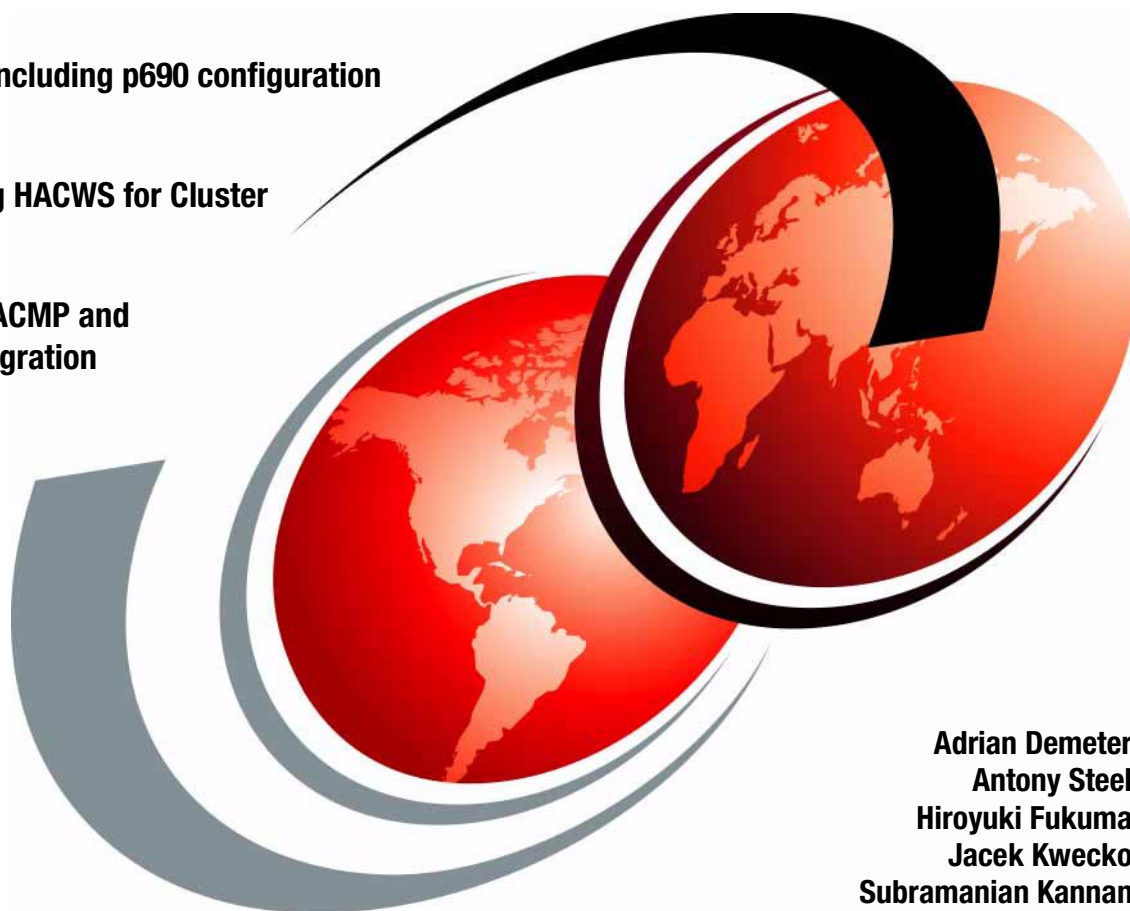


Configuring Highly Available Clusters Using HACMP 4.5

Examples including p690 configuration

Configuring HACWS for Cluster 1600

Explains HACMP and HAGEO integration



Adrian Demeter
Antony Steel
Hiroyuki Fukuma
Jacek Kwecko
Subramanian Kannan



International Technical Support Organization

**Configuring Highly Available Clusters Using
HACMP 4.5**

October 2002

Note: Before using this information and the product it supports, read the information in “Notices” on page xiii.

Second Edition (October 2002)

This edition applies to Version 4, Release 5, of High Availability Cluster Multi-Processing for AIX.

Note: This book is based on a pre-GA version of a product and may not apply when the product becomes generally available. We recommend that you consult the product documentation or follow-on versions of this redbook for more current information.

© Copyright International Business Machines Corporation 2002. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	vii
Tables	xi
Notices	xiii
Trademarks	xiv
Preface	xv
The team that wrote this redbook	xv
Become a published author	xvi
Comments welcome	xvii
Summary of changes	xix
October 2002, Second Edition	xix
Chapter 1. HACMP 4.5 overview	1
1.1 Introduction to HACMP	2
1.2 Requirements and prerequisites	3
1.2.1 Supported hardware	3
1.2.2 Required software levels	3
1.3 New features and functions	4
1.3.1 Usability enhancements	5
1.3.2 Administrative enhancements	22
1.3.3 Network enhancements	32
1.3.4 Device support	44
1.3.5 Application support	47
1.4 Installation and migration considerations	58
Chapter 2. Configuring highly available p690 clusters	61
2.1 LPAR	62
2.2 Hardware Management Console (HMC)	63
2.3 Planning considerations	63
2.3.1 System configuration	63
2.3.2 HMC high availability	68
2.3.3 Network	69
2.3.4 Storage	70
2.3.5 Software	71
2.4 Clustering with HACMP	72
2.4.1 Lab environment	72

2.4.2	System configuration.	72
2.4.3	Preparing the cluster for high availability	75
2.4.4	Define the LPARs configuration on both p690s	78
2.4.5	Configure the nodes	79
2.4.6	Installing HACMP	84
2.4.7	Scenario 1: Cluster with 2 Ethernet and SSA storage	87
2.4.8	Scenario 2: Using SP Switch/SP Switch2 adapter	116
2.4.9	Scenario 3: Dual SP Switch2 network.	128
2.4.10	Scenario 4: IP Aliasing	129
2.4.11	Scenario 5: Integrating ESS storage into HACMP	138
Chapter 3.	HACWS: An HACMP Application for Cluster 1600.	157
3.1	HACWS.	158
3.2	Definitions	159
3.3	Requirements	159
3.3.1	Hardware requirements.	159
3.3.2	Software requirements	161
3.4	Operation of HACWS	162
3.4.1	Components of HACWS	162
3.4.2	Planning the HACMP configuration.	165
3.4.3	Requirements imposed on the logical definitions	165
3.5	Configurations used in this document	168
3.5.1	SP frame only with no standby adapters on the CWS	169
3.5.2	SP frame only with standby adapters on the primary CWS	169
3.5.3	SP Frame and p690 with the CWS and HMC on same SPLAN	170
3.5.4	CWS and HMC on a private network other than SPLAN	171
3.5.5	CWS and HMC on a private network with standby adapters	172
3.5.6	IP labels and networks	173
3.6	Installing and configuring HACWS	174
3.6.1	Preparation	174
3.6.2	Configuration of the backup control workstation	175
3.6.3	Kerberos configuration on the backup control workstation	176
3.6.4	Install HACMP/ES on both control workstations	182
3.6.5	Install HACWS	182
3.6.6	Configure HACWS	183
3.6.7	Configure HACMP topology	187
3.6.8	Set up the HACWS configuration	208
3.6.9	Verify HACWS and hardware configuration	213
3.6.10	Reboot primary and start cluster services.	214
3.6.11	Verify operation of the primary control workstation	215
3.6.12	Start the backup control workstation	217
3.6.13	Starting of cluster services on the primary workstation.	217
3.6.14	Backups	218

3.6.15 Testing HACWS	218
3.7 Considerations	222
Chapter 4. HAGEO integration with HACMP cluster	223
4.1 HAGEO integration with HACMP	224
4.1.1 History	224
4.2 Planning	225
4.2.1 Hardware requirements	226
4.2.2 Software requirements	232
4.2.3 Configuration examples	233
4.3 New features of HAGEO 2.4	235
4.3.1 Integration with HACMP	235
4.3.2 TCP option for remote mirroring	242
4.3.3 Selection of temporal ordering policies	244
4.3.4 Support for 64-bit kernel environment	245
4.4 Clustering with HAGEO	246
4.4.1 Configure geographic topology	247
4.4.2 Configure GeoMirror devices	250
4.4.3 Managing the Geo Cluster	256
4.4.4 Performance considerations	258
4.4.5 Migration considerations	258
4.4.6 Troubleshooting	259
4.4.7 Maintenance considerations	271
Abbreviations and acronyms	273
Related publications	275
IBM Redbooks	275
Other resources	275
Referenced Web sites	276
How to get IBM Redbooks	276
IBM Redbooks collections	276
Index	277

Figures

1-1	New option to accept license agreements	4
1-2	Online Planning Worksheet	7
1-3	Creating a custom pager notification method	9
1-4	Send a Test Page	9
1-5	Example test page status (retry count was 3)	10
1-6	Dynamic node priority	11
1-7	User defined Dynamic node priority policy.	12
1-8	Change/Show Resource Group Processing Order	17
1-9	Acquisition and release order	17
1-10	Resource Group synchronization adding error notification entry	21
1-11	Change/Show Time Until Warning.	23
1-12	clstat with Web browser.	25
1-13	Display Event Summaries	26
1-14	Change/Show a Cluster Log Directory.	29
1-15	Application Availability Analysis	29
1-16	The output of the AAAT	30
1-17	Configuring a persistent IP alias	34
1-18	Add an Initial Interface	38
1-19	Quick configuration of IP Interfaces	39
1-20	Add Multiple Service IP Labels to a Network	41
1-21	Synchronous versus asynchronous operation	46
1-22	Creating a Workload Manager configuration	51
1-23	Define Workload Manager configuration name to HACMP	52
1-24	Adding Workload Manager classes to a resource group	53
1-25	Verification of Workload Manager by clverify	54
2-1	High availability cluster with one LPAR per p690 server	65
2-2	High availability cluster with two or more LPARs in one p690 server	66
2-3	High availability cluster with one LPAR in a p690 and a pSeries server	67
2-4	Highly available cluster with one LPAR in a p690 in a Cluster 1600	68
2-5	HW configuration of testing environment.	73
2-6	LPAR configuration as seen from the HMC console	78
2-7	SMIT TCPIP: Minimum Configuration & Startup	80
2-8	SMIT TCPIP: Change/Show a Standard Ethernet Interface	81
2-9	SMIT: Install Software	86
2-10	An example of a cluster for scenario 1.	88
2-11	Add a Cluster Definition	91
2-12	Add a Cluster Nodes	91
2-13	Add an Initial Interface	92

2-14	Add an Initial Interface	93
2-15	Add Multiple IP-based Interfaces	94
2-16	Add Multiple Service IP Labels to a Network	95
2-17	Change / Show an Interface / IP Label	96
2-18	Change / Show an Interface / IP Label	97
2-19	Add IP Labels Requiring Individual Configuration	98
2-20	Add IP Labels Requiring Individual Configuration	98
2-21	Add a Non IP-based Adapter	99
2-22	Add a Non IP-based Adapter	100
2-23	Add a Resource Group	101
2-24	Add a Resource Group	101
2-25	Add an Application Server	102
2-26	Add an Application Server	102
2-27	Change/Show Resources/Attributes for Resource Group	105
2-28	Change/Show Resources/Attributes for a Resource Group	106
2-29	Start Cluster Services	107
2-30	Start Cluster Services (SPOC)	108
2-31	Cluster status	109
2-32	Cluster status	112
2-33	Stop Cluster Services	115
2-34	Stop Cluster Services (SPOC)	116
2-35	SP Switch network addresses	118
2-36	SMIT HACMP - Add an IP-based Network	122
2-37	SMIT HACMP - Add an Initial Interface	123
2-38	SMIT HACMP menu - Change an IP-based Network	124
2-39	Change/Show Resources/Attributes for a Resource Group	125
2-40	Dual SP Switch2 configuration.	129
2-41	An example of a cluster for scenario 4.	130
2-42	Cluster status	135
2-43	ESS Interconnection diagram	139
2-44	Add a Volume Group with Data Path Devices	144
2-45	Add the shared volume group to the HACMP resource group.	148
2-46	SMIT - Add a Volume Group with Data Path Devices	151
2-47	SMIT - Add a Physical Volume	154
3-1	Highly available control workstation configuration	160
3-2	IP Label configuration	167
3-3	One frame with standby adapters on the control workstations	170
3-4	p690s on an SP administration network	171
3-5	p690s on a private network other than SPLAN	172
3-6	p690s on private network with optional standby adapters	173
3-7	Installing the HACWS fileset	183
3-8	SMIT chinet to set boot IP label on adapter.	185
3-9	Set cluster ID and name	187

3-10	Configuring cluster nodes	188
3-11	Add Initial Interfaces	189
3-12	Add boot IP label for the primary CWS	190
3-13	Add boot IP label for the backup CWS	191
3-14	Add boot IP label for the primary CWS on a private network.	192
3-15	Add boot IP label for the backup CWS on a private network.	193
3-16	Add standby IP label for the primary CWS	194
3-17	Add standby IP label for the backup CWS.	195
3-18	Add standby IP label for the primary CWS on a private network	196
3-19	Add standby IP label for the backup CWS on a private network	197
3-20	Discover IP Topology.	197
3-21	Adding a service IP label	198
3-22	Add service IP label on private network.	199
3-23	Adding a persistent IP label	200
3-24	Subnets for hacwsether	200
3-25	Synchronize Cluster Topology	201
3-26	Add a Non IP-base Adapter	204
3-27	Add an Application Server	205
3-28	Add a Resource Group	206
3-29	Adding a resource to hacws_group1	207
3-30	Synchronize and verify the cluster resources	208
3-31	Install and Configure HACWS	209
3-32	Output from the install_hacws command.	211
3-33	Add event scripts to HACMP database	212
3-34	spcw_addevents output	212
3-35	Script hacws_verify output	213
3-36	Script spcw_verify_cabling output	213
3-37	Starting cluster services on the primary control workstation	214
3-38	Setting cluster services on control workstation to start automatically .	217
3-39	Stop cluster services with takeover	218
4-1	General HAGEO design.	226
4-2	Two nodes at each site	233
4-3	Two nodes at the primary site and one node at the secondary site. . .	234
4-4	One node at each site	235
4-5	Define cluster nodes	237
4-6	SMIT HACMP - Add Site	237
4-7	Configure GeoPrimary network	238
4-8	Creating geographic resource groups	240
4-9	GMD replicated resources	241
4-10	SMIT HAGEO menu	242
4-11	SMIT HAGEO - Configure Global GeoMirror Properties	243
4-12	SMIT HAGEO - Temporal Ordering Policy	244
4-13	Our basic existing cluster.	246

4-14	Sample cluster with a remote site	247
4-15	Adding a remote node	247
4-16	Add an IP-based Network	248
4-17	Add an additional network adapter	248
4-18	Add an boot adapter for IPAT between sites	249
4-19	Add a Non IP-based Adapter	249
4-20	Adding a local site	250
4-21	Adding a remote site	250
4-22	Add a GeoMirror Device	252
4-23	Synchronize GeoMirror Devices	252
4-24	GeoMessage Utilities	253
4-25	Start ALL GeoMirror Devices	253
4-26	Another resource group configuration	254
4-27	Add a remote node to our resource group	255
4-28	Add GMD replicated resources	255
4-29	Resource group definition on previous HAGEO configuration	259

Tables

1-1	clavan.log variable portion	31
2-1	p690 hardware resources	73
2-2	p690 drawer configuration	74
2-3	LPAR configuration	75
2-4	IP addresses for node C37_P01 - netmask 255.255.255.224	76
2-5	IP addresses for node C38_P01 - netmask 255.255.255.224	76
2-6	Volume groups and file systems	77
2-7	SSA disk configuration	83
2-8	Cluster definition - scenario 1	89
2-9	Cluster Topology definition - scenario 1	89
2-10	Cluster Resources definition - scenario 1	90
2-11	Cluster Topology definition - scenario 4	131
2-12	Cluster Resources definition - scenario 4	131
3-1	IP labels	173
4-1	Summary of HAGEO releases	224

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described 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:

IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

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

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.


This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks

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

AIX®	NetView®	SP™
AIX 5L™	PAL®	SP2®
Enterprise Storage Server™	Perform™	Tivoli®
IBM®	pSeries™	VisualAge®
IBM.COM™	Redbooks™	zSeries™
IBM eServer™	Redbooks (logo)™ 	
MORE™	RS/6000®	

The following terms are trademarks of other companies:

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

C-bus is a trademark of Corollary, Inc. in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

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.

Preface

The objective of this IBM Redbook is to provide how-to technical information about configuring highly available clusters using HACMP Version 4.5, with a special focus on IBM @server pSeries 690 model 681 servers. It describes, in detail, the installation, customization, and configuration procedures of the new LPAR supported servers with HACMP for high availability. As an case study application for HACMP 4.5, this redbook investigates and explains the procedures for configuring an Highly Available Control Workstation (HACWS) in an Cluster 1600 configuration. HAGEO has been integrated to exploit the HACMP 4.5 features, and a special focus on HAGEO 2.4 is described in Chapter 4 of this redbook. Prior knowledge of HACMP will be useful for easy understanding of this document. This document provides various example scenarios and configurations and demonstrates high availability clustering using HACMP 4.5.

Chapter 2 of this redbook is also published as a separate Redpaper titled *Configuring Highly Available p690 Clusters using HACMP 4.5*, REDP0218.

Chapter 3 of this redbook is also published as a separate Redpaper titled *HACWS: An HACMP Application for IBM @server Cluster 1600*, REDP0303.

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, Poughkeepsie Center.

Adrian Demeter is a senior IT consultant and technical manager at IBM Business Partner GC System a.s., Prague, Czech Republic. He holds the Ing. academic degree in Electrical engineering and technical cybernetics from Czech Technical University (CVUT), Prague. He is responsible for project design, implementation, and support of high-end computer systems in Europe. He also teaches IBM courses on HACMP and Cluster 1600. He has 18 years experience in electronics and computers, of which 10 years is in UNIX.

Antony Steel (Red) is a Senior I/T Specialist in ITS Australia. He has 10 years experience in the UNIX field. predominately AIX and Linux. He holds an honors degree in Theoretical Chemistry from the University of Sydney. His areas of expertise include scripting, system customization, performance, networking, and

high availability. He has written and presented on LVM, TCP/IP, and high availability in Australia and AP. This is his second redbook.

Hiroyuki Fukuma is an I/T Specialist at the IBM Japan Systems Engineering Co., Ltd. in Makuhari, Japan (ISE). He has worked at IBM for five years, and currently works in the HACMP support team. He has taught in classes for HACMP for the update and/or intensive courses and currently supported Linux-HA, system as well as HACMP, at the Server Technology Group in ISE.

Jacek Kwecko is a Advisory I/T Specialist in Poland. He has 10 years of experience in the computer technology field. He worked at IBM for five years. He holds a MS degree in Electronics Engineering from the Technical University of Wroclaw. His areas of expertise include operating systems (AIX and Linux), high availability (HACMP and Cluster 1600) and backup solutions (Tivoli Storage Manager and Veritas NetBackup).

Tony Steel is a Project Leader at the ITSO Poughkeepsie Center. He has several years of experience working with RS/6000 SP and IBM @server Cluster 1600.

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

International Technical Support Organization, Poughkeepsie Center

Dave Bennin, Peter Bertolozzi, Margarita Hunt, Al Schwab

The project team sincerely appreciates, and extends a special thank-you to, the following people in the Poughkeepsie UNIX development Lab. Their complete support and cooperation was invaluable during the development of this document:

Chris Algozzine, Patrick A Buah, Michael K Coffey, Elaine Krakower, Christopher DeRobertis, Stephen J Tovcimak, Paul Moyer, Ganesan Narayanasamy

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our 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:

IBM Corporation, International Technical Support Organization
Dept. JN9B Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition may also include minor corrections and editorial changes that are not identified.


Summary of Changes
for SG24-6845-01
for Configuring Highly Available Clusters Using HACMP 4.5
as created or updated on November 11, 2002.

October 2002, Second Edition

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

- ▶ Chapter 4, “HAGEO integration with HACMP cluster” on page 223 discusses the IBM High Availability Geographic Cluster for AIX (HAGEO) software product and how it works with the IBM High Availability Cluster Multi-Processing (HACMP) licensed program product to provide automatic detection, notification, and recovery of an entire geographic site from failures. It is an entirely new chapter.



HACMP 4.5 overview

This chapter will outline the new features of High Availability Cluster Mult-Processing (HACMP) 4.5. It assumes that the reader has a passing familiarity with HACMP. For more details on HACMP concepts and operation, please see the references in “How to get IBM Redbooks” on page 276.

In this chapter, we:

- ▶ Define common terms.
- ▶ Examine the requirements and prerequisites.
- ▶ Look at each new feature in detail, giving examples for most.
- ▶ Discuss installation and migration considerations.

For further information about High Availability concepts, refer to Chapter 1 of the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306.

1.1 Introduction to HACMP

High Availability Cluster Multi-Processing (HACMP) is IBM's software for building highly available clusters on IBM Scalable POWERParallel systems and/or a combination of pSeries systems. It is supported by a wide range of IBM @server pSeries systems, including the new p690, storage systems, and network types. HACMP builds on the inherent reliability of the hardware to provide greater uptime for applications and enables upgrades and reconfiguration without interrupting operations.

The following terms will be used in this book

HAS	IBM High Availability Cluster Multi-Processing Classic for AIX.
HACMP/ES	IBM High Availability Cluster Multi-Processing for AIX Enhanced Scalability (sometimes called HA/ES).
Node Name	This is the name that HACMP uses to refer to the node; it does not have to be the same as the host name or the IP label that points to the host name. Note: This is an arbitrary string of 30 characters (alphanumeric and underscore, but cannot start with a number).
Application	A user application that is started, stopped, and monitored by HACMP. Also called an Application Server.
Fallover	Used to describe the user initiated process of moving resources from one node to another.
Failover	Used to describe the process when resources are moved from a node due to an error condition.
Resource Group	A resource group is a collection of often interdependent resources that will be controlled by HACMP as one unit. Can consist of network addresses, disks, volume groups, file systems, NFS resources, tape devices, applications, and priority policies
RSCT	The IBM RS/6000 Cluster Technology (RSCT) high availability services provide greater scalability, notify distributed subsystems of software failure, and coordinate recovery and synchronization among all subsystems in the software stack.
Topology Services	This RSCT facility generates heartbeats over multiple networks and provides information about adapter membership, node membership, and routing.

Topology

More than two nodes, network adapters on these nodes, and networks in which network adapters on these nodes exist are configured as part of the HACMP/ES cluster topology.

1.2 Requirements and prerequisites

For HACMP to work as designed, it is important that the following requirements be met:

- ▶ Supported hardware
- ▶ Required software levels

1.2.1 Supported hardware

HACMP 4.5 now includes support for IBM *@server* pSeries 690, both as a stand-alone server and also part of a IBM *@server* Cluster 1600. For a complete list of supported hardware, refer to the *HACMP for AIX 4.5 Installation Guide*, SC23-4278 or the relevant hardware manual for the latest information. IBM marketing in your country will be able confirm HACMP support for your hardware configuration.

1.2.2 Required software levels

HACMP 4.5 requires AIX 5L Version 5.1 and RSCT Version 2.2.1.0. All the work in this book is carried out using AIX 5L Version 5.1 Maintenance Level 02 and HACMP 4.5 PTF 1.

Note: The RSCT fileset supplied with AIX 5L Version 5.1 should be used.

The following software are also required:

- ▶ bos.adt.libm 5.1.0.0
- ▶ bos.adt.syscalls 5.1.0.0
- ▶ bos.data 5.1.0.0
- ▶ rsct.compat.basic.hacmp 2.2.1.0
- ▶ rsct.compat.clients.hacmp 2.2.1.0
- ▶ vacpp.msg.en_US.ioc.rte
- ▶ vacpp.ioc.aix50.rte

RSCT also requires:

- ▶ csm.client 1.1.0.25
- ▶ devices.chrp.base.ServiceRM 1.1.0.25

Before installing HACMP 4.5, the user is explicitly required to accept the license agreement in order for the product to be successfully installed. (See Figure 1-1).

Install Software

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

[Entry Fields]

* INPUT device / directory for software

.

* SOFTWARE to install

[_all_latest]

+

PREVIEW only? (install operation will NOT occur)

no

+

COMMIT software updates?

yes

+

SAVE replaced files?

no

+

AUTOMATICALLY install requisite software?

yes

+

EXTEND file systems if space needed?

yes

+

OVERWRITE same or newer versions?

no

+

VERIFY install and check file sizes?

no

+

Include corresponding LANGUAGE filesets?

yes

+

DETAILED output?

no

+

Process multiple volumes?

yes

+

ACCEPT new license agreements?

yes

+

Preview new LICENSE agreements?

no

+

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 1-1 New option to accept license agreements

1.3 New features and functions

HACMP 4.5 is the latest development of IBM's High Availability product line. A number of the new features have been included to improve HACMP's user interface and ease of administration. Further changes have been made to extend the functionality of HACMP to include new hardware and network options. We have broken these down as follows:

- ▶ Usability enhancements
- ▶ Administrative enhancements
- ▶ Network enhancements
- ▶ Device support
- ▶ Application support

1.3.1 Usability enhancements

The following enhancements extend the usability of both HACMP and HACMP/ES:

- ▶ Online planning worksheets
- ▶ Enhanced customer pager notification support
- ▶ Enhanced Concurrent Mode

The following apply only to HACMP/ES:

- ▶ 64-bit clinfo API.
- ▶ Dynamic node priority policies
- ▶ Resource group parallel processing (temporal ordering)
- ▶ Selective fallover triggered by volume group loss

Online Planning Worksheets

The Online Planning Worksheets (OLPW) can now be run as a stand-alone Java application and is currently supported on both AIX and Windows.

The following steps are all that are required to complete the planning worksheet, populate the HACMP configuration database, and perform a cluster verification and synchronization:

1. Copy the worksheet files to your PC or IBM @server pSeries server.
2. Execute the startup program (worksheets.bat for Windows or worksheets for AIX).
3. Enter your configuration data into the OLPW.
4. FTP your configuration data to one of the nodes.
5. Execute the `c1_opsconfig` command.

Prerequisites for the Online Planning Worksheet program client

To use the worksheets, you need Java 2 Runtime Environment (J2RE) Version 1.2 or higher as follows:

AIX 5L Version 5.1 or higher

AIX installs the J2RE by default.

MS Windows 95/98, NT, and 2000

If you run the program on MS Windows, check to see that the appropriate J2RE is installed.

Installing and running the OLWP on AIX and Windows

When you install the HACMP/ES software, you will find the OLWP files in `/usr/es/sbin/cluster/samples/worksheets` directory.

To install the OLPW, do the following procedures:

- | | |
|----------------|---|
| AIX | Requires worksheets (start script) and worksheets.jar (Java Archive file). Transfer these files to an appropriate directory on the RS/6000 client. If you move them from the default location of /usr/lpp/cluster/samples/worksheets, then you will need to modify the WORKSHEETS variable in the worksheets script file, using the full path of the new location of the Java archive file. |
| Windows | Requires worksheets.bat (start batch file) and worksheets.jar (Java Archive file). Transfer these files to an appropriate directory on the PC (if using FTP, remember to specify binary mode). If worksheets.bat and worksheets.jar are stored in different directories, then the path in worksheets.bat needs to be edited. |

There is also a readme - worksheets.html

Run the application by executing the above script/batch file, and you will see the worksheet panel shown in Figure 1-2 on page 7. Your HACMP configuration can now be created, following a similar logic to the planning worksheets.

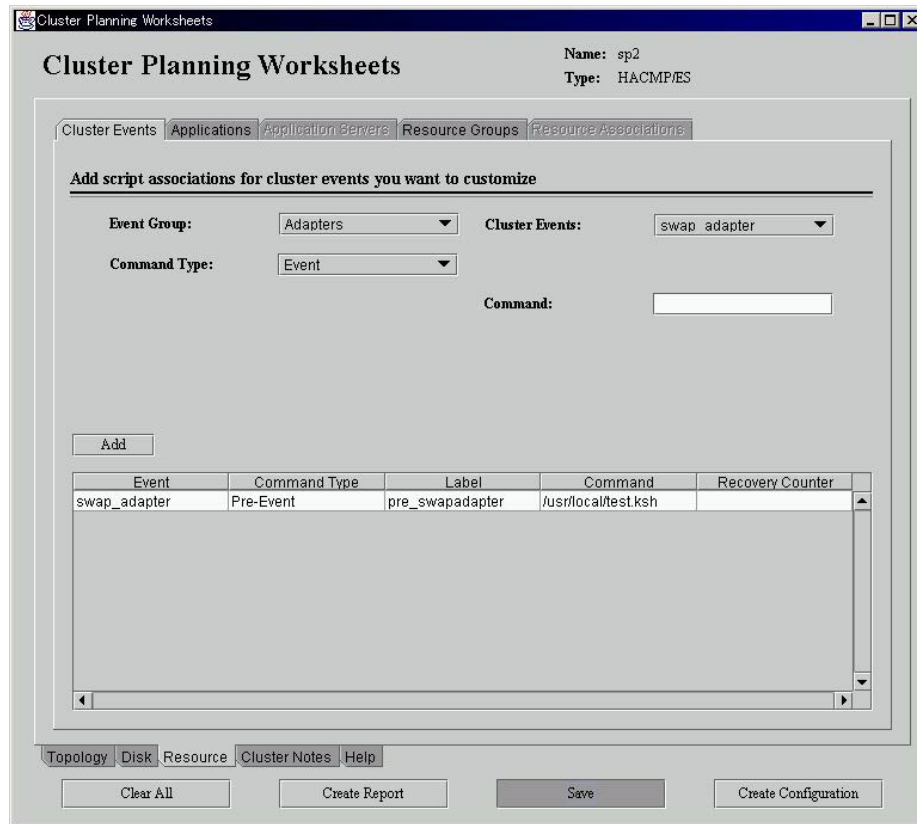


Figure 1-2 Online Planning Worksheet

For more information on entering cluster configuration data, see Appendix B of the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306.

Applying worksheet data to AIX cluster nodes

Once the configuration data has been entered, you need to create a file containing your cluster configuration data, which can be transferred back to one of the nodes. Press the Create Configuration button, located on the base panel, and save your configuration file (the default name for which is cluster.conf). Now use FTP to transfer the configuration data file to one of the AIX cluster nodes (Remember to use binary mode if transferring from Windows).

The HACMP script `cl_opsconfig` is then used to load the configuration data into the HACMP database. It is used as follows:

```
/usr/sbin/cluster/utilities/cl_opsconfig <your_config_file>
```

After loading your configuration, the **cl_opsconfig** command automatically performs a cluster synchronization, including a cluster verification, of the configuration.

Enhanced customer pager notification support

The ability for users to test their pager notification configuration has been added. Once a pager notification method has been configured, it can be tested by attempting to send a test page from the node which has been configured for the selected notification method.

Pager notification requires at least one node in the cluster with a free serial port, which can be connected to a hayes compatible modem. This node and tty combination is used to create what is called a node/port pair. If more than one is created, then pager notification methods can be configured to use one or more of these node/port pairs. On sending a page, this method will then work through the node names left to right.

Pager notification methods are created for particular cluster events, or combinations of events, using user defined message files.

Note: Only one pager notification method can be defined for each event.

Pager notification methods are reached from the SMIT HACMP Screen by selecting **Cluster Configuration -> Cluster Custom Modification -> Define Custom Pager Notification Method**, or by running **smitty hacmp** and selecting **RAS Support -> Define Custom Pager Notification Method** (see Figure 1-3 on page 9).

Add a Custom Pager Notification Method

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

	[Entry Fields]
* Method name	[Op_page]
Description	[Page Operator]
* Nodename(s)	[sp2n3] +
* Number to dial	[1800itsbroke]
* Filename	[/usr/es/sbin/cluster/s>
* Cluster event(s)	node_down_complete +
Retry counter	[3] #
TIMEOUT	[10] #

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 1-3 Creating a custom pager notification method

Once the pager notification method has been created, it can be tested from the SMIT Send a Test Page screen (see Figure 1-4).

Send a Test Page

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

	[Entry Fields]
Method name	Op_page
* EVENTNAME	+

EVENTNAME

Move cursor to desired item and press Enter.

node_down

F1=Help

F2=Refresh

F3=Cancel

F1 F8=Image

F10=Exit

Enter=Do

F5 /|=Find

n=Find Next

F9+

Figure 1-4 Send a Test Page

Once the Pager notification method and event name have been selected, SMIT runs the `/usr/es/sbin/cluster/utilities/clissuepage` command. This sends the message file defined for the given event, using the first available defined Node/Port pair. Progress messages are displayed (see Figure 1-5).

```

                                COMMAND STATUS

Command: failed          stdout: yes          stderr: no

Before command completion, additional instructions may appear below.

clissuepage: --->AT<0D>
clissuepage: --->AT<0D>
clissuepage: --->AT<0D>
clissuepage: Modem is not answering
clissuepage: sp2n3: Message was NOT sent, event - node_down
clissuepage: clissuepage exited with the return code = -1

F1=Help          F2=Refresh          F3=Cancel          F6=Command
F8=Image         F9=Shell           F10=Exit          /=Find
n=Find Next

```

Figure 1-5 Example test page status (retry count was 3)

Enhanced Concurrent Mode

Enhanced Concurrent Mode is an integral part of AIX 5L Version 5.1 and above. HACMP 4.5 with AIX 5L Version 5.1 extends the support for concurrent volume groups on all supported disk devices.

64-bit clinfo API

Cluster Information Program (clinfo) is the SNMP monitoring program, which runs on a client machine or on a cluster node. It queries the clsmuxpd daemon for up-to-date cluster information and then provides a simple display to the user. It shows information regarding the state of the HACMP cluster, nodes, and networks. The clinfo cluster.es.client.lib library now contains the libcl.a with both 32- and 64-bit objects.

The application must be compiled in a 64-bit environment for it to use this new API.

Dynamic node priority policies

It is possible for the cluster manager to determine a fallover node for a resource group, based on either a pre-defined or user defined policy. This feature is

controlled through the Change/Show Resources/Attributes for a Resource group SMIT menu (see Figure 1-6).

Change/Show Resources/Attributes for a Resource Group

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

[TOP]

Resource Group Name

Node Relationship

Site Relationship

Participating Node Names / Default Node Priority

Dynamic Node Priority

[Entry Fields]

sp2n123casc1

cascading

ignore

sp2n1 sp2n2 sp2n3

[c1_lowest_disk_busy] +

[BOTTOM]

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 1-6 Dynamic node priority

User defined dynamic node priority policies can be created, removed, and modified from the Configure Dynamic Node Priority Policies menu under Cluster Resources (see Figure 1-7 on page 12). HACMP 4.5 is supplied with three pre-configured policies, but also allows user defined policies to be created using the same or different RSCT variables. See the detailed description of the variables in the *RSCT: Event Management Programming Guide and Reference*, SA22-7354.

Add a Dynamic Node Priority Policy

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

[Entry Fields]

* Dynamic Node Priority Policy Name [avail_proc]
 Dynamic Node Priority Policy Description [Available processor time]

* Resource Variable +
 * Condition largest +

Resource Variable

Move cursor to desired item and press Enter.

IBM.PSSP.aixos.PagSp.totalfree	Available virtual memory paging space
IBM.PSSP.aixos.Disk.busy	Fraction of time disks are busy
IBM.PSSP.aixos.CPU.gldle	Available processor time

F1=Help
F2=Refresh
F3=Cancel

F1 F8=Image
F10=Exit
Enter=Do

F5 /=Find
n=Find Next

F9+

Figure 1-7 User defined Dynamic node priority policy

The three pre-defined policies are:

- cl_highest_free_mem** A node priority policy based on the highest percentage of free memory.
- cl_highest_idle_cpu** A node priority policy based on the node with the most idle time.
- cl_lowest_disk_busy** A node priority policy based on the disk that is least busy.

For resource groups where the dynamic node priority policy is enabled, the order of nodes on the takeover list is determined at the time of the actual event. The cluster manager only uses this facility for failures (node_down and rg_move events). For node_up and reconfig_resource events, the cluster manager uses the default node priority list.

The selection of policy for a resource group is independent of policies chosen for other resource groups, but the following items should be considered when implementing dynamic node priority policies:

- ▶ It works best for clusters of similarly configured nodes.
- ▶ There must be at least three nodes in the node list for the resource group.
- ▶ This does not apply to concurrent resource groups.

- ▶ A resource group will not fallover to a node that, at the time, has insufficient physical resources (for example, network adapters).
- ▶ Dynamic Reconfiguration (DARE) events will not follow the dynamic node priority policy, as the cluster manager only uses the resource variable calculation when it handles failure events
- ▶ If two nodes return the same value for the particular resource variable, the node with the higher priority in the node list will be selected.
- ▶ Resource groups will still initially reside on the first node listed in the node priority list.

Logs

There are a couple of entries in `clstrmgr.debug` that are related to the dynamic node priority policy. The first entry is added when the policy is added to the resource group and the cluster synchronized. At this point, a new resource is added called the `NODE_PRIORITY_POLICY`, and it is assigned the particular policy.

The other entry is when the actual takeover happens. The `clstrmgr.debug` file will show where the cluster manager determined which node should takeover the resource group. Example 1-1 shows the result of the cluster manager determining the priority of the nodes in its list, then confirming that sufficient physical resources are available.

Example 1-1 Cluster manager determines node priority policy

```
Sat Jun  8 15:27:48 Attempting to calculate node priority...
Sat Jun  8 15:27:48 Attempting to calculate node order.
Sat Jun  8 15:27:48 Node list order already known.
Sat Jun  8 15:27:48 Completed node order calculation.
Sat Jun  8 15:27:48 For Resource Group sp2n123cascl, BestNode got node order
Sat Jun  8 15:27:48 Got the following 2 node IDs:
Sat Jun  8 15:27:48 3 2
Sat Jun  8 15:27:48 The best node for group sp2n123cascl is sp2n3.
Sat Jun  8 15:27:48 RGPA got sp2n3 as highest priority node.
Sat Jun  8 15:27:48 HACMPnetwork alias stanza for [sp2n123ether] is [0]
Sat Jun  8 15:27:48 Testing possible allocation of [sp2n123cascl] on [sp2n3]
Sat Jun  8 15:27:48 Status of [sp2-n1-svc] on [sp2n3] is [0]
Sat Jun  8 15:27:48   Testing possible allocation of [sp2-n1-svc] on
[sp2-n3-boot]
Sat Jun  8 15:27:48       [sp2-n3-boot] is not up
Sat Jun  8 15:27:48   Testing possible allocation of [sp2-n1-svc] on
[sp2-n3-stby]
Sat Jun  8 15:27:48       Testing if [sp2-n1-svc] can be placed over
[sp2-n3-stby]
Sat Jun  8 15:27:48           Node [sp2n3] is a member of group [sp2n123cascl]
Sat Jun  8 15:27:48           Home node is [sp2n1]
```

Sat Jun 8 15:27:48
Sat Jun 8 15:27:48

OK: Node is not home, adapter is standby
and group supports IPAT on standbys

Resource group parallel processing (temporal ordering)

In order to give administrators greater control over the order in which multiple resource groups will be processed during a single event, HACMP now allows users to explicitly specify the order in which resource groups should be processed during acquisition and release.

Prior to this release, when multiple resource groups were to be processed for acquisition or release, they were handled consecutively in alphabetical order. Resource groups that required only the mounting or unmounting of NFS were processed in alphabetical order after all other resource groups.

Thus, the only way to control the order in which resource groups were processed, is to name them appropriately. The only exception to this behavior was that the following resource types were processed in parallel *within* a particular resource group to improve the overall efficiency of resource group acquisition and release:

- ▶ Parallel disk types
- ▶ Volume groups
- ▶ File systems for which the Resource Group Recovery Method selected in SMIT is parallel
- ▶ NFS mounts
- ▶ Application server start scripts

In this new version, users can now explicitly specify the order in which resource groups are acquired and released. All resource groups that are not so defined with an explicit acquisition or release order will still be processed in the default order.

Enhancements in HACMP Classic features

The default order for the processing of resource groups has not changed. Thus, resource groups that have not been specified for custom acquisition or release ordering will be processed consecutively in alphabetical order after the customized groups.

Resource group acquisition will occur in the following order:

1. Specified resource groups are processed serially.
2. Resource groups that are not specified will be processed consecutively in alphabetical order.

3. Specified resource groups only mounting NFS are processed serially.
4. Resource groups only mounting NFS are processed consecutively in alphabetical order.

Resource group release will occur in the following order:

1. Resource groups that are not specified will be released consecutively in alphabetical order.
2. Specified resource groups are released serially.
3. Resource groups only mounting NFS are released consecutively in alphabetical order.
4. Specified resource groups only mounting NFS are released serially.

HAES

The default order for HACMP/ES Version 4.5 is now parallel. However, this parallel processing of the resource groups does not take the form of multiple parallel threads of execution. Rather, those resource groups that are to be processed in parallel are treated as if they are consolidated into one resource group. Within this consolidated group, individual resource types are still processed consecutively. The greatest benefit of this design is exhibited when this consolidated group contains resources that are processed in parallel (for example, volume groups in the above list).

Parallel resource processing is managed by a new event script called `process_resources`. This script plays the same role for parallel processed resource groups that the `node_[up|down]_[local|remote]` scripts play for serially processed resource groups.

Output to `hacmp.out` has been enhanced to allow users to more easily isolate details related to a specific resource group and its resources. This change affects both parallel and serial resource groups.

Users modify the default behavior by specifying a serial acquisition list of resource groups, which may be none, some, or all of the defined resource groups. Resource group acquisition will occur in the following order:

1. Specified resource groups are processed serially.
2. Resource groups only mounting NFS are processed in parallel.
3. Resource groups not specified will be processed in parallel.

Resource group release will occur in the following order:

1. Resource groups not specified will be processed in parallel.
2. Specified resource groups are processed serially.
3. Resource groups only mounting NFS are processed in parallel.

The following restrictions apply to the parallel processing of resource groups:

- ▶ Parallel resource group processing does not occur during DARE configuration changes or DARE resource group migration.
- ▶ Parallel resource group processing does not occur during rg_move events
- ▶ If an error occurs during the acquisition or release of a resource group, recovery procedures will continue to be run after *all* other resource group processing has completed.
- ▶ An explicit ordering must be consistent across the cluster. There is no option for one node to process resource groups according to a different ordering than any other node.
- ▶ This control over the ordering applies only to the local node and does not apply for concurrent resource groups running on different nodes. It is thus possible for a resource group on one node to be configured before another resource group on the first node.
- ▶ If a resource group that has been included on the explicit ordering list is removed from the cluster, then the name of that resource group will be automatically removed from the list. If a resource group's name is changed, the resource group processing order list will be updated appropriately.

Migration issues

When migrating to HACMP 4.5, there is no issue, as the default order is alphabetical and consecutive for this version and earlier versions

When migrating to HACMP/ES 4.5, all the resources must be processed consecutively in alphabetical order and this order needs to be specified, or the resource groups will be processed in parallel.

Example (HACMP/ES)

If you have defined five resource groups (rg1, rg2, rg3, rg4, and rg5), and all five of the resource groups are being acquired by the same node, then, by default, they will be processed in parallel.

However, if the resource group rg4 must be brought up before rg2, and if resource group rg2 must be brought up before the remaining resource groups, then this can be done as shown in Figure 1-8 on page 17.

Change / Show Resource Group Processing Order

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

[Entry Fields]

Resource Groups Acquired in ParallelSerial Acquisition OrderNew Serial Acquisition Order

rg1 rg2 rg3 rg4 rg5[rg4 rg2]

+

Resource Groups Released in ParallelSerial Release OrderNew Serial Release Order

rg1 rg2 rg3 rg4 rg5[rg4 rg2]

+

F1=HelpF5=ResetF9=Shell

F2=RefreshF6=CommandF10=Exit

F3=CancelF7=EditEnter=Do

F4=ListF8=Image

Figure 1-8 Change/Show Resource Group Processing Order

This screen shows that rg4 and rg2 will be processed serially, and when finished, resource groups rg1, rg3, and rg4 will be consolidated and processed together.

After completing this screen, a detailed display of the resource group acquisition and release order is given (see Figure 1-9).

COMMAND STATUS

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

[TOP]

Acquisition Order:

Resource Group Serial Acquisition Order (2 RGs):
rg4 rg2

Resource Groups Acquired in Parallel After Serial Processing (3 RGs):
rg1 rg3 rg5

Release Order:

Resource Groups Released in Parallel Before Serial Processing (3 RGs):
rg1 rg3 rg5

Resource Group Serial Release Order (2 RGs):
rg4 rg2

Figure 1-9 Acquisition and release order

Changes to hacmp.out

The content of hacmp.out has been enhanced in order to make it easier to follow the flow of parallel resource group processing.

The event summary below shows the individual processing for the resource groups rg4 and rg2, then the processing of the consolidated resources of rg1, rg3, and rg5 (see Example 1-2).

Example 1-2 Event summary showing parallel processing

HACMP Event Summary		
Event: node_up sp2n3		
Start time: Sat Jun 8 18:10:20 2002		
End time: Sat Jun 8 18:12:09 2002		
Action:	Resource:	Script Name:

Acquiring resource group:	rg4	node_up_local
Search on:	Sat.Jun.8.18:10:23.EDT.2002.node_up_local.rg4.ref	
Acquiring resource:	192.168.6.42	cl_swap_IP_address
....		
Acquiring resource:	hdisk10	cl_disk_available
Search on:	Sat.Jun.8.18:11:13.EDT.2002.cl_disk_available.hdisk10.rg4.ref	
....		
Acquiring resource:	redvg	cl_activate_vgs
Search on:	Sat.Jun.8.18:11:18.EDT.2002.cl_activate_vgs.redvg.rg4.ref	
Resource online:	redvg	cl_activate_vgs
Search on:	Sat.Jun.8.18:11:36.EDT.2002.cl_activate_vgs.redvg.rg4.ref	
Acquiring resource:	/test/appl	cl_activate_fs
Search on:	Sat.Jun.8.18:11:41.EDT.2002.cl_activate_fs..test.appl.rg4.ref	
....		
Acquiring resource group:	rg2	node_up_local
Search on:	Sat.Jun.8.18:10:45.EDT.2002.node_up_local.rg2.ref	
Acquiring resource:	192.168.6.43	cl_swap_IP_address
....		
Acquiring resource group:	rg1	process_resources
Search on:	Sat.Jun.8.18:11:44.EDT.2002.process_resources.rg1.ref	
Acquiring resource group:	rg3	process_resources
Search on:	Sat.Jun.8.18:11:45.EDT.2002.process_resources.rg3.ref	
Acquiring resource group:	rg5	process_resources
Search on:	Sat.Jun.8.18:11:46.EDT.2002.process_resources.rg5.ref	
Acquiring resource:	hdisk9	cl_disk_available
Search on:	Sat.Jun.8.18:11:48.EDT.2002.cl_disk_available.hdisk9.rg1.ref	
Acquiring resource:	hdisk12	cl_disk_available
Search on:	Sat.Jun.8.18:11:48.EDT.2002.cl_disk_available.hdisk12.rg3.ref	
....		
Resource online:	greenvg	cl_activate_vgs
Search on:	Sat.Jun.8.18:11:58.EDT.2002.cl_activate_vgs.greenvg.rg3.ref	
Resource online:	bluevg	cl_activate_vgs
Search on:	Sat.Jun.8.18:11:59.EDT.2002.cl_activate_vgs.bluevg.rg1.ref	
Acquiring resource:	/test7	cl_activate_fs
Search on:	Sat.Jun.8.18:12:04.EDT.2002.cl_activate_fs..test7.rg1.ref	
Resource online:	/test7	cl_activate_fs
Search on:	Sat.Jun.8.18:12:04.EDT.2002.cl_activate_fs..test7.rg1.ref	

....

