failure Detection Rate that is appropriate for your cluster environment to discover a node failure. For more information about the Failure Detection Rate, see Section 7.3.4, “Changing the Failure Detection Rate” on page 202.

In this environment, the heartbeat messages must *not* be interrupted on this node for 20 seconds or more, because then the deadman switch (DMS) will be invoked on this node. This time limit for the heartbeat messages can be seen in Example 7-16 on page 209 by looking at the `trip  interval = 20 seconds inetrvall` parameter.

If, for some reason, the topsvcs daemon is not able to receive or send any keepalive packets to its cluster node neighbors for, for example, 17 seconds, the DMS will not be invoked, but that is enough time to log entries in the AIX Error Log file by topology services when this abnormal condition occurs. Example 7-17 shows the entries in the AIX Error Log file.

Example 7-17 errpt output

```
# errpt
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
2OCD2OE5   1106004301 U U topsvcs        Contact with a neighboring adapter lost
2OCD2OE5   1106004301 U U topsvcs        Contact with a neighboring adapter lost
EDOC4C29   1106004301 P U topsvcs        Late in sending heartbeat
EDOC4C29   1106004301 P U topsvcs        Late in sending heartbeat
4DD91286   1106004201 I S hats           DeadMan Switch (DMS) close to trigger
```

Example 7-18 shows a fragmented output of the TS_DEATH_TR entry in the AIX Error Log, which has been entered by the topology services, because the node austin could not receive any keepalive packets from its cluster neighbors due to high CPU activity on the node austin.

Example 7-18 Fragmented TS_DEATH_TR error output

```
LABEL:          TS_DEATH_TR
IDENTIFIER:     2OCD2OE5

Date/Time:       Tue Nov  6 OO:43:03 CST
Sequence Number: 2049
Machine Id:      OOO1615F4COO
Node Id:         austin
Class:           U
Type:            UNKN
Resource Name:   topsvcs
Resource Class:  NONE
Resource Type:   NONE
```

```
Location:        NONE
VPD:

Description
Contact with a neighboring adapter lost

Probable Causes
The neighboring adapter mal-functioned
Networking problem renders neighboring adapter unreachable

Failure Causes
The neighboring adapter mal-functioned
Problem with the network

        Recommended Actions
        Verify status of the faulty adapter
        Verify status of network
```

The TS_LATEHB_PE entry in the AIX Error Log file shows a fragmented output (Example 7-19), that the node austin was late in sending keepalive packets that are equal to or greater than the amount of time needed to consider the network adapter down.

Example 7-19 Fragmented TS_LATEHB_PE output

```
LABEL:          TS_LATEHB_PE
IDENTIFIER:     ED0C4C29

Date/Time:       Tue Nov  6 00:42:47 CST
Sequence Number: 2047
Machine Id:      0001615F4C00
Node Id:         austin
Class:           U
Type:            PERF
Resource Name:   topsvcs
Resource Class:  NONE
Resource Type:   NONE
Location:        NONE
VPD:

Description
Late in sending heartbeat

Probable Causes
Heavy CPU load
Severe physical memory shortage
Heavy I/O activities

Failure Causes
```

```
Daemon can not get required system resource

        Recommended Actions
        Reduce the system load
```

In Example 7-20, we see the TS_DMS_WARNING_ST output in the AIX Error Log; this shows that the system is in a state where it will soon crash, due to the DMS. This entry is just a warning, so the DMS will not be invoked if the topology services gets the CPU's attention before the time limit is exceeded. This shows that the system needs some tuning (see Section 7.3, "Deadman switch" on page 197).

Example 7-20 Fragmented TS_DMS_WARNING_ST output

```
LABEL:          TS_DMS_WARNING_ST
IDENTIFIER:     4DD91286

Date/Time:       Tue Nov  6 00:42:47 CST
Sequence Number: 2045
Machine Id:      0001615F4C00
Node Id:         austin
Class:           S
Type:            INFO
Resource Name:   hats

Description
DeadMan Switch (DMS) close to trigger

Probable Causes
Topology Services daemon cannot get timely access to CPU

User Causes
Excessive I/O load is causing high I/O interrupt traffic
Excessive memory consumption is causing high memory contention

        Recommended Actions
        Reduce application load on the system
        Change (relax) Topology Services tunable parameters
        Call IBM Service if problem persists

Failure Causes
Problem in Operating System prevents processes from running
Excessive I/O interrupt traffic prevents processes from running
Excessive virtual memory activity prevents Topology Services from making
progress

        Recommended Actions
        Examine I/O and memory activity on the system
        Reduce load on the system
```

```
Change (relax) Topology Services tunable parameters
Call IBM Service if problem persists
```

7.5 Node isolation and partitioned clusters

Node isolation occurs when all the 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 us 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 all the nodes in the cluster.

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 diagnostic group shutdown partition (DGSP) message.

7.6 The DGSP message

A Diagnostic Group Shutdown Partition (DGSP) message 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.

7.7 Troubleshooting SSA

In an HACMP cluster environment, there is almost always a twin-tailed disk storage implemented that is shared by two or more nodes in a cluster. SSA disk subsystem is an open standard disk technology that is very often implemented in a HACMP cluster environment.

SSA is based on a loop technology, which offers multiple data path to disk. There are two loops on a SSA adapter, loop A and loop B, and on each of these loops, there can be up to 48 disks, and to each disk a write and read can be done in parallel. Because of the loop technology, it is very easy to maintain this disk

subsystem on the fly because if the loop breaks, for example, when a disk is removed from the loop or one of the cables is disconnected from the loop, there are still connections to all the other devices in the loop if this loop has been properly configured.

To get a high availability in the SSA disk subsystem and remove single points of failure (SPOF), a redundancy of all SSA components must be available on all the HACMP nodes in the cluster. This means that two or more SSA adapters must be configured on each node, and those SSA adapters must be connected to a shared loop that will be configured through two or more 7133 disk cabinets. The SSA loop itself already has redundancy; if this is connected to two or more nodes, even though the loop breaks, all the devices except the failed device will be available. Figure 7-1 shows a two nodes configuration with two SSA adapters in each node, using two loops.

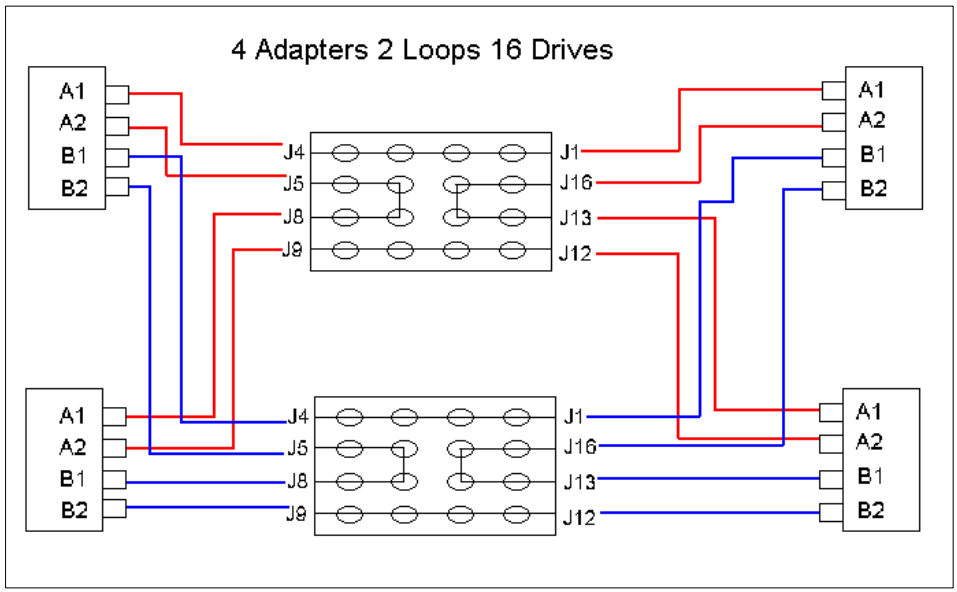


Figure 7-1 7133-D40/T40 loop configuration

For more information about supported SSA devices and configuration planning, refer to the redbook *Understanding SSA Subsystems in Your Environment*, SG24-5750.

7.7.1 SSA pdisk

A SSA physical disk is called a pdisk and a pdisk can be one of this three types:

- System disk

- Array Candidate Disk
- Hot Spare Disk

If a pdisk is configured as a system disk, it will be mapped to an hdisk, which is the logical device of the pdisk, for example, pdisk3 is mapped to hdisk5. The pdisk number and the hdisk number is almost never the same because the internal disks is most often already configured to hdisk0, hdisk1, and so on. The pdisk number starts with pdisk0. To change or show the use of a pdisk, use the **smit ssaraid** fast path and then select Change/Show Use of an SSA Physical Disk, as shown in Example 7-21.

Example 7-21 smit SSA RAID Arrays output

```
                              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
  List Status of Hot Spare Pools
  List Status of Hot Spare Protection for an SSA RAID Array
  List Components in a Hot Spare Pool
  Add a Hot Spare Pool
  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
  Change/Show/Delete a Hot Spare Pool
  Array Copy Services

+--------------------------------------------------------------------------+
|                            SSA RAID Manager                              |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   ssa0 Available 31-08 IBM SSA 160 SerialRAID Adapter (14109100)         |
|   ssa1 Available 2A-08 IBM SSA 160 SerialRAID Adapter (14109100)         |
|                                                                          |
|    F1=Help                   F2=Refresh          F3=Cancel               |
|    F8=Image                  F10=Exit            Enter=Do                |
|    /=Find                    n=Find Next                                 |
+--------------------------------------------------------------------------+
```

After you select one of the SSA adapters, a new menu with the pdisk that is connected to that SSA adapter that was selected will show up. Select the pdisk you want to change or show, as in Example 7-22. If you would like to change multiple pdisks at the same time, select Change Use of Multiple SSA Physical Disks, as in Example 7-21 on page 216.

Example 7-22 Change/Show Use of an SSA Physical Disk output

```
                Change/Show Use of an SSA Physical Disk

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

                                                        [Entry Fields]
  SSA RAID Manager                                     ssa0
  SSA Physical Disk                                    pdisk6
  CONNECTION address                                   0004AC51E2CB00D
  Current Use                                          System Disk
+--------------------------------------------------------------------------+
|                              Current User                                |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   Array Candidate Disk                                                   |
|   Hot Spare Disk                                                         |
|   System Disk                                                            |
|                                                                          |
| F1=Help                F+2=Refresh          F3=Cancel                    |
| F8=Image               F10=Exit             Enter=Do                     |
| /=Find                 n=Find Next                                       |
+--------------------------------------------------------------------------+
```

To find out what relationships there are between all the pdisks and hdisks on the systems, use the **smitty devices** fast path (see output in Example 7-23) and then select SSA Disks, as in Example 7-26 on page 219.

If the pdisk is an Array Candidate Disk, then one or multiple pdisks must be mapped to one hdisk as an SSA RAID. The output from the **ssaxlate -l hdisk9** command would show more than one pdisk if there was an SSA RAID mapped to hdisk9.

A Hot Spare Disk is a pdisk that is used to replace an failed pdisk from a RAID Array.

Example 7-23 smitty devices output

```
                                Devices

Move cursor to desired item and press Enter.
```

```
  Install/Configure Devices Added After IPL
  Printer/Plotter
  TTY
  Asynchronous Adapters
  PTY
  Console
  Fixed Disk
  Disk Array
  CD ROM Drive
  Read/Write Optical Drive
  Diskette Drive
  Tape Drive
  Communication
  Graphic Displays
  Graphic Input Devices
  Low Function Terminal (LFT)
  SCSI Initiator Device
  SCSI Adapter
  Asynchronous I/O
  Multimedia
  List Devices
  Configure/Unconfigure Devices
  Install Additional Device Software
  PCI Hot Plug Manager
  ISA Adapters
  SSA Adapters
  SSA Disks
  SSA RAID Arrays
```

As you can see in Example 7-24, you can select either disk, but we will select SSA Logical Disks.

Example 7-24 smitty menu output

```
                                SSA Disks

Move cursor to desired item and press Enter.

  SSA Logical Disks
  SSA Physical Disks
```

In Example 7-25, you select Show Logical to Physical SSA Disk Relationship. and the output from there will show the relationship between pdisk and hdisk, as can be seen in Example 7-26 on page 219.