The file hacmp.out also shows the JOB_TYPE equal to ACQUIRE at the beginning of an acquisition event for the processing of the consolidated resource groups. This variable is set to NONE for groups that are being processed serially. The variable RESOURCE_GROUPS will list the resource groups that are being acquired in parallel during this event. This is shown in Example 1-3

Example 1-3 New JOB_TYPE for parallel processing

```
:process_resources[1482] JOB_TYPE=NONE
:process_resources[1484] RC=0
:process_resources[1485] set +a
:process_resources[1487] [ 0 -ne 0 ]
:process_resources[1727] break
:process_resources[1737] exit 0
:node_up[367] set -a
:node_up[368] clsetenvres rg1 node_up
:node_up[368] eval PRINCIPAL_ACTION="ACQUIRE" ASSOCIATE_ACTION="NONE"
AUXILLIARY_ACTION="NONE" VG_RR_ACTION="ACQUIRE" SIBLING_NODES= FOLLOWER_ACTION="NONE"
NFS_HOST= DISK= CONCURRENT_VOLUME_GROUP= EXPORT_FILESYSTEM= AIX_CONNECTIONS_SERVICES=
AIX_FAST_CONNECT_SERVICES= SNA_CONNECTIONS= COMMUNICATION_LINKS= SHARED_TAPE_RESOURCES=
MOUNT_FILESYSTEM= TAKEOVER_LABEL= NFSMOUNT_LABEL= MISC_DATA= NFS_NETWORK=
SHARED_TAPE_RESOURCES= PPRC_REP_RESOURCE= GMD_REP_RESOURCE= APPLICATIONS="app1"
CASCADE_WO_FALLBACK="false" FILESYSTEM="/test/app1" FSCHECK_TOOL="fsck"
FS_BEFORE_IPADDR="false" INACTIVE_TAKEOVER="false" RECOVERY_METHOD="sequential"
SERVICE_LABEL="sp2-n3-svc" SSA_DISK_FENCING="false" VG_AUTO_IMPORT="false"
VOLUME_GROUP="redvg"
:node_up[368] PRINCIPAL_ACTION=ACQUIRE ASSOCIATE_ACTION=NONE AUXILLIARY_ACTION=NONE
VG_RR_ACTION=ACQUIRE SIBLING_NODES= FOLLOWER_ACTION=NONE NFS_HOST= DISK=
CONCURRENT_VOLUME_GROUP= EXPORT_FILESYSTEM= AIX_CONNECTIONS_SERVICES=
AIX_FAST_CONNECT_SERVICES= SNA_CONNECTIONS= COMMUNICATION_LINKS= SHARED_TAPE_RESOURCES=
MOUNT_FILESYSTEM= TAKEOVER_LABEL= NFSMOUNT_LABEL= MISC_DATA= NFS_NETWORK=
SHARED_TAPE_RESOURCES= PPRC_REP_RESOURCE= GMD_REP_RESOURCE= APPLICATIONS=app1
CASCADE_WO_FALLBACK=false FILESYSTEM=/test/app1 FSCHECK_TOOL=fsck FS_BEFORE_IPADDR=false
INACTIVE_TAKEOVER=false RECOVERY_METHOD=sequential SERVICE_LABEL=sp2-n3-svc
SSA_DISK_FENCING=false VG_AUTO_IMPORT=false VOLUME_GROUP=redvg
:node_up[369] set +a
:node_up[370] export GROUPNAME=rg1
rg1:node_up[375] [ sp2n3 = sp2n3 ]
rg1:node_up[377] clcallev node_up_local
....
:process_resources[1482] eval JOB_TYPE=ACQUIRE RESOURCE_GROUPS="rg1 rg3 rg5"
:process_resources[1482] JOB_TYPE=ACQUIRE RESOURCE_GROUPS=rg1 rg3 rg5
:process_resources[1484] RC=0
:process_resources[1485] set +a
:process_resources[1487] [ 0 -ne 0 ]
:process_resources[1693] set_resource_group_state ACQUIRING
rg1:process_resources[6] export GROUPNAME
rg1:process_resources[7] [ ACQUIRING != DOWN ]
rg1:process_resources[9] [ REAL = EMUL ]
```

```
rg1:process_resources[14] clchdaemons -d clstrmgr_scripts -t resource_locator  
-n sp2n3 -o rg1 -v ACQUIRING  
rg1:process_resources[15] [ 0 -ne 0 ]  
rg1:process_resources[26] [ ACQUIRING = ACQUIRING ]  
rg1:process_resources[28] cl_RMupdate acquiring rg1 process_resources  
Reference string: Sat.Jun.8.18:11:44.EDT.2002.process_resources.rg1.ref  
rg1:process_resources[29] continue  
rg3:process_resources[6] export GROUPNAME  
rg3:process_resources[7] [ ACQUIRING != DOWN ]  
rg3:process_resources[9] [ REAL = EMUL ]
```

Similarly, there is a JOB_TYPE of RELEASE for the release of the consolidated resource group.

Selective failover triggered by volume group loss

Greater granularity has been added to HACMP 4.5 by providing the facility for recovery of individual resource groups that are affected by the failure of particular volume group. Prior versions of HACMP had selective failover for:

- ▶ Network Adapter failures
- ▶ Local network failures
- ▶ Application failures
- ▶ X25 Communication link failure

In HACMP/ES 4.5, selective failover is also triggered when there is a loss of quorum (LVM_SA_QUORCLOSE). HACMP uses the AIX error notification facility to monitor volume groups for loss of quorum.

Note: If the AIX errdaemon is not running on a cluster node, HACMP has no means to detect the "loss of quorum" error in the AIX log file, and, therefore, cannot selectively move a resource group if it contains a failed volume group.

During cluster synchronization, when HACMP/ES synchronizes a resource group containing a shared volume group, an entry is also added into the errornotify ODM Class on the relevant nodes (See Figure 1-10 on page 21).

Verifying Configuration of Errnotify Stanzas

----- WARNING:

Error notification stanzas will be added
during synchronization for following:

Node : en_label : en_resource : en_class
sp2n3: : redvg : logical_volume : LVM_SA_QUORCLOSE
sp2n4: : redvg : logical_volume : LVM_SA_QUORCLOSE

Verification has completed normally.

Remember to redo automatic error notification if configuration has changed.
Updating ODM errnotify on node sp2n3.

Figure 1-10 Resource Group synchronization adding error notification entry

If the AIX error daemon detects a loss of quorum error (Error Label LVM_SA_QUORCLOSE), the cluster manager is updated. The cluster manager will then move the resource group that contains the affected volume group to another node in the cluster. The destination node will depend on the available nodes in the list of nodes for that resource group, and the cluster's resource group question policies, for example, if a dynamic node priority policy has been configured.

Details will be logged in both hacmp.out and clstrmgr.debug, and they include:

- ▶ AIX error label and ID
- ▶ The name of the affected resource group
- ▶ The node's name on which the error occurred

On the destination node, the resource group will continue to operate, as the varyon of the volume group will be forced to overcome the loss of quorum, and the Event Management daemon will not react until the next loss of quorum error.

If a volume group that is part of a concurrent resource group has failed on a node, HACMP/ES only brings the affected resource group offline on the affected node.

Prior to HACMP/ES 4.5, if you wanted the cluster manager to respond to a volume group failure, you could configure a customized error notification event for this error. This event might have caused a node_down event or moved the particular resource group to another node. If these methods already exist from a previous installation, they will still work with HACMP/ES 4.5, but may no longer be the best method for recovery.

Note: Users should not modify the error notification objects that are generated by HACMP for selective fallover. If the methods are modified, it is possible that they will not take the appropriate action.

For more information on how error notification works, see Chapter 13 of the *HACMP for AIX Enhanced Scalability Installation and Administration Guide*, SG23-4306.

1.3.2 Administrative enhancements

The following administrative enhancements have been added to both HACMP and HACMP/ES:

- ▶ User-specified time before warning
- ▶ Web interface to **clstat** command
- ▶ Enhanced Display of Event Summaries
- ▶ Troubleshooting

The following administrative enhancements have been added to HACMP/ES:

- ▶ Application availability analysis tool

User-specified time before warning

With HACMP 4.5, the user can now easily customize the amount of time between the start of an event and the calling of the `config_too_long` event. This is the event that appends the `config_too_long` warning message to the `hacmp.out` log every 30 seconds until the original event completes. By default, the cluster manager will allow an event to process for six minutes before it issues the `config_too_long` warning.

With previous versions of HACMP it was possible to modify this time delay by using the `chssys -s clstrmgr -a "-u delay_in_milliseconds"` command to change the default of six minutes.

HACMP 4.5 allows more precise control by dividing events into two classes:

- | | |
|--------------------|---|
| Fast events | Events that do not involve the acquisition or release of resources and therefore would normally take a shorter time to complete. |
| Slow events | Events that involve the acquisition or release of resources, the running of start or stop scripts for Application Servers, or site events for HAGEO. These events would normally take longer. |

By splitting events into these two groups, the time allowed for slower events can be customized without increasing the risk of too quickly detecting problems.

Similarly, you can avoid getting unnecessary warnings while keeping the `config_too_long` time at a reasonable time for most events.

These settings are modified through the SMIT menus **Cluster Configuration** -> **Advanced Performance Tuning Parameters** (see Figure 1-8 on page 17).

Change/Show Time Until Warning

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

[Entry Fields]

Event Duration (in seconds) [0]

Resource Group Duration (in seconds) [0]

NOTE: Changes made to this panel must be propagated to the other nodes by syncing Cluster Resources or Topology.

Figure 1-11 Change/Show Time Until Warning

The time allowed for “fast” events before the `config_too_long` event is modified through the Event Duration entry. The allowed time for “slow” events to run before the warning message is the sum of the Event Duration time and the Resource Group Duration time.

For clusters that have been upgraded to HACMP 4.5, the default for Event Duration is 360 and the default for Resource Group duration is 0. For new clusters, the default for Event Duration is also 360, but Resource Group Duration is 60.

There has also been a change to the frequency that the `config_too_long` messages are appended to the `hacmp.out` log. Prior to HACMP 4.5, these messages would appear every 30 seconds until the event completed. Now a throttle has been added to the time interval. It works as follows: the first five messages will appear every 30 seconds, then the time interval will double every five messages until it has been one hour since the `config_too_long` event started. The frequency will remain at once every hour until the `config_too_long` exits or is terminated on that node. Each time an event goes into `config_too_long`, the frequency is reset and the above throttle applied again.

A further change is that prior to HACMP 4.5, the cluster manager would send a SIGKILL (9) to the `config_too_long` event script, which could not be caught or ignored. Now the cluster manager uses a SIGQUIT (3), so `config_too_long` can trap the signal, print a message to the console and `hacmp.out`, then exit cleanly.

Web interface to clstat command

This release adds a Web interface, using CGI capability, to clstat. The clstat.cgi, when run on a properly configured Web server, can be used to view information about the cluster state on a remote HTML browser.

If this executable is called via a CGI capable Web server, it will return an HTML formatted version of the standard clstat display. This new display uses both text and colors to show the status of the cluster, nodes, and interfaces. It also makes it easier to monitor multiple clusters by providing information for multiple clusters on a single page with hyperlink access to each.

When you want to see your cluster status on a Web browser with this function, your machines must meet some requirements;

- ▶ To view the clstat display through a Web browser, you must have a Web server installed on a machine where clinfo is running and able to gather cluster information. This could be a client node or a server node.
- ▶ To configure a Web server on a cluster node, you need to move or copy clstat.cgi to the cgi-bin or script directory of the web server, for example, the default HTTP Server directory /usr/HTTPserver/cgi-bin.
- ▶ You can now view cluster status through a Web browser, as shown in Figure 1-12 on page 25, by typing in a URL in the following format:

`http://192.168.6.35/cgi-bin/clstat.cgi`

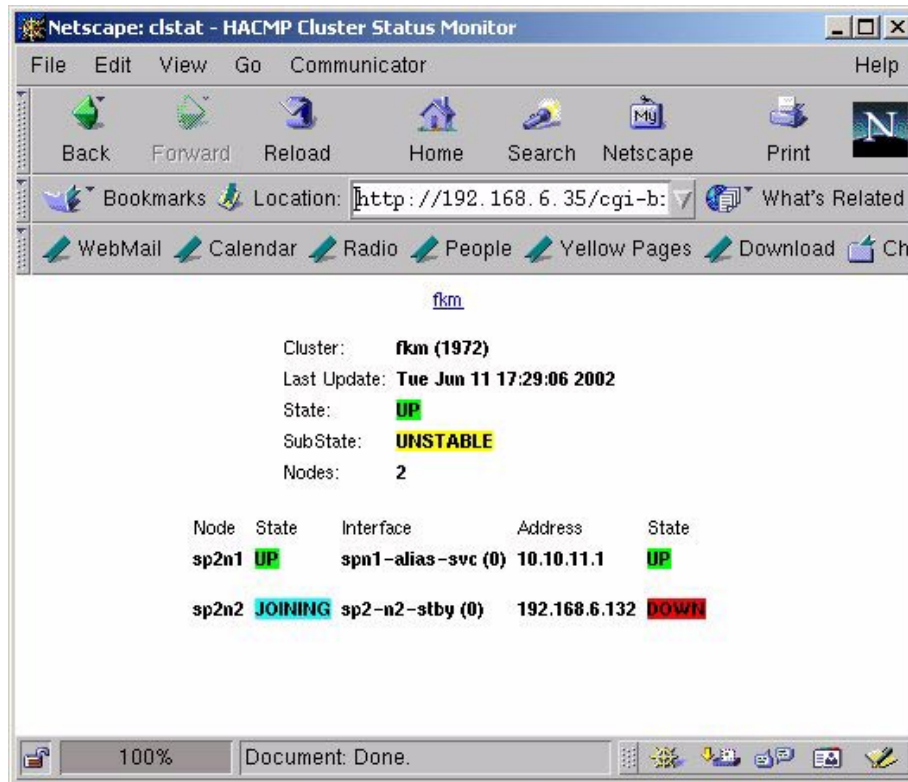


Figure 1-12 clstat with Web browser

Limitations and considerations

- ▶ To view the clstat display through a Web browser, the user must have a Web server installed on a machine where the clinfo daemon is running. This could be either an HACMP client node or an HACMP server node.
- ▶ The clstat.cgi program works with any Web sever that supports the CGI standard, which includes most currently-available Web servers for AIX. For example, a user might choose to use IBM HTTP Server, which is included on the Expansion Pack CD for AIX 5L.
- ▶ Full instructions for installing and configuring a Web server should be provided with the Web server software. Before attempting to use clstat.cgi for web access to cluster information, a user should first consult the user documentation for the Web server application and/or a local Web administrator.

- ▶ Because clstat.cgi is not run as root, there should be no immediate security threat of users gaining unauthorized access to HACMP/ES by accessing clstat.cgi from the Web server.
- ▶ Some administrators may wish to restrict access to clstat.cgi from the Web server, and can use methods built-in to the Web server to prevent access, such as password authentication or IP address blocking. HACMP does not provide any independent means of access restriction.
- ▶ This feature has been tested on the following browsers:
 - Netscape Navigator Version 6 for Windows and Version 4.75 for AIX.
 - Internet Explorer Versions 5.0, 5.5, and 6.0 for Windows.
 - Internet Explorer Version 5.5 has incomplete JavaScript support for automatic content refresh. For this reason, if a user has used a hyperlink to move to a particular part of the display, the refresh action in Internet Explorer Version 5.5 will cause a return to the top of the display.

Further information about using this feature can be found in Chapter 21, “Monitoring an HACMP/ES cluster”, in the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306 and Chapter 3, “Monitoring an HACMP cluster”, in the *HACMP for AIX 4.5 Administration Guide*, SC23-4279.

Enhanced display of event summaries

In HACMP 4.4.1, the cluster’s event summaries were added to hacmp.out, and file is renamed on a regular basis by the **clcycle** command. In HACMP 4.5, **clcycle** has been modified to append this cluster event summary information to a summary file called cl_event_summaries before hacmp.out is renamed. The default location of the cluster event summary file is located in /usr/es/sbin/cluster/etc.

A new SMIT panel called Display Event Summaries is provided as the interface to a new utility, called clevesummary, as shown in Figure 1-13.

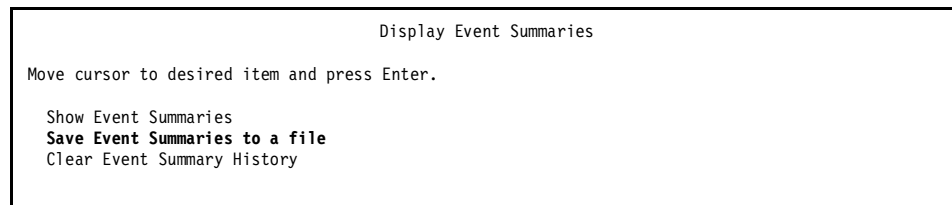


Figure 1-13 Display Event Summaries

This SMIT menu has three options:

- ▶ Show Event Summaries
- ▶ Save Event Summaries
- ▶ Clear Event Summary History

Both Show Event Summaries and Save Event summaries do the following:

- ▶ Gather the information from the clevsum.txt file
- ▶ Find the location of hacmp.out in the HACMP logs ODM file and extract any event summaries from hacmp.out
- ▶ For each resource group, run clfindres

then either send to standard output, or copy to the user specified file.

The Clear Events Summary History will set the clevsum.txt file to zero length.

Example 1-4 shows the HACMP event summary output, in hacmp.out file.

Example 1-4 HACMP event summary output

```
HACMP Event Summary
Event: node_up sp2n1
Start time: Tue Jun  4 17:35:45 2002

End time: Tue Jun  4 17:36:17 2002

Action:      Resource:      Script Name:
-----
Acquiring resource group:  rg  process_resources
Search on: Tue.Jun.4.17:35:49.EDT.2002.process_resources.rg.ref
Acquiring resource: 10.10.11.1  cl_swap_IP_address
Search on: Tue.Jun.4.17:35:52.EDT.2002.cl_swap_IP_address.10.10.11.1.rg.ref
Acquiring resource: hdisk2  cl_disk_available
Search on: Tue.Jun.4.17:35:57.EDT.2002.cl_disk_available.hdisk2.rg.ref
Resource online:    hdisk2  cl_disk_available
Search on: Tue.Jun.4.17:35:58.EDT.2002.cl_disk_available.hdisk2.rg.ref
Acquiring resource: fkmvg1  cl_activate_vgs
Search on: Tue.Jun.4.17:36:00.EDT.2002.cl_activate_vgs.fkmvg1.rg.ref
Resource online:    fkmvg1  cl_activate_vgs
Search on: Tue.Jun.4.17:36:05.EDT.2002.cl_activate_vgs.fkmvg1.rg.ref
Acquiring resource: /test1  cl_activate_fs

.....

End time: Tue Jun  4 19:16:13 2002

Action:      Resource:      Script Name:
-----
```

No resources changed as a result of this event

GroupName	Type	State	Location	Sticky Loc
casc1	cascading	UP	sp2n4	
casc2	cascading	UP	sp2n4	

Application Availability Analysis Tool

The Application Availability Analysis Tool (AAAT) measures the amount of time that any of the applications under the control of HACMP have been available. HACMP now collects the following information about the state of applications in a log file (clavan.log):

- ▶ An application monitor is defined, changed, or removed.
- ▶ An application starts, stops, or fails.
- ▶ A node fails or is shut down, or comes up.
- ▶ A resource group is taken offline or moved.
- ▶ Application monitoring is suspended or resumed.

The Application Availability Analysis Tool then uses this information to produce its report.

HACMP/ES can monitor applications that are defined to application servers in one of two ways:

- ▶ Process monitoring detects the death of a process, using RSCT event management capability.
- ▶ Custom monitoring monitors the health of an application based on a monitor method that you define.

Once an application server has been defined for a resource group, the above events will start to be recorded in the clavan.log file. More complete information about the availability of the application servers will be recorded if an application monitor is also configured for each application server.

As with other logs, the administrator should ensure that there is adequate space for clavan.log on the file system on which it is being written:

- ▶ The clavan.log file will require roughly 150 bytes of disk storage per outage.
- ▶ Cluster verification provides information about the space available in the file system to which clavan.log is being written.

The default location for clavan.log is /var/adm/. Using the existing Change/Show a Cluster Log Directory SMIT panel, shown in Figure 1-14 on page 29, users may redirect clavan.log to a non-default file system.

```
Change/Show a Cluster Log Directory

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

Cluster Log Name           [Entry Fields]
Cluster Log Description    clavan.log
Default Log Destination Directory  Generated by Applicati>
* Log Destination Directory /var/adm
Allow Logs on Remote Filesystems [/home/hiro]
                             false      +
```

Figure 1-14 Change/Show a Cluster Log Directory

Note: When clavan.log is redirected, the existing data is retained, but it is not moved to the new location. After redirection, the new file will contain only the data that has accumulated since the file was redirected.

To use the AAAT, use the Application Availability Analysis SMIT panel, selecting **Cluster System Management -> Cluster Applications -> Application Availability Analysis**, as shown in Figure 1-15.

```
Application Availability Analysis

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

* Select an Application           [Entry Fields]
* Begin analysis on YEAR (1970-2038) [app11]      +
* MONTH (01-12)                  [2002]      #
* DAY (1-31)                      [06]        #
* Begin analysis at HOUR (00-23)  [04]        #
* MINUTES (00-59)                 [00]        #
* SECONDS (00-59)                 [00]        #
* End analysis on YEAR (1970-2038) [2002]      #
* MONTH (01-12)                   [06]        #
* DAY (1-31)                       [07]        #
* End analysis at HOUR (00-23)    [12]        #
* MINUTES (00-59)                  [00]        #
* SECONDS (00-59)                  [18]        #
```

Figure 1-15 Application Availability Analysis

After pressing Enter on the Application Availability Analysis SMIT screen, the user will see output similar to Figure 1-16 on page 30.

```
COMMAND STATUS

Command: OK          stdout: yes          stderr: no

Before command completion, additional instructions may appear below.

[TOP]
Analysis begins:      Tuesday, 04-June-2002, 00:00
Analysis ends:        Friday, 07-June-2002, 11:50
Application analyzed: appl1

Total time:           3 days, 11 hours, 50 minutes, 30 seconds

Uptime:
  Amount:              3 days, 5 hours, 58 minutes, 29 seconds
  Percentage:          93.00%
  Longest period:      1 days, 16 hours, 48 minutes, 3 seconds

Downtime:
  Amount:              0 days, 5 hours, 52 minutes, 1 seconds
[MORE...14]
```

Figure 1-16 The output of the AAAT

The clavan.log file

The clavan.log file records changes in state of the applications managed by HACMP. It is designed to be readable, but also easily parsed so that the user can write their own analysis scripts.

Each record in clavan.log consists of one line, a fixed portion and a variable portion:

AAA: Ddd Mmm DD hh:mm:ss:YYYY: mnemonic:[data]:[data]: <var portion>

The fixed portion of clavan.log is in the following format:

AAA	A keyword
Ddd	The 3-letter abbreviation for the day of the week
YYYY	The 4-digit year
Mmm	The 3-letter abbreviation for month
DD	The 2-digit day of the month (01...31)
hh	The 2-digit hour of the day (00...23)
mm	The 2-digit minute within the hour (00...59)
ss	The 2-digit second within the minute (00...59)

The variable portion is shown in Table 1-1 on page 31.

Table 1-1 *clavan.log* variable portion

Mnemonic	Description	As used in <i>clavan.log</i> file
umtmonstart	Monitor started	umtmonstart:monitor_name:node:
umtmonstop	Monitor stopped	umtmonstop:monitor_name:node:
umtmonfail	Monitor failed	umtmonfail:monitor_name:node:
umtmonsus	Monitor suspended	umtmonsus:monitor_name:node:
umtmonres	Monitor resumed	umtmonres:monitor_name:node:
umtappstart	App server started	umtappstart:app_server:node:
umtappstop	App server stopped	umtappstop:app_server:node:
umtrgonln	Resource group online	umtrgonln:group:node:
umtrgoffln	Resource group offline	umtrgoffln:group:node:
umtlastmod	File last modified	umtlastmod:date:node:
umtnodefail	Node failed	umtnodefail:node:
umteventstart	Cluster event started	umteventstart:event:
umteventcomplete	Cluster event completed	umteventcomplete:event:

Example 1-5 shows the format of some typical entries in *clavan.log*.

Example 1-5 *clavan.log*

```

Jun  8 17:55:43 2002: umteventcomplete:network_down -1 n24ssa : Cluster event
network_down -1 n24ssa  completed
...
AAA: Sat Jun  8 17:56:37 2002: umtmonstart:app1:sp2n3: Application monitor app1
started on node sp2n3
AAA: Sat Jun  8 17:56:53 2002: umteventcomplete:network_up_complete sp2n3
n24ssa : Cluster event network_up_complete sp2n3 n24ssa  completed
AAA: Sat Jun  8 17:57:25 2002:
umteventstart:/usr/es/sbin/cluster/events/check_for_site_up sp2n4 : Cluster
event /usr/es/sbin/cluster/events/check_for_site_up sp2n4  started
...
AAA: Sat Jun  8 17:57:25 2002: umteventstart:node_up sp2n4 : Cluster event
node_up sp2n4  started
...
AAA: Sat Jun  8 17:59:19 2002:
umteventstart:/usr/es/sbin/cluster/events/check_for_site_down sp2n4 graceful :
Cluster event /usr/es/sbin/cluster/events/check_for_site_down sp2n4 graceful
started

```

```
...
AAA: Sat Jun  8 18:00:35 2002: umtmonstop:app1:sp2n3: Application monitor app1
stopped on node sp2n3
...
AAA: Sat Jun  8 18:00:41 2002: umtappstop:app2:sp2n3: Application app2 stopped
on node sp2n3
...
AAA: Sat Jun  8 18:00:58 2002: umtrgoffln:testcasc1:sp2n3: Resource group
testcasc1 offline on node sp2n3
...
AAA: Sat Jun  8 18:00:59 2002: umteventcomplete:node_down sp2n3 graceful :
Cluster event node_down sp2n3 graceful  completed
```

Limitations and requirements

- ▶ All nodes must be available when you run the tool to display the uptime and downtime statistics. Clocks on all nodes must be synchronized in order to get accurate readings.
- ▶ The AAAT treats an application in a concurrent resource group as available as long as the application is running on any of the nodes in the cluster. Only when the application has gone offline on *all* nodes in the cluster will the AAAT observe that the application is unavailable.
- ▶ The AAAT *cannot* detect availability from an end-user's point of view. For example, if the resource group and the application stay online on a node, but the connection between a client server and the cluster node has been severed, the application server will still be reported as being online during that period.
- ▶ * clavan.log is a cumulative file. The AAAT reports will be inconsistent if run over the period that the clavan.log has changed location.

1.3.3 Network enhancements

The following networking enhancements have been added to both HACMP and HACMP/ES:

- ▶ Persistent node IP alias
- ▶ WAN communication link support

The following networking enhancements have been added to HACMP/ES:

- ▶ Enhanced Network Discovery
- ▶ IP address takeover (IPAT) using IP aliasing

Persistent node IP alias

HACMP now includes the option of a persistent IP label that is available as long as the node has an operating network adapter and stays with that node. It can be

used to access cluster nodes for administrative purposes, for example, HATivoli used to maintain its own IP aliasing scheme using pre-event scripts, post-event scripts, and the `ipalias.conf` file. We recommend that a persistent IP label be configured for use by HATivoli.

A persistent IP label is an IP alias that share an interface with a boot or service address. If the underlying interface fails, the persistent IP label will fail over to another interface on the same network and the same node. This is handled by the `swap_adapter` event. However if the node fails, the label is not taken by another node.

Thus a persistent IP label is an address that:

- ▶ Always stays on the same node (is node-bound).
- ▶ Has only one persistent IP label per node per cluster network.
- ▶ Co-exists with service or boot labels present on an interface.
- ▶ Does not require installing an additional physical adapter on that node.
- ▶ Is *not* part of any resource group.
- ▶ Is applied to the nodes once the cluster configuration is synchronized and is then always available, even if the cluster manager is not running.
- ▶ Is supported on Ethernet, token ring, FDDI, and ATM LANE.
- ▶ Is not supported on SP switch or ATM Classic IP.

For non-aliased networks (that is, public networks in the cluster that use standard IP), the following subnet requirements apply:

- ▶ The subnet of the node's persistent IP label must be different from the subnet of the node's standby adapters.
- ▶ The subnet of the node's persistent IP label may be the same as the subnet of the node's boot and service adapters.

For aliased networks (see "IP address takeover through IP aliasing" on page 39), the following subnet requirement applies:

- ▶ The subnet of the node's persistent IP label must be different from the subnet of the node's boot and service adapters.

For both these network types, the node's persistent IP label will want to be on the same subnet as the TMR node for Tivoli cluster monitoring

Configuring a persistent IP alias address

A persistent IP alias label is configured using SMIT the same way you would to create a boot, service, or standby address. Once a boot adapter has been

defined for a node on a network, a new IP Label is created with the Interface / IP Function attribute set as persistent (Figure 1-17).

Change / Show an Interface / IP Label

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

	[Entry Fields]
IP Label	sp2-n1-per
New IP Label	[] +
* Network Type	[ether] +
* Network Name	[lan1] +
* Network Attribute	[public] +
* Interface / IP-Label Function	[persistent] +
IP Address	[192.168.6.43]
Hardware Address	[]
Node Name	[sp2n1] +
Netmask	[255.255.255.224] +
Interface Name	

Figure 1-17 Configuring a persistent IP alias

Bring up and down persistent IP address

Once the persistent IP alias has been configured and the configuration synchronized across the cluster, the address will be available. This process adds the persistent IP label into the HACMP ODM, creates the IP alias address on the boot adapter using the **ifconfig** command, and makes the following changes to /etc/inittab:

- ▶ Adds an entry for /usr/es/sbin/cluster/etc/harc.net.
- ▶ Ensures that rc.tcpip, rc.nfs, qdaemon, and writesrv are set to run level a and clinit and pst_clinit added at run level a.

The /usr/es/sbin/cluster/etc/harc.net script runs cl_configure_persistent_address to add the alias, as well as starting the above services. Therefore, the alias address will be re-added after a reboot and before HACMP starts, even though its definition does not exist for AIX. Example 1-6 shows the persistent alias address on the boot adapter.

When HACMP cluster services start, the boot IP label is removed, and the service IP label is added. This causes the persistent IP label be removed, so HACMP must then add it back again.

Example 1-6 netstat shows the persistent alias address

```
sp2-n1:/ netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 0.4.ac.49.c7.46 351265 0 337899 0 0
en0 1500 192.168.6.3 sp2-n1 351265 0 337899 0 0
en0 1500 192.168.6.9 sp2-n1-per 351265 0 337899 0 0
```


en1	1500	link#3	0.4.ac.5e.b8.ce	298021	0	296151	0	0
en1	1500	192.168.6.1	sp2-n1-boot2	298021	0	296151	0	0
lo0	16896	link#1		305311	0	305754	0	0
lo0	16896	127	loopback	305311	0	305754	0	0
lo0	16896	::1		305311	0	305754	0	0

Removing a persistent IP label

When you delete a persistent label from the cluster configuration, it is not automatically deleted from the interface on which it is aliased. To remove it, you need to remove the alias with the **ifconfig en0 delete sp2-n1-per** command or reboot the cluster node.

IPAT with persistent IP address

A persistent IP label is not part of any resource group, but in the event of an adapter failure, HACMP/ES moves the persistent IP address and the associated service addresses from the failed adapter to a standby adapter, if one is available. If the node was using an IP aliased network, then HACMP/ES moves this persistent IP address to an available boot adapter on the same node, as shown in Example 1-7, with the other alias addresses.

Example 1-7 Swap adapter with persistent IP address

sp2-n1:/ netstat -i								
Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
en0*	1500	link#2	0.4.ac.49.c7.46	356745	0	340890	0	0
en0*	1500	192.168.6.3	sp2-n1	356745	0	340890	0	0
en1	1500	link#3	0.4.ac.5e.b8.ce	303561	0	299128	0	0
en1	1500	192.168.6.1	sp2-n1-boot2	303561	0	299128	0	0
en1	1500	10.10.11	spn1-alias-svc	303561	0	299128	0	0
en1	1500	192.168.6.9	sp2-n1-per	303561	0	299128	0	0
lo0	16896	link#1		311333	0	311793	0	0
lo0	16896	127	loopback	311333	0	311793	0	0
lo0	16896	::1		311333	0	311793	0	0

Note: There is one specific exception to the rule that the persistent IP label will fail over to a standby adapter if the first adapter fails. The exception happens if the standby adapter has already been taken over by the service label of a cascading resource group not on its home node. In this case if the node's service or boot adapter fails, the persistent IP label will not move to the standby and will not be available.

WAN communication link support

Prior to HACMP 4.5, the only WAN connection that could be made highly available in HACMP/ES was an SNA (CS/AIX) connection configured over a LAN

adapter; for other configurations, the add-on HAWAN product had to be installed. In HACMP/ES 4.5, two additional connection options, SNA over X.25 and pure X.25, can be configured as highly available resources.

An X.25 link is always "aware" of whether or not it has a good connection to the X.25 network. By issuing the **x25status** command from a shell prompt, users can see the status of all links currently active on the node. However, this does not indicate whether or not there is a good end-to-end X.25 link; it only shows that there is a good connection at the network level to at least one other X.25 aware device. HACMP maintains the high availability of an X.25 link by using **x25status** to monitor the status of its network connection.

A stand-alone daemon, **clcomminkd**, is started during **node_up** on nodes that include communication link resources. This daemon is responsible for the following:

- ▶ Starting and stopping highly available X.25 links
- ▶ Monitoring X.25 link connectivity status
- ▶ Launching X.25 link failure recovery procedures

Further details can be found in Chapter 3, "Initial Cluster Planning", and Chapter 18, "Configuring an HACMP/ES cluster" of the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306.

X.25 link failure recovery

When a highly available X.25 link is to be started as part of processing a cluster event, the necessary information is passed to **clcomminkd** for it to determine an appropriate X.25 port for the link. Event utility scripts are then run, and if the link is started successfully, its details will be stored in the table of links that should be monitored. Periodically, **clcomminkd** runs **x25status** to determine the state of each of the links in its table.

If an X.25 link monitored by **clcomminkd** loses network connectivity, the daemon will carry out recovery procedures. These consist of:

- ▶ If there is another X.25 adapter port available on the same node, it will run X.25 link fallover procedures.
- ▶ If the node is running HA/ES and there is not another X.25 adapter port available, selected resource group fallover procedures will be launched.
- ▶ If the node is running HAS and there is not another X.25 adapter port available, then the node will be shut down gracefully with takeover.

Prerequisites for X.25 links

- ▶ Before configuring X.25 adapters in HACMP, they must be defined at the operating system level, as cluster verification will fail if any adapters used by highly available communication links are not recognized by the associated

nodes. The **lsdev -Cc adapter** command will give the list of adapters that are configured on that node.

- ▶ Similarly, the X.25 drivers for the adapters/ports must be recognized. This can be confirmed by the **lsdev -Cc driver** command, which will list the drivers that are currently configured

Note: IBM 2-port Multi protocol Adapters will have one instance of the HDLC driver per port. IBM Artic 960HX adapters will have one instance of the TWD driver per adapter.

- ▶ It is important to confirm exactly which driver instance corresponds with each adapter and/or port that is to be configured.

SNA link failure recovery

The availability of an highly available SNA link is dependent upon the availability of the underlying resource.

An SNA-over-LAN link is an SNA connection that runs over a cluster interface that also hosts a service IP label. If the underlying interface fails, causing the service label to move to a different interface, the SNA link will follow the service label to the new interface. Likewise, an SNA-over-X.25 link is an SNA connection that runs over an X.25 link that is now monitored by HACMP. If the underlying X.25 link fails, causing the X.25 link to move to a different X.25 adapter port, the SNA link will follow the X.25 link to the new port.

The SNA connection itself is not monitored directly. If an SNA link is successfully established, but it fails later for a reason that is not related to its underlying interface or X.25 link, HACMP will not notice the failure and will take no action to recover the SNA link.

Prerequisites for SNA-over-X.25 links

- ▶ An SNA-over-X.25 link is simply a combination of an SNA link and an X.25 link. It shares all of the pre-requisites of X.25 link configuration, and has some additional ones of its own.
- ▶ An SNA link *must* contain a DLC, and it may contain multiple ports and link stations. The DLC and all of the specified ports and link stations must exist in the SNA configuration on each node that may host the associated resource group.
- ▶ Before configuring HACMP, verify that any required DLCs exist on all nodes that will require them.
- ▶ Verify that any SNA ports that will be specified in an SNA link use the DLC that will be part of that link.

- Make sure that any link stations to be specified use one of the SNA ports that will be part of the SNA communication link.

Enhanced network discovery

Network Configuration Discovery has been enhanced for HACMP 4.5. It now runs significantly faster, and allows the creation or deletion of multiple adapters in a single operation.

It is still possible to step through the configuration manually and avoid using this feature. However, there is little reason to do it this way, when you consider the ease and performance of network discovery.

This is now the recommended flow of topology configuration:

1. Add initial interfaces

We found the easiest way was to configure an initial interface on each node through Configure Adapters (see Figure 1-18). Only one interface needs to be configured for each node, so long as they can all be reached when the network discovery process is run (that is, no service labels).

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[sp2-sn2-boot]	+
* Network Type	[ether]	+
Network Name	[lan1]	+
* Network Attribute	[public]	+
* Interface Function	[boot]	+
Interface IP Address	[]	
* Node Name	[sp2n2]	+
Netmask	[]	+

Figure 1-18 Add an Initial Interface

2. Discover IP topology

Now that each node can be reached, run Discover IP Topology (see Figure 1-19 on page 39).

Note: If you want to add or change an IP address and have previously run the automatic IP address discovery cluster-wide, then HACMP/ES filters out (from the list of IP addresses) those IP addresses that have already been added to the cluster configuration.

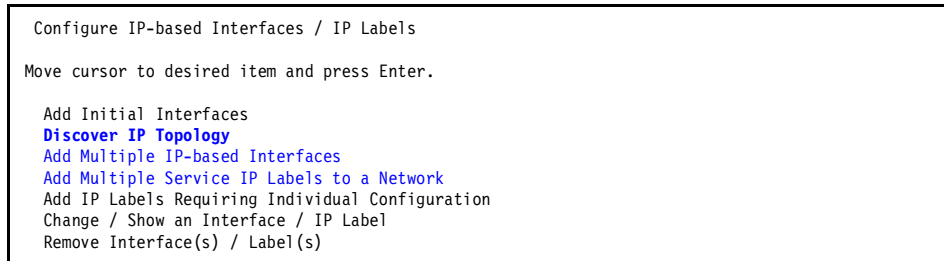


Figure 1-19 Quick configuration of IP Interfaces

3. Add multiple boot and standby interfaces

Next, define the remaining boot and standby IP Interfaces (Ether, FDDI, and so on) (see Figure 1-19). Multiple adapters with the same attributes on the same network can be defined in one operation (a space separated list of IP labels).

4. Add multiple shared service labels

The service IP Labels are added in a similar fashion (see Figure 1-19). Adapters with particular attributes need to be configured in the next step.

5. Add individual non-shared service labels, persistent labels, and those with HWAT.

Those adapters with individual characteristics, such as hardware address swapping or persistent node IP Labels, can be configured using the SMIT menu item Add IP Labels Requiring Individual Configuration (see Figure 1-19).

Note: Those users more familiar with configuring each adapter individually will have to ensure that the network information is correct - particularly subnet details - or resources may not fail over successfully. This should be done through the SMIT **Cluster Topology -> Configure Networks** menus

IP address takeover through IP aliasing

HACMP/ES 4.5 provides a facility for IP address takeover (IPAT) through IP aliasing. Instead of the service IP address replacing the boot IP address, the service address is added as an alias to the boot adapter. Also, IPAT through IP aliasing does not use standby adapters. All adapters on an aliased network should be created with their function defined as boot.

Note: Boot interfaces and service labels on an aliased network should be configured to use separate subnets.

When a new network is added to the cluster, its status as an aliased network will be discovered automatically.

Comparing traditional IPAT to IPAT through aliasing

1. Aliased networks decrease the time required for adapter and resource group recovery. In our test environment, we recorded the following times to completely execute the `acquire_service_addr` event:

IPAT through aliasing 5 seconds

Traditional IPAT 11 seconds

2. With IPAT through aliasing, the adapter has both the boot and service IP addresses configured, that is, the service address is added as an alias on the adapter. Unlike in traditional IPAT, the boot address is never removed from an adapter.
3. As with traditional IPAT, if a node fails on an aliased network, the takeover node acquires the failed node's service address as an alias on one of its boot adapters on the same HACMP network. This makes the failure transparent to clients using that specific service address.
4. Hardware address takeover is not supported by IPAT through aliasing.
5. The concept of boot and standby adapters becomes irrelevant and the alias will use the adapters in the order in which they were defined to HACMP. The `c11sif` command will show this order.
6. When a resource group with a aliased service address falls over to another node, the alias address will be added to the adapter that appears first in the HACMP definition, that is, the adapter that already has aliased addresses.

Configure adapters for aliased networks

Configuring an adapter for an aliased network that will use IPAT through IP Aliasing is simply a matter of configuring each adapter as a boot adapter and ensuring that the service addresses are resolvable and on a different subnet to the boot addresses.

The SMIT menu Add Multiple Service IP Labels to a Network is then used to define the alias labels (see Figure 1-20 on page 41).

Add Multiple Service IP Labels to a Network

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

	[Entry Fields]	
* IP Label(s)	[spn1-alias-svc	> +
Network Type	ether	
Network Attribute	public	
Network Name	lan1	
IP Label Function	service	

Figure 1-20 Add Multiple Service IP Labels to a Network

The configuration of the remaining topology is the same as for traditional IPAT. Once the cluster manager is started, the **netstat** output shows both the boot and alias IP labels (see Example 1-8).

Example 1-8 IPAT using IP aliasing

```
sp2-n1:/ netstat -i
```

Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
en0	1500	link#2	0.4.ac.49.c7.46	53343	0	44928	0	0
en0	1500	192.168.6.3	sp2-n1	53343	0	44928	0	0
en0	1500	10.10.11	spn1-alias-svc	53343	0	44928	0	0
en1	1500	link#3	0.4.ac.5e.b8.ce	9454	0	9063	0	0
en1	1500	192.168.6.1	sp2-n1-boot2	9454	0	9063	0	0
lo0	16896	link#1		37599	0	37843	0	0
lo0	16896	127	loopback	37599	0	37843	0	0
lo0	16896	::1		37599	0	37843	0	0

```
sp2-n1:/ netstat -in
```

Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
en0	1500	link#2	0.4.ac.49.c7.46	361035	0	345112	0	0
en0	1500	192.168.6.3	192.168.6.35	361035	0	345112	0	0
en0	1500	10.10.11	10.10.11.1	361035	0	345112	0	0
en1	1500	link#3	0.4.ac.5e.b8.ce	307831	0	303333	0	0
en1	1500	192.168.6.1	192.168.6.131	307831	0	303333	0	0
lo0	16896	link#1		315764	0	316236	0	0
lo0	16896	127	127.0.0.1	315764	0	316236	0	0
lo0	16896	::1		315764	0	316236	0	0

After a `swap_adapter` event, we see the same configuration with the boot IP addresses on each adapter, but with the alias now moved to the other adapter (see Example 1-9).

Example 1-9 netstat output after swap_adapter event in IP aliasing environment

```
sp2-n1:/ netstat -i
```

Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
en0*	1500	link#2	0.4.ac.49.c7.46	356745	0	340890	0	0

en0*	1500	192.168.6.3	sp2-n1	356745	0	340890	0	0
en1	1500	link#3	0.4.ac.5e.b8.ce	303561	0	299128	0	0
en1	1500	192.168.6.1	sp2-n1-boot2	303561	0	299128	0	0
en1	1500	10.10.11	spn1-alias-svc	303561	0	299128	0	0
lo0	16896	link#1		311333	0	311793	0	0
lo0	16896	127	loopback	311333	0	311793	0	0
lo0	16896	::1		311333	0	311793	0	0

After a fallover of a cascading resource group, we see that the IP alias address is now on the second adapter on the standby node, exhibiting the same behavior as traditional IPAT (see Example 1-10).

Example 1-10 netstat output on standby node after takeover

sp2-n2:/ netstat -i									
Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll	
en0	1500	link#2	0.4.ac.49.ba.46	27698	0	25171	0	0	
en0	1500	192.168.6.3	sp2-n2	27698	0	25171	0	0	
en0	1500	10.10.12	spn2-alias-svc	27698	0	25171	0	0	
en1	1500	link#3	0.20.35.12.40.e8	8358	0	7851	0	0	
en1	1500	192.168.6.1	sp2-n2-boot2	8358	0	7851	0	0	
en1	1500	10.10.11	spn1-alias-svc	8358	0	7851	0	0	
lo0	16896	link#1		7924	0	7982	0	0	
lo0	16896	127	loopback	7924	0	7982	0	0	
lo0	16896	::1		7924	0	7982	0	0	

Aliased networks are public networks in the cluster that are configured to use IPAT through IP Aliasing. HACMP/ES *automatically* recognizes a cluster network as an aliased network as long as the following apply:

- ▶ Standby adapters are not configured.
- ▶ Hardware address takeover (HWAT) for IP service labels is not defined in HACMP/ES.
- ▶ This network type supports Gratuitous ARP in HACMP/ES.
- ▶ Boot and service adapters are defined, and are placed on different subnets.

Thus, it can be seen that the only configuration difference between IPAT through aliasing and traditional IPAT is that, for aliased networks, you configure multiple boot adapters on different subnets instead of configuring boot and standby adapters on different subnets.

While using IPAT through aliasing, each node that can have its IP address taken over must have at least one interface configured for the boot function on the network; however, we strongly recommend that you have *two* boot interfaces to remove single points of failure. Sometimes this option is not always available, so

it is possible to configure only one boot adapter. As shown in Example 1-11, service IP alias address is configured on en0 after the failover.

Example 1-11 netstat for aliasing and one adapter

```
sp2-n2:/ netstat -i
```

Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
en0	1500	link#2	0.4.ac.49.ba.46	323739	0	311994	0	0
en0	1500	192.168.6.3	sp2-n2	323739	0	311994	0	0
en0	1500	10.10.11	spn1-alias-svc	323739	0	311994	0	0
en0	1500	10.10.12	spn2-alias-svc	323739	0	311994	0	0
lo0	16896	link#1		272826	0	273273	0	0
lo0	16896	127	loopback	272826	0	273273	0	0
lo0	16896	::1		272826	0	273273	0	0

Limitations and considerations

To enable IP Address Takeover through IP Aliasing, you should configure adapters according to the following requirements:

- ▶ Instead of standby adapters, you should have additional boot adapters configured on the node. These boot adapters should be defined on different subnets to the other boot adapters and the service IP addresses. If your configuration has only one boot adapter, you will receive the following warning during the synchronization of the cluster topology:

```
WARNING: There may be an insufficient number of boot adapters defined on
node sp2n1 and network lan1.
Multiple boot adapters are recommended for networks that will use IP
aliasing.
```

- ▶ At least one boot adapter label must be assigned to the service address.
- ▶ Standby adapters should not be defined on any cluster node on the HACMP network.
- ▶ Hardware Address Takeover should *not* be configured for any adapter on the HACMP network.
- ▶ A service address must be on a different subnet from ALL boot addresses defined on the cluster node. This requirement enables HACMP/ES to comply with the IP route striping functionality of AIX 5L Version 5.1, which allows multiple routes to the same subnet.
- ▶ Multiple service and boot labels can coexist as aliases on a given adapter on the cluster network.
- ▶ Old-style HPS networks are still supported in HACMP 4.5, but it is recommended that users switch to the new IPAT through IP aliasing network.

If you have more than two boot addresses (for example, boot_1 and boot_2) on each node, the alias service IP addresses may not always be configured on en0, as shown in Example 1-12

Example 1-12 Service address configured on boot_2

```
sp2-n1:/usr/es/sbin/cluster/utilities netstat -i
```

Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
en0	1500	link#2	0.4.ac.49.c7.46	142389	0	60912	0	0
en0	1500	192.168.6.3	boot_1	142389	0	60912	0	0
en1	1500	link#3	0.4.ac.5e.b8.ce	33483	0	28672	0	0
en1	1500	192.168.6.1	boot_2	33483	0	28672	0	0
en1	1500	10.10.11	spn1-alias-svc	33483	0	28672	0	0
lo0	16896	link#1		66281	0	68243	0	0
lo0	16896	127	loopback	66281	0	68243	0	0

The adapter that will get the alias addresses is the first adapter that appears in the output of the `c11sif` command. Using Example 1-12, the order according to `c11sif` is shown in Example 1-13.

Example 1-13 Adapter order as shown by the c11sif command

```
sp2-n1: /usr/es/sbin/cluster/utilities/c11sif -cSi sp2n1
sp2-n1-boot2:boot:lan1:ether:public:sp2n1:192.168.6.131::en1::255.255.255.224
sp2-n1:boot1:lan1:ether:public:sp2n1:192.168.6.35::en0::255.255.255.224
spn2-alias-svc:service:lan1:ether:public:sp2n1:10.10.11.2:::255.255.255.224
spn1-alias-svc:service:lan1:ether:public:sp2n1:10.10.11.1:::255.255.255.224
```

If, for any reason, it is required that the service alias addresses be configured on the first adapter, then removing and redefining the definition for boot_2 will change the order.

Note: Note the order of the adapters, as shown by the `c11sif` command, is the reverse of that shown by `odmget HACMPadapter`.

1.3.4 Device support

Device support for Fibre Channel tape drive has been added to both HACMP and HACMP/ES.

Fiber Channel tape drive support

In HACMP/ES 4.5, you can configure a Fibre Channel (as well as SCSI) tape drive as a cluster resource, making it highly available to multiple nodes in a cluster. Management of shared tape drives is simplified by HACMP functionality.

HACMP 4.4 added SCSI streaming tape devices to the range of resources that could be added to a cluster resource group. However, the tape device support SMIT screens were not initially made user-visible.

With HACMP 4.5, users can now configure HA tape resources via SMIT, and the range of supported devices has been extended to include Fiber Channel tape drives.

For more information on configuring and maintaining a shared tape drive as a cluster resource, refer to Chapter 5, “Planning Shared Disk and Tape Devices”, Chapter 11, “Checking installed software”, and Chapter 18, “Configuring an HACMP/ES cluster” of the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306.

Reserving and releasing tape drives

When a resource group with tape resources is activated, the tape drive is reserved to allow its exclusive use. This reservation is held until an application releases it or the node is removed from the cluster.

When the special file for the tape is closed, the default action is to release the tape drive. An application can open a tape drive with a "do not release on close" flag. HA/ES will not be responsible for maintaining the reservation after an application is started.

When a node does a graceful shutdown, the tape drive is released, allowing access from other nodes. However, if the node fails, a forced release is done by the node taking over the resource group containing the tape drive. The activation of that resource group then reserves the tape drive.

Starting and stopping tape resources

Tape drive acquisition and release procedures are highly dependent on the application accessing the tape drive. Rather than trying to predict likely scenarios and develop all the necessary procedures, HACMP provides for the execution of user defined tape device start and stop scripts.

Tape start and stop scripts are invoked when a resource group is activated (tape start) or when a resource group is deactivated (tape stop). The following sample start and stop scripts can be found in the `/usr/es/sbin/cluster/samples/tape` directory:

```
tape_resource_start_example  
tape_resource_stop_example
```

During tape start, HACMP reserves the tape drive, forcing a release if necessary, and then invokes the user-provided tape start script.

During tape stop, HACMP invokes the user-provided tape stop script, and then releases the tape drive.

Synchronous versus asynchronous operation

If a tape operation is in progress when a tape reserve or release is initiated, it may take many minutes before the reserve or release operation completes. HACMP allows synchronous or asynchronous reserve and release operations, as shown in Figure 1-21.

HACMP allows only two methods to handle the reserve/release operations:

- Synchronous operation**
- HACMP waits for the reserve or release operation to complete before continuing with the rest of the cluster event. This is the default value.
- Asynchronous operation**
- HACMP creates a child process to perform the reserve or release operation, including the execution of a user defined recovery procedure, and continues immediately.

Note: For asynchronous operation, the user must provide a way to notify the application server to wait until the reserve/release operation is complete

This can be configured through the SMIT menu Add a Tape Resource (see Figure 1-21).

Add a Tape Resource

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

* Tape Resource Name

Description

* Tape Device Name

Start Script

Stop Script

Start Processing Synchronous?

Stop Processing Synchronous?

[Entry Fields]

[ha_tape]

[/dev/rmt0]

[/usr/es/sbin/cluster/s>

[/usr/es/sbin/cluster/s>

Yes

Yes

+

+

Figure 1-21 Synchronous versus asynchronous operation

Requirements and limitations

When planning to use highly available tape resources, the following items should be noted:

- ▶ Support is limited to SCSI or Direct Fibre Channel tape drives that have hardware reserve and hardware reset/release functions.
- ▶ Tape loaders/stackers are not supported.

- ▶ No more than two cluster nodes can share the tape resource.
- ▶ Tape resources may not be part of concurrent resource groups.
- ▶ The tape drive must have the same name, for example, /dev/rmt0, on both nodes sharing the tape device.
- ▶ When a tape special file is closed, the default action is to release the tape drive. HACMP is not responsible for the state of the tape drive once an application has opened the tape.
- ▶ Tape drives with more than one SCSI interface are not supported. Therefore, only one connection can exist between a node and the tape drive. Adapter failover is not an option.

1.3.5 Application support

The following enhancements extend the application support of both HACMP and HACMP/ES:

- ▶ Integration with AIX Workload Manager
- ▶ Cluster administration and security via Tivoli

The following enhancements extend the application support of HACMP/ES:

- ▶ GPFS integration

Integration with AIX Workload Manager

Workload Manager (WLM) can be used to manage system resources with AIX Version 4.3.3 and above. It allows limits to be set on CPU time, physical memory usage, and disk I/O bandwidth usage for different process and applications. It is designed for better allocation of critical system resources at peak system loads. HACMP 4.5 allows Workload Manager classes to be added to cluster resource groups so that the starting and stopping of WLM and the active WLM configuration then under the control of the cluster.

When the cluster is started or reconfigured, HACMP will manage the WLM on each node so that only the requested classes are active. When the cluster is stopped, HACMP will return it to the state that it was in when the cluster was started, while at the same time preserving various configuration files for debugging purposes.

HACMP does not verify every aspect of your WLM configuration; therefore, it remains your responsibility to ensure the integrity of the WLM configuration files. After you add the WLM classes to an HACMP resource group, the clverify utility checks only whether the required WLM classes exist. Therefore, you must fully understand how WLM works and configure it carefully. Poorly planned but

consistent configuration parameters can reduce the productivity and availability of the system.

Workload Manager background

The Workload Manager distributes system resources among requesting processes according to the class to which these processes are assigned. The properties of a class include:

The class name	A unique alphanumeric string, up to 16 characters.
The class tier	A number from 0 to 9, representing the relative importance of the class from most important (tier 0) to least important (tier 9)
Number of shares	This applies to the allocation of CPU time and physical memory. The actual number of resources allocated to a class depends on the total number of shares in all classes. Thus, if two classes are defined on the system, one with two shares of the target CPU usage, and the other with three shares, then the first class will receive 2/5 of the CPU time, and the second 3/5.
Minimum and maximum percentages	These limits are used for setting limits on CPU time, physical memory, and disk I/O bandwidth accessible by the process

Class assignment rules are defined to tell Workload Manager how to classify all new process (as well as those already running when Workload Manager started) according to their UID, GID and full path name.

For complete information on how to set up and use Workload Manager, see the redbook *AIX 5L Workload Manager (WLM)*, SG24-5977.

Workload Manager reconfiguration

After Workload Manager classes have been added to an HACMP resource group and the cluster is synchronized, HACMP reconfigures the Workload Manager so that it will use the rules required by the classes associated with the node. In the case of a Dynamic Reconfiguration (DARE) event, Workload Manager will also be reconfigured in response to any changes made to Workload Manager classes associated with the resource group changes.

Workload Manager startup

Workload Manager startup occurs either when the node joins the cluster or when a DARE event takes place. The configuration is node specific and depends on the resource groups the node participates in. If the node cannot acquire any resource groups associated with Workload Manager classes, then WLM will not be started.

The next step depends on the type of resource group:

Cascading

The startup script will determine whether the resource group is running on its home node or not, and will add the corresponding WLM class assignment rules to the WLM configuration. The primary Workload Manager class is used on the *home* node, and the secondary class is used on all the other nodes in the resource group's node list.

Rotating or Concurrent

For each of the resource groups of this type that the node can acquire, the primary WLM class associated with the resource group will be placed in the WLM configuration. The corresponding rules will be added to the rules table.

If, however, the Workload Manager was running and not started by HACMP, the startup script will save the current configuration and then restart WLM using the HACMP specified configuration. When HACMP is stopped, Workload Manager is returned to its saved configuration.

Failure to start the Workload Manager will generate an error message in the hacmp.out log file, but the node startup or resource configuration will continue normally.

Note: Once HACMP starts on a node that has been configured for WLM, only those WLM rules that are associated with the classes in the resource groups associated with that node will be active on that node.

Workload Manager shutdown

Workload Manager shutdown occurs either when the node leaves the cluster or a DARE event. The shutdown script will:

- ▶ Do nothing if Workload Manager is not currently running.
- ▶ Check if Workload Manager was running prior to being started by HACMP. If Workload Manager was running, then:
 - If it was not running prior to being started by HACMP, it will be stopped.

- If it was running prior to being started by HACMP, it will be stopped and then restarted with its prior configuration.

Limitations and considerations

The following should be considered when planning your Workload Manager configuration:

- ▶ Some WLM configurations could adversely affect HACMP performance. Be careful when designing your classes and rules, and be sure you understand their implications and how they might affect HACMP.
- ▶ You can have no more than 27 non-default WLM classes across the cluster, since one configuration is shared across the cluster nodes.
- ▶ An HACMP Workload Manager configuration does not support sub-classes, even though WLM allows them in AIX 5L Version 5.1. If you configure sub-classes for any WLM classes that are placed in a resource group, a warning will be issued upon cluster verification, and the sub-classes will not be propagated to other nodes during synchronization.
- ▶ On any given node, only the rules for classes associated with resource groups that can be acquired by a node are active on that node.
- ▶ Proper configuration of the WLM can be a complex task that requires a detailed analysis of cluster resource using patterns under normal and extraordinary operating conditions.

Configuring Workload Manager with HACMP

Configuring the Workload Manager in HACMP classes involves four simple steps:

1. Configure WLM classes and rules using the appropriate SMIT panels.
Start by creating a Workload Manager configuration through SMIT by selecting **Performance and Resource Scheduling -> Workload Management -> Work on alternate configurations -> Create a configuration**. (The fast path is **smitty wlm**) (see Figure 1-22 on page 51). We recommend that you use the default name supplied by HACMP (HA_WLM_config).

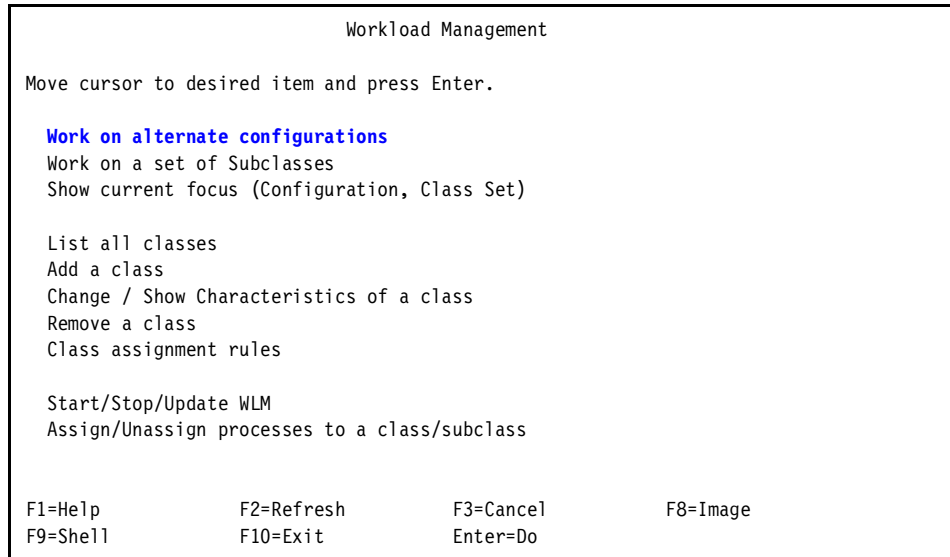


Figure 1-22 Creating a Workload Manager configuration

Classes now need to be added to the Workload Manager configuration that you have created. At the main SMIT menu, select **Performance and Resource Scheduling -> Workload Management -> Add a class and Change / Show Characteristics of a class**. These are the classes that will be added into the resource group definitions.

Note: It is not in the scope of this redbook to cover detailed scenarios of Workload Manager configurations.

2. Optionally, If you chose a configuration other than the default, HACMP must be configured to use it.

This is where you define, to HACMP, what Workload Manager configuration it should manage. If you have chosen a name other than the default, then this needs to be changed in the HACMP configuration. This is done using SMIT from the main HACMP menu, selecting **Cluster Configuration -> Cluster Resources -> Change/Show HACMP Workload Manager Run-time Parameters**, and adding in the new name (see Figure 1-23 on page 52).

Workload Manager Configuration

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

* Workload Manager Configuration

[Entry Fields]

[HA_WLM_config] +/

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

Figure 1-23 Define Workload Manager configuration name to HACMP

3. Assign the classes for this configuration to a resource group, selecting from the classes pick list.

The Workload Manager classes that were created in Step 2 can now be assigned to resource groups. From the main HACMP SMIT menu, select **Cluster Configuration -> Cluster Resources -> Change/Show Resource/Attributes for a Resource Group** (see Figure 1-24 on page 53).

Change/Show Resources/Attributes for a Resource Group

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

[TOP]	[Entry Fields]	
Resource Group Name	casc2	
Node Relationship	cascading	
Site Relationship	ignore	
Participating Node Names / Default Node Priority	s38 s37	
Dynamic Node Priority	[]	+
Service IP label	[c38srvc01]	+
.....		
Application Servers	[app2]	+
Communication Links	[]	+
Primary Workload Manager Class	[WLM_App1]	+
Secondary Workload Manager Class	[WLM_App2]	+

Primary Workload Manager Class

Move cursor to desired item and press Enter.

WLM_App1

WLM_App2

+

+

+

+

+

[B]

F1=Help

F2=Refresh

F3=Cancel

F1 F8=Image

F10=Exit

Enter=Do

F5 /-=Find

n=Find Next

F9+-----

Figure 1-24 Adding Workload Manager classes to a resource group

- After adding the Workload Manager classes to the resource groups and resource group configuration is complete, synchronize the cluster resources.

After making the above changes to the configuration of your resource groups, these changes need to be synchronized across the cluster. At this time, run the **clverify** command (see Figure 1-25 on page 54) It verifies the following items:

- Checks that, for each resource group, that has an associated WLM class, that there is an application server defined. While this is not required, it is expected, and clverify issues a warning if no application server is found.
- Checks that each WLM class defined in a resource group exists in the specified HACMP Workload Manager directory.
- Checks that a cascading resource group does not contain a secondary WLM class without a primary one being defined.
- Checks that concurrent and rotating resource groups only have primary Workload Manager classes defined.

```

                                COMMAND STATUS

Command: OK                      stdout: yes                      stderr: no

Before command completion, additional instructions may appear below.

[MORE...117]

Verifying Custom Disk Methods...

No custom disk methods to verify.

Verifying WLM settings for resource group: cascl.
-----

Verifying WLM settings for resource group: casc2.
-----

Verifying export files for resource group: cascl.
[MORE...105]

F1=Help          F2=Refresh          F3=Cancel          F6=Command
F8=Image         F9=Shell            F10=Exit           /=Find
n=Find Next

```

Figure 1-25 Verification of Workload Manager by clverify

Note: The **clverify** command cannot check class assignment rules, because it has no way of knowing the eventual UID, GID, or path name of the user application. The administrator is completely responsible for assigning user applications to Workload Manager classes when configuring the WLM class assignment rules.

Cluster verification can only look for obvious problems and cannot verify all aspects of the configuration. This should be planned well in advance.

Workload Manager entries in hacmp.out log

The hacmp.out log file records the cl_wlm_reconfig event, which is called by node_up. After that completes, node_up calls cl_wlm_start. Typical entries from the hacmp.out file can be seen in Figure 1-14.

Example 1-14 Workload Manager entries in hacmp.out

```

:node_up[149] [[ REAL = EMUL ]]
:node_up[157] rm -f /usr/es/sbin/cluster/etc/.hacmp_wlm_config_changed
:node_up[159] cl_wlm_reconfig node_up
:node_up[159] EMULATE=REAL

```

```

:cl_wlm_reconfig[287] [[ high = high ]]
:cl_wlm_reconfig[287] version=1.8
:cl_wlm_reconfig[288] :cl_wlm_reconfig[288] cl_get_path
HA_DIR=es
:cl_wlm_reconfig[289] SCD=/usr/es/sbin/cluster/etc/objrepos/stage
:cl_wlm_reconfig[290] ACD=/usr/es/sbin/cluster/etc/objrepos/active
:cl_wlm_reconfig[292] EMULATE=REAL
:cl_wlm_reconfig[294] CALLING_EVENT=node_up
:cl_wlm_reconfig[296] HA_WLM_CLASSES=
....
GROUP=casc1
:cl_wlm_reconfig[14] :cl_wlm_reconfig[14] cut -d: -f2
:cl_wlm_reconfig[14] echo casc1:cascading:ignore:c37 c38
TYPE=cascading
:cl_wlm_reconfig[15] NODES=c37 c38
:cl_wlm_reconfig[15] [[ -z casclcascadingc37 c38 ]]
:cl_wlm_reconfig[15] [[ cascading = cascading ]]
:cl_wlm_reconfig[15] [[ c37 = c37 ]]
:cl_wlm_reconfig[28] PRIMARY= casc1
:cl_wlm_reconfig[11] read line
:cl_wlm_reconfig[13] :cl_wlm_reconfig[13] cut -d: -f1
:cl_wlm_reconfig[13] echo casc2:cascading:ignore:c38 c37
GROUP=casc2
.....
:node_up[170] cl_wlm_start
:cl_wlm_start[47] [[ high = high ]]
:cl_wlm_start[47] version=1.4
:cl_wlm_start[48] :cl_wlm_start[48] cl_get_path
HA_DIR=es
:cl_wlm_start[51] :cl_wlm_start[51] awk BEGIN { FS = ":" } $1 !~ /^#.* / { print
$1 }
:cl_wlm_start[51] /usr/es/sbin/cluster/utilities/clwlmruntime -l
HA_WLM_CONFIG=HA_WLM_config
:cl_wlm_start[52] [[ -z HA_WLM_config ]]
:cl_wlm_start[61] wlmcntrl -q
WLM is stopped
:cl_wlm_start[62] WLM_IS_RUNNING=1
:cl_wlm_start[65] [[ ! -e /etc/wlm/HA_WLM_config/HA_prev_config_subdir ]]
:cl_wlm_start[67] echo
:cl_wlm_start[67] 1> /etc/wlm/HA_WLM_config/HA_prev_config_subdir
:cl_wlm_start[68] [[ 1 -eq 0 ]]
:cl_wlm_start[94] [[ 1 -eq 0 ]]
:cl_wlm_start[103] wlmcntrl -a -d HA_WLM_config
:cl_wlm_start[104] [ 0 -ne 0 ]
:cl_wlm_start[112] [[ -e /etc/wlm/HA_WLM_config/rules ]]
:cl_wlm_start[114] mv /etc/wlm/HA_WLM_config/rules
/etc/wlm/HA_WLM_config/rules.active
:cl_wlm_start[117] [[ -e /usr/es/sbin/cluster/etc/wlm/rules ]]

```

```
:cl_wlm_start[119] cp /usr/es/sbin/cluster/etc/wlm/rules
/etc/wlm/HA_WLM_config/rules
:cl_wlm_start[123] exit 0
:node_up[183] :node_up[183] cl_rrmethods2call ss_load
....
```

For more information on configuring WLM on HACMP, see Chapter 7, “Planning resource groups”, and Chapter 18, “Configuring in HACMP/ES cluster” of the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SG23-4306, or Chapter 7, “Planning Applications, Application Servers, and Resource Groups” in the *HACMP for AIX 4.5 Planning Guide*, SC23-4277.

Cluster administration and security via Tivoli

HACMP 4.4.0 introduced cluster monitoring via the Tivoli Framework interface, supporting the following:

- ▶ Cluster state and substate
- ▶ Configured networks and network state
- ▶ Participating nodes and node state
- ▶ Configured resource group location and state
- ▶ Individual resource location (not state).

In HACMP 4.5, users can now perform the following cluster administration tasks from within Tivoli as well:

- ▶ Start cluster services on specified nodes
- ▶ Stop cluster services on specified nodes
- ▶ Bring a resource group online (ES)
- ▶ Bring a resource group offline (ES)
- ▶ Move a resource group to another node. (ES)

After one of the new Tivoli cluster administration tasks has been run, if the output destination is set to `Display to Desktop`, Tivoli will display output from the cluster nodes indicating the success or failure of the operation.

Note: Tivoli is not yet fully supported on AIX 5L Version 5.1. The features detailed here will be shipped with HACMP 4.5, but they will not be fully tested or supported until complete Tivoli support is available for the required version of AIX.

Security enhancement in HATivoli/HATivoli

In previous releases of HATivoli, users could login to AIX as non-root users on the TMR node, start Tivoli, and then choose to open the SMIT panels on a managed node from within Tivoli. These SMIT panels would be opened with root

permissions, even though the user did not have root permission on the TMR node.

HACMP 4.5 adds root password verification to the process. Each time a non-root user runs the command to open a SMIT window for cluster management, the user will be required to enter the root password for the managed node before the SMIT panel will be displayed.

If the root user attempts to open a SMIT window via Tivoli, the window will open without additional authentication.

Although HAView does not share the same security weakness as HATivoli in regard to opening remote SMIT windows, it has been updated to include this security mechanism, in order to maintain uniformity of presentation between the two programs.

GPFS integration

GPFS Version 1.5 provides concurrent high speed file access to applications executing on multiple systems that are part of an HACMP/ES cluster.

Specifically, GPFS allows:

- ▶ Execution of parallel applications that require concurrent sharing of the same data from many nodes, including concurrent update of files.
- ▶ Parallel maintenance of metadata, giving higher scalability and availability.
- ▶ Failure recovery, protecting against the loss of data access for the surviving nodes when one node in the GPFS cluster fails.

HACMP/ES 4.5 and AIX 5L Version 5.1 provide the operating and administrative environment for GPFS in a cluster environment. GPFS uses the Group Services and Topology services components of RSCT.

RSCT services for GPFS Version 1.5 include:

- ▶ Coordinating GPFS daemon membership during operation.
- ▶ Handling recovery actions in case of failure.
- ▶ Handling GPFS Cluster topology: nodes, networks, and network adapters used by a GPFS cluster are configured as part of the larger HACMP/ES cluster. Only one GPFS cluster can be configured per HACMP/ES cluster.

Each node in the GPFS cluster is defined as belonging to a single GPFS nodeset, and must have access to all the directly-attached disks that form part of the GPFS cluster. A GPFS cluster can contain multiple nodesets. Nodes can be dynamically added to or removed from a GPFS nodeset.

One service-only network is defined for communication between GPFS daemons.

The nodes can continue to host HACMP/ES cluster resources as well as the GPFS resources.

GPFS resources are not defined to the HACMP/ES cluster; they are defined only for the GPFS cluster.

When you configure and maintain GPFS cluster, you can see more information in Appendix H, “Configuring a GPFS Cluster”, in the *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306.

Limitations and considerations

- ▶ Support is now available for the IBM General Parallel File System (GPFS) if run within an HACMP/ES cluster environment.
- ▶ GPFS uses an IP network to connect all of the nodes. This is typically a LAN with sufficient bandwidth (100Mbps) available for GPFS control traffic. The IP network and adapters must be configured within the HACMP/ES cluster in order for the GPFS system to function.
- ▶ When issuing GPFS commands, the host name or IP address for the node must refer to the adapter port over which the GPFS daemons communicate. Alias interfaces are not allowed, and GPFS cannot make use of IPAT or other methods of adapter failure recovery provided by HA/ES.
- ▶ GPFS requires that the SSA or Fibre Channel disk devices be configured on all nodes that will mount the file systems. The disk devices are not configured within the HACMP/ES cluster.

1.4 Installation and migration considerations

With HACMP 4.5, the prerequisites for installation of HACMP have changed.

HACMP 4.5 now requires AIX 5L Version 5.1 and RSCT Version 2.2.1.0. With this change, some components of the product that are considered obsolete have been removed. Installation of HACMP now requires users to explicitly accept the license agreement in order for the product to be successfully installed. Customers may need to install the required PTFs prior to running HACMP 4.5.

Features removed from HACMP 4.5

The following obsolete components have been removed from the product in HACMP 4.5:

- The VSM utility Xhacmpm

- Support for 9333 disks
- The Task Guides utility
- Support for HTY service labels
- The Quick Config utility
- Postscript version of the user documentation (docs are now shipped as HTML and PDF)

Required PTFs for HACMP/ES 4.5

You need APAR IY29784 for all versions of HACMP/ES 4.5. For HACMP/ES on the RS/6000 SP and Cluster 1600, you also need APAR numbers IY29490 and IY29870.

Concurrent access migration issues

When migrating concurrent volume groups from a previous version of HACMP (HAS) or HACMP/ES to HACMP/ES 4.5, certain factors must be considered:

- ▶ Enhanced concurrent mode is supported only on AIX 5L Version 5.1 and higher.
- ▶ SSA concurrent mode is not supported on 64-bit kernel.
- ▶ All nodes in a cluster must use the same form of concurrent mode for a given volume group.

If you have SSA disks in concurrent mode, you cannot run 64-bit kernel until you have converted all volume groups to enhanced concurrent mode.

Migration from earlier HACMP versions with HATivoli

During a cluster migration from earlier versions to HACMP/ES 4.5, if the cluster has HATivoli installed and has IP aliases configured, HACMP/ES integrates the existing HATivoli aliases by using them as persistent node IP labels, and removes the HATivoli post-event script entries from the HACMP event cluster ODM class.

Note that the post-event scripts remain in the file system. If you have modified any of the original HATivoli post-event scripts and want your modifications to continue to run as post-event scripts, you must move the code to a new script and add it as a new post-event script in HACMP/ES.

Migration issues about HACMP 4.5: New features

The migration issues are for Resource Group Temporal Ordering and Parallel Processing:

- ▶ In releases prior to 4.5, all resource groups were processed in strict alphabetical order during both acquisition and release.

- ▶ This order is maintained automatically on migration from earlier releases, because it represents the HAS default order. Users will not need to make any changes to the configuration, and no conversion is required.

The following migration issues are for IPAT through IP Aliasing:

- ▶ No changes to the existing cluster network configuration are required during migration from HAS or upgrade from a previous release of HA/ES.
- ▶ During migration or upgrade, all existing networks will be set to have the “Use IP Aliasing for IPAT” field disabled.
- ▶ Following the migration or upgrade to HA/ES 4.5, we recommend that existing HACMP networks on the SP switch be converted into IPAT through IP aliasing networks.



Configuring highly available p690 clusters

IBM @server pSeries 690 family of servers incorporates the advanced technologies available on the IBM @server line, as well as technology enhancements from IBM research divisions. The results are high-performance, high-availability servers that offer enhanced features and benefits.

The pSeries p690 is based on a modular design. It features a Central Electronics Complex (CEC) where memory and processors are installed, as well as power subsystem, I/O drawers, and the media drawer. Optional battery backups can provide energy for an emergency shutdown in case of a power failure. The POWER4 chip offers advanced microprocessors, with an SMP design, in a single silicon substrate.

Building on the IBM @server zSeries heritage, the pSeries 690 supports logical partitioning (LPAR). The Hardware Management Console (HMC) is used for defining resources for p690 LPARs. For more details about the IBM @server pSeries 690, refer to *IBM @server pSeries 690 System Handbook*, SG24-7040.

2.1 LPAR

The p690 features the logical partitioning (LPAR) that enables you to logically split one p690 system into separate sub-systems, each running its own instance of the operating system (AIX or Linux).

The p690 system can run in two possible modes:

- ▶ Full system partition
- ▶ LPAR

The Full system partition is the default configuration of the p690. It runs the system with one instance of the operating system. All the installed resources are available to the operating system.

The LPAR configuration of the p690 splits one p690 into several logically separate systems (LPARs). Every LPAR is running its own instance of the operating system and accesses only resources allocated to that particular LPAR. One resource cannot be allocated to more than one active LPAR. The allocation units of the partitions are based on 1 processor, 256 MB RAM units, and PCI slots.

Note: One PCI slot cannot be concurrently shared by different active LPARs in AIX 5L Version 5.1. If we want to assign, for example, disks to two different LPARs, then we need at least two separate I/O adapters, like SCSI, SSA, or FC.

The main controller logic is running in the Central Electronic Complex (CEC) unit and stores the LPAR configuration internally in the NVRAM non-volatile memory. Any system failure occurred in one LPAR instance does not influence other LPARs. Each time the p690 system is rebooted/initialized/restarted, the last configuration is used to set up the system.

For configuring a highly available cluster with p690 LPARs, you must consider the PCI slot allocation when defining the LPAR configuration. You must follow instructions regarding placement of adapters in the correct slots and you should configure these I/O slots during allocation of resources to particular LPARs. We have to consider the extended error handling (EEH) capabilities of the installed adapter.

Important: PCI adapter errors that occur on non-EEH capable adapters cannot be isolated and are propagated to the PCI host bridge (PHB). All adapters connected to this PHB may fail. This consideration is also important when planning HACMP clusters.

For detailed information on adapter placement, refer to the *PCI Adapter Placement Reference*, SA38-0538.

2.2 Hardware Management Console (HMC)

The HMC appears as a black-box computer that provides a graphical user interface (GUI) to configure and administer the hardware setup of the LPARs in the p690 system. The main purpose of the HMC is used to configure and administer LPARs and resource monitoring. The HMC stores the LPAR configuration, and a copy is maintained in the p690 CEC NVRAM.

2.3 Planning considerations

Many of the high availability features are already implemented in the base hardware design of the p690, so the planning of any p690 cluster consists of two main areas:

1. Understanding of the high availability design of p690 internal hardware RAS features
2. Design of the high availability clusters using p690 LPARs

The objective is to plan to eliminate single point of failures (SPOF) and focus on the HACMP high availability design, while keeping in mind the RAS features of the p690.

2.3.1 System configuration

In this section, we will describe the high availability planning considerations for hardware components.

p690 server

We stated earlier that the p690 server can be operated in a full system partition mode or in an LPAR mode. HACMP can be used for designing highly available clusters either in a full system partition mode or LPAR mode. In either case, it is seen as a separate node by the HACMP software. Hence, we need to understand the inherent RAS features of p690 hardware. The RAS features of the p690 are described in detail in the *IBM @server pSeries 690 Availability Best Practices* whitepaper, downloadable from:

http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_avail.html

The HACMP design adds and extends the availability features for the following events.

- ▶ Operating system failure
- ▶ Total system failure of the p690 rack
- ▶ System maintenance procedures, like upgrades or re-configurations
- ▶ Network and network adapter failures

From the HACMP point of view, we have to chose a basic layout of the cluster, which consists of the following:

- ▶ Basic cluster definitions
- ▶ Cluster nodes - an LPAR or entire p690
- ▶ Physical and logical networks
- ▶ Network adapters
- ▶ Disk adapters, connections, and layout
- ▶ Disk layout
- ▶ Other resources, like tape devices, WAN adapters and communication links, and so on
- ▶ Applications

Using LPAR in a highly available configuration

An LPAR configuration is a separate system and, as such, from the HACMP point of view, it could be one of the nodes to participate in the cluster. This means as for as HACMP is concerned, the cluster node can be an LPAR in the p690 or a standard pSeries server. However, there are few design considerations to note while selecting LPAR node in an high available clusters. They are:

- ▶ Clustering LPAR nodes on two or more separate p690 servers
- ▶ Clustering two or more LPARs within a p690 server
- ▶ Clustering LPARs on p690 server with any other pSeries servers
- ▶ Clustering LPARs that are part of a Cluster 1600 configuration

There may be several other possible ways LPAR nodes can be used in a highly available cluster configurations; however, we only discuss some of these scenarios in this chapter.

LPARs on two or more separate p690

The recommended configuration is clustering on separate p690 systems. This could be either the full system partition mode of p690 or in LPAR mode. This

configuration completely separates the cluster nodes and simplifies the cluster design while it gives the most flexibility to HACMP. From an HACMP point of view, each LPAR is a standard AIX node. This scenario is shown in Figure 2-1.

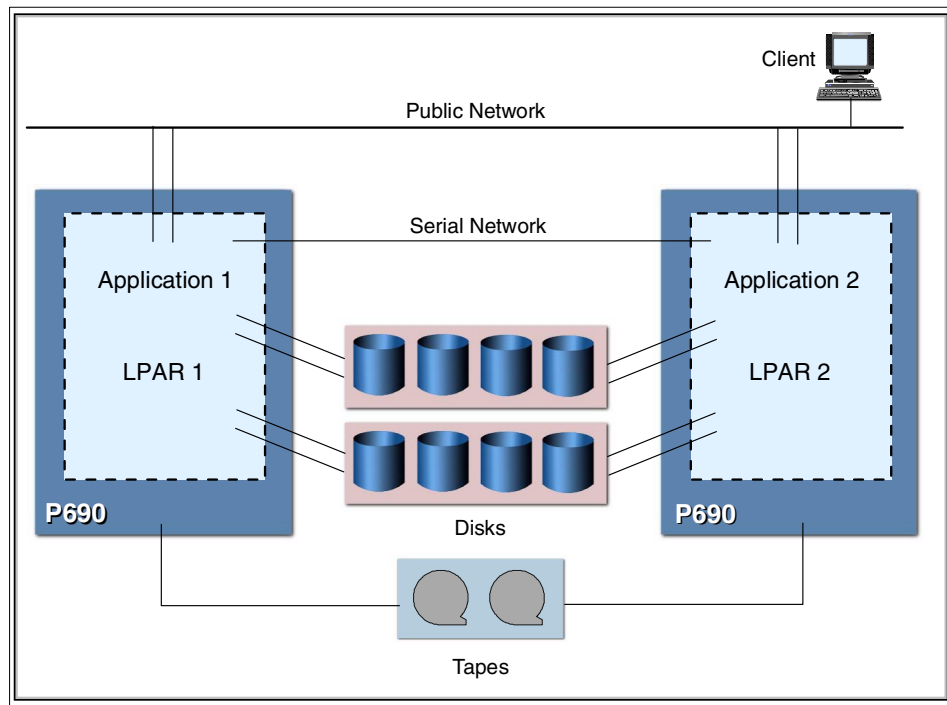


Figure 2-1 High availability cluster with one LPAR per p690 server

Two or more LPARs within the same p690

Since each LPAR is viewed as a separate node, you may consider HACMP clusters using the LPARs within one p690. While this configuration is similar to clustering two servers, this is not a recommended solution. This may be due to the fact that there may be a need to shut down the entire p690 server for, for example, firmware upgrades and maintenance procedures. However, this design provides the high availability for handling operating system failures, software upgrade procedures, application failover, and so on, within an LPAR. Also, in this configuration, the LPARs appear as standard AIX nodes from an HACMP point of view. This configuration is shown in Figure 2-2 on page 66.

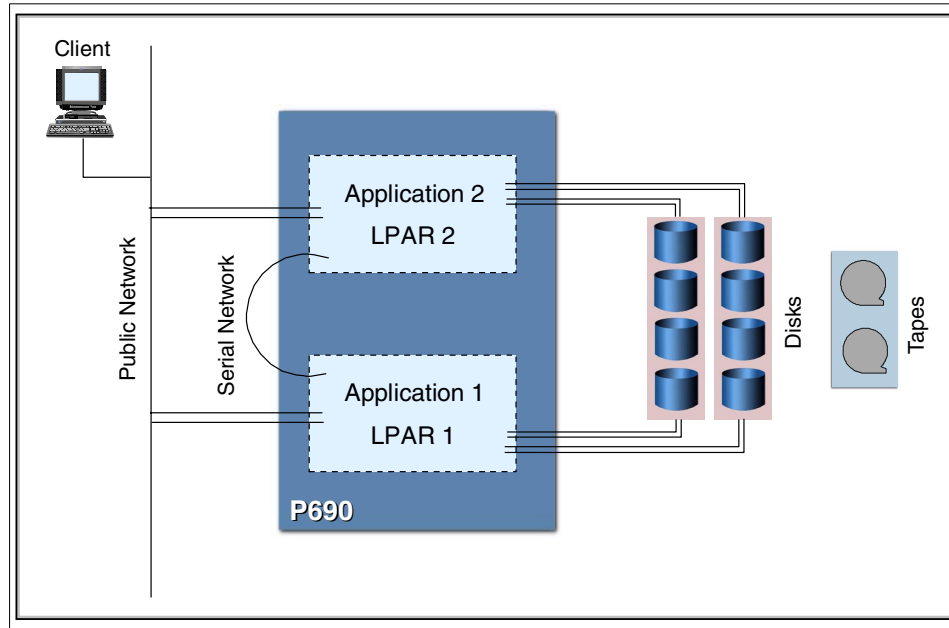


Figure 2-2 High availability cluster with two or more LPARs in one p690 server

LPARs on p690 with other pSeries systems

HACMP supports clustering of an LPAR node with other pSeries servers. From HACMP's point of view, the LPAR nodes are like any other pSeries server. In this case, you need to understand that the hardware configuration is not symmetrical, so you must consider performance, adapter support, and so on. This configuration is shown in Figure 2-3 on page 67.

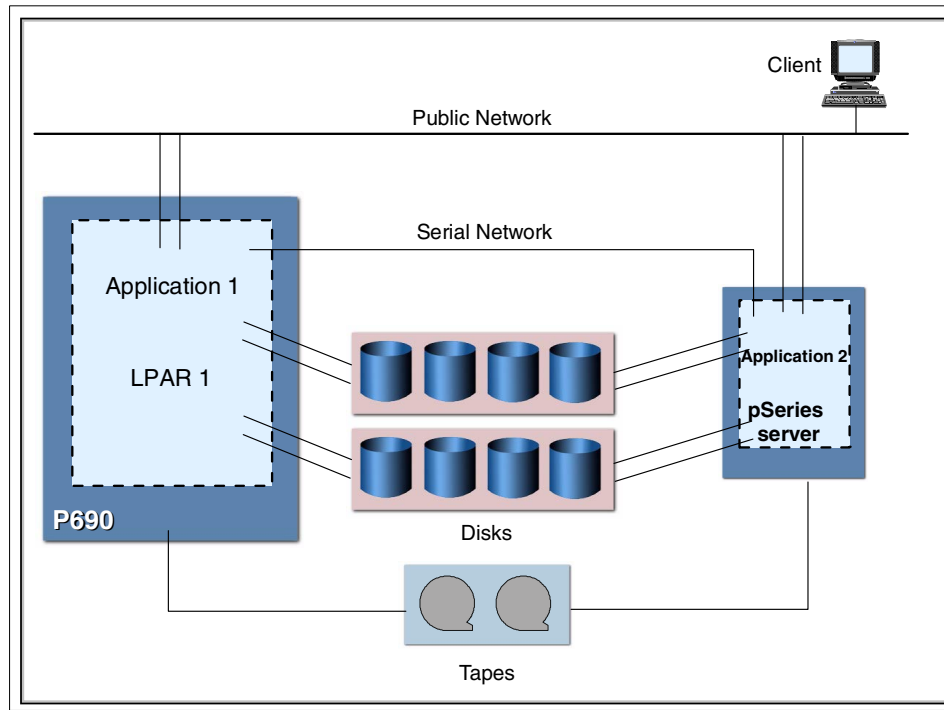


Figure 2-3 High availability cluster with one LPAR in a p690 and a pSeries server

Using LPARs that are part of a Cluster 1600 configuration

The Cluster 1600 environment is the high-end configuration of pSeries servers aimed to benefit the high performance of clustered servers. There are four main areas that must be considered when integrating any HACMP clusters into Cluster 1600:

- System management** The system management is centralized on the control workstation that is monitoring and controlling every node of the cluster. This function overlaps with some of the HACMP features. We must also understand the impact of the failure of this control workstation.
- SP Switch network** This network is used for high performance communication among cluster nodes. HACMP handles this network while considering its specific features.
- Cluster security** The Cluster 1600 implements the Kerberos 4 security structure by default. From HACMP point of view, we must handle the possible single point of failure of the Kerberos authentication server. To the other side, we

should integrate the Kerberos security into HACMP using the “enhanced security” of the HACMP.

RSCT Integration

The RSCT daemons, like the topology services, group services, or event management, are used by HACMP as well as by Cluster 1600. In the first release of HACMP 4.5, these daemons run separate for the HACMP and Cluster 1600.

This configuration is shown in Figure 2-4. We will focus more on this topic in Section 2.4.8, “Scenario 2: Using SP Switch/SP Switch2 adapter” on page 116 and Section 2.4.9, “Scenario 3: Dual SP Switch2 network” on page 128.

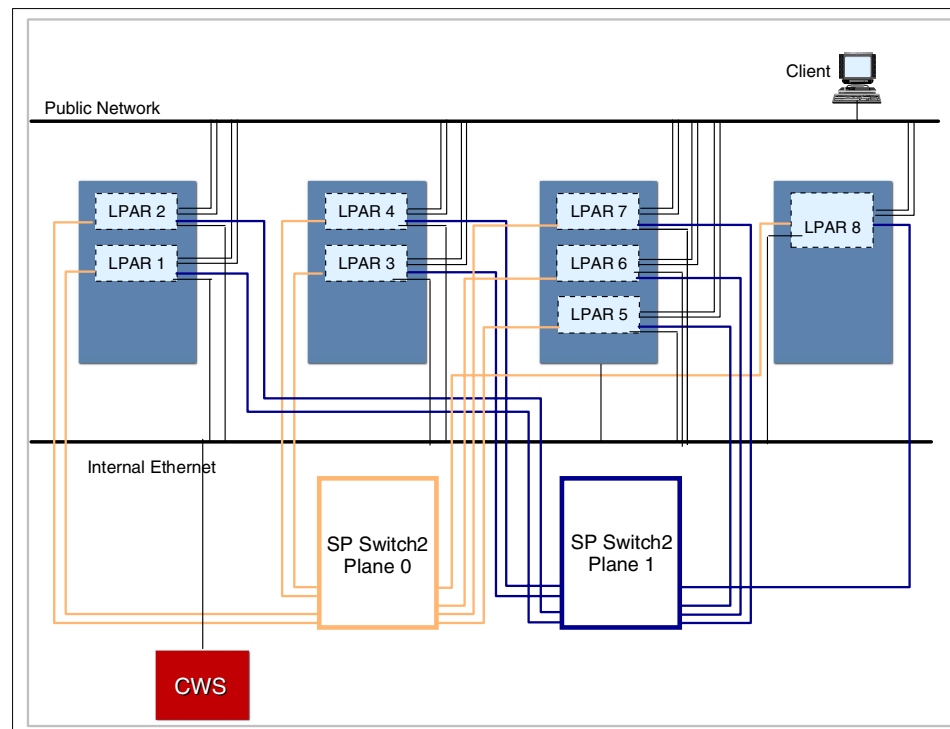


Figure 2-4 Highly available cluster with one LPAR in a p690 in a Cluster 1600

2.3.2 HMC high availability

To avoid having HMC become, potentially, a single point of failure, we recommend that you connect each p690 to two different HMCs. That way, in the event of an HMC failure, you are still capable of reaching p690 through the second HMC connection. Depending on how many p690 servers you have, additional 8-port async adapters may be required.

2.3.3 Network

HACMP configurations may require implementing networks between LPAR nodes. This section describes network considerations.

HMC network

Each LPAR of the p690 system and the HMC is interconnected by an Ethernet network (see *IBM @server pSeries 690 System Handbook*, SG24-7040). This network is used for system monitoring by HMC. For this reason, we do not recommend that you implement IP address takeover over this network interface. This network still can be defined for HACMP monitoring and we recommend this.

Cluster network

Each LPAR node should be connected to at least one public network for clients to connect to the nodes. For supported network adapters, please refer to the *HACMP for AIX 4.5 Planning Guide*, SC23-4277. The cluster network can be one of the following:

- ▶ A network with two adapters, one for boot/service and another for standby
- ▶ One network with one adapter only per node using the IP aliasing feature

For a cluster network, we recommend using two network adapters per node. We also recommend an additional private network, if you are planning for concurrent access configurations with a lock manager.

Serial network

Every cluster configuration must contain at least one serial network to send heartbeat messages through and monitor the node availability in order to eliminate a single point of failure of the IP protocol stack in the kernel. In an LPAR p690 configuration, there are no integrated serial ports available for configuring serial network. Hence, we suggest the following alternatives:

- ▶ Install an 8-port asynchronous adapter in each LPAR and use the RS232 ports to configure interconnect the LPARs.
- ▶ Configure target mode SSA between LPARs.
- ▶ Configure target mode SCSI between LPARs.

When configuring target mode SSA or SCSI, remember that large blocks of continuous data transfers to or from disks may arbitrate the SCSI bus or SSA loop for a relatively long time, and may produce `network_down` events. We recommend you use a dedicated SCSI bus or SSA loop for HACMP heartbeat monitoring through target mode devices and also enable AIX I/O pacing.

HACMP Cluster in a Cluster 1600

If the p690 is integrated into a Cluster 1600, you must have one Ethernet network dedicated for system management for PSSP software. The HMC monitoring can use this system management network and this network should not be handled for IP address takeover. In fact, we can use one network for both purposes, so the HMC network and the Cluster 1600 internal network is the same logical and physical network. This network should not be used for client communications and should be implemented on a separate physical network other than the HACMP public network, because of security reasons.

The high performance SP Switch network used in Cluster 1600 is a private network and is different from other networks. If an SP Switch/ network is used for an HACMP Cluster, we need to configure one SP Switch network adapter per LPAR, and we define one fixed address to the adapter for PSSP management, and the boot and service addresses are defined as IP aliases.

The SP Switch2 used in Cluster 1600 supports up to two SP Switch2 adapters per LPAR. If one SP Switch2 adapter is used for an HACMP Cluster, then one network address is used for PSSP management, and the boot and service addresses are defined as IP aliases. The second SP Switch2 adapter can be configured as standby. With two switch adapters, we can exploit the highly available features of the SP Switch2 by its device driver and also by HACMP. Example configurations are described in Section 2.4.8, “Scenario 2: Using SP Switch/SP Switch2 adapter” on page 116 and Section 2.4.9, “Scenario 3: Dual SP Switch2 network” on page 128.

2.3.4 Storage

While the nodes, networks, and adapters define the topology of the cluster, storage components are resources handled by HACMP event scripts using AIX Logical Volume Manager (LVM) commands and storage device driver utilities.

The storage devices in HACMP clusters can be:

- Shared
- Dedicated

Typical dedicated storage devices are the system disks in the rootvg volume group, where the AIX operating system resides. To eliminate single point of failure, use full mirroring of all disks in the rootvg.

The shared storage devices are used mainly to store the data that must be available to all nodes of the cluster that participate in the takeover. Every shared device must support the SCSI reservation and break independently for each

disk. In p690 environments, we recommend you implement the following shared disk subsystems:

- ▶ IBM Enterprise Storage Server (ESS, also known as Shark) for intensive I/O operations and flexibility in allocation of disk resources. An example configuration is described in Section 2.4.11, “Scenario 5: Integrating ESS storage into HACMP” on page 138.
- ▶ IBM SSA for standard disk use. The SSA disk subsystem features robustness and ease of use, while maintaining good performance. An example configuration is described in Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87.

To support these devices, make sure to define these adapters in the LPAR node configuration.