Example 7-25 SSA Logical Disks output

```
                            SSA Logical Disks
```

```
Move cursor to desired item and press Enter.

  List All Defined SSA Logical Disks
  List All Supported SSA Logical Disks
  Add an SSA Logical Disk
  Change/Show Characteristics of an SSA Logical Disk
  Remove an SSA Logical Disk
  Configure a Defined SSA Logical Disk
  Generate Error Report
  Trace an SSA Logical Disk
  Show Logical to Physical SSA Disk Relationship
  List Adapters Connected to an SSA Logical Disk
  List SSA Logical Disks Connected to an SSA Adapter
  Identify an SSA Logical Disk
  Cancel all SSA Disk Identifications
  Enable/Disable Fast-Write for Multiple Devices
```

The output in Example 7-26 will show the relationship between pdisk and hdisk.

Example 7-26 lspv output

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

hdisk6: pdisk1
hdisk7: pdisk2
hdisk8: pdisk6
hdisk9: pdisk7
hdisk10: pdisk9
hdisk11: pdisk10
hdisk12: pdisk14
hdisk13: pdisk15
hdisk14: pdisk0
hdisk15: pdisk3
hdisk16: pdisk4
hdisk17: pdisk5
hdisk18: pdisk8
hdisk19: pdisk12
hdisk20: pdisk13
hdisk21: pdisk11
```

If you want to see the relationship between disks, then use the **ssaxlate** command, as shown in Example 7-27 on page 220.

As shown in Example 7-26, you cannot tell if there are SCSI disks or SSA disks, only that they are logical devices, so to separate SCSI disks from SSA disk, do as in Example 7-27.

Example 7-27 hdisk output

```
# ssaxlate -l hdisk8
pdisk6
# ssaxlate -l pdisk5
hdisk17
```

7.7.2 SSA adapters

To list the SSA adapters, use the **lsdev -Cc adapter | grep ssa** command, as shown in Example 7-28.

Example 7-28 lsdev -Cc adapter |grep ssa output

```
# lsdev -Cc adapter |grep ssa
ssa0    Available 31-08    IBM SSA 160 SerialRAID Adapter (14109100)
ssa1    Available 2A-08    IBM SSA 160 SerialRAID Adapter (14109100)
```

For more information about SSA adapters and SSA disk subsystem, see Section 3.3.1, “SSA” on page 82.

7.7.3 SSA problem determination

If you start to experience problems with your SSA devices or want to verify that the SSA disk subsystems is in a stable state on your RS/6000 system, here are some troubleshooting guidelines:

- Run diagnostics on the SSA adapters and follow the Maintenance Analysis Procedures (MAPs) for any Service Request Numbers (SRNs).
- Check the AIX Error Log; if there is a SSA disk subsystem problem, you will find SSA error entries there. In a multi-host configuration, this should be done on every node.
- In a multi-host configuration, run diagnostics on all the hosts. Run **/usr/lpp/diagnostics/bin/run_ssa_healthcheck**; this will perform the SSA health check function, which will attempt to fix any problems on the SSA loop. This function is called every hour by a cron entry.
- To do a diagnostic on any SSA device, use the **diag -ecd <ssa_device>** command. If disks are to be diagnosed, then only pdiskX is a valid ssa_device, not hdiskX.
- Verify that the SSA devices have the same level of microcode and check that the microcode level is supported.

To check the SSA code levels:

- For the adapter microcode level, use **lscfg -vl ssaX**, for example, **lscfg -vl ssa2**
- For the disk drive microcode level, use **lscfg -vl pdiskX**, for example, **lscfg -vl pdisk1**
- For the device driver levels, use **lslpp -L | grep SSA**.

For more information about SSA microcode, see:

`http://techsupport.services.ibm.com/server/nav?fetch=hm`

7.7.4 Replace failure detected SSA disk device

Even though an SSA 7133 hdisk is a hot swappable device, before we replace the failed detected hdisk, we must deconfigure the failed hdisk and configure the new hdisk to the AIX system with AIX commands; in this example, hdisk7 is failed detected from Volume Group (VG) austin_vg. The necessary steps are:

1. Run **reducevg austin_vg hdisk7**.

   This removes the hdiskX from austin_vg.

2. Run **ssaxlate -l hdisk7**.

   This shows the failed detected pdisk2 that is mapped to your hdisk7, as shown in Example 7-26 on page 219.

3. Run **rmdev -l pdisk2 -d**.

   This removes the pdisk2 from the ODM.

4. Run **rmdev -l hdisk7 -d**.

   This removes the hdisk7 from ODM.

5. Replace the failed pdisk2.

6. Run **cfgmgr**.

   This configures the new disk.

7. Run **lspv**.

   This verifies that the new disk is available.

8. Verify that the microcode level for the new disk is at the same level as all the other disks on the system. If it is not, then you must upgrade the disk so that all the disks on the system has the same microcode level.

7.8 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, because 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 situation applies to NFS mounted file systems.

For more information on managing user and group accounts within a cluster, refer to the *HACMP for AIX, Version 4.4.1: Administration Guide*, SC23-4279.

7.9 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.

8

Cluster management and administration

As a system administrator of a highly available environment, you should strive to minimize cluster down time. Whenever it is possible, use the HACMP for AIX dynamic reconfiguration capability to change an active cluster without requiring any downtime. For those changes that require the cluster to be shut down, schedule the down time for off hours when there is minimal activity on the system.When you plan to take the system down, notify end users in advance and inform them when services will be restored.

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 if an subsequent failure occurs before the initial error condition is resolved.

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 Tivoli management interface gives you the capability to monitor a HACMP cluster. Tivoli can monitor the state of your cluster nodes and networks.
- 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
  - 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 4.4.1: 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 and stable. In Example 8-1, we can see that both nodes austin and boston in the HACMP cluster cl_hacmp441 are up and that the cluster is stable. For each node, the utility displays the IP label and address of each network interface attached to the node, and the status of the interface, which can be up or down.

Example 8-1 clstat stable output

```
clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 29 21:43:33 CST 2001
                State: UP               Nodes: 2
                SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

If one or more of the nodes is down but at least one node in the cluster is up, then the cluster would still be stable and the given node or nodes will have their network interface status marked down. In Example 8-2, the output shows that the node austin is down and also it has changed to its boot address and the cl_hacmp441 cluster is stable.

Example 8-2 clstat output

```
clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 29 21:49:19 CST 2001
               State: UP                Nodes: 2
               SubState: STABLE
        Node: austin            State: DOWN
           Interface: boot1 (0)                 Address: 192.168.1.11
                                                State:   DOWN
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN
```

If all of the nodes in the HACMP cluster is stopped, then, as can be seen in Example 8-3, the cl_hacmp441 cluster is down.

Example 8-3 cluster down output

```
clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 29 22:05:05 CST 2001
               State: DOWN              Nodes: 0
               SubState: UNKNOWN
```

clstat will also report whether a node is joining. In Example 8-4 on page 229, we can se that the node austin is joining the cl_hacmp441 cluster and the cl_hacmp441 cluster will remain unstable until the node has been successfully started or if it fails and goes down.

Example 8-4 clstat joining output

```
clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 29 22:13:40 CST 2001
               State: UP               Nodes: 2
               SubState: UNSTABLE
        Node: austin            State: JOINING
           Interface: boot1 (0)                 Address: 192.168.1.11
                                                State:   DOWN
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

It also reports whether a node is leaving, or reconfiguring, and the number of nodes in the cluster. Example 8-5 shows that the node austin is going down.Even though node austin's service network interface has the status up, you can see that the node austin is leaving the cl_hacmp441 cluster because the node state for the node austin is LEAVING. The cl_hacmp441 cluster status will remain unstable until that the node austin has been successfully shutdown.

Example 8-5 clstat leaving output

```
               clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 29 22:30:30 CST 2001
               State: DOWN             Nodes: 2
               SubState: UNSTABLE
        Node: austin            State: LEAVING
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP

        Node: boston            State: UP
           Interface: boston (0)                Address: 192.168.1.20
```

```
                                        State:   UP
Interface: boston_tty0 (1)              Address: 0.0.0.0
                                        State:   UP
Interface: boston_tmssa1 (2)            Address: 0.0.0.0
                                        State:   UP
```

Use the **man clstat** command for additional information about this utility; if this is not installed, it will be available on the HACMP software device.

If you monitor the HACMP cluster with **clstat** on a client machine, then you must edit the /usr/sbin/cluster/etc/clhosts file on that client so it will contain all boot and/or service names/addresses of the HACMP cluster that you want to be accessible through logical connections with this client node.

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 4.4.1: Administration Guide*, SC23-4279.

8.1.3 Monitoring HACMP cluster with Tivoli

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

- Cluster state and substate
- Configured networks and network state
- Participating nodes and node state

The Tivoli Desktop interface is shown in Figure 8-1.

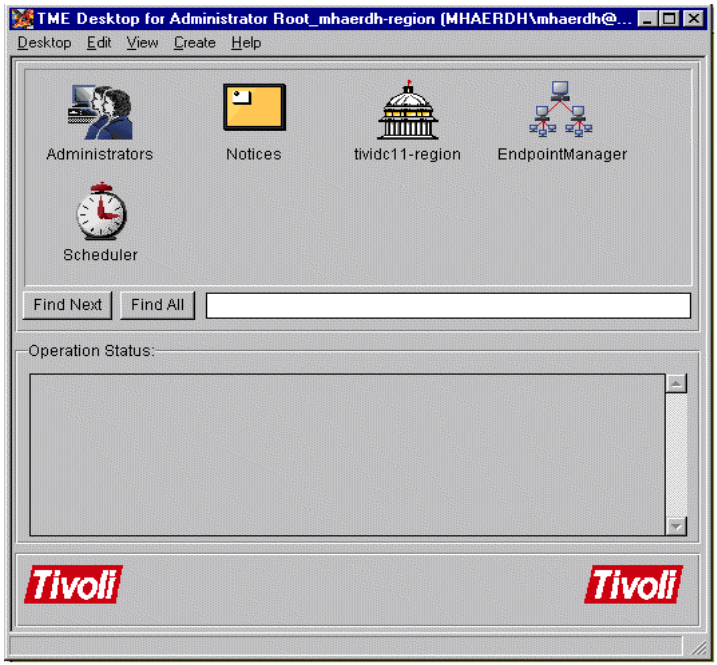


Figure 8-1 Tivoli output

Tivoli's thermometer icons (see Figure 8-1) provide a visual indication of whether components are up, down, in transition, or in an unknown error state. From the window for a selected Policy Region, you can go to a cluster's Indicator Collection window, which display thermometer icons indicating the state of all cluster components.

The cluster status information shown by the thermometer is updated every three minutes or at another interval you specify.

In order to integrate HACMP with Tivoli, you must configure your HACMP cluster nodes as subscriber (client) nodes to the Tivoli server or Tivoli Management Region. Each cluster node can then maintain detailed node information in it's local Tivoli database, which the Tivoli Management Region accesses for updated node information to display.

For full installations details, see Appendix A, "Installing and Configuring Cluster Monitoring with Tivoli", in the *HACMP for AIX 4.4.1:Installations Guide*, SC23-4278.

8.1.4 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 4.4.1: Troubleshooting Guide,* SC23-4280 for more information about these files.

/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 and the date when they occurred are appended to the file.

/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.

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

The /usr/sbin/cluster/history/cluster.mmdd file contains time-stamped, formatted messages generated by HACMP for AIX scripts. The system creates a cluster history file 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.

System error log

The system error log file contains time-stamped, 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.

/tmp/cm.log

Contains time-stamped, formatted messages generated by HACMP for AIX **`clstrmgr`** activity. This file is typically used by IBM Support personnel.

/tmp/cspoc.log

Contains time-stamped, 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.

/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.

Note: If you want the EMUL_OUTPUT variable to be permanent even after a system reboot, set the EMUL_OUTPUT variable in the /etc/environment file.

With HACMP/ES, because of its RSCT technology, there are three more log files you may want to watch. These are as follows:

/var/ha/log/grpsvcs.<filename>

Contains time-stamped 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.

/var/ha/log/topsvcs.<filename>

Contains time-stamped 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.

/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.

8.1.5 Manage the HACMP log files directory

It can sometimes be useful to change the directory for the HACMP log files, for example, by using a separate local or remote file systems for the HACMP log files.

Before you change the HACMP log files directory, there are some considerations to be aware about:

- The disk space needed for most cluster log files are 2 MB; 14 MB is recommended for hacmp.out.
- A local mount point has to be created manually on every node in the cluster before the HACMP log file directory changes. Do not forget to mount the file systems; if the file systems are not available, the log file will be in the directory that you created under / (root).
- If the file system is remote, it should not be on a shared NFS file system, because even though it is desirable in some situations, in rare cases, such actions may cause problems if the file systems need to unmount during a fallover event.
- The HACMP logs cannot be on a file system that is controlled by HACMP; that means it is not possible to set logs on a file system that is part of any volume group that is controlled by HACMP.

8.1.6 Change a HACMP log file directory

In this scenario, we will change the HACMP log file cspoc.log's default directory (/tmp) to a new directory (/local_halogfs). Before you perform the change, you must first verify that the mount point is the same and available on every node in the cluster.

To change a directory for a HACMP log file enter **smitty cl_admin** and select Cluster Log Management, as shown in Example 8-6.

Example 8-6 Cluster Log Management output

```
                        Cluster System Management

Move cursor to desired item and press Enter.