2.3.5 Software

The software prerequisites for a correct HACMP implementation are based on the following:

- ▶ The minimal AIX level required by p690 hardware is AIX 5L Version 5.1.0 with Maintenance Level 1 and APAR IY22854. The AIX software is delivered by ordering Feature Code 5765-C34. The Maintenance Level and APAR IY22854 is downloadable from <http://techsupport.services.ibm.com/>. The correct level of the AIX can be checked by the `oslevel` or `instfix` commands.
- ▶ RSCT is required at level 2.2.1.0. This software is delivered on the AIX 5L Version 5.1 distribution media. There may be RSCT updates in the future. The level of the RSCT can be checked by the `lslpp -l | grep rsct` command.
- ▶ HACMP is required to be installed at level 4.5 or higher.
- ▶ If the p690 server is intended to be part of a Cluster 1600 configuration, the PSSP software must be installed. The PSSP must be at level PSSP 3.4 or higher. For detailed information on PSSP installation in a Cluster 1600 environment, refer to the redbook *RS/6000 SP/Cluster: New Enhancements in PSSP 3.4*, SG24-6604.

Note: The p690 systems has usually only one CD-ROM drive per system, allocated to one LPAR, and it cannot be shared as a device among partitions.

To install HACMP software, you may consider one of the following:

- ▶ Allocate the CD-ROM device to LPARs for each installation. This requires reconfiguration of the LPARs.

- ▶ Allocate the CD-ROM to one LPAR and export its mount point through NFS. Other LPARs can mount the CD-ROM directory and use it as a standard NFS directory tree.
- ▶ Install the first LPAR directly from CD-ROM and configure it to be a Network Install Manager (NIM) server. Install other LPARs from this NIM server. AIX 5L Version 5.1 APAR IY22854 provides simple setup routines to configure the NIM environment.
- ▶ If the p690 server is part of Cluster 1600 environment, you may consider using the lppsource directory of CWS for installing the HACMP product.

2.4 Clustering with HACMP

During this project, we configured and tested various scenarios using two p690 servers with one LPAR each. Using these example scenarios, we describe the process of preparation, installation, and configuration of the HACMP cluster and testing. This section contains following scenarios:

- ▶ With two Ethernet adapters and SSA storage
- ▶ With one SP Switch2 adapter and SSA storage
- ▶ With two SP Switch2 adapters and SSA storage
- ▶ With IP aliasing over one Ethernet adapter
- ▶ With Ethernet adapter and ESS storage

2.4.1 Lab environment

The lab environment we used for testing HACMP 4.5 with p690 servers is in fact a Cluster 1600 configuration with p690 servers. The system configuration consists of two p690 servers, with one HMC and one control workstation. The Cluster configuration also consists of two SP Switch2 network adapters. The software environment consists of AIX 5L Version 5.1, PSSP 3.4, and HACMP 4.5.

2.4.2 System configuration

To test the scenarios, we configured following testing environment. It contains two p690 servers that have the same hardware configuration. In each server, we configured one LPAR for HACMP testing, but we did not use the other LPARs in the server. Figure 2-5 on page 73 shows our testing environment.

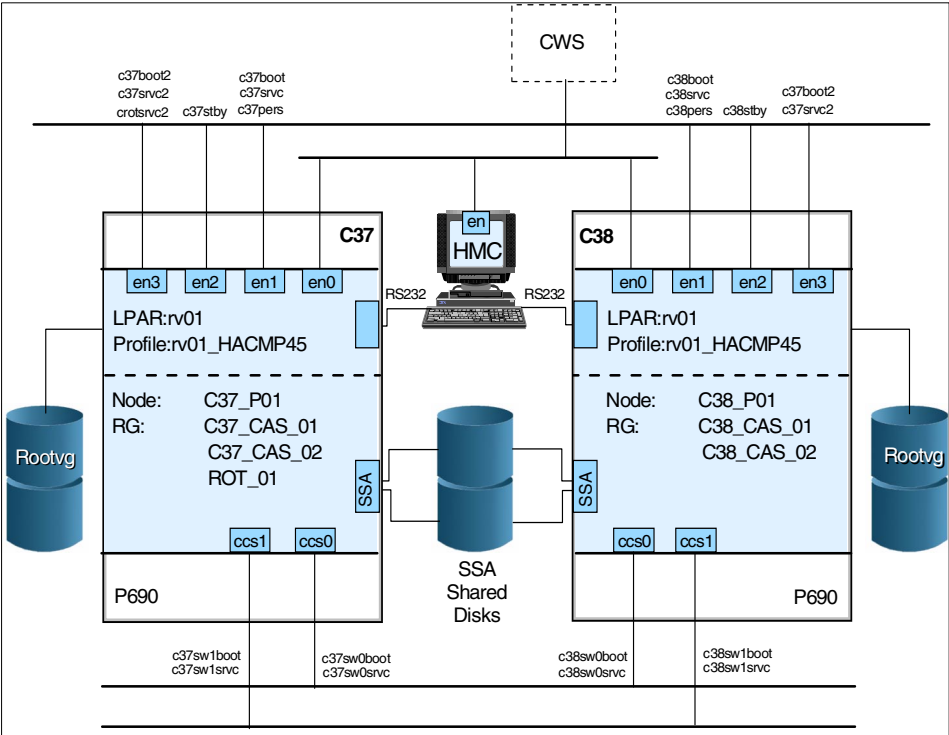


Figure 2-5 HW configuration of testing environment

Table 2-1 contains the complete hardware configuration of each p690. For our testing we used only one LPAR configuration only.

Table 2-1 p690 hardware resources

Resource	Configuration
CPU	16
Memory	32 GB
Ethernet Controllers	16
SP Switch 2 Controller	8
SSA Controller	1
SCSI Controller	16+1

We created two LPARs, one on each p690. Both LPARs have the same hardware configuration. We created an profile named HACMP45 for both LPARs using drawer U1.9.

The controllers are placed into four drawers. Table 2-2 shows the placement of controllers in the drawers.

Table 2-2 p690 drawer configuration

Slot	Drawers			
	U1.1	U1.5	U1.9	U1.13
P1-1	Ethernet	Ethernet	Ethernet	Ethernet
P1-2	Ethernet	Ethernet	Ethernet	Ethernet
P1-3	SP Switch2	SP Switch2	SP Switch2	SP Switch2
P1-4				
P1-5				
P1-6				
P1-7				
P1-8			SSA	
P1-9				
P1-10			SCSI	
P1-Z1	SCSI	SCSI	SCSI	SCSI
P1-Z2	SCSI	SCSI	SCSI	SCSI
P2-1	Ethernet	Ethernet	Ethernet	Ethernet
P2-2	Ethernet	Ethernet	Ethernet	Ethernet
P2-3	SP Switch2	SP Switch2	SP Switch2	SP Switch2
P2-4				
P2-5				
P2-6				
P2-7				
P2-8				
P2-9				

Slot	Drawers			
	U1.1	U1.5	U1.9	U1.13
P2-10				
P2-Z1	SCSI	SCSI	SCSI	SCSI
P2-Z2	SCSI	SCSI	SCSI	SCSI

Note: We recommend you use multiple drawers to avoid hardware SPOF.

Table 2-3 shows the requested resources for the LPAR. When an LPAR is activated using the HACMP45 profile, those resources are assigned.

Table 2-3 LPAR configuration

Resource	configuration
CPU	4
Memory	4GB
Ethernet Controllers	4
SP Switch 2 Controller	2
SSA Controller	1
SCSI Controller	2

2.4.3 Preparing the cluster for high availability

This section contains information on HACMP resources that we use in the described scenarios.

Cluster ID and name

We need to define the cluster ID and name for HACMP cluster. We assign the following data for our cluster configuration:

Cluster name:P690_2
Cluster ID:1

Nodes

We define two cluster nodes names, C37_P01 for one LPAR on the first p690 and C38_P01 for one LPAR on the second p690.

Adapters

Table 2-4 and Figure 2-5 on page 73 contain the planning information for IP addresses and IP labels for nodes C37_P01 and C38_P01, respectively.

Table 2-4 IP addresses for node C37_P01 - netmask 255.255.255.224

Type	IP address	IP label
Boot (1)	192.168.3.1	c37boot
Service (1)	192.168.3.11	c37srcv
Boot (2)	192.168.3.41	c37boot2
Service (2)	192.168.3.161	c37srcv2
Rotation	192.168.3.163	crotsrvc
Standby	192.168.3.71	c37stby
Persistent	192.168.3.131	c37pers
Boot for SP Switch2 (1)	192.168.13.1	c37sw0boot
Service for SP Switch2 (1)	192.168.13.11	c37sw0srcv
Boot for SP Switch2 (2)	192.168.13.231	c37sw1boot
Service for SP Switch2 (2)	192.168.13.241	c37sw1srcv

Table 2-5 IP addresses for node C38_P01 - netmask 255.255.255.224

Type	IP address	IP label
Boot (1)	192.168.3.2	c38boot
Service (1)	192.168.3.12	c38srcv
Boot (2)	192.168.3.42	c38boot2
Service (2)	192.168.3.162	c38srcv2
Standby	192.168.3.72	c38stby
Persistent	192.168.3.132	c38pers
Boot for SP Switch2 (1)	192.168.13.2	c38sw0boot
Service for SP Switch2 (1)	192.168.13.12	c38sw0srcv
Boot for SP Switch2 (2)	192.168.13.232	c38sw1boot
Service for SP Switch2 (2)	192.168.13.242	c38sw1srcv

Volume groups and file systems

The volume groups and all file systems are placed on shared SSA disks and configured on both nodes. Table 2-6 describes the volume group data.

Table 2-6 Volume groups and file systems

Node	Volume group	File systems
C37_P01	c37vg	/c37fs01 /c37fs02
C38_P01	c38vg	/c38fs01 /c38fs02

Application servers

The C37_APP_01 application server is a primary application for node C37_P01:

Start script: /usr/sbin/cluster/etc/scripts/start_c37app01.ksh

Stop script: /usr/sbin/cluster/etc/scripts/stop_c37app01.ksh

The C38_APP_01 application server is a primary application for node C38_P01:

Start script: /usr/sbin/cluster/etc/scripts/start_c38app01.ksh

Stop script: /usr/sbin/cluster/etc/scripts/stop_c38app01.ksh

Note: The scripts must exist on the both nodes and have execution permission.

Resource groups

The C37_CAS_01 resource group contains:

C37_P01 as primary node
C38_P01 as backup node
c37src01 service interface
c37vg volume group
/c37fs01 and /c37fs02 filesystems
C37_APP_01 application server

The C38_CAS_01 resource group contains:

C38_P01 as primary node
C37_P01 as backup node
c38src01 service interface
c38vg volume group
/c38fs01 and /c38fs02 filesystems
C38_APP_01 application server

2.4.4 Define the LPARs configuration on both p690s

We used HMC to define LPARs, to define LPAR's profiles, and to assign resources according to the cluster configuration. For more information on assigning resources for LPARs, refer to the *IBM @server pSeries 690 System Handbook*, SG24-7040. Figure 2-6 shows the LPAR configuration as seen from the HMC console.

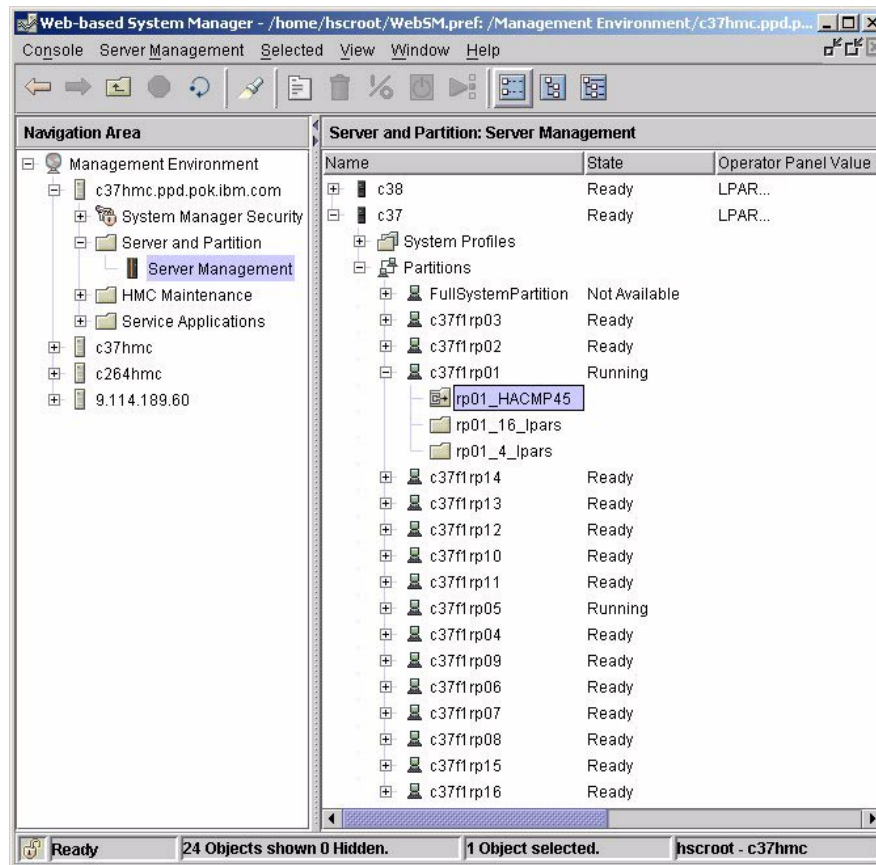


Figure 2-6 LPAR configuration as seen from the HMC console

Installation of AIX onto both LPARs

We assigned the CD-ROM to the LPAR, and the AIX operating system is installed using the standard AIX installation procedure from distribution CDs.

2.4.5 Configure the nodes

After the AIX installation, we need:

- ▶ To configure boot interfaces and standby interfaces on the both nodes
- ▶ To define tmssa devices for SSA serial link
- ▶ To configure shared volume groups and file systems

Note: To avoid a single point of failure (SPOF) we strongly recommend you mirror the system volume group **rootvg** and all shared volume groups, including all file system and logical volumes on each cluster node.

Configure TCP/IP onto both LPARs

On the TCPIP SMIT screen, select Minimum Configuration & Startup.

Configure the boot interface on the first node using the information from Table 2-4 on page 76 and Table 2-5 on page 76.

Choose interface en1 and fill in the following fields:

Internet ADDRESS (dotted decimal)	Enter the IP address for the boot adapter for the first node.
Network MASK (dotted decimal)	Enter the proper network mask for the chosen subnet.

Figure 2-7 on page 80 shows the SMIT screen for this step.

Minimum Configuration & Startup

To Delete existing configuration data, please use Further Configuration menus

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

[Entry Fields]

* HOSTNAME

[c37f1rp01]

* Internet ADDRESS (dotted decimal)

[192.168.3.1]

Network MASK (dotted decimal)

[255.255.255.224]

* Network INTERFACE

en1

NAMESERVER

Internet ADDRESS (dotted decimal)

DOMAIN Name

Default Gateway

Address (dotted decimal or symbolic name)

[9.114.189.62]

Cost

[0]

Do Active Dead Gateway Detection?

no

Your CABLE Type

N/A

START Now

no

#

+

+

+

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-7 SMIT TCPIP: Minimum Configuration & Startup

On the SMIT TCPIP screen, select **Further Configuration -> Network Interfaces -> Network Interface Selection -> Change / Show Characteristics of a Network Interface**.

Configure the boot interface on the first node using information given in Table 2-4 on page 76 and Table 2-5 on page 76.

Choose interface en2 and fill in or change the following fields:

- Internet ADDRESS (dotted decimal)

Enter the IP address for the standby adapter for the first node.
- Network MASK (dotted decimal)

Enter the proper network mask for the chosen subnet.
- Current STATE

Change the status to up.

Figure 2-8 on page 81 shows the SMIT screen for this step.

Change / Show a Standard Ethernet Interface

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

[Entry Fields]

Network Interface Nameen2

INTERNET ADDRESS (dotted decimal)[192.168.3.71]

Network MASK (hexadecimal or dotted decimal)[255.255.255.224]

Current STATEup+

Use Address Resolution Protocol (ARP)?yes+

BROADCAST ADDRESS (dotted decimal)

F1=HelpF2=RefreshF3=CancelF4=List

F5=ResetF6=CommandF7=EditF8=Image

F9=ShellF10=ExitEnter=Do

Figure 2-8 SMIT TCPIP: Change/Show a Standard Ethernet Interface

To configure the second boot adapter, we use the same SMIT screen using the en3 interface and the appropriate IP address.

For the second node, we configure all the adapters using the same SMIT screens with information from the worksheets.

Configure devices for the SSA serial link onto the both LPARs

We want to use the serial network for sending keepalive packets using target mode SSA. For this purpose, we require AIX fileset devices.ssa.tm.rte. It must be installed on both nodes. We can verify this using the `lslpp` command. If the fileset is not installed, we should install it from AIX using `smit install` fast path or the `installp` command. Example 2-1 shows you how to do it.

Example 2-1 Check and install fileset for Target Mode SSA

```
[c37f1rp01]:/ # lslpp -l devices.ssa.tm.rte
lslpp: 0504-132 Fileset devices.ssa.tm.rte not installed.
[c37f1rp01]:/ # installp -aQcNgXd '/usr/sys/inst.images' devices.ssa.tm
...
Installation Summary
-----
Name                                Level      Part      Event      Result
-----
devices.ssa.tm.rte                  5.1.0.10   USR        APPLY      SUCCESS
devices.ssa.tm.rte                  5.1.0.10   ROOT       APPLY      SUCCESS
[c37f1rp01]:/ # lslpp -l devices.ssa.tm.rte
Fileset                             Level     State      Description
-----
Path: /usr/lib/objrepos
```

devices.ssa.tm.rte	5.1.0.10	COMMITTED	Target Mode SSA Support
Path: /etc/objrepos			
devices.ssa.tm.rte	5.1.0.10	COMMITTED	Target Mode SSA Support

The `node_number` attribute must be unique and different for ssar devices on both nodes. The default value is zero. Change the `node_number` attribute using the **chdev -l ssar -a node_number=<value>** command and run the **cfgmgr** command.

For example:

- On the first node: **chdev -l ssar -a node_number=38;cfgmgr**
- On the second node: **chdev -l ssar -a node_number=37;cfgmgr**

You must run **cfgmgr** again on the first node to complete the device's configuration.

Check devices on both systems using the **lsdev -Cctmssa** command.

Example 2-2 contains a step-by-step procedure to configure target mode SSA devices.

Example 2-2 Configure tmssa devices

On first node:

```
[c37f1rp01]:/ # lsattr -El ssar
node_number 0 SSA Network node number True
[c37f1rp01]:/ # chdev -l ssar -a node_number=38
node_number 38 SSA Network node number True
[c37f1rp01]:/ # cfgmgr
[c37f1rp01]:/ # lsdev -C | grep ssa
ssa0          Available 31-08          IBM SSA 160 SerialRAID Adapter (14109100)
ssar          Defined                  SSA Adapter Router
tmssar        Available                Target Mode SSA Router
```

On second node:

```
[c38f1rp01]:/ # lsattr -El ssar
node_number 0 SSA Network node number True
[c38f1rp01]:/ # chdev -l ssar -a node_number=37
node_number 37 SSA Network node number True
[c38f1rp01]:/ # cfgmgr
[c38f1rp01]:/ # lsdev -C | grep ssa
ssa0          Available 31-08          IBM SSA 160 SerialRAID Adapter (14109100)
ssar          Defined                  SSA Adapter Router
tmssar        Available                Target Mode SSA Router
tmssa38       Available                Target Mode SSA Device
```



```
On first node:
[c37f1rp01]:/ # cfmgr
[c37f1rp01]:/ # lsdev -C | grep ssa
ssa0      Available 31-08      IBM SSA 160 SerialRAID Adapter (14109100)
ssar      Defined          SSA Adapter Router
tmssar    Available          Target Mode SSA Router
tmssa37   Available          Target Mode SSA Device
```

For more information on this topic, refer to the *HACMP for AIX 4.4.1 Enhanced Scalability Installation and Administration Guide*, SC23-4306 or *HACMP for AIX 4.4.1 Installation Guide*, SG23-4278.

Configure shared volume groups and file systems

According to Table 2-4 on page 76 and Table 2-5 on page 76, we need to create two shared volume groups and four file systems and configure them on both nodes. All configuration steps we perform on the first node can be imported to the second node from the first node.

We have 16 SSA disks, eight per loop. Table 2-7 shows which disk belongs to which loop.

Table 2-7 SSA disk configuration

SSA loop	hdisk numbers
A	9, 10, 11, 12, 13, 14, 15, and 16
B	1, 2, 3, 4, 5, 6, 7, and 8

The first volume group, named s37vg, contains disks number 7, 8, 9, and 10. The second one, named s38vg, contains disks number 5, 6, 11, and 12. Both volume groups contain two disks from each loop to avoid SPOF and for mirroring file systems. The volume groups have to have the auto-varyon and quorum options turned off.

s37vg is spread out between two mirrored file systems, /s37fs01 and /s37fs02. The volume group s37vg and its file systems belong to the C37_CAS_01 resource group.

s38vg is also spread out between two mirrored file systems, /s38fs01 and /s38fs02. The volume group s38vg and its file systems belong to the C38_CAS_01 resource group.

Example 2-3 on page 84 shows the configuration and import procedure using an AIX command.

Example 2-3 Create volume groups and file systems across the cluster nodes

On first node:

```
[c37f1rp01]:/ # mkvg -y s37vg -s 64 -V 39 hdisk7 hdisk8 hdisk9 hdisk10
[c37f1rp01]:/ # mklv -c 2 -y s37lv01 s37vg 10 hdisk7 hdisk9
[c37f1rp01]:/ # mklv -c 2 -y s37lv02 s37vg 10 hdisk8 hdisk10
[c37f1rp01]:/ # crfs -v jfs -d s37lv01 -m /s37fs01
[c37f1rp01]:/ # crfs -v jfs -d s37lv02 -m /s37fs02
[c37f1rp01]:/ # chvg -a n -Q n s37vg
[c37f1rp01]:/ # varyoffvg s37vg
[c37f1rp01]:/ # mkvg -y s38vg -s 64 -V 40 hdisk5 hdisk6 hdisk11 hdisk12
[c37f1rp01]:/ # mklv -c 2 -y s38lv01 s38vg 10 hdisk5 hdisk11
[c37f1rp01]:/ # mklv -c 2 -y s38lv02 s38vg 10 hdisk6 hdisk12
[c37f1rp01]:/ # crfs -v jfs -d s38lv01 -m /s38fs01
[c37f1rp01]:/ # crfs -v jfs -d s38lv02 -m /s38fs02
[c37f1rp01]:/ # chvg -a n -Q n s38vg
[c37f1rp01]:/ # varyoffvg s38vg
```

On second node:

```
[c38f1rp01]:/ # cfgmgr
[c38f1rp01]:/ # importvg -y s37vg -V 39 hdisk7
[c38f1rp01]:/ # chvg -a n -Q n s37vg
[c38f1rp01]:/ # varyoffvg s37vg
[c38f1rp01]:/ # importvg -y s38vg -V 40 hdisk5
[c38f1rp01]:/ # chvg -a n -Q n s38vg
[c38f1rp01]:/ # varyoffvg s38vg
```

For more details, refer to *AIX 5L Version 5.1 System Management Guide: Operating System and Devices* and *HACMP for AIX 4.4.1 Enhanced Scalability Installation and Administration Guide*, SC23-4306.

2.4.6 Installing HACMP

HACMP installation is a standard installation process using the **installp** command. It is very well documented in *HACMP for AIX 4.4.1 Enhanced Scalability Installation and Administration Guide*, SC23-4306 or *HACMP for AIX 4.4.1 Installation Guide*, SG23-4278. This section describes important information on how to prepare and perform HACMP installation.

Prerequisites

The HACMP/ES software has the following prerequisites:

- ▶ Each cluster node must have AIX 5L Version 5.1.0.10 installed.
- ▶ Version 3 Release 4 of the PSSP (AIX Parallel System Support Programs) or greater must be installed on the SP control workstation and SP nodes.

- ▶ The RS/6000 Cluster Technology (RSCT) images must be installed on each node prior to installing HACMP/ES. These are now included in AIX 5L Version 5.1 rather than in the HACMP/ES software. RSCT 2.2.1.0 or higher is required.
- ▶ The following AIX optional bos components are mandatory for HACMP/ES:
bos.adt.lib, bos.adt.libm, bos.adt.syscalls, bos.net.tcp.client,
bos.net.tcp.server, bos.rte.SRC, bos.rte.libc, bos.rte.libcfg, bos.rte.libcur,
bos.rte.libpthreads, and bos.rte.odm.
- ▶ If you are installing Concurrent Resource Manager, you have
bos.rte.lvm.usr5.1.0.25 or higher and bos.clvm.enh.
- ▶ The /usr directory must have 52 MB of free space for a full install (for optional software, you can plan for more space).
- ▶ The / (root) directory must have 500 KB of free space (beyond any need to extend the /usr directory).

Note:

- ▶ The root user must perform the installation.
- ▶ You must accept the license agreement as you install.
- ▶ Each cluster node requires its own HACMP/ES software license.

Installation choices

To install the HACMP/ES software on all the nodes, you have the following choices:

- ▶ Installing from an installation server
- ▶ Installing from the installation media
- ▶ Installing from a disk

We used the last option. The HACMP/ES code was saved on the /usr/sys/inst.images directory. For the second node installation, we NFS-mounted this directory from the first node. Figure 2-9 on page 86 shows the SMIT installation screen that is used to perform the HACMP installation.

Install Software			
Type or select values in entry fields. Press Enter AFTER making all desired changes.			
	[Entry Fields]		
* INPUT device / directory for software	/usr/sys/inst.images		
* SOFTWARE to install	[_all_latest]		+
PREVIEW only? (install operation will NOT occur)	no		+
COMMIT software updates?	yes		+
SAVE replaced files?	no		+
AUTOMATICALLY install requisite software?	yes		+
EXTEND file systems if space needed?	yes		+
OVERWRITE same or newer versions?	no		+
VERIFY install and check file sizes?	no		+
Include corresponding LANGUAGE filesets?	yes		+
DETAILED output?	no		+
Process multiple volumes?	yes		+
ACCEPT new license agreements?	yes		+
Preview new LICENSE agreements?	no		+
F1=Help	F2=Refresh	F3=Cancel	F4=List
F5=Reset	F6=Command	F7=Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Figure 2-9 SMIT: Install Software

On the SMIT screen, select **Software Installation and Maintenance -> Install and Update Software -> Install Software** or use the `smit install_latest` fast path.

Enter the path that contains HACMP/ES software (in our case, /usr/sys/inst.images)

Change the following fields:

SOFTWARE to install

Leave as default if you want to install all filesets or press F4 if you want to install only several components.

Note: Usually, the “SOFTWARE to install” field is used to install a products default value. But the HACMP/ES distribution contains parts for NetView and Tivoli. If those products are not installed, the HACMP/ES component for those products will fail due to certain prerequisites. We recommend that you choose the filesets from the list.

Choose the software to install from the following list:

- ▶ cluster.adt.es
- ▶ cluster.doc.en_US.es
- ▶ cluster.es
- ▶ cluster.es.clvm
- ▶ cluster.es.spoc
- ▶ cluster.es.plugins
- ▶ cluster.license
- ▶ cluster.man.en_US.es

ACCEPT new license agreements?

Change to yes if you accepted the licence agreement; otherwise, the installation fails.

For more details, refer to *HACMP for AIX 4.4.1 Enhanced Scalability Installation and Administration Guide*, SC23-4306 or *HACMP for AIX 4.4.1 Installation Guide*, SG23-4278.

2.4.7 Scenario 1: Cluster with 2 Ethernet and SSA storage

Figure 2-10 on page 88 shows the cluster environment for this scenario and the resources we have to configure.

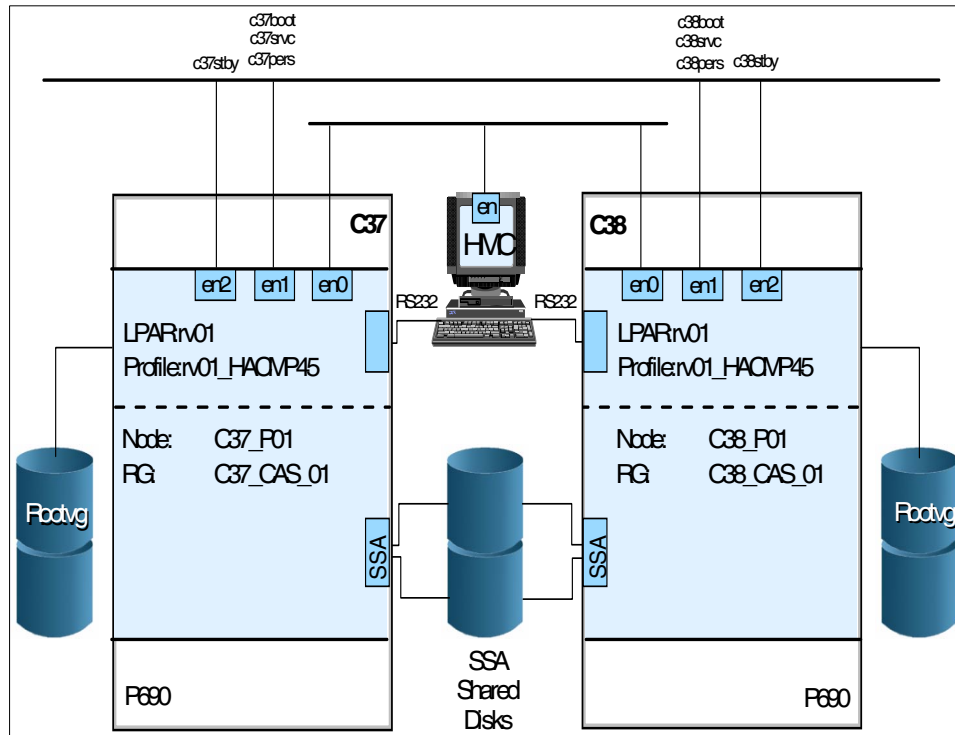


Figure 2-10 An example of a cluster for scenario 1

Description and assumptions

In this scenario, we will configure two node clusters as mutual takeover HACMP clusters, including IP Address Takeover (IPAT) and Hardware Address Takeover (HWAT).

We assume the following items in this configuration:

- ▶ Two nodes with:
 - Two boot labels
 - Two service labels
 - Two standby labels
 - Two persistent labels
- ▶ Serial link using target mode SSA
- ▶ Two shared volume groups with four mirrored file systems
- ▶ Two cascading resource groups with two application servers

We assume following for this test:

► Node_failure (takeover)

We expect that:

- The IP service label will go to the backup node.
- The volume group and all file systems will go to the backup node.
- Application 1 will restart on the backup node.

► Network adapter failure (swap_adapter)

We expect that the IP service and persistent label will go to the standby adapter on the service node.

Resources

According to Section 2.4.3, “Preparing the cluster for high availability” on page 75, we have to use the resources represented in Table 2-8, Table 2-9, Table 2-10 on page 90.

Table 2-8 Cluster definition - scenario 1

Name	Value
Cluster ID	1
Cluster Name	P690_2
Nodes Name	C37_P690 C38_P690

Table 2-9 Cluster Topology definition - scenario 1

Name	Node Name	
	C37_P690	C38_P690
Boot adapter	c37boot	c38boot
Service adapter	c37srcv	c38srcv
Standby adapter	c37stby	c38stby
Persistent adapter	c37pers	c38pers
Resource group	C37_CAS_01	C38_CAS_01

Note: We add all the adapters to the PUBL_NET_1 public network definition.

Table 2-10 Cluster Resources definition - scenario 1

Name	Resource Group Name	
	C37_CAS_01	C38_CAS_01
Participating nodes	C37_P690 (primary) C38_P690 (backup)	C38_P690 (primary) C37_P690 (backup)
Service addresses	c37svc	c38svc
File systems	/c37fs01 /c37fs02	/c38fs01 /c38fs02
Volume groups	c37vg	c38vg
Application servers	C37_APP_01	C38_APP_01

Configuration steps

The HACMP cluster configuration has two main steps. The first is to configure the cluster topology. The second is to configure the cluster resources.

Define the cluster topology

You only need to perform these steps on one node. The definition will be imported to the other nodes during the synchronization of the cluster topology function.

Complete the following steps to define the cluster topology:

1. Give the cluster a name and an ID.

The cluster name and ID must be unique.

On the SMIT HACMP screen (see Figure 2-11 on page 91), select **Cluster Configuration -> Cluster Topology -> Configure Cluster -> Add a Cluster Definition**.

Enter the following data:

Cluster ID	Enter an unique positive integer in the range 1-999999.
Cluster Name	Enter ASCII text string up to 32 characters (space is not allowed).

Add a Cluster Definition

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

[Entry Fields]

**NOTE: Cluster Manager MUST BE RESTARTED
in order for changes to be acknowledged.**

* Cluster ID

[1]

#

* Cluster Name

[P690_2]

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-11 Add a Cluster Definition

2. Define the nodes.
- On the SMIT HACMP screen (see Figure 2-12), select **Cluster Configuration -> Cluster Topology -> Configure Nodes -> Add Cluster Nodes**.
- Fill in the following field:
- Node Names**

Enter all node's names separate by space; there must be an unique ASCII text string up to 32 characters for each node.

Add Cluster Nodes

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

[Entry Fields]

* Node Names

[C37_P690 C38_P690]

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-12 Add a Cluster Nodes

3. Configure the adapters.
- Configure the boot interface for the first node.

On the SMIT HACMP screen (see Figure 2-13), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Choose Add an Adapter on a new Network.

For the first node, fill in the following fields:

IP Label	Press F4 and choose the boot interface for the first node.
Network Type	Press F4 and choose the ether type of network.
Network Name	Enter the symbolic name of the network.
Network Attribute	Leave as default (public).
Interface Function	Leave as default (boot).
Interface IP Address	Leave blank; HACMP will take the IP address from /etc/hosts, according to the IP label.
Node Name	Press F4 and choose the related node name.
Netmask	Enter the netmask for the interface.

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[c37boot]	+
* Network Type	[ether]	+
Network Name	[PUBL_ENT_1]	+
* Network Attribute	[public]	+
* Interface Function	[boot]	+
Interface IP Address	[]	
* Node Name	[C37_P690]	+
Netmask	[255.255.255.224]	+

F1=Help
F5=Reset
F9=Shell

F2=Refresh
F6=Command
F10=Exit

F3=Cancel
F7=Edit
Enter=Do

F4=List
F8=Image

Figure 2-13 Add an Initial Interface

- Configure the boot interface for the second node.

On the SMIT HACMP screen (see Figure 2-14 on page 93), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Choose PUBL_NET_1 (), which was defined in the previous step.

For the second node, enter the following data:

IP Label	Press F4 and choose the boot interface for the second node.
Interface Function	Leave as default (boot).
Interface IP Address	Leave blank; HACMP will take the IP address from /etc/hosts, according to the IP label.
Node Name	Press F4 and choose the related node name.
Netmask	Enter the netmask for the interface.

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[c38boot]	+
Network Type	ether	
Network Name	PUBL_NET_1	
* Interface Function	[boot]	+
Interface IP Address	[]	
* Node Name	[C38_P690]	+
Netmask	[255.255.255.224]	+

F1=Help

F5=Reset

F9=Shell

F2=Refresh

F6=Command

F10=Exit

F3=Cancel

F7=Edit

Enter=Do

F4=List

F8=Image

Figure 2-14 Add an Initial Interface

- Perform Discover IP Topology to gather information about the current cluster topology.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Discover IP Topology**.

Discovered information is saved in the /usr/sbin/cluster/etc/config/clip_config file and is required by the next steps.

- Configure standby adapters for both nodes.

On the SMIT HACMP screen (see Figure 2-15 on page 94), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Multiple IP-based Interface**.

Choose ether.

Fill in the following fields:

Interface IP Label(s) Press F4 and choose standby interfaces for both nodes using the F7 highlight choices.

Network Name Press F4 and choose the network name.

Interface Function Press F4 and choose standby.

Add Multiple IP-based Interfaces

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

	[Entry Fields]	
* Interface IP Label(s)	[c37stby]	> +
Network Name	[PUBL_NET_1]	+
Network Type	ether	
Interface Function	[standby]	+
Network Attribute	[public]	+

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

Figure 2-15 Add Multiple IP-based Interfaces

- Configure the service adapters for both nodes.

On the SMIT HACMP screen (see Figure 2-16 on page 95), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Multiple Service IP Labels to a Network**.

Choose PUBL_NET_1.

Enter the following data:

IP Label(s) Press F4 and choose standby interfaces for both nodes using the F7 highlight choices.

Add Multiple Service IP Labels to a Network

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

	[Entry Fields]	
* IP Label(s)	[c37src	> +
Network Type	ether	
Network Attribute	public	
Network Name	PUBL_NET_1	
IP Label Function	service	

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-16 Add Multiple Service IP Labels to a Network

- Bind the service adapters with the related nodes and assign the alternate HW addresses.

On the SMIT HACMP screen (see Figure 2-17 on page 96), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Change / Show an Interface / IP Label**.

Choose the service address for the first node.

Enter the following data for the first service adapter:

Hardware Address	Enter an alternate HW address for HW address Takeover.
Node Name	Press F4 and choose the related node name.

Change / Show an Interface / IP Label

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

	[Entry Fields]	
IP Label	c37srcv	
New IP Label	[]	+
* Network Type	[ether]	+
* Network Name	[PUBL_NET_1]	+
* Network Attribute	[public]	+
* Interface / IP-Label Function	[service]	+
IP Address	[192.168.3.11]	
Hardware Address	[0002556A3737]	
Node Name	[C37_P690]	+
Netmask	[255.255.255.224]	+
Interface Name		

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-17 Change / Show an Interface / IP Label

Use the same SMIT screen (see Figure 2-18 on page 97) to enter the data for the second service adapter.

Change / Show an Interface / IP Label			
Type or select values in entry fields. Press Enter AFTER making all desired changes.			
	[Entry Fields]		
IP Label	c38srcv		
New IP Label	[]		+
* Network Type	[ether]		+
* Network Name	[PUBL_NET_1]		+
* Network Attribute	[public]		+
* Interface / IP-Label Function	[service]		+
IP Address	[192.168.3.12]		
Hardware Address	[0x0002556a3838]		
Node Name	[C38_P690]		+
Netmask	[255.255.255.224]		+
Interface Name			
F1=Help	F2=Refresh	F3=Cancel	F4=List
F5=Reset	F6=Command	F7=Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Figure 2-18 Change / Show an Interface / IP Label

- Configure persistent adapters for both nodes.

On the SMIT HACMP screen (see Figure 2-19 on page 98), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add IP Labels Requiring Individual Configuration**.

Choose PUBL_NET_1.

Enter the following data:

IP Label	Press F4 and choose persistent interface.
IP Label Function	Press F4 and choose persistent.
IP Address	Leave blank; HACMP will take the IP address from /etc/hosts, according to the IP label.
Hardware Address	Leave blank.
Node Name	Press F4 and choose the related node name.
Netmask	Enter the proper netmask for the interface.

Add IP Labels Requiring Individual Configuration

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

	[Entry Fields]	
* IP Label	[c37pers]	+
Network Type	ether	
Network Name	PUBL_NET_1	
* IP Label Function	[persistent]	+
IP Address	[]	
Hardware Address	[]	
Node Name	[C37_P690]	+
Netmask	[255.255.255.224]	+

F1=Help
F2=Refresh
F3=Cancel
F4=List

F5=Reset
F6=Command
F7=Edit
F8=Image

F9=Shell
F10=Exit
Enter=Do

Figure 2-19 Add IP Labels Requiring Individual Configuration

Use the same SMIT screen (see Figure 2-20) to enter the data for the second persistent adapter.

Add IP Labels Requiring Individual Configuration

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

	[Entry Fields]	
* IP Label	[c38pers]	+
Network Type	ether	
Network Name	PUBL_NET_1	
* IP Label Function	[persistent]	+
IP Address	[]	
Hardware Address	[]	
Node Name	[C38_P690]	+
Netmask	[255.255.255.224]	+

F1=Help
F2=Refresh
F3=Cancel
F4=List

F5=Reset
F6=Command
F7=Edit
F8=Image

Figure 2-20 Add IP Labels Requiring Individual Configuration

- Configure serial adapters for both nodes using target mode SSA.
- On the SMIT HACMP screen (see Figure 2-21 on page 99), select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure Adapters on Non IP-based networks -> Add an Adapter.**

Choose Add an Adapter on a new Network.

For the first adapter, enter the following data:

Adapter Label	Enter the symbolic name of the serial adapter.
Network Type	Press F4 and choose the type of serial network.
Network Name	Enter the symbolic name of the serial network.
Device Name	Enter the AIX device name for the serial adapter.
Node Name	Press F4 and choose the related node.

Add a Non IP-based Adapter

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

	[Entry Fields]	
* Adapter Label	[tmssa37]	
Network Type	[tmssa]	+
* Network Name	[TM_SSA_1]	+
* Device Name	[/dev/tmssa37]	
* Node Name	[C37_P690]	+

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

Figure 2-21 Add a Non IP-based Adapter

Use the same SMIT screen to enter the data for the second node.

On the SMIT HACMP screen (see Figure 2-22 on page 100), select
Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure Adapters on Non IP-based networks -> Add an Adapter.

Choose TM_SSA_1.

For the first adapter, enter the following data:

Adapter Label	Enter the symbolic name of the serial adapter.
Device Name	Enter the AIX device name for the serial adapter.
Node Name	Press F4 and choose the related node.

Add a Non IP-based Adapter

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

* Adapter Label

Network Type

Network Name

* Device Name

* Node Name

[tmssa38]

tmssa

TM_SSA_1

[/dev/tmssa38]

[C38_P690]

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-22 Add a Non IP-based Adapter

4. Synchronize cluster topology

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Synchronize Cluster Topology**.

Press the Enter key to run cluster verification and synchronization. The cluster synchronization process will be done when the verification process passes without errors.

Define the Cluster resources

Follow these steps to define the Cluster resources:

1. Configure resource groups for the both nodes.

On the SMIT HACMP screen (see Figure 2-23 on page 101), select **Cluster Configuration -> Cluster Resources -> Define Resource Groups -> Add a Resource Group**.

For the first resource group, enter the following data:

Resource Group Name	Enter the resource group name (up to 32 characters).
Node Relationship	Leave as default (cascading).
Site Relationship	Leave as default (ignore).
Participating Node Names / Default Node Priority	Press F4 and choose the node which participates with the resource group. The default node priority is decreases from left to right.

Add a Resource Group

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

[Entry Fields]			
* Resource Group Name	[C37_CAS_01]		
* Node Relationship	cascading		+
* Site Relationship	ignore		+
* Participating Node Names / Default Node Priority	[C37_P690 C38_P690]		+

F1=Help
F5=Reset
F9=Shell

F2=Refresh
F6=Command
F10=Exit

F3=Cancel
F7=Edit
Enter=Do

F4=List
F8=Image

Figure 2-23 Add a Resource Group

Use the same SMIT screen (see Figure 2-24) to enter the data for the second resource group. Please to pay attention to the order in Participating Node Name; it is the default method of node priority.

Add a Resource Group

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

[Entry Fields]			
* Resource Group Name	[C38_CAS_01]		
* Node Relationship	cascading		+
* Site Relationship	ignore		+
* Participating Node Names / Default Node Priority	[C38_P690 C37_P690]		+

F1=Help
F5=Reset
F9=Shell

F2=Refresh
F6=Command
F10=Exit

F3=Cancel
F7=Edit
Enter=Do

F4=List
F8=Image

Figure 2-24 Add a Resource Group

2. Configure application servers for the both nodes.

On the SMIT HACMP screen (see Figure 2-25 on page 102), select **Cluster Configuration -> Cluster Resources -> Define Application Servers -> Add an Application Server**.

For the first application server, fill in the following fields:

Server Name	Enter the symbolic name of the application server.
Start Script	Enter the start script that is used when the node joins the cluster. (The full path is required.)

Stop Script

Enter the stop script that is used when the node leaves the cluster.

Note: The scripts should have executable permissions, be entered with an absolute path, and must exist on both nodes in the same path.

Add an Application Server

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

* Server Name

* Start Script

* Stop Script

[Entry Fields]

[C37_APP_01]

<pts/start_c37app01.ksh]

[/usr/sbin/cluster/etc/>

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-25 Add an Application Server

Use the same SMIT screen (see Figure 2-26) to enter the data for the second application server.

Add an Application Server

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

* Server Name

* Start Script

* Stop Script

[Entry Fields]

[C38_APP_01]

<pts/start_c38app01.ksh]

[/usr/sbin/cluster/etc/>

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-26 Add an Application Server

3. Perform discovery of the volume groups to gather information about current disk configuration on the both nodes.
Collect volume groups data from the local node.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Resources -> Discover Current Volume Group Configuration -> Local Configuration**.

Collect the shared volume groups data from both nodes.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Resources -> Discover Current Volume Group Configuration -> Cluster-wide Configuration**.

Discovered information is saved in the `/usr/sbin/cluster/etc/config/clip_config` file and is required by the next steps.

Note: We recommend you do local configuration first to collect information from the local node (volume groups created in previous section), then do cluster-wide configuration to collect information from others nodes in the cluster.

4. Define and assign resources into resource groups for both nodes.

On the SMIT HACMP screen (see Figure 2-27 on page 105), select **Cluster Configuration -> Cluster Resources -> Change/Show Resources/Attributes for a Resource Group**.

Choose C37_CAS_01.

For the first application server, fill in the following fields:

Dynamic Node Priority	Leave blank.
Service IP label	Press F4 and choose the appropriate service adapter for the resource group.
Filesystems (default is All)	Press F4 and choose the appropriate file systems for the resource group.
Filesystems Consistency Check	Leave as default (fsck).
Filesystems Recovery Method	Leave as default (sequential).
Filesystems/Directories to Export	Press F4 and choose the appropriate file systems to export for the resource group.
Filesystems/Directories to NFS mount	Press F4 and choose the appropriate file systems to NFS cross mount via cluster for the resource group.
Network For NFS Mount	Leave blank.

Volume Groups

Press F4 and choose appropriate volume groups for the resource group.

Concurrent Volume groups, Raw Disk PVIDs, Connections Services, Fast Connect Services, Tape Resources, Application Servers, Communication Links, Primary Workload Manager Class, Secondary Workload Manager Class, Miscellaneous Data

Leave default or empty.

Automatically Import Volume Groups, Inactive Takeover Activated, Cascading Without Fallback Enabled, Disk Fencing Activated, Filesystems mounted before IP configured

Leave default.

Change/Show Resources/Attributes for a Resource Group

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

	[Entry Fields]	
Resource Group Name	C37_CAS_01	
Node Relationship	cascading	
Site Relationship	ignore	
Participating Node Names / Default Node Priority	C37_P690 C38_P690	
Dynamic Node Priority	<input type="checkbox"/>	+
Service IP label	[c37srcv]	+
Filesystems (default is All)	[/s37fs01 /s37fs02]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems/Directories to Export	[/s37fs02]	+
Filesystems/Directories to NFS mount	[/s37fs02]	+
Network For NFS Mount	<input type="checkbox"/>	+
Volume Groups	[s37vg]	+
Concurrent Volume groups	<input type="checkbox"/>	+
Raw Disk PVIDs	<input type="checkbox"/>	+
Connections Services	<input type="checkbox"/>	+
Fast Connect Services	<input type="checkbox"/>	+
Tape Resources	<input type="checkbox"/>	+
Application Servers	<input type="checkbox"/>	+
Communication Links	<input type="checkbox"/>	+
Primary Workload Manager Class	<input type="checkbox"/>	+
Secondary Workload Manager Class	<input type="checkbox"/>	+
Miscellaneous Data	<input type="checkbox"/>	
Automatically Import Volume Groups	false	+
Inactive Takeover Activated	false	+
Cascading Without Fallback Enabled	false	+
Disk Fencing Activated	false	+
Filesystems mounted before IP configured	false	+

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-27 Change/Show Resources/Attributes for Resource Group

Use the same SMIT screen (see Figure 2-28 on page 106) to enter the data for the second application server.