  Cluster Users & Groups
  Cluster Logical Volume Manager
  Cluster Concurrent Logical Volume Manager
  Cluster Physical Volume Manager
  Cluster Resource Group Management
  Cluster Log Management
  HACMP Cluster Services
  Taskguide for Creating a Shared Volume Group
  Cluster Communications Adapter Management
  Suspend/Resume Application Monitoring
```

Select the cspoc.log directory (see Example 8-7 on page 235).

Example 8-7 Cluster Log Management output

```
                        Cluster Log Management

Move cursor to desired item and press Enter.

  Change/Show a Cluster Log Directory

+--------------------------------------------------------------------------+
|                      Select a Cluster Log Directory                      |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   clstrmgr.debug   - Generated by the clstrmgr daemon                    |
|   cluster.log      - Generated by cluster scripts and daemons            |
|   cluster.mmddyyyy - Cluster history files generated daily               |
|   cl_sm.log        - Generated by the cluster Shared Memory library      |
|   cspoc.log        - Generated by CSPOC commands                         |
|   dms_loads.out    - Generated by deadman's switch activity              |
|   emuhacmp.out     - Generated by the event emulator scripts             |
|   hacmp.out        - Generated by event scripts and utilities            |
|   clumt.log        - Generated by Uptime Management Tool                 |
+--------------------------------------------------------------------------+
```

In Example 8-8, we see that the default directory for the cspoc.log files is the /tmp directory, and we will change the directory to /local_halogfs.

Example 8-8 Change/Show a Cluster Log Directory output

```
                          Change/Show a Cluster Log Directory

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

                                      [Entry Fields]
  Cluster Log Name                    cspoc.log
  Cluster Log Description             Generated by CSPOC commands
  Default Log Destination Directory   /tmp
* Log Destination Directory           [/local_halogfs]
  Allow Logs on Remote Filesystems    false
```

When the changes for the cpsoc.log file directory has been done, you must synchronize the cluster resources from the same node where you did the changes for the cspoc.log file (see Example 8-9 on page 236).

Example 8-9 smit output

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

After all log directory changes have been made, please perform
a cluster resource synchronization from this node. This will
ensure that all desired changes have been propagated throughout
the cluster.
```

The new cspoc.log directory changes will take affect after you synchronize cluster resources, or if the cluster is not up, the changes will be effective the next time the cluster is restarted.

To test the new directory /local_halogfs/cspoc.log, we must run a C-SPOC command so that the C-SPOC command will log the output into the file /local_halogfs/cspoc.log. Even though the changes for the C-SPOC log file is synchronized and active, there will not be any cspoc.log files created in the /local_hafs directory until a C-SPOC command has been executed. We will use C-SPOC to create the user austin1 for the HACMP cluster. To do this, enter **smitty cl_admin**, and from there select Cluster Users & Groups. The user austin1 will be connected to Resource Group austin_rg (see Example 8-11 on page 237 and Example 8-10).

Example 8-10 Add a User to the Cluster output

```
                    Add a User to the Cluster

Type or select a value for the entry field.
Press Enter AFTER making all desired changes.

[Entry Fields]
  Select nodes by Resource Group                    []
   *** No selection means all nodes! ***
+--------------------------------------------------------------------------+
|                   Select nodes by Resource Group                         |
|                *** No selection means all nodes! ***                     |
|             Move cursor to desired item and press Enter.                  |
|                                                                          |
|   austin_rg                                                              |
|   boston_rg                                                              |
|                                                                          |
| F1=Help                F2=Refresh             F3=Cancel                  |
| F8=Image               F10=Exit               Enter=Do                   |
| /=Find                 n=Find Next                                       |
```

```
+--------------------------------------------------------------------------+
```

Example 8-11 shows how to add a user.

Example 8-11 Add a User to the Cluster output

```
                      Add a User to the Cluster

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

[TOP]                                                   [Entry Fields]
  Select nodes by Resource Group                       austin_rg
   *** No selection means all nodes! ***

* User NAME                                            [austin1]
  User ID                                             []
  ADMINISTRATIVE USER?                                 false
  Primary GROUP                                       []
  Group SET                                           []
  ADMINISTRATIVE GROUPS                               []
  Another user can SU TO USER?                         true
  SU GROUPS                                           [ALL]
  HOME directory                                      []
```

After the user is created, we look in the /local_halogfs directory and verify that the cspoc.log files has been created there (see Example 8-12).

Example 8-12 ls /local_halogfs/ output

```
# ls /local_halogfs/
cspoc.log   lost+found
```

Note: When C-SPOC is used, it will only log output to its log file (cspoc.log) on the node where the C-SPOC command was executed, so to test the new directory for the cspoc.log file, a C-SPOC command must be executed on each node in the cluster.

If the changes for a HACMP log file is remote, making the changes is the same as for local with a exception for the attribute Allow Logs on Remote Filesystems, which must be set to true (see Example 8-8 on page 235).

8.2 Starting and stopping HACMP on a node or a client

This section 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.

Before the you start or stop nodes in the HACMP cluster, verify on which HACMP node you are inlogged on by type the **`/usr/sbin/cluster/utilities/get_local_nodename`** command.

Note: In ES the directory /usr/es/sbin/cluster and subdirectories have symbolic links to the /usr/bin/cluster directory and subdirectories. However files in these directory are not linked as they were in releases prior to 4.3.1.

8.2.1 HACMP daemons

The following lists the required and optional HACMP for AIX daemons.

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. It also centralizes the storage of and publishes updated information about HACMP-defined resource groups. The Cluster Manager on each node coordinates information that are gathered by

HACMP global ODM, and other Cluster Manager in the cluster to maintain updated information about the content, location, and status of the all HACMP resource groups. This information is updated and synchronized among all the nodes whenever an event occurs that affects resource group configuration, status, and location. All cluster nodes must run the clstrmgr daemon.

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

To see if clstrmgr and clsmuxpd are active, type the **`lssrc -g cluster`** command. Example 8-13 shows that clstrmgr and clsmuxpd are active and also shows their PIDs; all of this was listed because they are present in the cluster group.

Example 8-13 lssrc cluster output

```
# lssrc -g cluster
Subsystem         Group            PID     Status
 clstrmgrES       cluster          36346   active
 clsmuxpdES       cluster          38796   active
 clinfoES         cluster          33820   active
```

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 (that is, the ports are already in use), the cluster manager logs an error message and dies.

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. To see if the clinfo daemon is active, type the **`lssrc -s clinfo`** command, which will list the clinfo

daemon only, or use the **lssrc -g cluster** command (Example 8-13), which will show all of the active daemons that are members of the cluster group. For those HACMP cluster daemons that are not active, those daemons will not be listed unless they are started by the **lssrc** command.

With RSCT (RISC System Cluster Technology) on HACMP/ES, there are several more daemons.

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. The **lssrc -s topsvcs** command will show if it is active.

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.The **lssrc -s emsvcs** command will show if it is active, or, to see if both emsvcsd and emaixos are active, use the **lssrc -g emsvcs** command (see Example 8-14).

Example 8-14 lssrc -g emsvcs output

```
# lssrc -g emsvcs
Subsystem         Group            PID     Status
 emsvcs           emsvcs           21146   active
 emaixos          emsvcs           19686   active
```

Event Management AIX OS Resource Monitor (emaixos)

This daemon acts as a resource monitor for the event management subsystem and provides information about the operating system (OS) characteristics and utilization. The emaixos demon is started automatically by Event Management. The **lssrc -s emsvcs** command will show if it is active (see Example 8-14).

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. The **lssrc -s grpsvcs** command will show if it is active (see the Example 8-15 to list the grpsvcs group).

Example 8-15 lssrc -g grpsvcs output

```
# lssrc -g grpsvcs
Subsystem         Group            PID     Status
 grpsvcs          grpsvcs          32384   active
```

```
grpglsm          grpsvcs          25280   active
```

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.The **`lssrc -s grpglsm`** command will show if it is active, or, to list the grpsvcs group, see Example 8-15.

To list all of the HACMP cluster daemons that are active, use the **`clshowres -a`** command on each node in the cluster. Example 8-18 on page 243 shows you an output of this command.

8.2.2 Starting cluster services on a node

The following section describes how to start the HACMP cluster on a single node. The C-SPOC utility gives you the possibility to start up multiple nodes from one node in the cluster.

Note: If there is a console connected to an tty port on a node with HACMP/ES, that console must be powered on or that node cannot start up HACMP/ES. This can be changed by editing the /usr/es/sbin/cluster/etc/rc.cluster file, finding the scripts that ends with 2>/dev/console, and change this to reflect whatever behavior is desired. For example, you can redirect the output to a another tty. Be aware that this redirects the startup messages on the 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 (see Example 8-16). 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.

Example 8-16 Start Cluster Services output

```
                        Start Cluster Services

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

                                                      [Entry Fields]
* Start now, on system restart or both                now

  BROADCAST message at startup?                       true
```

```
  Startup Cluster Lock Services?                    true
  Startup Cluster Information Daemon?               true
```

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 or by using the C-SPOC services menu SMIT HACMP (see Example 8-17 on page 242). 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.

Example 8-17 Start Cluster Services output

```
                        Start Cluster Services

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

                                                      [Entry Fields]
* Start now, on system restart or both                now
  Start Cluster Services on these nodes               []
  BROADCAST message at startup?                        true
  Startup Cluster Lock Services?                       true
  Startup Cluster Information Daemon?                  true

+--------------------------------------------------------------------------+
|                  Start Cluster Services on these nodes                   |
|                                                                          |
| Move cursor to desired item and press Esc+7.                             |
|     ONE OR MORE items can be selected.                                   |
| Press Enter AFTER making all selections.                                 |
|                                                                          |
|   austin                                                                 |
|   boston                                                                 |
|                                                                          |
| F1=Help                F2=Refresh             F3=Cancel                  |
| F7=Select              F8=Image               F10=Exit                   |
| Enter=Do               /=Find                 n=Find Next                |
+--------------------------------------------------------------------------+
```

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: In ES, if you list the daemons in the AIX System Resource Controller (SRC), you will see ES appended to their names (see Example 8-18).

Example 8-18 clshowres -a output

```
# /usr/es/sbin/cluster/utilities/clshowsrv -a
Subsystem         Group            PID     Status
 clstrmgrES       cluster          43826   active
 clinfoES         cluster          55020   active
 clsmuxpdES       cluster          21094   active
 cllockdES        lock             27376   active
```

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.

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 (see Example 8-19 on page 244). 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**.

Example 8-19 Stop Cluster Services output

```
                        Stop Cluster Services

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

                                                        [Entry Fields]
* Stop now, on system restart or both                  now

  BROADCAST cluster shutdown?                          true
* Shutdown mode                                        graceful
          (graceful or graceful with takeover, forced)

+--------------------------------------------------------------------------+
|                            Shutdown mode                                 |
|                 (graceful or graceful with takeover, forced)             |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   graceful                                                               |
|   takeover                                                               |
|   forced                                                                 |
|                                                                          |
| F1=Help                F2=Refresh             F3=Cancel                  |
| F8=Image               F10=Exit               Enter=Do                   |
| /=Find                 n=Find Next                                       |
+--------------------------------------------------------------------------+
```

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.

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.

Shutdown mode

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 (see Example 8-19 on page 244):

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.

When you use the clstop script you can decide how to stop the cluster with the following options:

-f Forced shutdown

-g Graceful shutdown with no takeover

-gr Graceful shutdown with takeover

For an example, to shutdown HACMP cluster services on a local node with graceful takeover, type the **/usr/sbin/cluster/utilities/clstop -gr** command.

Note: You can now use **clstop** with the -f option on HACMP ES 4.4.0 and ESCRM 4.4.0. This was previously available only with HACMP HAS (High Availability Subsystem) and Concurrent Resource Manager (CRM).

Note: When AIX operating system is shut down with the **shutdown** command, HACMP is stopped gracefully without takeover.

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. Type the **man clexit.rc** command for additional information. If the man pages for clexit.rc are not available, then install the man pages from the HACMP software device.

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
```

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 host names 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 HACMP 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 host names/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 4.4.1: 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 an HACMP node:

- Calculate the amount of RAM that the replacing system will need depending on what type of workload this HACMP node will have, and other workloads that can be taken over from other HACMP nodes in the cluster.
- 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.
- Verify that there are enough slots in the HACMP node for all current an planned adapters for the HACMP node.
- 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 on a HACMP node, you must check if the hardware on that HACMP node supports hot swap adapters, and if this HACMP node is supported, then you also must verify that the adapter that is to be added or replaced also is supported for hot swap adapters on that hardware.

If the HACMP node or adapter do *not* support hot swap adapters, then the HACMP node must be powered off. This means down time for the applications usually running on this HACMP node, at least during the takeover to other HACMP nodes.

Consider the following points when replacing or adding adapters in a HACMP 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, because when a new adapter will be added to a SCSI bus, the default value to that adapter will be SCSI ID 7, which will cause an SCSI ID conflict on that SCSCI bus. Remove the internal terminations on the SCSI card and add a external termination to the adapter if this adapter is not connected to a device in both ends of the Y-cable. For more information about SCSI adapters, see “Supported SCSI adapters” on page 33.

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?

- Is the type of disk you are using hot-swappable?

SSA/SCSI disk replacement (RAID)

RAID arrays are typically designed for concurrent maintenance.You can use the SMIT menu 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.

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**, or **diags**.
2. Remove all LV copies from the failed disk (use **rmlvcopy**).
3. Remove the disk from the VG (use **reducevg**).
4. Logically remove the disk from the system (use **rmdev -l hdiskX -d** or **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 (use **mkdev** or **cfgmgr**) on all nodes.
7. Add the disk to the shared volume group (use **extendvg**).
8. Increase the number of LV copies to span across the new disk (use **mklvcopy**).
9. Synchronize the volume group (use **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 in its normal node and start the application(s).

Disk replacement (non-RAID) with HACMP Version 4.3

With the HACMP Version 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`**, or **`diag`**.
2. Remove all LV copies from the failed disk (use **`smit cl_lvsc`**).
3. Remove the disk from the VG (use **`smit cl_vgsc`**).
4. Logically remove the disk from the system (use **`rmdev -l hdiskX -d`** or **`rmdev -l pdiskY -d`**, if it is an SSA disk).
5. Physically remove the failed disk and replace it with a new disk.
6. Add the new disk to the ODM (use **`mkdev`** or **`cfgmgr)`**.
7. Add the new disk to the sharedvg (use **`smit cl_vgsc`**).
8. Increase the number of LV copies to span across the new disk (use **`smit cl_lvsc`**).
9. Sync the volume group (use **`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 (for example, increasing the size of a file system) are made, the information about the volume group in ODM and in the VGDA on the disks is 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 4.4.1: 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, It is *not* possible to remove a VG on all of the cluster nodes with C-SPOC, TaskGuide, or Lazy Update.

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 Section 8.8, “User management” on page 277).
- Maintain shared Logical Volume Manager (LVM) components.
- Control HACMP services on a cluster-wide basis (see Section 8.2, “Starting and stopping HACMP on a node or a client” on page 238).

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 4.4.1: Administration Guide*, SC23-4279 for detailed information on using C-SPOC SMIT options.

With C-SPOC, you can perform the following tasks:

Note: The C-SPOC utility only operates on both shared and concurrent LVM components that are defined as part of an HACMP/ES resource.

- Shared volume groups
  - Create a shared volume group
  - Import a volume group
  - Extend a volume group
  - Reduce a volume group
  - Mirror a volume group

  - Unmirroring a volume group
  - Synchronize volume group mirrors
  - List all shared volume groups
  - List all active shared volume group
  - Display characteristics of a shared volume group
- Shared logical volumes
  - Create a shared logical volume
  - List all logical volumes by volume group
  - Make a copy of a logical volume
  - Remove a copy of a logical volume
  - Change or view the characteristics of a logical volume (name and size)
  - Remove a logical volume
- Shared file systems (only applicable for non-concurrent VG's)
  - Create a shared file systems
  - 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:

- A Volume Group that has been created with C-SPOC must be defined in a resource group, and cluster resources must be synchronized prior to using C-SPOC to manage it.

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.

TaskGuide requirements

Before you start the TaskGuide, make sure that:

- You have a configured HACMP cluster in place.
- You have the TaskGuide filesets installed.
- You are on a graphics capable terminal
- You have set the display to our machine using your IP address or an alias, for example:

  ```
  # export DISPLAY=<your- IP adress>:0.0
  ```

- You have set the variable TERM. For example, enter:

  ```
  export TERM=xterm
  ```

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 cl_admin**.
2. From the SMIT main menu, choose 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 four types: cascading, cascading without fallback (CWOF), 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 (see Example 8-20), use **smit cm_add_grp**.

Example 8-20 Add a Resource Group output

```
                        Add a Resource Group

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

                                                      [Entry Fields]
* Resource Group Name                                [boston_vg]
* Node Relationship                                   cascading
* Participating Node Names / Default Node Priority   []

+--------------------------------------------------------------------------+
|         Participating Node Names / Default Node Priority                 |
|                                                                          |
| Move cursor to desired item and press Esc+7.                             |
| ONE OR MORE items can be selected.                                       |
| Press Enter AFTER making all selections.                                 |
|                                                                          |
| austin                                                                   |
| boston                                                                   |
+--------------------------------------------------------------------------+
```

- To change or remove a resource group (see Example 8-21), use **smit cm_add_res**.

Example 8-21 Define Resource Groups output

```
                        Define Resource Groups

Move cursor to desired item and press Enter.

  Add a Resource Group
  Change / Show a Resource Group
  Remove a Resource Group
```

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 (see Example 8-22 on page 258).

Note: Before you synchronize the cluster configuration, verify that all nodes are powered on and, if you using HACMP HAS or HACMP/ES with Standard security mode, that the /etc/hosts and /.rhosts include all the nodes IP labels. If you use Enhanced security mode in HACMP/ES, the /.klogin files on all nodes participating in the service have an entry for each service principal configured for Kerberos.

Ignore cluster verification errors

The cluster verification utility will be run before the information is synchronized to all cluster nodes. The verification utility will verify that the cluster topology and/or cluster resources are properly configured. Under certain circumstances, it may be necessary to perform the synchronization even if the verification routines report an error; in these cases, this parameter must be set to `true`, (the default is `no`). Please be advised that the verification should be ignored only under conditions well understood by the cluster administrator (see Example 8-22 on page 258).

Un/configure cluster resources

By default, this is set to `Yes`, and when there has been a change to the cluster resources, the affected resources may be unconfigured and perhaps reconfigured during the period of synchronization via a set of scripts: reconfig_resource_release, reconfig_resource_acquire, and reconfig_resource_complete. However, if the cluster administrator so desires, setting this flag to `No` will cause the affected resources to be removed from the HACMP configuration, but will not cause any scripts to be run, which would configure/unconfigure a resource (see Example 8-22).

Emulate or actual

When you synchronize cluster resources, HACMP determines what configuration changes have taken place and checks for various errors before changing the configuration on any node. If you choose to emulate synchronization, then it does not effect the Cluster Manager, but if it is set to `Actual`, then the new configuration will take effect. The default is `Actual` (see Example 8-22).

Skip cluster verification

If this has been set to `Yes`, the Skip Cluster Verification option allows you to skip Cluster Verification while synchronizing resources or topology. If any of the nodes in the cluster is active, then Cluster Verification will not be skipped. The default is `No` (see Example 8-22).

Example 8-22 Synchronize Cluster Resources output

```
                    Synchronize Cluster Resources

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

                                                      [Entry Fields]
  Ignore Cluster Verification Errors?                 [No]
  Un/Configure Cluster Resources?                     [Yes]
* Emulate or Actual?                                  [Actual]
* Skip Cluster Verification                           [No]

  Note:
  Only the local node's default configuration files
  keep the changes you make for resource DARE
  emulation. Once you run your emulation, to
  restore the original configuration rather than
  running an actual DARE, run the SMIT command,
  "Restore System Default Configuration from Active
  Configuration."
  We recommend that you make a snapshot before
  running an emulation, just in case uncontrolled
  cluster events happen during emulation.
```

Note: In HACMP for AIX, the event customizations 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 Default Configuration Directory (DCD; for more information, see Chapter 3 in the *HACMP for AIX, Version 4.4.1: 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 Dynamic Reconfiguration (DARE) Resource Migration utility is an administration cluster management tool that is used to change status and locations for resource groups without stopping the cluster services.This utility provides improved cluster management by allowing you to do following:

- Bring a resource group up/online (this option also restores a resource group to its default setting by removing any sticky markers for the given resource group. Sticky and non-sticky is discussed later in this chapter).
- Bring a resource group down/offline.
- Move resource group to a new location.
- Perform maintenance on a node without losing access to the node's resources.
- Relocate resource groups for enhanced performance.

The **cldare** 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.

If you using the SMIT interface, you can only change one resource group at a time.

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.

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.

Because 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.

A cascading resource group's fallback behavior is depending on if the Cascading Without Fallback Enabled (CWOF) flag is set to `true` or `false` and the if Inactive Takeover flag is set to `true` or `false`**.** For example, if the Inactive Takeover is set to `false`, the first node with the highest priority will acquire the resource group only if it is a the sticky node.

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 bounce 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.

If a cascading resource group has the CWOF flag set to `false`, it cannot be moved by an non-sticky migration, because it is already located on the node with the highest priority for that resource group.

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.

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 requisition.

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 requisition even after a failover or reintegration.

8.5.4 CWOF versus other resource groups policies

In this section we will describe the difference between cascading without fallback (CWOF) and rotating cascading resource group policies.

CWOF versus rotating resource group

Even though a CWOF resource group has failed over to another node and its primary node has rejoin the cluster, the CWOF resource group will still be located on that node it failed over to until a manual management or a failure of the node occurs. But unlike an rotating group CWOF has a primary node, and when a failover occurs it will require one of the failover nodes standby adapter, unlike a rotating resource group that will require a standby nodes boot adapters. If

several rotating resource groups share a network, only one rotating resource group can be up and running on a given node at any time. This varies, as it can be one CWOF resource group for every service and standby adapter (that has been required of an CWOF resource group) at a given node at any time.

CWOF versus a DARE sticky move

If an failover for a CWOF resource group occurs, it will not change the node priority, even though it must be manually managed to bring back the CWOF resource group to its primary node. A DARE sticky move makes the node, which the resource group moves to, the highest priority node for that resource group until another DARE migration changes this (DARE to another node, DARE to stop or a DARE to default).

DARE migration has been enhanced in a CWOF resource group. A CWOF resource group supports both sticky or non-sticky DARE migrations, unlike a cascading resource group, which only support sticky DARE migrations.

Note: If you are doing maintenance on a CWOF owner's node, it can be helpful to use an non-sticky DARE migration, because if the default node fails, the CWOF resource group will fallover to its owner node, if that is available.

8.5.5 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.

Startup a cascading resource group with a down primary node

In this DARE migration scenario, resource group boston_rg's primary node is boston, which is not started. Resource group boston_rg will be sticky DARE, migrated to be up and running on node austin.

Example 8-23 on page 265 shows that node boston is down and also the resource group boston_rg is not available (see Example 8-24 on page 265).

Example 8-23 clstat output

```
                clstat - HACMP Cluster Status Monitor
              -------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 22 15:52:29 CDT 2001
               State: UP                Nodes: 2
               SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN

        Node: boston            State: DOWN
           Interface: boot2 (0)                 Address: 192.168.1.21
                                                State:   DOWN
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN
```

Example 8-24 shows that resource group boston_rg not is available in the cluster.

Example 8-24 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading     DOWN    N/A
```

We will now startup resource group boston_rg, but not on the primary node boston, which is down; instead, we are going to start up resource group boston_rg on node austin by using *sticky DARE migration* (see Example 8-25).