Choose C38_CAS_01.

Change/Show Resources/Attributes for a Resource Group

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

	[Entry Fields]	
Resource Group Name	C38_CAS_01	
Node Relationship	cascading	
Site Relationship	ignore	
Participating Node Names / Default Node Priority	C38_P690 C37_P690	
Dynamic Node Priority	<input type="checkbox"/>	+
Service IP label	[c38srvc]	+
Filesystems (default is All)	[/s38fs01 /s38fs02]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems/Directories to Export	<input type="checkbox"/>	+
Filesystems/Directories to NFS mount	<input type="checkbox"/>	+
Network For NFS Mount	<input type="checkbox"/>	+
Volume Groups	<input type="checkbox"/>	+
Concurrent Volume groups	<input type="checkbox"/>	+
Raw Disk PVIDs	<input type="checkbox"/>	+
Connections Services	<input type="checkbox"/>	+
Fast Connect Services	<input type="checkbox"/>	+
Tape Resources	<input type="checkbox"/>	+
Application Servers	[C38_APP_01]	+
Communication Links	<input type="checkbox"/>	+
Primary Workload Manager Class	<input type="checkbox"/>	+
Secondary Workload Manager Class	<input type="checkbox"/>	+
Miscellaneous Data	<input type="checkbox"/>	
Automatically Import Volume Groups	false	+
Inactive Takeover Activated	false	+
Cascading Without Fallback Enabled	false	+
Disk Fencing Activated	false	+
Filesystems mounted before IP configured	false	+

F1=Help
F2=Refresh
F3=Cancel
F4=List

F5=Reset
F6=Command
F7=Edit
F8=Image

F9=Shell
F10=Exit
Enter=Do

Figure 2-28 Change/Show Resources/Attributes for a Resource Group

5. Synchronize cluster resources

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Resources -> Synchronize Cluster Resources**.

Press the Enter key to run cluster verification and synchronization. The cluster synchronization process will be done when the verification process passes without errors.

Start cluster

We have a few choices for how we want to start the cluster on the nodes:

- Start SMIT HACMP on all nodes (see Figure 2-29) and select **Cluster Services** -> **Start Cluster Services** or use the `smit c1start` fast path.

Fill in the following field:

Start now, on system restart or both Choose the type of starting daemons.

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?	true	+
Cluster to re-acquire resources after forced down?	false	+

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-29 Start Cluster Services

- Start SMIT HACMP on all nodes from one node (see Figure 2-30 on page 108) and select **Cluster System Management** -> **HACMP Cluster Services** -> **Start Cluster Services** or use the `smit c1_c1start.dialog` fast path.

Enter the following data:

Start now, on system restart or both Choose the type of starting daemons.

Start Cluster Services on these nodes Press F4 to get list of cluster nodes and choose all nodes using F7 or enter the node's name using a comma as a separator.

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

[C37_P690,C38_P690]

+

BROADCAST message at startup?

true

+

Startup Cluster Lock Services?

false

+

Startup Cluster Information Daemon?

true

+

Cluster to re-acquire resources

false

+

after forced down?

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

Figure 2-30 Start Cluster Services (SPOC)

Cluster status

The cluster is up and running and all resources are up on their own nodes. Verify the HACMP subsystems using the `lssrc -g cluster` command (see Example 2-4).

Example 2-4 Output of `lssrc -g cluster` command

[c37f1rp01]:/ # lssrc -g cluster			
Subsystem	Group	PID	Status
clstrmgrES	cluster	23206	active
clsmuxpdES	cluster	20600	active
clinfoES	cluster	35276	active

The cluster daemons are started. Check the cluster status using the `clstat` command, as shown in Figure 2-31 on page 109.

```

clstat - HACMP Cluster Status Monitor
-----
Cluster: P690_2 (1)          Tue Jun  4 13:24:38 EDT 2002
      State: UP              Nodes: 2
      SubState: STABLE
      Node: C37_P690         State: UP
        Interface: c37srcv   (0)      Address: 192.168.3.11
                                   State:  UP
        Interface: tmssa37   (2)      Address: 0.0.0.0
                                   State:  UP

      Node: C38_P690         State: UP
        Interface: c38srcv   (0)      Address: 192.168.3.12
                                   State:  UP
        Interface: tmssa38   (2)      Address: 0.0.0.0
                                   State:  UP

***** f/forward, b/back, r/refresh, q/quit *****

```

Figure 2-31 Cluster status

Check the network interfaces on the both nodes using the **netstat -i** command (see Example 2-5). As you can see, the service, persistent, and standby IP labels are configured correctly and up.

Example 2-5 Output of netstat -i command

```

on first node
[c37f1rp01]:/ # netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en1 1500 link#4 0.2.55.6a.37.37 1712 0 1743 0 0
en1 1500 192.168.3 c37srcv 1712 0 1743 0 0
en1 1500 192.168.3.1 c37pers 1712 0 1743 0 0
en2 1500 link#3 0.2.55.6a.b2.d 8010 0 7128 0 0
en2 1500 192.168.3.6 c37stby 8010 0 7128 0 0
lo0 16896 link#1 27144 0 27399 0 0
lo0 16896 127 loopback 27144 0 27399 0 0
lo0 16896 ::1 27144 0 27399 0 0

on second node
[c38f1rp01]:/ # netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en1 1500 link#3 0.2.55.6a.38.38 1450 0 1310 0 0
en1 1500 192.168.3 c38srcv 1450 0 1310 0 0
en1 1500 192.168.3.1 c38pers 1450 0 1310 0 0
en2 1500 link#4 0.2.55.6a.a8.40 3244 0 2434 0 0
en2 1500 192.168.3.6 c38stby 3244 0 2434 0 0

```

lo0	16896	link#1		15799	0	15976	0	0
lo0	16896	127	loopback	15799	0	15976	0	0
lo0	16896	::1		15799	0	15976	0	0

Check the volume groups and file systems on both nodes by using the **lsvg -o** and **df -k** commands (see Example 2-6). As you can see, the volume groups are active and the file systems are mounted.

Example 2-6 Output of lsvg -o and df -k commands

on first node:								
[c37f1rp01]:/ # lsvg -o								
s37vg								
rootvg								
[c37f1rp01]:/ # df -k								
Filesystem	1024-blocks		Free %Used		Iused %Iused		Mounted on	
...								
/dev/s37lv01	655360		634736 4%		18 1%		/s37fs01	
/dev/s37lv02	655360		626988 5%		18 1%		/s37fs02	
on second node								
[c38f1rp01]:/ # lsvg -o								
s38vg								
rootvg								
[c38f1rp01]:/ # df -k								
Filesystem	1024-blocks		Free %Used		Iused %Iused		Mounted on	
...								
/dev/s38lv01	655360		634736 4%		18 1%		/s38fs01	
/dev/s38lv02	655360		634740 4%		17 1%		/s38fs02	

Check the resource group using the **clfindres** command (see Example 2-7). As you can see, the resource groups are up and acquired by the owner nodes.

Example 2-7 Output of clfindres command

[c37f1rp01]:/ # clfindres				
GroupName	Type	State	Location	Sticky Loc
-----	-----	-----	-----	-----
C37_CAS_01	cascading	UP	C37_P690	
C38_CAS_01	cascading	UP	C38_P690	

Check the application logs to see if the application is started on both nodes (see Example 2-8).

Example 2-8 Output of tail command on the log files

for application 1	
[c37f1rp01]:/s37fs01 # tail c37_app_01.log	

```

...
Tue Jun  4 13:20:40 EDT 2002
Start C37 application 01

for application 2
[c37f2rp01]:/s38fs01 # tail c38_app_01.log
...
Tue Jun  4 13:22:30 EDT 2002
Start C38 application 01

```

Cluster status after Ethernet adapter failure occurs

When the adapter failure occurs on the C37_P690 node, we can see that the service and persistent IP labels have been taken over by the standby adapter (see the `netstat` command output in Example 2-9).

Example 2-9 Output of `netstat -i` command

[c37f1rp01]:/s37fs01 # netstat -i								
Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
...								
en1	1500	link#4	0.2.55.6a.0.37	31	0	21	0	0
en1	1500	192.168.3.6	c37stby	31	0	21	0	0
en2	1500	link#3	0.2.55.6a.37.37	36	0	29	0	0
en2	1500	192.168.3	c37srvc	36	0	29	0	0
en2	1500	192.168.3.1	c37pers	36	0	29	0	0
...								

Example 2-10 shows the events handled by cluster manager when the adapter failure occurs. After a failure, HACMP reconfigures the network adapters.

Example 2-10 `Swap_adapter` event in `cluster.log` file

...	
Jun 4 13:52:58 c37f1rp01 HACMP for AIX: EVENT START: swap_aconn_protocols	en2 en1
Jun 4 13:52:58 c37f1rp01 HACMP for AIX: EVENT COMPLETED:	swap_aconn_protocols en2 en1
Jun 4 13:52:59 c37f1rp01 HACMP for AIX: EVENT COMPLETED: swap_adapter	C37_P690 PUBL_NET_1 192.168.3.71 192.168.3.11
Jun 4 13:52:59 c37f1rp01 HACMP for AIX: EVENT START: swap_adapter_complete	C37_P690 PUBL_NET_1 192.168.3.71 192.168.3.11
Jun 4 13:52:59 c37f1rp01 HACMP for AIX: EVENT COMPLETED:	swap_adapter_complete C37_P690 PUBL_NET_1 192.168.3.71 192.168.3.11
Jun 4 13:53:03 c37f1rp01 HACMP for AIX: EVENT START: fail_standby C37_P690	192.168.3.71
Jun 4 13:53:03 c37f1rp01 HACMP for AIX: EVENT COMPLETED: fail_standby	C37_P690 192.168.3.71

...

Cluster status after node failure occurs

When the node C37_P01 crashes, we can see that all the resources for that node have been taken over by node C38_P01. Verify the cluster status (see Figure 2-32). We can see that the cluster is up and stable, but node C37_P690 is down.

```

                                clstat - HACMP Cluster Status Monitor
                                -----
Cluster: P690_2 (1)                Tue Jun  4 14:24:38 EDT 2002
      State: UP                      Nodes: 2
      SubState: STABLE
Node: C37_P690                      State: DOWN
  Interface: c37boot (0)              Address: 192.168.3.1
                                      State:  DOWN
  Interface: tmssa37 (2)              Address: 0.0.0.0
                                      State:  DOWN

Node: C38_P690                      State: UP
  Interface: c38srvc (0)              Address: 192.168.3.12
                                      State:  UP
  Interface: tmssa38 (2)              Address: 0.0.0.0
                                      State:  DOWN

***** f/forward, b/back, r/refresh, q/quit*****

```

Figure 2-32 Cluster status

Check the resource group using the **clfindres** command (see Example 2-11). Node C38_P690 acquired resource group C37_CAS_01 from the failed node.

Example 2-11 Output of clfindres command

[c38f1rp01]:/s38fs01 # clfindres				
GroupName	Type	State	Location	Sticky Loc
-----	-----	-----	-----	-----
C37_CAS_01	cascading	UP	C38_P690	
C38_CAS_01	cascading	UP	C38_P690	

Check the network interfaces on both nodes using the **netstat -i** command (see Example 2-12 on page 113). The IP standby adapter on the backup node is reconfigured to be the IP service adapter on the failed node.

Example 2-12 Output of netstat -i command

[c37f2rp01]:/s38fs01 # netstat -i									
Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll	
...									
en1	1500	link#3	0.2.55.6a.0.38	7852	0	6841	0	0	
en1	1500	192.168.3	c38srvc	7852	0	6841	0	0	
en1	1500	192.168.3.1	c38pers	7852	0	6841	0	0	
en2	1500	link#4	0.2.55.6a.0.37	365	0	163	0	0	
en2	1500	192.168.3	c37srvc	365	0	163	0	0	
...									

Check the volume groups and file systems on both nodes by using the **lsvg -o** and **df -k** commands (see Example 2-13). The volume group and file system from the failed node was acquired correctly by the backup node.

Example 2-13 Output of lsvg -o and df -k command

[c37f2rp01]:/s38fs01 # lsvg -o									
s37vg									
s38vg									
rootvg									
[c37f2rp01]:/s38fs01 # df -k									
Filesystem	1024-blocks	Free	%Used	Iused	%Iused	Mounted on			
...									
/dev/s38lv01	655360	634736	4%	18	1%	/s38fs01			
/dev/s38lv02	655360	634740	4%	17	1%	/s38fs02			
/dev/s37lv01	655360	634736	4%	18	1%	/s37fs01			
/dev/s37lv02	655360	626988	5%	18	1%	/s37fs02			

Check the application log (see Example 2-14), which indicates that application 1 was restarted on the backup node as well.

Example 2-14 Output of tail /s37fs01/c37_app_01.log

[c37f2rp01]:/s38fs01 # tail /s37fs01/c37_app_01.log									
Tue Jun 4 13:20:40 EDT 2002									
Start C37 application 01									
Tue Jun 4 14:22:32 EDT 2002									
Start C37 application 01									

Example 2-15 on page 114 shows the events handled by the cluster manager when node failure occurs. After the C37_P690 node crashed, all the resources were acquired by the backup node.

Example 2-15 Node_down event in cluster.log file

```
Jun  4 14:21:30 c37f2rp01 HACMP for AIX: EVENT START: node_down C37_P690
Jun  4 14:21:34 c37f2rp01 HACMP for AIX: EVENT START: acquire_takeover_addr
Jun  4 14:21:41 c37f2rp01 HACMP for AIX: EVENT COMPLETED:
acquire_takeover_addr
Jun  4 14:22:30 c37f2rp01 HACMP for AIX: EVENT COMPLETED: node_down
C37_P690
Jun  4 14:22:30 c37f2rp01 HACMP for AIX: EVENT START: node_down_complete
C37_P690
Jun  4 14:22:32 c37f2rp01 HACMP for AIX: EVENT START: start_server
C37_APP_01
Jun  4 14:22:33 c37f2rp01 HACMP for AIX: EVENT COMPLETED: start_server
C37_APP_01
Jun  4 14:22:34 c37f2rp01 HACMP for AIX: EVENT COMPLETED:
node_down_complete C37_P690
Jun  4 14:22:39 c37f2rp01 HACMP for AIX: EVENT START: network_down -1
TM_SSA_1
Jun  4 14:22:39 c37f2rp01 HACMP for AIX: EVENT COMPLETED: network_down -1
TM_SSA_1
Jun  4 14:22:39 c37f2rp01 HACMP for AIX: EVENT START: network_down_complete
-1 TM_SSA_1
Jun  4 14:22:39 c37f2rp01 HACMP for AIX: EVENT COMPLETED:
network_down_complete -1 TM_SSA_1
```

Stop cluster

We have a few choices for stopping a cluster on the nodes:

- Start SMIT HACMP on all nodes (see Figure 2-33 on page 115) and select **Cluster Services -> Stop Cluster Services** or use the **smit clstop** fast path.

Stop now, on system restart or both Choose the type of stopping daemons.

Shutdown mode Press F4 to choose the mode to release the resources.

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)

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-33 Stop Cluster Services

- Start SMIT HACMP on one node (see Figure 2-34 on page 116) and select **Cluster System Management -> HACMP Cluster Services -> Stop Cluster Services** or use the `smit c1_c1stop.dialog` fast path.

Fill in the following fields:

Stop now, on system restart or both	Choose the type of starting daemons.
Stop Cluster Services on these nodes	Press F4 to get a list of the cluster nodes and choose all the nodes by using F7 or by entering the nodes name using the comma as a separator.
Shutdown mode	Press F4 to choose the mode to release resources.

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

+

Stop Cluster Services on these nodes

+

BROADCAST cluster shutdown?

true

+

* Shutdown mode

graceful

+

(graceful or graceful with takeover, forced)

F1=Help

F2=Refresh

F3=Cancel

F4=List

F5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure 2-34 Stop Cluster Services (SPOC)

2.4.8 Scenario 2: Using SP Switch/SP Switch2 adapter

In this scenario, we will use the SP Switch2 instead of the Ethernet adapter used in the scenario described in Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87. All examples in this scenario assume the already existing configuration from Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87. We do not describe configuration details that are already described, and we focus only on differences and additional configurations required for using the SP Switch2.

Using SP Switch/SP Switch2 networks in an HACMP requires implementation of the p690 systems in a Cluster 1600 environment. That requires implementation of Parallel System Support Program (PSSP) at level PSSP 3.4. For a detailed explanation of the SP Switch/SP Switch2 implementation in a Cluster 1600, see the redbook *RS/6000 SP/Cluster: New Enhancements in PSSP 3.4*, SG24-6604.

SP Switch/SP Switch2 considerations

HACMP 4.5 has a new feature, IP address takeover using IP aliasing. This new feature works on most of the IP network types, including the SP Switch/SP Switch2. The SP Switch/SP Switch2 can operate in two modes:

- ▶ Traditional IPAT mode, as in previous releases
- ▶ IP address takeover, using the new IP aliasing mode supported in HACMP 4.5

In traditional IPAT on the SP Switch/SP Switch2, the boot address defined to HACMP is itself an alias address and not the AIX boot time or "base" address of the adapter. The alias boot address is added to the adapter when you start

cluster services. When IPAT occurs, the alias boot address is removed and the service address is added as an alias. The base address is not defined to HACMP and is not changed during IPAT

In the case of IPAT via aliasing the SP Switch/SP Switch2, the boot address defined to HACMP is the AIX boot time address of the adapter (the base address). There is no longer a need for the "alias boot" adapter. When IPAT occurs, the base or boot address remains unchanged, and the service address is added or removed as an alias.

Considerations for the IP network schema we used in this scenario:

- ▶ We assume use of one SP Switch/SP Switch2 network.
- ▶ We assume use of one SP Switch/SP Switch2 adapter per LPAR.
- ▶ We use the traditional IPAT mode.
- ▶ We assume the SP Switch/SP Switch2 base addresses are already defined in the PSSP environment and the ARP is enabled for the interfaces. The IP netmask is defined.

We recommend that the IP addresses for the HACMP IP labels be defined to a separate IP network than the one for the base addresses, but with the same netmask. Addresses defined to the same IP logical subnet as the base address should work from the HACMP point of view, but they may have problems because of the IP striping introduced in AIX 5L Version 5.1.

Note: The IP alias addresses used for HACMP do not need to be configured by AIX, because the HACMP startup will check for the interfaces and if they are not present, HACMP will define them using the `/usr/lpp/ssp/css/ifconfig css0 alias` command. However, HACMP expects that all IP labels and IP addresses be resolved through `/etc/hosts` or DNS, and it also expects the remote commands (`rsh`, `rnp`, and so on) to run through these interfaces.

During the configuration, we should consider the following items:

- ▶ ARP must be enabled for the SP Switch/SP Switch2 interfaces so the IP address takeover can work.
- ▶ Use of the AIX error notification should be considered in order to handle SP Switch/SP Switch2 adapter errors.
- ▶ The SP Switch network must be defined as an HACMP private network.

Planning for HACMP resources

Here we assume that the cluster is already configured, as in Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87. We intend to add additional resources that are related to the Cluster 1600 and the SP Switch/SP Switch2 environment only.

Figure 2-35 describes the network topology and resources of the Cluster 1600. The diagram will help you to better understand the planning for HACMP resources.

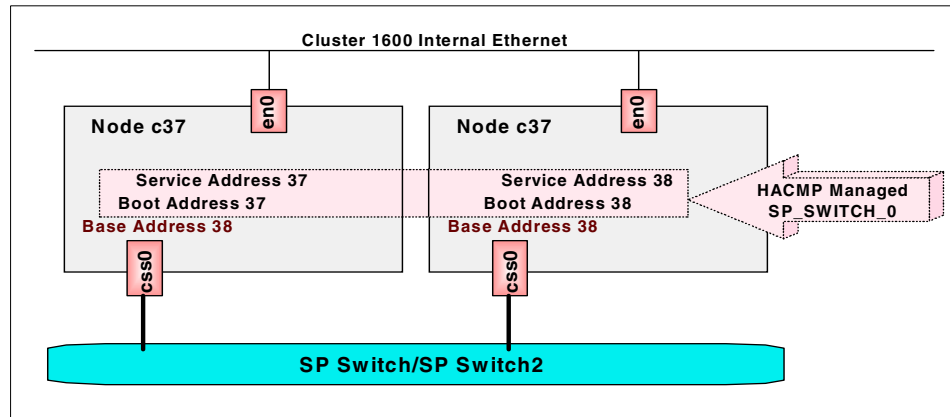


Figure 2-35 SP Switch network addresses

Network considerations and design

Let us focus on the networks of the Cluster 1600 environment. We already know that the internal Ethernet network of the Cluster 1600 should not be used for HACMP IP address takeover, but can be used for HACMP monitoring. We can define the HACMP networks in the following way:

Internal Ethernet This network is already defined from the Cluster 1600 PSSP configuration. We define this network in HACMP as a private network with only one service interface per node and with no boot and no standby interfaces. HACMP understands this configuration as a network used only for monitoring.

Base net for Switch The SP Switch/SP Switch2 base addresses are required for the monitoring and control of the SP Switch/SP Switch2 high performance network and, like the Internal Ethernet, these addresses should not be used for HACMP IP address takeover, but can be used for HACMP monitoring.

Switch network The HACMP managed SP Switch/SP Switch2 network is not a separate network. It is just a logical network on the same SP Switch with interfaces created by IP aliasing to the SP Switch base address.

Check the IP addresses and corresponding labels and set up name resolution in **/etc/hosts** on every node of the cluster and the control workstation, DNS, or both, (see Example 2-16).

Example 2-16 /etc/hosts file for SP Switch2 interfaces

```
/etc/hosts:
# Cluster 1600 Internal Ethernet:
9.114.189.1      c37f1rp01.pps.pok.ibm.com c37f1rp01
9.114.189.17    c37f2rp01.pps.pok.ibm.com c37f2rp01
# SP Switch base network:
9.114.189.65    c37sw0base.pps.pok.ibm.com c37sw0base
9.114.189.81    c38sw0base.pps.pok.ibm.com c38sw0base
# SP Switch HACMP managed interfaces
192.168.13.1    c37sw0boot.ppd.pok.ibm.com c37sw0boot
192.168.13.2    c38sw0boot.ppd.pok.ibm.com c38sw0boot
192.168.13.11   c37sw0srvc.ppd.pok.ibm.com c37sw0srvc
192.168.13.12   c38sw0srvc.ppd.pok.ibm.com c38sw0srvc
```

Place configuration information into the control workstation

The Cluster 1600 control workstation centrally stores configuration information in a system database called the System Data Repository (SDR). The SDR comes together with the PSSP installation procedures. As we have assumed the Cluster 1600 environment is already configured, verify the following:

1. Check the correct definition of the Cluster 1600 Internal Ethernet on the control workstation
2. Check or define the SP Switch/SP Switch2 network and base interfaces on the control workstation according to the SP Switch installation procedures.
3. Optionally define HACMP boot and service interfaces in the SDR by running SMIT and selecting **RS/6000 SP System Management -> RS/6000 SP Configuration Database Management -> Enter Database Information -> Node Database Information -> Additional Adapter Information**.

Configure Kerberos security

Security is a requirement of Cluster 1600 environments that cannot be omitted. Kerberos Version 4 is the recommended security protocol used in Cluster 1600. HACMP supports the use of the Kerberos protocol, but some manual configuration steps should be performed in order to get Kerberos running for HACMP interfaces.

The PSSP utilities do not create the Kerberos principals for the HACMP interfaces. HACMP provides the **c1_setup_kerberos** tool, which helps properly set up Kerberos. This tool checks HACMP configuration and creates the Kerberos definitions. However, we recommend that you know the manual procedure to create the Kerberos definitions.

The manual configuration of the Kerberos definitions requires you to create Kerberos instances of the rcmd principal for every IP label (in our case, rcmd.c37sw0boot, rcmd.c38sw0boot, rcmd.c37sw0srcv, and rcmd.c38sw0srcv). These instances can be added from the control workstation by the kadmin or the kdb_edit utility. Example 2-17 shows adding one Kerberos rcmd instance using the kadmin utility.

Example 2-17 Defining the Kerberos principal for HACMP

```
/usr/lpp/ssp/kerberos/bin/kadmin
Welcome to the Kerberos V4 Administration Program, version2
Type "help" if you need it.
admin: ank rcmd.c37sw0boot
Admin password: <Kerberos password of root.admin>
Password for rcmd.c37sw0boot: <password>
Verifying, please re-enter Password for rcmd.c37sw0boot: <password>
rcmd.c37sw0boot added to database.
admin: q
Cleaning up and exiting.
```

Important: The password chosen for the rcmd instance is not referenced anymore but must not be null.

After the rcmd instances are defined, you can extract the corresponding Kerberos keys for each rcmd instance (on the control workstation) using the ext_srvtab utility and append the results to the /etc/krb-srvtab file on the corresponding nodes (see Example 2-18).

Example 2-18 Upgrading /etc/krb-srvtab file

```
cws> /usr/lpp/ssp/kerberos/etc/ext_srvtab -n c37sw0boot c37sw0srcv
cws> /usr/lpp/ssp/kerberos/etc/ext_srvtab -n c38sw0boot c38sw0srcv
cws> ls -l *-new-srvtab
cws> rcp c37sw0boot-new-srvtab c37sw0srcv c37:/tmp/
cws> rcp c38sw0boot-new-srvtab c38sw0srcv c38:/tmp/
cws> rsh c37 "cat /tmp/c37sw0boot-new-srvtab /tmp/c37sw0srcv-new-srvtab >>
/etc/krb-srvtab"
cws> rsh c37 "ksrvutil list"
cws> rsh c38 "cat /tmp/c38sw0boot-new-srvtab /tmp/c38sw0srcv-new-srvtab >>
/etc/krb-srvtab"
```

```
cws> rsh c38 "ksrvutil list"
```

Add the rcmd principal instances to the /.klogin file, which allows access to the node. Edit the files locally or edit the file on the control workstation and distribute it to cluster nodes (see Example 2-19).

Example 2-19 Updating .klogin files in nodes

```
vi /.klogin
root.admin@PPD.POK.IBM.COM
...
rcmd.c37sw0boot@PPD.POK.IBM.COM
rcmd.c37sw0svrc@PPD.POK.IBM.COM
rcmd.c38sw0boot@PPD.POK.IBM.COM
rcmd.c38sw0svrc@PPD.POK.IBM.COM

pcp -w c37,c38 /.klogin /.klogin
```

Check connectivity

Check that the SP Switch/SP Switch2 interfaces are up and running on all HACMP nodes (see Example 2-20).

Example 2-20 Testing the switch interface

```
netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
...
css0* 65504 link#5 12757 0 13220 0 0
css0* 65504 9.114.189.6 c38sw0base 12757 0 13220 0 0
ifconfig -a
...
css0: flags=800843<UP,BROADCAST,RUNNING,SIMPLEX>
        inet 9.114.189.81 netmask 0xffffffc0 broadcast 9.114.189.127
ping <SP Switch base addresses>
```

Configure SP Switch/SP Switch2 in HACMP

The following steps describe configuring the SP Switch/SP Switch2 network in HACMP:

1. Configure the HACMP Network.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Networks -> Configure IP based Networks -> Add a Network**.

Add an IP-based Network

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

* Network Name

* Network Attribute

Network Type

Subnet(s)

[Entry Fields]

[SP_SWITCH_0]

private

[hps]

[192.168.13.0/26]

+

+

+

F1=Help

F2=Refresh

F3=Cancel

F4=List

Esc+5=Reset

Esc+6=Command

Esc+7=Edit

Esc+8=Image

Esc+9=Shell

Esc+0=Exit

Enter=Do

Figure 2-36 SMIT HACMP - Add an IP-based Network

Note: Because the HACMP handles SP Switch/SP Switch2 interfaces in a different way, the Discover IP Topology menu does not discover the aliased SP Switch/SP Switch2 networks. Add the network information by typing it in manually.

- Network Name

Select a symbolic name that HACMP will use to refer to this configuration, for example: SP_SWITCH_0.
- Network Attribute

Always select private for the SP Switch network.
- Network Type

Use List to select the network type. For SP Switch/SP Switch2, use hps.
- Subnet(s)

Type in the IP network address your HACMP boot and service addresses belong to and finish the address with the “/” (slash) character and the netmask length (number of 1s in the binary representation of the netmask), for example, 192.168.13.0/26.

2. Configure HACMP Adapters

In the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP based Interfaces / IP Labels -> Add Initial Interfaces.**

Select the SP Switch Network defined previously (SP_SWITCH_0).

You will see a screen similar to the one in Figure 2-37 on page 123.

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[c37sw0boot]	+
Network Type	hps	
Network Name	SP_SWITCH_0	
* Interface Function	[boot]	+
Interface IP Address	[]	
* Node Name	[C37_P690]	+
Netmask	[]	+

F1=Help

F2=Refresh

F3=Cancel

F4=List

Esc+5=Reset

Esc+6=Command

Esc+7=Edit

Esc+8=Image

Esc+9=Shell

Esc+0=Exit

Enter=Do

Figure 2-37 SMIT HACMP - Add an Initial Interface

Note: Repeat this menu for every IP label (c37sw0boot, c37sw0srcv, c38sw0boot, and c38sw0srcv).

Fill in the following fields

- IP Label

The IP label of the HACMP handled interface. The IP label must be correctly resolved from /etc/hosts file or DNS.
- Network Type

Information imported from the network definition.
- Network Name

The SP Switch network HACMP reference (as selected).
- Interface Function

Use List to select boot or service according to the planning diagram or sheet.
- Interface IP Address

Leave this line empty. HACMP resolves the correct IP address that corresponds to the IP label from system name resolution.
- Node Name

For every boot interface and service interface used for Cascading resource groups, use List to select the HACMP node name.

For the service interface (if used in rotating resource group), leave this item empty.

Netmask Leave this line empty. HACMP resolves the correct IP netmask from the HACMP network definition.

If you check the SMIT menu for changing the HACMP network, there is an additional configurable item concerning the use of the IP aliasing (see Figure 2-38).

Change an IP-based Network

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

	[Entry Fields]	
Network Name	SP_SWITCH_0	
New Network Name	<input type="text"/>	
* Network Attribute	private	+
* Network Type	[hps]	+
Subnet(s)	192.168.13.0/26	
Add Subnets	<input type="text"/>	+
Remove Subnets	<input type="text"/>	+
Use IP aliasing for IPAT	unsupported	+

F1=Help F2=Refresh F3=Cancel F4=List
Esc+5=Reset Esc+6=Command Esc+7=Edit Esc+8=Image
Esc+9=Shell Esc+0=Exit Enter=Do

Figure 2-38 SMIT HACMP menu - Change an IP-based Network

The Use IP aliasing item must be set to unsupported. If the item is set to anything else, it will be automatically changed back to unsupported.

3. Synchronize HACMP topology.

In the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Synchronize Cluster Topology**.

The synchronization can be performed while the cluster is running. In a running cluster, the DARE will handle the new topology and refreshes the cluster subsystems.

Associate Switch service addresses with resource groups

Follow these steps to associate Switch service addresses with resource groups:

1. In the SMIT HACMP scree, add the SP Switch/SP Switch2 service IP labels to corresponding resource groups by selecting **Cluster Configuration -> Cluster Resources -> Change/Show Resources/Attributes for a Resource Group**.

Select the resource group where the service IP label should be added (C37_CAS01).

Testing the configuration

For testing the HACMP function using SP Switch/SP Switch2 network, perform the steps in this section.

Prior to HACMP startup

When HACMP is not running on the node, the IP addresses are available as listed:

Base address	The SP Switch/SP Switch2 base address is always available because the PSSP monitoring and control uses this interface and is not handled by HACMP IP address takeover.
Boot address	The SP Switch/SP Switch2 boot address is defined in the HACMP. It does not need to be defined to the css0 interface but may be available. If HACMP did not run since the last system reboot, the boot address is usually not available.
Service address	The SP Switch/SP Switch2 service address is not available and should not be on this node.

Example 2-21 shows the **netstat** and **ifconfig** commands for the state of the SP Switch2 interfaces.

Example 2-21 Output of netstat -i for css network

```
netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
...
css0* 65504 link#5 12757 0 13220 0 0
css0* 65504 9.114.189.6 c38sw0base 12757 0 13220 0 0
ifconfig -a
...
css0: flags=800843<UP,BROADCAST,RUNNING,SIMPLEX>
      inet 9.114.189.81 netmask 0xffffffc0 broadcast 9.114.189.127
```

After HACMP is started

After HACMP is started on the node, the startup procedure tests the presence of the boot address and defines it to the css0 interface. This interface is needed only internally for HACMP startup and is replaced by the service address. The boot address is deleted and the service address is allocated to the css0 interface. When HACMP startup is finished and the HACMP is running, the following interfaces are available:

Base address	Not handled by HACMP; therefore, it is always available.
Boot address	Not available.

Service address Available and used by started applications.

Example 2-22 shows output of the **netstat** command on the nodes.

Example 2-22 Output of netstat -i command after HACMP startup

netstat -i								
Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
...								
css0	65504	link#5		12757	0	13220	0	0
css0	65504	9.114.189.6	c38sw0base	12757	0	13220	0	0
css0	65504	9.114.189.6	c38sw0srcv12757	0	13220	0	0	
ifconfig -a								
...								
css0:	flags=800843<UP,BROADCAST,RUNNING,SIMPLEX>							
	inet	9.114.189.81	netmask 0xffffffff	broadcast	9.114.189.127			
	inet	192.168.13.12	netmask 0xffffffc0	broadcast	9.114.189.127			

After adapter failure

SP Switch/SP Switch2 adapter failure is a single point of failure in one switch configuration. To eliminate this failure, we recommend you use two SP Switch2 configurations, as described in Section 2.4.9, “Scenario 3: Dual SP Switch2 network” on page 128.

The SP Switch adapter failure is handled by HACMP by promoting this failure to a resource group move. The resource group containing the failed SP Switch service IP label is forced to move to the next node defined in the HACMP resource group configuration. Other resources, the ones not containing the failed SP Switch IP label, remain on the node.

After node failure

After a node fails, the resource groups are activated on the second node. The SP Switch IP label is allocated on the secondary node as an additional IP alias.

Example 2-23 is the output of the **netstat** and **ifconfig** commands on node c37 that took over the resource group from node c38 after a node failure. Node c38 is powered off.

Example 2-23 Output of netstat -i after node failure

netstat -i								
Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
...								
css0	65504	link#5		12757	0	13220	0	0
css0	65504	9.114.189.6	c37sw0base	12757	0	13220	0	0
css0	65504	192.168.13	c37sw0srcv12757	0	13220	0	0	
css0	65504	192.168.13	c38sw0srcv12757	0	13220	0	0	
...								
ifconfig -a								
...								

```
css0: flags=800843<UP,BROADCAST,RUNNING,SIMPLEX>  
    inet 9.114.189.65 netmask 0xffffffc0 broadcast 9.114.189.127  
    inet 192.168.13.11 netmask 0xffffffc0 broadcast 192.168.13.63  
    inet 192.168.13.12 netmask 0xffffffc0 broadcast 192.168.13.63
```

2.4.9 Scenario 3: Dual SP Switch2 network

The new features implemented in the SP Switch2 architecture brought new possibilities for implementing dual SP Switch2 networks. This feature is not available in the previous SP Switch architecture. The supported SP Switch2 configurations are:

- ▶ One SP Switch2 network per one Cluster 1600 and one SP Switch2 adapter per node.
- ▶ Two SP Switch2 networks per one Cluster 1600 and two SP Switch2 adapters per node, each connected to different SP Switch2 networks.

Restriction: Configurations of two SP Switch2 adapters per node connected to one SP Switch2 are not supported.

The implementation of the dual SP Switch2 network with two SP Switch2 adapters per node creates two devices per node. These two devices are `css0` and `css1`. The PSSP software also creates a special pseudo-device `ml0`, which is used to create load balancing and high availability above the `css0` and `css1` adapter. The `ml0` pseudo-device is not supported by HACMP 4.5.

Dual SP Switch2 configuration

As the SP Switch2 network consists of two separate physical networks, the IP addresses must be selected from different IP subnets. We should also remember that the HACMP IPAT-handled addresses should be from a different subnet than the base addresses.

Note: A correct design of two SP Switch2 networks handled by HACMP IP address takeover requires you to allocate four IP subnets, or even more for more complex configurations.

The network layout of the two SP Switch2 configuration appears to the HACMP as two separate networks, and HACMP expects that these two networks will be defined in two different HACMP networks (Figure 2-40 on page 129).

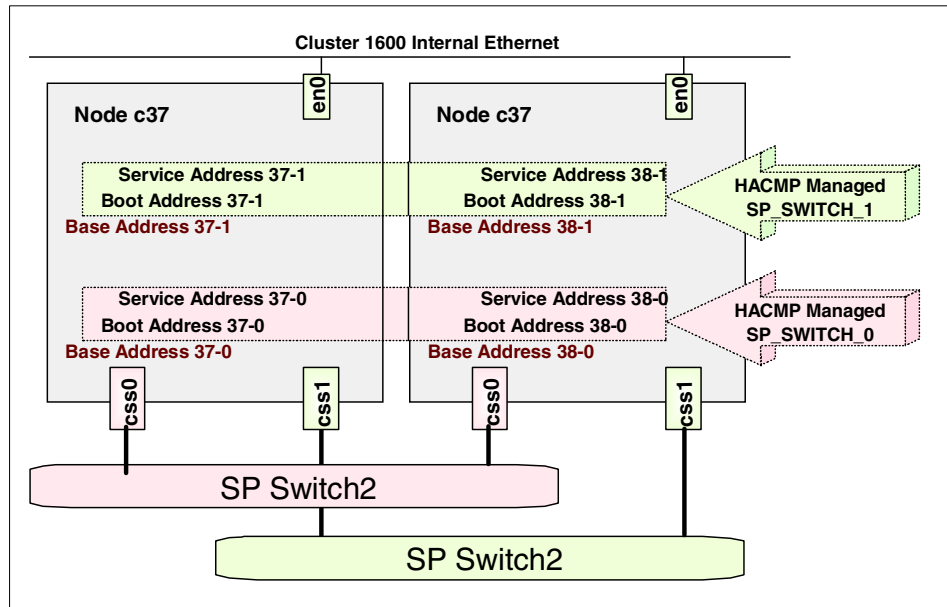


Figure 2-40 Dual SP Switch2 configuration

Important: The HACMP network_down event does not swap the networks in any way; for this event, only a log entry is performed. It is the user's responsibility to create post-event scripts to swap networks.

2.4.10 Scenario 4: IP Aliasing

Figure 2-41 on page 130 shows a cluster environment for Scenario 4 and the additional resources we will configure.

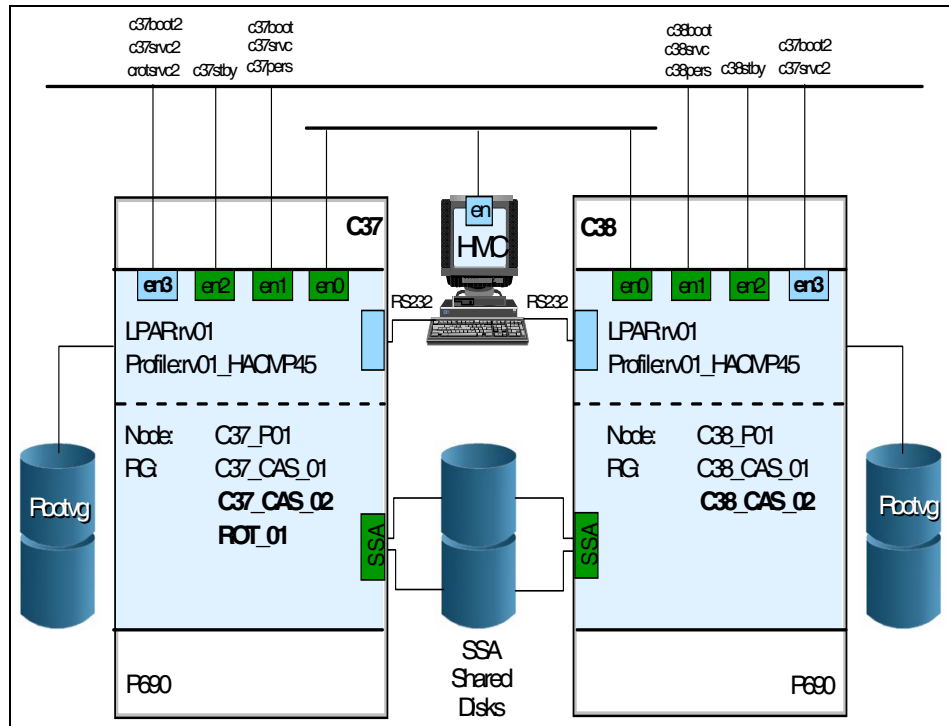


Figure 2-41 An example of a cluster for scenario 4

Description and assumptions

In this scenario, we will expand the cluster configuration from Scenario 1 to show how to configure additional adapters using the IP Aliasing feature.

We assume further configuration to the existing configuration:

- ▶ Additional adapters
 - Two secondary boot labels
 - Two secondary service labels
 - One rotating service label
- ▶ Two additional cascading resource groups
- ▶ One rotating resource group

We assume the following configuration:

- ▶ Network adapter failure (swap_adapter) expectation: IP service and rotating labels will go to adapter on backup node as alias. Related resource groups will move to backup node.

Used resources

According to Section 2.4.3, “Preparing the cluster for high availability” on page 75, we have to use the resources represented in Table 2-11 and Table 2-12.

Note: Scenario 4 is an extension of Scenario 1. We need to use the information contained in Table 2-8 on page 89, Table 2-9 on page 89, and Table 2-10 on page 90.

Table 2-11 Cluster Topology definition - scenario 4

Name	Node name	
	C37_P690	C38_P690
Secondary boot adapter	c37boot2	c38boot2
Service adapter	c37svc2	c38svc2
Rotating service	crotsrv	
Resource group	C37_CAS_02	C38_CAS_02

Note: All adapters will be added to the PUBL_NET_2 public network definition.

Table 2-12 Cluster Resources definition - scenario 4

Name	Resource group name		
	C37_CAS_02	C38_CAS_02	ROT_01
Participating nodes	C37_P690 (primary) C38_P690 (backup)	C38_P690 (primary) C37_P690 (backup)	C38_P690 C37_P690
Service addresses	c37svc2	c38svc2	crotsrv

Configuration steps

The HACMP cluster configuration is accomplished by doing two main steps. The first is to configure the cluster topology. The second is to configure the cluster resources.

Define the Cluster topology

You only need to perform these steps on one node. The definition will be copied to the other nodes after you synchronize the cluster topology. Refer to “Used resources” on page 131. This scenario is a extension of Scenario 1 and we

assume that all steps from Scenario 1 are performed before the steps described below.

Complete the following steps to define the cluster topology:

1. Configure Additional Adapters

- Configure second boot interface for the first node.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Choose Add an Adapter on a new Network.

For the first node, fill in the following fields:

IP Label Press F4 and choose c37boot2.

Network Type Press F4 and choose ether.

Network Name Enter PUBL_NET_2.

Network Attribute Leave as default (public).

Interface Function Leave as default (boot).

Interface IP Address Leave blank.

Node Name Press F4 and choose C37_P690.

Netmask Enter 255.255.255.224.

- Configure the second boot interface.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Choose PUBL_NET_2 () and fill in following fields for the second node:

IP Label Press F4 and choose c38boot2.

Interface Function Leave as default (boot).

Interface IP Address Leave blank.

Node Name Press F4 and choose C38_P690.

Netmask Enter 255.255.255.224.

- Configure service adapters for both nodes and rotating service adapter

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Multiple Service IP Labels to a Network**.

Choose PUBL_NET_2 and fill in the following field:

IP Label(s) Press F4 and choose c37srcv2, c38srcv2, and crotsrv.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Change / Show an Interface / IP Label**.

Choose c37srcv2 and fill in the following field:

Node name Press F4 and choose C37_P690.

Do it again, but choose c38srcv2 instead, and fill in the node name field with C38_P690.

2. Synchronize the cluster topology.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Topology -> Synchronize Cluster Topology**.

Press Enter to run cluster verification and synchronization. The cluster synchronization process will be done when the verification process passes without errors.

Define the cluster resources

Follow these steps to define the cluster resources:

1. Configure additional cascading resource groups for both nodes and the rotating group for the rotating service adapter.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Resources -> Define Resource Groups -> Add a Resource Group**.

For the C37_CAS_02 resource group, fill in the following fields:

Resource Group Name Enter C37_CAS_02.

Node Relationship Leave default (cascading).

Site Relationship Leave default (ignore).

Participating Node Names /

Default Node Priority Press F4 and choose C37_P690 and C38_P690.

Use the same SMIT screen to create the data for the C38_CAS_02 resource group. Please pay attention to the order in the Participating Node Name field. It is the default method of node priority. Fill in this field with C38_P690 and C37_P690.

Use the same SMIT screen to create the ROT_01 rotating resource group.

Fill in the following fields:

Resource Group Name	Enter ROT_01.
Node Relationship	Leave default (rotating).
Site Relationship	Leave default (ignore).
Participating Node Names / Default Node Priority	Press F4 and choose C37_P690 and C38_P690.

2. Define and assign resources into resource groups.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Resources -> Change/Show Resources/Attributes for a Resource Group**.

Choose C37_CAS_02 to enter the data for the C37_CAS_02 cascading resource group.

Fill in the following fields:

Dynamic Node Priority	Leave blank.
Service IP label	Press F4 and choose c37srcv2.

Leave the rest of fields as default or blank.

Use the same SMIT screen to enter the data for C38_CAS_02 cascading resource group.

Choose C38_CAS_02 and fill in the following field:

Service IP label	Press F4 and choose c38srcv2
-------------------------	------------------------------

Use the same SMIT screen to enter the data for ROT_01 rotating resource group.

Choose **ROT_01** and fill:

Service IP label	Press F4 and choose crotsrv.
-------------------------	------------------------------

3. Synchronize cluster resources.

On the SMIT HACMP screen, select **Cluster Configuration -> Cluster Resources -> Synchronize Cluster Resources**.

Press Enter to run cluster verification and synchronization. The cluster synchronization process will be done when the verification process passes without errors.

Testing the IP alias feature

For this test, we performed the following steps:

- Start the cluster

- ▶ Cause a network adapter failure
- ▶ Stop the cluster

We start the cluster using the starting procedure from Scenario 1.

Cluster status

The cluster is up and running and all resources are up on their own nodes. Verify the HACMP subsystems by using the `lssrc -g cluster` command. Example 2-24 shows how the cluster daemons are started.

Example 2-24 Output of lssrc -g cluster command

[c37flrp01]:/ # lssrc -g cluster			
Subsystem	Group	PID	Status
clstrmgrES	cluster	23206	active
clsmuxpdES	cluster	20600	active
clinfoES	cluster	35276	active

Check the cluster status using the `clstat` command (see Figure 2-42).

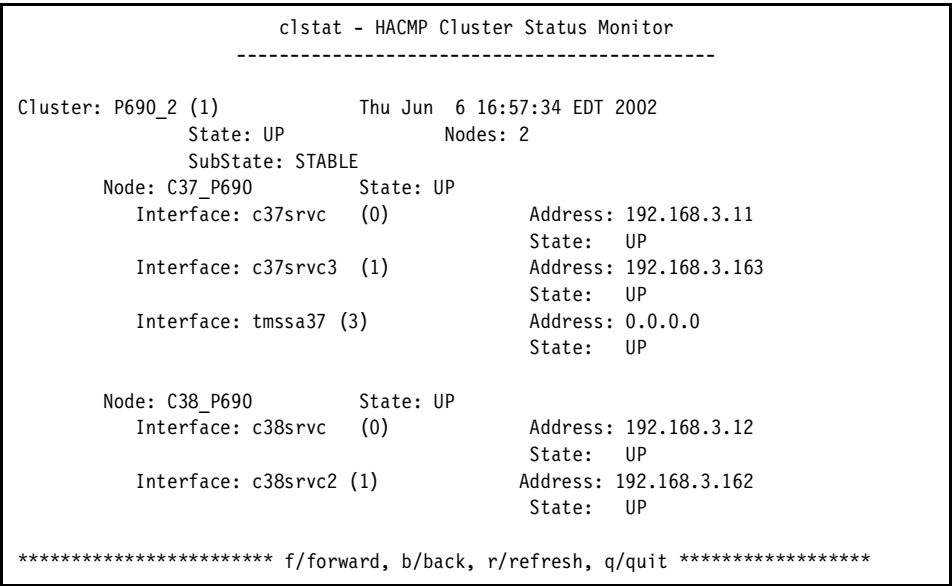


Figure 2-42 Cluster status

Check the network interfaces on the both nodes using the `netstat -i` command (see Example 2-25 on page 136). Services and rotating IP labels are configured correctly as alias on the boot adapter.