Example 8-25 Fragmented DARE migration output

```
# cldare -M boston_rg:austin:sticky
Performing preliminary check of migration request...
Migration request passed preliminary check for compatibility with
current cluster configuration and state.
Verification to be performed on the following:
        Cluster Topology
        Resources
cldare: Requesting a refresh of the Cluster Manager...
0513-095 The request for subsystem refresh was completed successfully.
```

```
...completed.
Waiting for migrations to occur........ completed.
Committing location information to ODM on all nodes..... completed.
Performing final check of resource group locations:
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
boston_rg   cascading      UP     austin     austin
-----------------------------------------------------
Requested migrations succeeded.
```

Note: You must start the DARE migration from a node in the cluster that has the Cluster Manager active or the DARE migration will not occur and you get an message output (see Example 8-26).

Example 8-26 cldare output

```
cldare: An active Cluster Manager was detected elsewhere in the
cluster. This command must be run from a node with an active Cluster
Manager process in order for the Dynamic Reconfiguration to proceed.
The new configuration has been propagated to all nodes for your convenience.
```

As we can see in Example 8-27, node boston is still down, but the resource group boston_rg is now up and running on node austin which now has the highest priority for the resource group boston_rg. By using the **clfindres** command, if an resource group has been DARE migrated with sticky, it will show under the sticky label, as shown in Example 8-28 on page 267.

Example 8-27 clstat output

```
                clstat - HACMP Cluster Status Monitor
               -------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 22 16:16:40 CDT 2001
                State: UP               Nodes: 2
                SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   DOWN

        Node: boston            State: DOWN
           Interface: boot2 (0)                 Address: 192.168.1.21
                                                State:   DOWN
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   DOWN
```

```
          Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                               State:   DOWN
```

Example 8-28 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     austin     austin
```

We will now start up node boston. Even though node boston is the owner of the resource group boston_rg, it no longer has the highest priority, because we did a sticky DARE migration of resource group boston_rg to node austin earlier and therefore node austin will not release the resource group boston_rg when node boston rejoins the cluster.

Example 8-29 shows that the node boston has rejoined the cluster. By using the **clfindres** command, we can see that resource group boston_rg is still located on node austin (see Example 8-30 on page 268).

Example 8-29 clstat output

```
               clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 22 16:40:46 CDT 2001
               State: UP                Nodes: 2
               SubState: STABLE
        Node: austin           State: UP
           Interface: austin (0)               Address: 192.168.1.10
                                               State:   UP
           Interface: austin_tty0 (1)          Address: 0.0.0.0
                                               State:   UP
           Interface: austin_tmssa1 (2)        Address: 0.0.0.0
                                               State:   UP

        Node: boston           State: UP
           Interface: boot2 (0)                Address: 192.168.1.21
                                               State:   UP
           Interface: boston_tty0 (1)          Address: 0.0.0.0
                                               State:   UP
           Interface: boston_tmssa1 (2)        Address: 0.0.0.0
                                               State:   UP
```

The resource group boston_rg is still located on node austin, as can be seen in Example 8-30 on page 268.

Example 8-30 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     austin     austin
```

To bring back resource group boston_rg to its primary node boston, we must remove the sticky from the resource group boston_rg to undo node austin to have the highest priority for the resource group boston_rg. All this is done by using the **cldare -M boston_rg:default** command (see Example 8-31).

Example 8-31 Fragmented cldare output

```
# cldare -M boston_rg:default
Performing preliminary check of migration request...
Migration request passed preliminary check for compatibility with
current cluster configuration and state.
Verification to be performed on the following:
        Cluster Topology
        Resources
cldare: Requesting a refresh of the Cluster Manager...
0513-095 The request for subsystem refresh was completed successfully.
...completed.
Waiting for migrations to occur..... completed.
Committing location information to ODM on all nodes..... completed.
Performing final check of resource group locations:
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
boston_rg   cascading      UP     boston
-----------------------------------------------------
Requested migrations succeeded.
```

To verify that the resource group boston_rg is now non-sticky and that it is located on its primary node boston, use the **clfindres** command (see output in Example 8-32).

Example 8-32 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

Stopping resource groups

If the location field of a migration contains the keyword `stop` instead of an actual node name, 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.

To stop a cascading resource group with sticky

The resource group boston_rg is up on its primary node boston, as can be seen in Example 8-32 on page 268 and Example 8-31 on page 268. To stop resource group boston_rg (and we want it to stay down, even if the node boston will be restarted), use the **`cldare -M boston_rg:stop:sticky`** command (see Example 8-33).

Example 8-33 Fragmented cldare output

```
# cldare -M boston_rg:stop:sticky
Performing preliminary check of migration request...
Migration request passed preliminary check for compatibility with
current cluster configuration and state.
Verification to be performed on the following:
        Cluster Topology
        Resources
cldare: Requesting a refresh of the Cluster Manager...
0513-095 The request for subsystem refresh was completed successfully.
...completed.
Waiting for migrations to occur.................. completed.
Committing location information to ODM on all nodes..... completed.
Performing final check of resource group locations:
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
boston_rg   cascading     DOWN    N/A        STOP
-----------------------------------------------------------
Requested migrations succeeded.
```

After we brought down the resource group boston_rg with sticky (see Example 8-34), it is only the resource group boston_rg that has been stopped, not the node boston.

Example 8-34 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading     DOWN    N/A        STOP
```

By issuing the **clstat** command, we can see that the node boston is still up and running, as shown in Example 8-35.

Example 8-35 clstat output

```
                clstat - HACMP Cluster Status Monitor
               ---------------------------------------------

Cluster: cl_hacmp441    (1)             Mon Oct 22 17:41:22 CDT 2001
                State: UP               Nodes: 2
                SubState: STABLE
        Node: austin            State: UP
           Interface: austin (0)                Address: 192.168.1.10
                                                State:   UP
           Interface: austin_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: austin_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP

        Node: boston            State: UP
           Interface: boot2 (0)                 Address: 192.168.1.21
                                                State:   UP
           Interface: boston_tty0 (1)           Address: 0.0.0.0
                                                State:   UP
           Interface: boston_tmssa1 (2)         Address: 0.0.0.0
                                                State:   UP
```

We have restarted node boston and also started up the HACMP on node boston and as we can see in Example 8-36, the resource group boston_rg is still down, and this is because we use a sticky. If we only had stopped the resource group boston_rg with non-sticky DARE migration, then the resource group boston_rg would also been up on node boston.

Example 8-36 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
```

```
austin_rg   cascading      UP     austin
boston_rg   cascading     DOWN    N/A        STOP
```

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 4.4.1: Administration Guide*, SC23-4279, for the syntax and typical output of the **clfindres** command.

Removing sticky markers when the cluster is down

Sticky location markers are stored in the HACMP resource 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.)

Example 8-36 on page 270 shows that the resource group boston_rg is sticky and down, because the primary node boston which owns the resource group boston_rg is up. We must use the default option to bring up resource group boston_rg and remove its sticky (see Example 8-37).

Example 8-37 cldare output

```
# cldare -M boston_rg:default
Performing preliminary check of migration request...

Migration request passed preliminary check for compatibility with
current cluster configuration and state.
Verification to be performed on the following:
        Cluster Topology
```

```
        Resources
clsnapshot: Succeeded creating Cluster Snapshot: active.0.
cldare: Requesting a refresh of the Cluster Manager...
0513-095 The request for subsystem refresh was completed successfully.
...completed.
Waiting for migrations to occur................... completed.
Committing location information to ODM on all nodes..... completed.
Performing final check of resource group locations:
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
boston_rg   cascading      UP     boston
-----------------------------------------------------
Requested migrations succeeded.
```

We also verify, with the **clfindres** command, that the resource group boston_rg is non-sticky and that it is up and running on its primary node boston (see Example 8-38).

Example 8-38 clfindres output

```
# /usr/sbin/cluster/utilities/clfindres
GroupName     Type       State    Location   Sticky Loc
---------   ----------   ------   --------   ----------
austin_rg   cascading      UP     austin
boston_rg   cascading      UP     boston
```

8.6 Software maintenance for 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. Back up the HACMP cluster environment, and verify that the backup has been successfully.
2. Use the **smit clstop** fast path 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 the cluster with takeover so that the takeover node will continue to provide these resources to users.

3. Apply the software maintenance to this node using the procedure described in the documentation distributed with the PTF.
4. 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.
5. 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.
6. Restart the HACMP for AIX software on the node using the **smit clstart** fast path and verify that the node successfully joined the cluster.
7. Repeat Steps 1 through 5 on the remaining cluster nodes.
8. When all the nodes in the HACMP cluster have the new HACMP maintenance level successfully installed, then back up the HACMP cluster environment, and verify that the backup has been successfully.

Figure 8-2 on page 274 below shows the procedure.

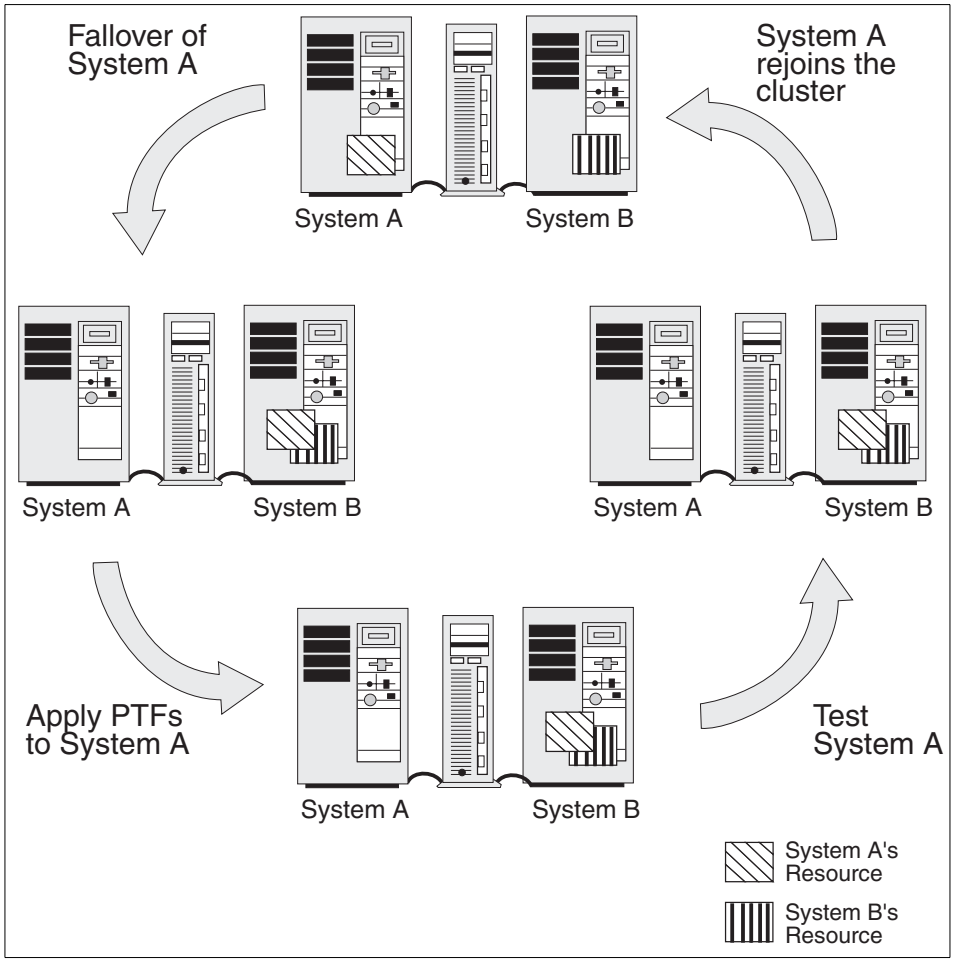


Figure 8-2 Applying a PTF to a cluster node

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 example, 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 is actually very important which of the several backup commands you are using and also what your strategy is. For the features and/or restrictions of backup commands, such as **mksysb**, **pax**, **savevg**, **rdump**, **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 mark 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.

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 volume 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, that is, 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 resources 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/ES Customization Examples*, SG24-4498.

8.8 User management

As Section 2.8, “User ID planning” on page 62 described, on an HACMP cluster, the administrator has to take care of user and group IDs throughout the cluster. If they do not match, the user will not get anything done after a failover happens. 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.

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 cluster node 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 C-SPOC password enhancement

Since HACMP Version 4.4, C-SPOC has been enhanced to set password for users accounts for the whole cluster or per resource group, instead of individually setting the password for user accounts on each node in the cluster with the **passwd** command.

Note: In HACMP Version 4.4, this password is only for setting up initial passwords and the users will be prompt to change there password when they log in.

Since HACMP Version 4.4.1, C-SPOC has been enhanced to set not only the initial password, but you can also choose to not set the “force change” flag into the file /etc/security/passwd; by doing this, the users do not have to change their passwords when they log in to the system.

Use **smit cl_passwd** or, in HACMP HAS, the **/usr/sbin/cluster/cspoc/fix_args nop cl_chpasswd** command, or, in HACMP/ES, **/usr/es/sbin/cluster/cspoc/fix_args nop cl_chpasswd command**.

You must be the root user to change a user's password. When changing a user's password, you must follow the password restrictions and conventions for the system as specified in each user's stanza in the /etc/security/user configuration file.

For security reasons, the clear password is never transmitted over the network, only the encrypted password from the file /etc/security/passwd.

Note: Do not use the **cl_chpasswd** command if you have a Network Information Service (NIS) database installed on any node in the cluster. The command in this environment can cause NIS database inconsistence.

8.8.4 Set or change a password using C-SPOC

To configure a password, use **smit cl_passwd** or the **cl_chpasswd** command.

If you do not select any nodes by resource groups for the user account, that is, to be configured, as shown in Example 8-39, then the password configuration will be changed for all nodes in the cluster.

Example 8-39 Change a User's Password in the Cluster output

```
                  Change a User's Password in the Cluster

Type or select a value for the entry field.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Select nodes by Resource Group                    []
    *** No selection means all nodes! ***
```

Example 8-40 shows how to select the user that will have the password set or configured. If the password is set for the first time after the user account has been created, or if this is to change an user's password, the **smit cl_passwd** or smit **cl_chpasswd** can be used for both issues. You will be prompted to enter the user's new password and then prompted again to re-enter the new password.

The user *must* change the password on first login. Here you specify whether the force change flag should be set in the /etc/security/passwd file. By setting this to `true`, it will require the user to change the password on each node at the next login; this is the default and also the AIX default behavior. If set to `false`, the user will not be required to change the password on the next log in.

Example 8-40 cl_passwd output

```
                Change a User's Password in the Cluster

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

                                                        [Entry Fields]
  Selection nodes by resource group
   *** No selection means all nodes! ***

* User NAME                                             []
  User must change password on first login?             true
```

Note: If any node in the cluster is a member of an SP with the usermgmt_config flag set to `true`, the **cl_chpasswd** command fails.

This command uses the AIX **rsh** facility to propagate commands to other nodes, and therefore requires the proper /.rhosts access to all nodes (unless you are using Kerberos on your system). Thus, each node must have a /.rhosts file that includes references to all boot and service interfaces for each cluster node.

8.8.5 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.

8.8.6 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.

8.8.7 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, which are very similar 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 4.4.1: Administration Guide*, SC23-4279.

8.8.8 C-SPOC log

Because these commands are running and executing while distributed among the cluster, it could happen that something does not 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 does not have to be the same in all cases, so the log file might be present on different cluster nodes, and does not contain the same data.

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. This section 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 (CWS) 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 necessary hardware (see Figure 9-1 on page 285).

Both the tty network and the RS/6000 SP internal ethernet are extended to the backup CWS. In contrast to standard HACMP, you do not 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.

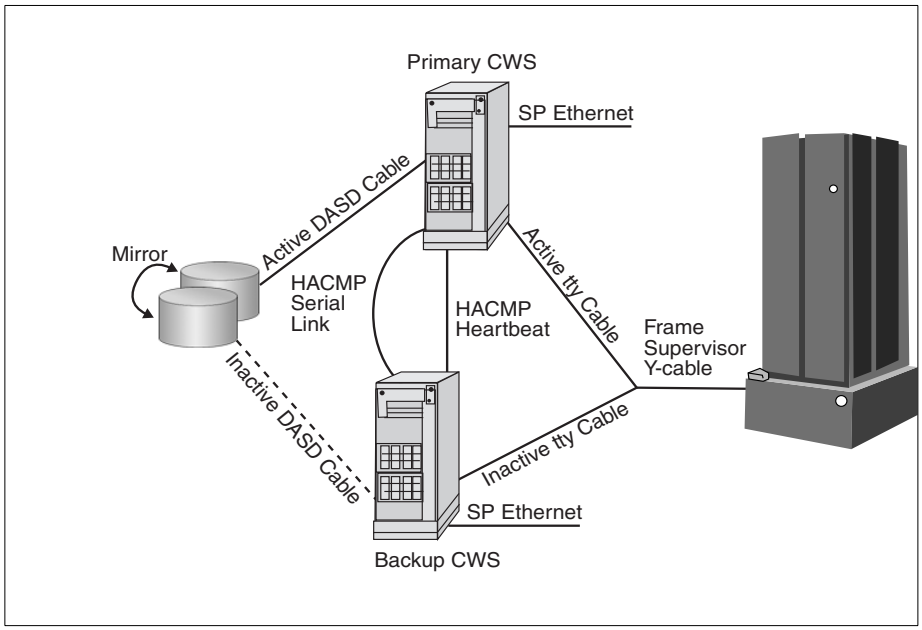


Figure 9-1 A simple HACWS environment

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 HACMP on the same level as well. For example, if you want to use HACMP Version 4.4.1 for AIX, you have to use PSSP Version 3.2 and AIX Version 4.3.3 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 4.4.1: Installation Guide*, SC23-4278. Verification, as described in Chapter 10 of the *HACMP for AIX, Version 4.4.1: 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, and 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, 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 the 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 described in Chapter 7 of the *HACMP for AIX, Version 4.4.1: 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 4.4.1: Installation Guide*, SC23-4278-02, 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 host name 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 host name, in order to finish the configuration and check that everything is in order. Check if the primary CWS can address the backup CWS by its host name and vice versa. If not, use the **ifconfig** command to temporarily set the interface to the host name 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 two 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 The process by which an entity tells another who it is.

Authentication The process by which the other entity verifies this identity.

Authorization 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.

The following is 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 service's servers. Tickets have a finite life, known as the ticket life span.

In Kerberos terms, making a Kerberos authenticated service provider work on the 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

With HACMP 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, for example, godm, an IP label, such as 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 the **clverify** command, which goes out to the nodes and checks the consistency of files.

9.2.2 Enhanced security option in PSSP 3.2

In PSSP Version 3.2, an enhanced level of security has been available, which makes it possible to use DCE authentication instead of Kerberos 4 authentication. However, this option may effect your HACMP functionality. Read the following before you decide your cluster security planning.

PSSP Version 3.2 enhanced security level

The enhanced security options in PSSP Version 3.2 removes the ability for PSSP to internally use **rsh** and **rcp** commands as root from a node.This security mode can be used with Kerberos 4 or DCE, but when HACMP is configured on a SP complex, DCE should not be used as the security management protocol for those partitions where HACMP is implemented; the same is also true for HACWS.

Note: If PSSP Version 3.2 is used with HACMP, then you must use HACMP Version 4.4.0 or newer versions of HACMP. Also, HACMP and HACMP/ES can coexist in the same partition, but not in the same HACMP cluster.

Only one version of HACMP can be installed on a node.

Because the root authority is mandatory for the administrator to run HACMP, and if you want to enable the enhanced security options, you must resolve some issues to make it possible to run your HACMP cluster together with this feature. To get more information about PSSP Version 3.2 enhanced security, see the redbook *Exploiting RS/6000 SP Security: Keeping it Safe*, SG24-5521.

To authorize a root user to run **rsh** and **rcp** commands on other nodes (if the enhanced security option has been enabled), do the following:

1. Before you alter or synchronize your HACMP cluster, you must create/update, on each node in the HACMP cluster, the root user **rsh** authorization file (either /.rhosts or /.klogin) for those nodes that need root user authorization to run the **rsh** command on this node. Also, the appropriate AIX remote command authorization method must be enabled on the nodes in the HACMP cluster.
2. Perform any desired alteration, verification, and synchronization of the HACMP cluster configuration.
3. Optional: If, after step 2, you do not need the root user authorization for the **rsh** command, then the administrator can remove the entries in the authorization files.

Note: If any resource group is configured to use cascading without fallback, then the root user must have **rsh** authorization on those nodes that are part of the CWOF resource group.

9.2.3 Configure HACMP cluster security mode

To use Kerberos authentication in HACMP, you must first have set up and configured Kerberos on all nodes in the cluster prior to changing the security option in HACMP.

Before you change the cluster to use Kerberos authentication, the /.rhosts file on every node in the cluster must be removed. The HACMP cluster security mode is a part of HACMP cluster ODM, so the changes will alter the cluster topology settings. Therefore, after the HACMP cluster is changed to use Kerberos, the cluster topology must be synchronized to all the nodes in the cluster, for the same reason the DARE cannot be used to change the security mode.

To set the security mode to enhanced and synchronize it to all the other nodes in the cluster, do the following:

1. Use **smit hacmp** and select Cluster Configuration. You will then select Cluster Security, as can be seen in Example 9-1.

Example 9-1 Cluster Configuration output

```
                          Cluster Configuration

Move cursor to desired item and press Enter.

  Cluster Topology
  Cluster Security
  Cluster Resources
  Cluster Snapshots
  Cluster Verification
  Cluster Custom Modification
  Restore System Default Configuration from Active Configuration
  Advanced Performance Tuning Parameters
```

2. Set the security mode to enhanced (see Example 9-2).

Example 9-2 Change / Show Cluster Security output

```
                    Change / Show Cluster Security

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

                                                        [Entry Fields]
* Cluster Security Mode                                 Enhanced
  WARNING: The /.rhosts file must be removed
  from ALL nodes in the cluster when the
  security mode is set to 'Enhanced'.
  Failure to remove this file makes it
  possible for the authentication server
  to become compromised. Once the server
  has been compromised, all authentication
  passwords must be changed.

  Changes to the cluster security mode
  setting alter the cluster topology
  configuration, and therefore need to be
  synchronized across cluster nodes. Since
  cluster security mode changes are seen as
  topology changes, they cannot be performed
  along with dynamic cluster resource
  reconfigurations.
```

3. After the security mode configuration has been successfully, you must synchronize the cluster topology. This can be done through the SMIT HACMP menu or by issuing the **`/usr/sbin/cluster/diag/clverify cluster topology sync`** command.

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 Disks (VSDs)

Virtual Shared Disk (VSD) allows data in logical volumes on disks physically connected to one node to be transparently accessed by other nodes. More 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.