Example 2-25 Output of netstat -i command with IP alias

```
on first node
[c37rp01]:/ # netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
...
en3 1500 link#6 0.2.55.6a.bd.39 63404 0 47982 0 0
en3 1500 192.168.3.3 c37boot2 63404 0 47982 0 0
en3 1500 192.168.3.1 c37srcv2 63404 0 47982 0 0
en3 1500 192.168.3.1 crotsrvc 63404 0 47982 0 0
...

on second node
[c38rp01]:/ # netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en3 1500 link#6 0.2.55.6a.c1.47 63737 0 47598 0 0
en3 1500 192.168.3.3 c38boot2 63737 0 47598 0 0
en3 1500 192.168.3.1 c38srcv2 63737 0 47598 0 0
...
```

Check the resource group using the **clfindres** command (see Example 2-26). The resource groups are up and acquired by the owner nodes.

Example 2-26 Output of clfindres command

```
[c37rp01]:/ # clfindres
GroupName Type State Location Sticky Loc
-----
C37_CAS_01 cascading UP C37_P690
C37_CAS_02 cascading UP C37_P690
C38_CAS_01 cascading UP C38_P690
C38_CAS_02 cascading UP C38_P690
ROT_01 rotating UP C37_P690
```

Cluster status after Ethernet adapter failure occurs

When the adapter failure occurred on the C37_P690 node, we noticed that service addresses from node C37_P690 linked to the en3 adapter have been taken over by the backup node and added as aliases. The **netstat** command output is given (Example 2-27) when this event occurs.

Example 2-27 Output of netstat -i command after adapter failure

```
on first node
[c37rp01]:/ # netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
...
en3 1500 link#6 0.2.55.6a.bd.39 65103 0 49533 0 0
en3 1500 192.168.3.3 c37boot02 65103 0 49533 0 0
```

...

on second node

[c38rp01]:/s37fs01 # netstat -i

Name	Mtu	Network	Address	Ipkts	Ierrs	Opkts	Oerrs	Coll
...								
en3	1500	link#6	0.2.55.6a.c1.47	65379	0	49111	0	0
en3	1500	192.168.3.3	c38boot02	65379	0	49111	0	0
en3	1500	192.168.3.1	c38srcv02	65379	0	49111	0	0
en3	1500	192.168.3.1	c37srcv02	65379	0	49111	0	0
en3	1500	192.168.3.1	c37srcv03	65379	0	49111	0	0
...								

Check the resource group using the **clfindres** command (see Example 2-28). Note that the resource groups linked to the mentioned addresses were moved to the backup node. We stopped the cluster using the stopping procedure from Scenario 1.

Example 2-28 Output of clfindres command

GroupName	Type	State	Location	Sticky Loc
-----	-----	-----	-----	-----
C37_CAS_01	cascading	UP	C37_P690	
C37_CAS_02	cascading	UP	C38_P690	
C38_CAS_01	cascading	UP	C38_P690	
C38_CAS_02	cascading	UP	C38_P690	
ROT_01	rotating	UP	C38_P690	

Example 2-29 shows the events handled when adapter failure occurs.

Example 2-29 Network_down event in cluster.log file

```
...
Jun  6 17:14:00 c37f1rp01 HACMP for AIX: EVENT START: network_down C37_P690
PUBL_NET_2
...
Jun  6 17:14:08 c37f1rp01 HACMP for AIX: EVENT START: rg_move C37_P690 3
...
Jun  6 17:14:10 c37f1rp01 HACMP for AIX: EVENT START: release_service_addr
c37srcv02
...
Jun  6 17:14:20 c37f1rp01 HACMP for AIX: EVENT COMPLETED: rg_move_complete
C37_P690 3
Jun  6 17:14:28 c37f1rp01 HACMP for AIX: EVENT START: rg_move C37_P690 5
...
Jun  6 17:14:29 c37f1rp01 HACMP for AIX: EVENT START: release_service_addr
c37srcv03
```

2.4.11 Scenario 5: Integrating ESS storage into HACMP

The IBM 2105 Enterprise Storage Server (ESS), also known as “Shark”, is a high-end disk storage server. Its main feature is its performance in accessing the disks, using a huge RAM memory as a cache, and resource allocation flexibility to different servers; by mapping physical disks to logical disks, they can be reallocated to different servers by running simple commands.

The high availability of such a system is very important and the redundancy of the ESS internal components are handled by an ESS internal design.

Disk failure	Disks are handled internally by the ESS internal design.
Disk controller	There are two internal controllers implemented in every ESS that can overtake each other's function. This is also implemented in the ESS internal design.
Host adapters	The adapters are handled internally by the ESS internal design.
Connection path	The connection path from the ESS to the server needs to be configured with redundancy. This is the responsibility of the system architect to ensure that there are at least two independent paths from the server, or each server, to the ESS. While SCSI configurations are quite straightforward, the FC connections need additional considerations.
Device driver	The ESS device driver for the AIX operating system is delivered together with the IBM Subsystem Device Driver (SDD) product used for load balancing and high availability of the connection paths.

Note: The SDD device driver handles connection path failures to the ESS only. It introduces a new disk device (*vpath*). The administration of these devices are different from the administration of the standard disk device (*hdisk*).

See Figure 2-43 on page 139 for an example of interconnecting the ESS in a two-node environment. The connection paths are marked with the numbers 1, 2, 3, and 4. Connection paths 1 and 2 go from the ESS to the Node c37 and paths

3 and 4 goes to node c38. Each node in this case has two connection paths to each disk.

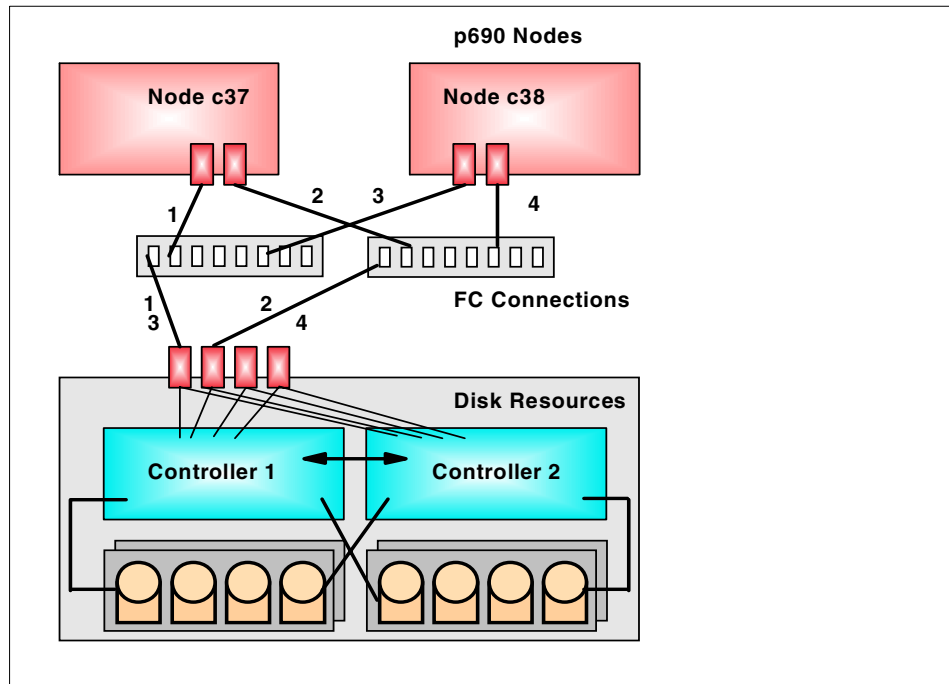


Figure 2-43 ESS Interconnection diagram

Planning

The intention of the planning is to have available shared disk resources configured in the ESS and allocated to two nodes in two fibre channel paths. The diagram in Figure 2-43 shows our scenario. The ESS is configured with logical disks visible to both servers. No other server can access these logical disks - this is done by FC zoning on the FC switches. We use the configuration of HACMP topology, as described in Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87.

Our goal is to give an example on how to integrate an ESS disk resource in the p690 HACMP cluster and how to perform a simple administrative task, like adding a new logical disk to a shared volume group.

Installation

We assume the installation of AIX and HACMP is already done, as described in Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87.

The IBM ESS requires the following filesets to be installed:

- ▶ devices.fc.*
- ▶ ibm2105.rte
- ▶ ibmSdd_510nchacmp.rte

Installation of base device support

Installation of the base device support is performed by the **cfgmgr -i** command from the base AIX media and from the supplied AIX Maintenance Level media. The device drivers are installed automatically for devices detected by the **cfgmgr** command. If the device drivers are installed correctly, we can see the Fibre Channel "fcs", "fscsi", "hdisk", and disk devices. However, the ESS disks may appear as "Other FC disk".

Installation of the Host Attachment scripts

The Host Attachment scripts add definitions for the ESS disks are in the operating system. The script can be downloaded from the official ESS support Web site:

<http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/2105>

Follow the link Downloads - Host Attachment scripts to obtain the script. After you download the file, read the associated instructions for installing the fileset. You need to un-tar the file to a temporary directory and install the **ibm2105.rte** fileset by using the **installp** command.

Example 2-30 shows the **lslpp -L** and **lsdev -Cc disk** command outputs before the host attachment scripts are installed.

Example 2-30 Lslpp and lsdev commands before ibm2105.rte installed

```
lslpp -L ibm2105.rte
lslpp: 0504-132 Fileset ibm2105.rte not installed
lsdev -Cc disk
hdisk2 Available 20-58-01 Other FC SCSI Disk Drive
```

Example 2-31 shows the **lslpp -L** command output after the host attachment scripts are installed.

Example 2-31 Check the ibm2105.rte fileset after installation

```
lslpp -L ibm2105.rte
  Fileset                                Level  State  Type  Description (Uninstaller)
-----
  ibm2105.rte                          32.6.100.9   C    F    IBM 2105 Disk Device
lsdev -Cc disk
hdisk2 Available 20-58-01      IBM FC 2105F20
```

Installation of the Subsystem Device Driver (SDD)

The SDD is the device driver responsible for FC or SCSI path load balancing and failover. The hdisk devices we have put in the operating system by installing the device drivers see each disk as many times as many paths exist to the disk. The SDD maps the individual hdisk devices to one vpath device. The vpath device driver is created by the SDD. If the SDD is not installed, you cannot see the vpath devices. The SDD can be downloaded from the official ESS support Web site:

<http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/2105>

Follow the link Downloads - Subsystem Device Driver to obtain the SDD.

Note: There are two versions of the SDD for each version of AIX. One is for non-cluster environments and concurrent clusters and the second is for non-concurrent clusters. Read the instructions on the Web site and download the correct version of the SDD.

The installation procedure involves un-tarring of the file to a temporary directory and installing it by using the **installp** command.

Example 2-32 shows the **lslpp**, **lsdev**, and **lspv** commands' output after the SDD is installed.

Example 2-32 SDD installation check

```
# lslpp -L ibmSdd_510nchacmp.rte
Fileset              Level State Type Description (Uninstaller)
-----
ibmSdd_510nchacmp.rte 1.3.1.3  C   F   IBM Subsystem Device Driver AIX
                               V51 for non-concurrent HACMP

# lsdev -Cc disk
...
hdisk2  Available 20-58-01      IBM FC 2105F20
hdisk3  Available 20-58-01      IBM FC 2105F20
hdisk4  Available 20-58-01      IBM FC 2105F20
...
vpath0  Available                Data Path Optimizer Pseudo Device Driver
vpath1  Available                Data Path Optimizer Pseudo Device Driver

# lspv
...
hdisk2      00600854728c8b77      None
hdisk3      00600854728c8c77      None
hdisk4      none                  None
...
vpath0      none                  None
vpath1      none                  None
```

...

Create a volume group for use by HACMP

The installation of the SDD is not enough, so additional configuration tasks should be performed. As you can see in Example 2-33, hdisk2 and hdisk10 have the same PVID. Note that none of the vpath devices have an assigned PVID. There are SDD tools to create, modify, and transform hdisk PVIDs to vpath PVIDs. However, this affects HACMP planning, installation, and administration.

Verify the configuration of the FC and disk environment first, as shown in Example 2-33.

Example 2-33 Verification of vpath devices

```
# datapath query adapter
Active Adapters :2
```

Adpt#	Adapter Name	State	Mode	Select	Errors	Paths	Active
0	fscsi0	NORMAL	ACTIVE	0	0	12	0
1	fscsi1	NORMAL	ACTIVE	0	0	12	0

```
# datapath query device
```

```
DEV#: 0  DEVICE NAME: vpath0  TYPE: 2105F20  SERIAL: 700FCA16
POLICY:  Optimized
```

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk2	CLOSE	NORMAL	0	0
1	fscsi0/hdisk10	CLOSE	NORMAL	0	0
2	fscsi1/hdisk14	CLOSE	NORMAL	0	0
3	fscsi1/hdisk22	CLOSE	NORMAL	0	0

```
DEV#: 1  DEVICE NAME: vpath1  TYPE: 2105F20  SERIAL: 701FCA16
POLICY:  Optimized
```

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk3	CLOSE	NORMAL	0	0
1	fscsi0/hdisk11	CLOSE	NORMAL	0	0
2	fscsi1/hdisk15	CLOSE	NORMAL	0	0
3	fscsi1/hdisk23	CLOSE	NORMAL	0	0

```
DEV#: 2  DEVICE NAME: vpath2  TYPE: 2105F20  SERIAL: 702FCA16
POLICY:  Optimized
```

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk4	CLOSE	NORMAL	0	0
1	fscsi0/hdisk12	CLOSE	NORMAL	0	0
2	fscsi1/hdisk16	CLOSE	NORMAL	0	0

```

3          fscsil/hdisk24    CLOSE    NORMAL    0          0
...

# lsdev -Cc disk
hdisk0 Available 10-60-00-0,0 16 Bit SCSI Disk Drive
hdisk1 Available 10-60-00-1,0 16 Bit SCSI Disk Drive
hdisk2 Available 10-70-01      IBM FC 2105F20
hdisk3 Available 10-70-01      IBM FC 2105F20
...
hdisk14 Available 20-58-01      IBM FC 2105F20
hdisk15 Available 20-58-01      IBM FC 2105F20
...
vpath0 Available                Data Path Optimizer Pseudo Device Driver
vpath1 Available                Data Path Optimizer Pseudo Device Driver
...

# lspv
hdisk0          006017150625de18          rootvg
hdisk1          00601457005321e7          rootvg
hdisk2          none                     None
hdisk3          none                     None
hdisk4          none                     None
...
vpath0          none                     None
vpath1          none                     None
vpath2          none                     None
...

```

The list in Example 2-33 on page 142 is the same on both nodes, except the hdisk0 and hdisk1 are local to each node and contain the rootvg. From this listing of disks, we have to chose disks participating in the volume group.

The best place to check the available devices is the **datapath query device** output, in order to see the relations between vpath and hdisk devices, and the lspv output from both nodes, to check which PVIDs are on which disks.

Note: Check that the PVID are on vpath or hdisk devices, but not on both. If you removed all hdisks and vpaths of the ESS and rebooted the system or run the **cfgmgr** command for each FC adapter, the PVIDs appear on the hdisk devices. When we create a volume group, we need to select vpath devices as physical volumes belonging to the volume group. This is done by special commands.

As an example, we will select vpath0 to create a volume group named vgha05. We use SMIT to create the volume group, but the same procedure can be done by using the **mkvg4vp** command.

Run the `smit vg` fast path and select the menu in Figure 2-44.

Add a Volume Group with Data Path Devices

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

VOLUME GROUP name

Physical partition SIZE in megabytes

* PHYSICAL VOLUME names

Activate volume group AUTOMATICALLY
at system restart?

Volume group MAJOR NUMBER

[Entry Fields]

[vgha05]

4

[vpath0] +

no +

[105]

+#

Esc+1=Help

Esc+2=Refresh

Esc+3=Cancel

Esc+4=List

Esc+5=Reset

Esc+6=Command

Esc+7>Edit

Esc+8=Image

Esc+9=Shell

Esc+0=Exit

Enter=Do

Figure 2-44 Add a Volume Group with Data Path Devices

Fill in the following menu items:

VOLUME GROUP name	Select a unique name for the volume group.
Physical partition SIZE in megabits	The value does not affect the HACMP configuration. Choose any value or refer to the base AIX documentation.
PHYSICAL VOLUME names	Use the List function and select the correct vpath devices. In our example, it is vpath0.
Activate volume group AUTOMATICALLY at system restart	Select no. This is a requirement of the HACMP, because HACMP activates the volume group and not AIX.
Volume group MAJOR NUMBER	Select a short positive integer number. In our example, it is 105.

Create logical volumes and file systems

When creating the logical volumes and file systems, follow the HACMP rules for creating shared storage (refer to *HACMP for AIX 4.5 Administration Guide*, SC23-4279). The procedure is the same for p690 and ESS (like any other AIX system). We created two JFS logs and two file systems to demonstrate JFS and JFS2 in an example. See the `1svg -1 vgha05` output in Example 2-34 on page 145.

Example 2-34 Volume group listing

```
# lsvg -l vgha05
vgha05:
LV NAME          TYPE      LPs   PPs   PVs   LV STATE      MOUNT POINT
logha0501        jfslog    2     2     1     closed/syncd  N/A
lvha0501         jfs       10    10    1     closed/syncd  /ha0501
logha0502        jfs2log   2     2     1     closed/syncd  N/A
lvha0502         jfs2      10    10    1     closed/syncd  /ha0502
```

Synchronize volume group information

The initial synchronization of the volume group information differs when implementing ESS vpath devices in a cluster. The difference is that the PVIDs are allocated in the ODM by default to the hdisk devices, but the volume group expects them on the vpath devices. However, there is a workaround to solve this difference:

1. Remove hdisk and vpath devices they are associated with the physical volumes containing the shared volume group. The reason for this step is to update the PVID of the disks, as we assume we changed the PVID of the disks from none to a new value. As this is a new installation, we remove and configure all disks in Example 2-35.

Example 2-35 Simple script to remove all hdisk and vpath devices

```
for i in `lsdev -Cc disk | grep -v rootvg | awk '{print $1}'`
do
    rmdev -dl $i
done
```

2. Configure the hdisk and vpath devices on the secondary node by rebooting or using the **cfgmgr** command. If using the **cfgmgr** command, run it for each FC adapter and then run it globally, as shown in Example 2-36.

Example 2-36 Configure hdisk and vpath devices

```
# cfgmgr -l fcs0
# cfgmgr -l fcs1
# cfgmgr
# lspv
...
hdisk2          00600854728c8b77          None
...
hdisk10         00600854728c8b77          None
...
hdisk14         00600854728c8b77          None
...
hdisk22         00600854728c8b77          None
...
```

vpath0	none	None
...		

3. Check that the paths are available from the hdisk devices to the vpath device (see Example 2-37). The state of the paths is in the CLOSE state, because the volume group is not varied on this node. If you miss any hdisk device that also belongs to the vpath, repeat the steps from Example 2-36 on page 145.

Example 2-37 Check paths from vpath to hdisks

```
# datapath query device
```

```
DEV#:    0  DEVICE NAME: vpath0  TYPE: 2105F20  SERIAL: 700FCA16
POLICY:   Optimized
```

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk2	CLOSE	NORMAL	0	0
1	fscsi0/hdisk10	CLOSE	NORMAL	0	0
2	fscsi1/hdisk14	CLOSE	NORMAL	1	0
3	fscsi1/hdisk22	CLOSE	NORMAL	0	0

4. Import the volume group from any hdisk that belongs to the volume group by using the **importvg** command or SMIT. Check that the PVIDs for these disks on the secondary node correspond to the PVIDs on the primary node for the shared disks. Use the **lspv** command from Example 2-36 on page 145 to check the PVIDs; see Example 2-38 for information on importing the volume group. Remember to change the volume group to be not activated automatically at system restart.

Example 2-38 Import the shared volume group

```
# importvg -y havg05 -V 105 hdisk2
# chvg -a n vgha05
# varyoffvg vgha05
# lspv
```

hdisk2	00600854728c8b77	havg05
...		
hdisk10	00600854728c8b77	havg05
...		
hdisk14	00600854728c8b77	havg05
...		
hdisk22	00600854728c8b77	havg05
...		
vpath0	none	None
...		

5. Reallocate the volume group from the hdisk devices to the vpath device using the hd2vp utility from the SDD device driver. Also, check that the volume group correctly uses the paths of the **datapath query device** command; you should see the state of OPEN to each disks for the vpath device. See Example 2-39 on the volume group reallocation.

Example 2-39 Reallocate PVID using hd2vp command

```
# hd2vp havg05
# lspv
...
hdisk2          none                               None
...
hdisk10         none                               None
...
hdisk14         none                               None
...
hdisk22         none                               None
...
vpath0          00600854728c8b77                  vgha05
...
# varyonvg vgha05
# datapath query device

DEV#:  0  DEVICE NAME: vpath0  TYPE: 2105F20  SERIAL: 700FCA16
POLICY:  Optimized
=====
Path#      Adapter/Hard Disk  State  Mode  Select  Errors
  0         fscsi0/hdisk2     OPEN   NORMAL  10      0
  1         fscsi0/hdisk10    OPEN   NORMAL  12      0
  2         fscsi1/hdisk14    OPEN   NORMAL  16      0
  3         fscsi1/hdisk22    OPEN   NORMAL  14      0
# varyoffvg vgha05
```

At this time, the shared volume group is ready to be defined in the HACMP configuration.

Configuration in the HACMP

When configuring HACMP we assume the environment is already configured, as described in Section 2.4.7, “Scenario 1: Cluster with 2 Ethernet and SSA storage” on page 87, and we are only adding the shared volume group of the ESS to the HACMP.

For the HACMP configuration, there is only one entry we need to consider - the HACMP resource group. However, this entry requires more steps to be performed.

Perform the following steps:

1. Discover the new volume group configuration by using the SMIT HACMP interface. Select the menus in the following order: **Cluster Configuration -> Cluster Resources -> Discover Current Volume Group Configuration -> Cluster-wide Configuration**.
2. Add the ESS shared volume group to the HACMP resource group by using SMIT HACMP interface. Select the menus in the following order: **Cluster Configuration -> Cluster Resources -> Change/Show Resources/Attributes for a Resource Group**. Select the resource group you want to add the shared volume group to.

You will receive the well-known menu for configuring the resource group attributes, where you need to fill in the Volume Groups item. Use the List function of SMIT to generate a list of shared volume groups and select the volume group (see Figure 2-45).

Change/Show Resources/Attributes for a Resource Group

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

[MORE...11]

[Entry Fields]

Filesystems/Directories to NFS mount	<input type="checkbox"/>	+
Network For NFS Mount	<input type="checkbox"/>	+
Volume Groups	[vgha05]	+
Concurrent Volume groups	<input type="checkbox"/>	+
Raw Disk PVIDs	<input type="checkbox"/>	+
Connections Services	<input type="checkbox"/>	+

+-----++

Volume Groups

+-----+

Move cursor to desired item and press Esc+7.
ONE OR MORE items can be selected.
Press Enter AFTER making all selections.

+-----+

> vgha05

[M]

Es| Esc+1=Help

Esc+2=Refresh

Esc+3=Cancel

Es| Esc+7=Select

Esc+8=Image

Esc+0=Exit

Es| Enter=Do

/=Find

n=Find Next

Es+-----+

Figure 2-45 Add the shared volume group to the HACMP resource group

3. Synchronize the cluster resources by using the `cldare` command or SMIT, and start the cluster by using the `rc.cluster` script or SMIT if the cluster is not running. Adding a resource, for example a volume group, can be done while the cluster is running (see Example 2-40 on page 149).

Example 2-40 Synchronizing and starting the cluster

```
# /usr/es/sbin/cluster/utilities/cldare -r
...
# /usr/es/sbin/cluster/etc/rc.cluster -boot -N -i
...
```

At this point the ESS shared volume group is fully integrated in the HACMP environment.

Testing the configuration

Testing the functions of the ESS shared volume group integration into HACMP is done by testing the scenario of the resource group move, node failure, adapter failure, and disk failure. In addition, we need to also test the administration of the shared volume group through the C-SPOC.

Resource group move

The resource group move, or cluster shutdown with takeover, behaves from the user's view as a classic volume group (described in Section 2.4.7, "Scenario 1: Cluster with 2 Ethernet and SSA storage" on page 87). However, there are additional tests for the vpath devices during takeover, as they need the SCSI_RESERVE to be handled differently. There is no need to do any SCSI_RESET on the SCSI_RESERVE, because the volume group is released correctly during the resource group move.

Node failure

The node failure is detected and handled by activating the resource group on the secondary node. In this case, the disk resources must be reset, as the SCSI_RESERVE could not be cleared by the crashed node. The takeover scripts detects that the volume group is defined on vpath devices and handles the vpaths using special utilities.

Adapter failure

Adapter failures are handled by the SDD device driver and HACMP does not need to do anything about the failure of the adapter. Example 2-41 shows a failing FC adapter.

Example 2-41 Example of datapath query device output

```
DEV#: 0  DEVICE NAME: vpath0  TYPE: 2105F20  SERIAL: 700FCA16
POLICY:  Optimized
=====
Path#      Adapter/Hard Disk  State   Mode    Select  Errors
  0          fscsi0/hdisk2    OPEN   NORMAL   17850     0
  1          fscsi0/hdisk10  OPEN   NORMAL   17976     0
  2          fscsi1/hdisk14  OFFLINE FAILED   17823   255
```

In case all FC adapters fail, the HACMP detects the loss of the volume group and initiates a move of the resource group to the backup node.

Adding a disk to shared volume group

Adding an ESS disk to a shared volume group involves a different procedure from the standard procedure. The difference is that we have to add a vpath device to the volume group, but the vpath devices does not have PVIDs until a volume group is configured on them using the SDD utilities. First, we have to configure the PVID on the vpath device and then we can use the vpath devices for HACMP. If the PVID is defined once on the vpath device, the vpath device is seen by HACMP as a regular disk device. We suggest the following procedure:

1. Check the configuration of the hdisk to vpath mappings by using the **datapath query device** and **lspv** commands, as shown in Example 2-42. We can see that the vpath1 device is not allocated yet; let us use it for our example.

Example 2-42 Output of datapath query device command

```
# datapath query device
DEV#:  0  DEVICE NAME: vpath0  TYPE: 2105F20   SERIAL: 700FCA16
POLICY:   Optimized
=====
Path#          Adapter/Hard Disk   State   Mode    Select   Errors
  0              fscsi0/hdisk2     OPEN   NORMAL   26913     0
  1              fscsi0/hdisk10    OPEN   NORMAL   26870     0
  2              fscsil/hdisk14    OPEN   NORMAL   26716     0
  3              fscsil/hdisk22    OPEN   NORMAL   26729     0

DEV#:  1  DEVICE NAME: vpath1  TYPE: 2105F20   SERIAL: 701FCA16
POLICY:   Optimized
=====
Path#          Adapter/Hard Disk   State   Mode    Select   Errors
  0              fscsi0/hdisk3     CLOSE  NORMAL     0         0
  1              fscsi0/hdisk11    CLOSE  NORMAL     0         0
  2              fscsil/hdisk15    CLOSE  NORMAL     0         0
  3              fscsil/hdisk23    CLOSE  NORMAL     0         0

...
# lspv
hdisk2          none                      None
hdisk3          none                      None
...
hdisk10         none                      None
hdisk11        none                      None
...
hdisk14         none                      None
hdisk15        none                      None
```

```

...
hdisk22      none                      None
hdisk23     none                     None
...
vpath0       00600854728c8b77         vgha05
vpath1     none                     None
...

```

2. Create a temporary volume group (*tempvg*) using the SDD utility `mkvg4vp` or the SMIT. Start the SMIT using the `smit vg` fast path and select the Add a Volume Group with Data Path Devices menu. Fill in any name as the volume group name and select `vpath1` as the Physical volume name (see Figure 2-46).

Add a Volume Group with Data Path Devices

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

	[Entry Fields]	
VOLUME GROUP name	[tempvg]	
Physical partition SIZE in megabytes	4	+
* PHYSICAL VOLUME names	[vpath1]	+
Activate volume group AUTOMATICALLY at system restart?	yes	+
Volume group MAJOR NUMBER	[]	+#

Esc+1=Help

Esc+5=Reset

Esc+9=Shell

Esc+2=Refresh

Esc+6=Command

Esc+0=Exit

Esc+3=Cancel

Esc+7=Edit

Enter=Do

Esc+4=List

Esc+8=Image

Figure 2-46 SMIT - Add a Volume Group with Data Path Devices

3. Verify that the PVID is assigned to the `vpath1` device for the `tempvg` and that all the paths are opened correctly. Use the `datapath query device` and the `lspv` command (see Example 2-43).

Example 2-43 Output of `datapath query device` command

```

# datapath query device
DEV#: 0  DEVICE NAME: vpath0  TYPE: 2105F20  SERIAL: 700FCA16
POLICY:  Optimized
=====
Path#      Adapter/Hard Disk  State   Mode    Select  Errors
  0         fscsi0/hdisk2     OPEN   NORMAL   26913    0
  1         fscsi0/hdisk10  OPEN   NORMAL   26870    0
  2         fscsi1/hdisk14  OPEN   NORMAL   26716    0
  3         fscsi1/hdisk22  OPEN   NORMAL   26729    0

```

```

DEV#: 1  DEVICE NAME: vpath1  TYPE: 2105F20  SERIAL: 701FCA16
POLICY:  Optimized
=====
Path#          Adapter/Hard Disk  State  Mode  Select  Errors
  0             fscsi0/hdisk3    OPEN  NORMAL    0        0
  1             fscsi0/hdisk11    OPEN  NORMAL    0        0
  2             fscsi1/hdisk15    OPEN  NORMAL    0        0
  3             fscsi1/hdisk23    OPEN  NORMAL    0        0
...
# lspv
hdisk2          none                      None
hdisk3          none                      None
...
hdisk10         none                      None
hdisk11        none                      None
...
hdisk14         none                      None
hdisk15        none                      None
...
hdisk22         none                      None
hdisk23        none                      None
...
vpath0          00600854728c8b77              vgha05
vpath1        00600882614b9fbc          tempvg
...

```

4. Export the tempvg using the **exportvg** command, as shown in Example 2-44.

Example 2-44 Use of the varyoffvg and exportvg commands.

```

# varyoffvg tempvg
# exportvg tempvg

```

5. On the secondary node, remove the hdisk definitions belonging to the vpath1, that is, hdisk3, hdisk11, hdisk15, and hdisk23, and then remove the vpath1 definition using the **rmdev** command, as shown in the Example 2-45. After that, configure the devices running the **cfgmgr** command for each FC adapter and for the vpaths, as shown in Example 2-45.

Example 2-45 Reconfigure the hdisk and vpath devices on secondary node

```

# rmdev -dl hdisk3
# rmdev -dl hdisk11
# rmdev -dl hdisk15
# rmdev -dl hdisk23

# cfgmgr -l fcs0
# cfgmgr -l fcs1

```

```
# cfmgr
```

6. Verify the PVID appears on the hdisk devices and the vpath devices have PVID none (see Example 2-46).

Example 2-46 Output of the lspv command after cfmgr was run

hdisk3	00600882614b9fbc	None
...		
hdisk11	00600882614b9fbc	None
...		
hdisk15	00600882614b9fbc	None
...		
hdisk23	00600882614b9fbc	None
...		
vpath1	none	None
...		

7. Import the volume group from any of hdisk device with the correct PVID and check the output of the **lspv** command, as shown in Example 2-47

Example 2-47 Import of the temporary volume group

```
# importvg -y tempvg hdisk3
# lspv
```

hdisk3	00600882614b9fbc	tempvg
...		
hdisk11	00600882614b9fbc	tempvg
...		
hdisk15	00600882614b9fbc	tempvg
...		
hdisk23	00600882614b9fbc	tempvg
...		
vpath1	none	None
...		

8. Convert the PVID from hdisk devices to the vpath device using the **hd2vp** command and verify the result by using the **lspv** command, as shown in Example 2-48.

Example 2-48 PVID conversion by the hd2vp command

```
# hd2vp tempvg
# lspv
```

hdisk3	none	None
...		
hdisk11	none	None
...		
hdisk15	none	None

...		
hdisk23	none	None
...		
vpath1	00600882614b9fbc	tempvg
...		

9. Varyoff and export the volume group using the **exportvg** command. Notice that the PVID remains on the vpath device and does not move back to the hdisk device. The vpath device is ready for HACMP now.
10. Discover the new volume group configuration using the **clharvest_vg -w** HACMP command or use SMIT HACMP and select the SMIT menu items **Cluster Configuration -> Cluster Resources -> Discover Current Volume Group Configuration -> Cluster-wide Configuration**. Run this task on both nodes of the cluster.
 - a. Use HACMP C-SPOC on the node where the resource group is active to add the vpath to the vgha05 volume group. Use the SMIT List function to select the PVID of the vpath.

In the SMIT HACMP screen, select the following menu items: **Cluster System Management -> Cluster Logical Volume Manager -> Shared Volume Groups -> Set Characteristics of a Shared Volume Group -> Add a Physical Volume to a Shared Volume Group**.
 - b. From a list, select the volume group you want to add the vpath to.
 - c. From the next list, select the vpath1 we want to add to the volume group.
 - d. Finally, confirm the SMIT screen shown in Figure 2-47. Ignore the warning messages.

Add a Physical Volume to a Shared Volume Group

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

	[Entry Fields]
Resource Group Name	ESS_RG_05
VOLUME GROUP name	vgha05
Reference node	C49N05
PHYSICAL VOLUME names	vpath1

Esc+1=Help	Esc+2=Refresh	Esc+3=Cancel	Esc+4=List
Esc+5=Reset	Esc+6=Command	Esc+7=Edit	Esc+8=Image
Esc+9=Shell	Esc+0=Exit	Enter=Do	

Figure 2-47 SMIT - Add a Physical Volume

At this point, we are ready to add a new vpath to the shared volume group, and have performed all the necessary steps from the HACMP point of view.

Other tasks performed on the system of HACMP, p690, and ESS

We performed also other administrative tasks using C-SPOC, like removing vpaths, extending JFS and JFS2 file systems, and so on. All these tasks behaved as though performed on non-ESS or non-p690 systems. For this reason, we do not describe them.



HACWS: An HACMP Application for Cluster 1600

In this chapter, we discuss the steps to configure High Availability Control Workstation (HACWS) Version 3.4 with AIX 5L Version 5.1 and PSSP 3.4 and HACMP 4.5. We examine this from the perspective of a “best practice” procedure, examining, in detail, a number of common configurations and taking the reader through the configuration steps in detail.

Important: The APAR number required to support HACWS on HMC-enabled servers in an IBM @server Cluster 1600 configuration (that is, Regatta attached nodes) is *APAR IY33906*.

3.1 HACWS

HACWS is designed to remove the control workstation (CWS) as a single point of failure in an IBM eServer pSeries Cluster 1600 System. With HACMP 4.5, this functionality has been extended to include SP Systems with attached p690 servers

Planning for a highly available control workstation requires planning for both the hardware and the software. There are a number of ways to increase the availability of the control workstation, but with the HACWS option, all hardware and software components are redundant, which allows for recovery from any single failure.

The HACWS component is based on the HACMP licensed program. It uses HACMP running on two control workstations in a two node rotating resource group configuration. It requires an external DASD that can be accessed non-concurrently by each of the control workstations for storage of the SP-related data. There is also a feature that allows both of the control workstations to be connected to the frame supervisor card in each SP frame. HACMP provides the usual detection, notification, and recovery of any single failure of a component of the control workstation.

The recommended method for installing HACWS is to first install and configure your SP complex with a single control workstation, then add the backup control workstation to the configuration. This ordered approach allows testing of each step along the way, and an easy back-out strategy should there be any problems. However, there are some planning decisions that can be made at the initial install of the SP complex that will facilitate the integration of the backup control workstation.

Note: While the inactive control workstation may be used to run other applications, this is not recommended, and it should be thoroughly tested so that these applications do not hinder the inactive control workstation's ability to take over from the active control workstation in a time of failure.

There are some restrictions on the use of high availability control workstations (See Section 3.7, "Considerations" on page 222). The *PSSP Read this First* document, downloadable from http://www.ibm.com/servers/eserver/pseries/library/sp_books/pssp.html, has a list of servers that are supported as control workstations. Also, check with IBM Marketing for the latest information concerning what servers are supported for HACMP.

HACWS is supported for classic SP Frames and p690 attached servers only, as there is no method available for connecting SP attached servers (like S70/H80 type of servers) to the backup control workstation. See Section 3.7, “Considerations” on page 222 for more information.

3.2 Definitions

The following definitions are required to fully understand the operation of HACWS:

Primary control workstation	The control workstation that was installed initially with the SP complex.
Backup control workstation	The control workstation that was added during the install of HACWS.
Active control workstation	The control workstation in the HACWS cluster that is currently running the HACWS rotation resource group, which includes the PSSP applications, including the SDR and hardmon.
Inactive control workstation	The workstation in the HACWS cluster that is acting as the “Hot Standby”.

3.3 Requirements

This section describes the hardware and software requirements for planning a HACWS for a Cluster 1600 configuration.

3.3.1 Hardware requirements

The design logic behind HACWS is to remove single points of failure from the SP control workstation. The SP system looks similar, except that there are now two control workstations connected to the SP Ethernet network and frame supervisor. The frame supervisor or tty network is modified to add a standby link for the backup control workstation (see Figure 3-1 on page 160).

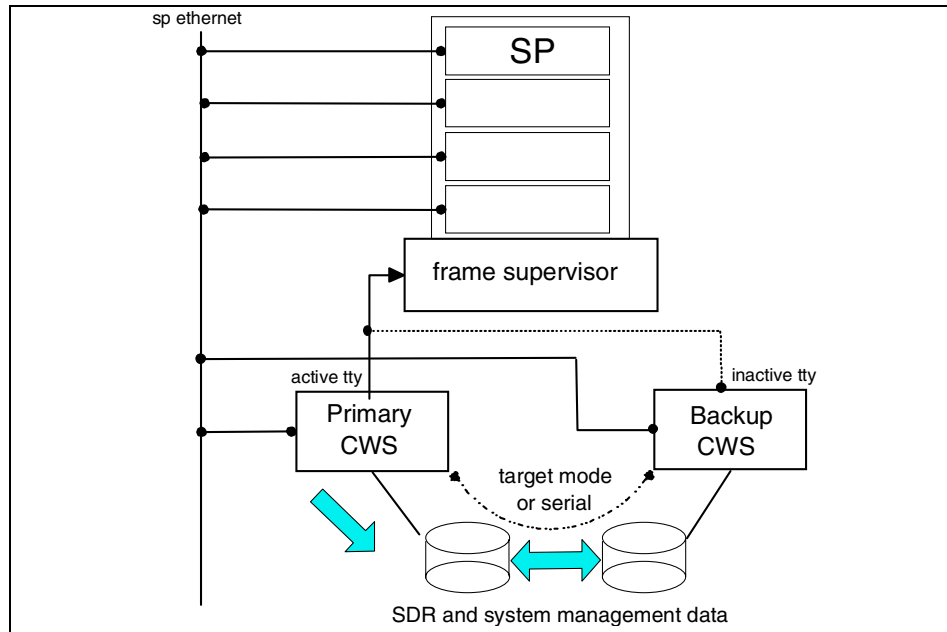


Figure 3-1 Highly available control workstation configuration

The following points need to be considered in relation to the hardware configuration:

As usual, mirroring data volumes does not negate the need to backup.

Data availability

To ensure that the data files are available to both control workstations, shared external storage is required, preferably mirrored across dual adapters for greater availability.

Frame supervisor tty link

The hardware feature #1245 provides a Y-cable option that must be installed so that both control workstations can connect to the frame supervisor card. This cable is not symmetric, so it is important to connect the primary control workstation to the correct side.

Non-IP HA network

HACMP requires a non-IP connection between the control workstations. This can either be a serial connection (if there are free ports), a target mode SCSI link, or a target mode SSA link (depending on the type of external storage used).

Standby network adapters

If there are slots available in each control workstation, then it is recommended that at least

the primary control workstation be configured with a standby network adapter.

Cluster with p690 servers The integration of the p690 into a cluster controlled by HACWS was free of complications. The only issue worth noting is that some customers have expressed concern about the security implications of having the control workstation open to HMC communications on the public SP administration network. In this case, a private network can be configured.

SP attached servers There are a number of functions that cannot be performed on an SP attached server from an active backup control workstation (see Chapter 3.7, “Considerations” on page 222).

For details on planning the hardware, see *RS/6000 SP: Planning, Vol. 1, Hardware and Physical Environment*, GA22-7280. This document describes the hardware components and cabling you need to install and run the HACWS software successfully.

3.3.2 Software requirements

The software requirements are:

- ▶ Two licenses for IBM C for AIX or the batch C and C++ compiler and run-time libraries of VisualAge C++ Professional 5.0.2.0 for AIX or later.
- ▶ Two licenses for HACMP 4.4.1, HACMP 4.5, or HACMP/ES 4.5. HACMP has its own requisite software.
- ▶ PSSP 3.4 optional component got HACWS. This comes with PSSP 3.4 as an optional component (the ssp.hacws fileset) and needs to be installed on both control workstations.

Note: While HACWS will work with all versions of AIX and PSSP, we chose AIX 5L Version 5.1 and PSSP 3.4, primarily to allow the inclusion of the p690 into the cluster.

3.4 Operation of HACWS

HACWS is really just a two node HACMP cluster with a rotating resource group. If the primary workstation fails, then HACMP moves the following resources to the backup control workstation:

- ▶ Shared DASD
- ▶ Service IP address
- ▶ PSSP control workstation applications (SDR, hardware monitoring, and so on)

Figure 3-1 on page 160 shows the layout of this configuration.

If the active control workstation fails, resources will fail over to the inactive control workstation. This will be disruptive and there will be a short interruption of service. The following events occur as part of the fail-over process:

- ▶ IP label for the active control workstation is configured on the inactive control workstation.
- ▶ If configured, the hardware address from the active control workstation will be configured on the inactive control workstation.
- ▶ Access to external storage is moved to the inactive control workstation and the file systems are mounted.
- ▶ PSSP applications started on the inactive control workstation and hardware monitoring is resumed.
- ▶ Clients can obtain services and data from the new active control workstation.

3.4.1 Components of HACWS

HACWS consists of the following HACMP definitions:

- ▶ Resource group
- ▶ Application server
- ▶ Custom cluster events

Rotating resource group - hacws_group1

The following resources are part of this resource group:

- ▶ Service label (two if using private network for HMC)
- ▶ The /spdata file system
- ▶ The datavg volume group (the volume group name is chosen by the user.)
- ▶ The hacws_apps application server (the application server name is chosen by the user.)

The hacws_apps application server

The application server (hacws_apps) uses the following scripts to start and stop the control workstation resources:

- ▶ Start: Use `/usr/sbin/hacws/spcw_apps -ua`, which sets this node as the active control workstation and starts the control workstations applications as follows:
 - The `/spdata/sys1/hacws/rc.syspar_aliases -add` adds any required aliases.
 - The sdr daemon is started.
 - The syspar_ctrl daemons are started.
 - The `/etc/rc.sp` script is run.
 - Stops and restarts sysctld.
 - Starts hardmon.
 - Starts spmgr and swt, if present.
 - Runs setup_server.
- ▶ Stop: Use `/usr/sbin/hacws/spcw_apps -di`, which sets this node to be the inactive control workstation and then stops the control workstation applications as follows:
 - Stops spmgr, syslogd, hardmon, sysctld, supfilesrv and sp_configd.
 - spspdm is stopped, if running.
 - The groups swt, emon, pman, hr, haem, hags, hb, hats and sdr are stopped.
 - Any NFS exports set by setup_server are removed
 - The `/spdata/sys1/hacws/rc.syspar_aliases -delete` script is run to remove an IP aliases.

Custom cluster events

As part of the HACWS installation, pre- and post-event scripts are configured in HACMP. A number of scripts are also copied into `/usr/sbin/hacws` and `/usr/sbin/hacws/events`:

- ▶ hacws_post_event: `/usr/sbin/hacws/hacws_post_event`
- ▶ hacws_pre_event: `/usr/sbin/hacws/hacws_pre_event`

These are generic scripts that are configured as pre and post events for all cluster events. They look in `/usr/sbin/hacws/events` for files with names in the following format and execute them if they exist.

hacws_pre_event looks for files that match:

- ▶ cluster_event_name.pre_pre_event in /var/adm/hacws/events
- ▶ cluster_event_name.pre_event in usr/sbin/hacws/events
- ▶ cluster_event_name.post_pre_event in /var/adm/hacws/events

hacws_post_event looks for files that match:

- ▶ cluster_event_name.pre_post_event in /var/adm/hacws/events
- ▶ cluster_event_name.post_event in usr/sbin/hacws/events
- ▶ cluster_event_name.post_post_event in /var/adm/hacws/events

The current version of ssp.hacws fileset installs the following scripts in the /usr/sbin/hacws/events directory:

acquire_service_addr.pre_event	This script checks if the node is the backup control workstation, and if it is, calls /etc/rc.backup_cw_alias -delete.
acquire_service_addr.post_event	This script checks if the node is the backup control workstation, and if it is, calls /etc/rc.backup_cw_alias -add.
network_down.post_event	This is run after a network_down event. It checks that it is running on the active control workstation and that the backup control workstation exists. If these conditions are met, it causes a failover to the other control workstation.
node_up_complete.post_event	This script checks if the local node that just came up is an inactive control workstation, and if it is, it runs /usr/sbin/hacws/spcw_apps -u, which starts the backup control workstation applications (see below).
release_service_addr.pre_event	This script checks if the node is the backup control workstation, and if it is, runs /etc/rc.backup_cw_alias -delete.
release_service_addr.post_event	This script checks if the node is the backup control workstation, and if it is, runs /etc/rc.backup_cw_alias -add.

The spcw_apps script is called by node_up_complete on the inactive workstation, and it does the following:

- ▶ Checks that the SDR is available; if it is not, exit.

- ▶ Runs `/etc/rc.sp`.
Checks if file collections has been specified. If it is not, then remove any configuration entries. If it is, then the inactive control workstation will be configured as a file collections client. File collections are run and the entry `/usr/sbin/hacws/spcw_filec_update` is put in the crontab to run every 15 minutes.
- ▶ Checks if the inactive control workstation needs to run as an NTP client.
- ▶ Stops and restarts `sysctld`.

3.4.2 Planning the HACMP configuration

When planning your HACWS installation, it is important to keep in mind that it is just a simple HACMP implementation with the following resources defined in HACMP:

- ▶ One rotating resource group label, which must be called `hacws_group1`
- ▶ One service address, which is the IP address that the SP knows as the CWS
- ▶ One application server, which will control the SP daemons, SDR, hardware monitoring, and so on

There are also some planning decisions that can be made at the initial install of the SP complex that will facilitate the integration of the backup control workstation. These include:

- ▶ Planning networks
- ▶ The `/spdata` file system

Planning networks

Address ranges need to be assigned, as required, for boot adapters and standby adapters. If a private network is being used for the HMC communications, it needs to be cabled separately and the address ranges assigned.