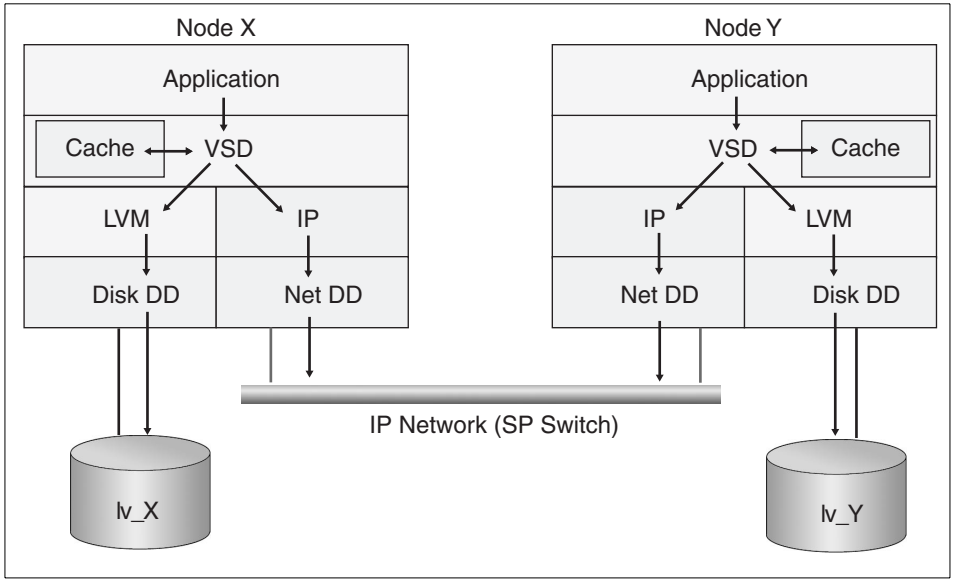


Figure 9-2 VSD architecture

With reference to Figure 9-2 on page 293, 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. The VSD cache is shared by all VSDs configured on a node. Data is stored in 4 KB blocks, a size optimized for Oracle Parallel Server. If your I/O patterns involve I/O operations larger than 4 KB, we recommend disabling the 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 good 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 *Managing Shared Disks*, SA22-7349.

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), without 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 9-3.

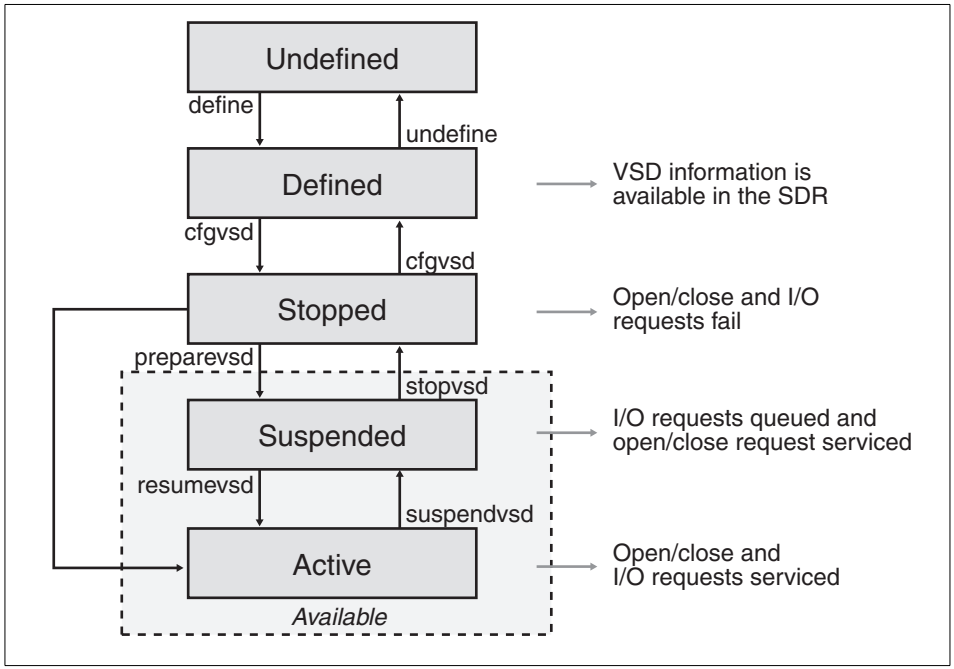


Figure 9-3 VSD state transitions

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.

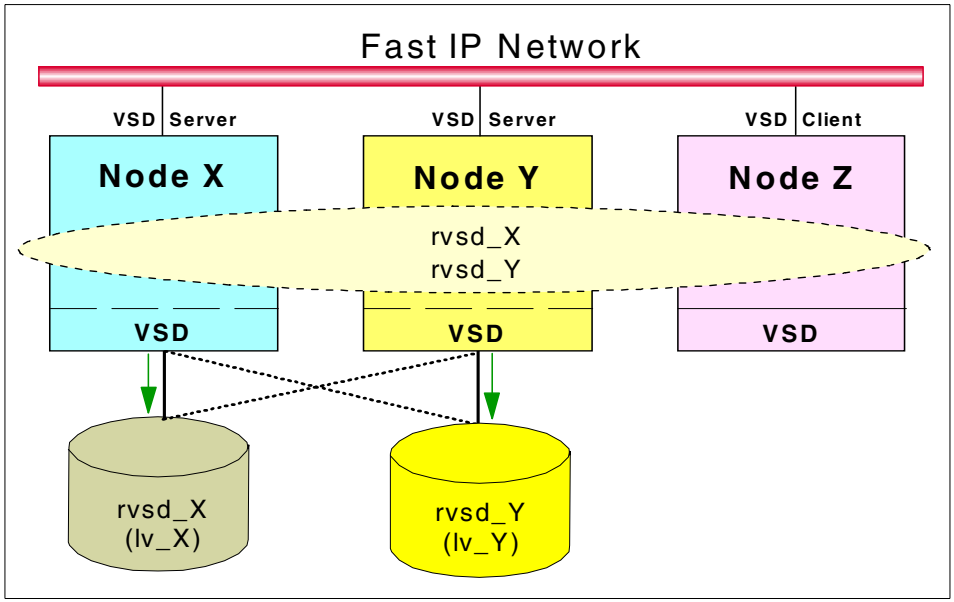


Figure 9-4 RVSD function

In reference to Figure 9-4, 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 9-5 on page 298. The RVSD daemon controls recovery. It invokes the recovery scripts whenever there is a change in the group membership, which it recognizes 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.

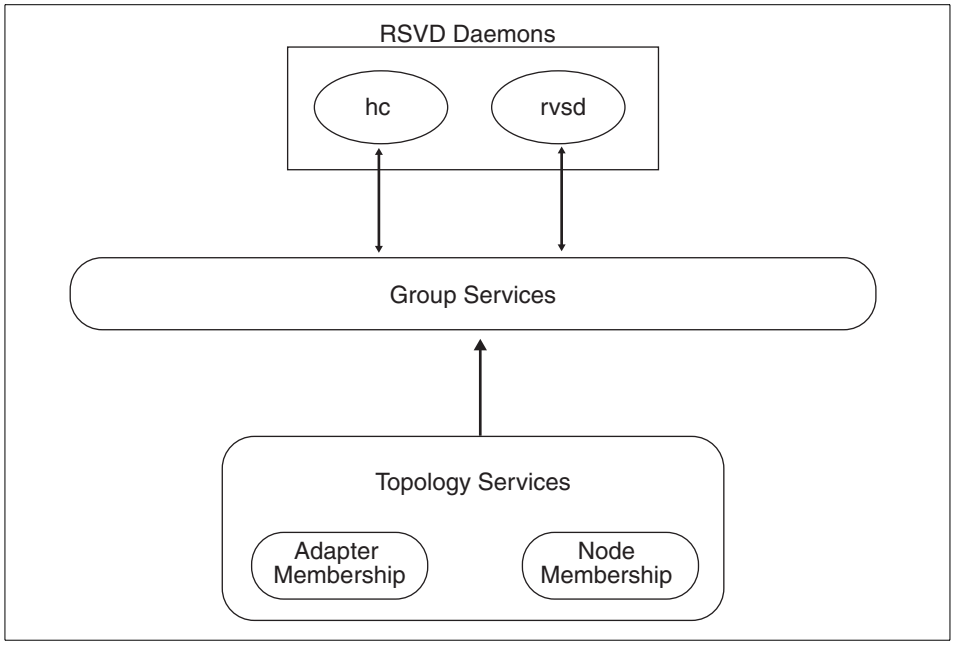


Figure 9-5 RVSD subsystem and HA infrastructure

9.3.3 Concurrent Virtual Shared Disks instead of RVSD

Virtual Shared Disks (VSD) includes concurrent disk access, which allows you to use multiple servers to satisfy disk requests by taking advantage of the concurrent disk access environment. VSD uses the services of the Concurrent Logical Volume Manager (CLVM), which provides the synchronization of LVM and the management of concurrency for system administration services.

Concurrent disk access extends the physical connectivity of multi-tailed concurrent disks beyond their physical boundaries. You can configure volume groups (VG) with a list of Virtual Disk servers. Nodes that are not locally attached have their I/O distributed across these servers (see Figure 9-6 on page 299).

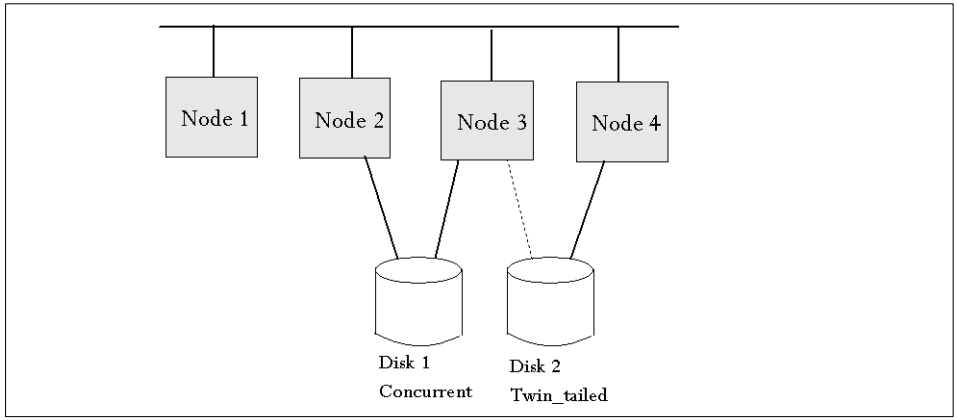


Figure 9-6 CVSD and RSVD I/O example

When you are using IBM CVSD, recovery from node failure is faster because the failed node is marked unavailable to all the other nodes and its access to the physical disks is fenced. This procedure is faster than the recovery procedure followed in the twin-tailed environment. An additional benefit from multiple VSD servers is that the disks services can be spread across multiple servers.

CVSD recommendations

There are some issues of CVSD that you should be aware of before you plan your CVSD environment:

- Only 2-way CVSD is supported.
  - This is a current limitation due to CLVM.
  - There is no mirror write consistency; we suggest RAID 5.
- CVSD is only supported for SSA.
- CVSDs do not support VSD caching.
- HACMP and VSD both use the CLVM, and if both are installed, then only one product is permitted to provide concurrent disk access.

CVSD node down and node reintegration

RSVD scripts has been changed with the implementation of CVSD.

Node down

Normally, when a node goes down, it will release the VGs that it has been serving and a backup node will take over the VGs and their I/Os to the VSDs, which are resumed on the backup node.

But with CVSD, the VGs are already online and available to another server, so the I/O can simply be resumed to the active node. To handle false failures and maintain data integrity, the CVSD will fence off the disks (using SSA fencing) so the node that went down not will be able to use the I/O.

Node reintegration

Normally, when a primary node comes back up, the VGs are varied back online to it, and the I/O is resumed to its primary node. In the case of CVSD, when any servers can come back up, the VGs will be concurrently varied online to it, and I/O will be resumed using all the CVSD servers.

Note: In both cases, all of the VSD's served by the new server node will be made active on the new server node, and also the new node will be unfenced, if it was fenced, so that the new node can access the SSA disks.

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, and the current supported type of switch for HACMP at customer sites is SP Switch, also known as the TB3 switch.

9.4.1 SP Switch support

In an SP system with the SP Switch, nodes can run different levels of PSSP, including PSSP Version 3.2; also, the system can be partitioned.

This type of switch has a good availability design from the hardware point of view, so the SP Switch was designed to do all actions regarding the switch fabric on the link level. For example, if the **`Estart`** command was used to get new nodes up on the switch, only the selected node is affected. All the others continue working without even noticing that something has happened on the switch network.

9.4.2 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, and each css0 service adapter can take over up to seven additional IP addressers by using aliasing. For SP Switch, 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, that is, 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.

9.4.3 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 HACMP Version 4.3, so if you upgrade to HACMP Version 4.4 or later on an RS/6000 SP and also upgrade your HPS switch to an SP switch, then you must first upgrade your SP Switch before you install HACMP Version 4.4 or later. If you are currently running HACMP Eprimary management with an HPS switch, you should first run the HACMP for AIX script to unmanage the Eprimary before you upgrade the switch.