The `/spdata` file system

It is best if the `/spdata` file system is initially installed on external storage devices, but it is possible to move `/spdata`, if it already exists on internal storage devices, during the HACWS configuration steps.

3.4.3 Requirements imposed on the logical definitions

We have already looked at the hardware options to remove single points of failure. We need to recognize that there are some requirements imposed on the logical configuration by availability considerations and the design of HACWS.

These areas are:

- ▶ Network design
- ▶ HACMP startup design

Network design

There are a number of features of the control workstation and PSSP that require that the primary workstation can always communicate with the IP label corresponding to the host name of the backup workstation, even while it is the active control workstation.

They are:

Kerberos

The primary control workstation is the primary Kerberos Version 4 authentication server. The recommended configuration has the backup control workstation as the secondary Kerberos Version 4 authentication server. If the primary control workstation is running after the PSSP resources have failed over to the backup control workstation, the primary still uses the backup control workstation's IP label.

File Collections/supper

While the primary control workstation is the inactive control workstation, it collects supper updates from the active backup control workstation every 15 minutes, if file collections is configured.

For this reason, the IP label configuration in Figure 3-2 on page 167 is recommended. The boot IP label on the backup control workstation is not known to Kerberos or the PSSP services, as it will be replaced by the service IP label when the HACWS resource group is taken over by the backup control workstation. The boot IP label on the primary control workstation has to be added to the Kerberos database as the backup control workstation needs to communicate with the running primary control workstation.

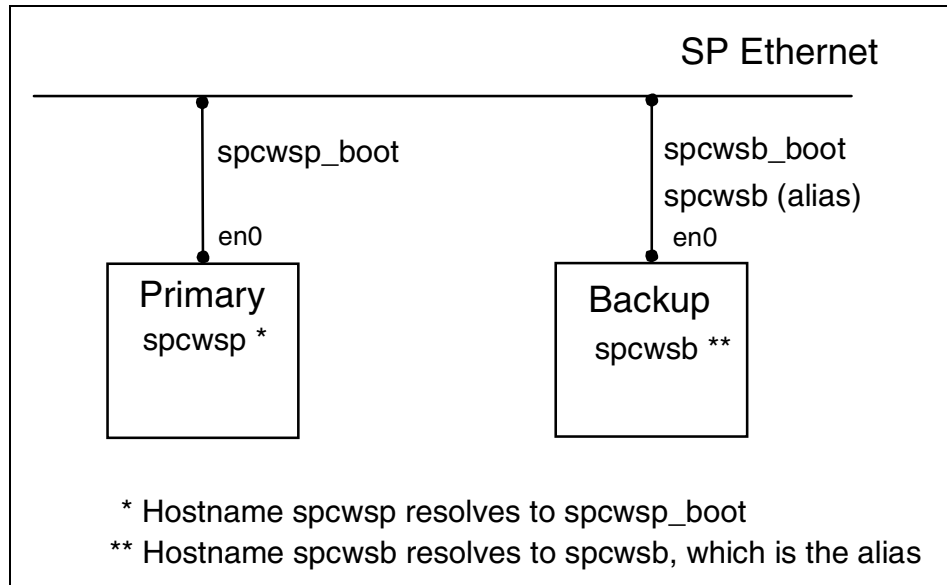


Figure 3-2 IP Label configuration

There are some situations where the control workstation can be configured with more than one network label on the SP administration Ethernet. Some possible reasons for this are:

- ▶ A large number of SP nodes with multiple adapters on the control workstation.
- ▶ A separate network is desired for each partition.

If you have enough adapters in the control workstations, this will not be a problem; however, if you are using aliasing to put multiple addresses on the adapter on the primary control workstation, we recommend the following configuration, which takes advantage of IPAT through IP aliasing:

- ▶ Each control workstation has a boot label on the boot adapter.
- ▶ Multiple service alias addresses are defined and added to the HACWS resource group.
- ▶ A persistent alias is added to the primary control workstation so that it can still communicate with the service alias IP labels once the HACWS resource group fails over to the backup control workstation. This is because the IP alias labels must be on a different subnet to the boot IP labels, and the primary control workstation will not have a label on the service alias subnet after failover.

HACMP startup design

It is recommended that HACMP not be automatically started on system boot. This is because the rotating resource group will start on the first control workstation that starts the cluster manager, and when there are two control workstations booting simultaneously, there is a potential race condition. Without the cluster manager starting automatically, the control workstations can be booted in any order, and the cluster manager is then started on the workstation that is intended to become the active workstation.

It is possible to configure HACMP to start automatically on the primary control workstation which means that it will only start the application if the backup control workstation is not active. This is discussed in detail Section 3.6.13, “Starting of cluster services on the primary workstation” on page 217.

3.5 Configurations used in this document

This chapter has two objectives:

- ▶ Testing HACWS with PSSP 3.4 and HACMP 4.5 and describing the configuration steps in detail
- ▶ Testing HACWS with a p690 integrated in a Cluster 1600

In planning and testing HACWS with HACMP, we choose a configuration that will take advantage of one of the new features of HACMP 4.5: the persistent alias feature. We will now show you how to configure your HACWS cluster using this configuration; however, if you wish to use the traditional network configuration, we also describe what extra steps should be followed.

We considered that the following configurations would also allow us to test both the operation of HACWS with PSSP 3.4 and HACMP 4.5, both on a standard SP configuration as well as in an Cluster 1600 with p690 servers:

- ▶ SP frame only, with no standby adapters on the primary CWS
- ▶ SP frame only, with standby adapters on the primary CWS
- ▶ SP frame and p690, with both the CWS and HMC on a single administrative network
- ▶ SP frame and p690, with both the CWS and HMC on a private network other than SPLAN
- ▶ SP frame with p690, with both the CWS and HMC on a private network with standby adapters on the primary CWS

Note: We realized that the same HACWS configuration will work for all the above scenarios with a few additional definitions. Here we present a single configuration and give options on the way to cover all scenarios.

3.5.1 SP frame only with no standby adapters on the CWS

The scenario of a single SP frame with two control workstations is shown in Figure 3-1 on page 160. This is often called the classic configuration.

There is one single point of failure in each control workstation: the single network card. HACWS overcomes this by forcing a failover to the backup control workstation if the adapter fails. This is a common setup despite of this restriction, as there are often not enough free adapter slots for the extra network card.

3.5.2 SP frame only with standby adapters on the primary CWS

This configuration removes one more single point of failure and, in the case of a network adapter failure, the application does not have to fail over to the backup control workstation; the service IP label is moved to the standby adapter. If adapter slots are available, the standby adapter configuration is recommended.

Figure 3-3 on page 170 shows this configuration.

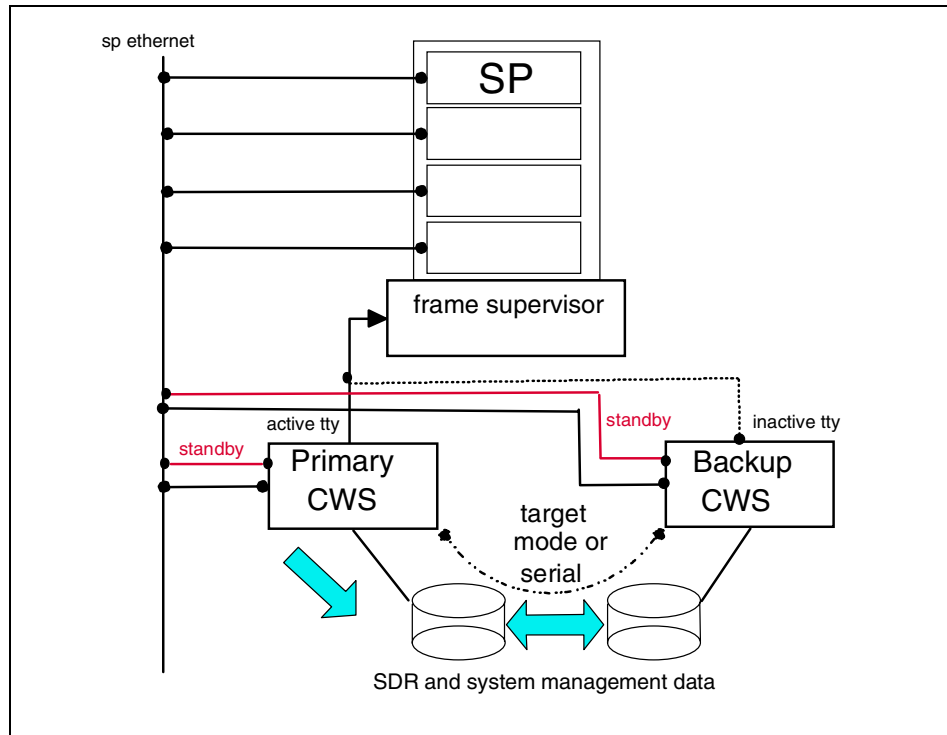


Figure 3-3 One frame with standby adapters on the control workstations

3.5.3 SP Frame and p690 with the CWS and HMC on same SPLAN

The configuration in Figure 3-4 on page 171 shows the layout for the integration of the p690 into the SP complex. The Redpaper *Configuring p690 in an @server Cluster 1600*, REDP0187 describes, in detail, the steps needed to add the p690 to the SP complex.

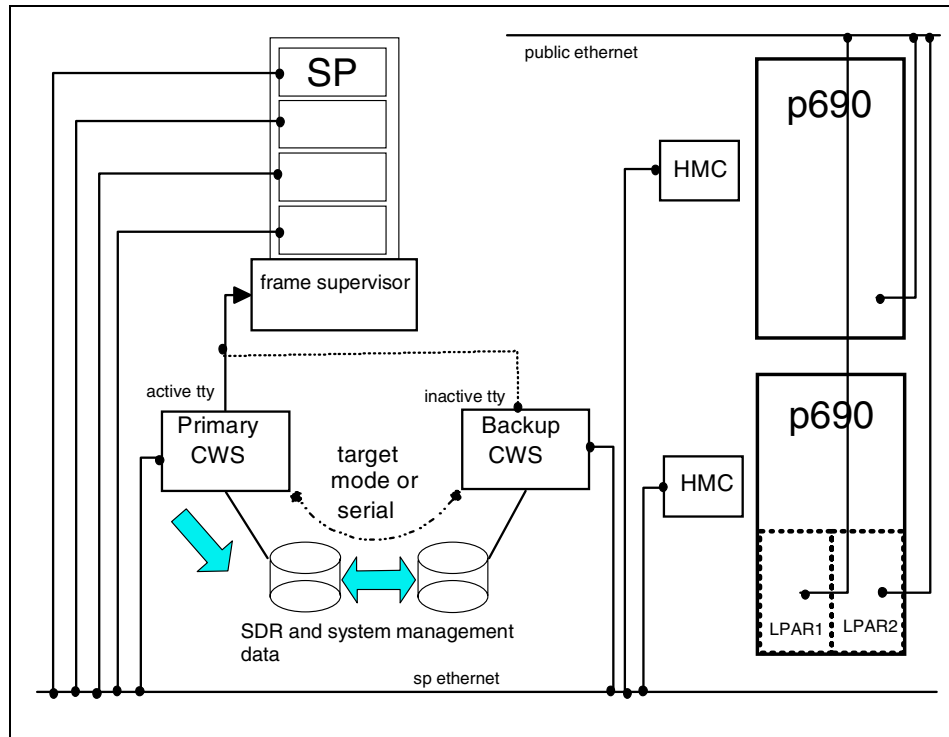


Figure 3-4 p690s on an SP administration network

No changes are required in the HACWS configuration, as the HMC has not introducing any further components for HACMP to control. In particular, the SPLAN is already under the control of HACWS.

3.5.4 CWS and HMC on a private network other than SPLAN

This configuration consists two control workstations and the SP frame on one network. Each control workstation has a second adapter on a private network with the HMC.

Figure 3-5 on page 172 shows this configuration.

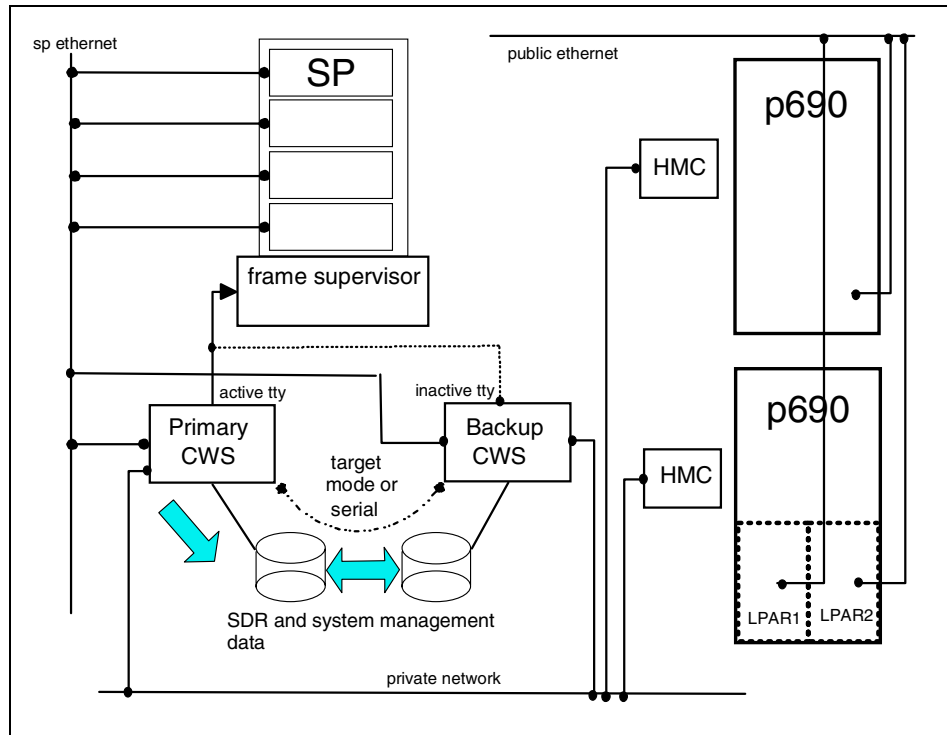


Figure 3-5 p690s on a private network other than SPLAN

The additional configuration required is to add a private network to the HACMP configuration, that is, defining boot and service adapter configurations for a private network to HACWS.

3.5.5 CWS and HMC on a private network with standby adapters

This configuration consists of two control workstations and the SP frame on one network with standby adapters. Each control workstation has a second pair of adapters on a private network with the HMC.

Figure 3-6 on page 173 shows this configuration. It is the same as the previous scenario, with additional standby adapters for each network.

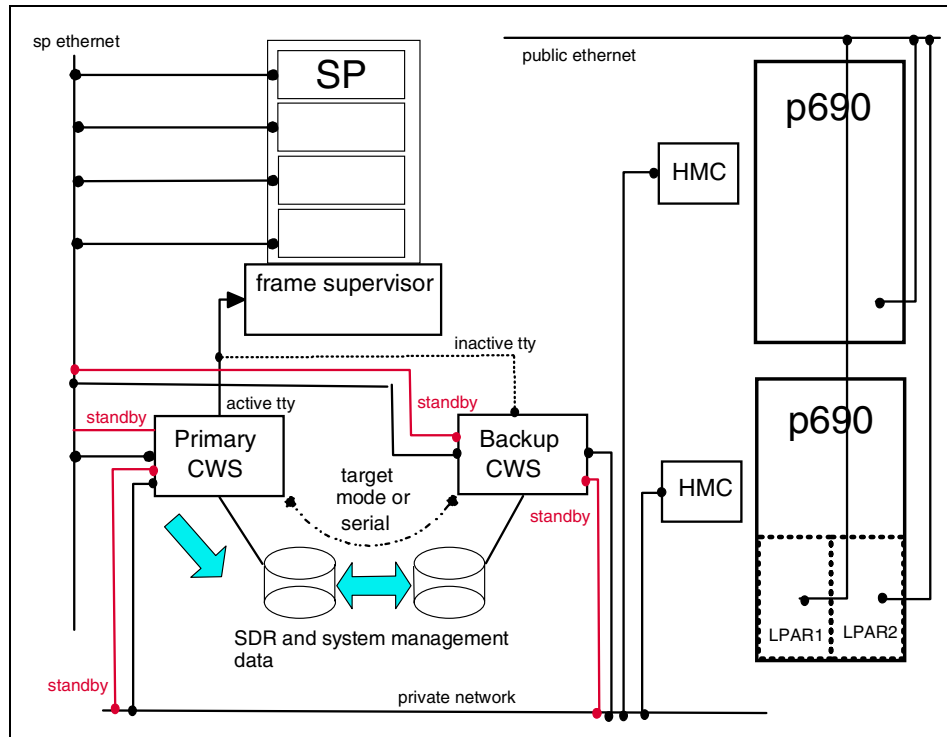


Figure 3-6 p690s on private network with optional standby adapters

Note: As it is preferred that the HACWS resources run on the primary control workstation, it is not necessary to have standby adapters in the backup control workstation.

3.5.6 IP labels and networks

We used the configuration as shown in Table 3-1 for IP labels and networks.

Table 3-1 IP labels

Function	IP label	Address	Netmask
Primary boot	spcwsp_boot	192.168.6.50	255.255.255.224
Backup boot	spcwsb_boot	192.168.6.51	255.255.255.224
Persistent alias	spcwsb	192.168.6.34	255.255.255.224
Service address	spcwsp	192.168.6.33	255.255.255.224

Function	IP label	Address	Netmask
Primary standby	spcwsp-stby	192.168.6.200	255.255.255.224
Backup standby	spcwsp-stby	192.168.6.201	255.255.255.224
HMC on SP Admin	spcwsp-hmc	192.168.6.45	255.255.255.224
HMC on private	spcwsp-priv-hmc	192.168.6.230	255.255.255.224
Primary private boot	spcwsp-priv_boot	192.168.6.228	255.255.255.224
Backup private boot	spcwsp-priv_boot	192.168.6.229	255.255.255.224
Primary private standby	spcwsp-priv_stby	192.168.6.170	255.255.255.224
Backup private standby	spcwsp-priv_stby	192.168.6.171	255.255.255.224
Service private	spcwsp-priv	192.168.6.231	255.255.255.224

3.6 Installing and configuring HACWS

This section gives a step-by-step procedure for installing HACWS, including options covering all the configurations above.

3.6.1 Preparation

Preparation for HACWS installation consists of the following steps:

1. Perform complete SP System installation with primary CWS
2. Install AIX on the backup control workstation
3. Configure shared storage
4. Configure RS-232 control lines
5. Configure non-IP network

Install the SP system

Use the method outlined in Chapter 2, “Installing and configuring a new RS/6000 SP system” in the *PSSP for AIX: Installation and Migration Guide*, GA22-7347.

For attaching p690 servers, follow these steps:

1. Connect the HMC to either the SP Administration network or a private network with just the control workstations and the HMC.
2. Add the p690 definition into the SDR. This procedure is covered in the Redpaper *Configuring p690 in an @server Cluster 1600*, REDP0187.

Install AIX on the backup control workstation

We installed AIX 5L Version 5.1 on the backup control workstation. We then made a mksysb backup image of both the primary and backup control workstations and a backup of /spdata, in case we needed to recover.

Configure shared storage

As we always intended to use the control workstation as part of an HACWS cluster, we configured the system initially with /spdata on external mirrored disks. If /spdata is already configured on internal storage, then external devices, which can be access non-concurrently by both control workstations, need to be added at this point.

Configure RS-232 control lines

Each SP frame requires a serial port on the control workstation. You need an Y cable, FC 1245 to connect the CWS with the SP frame. The primary control workstation is ready be connected. The backup control workstation must be connected to each frame using the same tty number as on the primary. If this is not done correctly, the hardware monitoring will not function.

Configure non-IP network

HACMP requires a non-IP connection between the nodes in the cluster. This can either be RS-232, Target mode SCSI, or Target mode SSA. We chose Target mode SSA for our testing; see Chapter 11 of the *HACMP for AIX 4.5 Enhanced Scalability Installation and Administration Guide*, SC23-4306.

3.6.2 Configuration of the backup control workstation

To set up the backup control workstation:

- ▶ Install PSSP on the backup control workstation.
- ▶ Tune parameters on the backup control workstation.
- ▶ Configure authentication on the backup control workstation.

Install PSSP on the backup control workstation

The same version and PTF level of PSSP must now be installed on the backup control workstation. The same RSCT code also needs to be installed. Do not install the ssp.hacws files at this time.

Important: PSSP should only be installed at this time and NOT configured; that is, do not run the `install_cw` command.

Tune parameters on the backup control workstation

The *PSSP for AIX: Installation and Migration Guide*, GA22-7347 talks about configuring a number of parameters on the control workstation. The same changes must now be made to the backup control workstation. These are as follows:

Network tuning	Configure the transmit queue size (using chdev to modify xmt_queue_size).
Max default processes	Set the maximum number of default processes allowed per user.
no tunable values	Network options (no) parameters need to be added in rc.net . Remember that thewall is not tunable in AIX 5L 5.1 and is for information only.

Configure authentication on the backup control workstation

The same authentication methods should be configured on the backup CWS. The **lsauthent** command will show the primary control workstation setting and **chauthent** should be used to configure the backup control workstation (see Example 3-1).

Example 3-1 Setting the authentication method on the backup control workstation

```
On primary
spcws:/# /usr/bin/lsauthent <Enter>
Kerberos 4
Standard Aix
...
On backup
spcwsb:/# /usr/bin/chauthent -k4 -std <Enter>
spcwsb:/#
```

3.6.3 Kerberos configuration on the backup control workstation

For the purpose of this document, the primary control workstation is the primary Kerberos Version 4 authentication server, and the backup control workstation is the secondary Kerberos Version 4 authentication server.

If the primary control workstation is a Kerberos Version 4 authentication client, then the backup control workstation also needs to be configured as a client.

Note: In this example, we only had one boot IP Label for each control workstation. If you have multiple labels, then each must be added to the Kerberos database.

A number of changes must be made to update Kerberos so that the primary workstation will continue to act as the Kerberos Version 4 primary authentication server, and that the backup will be the backup authentication server. These are:

- ▶ Add the boot address(es) of the primary control workstation.
- ▶ Add the Kerberos Version 4 rcmd service(s) key.
- ▶ Configure the secondary authentication server.
- ▶ Copy the Kerberos Version 4 keys to the backup control workstation.
- ▶ Verify the Kerberos Version 4 database.

Add the boot address(es) of the primary control workstation

When the backup control workstation is the active control workstation, it still needs to communicate with the primary authentication server. The primary control workstation will be on its boot IP Label on each adapter (spcwsp_boot in our example), so an instance for each of these boot IP labels needs to be added to the Kerberos database.

First, run `/usr/kerberos/etc/kdb_edit` to create a new rcmd principal for each boot IP label. You will be asked for the Kerberos master key; then, enter rcmd for the principle name, and the boot label for the instance. This is shown in Example 3-2.

Example 3-2 Adding principal for boot label using kdb_edit

```
spcwsp:/# /usr/kerberos/etc/kdb_edit <Enter>
Opening database...
```

```
Enter Kerberos V4 master key: ***** <Enter>
```

```
Previous or default values are in [brackets];
press <enter> to leave the same, or new value.
```

```
Principal name: rcmd <Enter>
Instance: spcwsp_boot <Enter>
```

```
<not found>, Create [yes] ? yes <Enter>
```

```
Principal: rcmd, Instance: spcwsp_boot, kdc_key_ver: 1
New Password: ***** <Enter>
Verifying, please re-enter
New Password: ***** <Enter>
```

```
Principal's new key version = 1
Expiration date (enter yyyy-mm-dd) [ 2037-12-31 ] ? <Enter>
Max ticket lifetime [ 255 ] ? <Enter>
Attributes [ 0 ] ? <Enter>
Edit O.K.
Principal name: <Enter>
```

```
spcwsp:/#
```

Add the Kerberos Version 4 rcmd service(s) key

The rcmd service key must be added for each boot IP label on the primary control workstation. Use `/usr/lpp/ssp/kerberos/bin/ksrvutil add` to add the rcmd service key for each boot IP label instance in your realm. This is shown in Example 3-3.

Example 3-3 Adding service key for boot label using ksrvutil add

```
spcwsp:/# /usr/lpp/ssp/kerberos/bin/ksrvutil add
Name: rcmd <Enter>
Instance: spcwsp_boot <Enter>
Realm: SPCWSP <Enter>
Version number: 1
New principal: rcmd.spcwsp_boot@SPCWSP; version 1
Is this correct? (yes,no) [yes] yes <Enter>
Password: ***** <Enter>
Verifying, please re-enter Password: ***** <Enter>
Key successfully added.
Would you like to add another key? (yes,no) [yes] no <Enter>
Old key file in /etc/krb-srvtab.old.
spcwsp:/#
```

Configure the secondary authentication server

The following steps are used to configure the backup control workstation as the secondary Kerberos Version 4 authentication server:

1. Add a line to `/etc/krb.conf` on the primary control workstation, listing the backup control workstation as a secondary server for the realm (see Example 3-4).

Example 3-4 /etc/krb.conf

```
spcwsp:/# cat /etc/krb.conf <Enter>
SPCWSP
SPCWSP spcwsp admin server

... Add new line:

spcwsp:/# cat /etc/krb.conf <Enter>
SPCWSP
SPCWSP spcwsp admin server
SPCWSP spcwsp
```

2. Copy `/etc/krb.conf` from the primary control workstation to the backup control workstation.

3. Copy `/etc/krb.realms` from the primary control workstation to the backup control workstation.
4. On the backup control workstation, run **setup_authent**.
 The **setup_authent** command will require you to log in as the admin user (root) and use the admin password used on the primary authentication server.
 The output from **setup_authent** will look like Example 3-5. The important message reporting on the success of the operation may be hidden in the middle of other messages concerning daemons being created.

Example 3-5 Running setup_authent on secondary authentication server

```
spcwsb:/# setup_authent <Enter>
*****
                Logging into Kerberos V4 as an admin user

You must assume the role of a Kerberos V4 administrator <user>.admin to
complete the initialization of Kerberos V4 on the local system. The k4init
command is invoked and will prompt you for the password. If you are
setting up your primary server here, you have just defined it. If you
have defined multiple administrative principals, or if your primary
authentication server is on another system, you must first enter the
name of an administrative principal who has root privilege (UID 0). You
need to be authenticated as this administrator so that this program
can create the principals and service key files for the authenticated
services that run on the SP system.

For more information, see the k4init man page.
*****
setup_authent: Enter name of admin user: root <Enter>
Kerberos V4 Initialization for "root.admin"
Password: **** <Enter>
add_principal: 2502-037 root.SPbgAdm already exists in database.
0513-004 The Subsystem or Group, kpropd, is currently inoperative.
0513-083 Subsystem has been Deleted.
sp2cws: sp3cws: success.
sp2cws: sp3cws: Succeeded
0513-071 The kpropd Subsystem has been added.
0513-059 The kpropd Subsystem has been started. Subsystem PID is 7426.
0513-004 The Subsystem or Group, Kerberos, is currently inoperative.
0513-083 Subsystem has been Deleted.
0513-071 The Kerberos Subsystem has been added.
0513-059 The Kerberos Subsystem has been started. Subsystem PID is 20474.
/usr/lpp/ssp/bin/SDRChangeAttrValues: 0025-001 A read-only SDR session was
obtained. Operations that create or change data are not allowed.
/usr/lpp/ssp/bin/SDRChangeAttrValues: 0025-001 A read-only SDR session was
obtained. Operations that create or change data are not allowed.
0513-044 The splogd Subsystem was requested to stop.
```

```
0513-059 The hardmon Subsystem has been started. Subsystem PID is 11760.
0026-413 Cannot obtain the authentication methods in use on the local host:
spsec_get_ts_authent() was unsuccessful (-1).
hmcmds: 0026-669 Cannot obtain a service ticket for the Hardware Monitor.
0513-059 The splogd Subsystem has been started. Subsystem PID is 20138.
hmreinit: 0037-262 The logging daemon timed out while attempting to connect to
the hardware.
spcwsb:/#
```

The **k4list** command should now give an output similar to Example 3-6.

Example 3-6 Output of k4list on secondary authentication server

```
spcwsb:/# k4list <Enter>
Ticket file:      /tmp/tkt0
Principal:        root.admin@SPCWSP

    Issued          Expires          Principal
Jun 18 14:26:44   Jul 18 14:26:44   krbtgt.SP2CWS@SPCWSP
Jun 18 14:26:46   Jul 18 14:26:46   rcmd.spcwsp@SPCWSP
Jun 18 15:07:47   Jul 18 15:07:47   rcmd.spcwsb@SPCWSP

spcwsb:/#
```

5. After running **setup_authent**, add an entry for the secondary authentication server to the `/etc/krb.conf` file on each of the nodes. This is the same format as the entry on the primary authentication server, as shown in Example 3-4 on page 178.
6. Add the crontab entry to propagate Kerberos changes. If this is the first secondary authentication server created for the realm, then a script will need to be run on the primary authentication server to periodically propagate the Kerberos database changes from the primary to the secondary. For example, the following entry in root's crontab file will propagate the updates at 01:00 every day:

```
0 1 * * * /usr/kerberos/etc/push-kprop >/tmp/kprop.erc 2>&1
```

The push-kprop script will e-mail root each time it runs.

Copy the Kerberos Version 4 keys to the backup CWS

When the HACWS resource group moves between the two control workstations, it requires the service keys to be the same on both primary and backup workstations. This is done by adding the contents of the `/etc/krb-srvtab` file on the primary to the `/etc/krb-srvtab` file on the backup workstation.

On the backup control workstation:

- Make a copy of krb-srvtab, preserving permissions and modification time:

```
cp -p /etc/krb-srvtab /etc/krb-srvtab.save
```

- Copy the krb-srvtab file from the primary:

```
/usr/lpp/ssp/rcmd/bin/rcp -p spcwsb:/etc/krb-srvtab /etc/krb-srvtab.pm
```

- Add the primary's keys to the backup control workstation's file:

```
cat /etc/krb-srvtab.pm >> /etc/krb-srvtab
```

Important: This update must be repeated whenever the keys on the primary authentication server are changed.

Verify the Kerberos Version 4 database

Now confirm that there is a Kerberos principal and rcmd service key that matches the IP label for the host name of the backup control workstation. This is done by running `/usr/lpp/ssp/kerberos/bin/kadmin` and using the `get_entry` option, as shown in Example 3-7, to confirm each principal.

Example 3-7 kadmin get_entry confirms principal

```
spcwsb:/# /usr/lpp/ssp/kerberos/bin/kadmin <Enter>
Welcome to the Kerberos V4 Administration Program, version 2
Type "help" if you need it.
admin: get_entry rcmd.spcwsb <Enter>
Admin password: ***** <Enter>
Info in Database for rcmd.spcwsb:
    Max Life: 255    Exp Date: Thu Dec 31 23:59:59 2037

    Attribs: 00    key: 0 0.
admin:

admin: quit <Enter>
Cleaning up and exiting.
spcwsb:/#
```

Next, run the `/usr/lpp/ssp/kerberos/bin/ksrvutil list` command on the backup control workstation to verify the key. Example 3-8 shows the output of this command.

Example 3-8 ksrvutil list output

```
spcwsb:/# /usr/lpp/ssp/kerberos/bin/ksrvutil list <Enter>
Version    Principal
  1      hardmon.spcwsb@SPCWSP
  1      rcmd.spcwsb@SPCWSP
```

```
1      hardmon.spcwsp-hmc@SPCWSP
1      rcmd.spcwsp-hmc@SPCWSP
1      root.SPbgAdm@SPCWSP
1      rcmd.spcwsp@SPCWSP
1      hardmon.spcwsp@SPCWSP
1      root.SPbgAdm@SPCWSP
1      rcmd.spcwsp_boot@SPCWSP
spcwsb:/etc#
```

3.6.4 Install HACMP/ES on both control workstations

HACMP/ES is installed by following the steps in Chapter 10 of *HACMP for AIX 4.5 Enhanced Scalability Installation and Administration Guide*, SC23-4306.

The installation should be checked using `/usr/bin/lppchk` (see Example 3-9).

Example 3-9 Verification of HACMP/ES software install with lppchk

```
spcwsb:/# lppchk -v <Enter>
spcwsb:/#
```

3.6.5 Install HACWS

The HACWS image now needs to be installed on both control workstations.

Install the PSSP fileset `ssp.hacws` on each control workstation. This fileset installs the `hacws` scripts in `/usr/sbin/hacws`, as shown in Figure 3-7 on page 183.

```

                                Install Software
Ty+-----+
Pr|                SOFTWARE to install                |
|
| Move cursor to desired item and press F7. Use arrow keys to scroll.
| * ONE OR MORE items can be selected.
| * Press Enter AFTER making all selections.
|
| [MORE...23]
|   @ 3.4.0.0  SP PERL Distribution Package
|   @ 3.4.0.0  SP Supervisor Microcode Package
|   @ 3.4.0.0  SP System Partitioning Aid
|
|   ssp.hacws                                     ALL
|   + 3.4.0.0  SP High Availability Control Workstation
|
|   ssp.resctr                                     ALL
|   [MORE...47]
|
| F1=Help          F2=Refresh          F3=Cancel
F7| F7=Select      F8=Image             F10=Exit
F5| Enter=Do       /=Find               n=Find Next
F9+-----+

```

Figure 3-7 Installing the HACWS fileset

3.6.6 Configure HACWS

The following steps are required to configure HACWS:

- ▶ Stop the SP services on the primary control workstation
- ▶ Configure boot adapters on boot IP labels
- ▶ Configure shared storage
- ▶ Additional administrative steps

Stop the SP services on the primary CWS

Before we can continue with the configuration of the control workstations, the primary control workstation can no longer be running as the primary control workstation, so we need to stop the PSSP services. The `spcw_apps` script, which was installed by the `ssp.hacws` fileset, is used to stop the control workstation services.

The format of the command is:

```
/usr/sbin/hacws/spcw_apps -d
```

Example 3-10 on page 184 shows the output of this command.

Example 3-10 Stopping control workstation services with spcw_apps -d

```
spcwsp:/# /usr/sbin/hacws/spcw_apps -d <Enter>
0513-044 The spmgr Subsystem was requested to stop.
0513-044 The splogd Subsystem was requested to stop.
0513-044 The hardmon Subsystem was requested to stop.
0513-044 The sysctld Subsystem was requested to stop.
0513-044 The supfilesrv Subsystem was requested to stop.
0513-044 The sp_configd Subsystem was requested to stop.
0513-044 The swtadmd Subsystem was requested to stop.
0513-044 The swtlog Subsystem was requested to stop.
0513-044 The pman.sp2cws Subsystem was requested to stop.
0513-044 The pmanrm.sp2cws Subsystem was requested to stop.
0513-044 The hr.sp2cws Subsystem was requested to stop.
0513-044 The haem.sp2cws Subsystem was requested to stop.
0513-044 The haemaixos.sp2cws Subsystem was requested to stop.
0513-044 The hagsglsm.sp2cws Subsystem was requested to stop.
0513-044 The hats.sp2cws Subsystem was requested to stop.
0513-044 The sdr.sp2cws Subsystem was requested to stop.
exportfs: 1831-184 unexported /spdata/sys1/install/aix51c145b/lppsource
exportfs: 1831-184 unexported /spdata/sys1/install/pssplpp
spcwsp:/#
```

The purpose of this script is discussed in Section 3.4.1, “Components of HACWS” on page 162. However, without the -i flag, it does not set this control workstation as the inactive control workstation.

Configure boot adapters on boot IP labels

Each control workstation needs to be configured so that they reboot with the boot IP label on the network adapters. This is done using the SMIT command:

```
smitty chinet
```

Figure 3-8 on page 185 shows how this is done.

Change / Show a Standard Ethernet Interface

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

	[Entry Fields]	
Network Interface Name	en0	
INTERNET ADDRESS (dotted decimal)	[192.168.6.50]	
Network MASK (hexadecimal or dotted decimal)	[255.255.255.224]	
Current STATE	up	+
Use Address Resolution Protocol (ARP)?	yes	+
BROADCAST ADDRESS (dotted decimal)	[]	

Figure 3-8 SMIT chinet to set boot IP label on adapter

After changing the IP label, it is important to refresh the inetd daemon or you will get errors trying to configure the HACMP cluster information, in particular, GODM errors. Use the following command:

```
/usr/bin/refresh -s inetd
```

Important: The control workstations should not be rebooted until HACWS has been fully configured.

Configure shared storage

When we initially configured our primary workstation, /spdata was created on external mirrored disks. If you have not already done this, then you should now follow these steps to copy /spdata to external storage:

1. Note the size of the logical volume that /spdata uses.

Note: If /spdata has not been configured as a separate file system, then you will need to mount the new file system created in Step 6 over /mnt and copy the data there.

2. Make a backup of /spdata.
3. Unmount /spdata.
4. Mount /spdata on a temporary mount point, for example, /mnt.
5. Create a volume group on the external disks, using a major number that is free on both control workstations.
6. Create a logical volume on the new volume group. Preferably, this logical volume should have at least two copies, each on separate disks, accessible by different adapters.

7. Add a file system to the logical volume; /spdata will be the mount point, and automatically mount should be set to false.
8. Change the activate automatically characteristic for the volume group to no.
9. It is recommended that you leave quorum for the volume group on and use the new feature of HACMP 4.5 to fail over the resource group on loss of quorum.
10. Mount /spdata.
11. Copy the files from the old /spdata (now /mnt) to the newly created file system:

```
cd /mnt
find . -print | cpio -pv dum /spdata
```
12. Unmount both /spdata and /mnt.
13. Vary off the new spdata volume group and import the volume group onto the backup control workstation, using the same major number as before.
14. Set the volume group characteristic Active Automatically to no and set quorum as before.
15. Vary off the volume group.
16. Vary on the volume group on the primary control workstation and mount /spdata.

Additional administrative steps

The *HACMP for AIX 4.5 Enhanced Scalability Installation and Administration Guide*, SC23-4306 recommends a number of additional AIX administration tasks to be done on each node in the HACMP cluster. We recommend the following tasks be done for each control workstation:

- ▶ Check that users and password are consistent on both control workstations.
- ▶ The thewall parameter is no longer a concern in AIX 5L Version 5.1.
- ▶ /etc/rc.net needs to be modified to add no -o routerevalidate=1.
- ▶ AIX keeps a cache that could cause delays on failing over Adapter IP Labels.
- ▶ Confirm that /etc/hosts on each machine has entries for all adapter addresses.
- ▶ Ensure that there is a /.rhosts file on each control workstation, with the boot and service addresses for both nodes.

Note: I/O pacing is automatically set by the install_hacws script.

3.6.7 Configure HACMP topology

The following HACMP administration tasks need to be done:

- ▶ Define the cluster ID and name.
- ▶ Define the nodes to HACMP.
- ▶ Define the adapters to HACMP.
- ▶ Check the subnet.
- ▶ Synchronize the cluster Topology.
- ▶ Configure the non-IP network.
- ▶ Add adapters for the serial network.
- ▶ Avoiding false adapter failures.
- ▶ Configure the HACWS application server.
- ▶ Configure the HACWS resource group.
- ▶ Synchronize cluster resources.

These tasks can be performed on either control workstations, as the data will be synchronized between the two of them. Traditionally, these tasks are performed on the primary control workstation.

Define the cluster ID and name

Choose a unique cluster ID and name for your cluster by running **smitty hacmp** and selecting **Cluster Configuration -> Cluster Topology -> Configure Cluster -> Add Cluster Definition** (see Figure 3-9).

The following data should be entered:

Cluster ID	A unique integer for your site between 1 and 99 999.
Cluster Name	A 30 character name, alphanumeric, with dash or underscore, but not a leading number.

Add a Cluster Definition

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

[Entry Fields]

****NOTE: Cluster Manager MUST BE RESTARTED
in order for changes to be acknowledged.****

* Cluster ID [1956] #*
Cluster Name [Mordor]

Figure 3-9 Set cluster ID and name

Define the nodes to HACMP

Add the node definitions by running **smitty hacmp** and selecting **Cluster Configuration -> Cluster Topology -> Configure Nodes -> Add Cluster Nodes**.

Add the node names for the cluster. HACWS configuration requires that they match the host names (see Figure 3-10).

Add Cluster Nodes

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

* Node Names

[Entry Fields]

[spcwsb spcwsb]

Figure 3-10 Configuring cluster nodes

Define the adapters to HACMP

This section discusses the configuration of all the adapters required by the various configurations given at the beginning of this book. Those marked as optional depend on what configuration you are using.

The following adapters need to be configured:

- ▶ Boot IP labels
- ▶ Standby IP labels (optional)
- ▶ Discover network topology
- ▶ Service IP labels
- ▶ Persistent IP label (optional)

Boot IP labels

The following boot IP labels need to be configured

- ▶ Boot IP label for the primary control workstation
- ▶ Boot IP label for the backup control workstation
- ▶ Boot IP label on a private network for the primary control workstation (optional)
- ▶ Boot IP label on a private network for the backup control workstation (optional)

The boot adapters are added by running **smitty hacmp** and selecting **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Select Add an Adapter to a new network is shown in Figure 3-11.

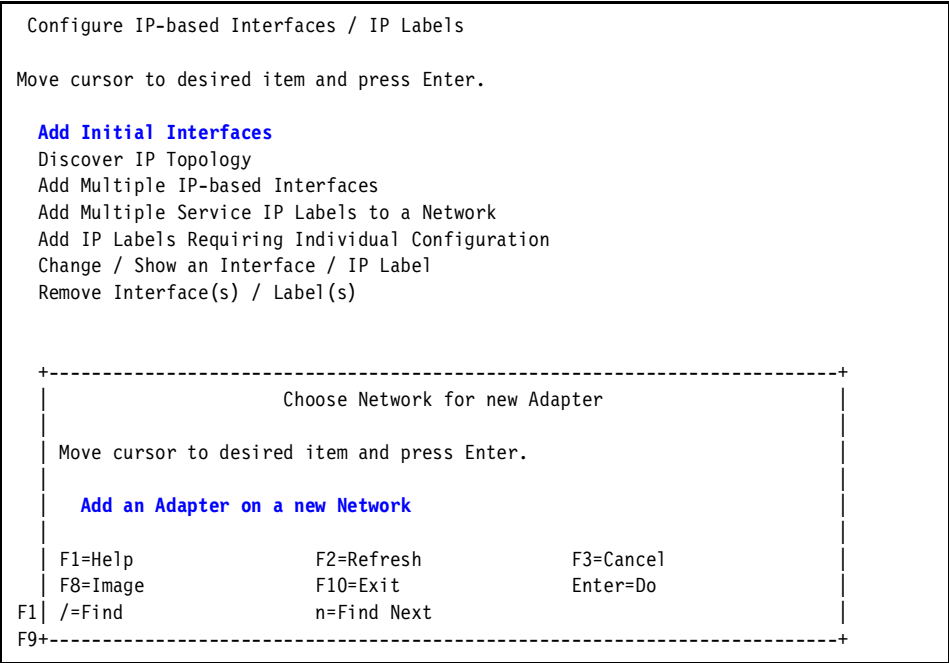


Figure 3-11 Add Initial Interfaces

Boot IP label for the primary control workstation

The following information now needs to be entered to create the boot label for the primary control workstation, as shown in Figure 3-12 on page 190:

IP Label	Primary control workstation boot label
Network type	ether
Network Name	Your name for the ethernet network
Network Attribute	public
Interface Function	boot
Interface IP address	Decimal address for the boot label
Node Name	Name of the primary control workstation
Netmask	Your network netmask

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[spcwsp_boot]	+
* Network Type	[ether]	+
Network Name	[hacwsether]	+
* Network Attribute	[public]	+
* Interface Function	[boot]	+
Interface IP Address	[192.168.6.50]	
* Node Name	[spcwsp]	+
Netmask	[255.255.255.224]	+

Figure 3-12 Add boot IP label for the primary CWS

Fill out the same information for the boot IP label for the backup control workstation, as shown in Figure 3-13 on page 191:

IP Label	Backup control workstation boot label
Network type	ether
Network Name	Your name for the Ethernet network
Network Attribute	public
Interface Function	boot
Interface IP address	Decimal address for the boot label
Node Name	Name of the backup control workstation
Netmask	Your network netmask

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[spcwsb_boot]	+
* Network Type	[ether]	+
Network Name	[hacwsether]	+
* Network Attribute	[public]	+
* Interface Function	[boot]	+
Interface IP Address	[192.168.6.51]	
* Node Name	[spcwsb]	+
Netmask	[255.255.255.224]	+

Figure 3-13 Add boot IP label for the backup CWS

Boot IP label on a private network for the primary control workstation

If you are using a private network for the HMC, you will need to add the boot addresses on a private network, selecting Add an Adapter on a new Network, as shown in Figure 3-11 on page 189.

Enter the following data, as shown in Figure 3-14 on page 192:

IP Label	Primary control workstation boot label on a private network
Network type	ether
Network Name	Your name for the private Ethernet network
Network Attribute	public
Interface Function	boot
Interface IP address	Decimal address for the boot label
Node Name	Name of the primary control workstation
Netmask	Your network netmask

Add an Initial Interface

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

	[Entry Fields]		
* IP Label	[spcwsp-priv_boot]	+	
* Network Type	[ether]	+	
Network Name	[hacwsether-priv]		+
* Network Attribute	[public]	+	
* Interface Function	[boot]	+	
Interface IP Address	[192.168.6.228]		
* Node Name	[spcwsp]	+	
Netmask	[255.255.255.224]	+	

Figure 3-14 Add boot IP label for the primary CWS on a private network

Boot IP label on a private network for the backup control workstation

If you are using a private network for the HMC, add the boot IP label for the backup control workstation on a private network, as shown in Figure 3-15 on page 193:

IP Label	Backup control workstation boot label on the private network
Network type	ether
Network Name	Your name for the private Ethernet network
Network Attribute	public
Interface Function	boot
Interface IP address	Decimal address for the boot label
Node Name	Name of the backup control workstation
Netmask	Your network netmask

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[spcwsb-priv_boot]	+
* Network Type	[ether]	+
Network Name	[hacwsether-priv]	+
* Network Attribute	[public]	+
* Interface Function	[boot]	+
Interface IP Address	[192.168.6.229]	
* Node Name	[spcwsb]	+
Netmask	[255.255.255.224]	+

Figure 3-15 Add boot IP label for the backup CWS on a private network

Standby IP labels (optional)

If standby adapters are being used on either your SP administration network or a private HMC network, they should be configured now.