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 does not have to take care of it any more.

9.4.4 Switch failures

As mentioned before, a node with a SP Switch is restricted to having 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 “Single point of failure hardware component recovery” on page 61. 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 shutdown 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.

10

HACMP classic versus HACMP/ES

This chapter will cover the differences between the HACMP versions that will be discussed in this section. This will give you some idea of what kind of HACMP version that would perhaps be the solution for your environment.

The subsections that follow will need a brief description of HACMP and HACMP/ES and also the features that comprise the HACMP and HACMP/ES.

The following options are referred to as:

HACMP classic

- High Availability Subsystem (HAS)
- Concurrent Resource Manager (CRM)
- High Availability Network File System (HANFS); this is included in HACMP and HACMP/ES since Version 4.4.0

HACMP/ES

- Enhanced Scalability (ES)
- Enhanced Scalability Concurrent Resource Manager (ESCRM)

10.1 HACMP classic

HAS and CRM

High Availability Subsystem uses the global ODM to store information about the cluster configuration and can have up to eight HACMP nodes in a HAS cluster. HAS provides the base services for cluster membership, system management, configuration intergrity, and control, failover, recovery, cluster status, and monitoring facilities are also there for the programmer and system administrator.

The Concurrent Resource Manager feature optionally adds the concurrent shared-access management for supported RAID and SSA disk subsystem. Concurrent access is provided at the raw logical volume level and the applications that use CRM must be able to control access to the shared data. The CRM includes the HAS, which provides distributed locking facility to support access to shared data.

Concurrent access configurations do not support journaled file systems.

In an HACMP Version 4.4.1 environment, concurrent access is available using only an IBM 7135-110 or 7135-210 disk array, an IBM 7137 disk array, IBM 2105-B09 and 2105-100 Versatile Storage Servers or IBM 2105 E-10 and E-20 Enterprise Storage Servers, an IBM 7133 SSA disk subsystem or an IBM 9333 disk subsystem. RAID devices from other manufacturers may not support concurrent access.

10.2 HANFS

Before HACMP Version 4.4.0, if there was a need for a system to have high availability on a network file system (NFS), then the system had to use high availability for network file system (HANFS). HANFS Version 4.3.1 and earlier 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.

Since HACMP Version 4.4.0, the HANFS features is included in HACMP and therefore the HANFS is no longer an separate software. For more information about high availability for NFS from HACMP Version 4.4, see Section 5.4.2, “Exporting NFS file systems and directories” on page 156.

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.

10.3 HACMP/ES and ESCRM

This was originally only available for SP environment, where tools were already in place with Parallel Systems Support Program (PSSP) to manage larger clusters. As of AIX Version 4.3.2 and PSSP Version 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 since Version 4.3 by using the rsct.core fileset.

Since HACMP/ES and ESCRM Version 4.4, application monitoring is available. This provides a process application monitoring and user-defined application monitoring of one or multiple processes for an application. This feature determines the status of the process or processes, and if those processes would fail they can be restarted, the system can be notified, and the resource group to which the monitored applications application server is present in can also fall over to another node that is a participating in that resource group.

HACMP/ES and ESCRM builds on the Event Management and Group Services facilities of RSCT for AIX to scale HACMP up to 32 HACMP nodes.

ESCRM optionally adds concurrent shared-access management for supported RAID and SSA disk subsystems. Concurrent access is provided at the raw disk level. The application must support some mechanism to control access to the shared data, such as locking. The ESCRM components includes the HACMP/ES components and the HACMP distributed lock manager.

10.3.1 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. But the RSCT that is included with AIX, is not the same RSCT that is included with PSSP.

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.

HACMP classic and HACMP/ES differ 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 (NIM).

Network Interface Modules (NIM)

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, by providing a link to other nodes on the network it monitors, and by initiating adapter swaps on certain networks.

- See the *HACMP for AIX 4.4.1: Enhanced Scalability & Administration Guide*, SC23-4284, for more information on these services.

10.3.2 Enhanced cluster security

With HACMP Version 4.3 comes an option to switch the 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 this file 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 it claims to be. This is done by granting access on the basis of tickets, which are provided only to those hosts having the correct identification.

For more information, see Section 9.2.2, “Enhanced security option in PSSP 3.2” on page 290.

10.4 Similarities and differences

All five 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 HAS and HACMP/ES. HACMP HAS 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 heart beat. Since Version 4.3, the restriction to run HACMP/ES on RS/6000 SP systems has been withdrawn. However, if you run it on an RS/6000 SP, you need to have PSSP Version 3.1 or later installed. As the HPS Switch is no longer supported with PSSP Version 3.1 or HACMP/ES Version 4.3.1, you need to upgrade to the SP Switch, in case you have not 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.

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.

For switchless RS/6000 SP systems or SPs with the SP Switch, the decision will be based on a more functional level.

Event management (EM) is much more flexible in HACMP/ES, since you can define custom events. Also, in AIX5L, the EM is part of the rsct.core fileset. 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 and *AIX 5L Performance Tools Handbook*, SG24-6039.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 311.

- *AIX 5L Performance Tools Handbook*, SG24-6039
- *Bullet-Proofing Your Oracle Database with HACMP: A Guide to Implementing AIX Databases with HACMP*, SG24-4788
- *Exploiting HACMP 4.4: Enhancing the Capabilities of Cluster Multi-Processing*, SG24-5979
- *Exploiting RS/6000 SP Security: Keeping it Safe*, SG24-5521
- *An HACMP Cookbook*, SG24-4553
- *HACMP Enhanced Scalability Handbook*, SG24-5328
- *HACMP Enhanced Scalability: User-Defined Events*, SG24-5327
- *HACMP/ES Customization Examples*, SG24-4498
- *Highly Available IBM eNetwork Firewall: Using HACMP or eNetwork Dispatcher*, SG24-5136
- *Inside the RS/6000 SP*, SG24-5145
- *Migrating to HACMP/ES*, SG24-5526
- *Monitoring and Managing IBM SSA Disk Subsystems*, SG24-5251
- *A Practical Guide to Serial Storage Architecture for AIX*, SG24-4599
- *Understanding SSA Subsystems in Your Environment*, SG24-5750

Other resources

These publications are also relevant as further information sources:

- *7133 Models 010 and 020 SSA Disk Subsystems: Installation Guide*, GA33-3260

- *7133 Models 500 and 600 SSA Disk Subsystems: Installation Guide*, GA33-3263
- *7133 SSA Disk Subsystem: Operator Guide*, GA33-3259
- *7133 SSA Disk Subsystems: Service Guide*, SY33-0185
- *7133 SSA Disk Subsystems for Open Attachment: Installation and User's Guide*, SA33-3273
- *7133 SSA Disk Subsystems for Open Attachment: Service Guide*, SY33-0191
- *HACMP for AIX 4.4.1: Administration Guide*, SC23-4279
- *HACMP for AIX 4.4.1: Concepts and Facilities*, SC23-4276
- *HACMP for AIX 4.4.1: Enhanced Scalability & Administration Guide*, SC23-4284
- *HACMP for AIX 4.4.1: Installation Guide*, SC23-4278
- *HACMP for AIX 4.4.1: Planning Guide*, SC23-4277
- *HACMP for AIX 4.4.1: Trouble Shooting Guide*, SC23-4280
- *IBM Parallel System Support Programs for AIX Installation and Migration Guide*, GA22-7347
- *IBM RS/6000 SP Planning Volume 2,Control Workstation and Software Environment*, GA22-7281
- *Managing Shared Disks*, SA22-7349
- *PCI Adapter Placement Reference Guide*, SA38-0538
- *RS/6000 Adapters, Devices and Cable Information for Micro Channel Bus Systems*, SA38-0533
- *RS/6000 Adapters, Devices and Cable Information for Multiple Bus Systems*, SA38-0516
- *Enterprise Storage Server Fibre Channel Attachement Version 6.0*, found at:

  `http://www.storage.ibm.com/hardsoft/products/ess/whitepaper.htm`

Referenced Web sites

These Web sites are also relevant as further information sources (some of them are IBM intranet Web sites):

- IBM Certification Home page:

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

- IBM Enterprise Storage Server White Papers

  `http://www.storage.ibm.com/hardsoft/products/ess/whitepaper.htm`
- SSA Support site

  `http://www.storage.ibm.com/hardsoft/products/ssa/index.html`
- Documentation Library

  `http://www.rs6000.ibm.com/resource/aix_resource/Pubs/`
- The HA-DASD Home Page

  `http://hacmp.aix.dfw.ibm.com/`
- Download microcodes

  `http://techsupport.services.ibm.com/server/nav?fetch=hm`

How to get IBM Redbooks

Search for additional Redbooks or Redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

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 Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Special notices

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, North Castle Drive, Armonk, NY 10504-1785.

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 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.

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere.,TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

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

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

Other company, product, and service names may be trademarks or service marks of others

Abbreviations and acronyms

DMSITSMTBF

AIX	Advanced Interactive eXecutive
APA	All Points Addressable
APAR	Authorized Program Analysis Report
ARP	Address Resolution Protocol
ASCII	American Standard Code for Information Interchange
CHRP	Common Hardware Reference Platform
CLM	Cluster Lock Manager
CLVM	Concurrent Logical Volume Manager
CPU	Central Processing Unit
CRM	Concurrent Resource Manager
CVSD	Concurrent Virtual Shared Disks
CWOF	Cascading Without Fallback
CWS	Control Work Station
DCD	Default Configuration Directory
DE	Differential Ended
DLC	Data Link Control
DNS	Domain Name Services
DMS	Deadman Switch
DSMIT	Distributed System Management Interface Tool
EM	Event Manager
ES	Enhanced Scalability
ESCRM	Enhanced Scalability Concurrent Resource Manager
ESS	Enterprise Storage Server
FC-AL	Fibre Channel - Arbitrated Loop
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
HACWS	High Availability Control Work Station
HANFS	High Availability Network File System
HAS	High Availability Subsystem
HCON	Host Connection Program
I/O	Input/Output
IBM	International Business Machines Corporation
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
LIP	Loop Initialization Process
LU	Logical Unit (SNA definition)
LUN	Logical Unit (RAID definition)
LVM	Logical Volume Manager

MAP	Maintenance Analysis Procedure
MIB	Management Information Base
MB	Megabyte
MTBF	Mean Time Between Failure
NETBIOS	Network Basic Input/Output System
NFS	Network File System
NIM	Network Interface Module
NIS	Network Information Service
ODM	Object Data Manager
PSSP	Parallel System Support Program
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
RSCT	RISC System Cluster Technology
RVSD	Recoverable Virtual Shared Disks
SCSI	Small Computer Systems Interface
SLIP	Serial Line Interface Protocol
SMIT	System Management Interface Tool
SMP	Symmetric Multi-Processor
SMUX	SNMP (see below) Multiplexor
SNMP	Simple Network Management Protocol
SOCC	Serial Optical Channel Converter
SPOF	Single Point of Failure
SSA	Serial Storage Architecture
SRC	System Resource Controller
SRN	Service Request Number
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Interface Protocol
UDP	User Datagram Protocol
UPS	Uninterruptable Power Supply
VGSA	Volume Group Status Area
VGDA	Volume Group Descriptor Area
VSD	Virtual Shared Disks
WAN	Wide Area Network
WWNN	World Wide Node Name
WWPN	World Wide Port Name
XRC	Extended Remote Copy

Index

C

D

E

M

N

O

P

Q

R

S

IBM @server Certification Study Guide - pSeries HACMP for AIX

IBM

IBM @server Certification Study Guide - pSeries HACMP for AIX

Redbooks

Update new features for the latest version of HACMP for AIX

Valuable guide for HACMP for AIX administrators

Get ready for pSeries HACMP for AIX certification

This redbook is designed as a study guide for professionals wishing to prepare for the certification exam to achieve IBM @server Certified Systems Expert - pSeries HACMP for AIX. The pSeries HACMP for AIX certification validates the skills required to successfully plan, install, configure, and support an HACMP for AIX cluster installation.

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.

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 pSeries HACMP for AIX certification exam, or just want to validate your HACMP skills, this redbook is for you.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6187-00 ISBN 0738423793