The following standby adapters can be configured:

- ▶ Standby IP label for the primary control workstation (optional)
- ▶ Standby IP label for the backup control workstation (optional)
- ▶ Standby IP label on a private network for the primary control workstation (optional)
- ▶ Standby IP label on a private network for the backup control workstation (optional)

The standby adapters are added by running **smitty hacmp** and selecting **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Standby IP label for the primary control workstation

If required, add the standby IP label for the primary control workstation, as shown in Figure 3-16 on page 194:

IP Label	Primary control workstation standby label
Network type	ether
Network Name	Your name for the Ethernet network
Network Attribute	public
Interface Function	standby

Interface IP address Decimal address for the standby label
Node Name Name of the primary control workstation
Netmask Your network netmask

Add an Initial Interface

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

[Entry Fields]

* IP Label

[spcwsp_stby]

+

* Network Type

[ether]

+

Network Name

[hacwsether]

+

* Network Attribute

[public]

+

* Interface Function

[standby]

+

Interface IP Address

[192.168.6.200]

* Node Name

[spcwsp]

+

Netmask

[255.255.255.224]

+

Figure 3-16 Add standby IP label for the primary CWS

Standby IP label for the backup control workstation

If required, add the standby IP label for the backup control workstation, as shown in Figure 3-17 on page 195:

IP Label Backup control workstation standby label
Network type ether
Network Name Your name for the Ethernet network
Network Attribute public
Interface Function standby
Interface IP address Decimal address for the standby label
Node Name Name of the backup control workstation
Netmask Your network netmask

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[spcwsb_stby]	+
* Network Type	[ether]	+
Network Name	[hacwsether]	+
* Network Attribute	[public]	+
* Interface Function	[standby]	+
Interface IP Address	[192.168.6.201]	
* Node Name	[spcwsb]	+
Netmask	[255.255.255.224]	+

Figure 3-17 Add standby IP label for the backup CWS

Standby IP label on a private network for the primary control workstation

If required, add the standby IP label for the primary control workstation, as shown in Figure 3-18 on page 196:

IP Label	Primary control workstation standby label on a private network
Network type	ether
Network Name	Your name for the private Ethernet network
Network Attribute	public
Interface Function	standby
Interface IP address	Decimal address for the standby label
Node Name	Name of the primary control workstation
Netmask	Your network netmask

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[spcwsp-priv_stby]	+
* Network Type	[ether]	+
Network Name	[hacwsether-priv]	+
* Network Attribute	[public]	+
* Interface Function	[standby]	+
Interface IP Address	[192.168.6.170]	
* Node Name	[spcwsp]	+
Netmask	[255.255.255.224]	+

Figure 3-18 Add standby IP label for the primary CWS on a private network

Standby IP label on a private network for the backup CWS

If required, add the standby IP label for the backup control workstation on a private network, as shown in Figure 3-19 on page 197:

IP Label	Backup control workstation standby label on private network
Network type	ether
Network Name	Your name for the private Ethernet network
Network Attribute	public
Interface Function	standby
Interface IP address	Decimal address for the standby label
Node Name	Name of the backup control workstation
Netmask	Your network netmask

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[spcwsb-priv_stby]	+
* Network Type	[ether]	+
Network Name	[hacwsether-priv]	+
* Network Attribute	[public]	+
* Interface Function	[standby]	+
Interface IP Address	[192.168.6.171]	
* Node Name	[spcwsb]	+
Netmask	[255.255.255.224]	+

Figure 3-19 Add standby IP label for the backup CWS on a private network

Discover network topology

Running the SMIT option Discover IP Topology will update the HACMP database with the networks we have created, including the subnet information. This is run by running `smitty hacmp` and selecting **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Discover IP Topology**.

Figure 3-20 shows the output of the network discovery option.

COMMAND STATUS

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

Discovering IP Network Connectivity
Discovering IP Network Connectivity
IP Network Discovery completed normally

Figure 3-20 Discover IP Topology

Service IP labels

A service IP label needs to be added for each IP network defined. Run `smitty hacmp` and select **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure IP-based Interfaces / IP Labels -> Add IP Labels Requiring Individual Configuration**.

The following service IP labels need to be defined:

- ▶ Service IP label on the SP administration network
- ▶ Service IP label on a private network (optional)

Service IP label on the SP administration network

This is the IP label that matches the primary control workstations host name and is the IP label that the nodes use to communicate with the control workstation. The following information is entered, as shown in Figure 3-21:

IP Label	Service label.
Network type	ether.
Network Name	Your name for the Ethernet network.
IP Label Function	service.
IP address	Decimal address for the service label.
Hardware Address	Enter the hardware address for hardware address takeover. See the HACMP documentation for details on configuring hardware address takeover.
Node Name	Blank for the rotation resource group.
Netmask	Your network netmask.

Add IP Labels Requiring Individual Configuration

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

	[Entry Fields]	
* IP Label	[spcwsp]	+
Network Type	ether	
Network Name	hacwsether	
* IP Label Function	[service]	+
IP Address	[192.168.6.33]	
Hardware Address	[0004AC4b11b0]	
Node Name	[]	+
Netmask	[255.255.255.224]	+

Figure 3-21 Adding a service IP label

Service IP label on a private network (optional)

If using a private network for the HMC, an IP label with the following details is created (see Figure 3-22 on page 199):

IP Label	Service label
Network type	ether

Add IP Labels Requiring Individual Configuration

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

	[Entry Fields]	
* IP Label	[spcwsb]	+
Network Type	ether	
Network Name	hacwsether	
* IP Label Function	[persistent]	+
IP Address	[192.168.6.34]	
Hardware Address	[]	
Node Name	[spcwsb]	+
Netmask	[255.255.255.224]	+

Figure 3-23 Adding a persistent IP label

Check the subnet

Check the subnets for the defined network by running `smitty hacmp` and selecting **Cluster Configuration -> Cluster Topology -> Configure Networks -> Configure IP Based Networks -> Change a Network**. Confirm that the correct subnets are listed, as shown in Figure 3-24. This will vary, depending on whether you have standby adapters. If you are using a private network for the HMC, check it also.

Change an IP-based Network

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

	[Entry Fields]	
Network Name	hacwsether	
New Network Name	[]	
* Network Attribute	public	+
* Network Type	[ether]	+
Subnet(s)	192.168.6.32/27	
Add Subnets	[]	+
Remove Subnets	[]	+
Use IP aliasing for IPAT	unsupported	+

Figure 3-24 Subnets for hacwsether

Synchronize the cluster topology

Now that we have created the cluster topology definitions on the primary control workstation, these need to be verified and synchronized with the backup control

workstation. Run **smitty hacmp** and select **Cluster Configuration -> Cluster Topology -> Synchronize Cluster Topology** (see Figure 3-25).

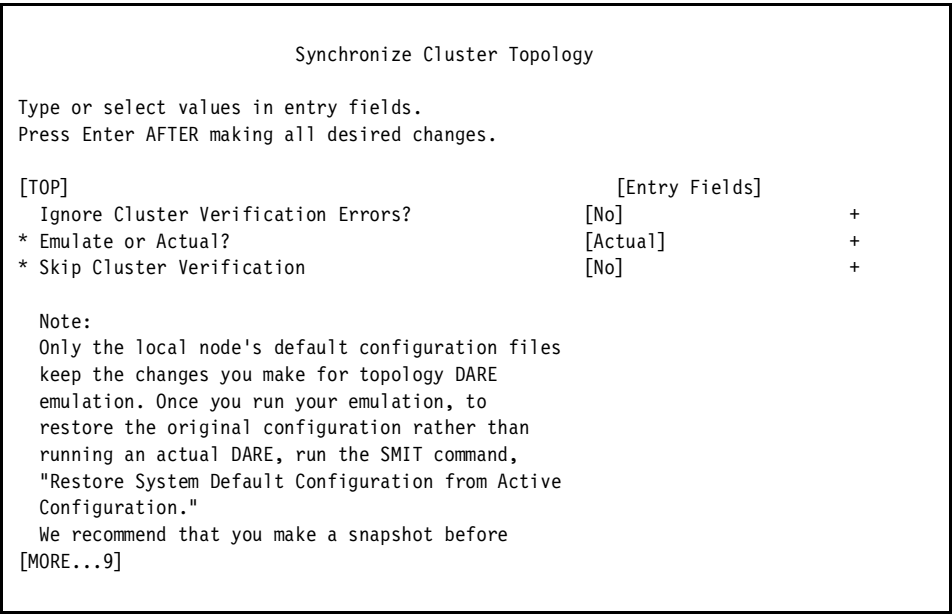


Figure 3-25 Synchronize Cluster Topology

After successfully synchronizing the topology, the persistent alias will be configured on the backup control workstation, as shown by `ifconfig` (see Example 3-11).

Example 3-11 Persistent alias

```
spcwsb:/ ifconfig -a
en0: flags=e080863<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
      inet 192.168.6.51 netmask 0xffffffff broadcast 192.168.6.63
      inet 192.168.6.34 netmask 0xffffffff broadcast 192.168.6.63
en1: flags=4e080863<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,PSEG>
.....
```

Configure the non-IP network

HACMP needs a non-IP network connecting the two control workstations for heartbeat traffic.

In our configuration, we used target mode SSA for the serial network. The *HACMP for AIX 4.5 Enhanced Scalability Installation and Administration Guide*, SC23-4306 recommends that the HACMP node ID be used for the SSA node ID. However, HACMP will not assign this value until the topology has been

synchronized. For this reason, we create the IP topology, synchronize the cluster, create the target mode SSA devices, add their definition into HACMP, then synchronize again. This ordering is not required if you are using a serial or a target mode SCSI network:

The following steps configure and test the target mode SSA devices.

1. Install the SSA target mode support filesets (devices.ssa.tm.rte).
2. Find the HACMP/ES internal number for each node using the command:

```
odmget -q "name = <node_name>" HACMPnode | grep node_id
```

 where `node_name` is the name of each node in HACMP.
 If any other systems are sharing the SSA loop with the control workstations, confirm that they are not using the same node numbers.
3. Change the SSA node number to match the HACMP `node_id`:

```
chdev -l ssar -a node_number=<node_id>
```

 where `node_id` is the HACMP `node_id`.
4. The `lsattr -El ssar` command will confirm that node number has been set (see Example 3-12).

Example 3-12 Command `lsattr` shows node number

```
spcwsb:/# lsattr -El ssar
node_number 4 SSA Network node number True
spcwsb:/#
```

5. Now run `cfgmgr` to create the target and initiator devices. If successful, the output from `lsdev -C | grep tmssa` should look like Example 3-13.

Example 3-13 Target mode SSA devices created

```
spcwsb:/# lsdev -C | grep tmssa
tmssar      Available      Target Mode SSA Router
tmssa4     Available      Target Mode SSA Device
spcwsb:/#
```

6. The devices should now be tested; on one node, direct the output from the target mode device to the screen, and on the other node cat a file to the initiator target mode device, as shown in Example 3-14.

Example 3-14 Testing target mode SSA device

On primary control workstation:

```
spcwsp:/etc/objrepos# cat < /dev/tmssa3.tm <Enter>
(Command will hang)
```



```
Now on backup control workstation:
spcwsb:/ cat /etc/motd > /dev/tmssa4.im <Enter>
spcwsb:/
```

And back on primary control workstation:

```
*****
*                                                                 *
*                                                                 *
* Welcome to AIX Version 5.1!                                     *
*                                                                 *
*                                                                 *
* Please see the README file in /usr/lpp/bos for information pertinent to *
* this release of the AIX Operating System.                       *
*                                                                 *
*                                                                 *
*****
<Ctrl-C> (to exit)
```

Add adapters for the serial network

The serial network devices are added by running **smitty hacmp** and selecting **Cluster Configuration -> Cluster Topology -> Configure Adapters -> Configure Adapters on non IP-based networks -> Add an Adapter**.

Add an adapter for each control workstation on the non-IP based network, as shown in Figure 3-26 on page 204:

Adapter Label	Label for an adapter on the serial network
Network Type	tmssa
Network Name	Cluster name for the network
Device Name	Name of the target mode device
Node Name	Control workstation node name

Add a Non IP-based Adapter

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

	[Entry Fields]	
* Adapter Label	[tmssa1]	
Network Type	[tmssa]	+
* Network Name	[hacwssa]	+
* Device Name	[/dev/tmssa3]	
* Node Name	[spcwsp]	+

Figure 3-26 Add a Non IP-base Adapter

After adding the non-IP based adapters, the cluster topology will need to be synchronized again, as in Figure 3-25 on page 201

Avoiding false adapter failures

In cases of clusters with single adapters, HACMP recommends that a list of devices that will respond to ICMP ECHO requests be constructed. The cluster manager will then attempt to ping these devices to confirm an adapter failure. This feature is configured by creating /usr/sbin/cluster/netmon.cf, which contains a list of IP labels or addresses, one per line, as shown in Example 3-15. For HACMP/ES, a symbolic link can be created to /usr/es/sbin/cluster.

Example 3-15 netmon.cf

```
spcwsp:/usr/sbin/cluster# cat netmon.cf <Enter>
sp2-n3
sp2-n4
lpar1
lpar2
spcwsp:/usr/sbin/cluster#
```

Configure HACWS application server

We now create the application server that will control the resources to be brought up on the active control workstation.

Run `smitty hacmp` and select **Cluster Configuration -> Cluster Resource -> Define Application Servers -> Add an Application Server.**

The following information is to be entered, as shown in Figure 3-27:

Server Name	Name for the application server
Start Script	/usr/sbin/hacws/spcw_apps -ua
Stop Script	/usr/sbin/hacws/spcw_apps -di

Add an Application Server

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

* Server Name

* Start Script

* Stop Script

[Entry Fields]

[**hacws_app**]

[**/usr/sbin/hacws/spcw_a**>

<**in/hacws/spcw_apps -di**]

Figure 3-27 Add an Application Server

Configure HACWS resource group

As discussed during the planning stage, we need to create a rotating resource group called hacws_group1. To do this we:

- ▶ Add the resource group.
- ▶ Add resources to the resource group.

Add the resource group

The resource group is added by running **smitty hacmp** and selecting **Cluster Configuration -> Cluster Resource -> Define Resource Groups -> Add a Resource Group**.

The following details need to be entered, as shown in Figure 3-28 on page 206:

Resource group name	hacws_group1
Node Relationship	rotating
Participating Node Names / Default Node Priority	spcwsp spcwsb

Chapter 3. HACWS: An HACMP Application for Cluster 1600 205

Add a Resource Group

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

	[Entry Fields]	
* Resource Group Name	[hacws_group1]	
* Node Relationship	rotating	+
* Site Relationship	ignore	+
* Participating Node Names / Default Node Priority	[spcwsp spcwsb]	+

Figure 3-28 Add a Resource Group

Add resources to the resource group

To add resource to the resource group, run **smitty hacmp** and select **Cluster Configuration -> Cluster Resources -> Change/Show Resources Attributes for a Resource Group**. Select **hacws_group1**; the following items need to be configured:

- Service IP label** spcwsp (and the service label for HMC private network, if it is being used)
- Filesystems** /spdata
- Volume Groups** datavg
- Application Servers** hacws_app

This is shown in Figure 3-29 on page 207.

Change/Show Resources/Attributes for a Resource Group

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

[TOP]	[Entry Fields]	
Resource Group Name	hacws_group1	
Node Relationship	rotating	
Site Relationship	ignore	
Participating Node Names / Default Node Priority	spcws spcwsb	
Dynamic Node Priority	<input type="checkbox"/>	+
Service IP label	[spcws]	+
Filesystems (default is All)	[/spdata]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems/Directories to Export	<input type="checkbox"/>	+
Filesystems/Directories to NFS mount	<input type="checkbox"/>	+
Network For NFS Mount	<input type="checkbox"/>	+
Volume Groups	[datavg]	+
Concurrent Volume groups	<input type="checkbox"/>	+
Raw Disk PVIDs	<input type="checkbox"/>	+
Connections Services	<input type="checkbox"/>	+
Fast Connect Services	<input type="checkbox"/>	+
Tape Resources	<input type="checkbox"/>	+
Application Servers	[hacws_app]	+
Communication Links	<input type="checkbox"/>	+
Primary Workload Manager Class	<input type="checkbox"/>	+
Secondary Workload Manager Class	<input type="checkbox"/>	+
Miscellaneous Data	<input type="checkbox"/>	
Automatically Import Volume Groups	false	+
Inactive Takeover Activated	false	+
Cascading Without Fallback Enabled	false	+
Disk Fencing Activated	false	+
Filesystems mounted before IP configured	false	+
[BOTTOM]		
F1=Help	F2=Refresh	F3=Cancel
F5=Reset	F6=Command	F7=Edit
F9=Shell	F10=Exit	Enter=Do
		F4=List
		F8=Image

Figure 3-29 Adding a resource to hacws_group1

Synchronize cluster resources

Now that we have created the cluster resource definitions on the primary control workstation, these need to be verified and synchronized with the backup control workstation. Run **smitty hacmp** and select **Cluster Configuration ->**

Cluster Resources -> Synchronize Cluster Resources. This is shown in Figure 3-30 on page 208.

```
Synchronize Cluster Resources

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

[TOP]                                     [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."
[MORE...3]
```

Figure 3-30 Synchronize and verify the cluster resources

3.6.8 Set up the HACWS configuration

The following steps will set up the HACWS configuration

- Make control workstations addressable by their host name.
- Install and configure HACWS.
- Customize cluster event processing.
- Add IP Alias.

Make control workstations addressable by their host name

In an earlier step, we had configured each machine on its boot IP label; for the next few steps, we need to be able to address each control workstation by its host name.

First, we need to put the primary control workstation onto the service address.

Note: Do not change the IP address using SMIT, as we do not want this change to be permanent.

On the primary control workstation, use the **ifconfig** command to change the IP address, as shown in this example:

```
ifconfig <adapter> <host_name> netmask <netmask> up
```

For example:

```
ifconfig en0 spcwsp netmask 255.255.255.224 up
```

If you are using the persistent alias feature of HACMP 4.5 for the backup control workstation, the backup control workstation will already have the alias that corresponds to its host name on the boot adapter. This would have occurred when the persistent alias was created and the cluster topology synchronized.

If, however, you have chosen to use the classic design, you will need to manually add that alias now, as follows:

```
ifconfig <adapter> alias <host_name> netmask <netmask> up
```

For example:

```
ifconfig en0 alias spcwsb netmask 255.255.255.224 up
```

Install and configure HACWS

We can now configure both the primary and backup control workstations as an HACWS configuration. This is done by running **smitty hacws** and selecting **Install and Configure HACWS**.

The following data is entered, as shown in Figure 3-31:

HOSTNAME of primary control workstation	Primary host name
HOSTNAME of backup control workstation	Backup host name
Execute on both primary and backup?	yes

Install and Configure HACWS

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

* HOSTNAME of primary control workstation

* HOSTNAME of backup control workstation

Execute on both primary and backup?

[Entry Fields]

[spcwsp]

[spcwsb]

yes

+

Figure 3-31 Install and Configure HACWS

The following script is executed:

```
/usr/sbin/hacws/install_hacws -p spcws -b spcwsb -s
```

This script does some configuration sanity checking: host names, file systems, and so on, and the following items:

- ▶ Configures HACWS environmental variables.
- ▶ Tidies up SMIT stanzas.
- ▶ Updates hardmon, sdr and sdrprot.
- ▶ Modifies /etc/rc.sp and creates /etc/rc.sp2.
- ▶ Checks the sdr for the default partition.
- ▶ Checks the hardmon configuration.
- ▶ Create the destination information file and copies it to the backup control workstation.
- ▶ Runs the SDR sanity check (init_ssp_envs).
- ▶ Copy the switch files into /etc/SP.
- ▶ Configures the spmon logging daemon and stops splogd on backup control workstation.
- ▶ Runs install_spmgr, if it exists.
- ▶ Runs install_swt, if it exists.
- ▶ Confirms that /tftpboot/tuning.cust exists on both control workstations and is the same.
- ▶ If the commands **t~~s~~m** and **s~~u~~** have been replaced by a link to SP commands, then the link is removed, and they are put back to the original AIX commands.
- ▶ Updates sysctld.conf.
- ▶ Backs up /etc/inittab and then remove hd, sp, sdrd, hr, splogd, hardmon, hmon, hats, hags, haem, pman, sp_configd, spmgr, swtlog, swtadmd and swt.
- ▶ Removes heartbeat services if the control workstation was updated from a previous version of PSSP.
- ▶ Creates /etc/rc.hacws and add it to /etc/inittab.
- ▶ Creates /etc/rc.backup_cw_alias.
- ▶ Adds HACWS to SDR.
- ▶ Remove files in /spdata/sys1 on the backup control workstation.
- ▶ Configures I/O pacing.

Figure 3-32 on page 211 shows the output of this script.


```
COMMAND STATUS

Command: running      stdout: yes      stderr: no

Before command completion, additional instructions may appear below.

spcwsp: 0518-307 odmdelete: 0 objects deleted.
spcwsp: 0513-059 The sdr.spcwsp Subsystem has been started. Subsystem PID is 6734.
spcwsp: 0513-059 The hardmon Subsystem has been started. Subsystem PID is
18352.spcwsp: 0513-004 The Subsystem or Group, splogd, is currently inoperative.
spcwsp: 0513-083 Subsystem has been Deleted.
spcwsp: 0513-071 The splogd Subsystem has been added.
spcwsp: 0513-059 The splogd Subsystem has been started. Subsystem PID is 39454.
spcwsp: Stopping the swtadmd subsystem is currently running
spcwsp: Stopping the swtlog subsystem is currently running
spcwsp: Stopping the swtadmd2 subsystem is currently running
spcwsp: Stopping the emaster subsystem is currently running
spcwsp: 0518-307 odmdelete: 0 objects deleted.
spcwsp: 0513-044 The splogd Subsystem was requested to stop.
spcwsp: 0513-083 Subsystem has been Deleted.
spcwsp: 0513-071 The splogd Subsystem has been added.
spcwsp: 0513-059 The splogd Subsystem has been started. Subsystem PID is 12210.
spcwsp: 0513-044 The splogd Subsystem was requested to stop.
spcwsp: Stopping the swtadmd subsystem is currently running
spcwsp: Stopping the swtlog subsystem is currently running
spcwsp: Stopping the swtadmd2 subsystem is currently running
spcwsp: Stopping the emaster subsystem is currently running
[BOTTOM]
```

Figure 3-32 Output from the install_hacws command

Customize cluster event processing

In the introduction, we discussed the pre- and post-event scripts used by HACWS. In this step, the script `/usr/sbin/hacws/spcw_addevents` will add these pre- and post-events into the HACMP definition, without the user having to step through each change.

This is done by running **smitty hacws** and selecting **Identify Event Scripts to HACMP** (see Figure 3-33 on page 212).

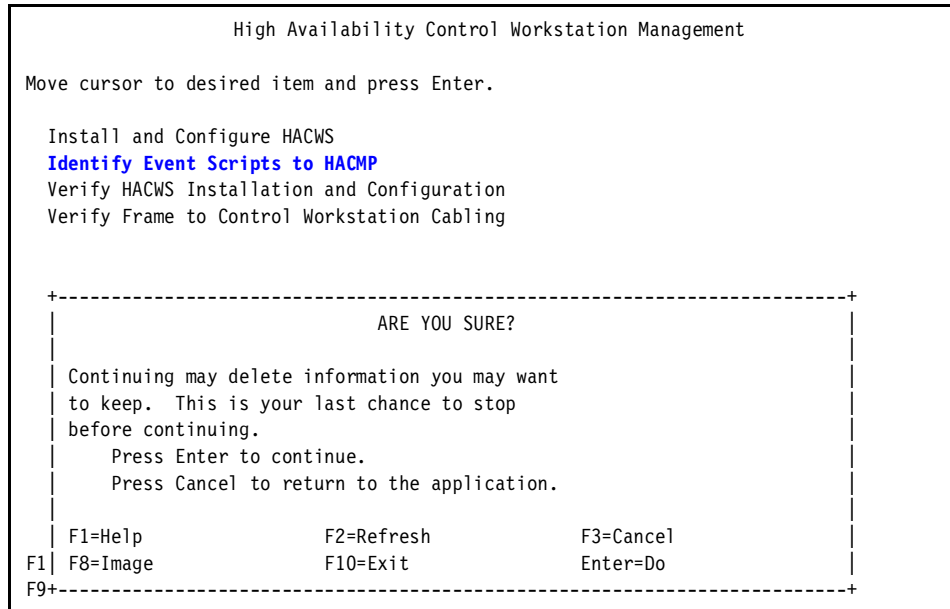


Figure 3-33 Add event scripts to HACMP database

The output of the `spcw_addevents` is shown in Figure 3-34.

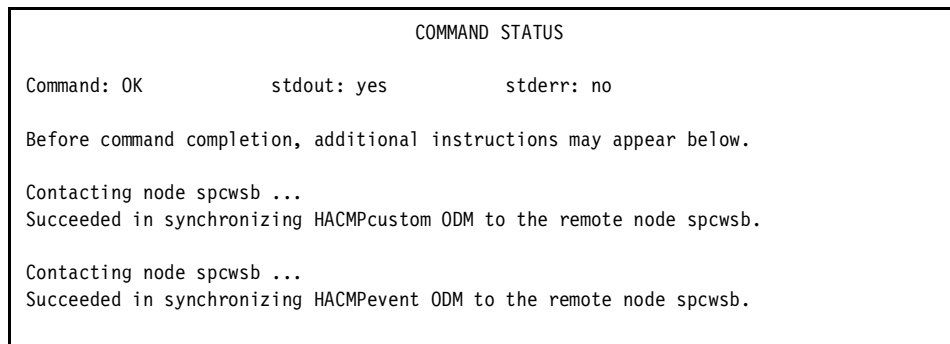


Figure 3-34 `spcw_addevents` output

Add IP Alias

If we are not using the persistent alias feature, you need to add the alias to the boot adapter on the backup control workstation and make this alias permanent.

To make sure this alias is created on boot, add the following line to /etc/rc.backup_cw_alias on the backup control workstation:

```
ifconfig <adapter> alias 'hostname' netmask <netmask> up
```

For example:

```
ifconfig en0 alias 'hostname' netmask 255.255.255.224 up
```

The following entry also needs to be added at the end of /etc/rc.net:

```
if [ -f /etc/rc.backup_cw_alias ]; then
/etc/rc.backup_cw_alias
fi
```

3.6.9 Verify HACWS and hardware configuration

The HACWS configuration now needs to be verified. The script /usr/sbin/hacws/hacws_verify is used, or you can run **smitty hacws** and select **Verify HACWS Installation and Configuration** (see Figure 3-35).

```

                                COMMAND STATUS
Command: OK                      stdout: yes                      stderr: no

Before command completion, additional instructions may appear below.
```

Figure 3-35 Script hacws_verify output

Similarly, the hardware configuration can also be verified using the supplied script /usr/sbin/hacws/spcw_verify_cabling or by running **smitty hacws** and selecting **Verify Frame to Control Workstation Cabling**, as shown in Figure 3-36.

```

Command: OK                      stdout: no                      stderr: no

Before command completion, additional instructions may appear below.

F1=Help      F2=Refresh      F3=Cancel      F6=Command
F8=Image     F9=Shell       F10=Exit      /=Find
n=Find Next
```

Figure 3-36 Script spcw_verify_cabling output

Note: If hardmon is still running, spcw_verify_cabling will not run, so just stop hardmon using `stopsrc -s hardmon`.

3.6.10 Reboot primary and start cluster services

Now we can reboot the two control workstations. After they have both rebooted, check that they are on their boot IP labels and that the persistent alias is configured on the backup control workstation. Use `ifconfig -a` on each control workstation to confirm this.

Cluster services should now be started on the primary control workstation. This is done using the `smitty clstart` fast path and selecting the options shown in Figure 3-37.

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?	true	+
Cluster to re-acquire resources after forced down?	false	+

Figure 3-37 Starting cluster services on the primary control workstation

Monitoring `/tmp/hacmp.out` will show if there are any errors. It will also show when the cluster services have started; you should see the message shown in Example 3-16.

Example 3-16 `hacmp.out` showing cluster services started on primary

```
+ exit 0
                                     HACMP Event Summary
Event: /usr/es/sbin/cluster/events/check_for_site_up_complete spcws
Start time: Thu Jun 13 17:18:24 2002

End time: Thu Jun 13 17:18:26 2002

Action:          Resource:          Script Name:
-----
No resources changed as a result of this event
```

```
-----
....

allnimres: Node 17 (lpar1) prepared for operation: disk.
allnimres: Node 18 (lpar2) prepared for operation: disk.
Tickets destroyed.
setup_server: Processing complete (rc= 0).
SPCW_APPS COMPLETE at Fri Jun 7 18:17:35 EDT 2002
```

Note: You may see network down events in /tmp/hacmp.out for the serial network. This is no cause for alarm, as the cluster services have not started on the backup control workstation. Once cluster services start on the backup control workstation, the serial network should come up.

3.6.11 Verify operation of the primary control workstation

You should now verify that the control workstation applications are successfully running on the now active primary control workstation. For example, the following tests should be run, as shown in Example 3-17:

- ▶ The service address is active on the primary control workstation. This can be verified by running **netstat -i**.
- ▶ The file system /spdata is mounted and accessible on the primary control workstation. This can be verified by running **df -kI**.
- ▶ The location of the resource group hacws_group1 can be shown by the **/usr/sbin/cluster/utilities/clfindres** command.
- ▶ The SDR is available, this can be verified by running **/usr/lpp/ssp/bin/SDRGetObjects** Frame.
- ▶ The hardmon daemon is operational; this can be verified by running **/usr/lpp/ssp/bin/spmon -d**.

Example 3-17 Output on primary CWS

```
spcwsb:/ netstat -i
Name Mtu Network Address Ipkts Ierrs Opkts Oerrs Coll
en0 1500 link#2 0.4.ac.49.c6.41 631568 0 564596 0 0
en0 1500 192.168.6.3 spcwsb 631568 0 564596 0 0
en2 1500 link#4 0.4.ac.5e.69.36 797120 0 202547 0 0
en2 1500 192.168.6.1 spcwsb-priv 797120 0 202547 0 0
...
spcwsb:/ df -k
Filesystem 1024-blocks Free %Used Iused %Iused Mounted on
...
/dev/ulv01 7520256 1624888 79% 17715 1% /spdata
...
```

```

spcwsb:/# /usr/sbin/cluster/utilities/clfindres
GroupName      Type      State      Location      Sticky Loc
-----
hacws_group1  rotating    UP      spcwsb
....
spcwsb:/ SDRGetObjects Frame
frame_number tty      frame_type  MACN      backup_MACN slots
frame_in_config snn_index switch_config hardware_protocol sl_tty
control_ipaddr domain_name
          1 /dev/tty0 switch      spcwsb      spcwsb      16 1
0          0 SP      ""      ""      "" 2 ""      ""
spcwsb      spcwsb      16 "" ""      ""      HMC
""          192.168.6.202 ITSOP690
...
spcwsb:/ spmon -d
1. Checking server process
   Process 35782 has accumulated 0 minutes and 0 seconds.
   Check successful

2. Opening connection to server
   Connection opened
   Check successful

3. Querying frame(s)
   2 frames
   Check successful

4. Checking frames

   This step was skipped because the -G flag was omitted.

5. Checking nodes
----- Frame 1 -----
Slot Node Type  Power Host      Switch  Key      Env  Front Panel  LCD/LED
      Responds Responds Switch  Error LCD/LED      Flashes
-----
   1   1  wide   on    no      notcfg  N/A    no  LCDs are blank  no
   3   3  wide   on    no      no      N/A    no  LCDs are blank  no
   5   5  wide   on    no      no      N/A    no  LCDs are blank  no
   7   7  wide   on    no      no      N/A    no  LCDs are blank  no

----- Frame 2 -----
Slot Node Type  Power Host      Switch  Key      Env  Front Panel  LCD/LED
      Responds Responds Switch  Error LCD/LED      Flashes
-----
   1   17 thin   on    no      notcfg  N/A    N/A  LCDs are blank  N/A
   2   18 thin   on    no      notcfg  N/A    N/A  LCDs are blank  N/A

```

3.6.12 Start the backup control workstation

Once the cluster services have been started and verified on the primary control workstation successfully, then the cluster services can now be started on the backup control workstation. This is done as before, by running **smitty clstart**. Monitor the `/tmp/hacmp.out` file for any errors.

3.6.13 Starting of cluster services on the primary workstation

As discussed above, it is recommended that cluster services not be automatically started on both control workstations. However, you can automate the start of cluster services by adding an entry into the `inittab` file only on the primary control workstation.

This will mean that:

- ▶ If the primary control workstation is rebooted and the backup control workstation is the active control workstation, nothing will happen, except the supper updates will start
- ▶ If the primary control workstation is rebooted and the backup control workstation is not active, then the primary control workstation will become the active control workstation.

To make this change on the primary control workstation, run **smitty clstart** and change the option to restart, as shown in Figure 3-38.

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	restart	+
BROADCAST message at startup?	true	+
Startup Cluster Lock Services?	false	+
Startup Cluster Information Daemon?	true	+
Cluster to re-acquire resources after forced down?	false	+

Figure 3-38 Setting cluster services on control workstation to start automatically

This will add the following line to `/etc/inittab`:

```
hacmp:2:wait:/usr/sbin/etc/rc.cluster -boot> /dev/console 2>&1 # Bring up Cluster
```

3.6.14 Backups

Make a mksysb backup of each control workstation, and a backup of the shared data volume group. The backup control workstation should be added to the site's backup cycle.

3.6.15 Testing HACWS

As noted in the discussion of the various scenarios, the addition of the HMC on the SPLAN will introduce nothing further in the HACMP configuration, so we only examined the case of the HMC on a private network.

The following tests were conducted:

- ▶ Testing failover and the operation of the backup control workstation
- ▶ Testing adapter failure (optional)
- ▶ Testing failure of private network for CWS and HMC

Testing failover and the operation of the backup CWS

Once cluster services has been successfully started on both control workstations, we can test the failover of the resources to the backup control workstation. This is done by stopping the cluster services on the primary control workstation with the takeover option set. Use **smitty clstop**, as shown in Figure 3-39.

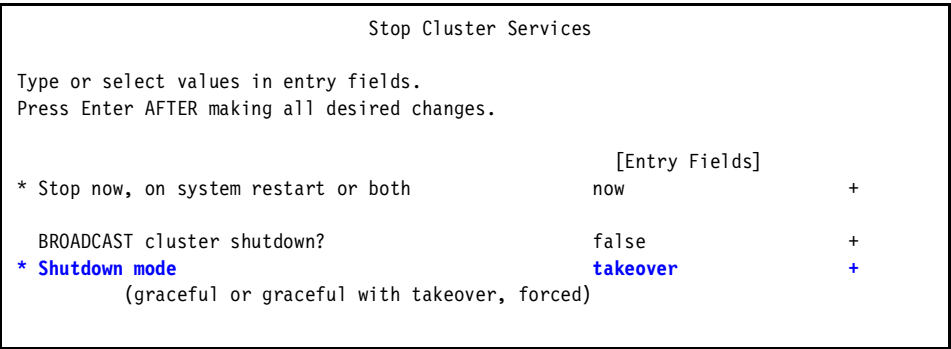


Figure 3-39 Stop cluster services with takeover

Attention: The first time the backup control workstation becomes the active control workstation, it will take some time. This is because the SPOT has to be rebuilt on the backup control workstation. Again, check /tmp/hacmp.out for the line SPCW_APPS COMPLETE at Fri Jun 7 18:17:35 EDT 2002.

Again, monitor the file /tmp/hacmp.out on the backup control workstation to confirm that the resources were taken over without any error. The same tests can be conducted to confirm the operation of the control workstation were conducted on the primary control workstation when it was active.

Once testing is complete, cluster services can be restarted on the primary control and the resource group can either be:

- ▶ Failed back by stopping cluster services on the backup control workstation with takeover.
- ▶ Moved back by moving the resource group back to the primary control workstation using **smitty cl_resgrp**.

Testing adapter failure (optional)

If you have configured your cluster with standby adapters, the following test can be conducted.

1. Ensure that the cluster services are running on both control workstations.
2. Ensure that one workstation is the active control workstation.
3. On the primary control workstation, remove the network cable from the adapter with the service IP label and monitor the file /tmp/hacmp.out.

Example 3-18 shows the entries in /tmp/hacmp.out.

Example 3-18 Swap adapter events in hacmp.out

```
Jun 13 18:26:10 EVENT START: swap_adapter spcws spcwsether 192.168.6.200
192.168.6.33

:swap_adapter[115] [[ high = high ]]
:swap_adapter[115] version=1.33.1.18
:swap_adapter[116] :swap_adapter[116] cl_get_path
HA_DIR=es
:swap_adapter[118] STATUS=0
:swap_adapter[119] [ ! -n ]
:swap_adapter[121] EMULATE=REAL
:swap_adapter[124] [ 4 -ne 4 ]
:swap_adapter[131] [ spcws != spcws ]
:swap_adapter[138] export NSORDER=local
:swap_adapter[144] :swap_adapter[144] name_to_addr 192.168.6.200
:swap_adapter[2] cllsif -cSn 192.168.6.200
.....
:swap_adapter[363] [ 0 -ne 0 ]
:swap_adapter[369] [ 0 -ne 0 ]
:swap_adapter[374] exit 0
Jun 13 18:26:34 EVENT COMPLETED: swap_adapter spcws spcwsether 192.168.6.200
192.168.6.33
```

HACMP Event Summary

Event: swap_adapter spcwspp hacwsether 192.168.6.200 192.168.6.33

Start time: Thu Jun 13 18:26:10 2002

End time: Thu Jun 13 18:26:34 2002

Action:	Resource:	Script Name:

Acquiring resource:	192.168.6.33	cl_swap_IP_address
Search on:	Thu.Jun.13.18:26:11.EDT.2002.cl_swap_IP_address.192.168.6.33.ref	
Resource online:	192.168.6.33	cl_swap_IP_address
Search on:	Thu.Jun.13.18:26:32.EDT.2002.cl_swap_IP_address.192.168.6.33.ref	

Jun 13 18:26:34 EVENT START: swap_adapter_complete spcwspp hacwsether
192.168.6.200 192.168.6.33

```
:swap_adapter_complete[111] [[ high = high ]]  
:swap_adapter_complete[111] version=1.17.1.14  
:swap_adapter_complete[112] :swap_adapter_complete[112] cl_get_path  
HA_DIR=es  
:swap_adapter_complete[115] export NSORDER=local  
:swap_adapter_complete[118] [ 4 -lt 3 ]  
:swap_adapter_complete[124] [ ! -n ]  
:swap_adapter_complete[126] EMULATE=REAL  
:swap_adapter_complete[130] :swap_adapter_complete[130] cllsif -Scn  
192.168.6.200  
.....
```

HACMP Event Summary

Event: swap_adapter_complete spcwspp hacwsether 192.168.6.200 192.168.6.33

Start time: Thu Jun 13 18:26:34 2002

End time: Thu Jun 13 18:26:35 2002

Action:	Resource:	Script Name:

No resources changed as a result of this event		

Note: If your control workstations do not have standby adapters, then the above test should result in the application being failed over to the other control workstation

Testing failure of the private network for CWS and HMC

If the adapter on a private network fails, then failover to the inactive CWS occurs. Example 3-19 shows the particular events in hacmp.out on the active backup control workstation that illustrate:

- ▶ Network down for private network
- ▶ HACWS post-event script called
- ▶ HACWS post-event issuing a **c1stop**
- ▶ Node down complete

Example 3-19 hacmp.out showing network down becoming node down

Jun 21 11:01:52 EVENT START: network_down spcws hacwsether_priv

```
:network_down[62] [[ high = high ]]
:network_down[62] version=1.21
:network_down[63] :network_down[63] c1_get_path
HA_DIR=es
:network_down[65] [ 2 -ne 2 ]
:network_down[77] :network_down[77] c1_rrmethods2call net_cleanup
....
+ exit 0
:network_down.post_event[68] EVENT_NAME=network_down
:network_down.post_event[69] EVENT_EXIT_STATUS=0
:network_down.post_event[70] NODE_NAME=spcws
:network_down.post_event[71] NETWORK_NAME=hacwsether_priv
....
:network_down.post_event[149] :network_down.post_event[149]
/usr/lpp/ssp/install/bin/node_number
node_num=0
:network_down.post_event[151] [[ 0 != 0 ]]
:network_down.post_event[159] :network_down.post_event[159] lshacws
HACWS_STATE=2
:network_down.post_event[160] [[ 2 != 2 ]]
:network_down.post_event[166] exec c1stop -N -gr -y -s
+ [[ high = high ]]
+ version=1.2.2.36
+ + c1_get_path
...
```

HACMP Event Summary

Event: node_down_complete spcws

Start time: Fri Jun 21 11:04:30 2002

End time: Fri Jun 21 11:04:53 2002

Action:	Resource:	Script Name:
---------	-----------	--------------

Resource group offline:	hacws_group1	process_resources
Search on:	Fri.Jun.21.11:04:33.EDT.2002.process_resources.hacws_group1.ref	

3.7 Considerations

The following restrictions apply to support of the high availability control workstations:

- ▶ DCE is not supported.
- ▶ There are no problems in the authentication of both control workstations on the nodes with restricted root access option enabled. However, you have to copy the authorization files `/.rhosts` and `/.klogin` to the backup control workstation. It is also important that restricted root access from the active primary control workstation is activated, since it is the Kerberos Version 4 master.
- ▶ You cannot specify none for the **chauthent** command when using HACWS.
- ▶ IPv6 is not supported.
- ▶ The load cannot be spread across the two control workstations, as this is a hot standby configuration.
- ▶ Each control workstation must be included in the list of supported servers for HACMP.
- ▶ The pair of control workstations can only support one SP system.
- ▶ Full control workstation functions from the backup control workstation is not supported for SP-attached servers, due to limitations on the serial terminal support connections.

Configuration changes

It is recommended that configuration changes are only made on the SP system when the primary control workstation is active. Most configuration changes are allowed when the backup control workstation is active. However, some changes require the updating of configuration files, and these updates need to be pushed back to the primary control workstation; this is the opposite direction to the normal operating situation, and it is easier to manage the system if these updates only go in one direction.



HAGEO integration with HACMP cluster

The IBM High Availability Geographic Cluster for AIX (HAGEO) software product provides a flexible, reliable platform for building disaster-tolerant computing environments. HAGEO components provide the capability for mirroring data across TCP/IP point-to-point networks over an unlimited distance from one geographic site to another. It works with the IBM High Availability Cluster Multi-Processing (HACMP) licensed program product to provide automatic detection, notification, and recovery of an entire geographic site from failures. This chapter discusses some basic disaster recovery concepts and explains what the HAGEO software can do to help you prepare for disaster recovery and carry it out.

4.1 HAGEO integration with HACMP

HAGEO 2.4 integrates and uses the HACMP high availability infrastructure for managing geographic clusters for disaster recovery. The High Availability Cluster Multi-Processing for AIX (HACMP) software product addresses recovery from the failure of a computer, an adapter, or a local area network within a computing complex at a single site. The HAGEO software extends the HACMP cluster to encompass two physically separate data center sites. Data entered at one site is sent across a point-to-point TCP/IP network and is mirrored at a second, geographically distant location. Because HAGEO does not require the use of fiber optic cable, there is no limit to how far apart these locations can be.

This chapter will describe the HAGEO 2.4 new features and give examples of its configuration. We expect the reader to already understand the HACMP 4.5 architecture and have basic knowledge of previous versions of HAGEO. For detailed information on previous versions of HAGEO, see *Disaster Recovery Using HAGEO and GeoRM*, SG24-2018.

4.1.1 History

Before we start describing the new functions introduced in HAGEO 2.4, let us briefly look at the previous releases of the product. See Table 4-1 for a summary of the HAGEO releases.

Table 4-1 Summary of HAGEO releases

HAGEO releases	New features
2.1	<ul style="list-style-type: none">▶ First IBM “badged” release of HAGEO▶ Synchronous, asynchronous, and MWC mirroring▶ Support for cascading, rotating, and concurrent resource groups▶ Several configuration restrictions▶ HACMP “Classic” support only

HAGEO releases	New features
2.2	<ul style="list-style-type: none"> ▶ First IBM owned release of HAGEO ▶ Geographic Remote Mirroring (GeoRM) introduced. GeoRM is a disaster recovery software that includes point-to-point mirroring of critical data without the automatic takeover of applications on the takeover site. ▶ Configuration from a single point ▶ Tunable parameters added ▶ Improved tracing and logging ▶ NLS support ▶ Support for HACMP ES
2.3	<ul style="list-style-type: none"> ▶ Performance and management improvements ▶ Support for large GeoMirror Device (GMD) sizes ▶ Ability to resize geo-mirrors dynamically ▶ Simpler and faster configuration ▶ GMD state is preserved during resynchronization ▶ Local peers in asynchronous mode ▶ HAGEO operations log to HACMP log files ▶ More flexible data transmission rates
2.4	<ul style="list-style-type: none"> ▶ Latest release of HAGEO ▶ Integration of HAGEO with HACMP ES ▶ Dropped support for HACMP HAS (Classic) ▶ GeoManager removed from HAGEO ▶ Option to use TCP instead of UDP protocol for mirroring ▶ 64-bit kernel support

New features of HAGEO 2.4 are described in Section 4.3, “New features of HAGEO 2.4” on page 235.

4.2 Planning

In this section, we will describe the hardware and software requirements and show the general planning approach to an HAGEO geographical cluster using HAGEO 2.4. An example scenario is presented in Section 4.4, “Clustering with HAGEO” on page 246.

4.2.1 Hardware requirements

The planning of an HAGEO environment is focused on disaster recovery. This requires that the computing resources must be spread geographically to distant locations (sites). The sites must be interconnected by a WAN network that has significant impact on the behavior of the design. Each site should consider designing the local high availability to use HACMP for local failover. See Figure 4-1 for a general diagram of an HAGEO cluster design. The figure shows two sites (site A and site B) interconnected by a WAN network.

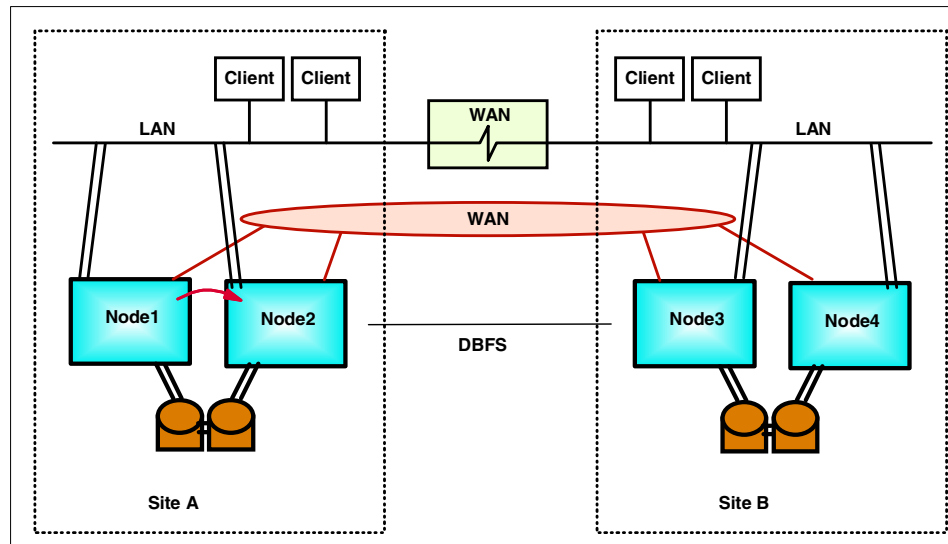


Figure 4-1 General HAGEO design

The hardware planning consists of the following:

- ▶ Site planning
- ▶ Local high availability
- ▶ Data space design
- ▶ Network planning

Site planning

Selection of the sites is a critical factor. The main focus is the distance of the sites. While we decrease the chances that both sites will experience a disaster in the same time with increasing distance, the farther the sites are, the more complicated and expensive it is to have sufficient network connections.

The usual configuration is to have two sites (site A and site B). Each site has a local HACMP cluster configuration, consisting of two or more nodes at each site

(node1 and node2 at site A, node3 and node4 at site B). All these nodes are defined into one HACMP cluster.

The selection of sites is related to the location of the clients and the accessibility of the site. If most of the clients are located at the primary site and the physical access to the secondary site is restricted, we usually plan for one of them to be the active site (the primary) and the other to be the hot standby (the backup site). If the clients are located at both sites, we may consider a mutual takeover configuration where the first site runs one group of applications and the second site runs another group of applications. Each site backs up the other. For this mutual takeover design, we should consider the performance issues of the network during normal operation and the server capacity when all applications reside at the same site, in the event of a disaster.

Local high availability

In all high availability scenarios, component failures should be solved locally. Eliminating single point of failures locally is more simple, more efficient, and less expensive. The local high availability should be implemented using HACMP architecture, which eliminates single point of failures locally. Use the HACMP planning procedures, as described in *HACMP for AIX 4.4.1 Planning Guide*, SC23-4277, to design the local cluster.

The local cluster planning (HACMP planning) involves the following steps:

- | | |
|---------------------------|---|
| Cluster topology | Define the cluster topology, like cluster ID, nodes, local networks, and network adapters. |
| Cluster resources | Define shared resources the cluster will handle and group them into resource groups. Select resources, such as disks, volume groups or file systems, NFS mounts, NFS exports, shared tapes, application servers, and so on. |
| Custom definitions | Define error log events, custom events, custom snapshots, and so on. |

In Figure 4-1 on page 226, site A and site B have local sub-clusters configured that consist of Node1 and Node2 at Site A and Node3 and Node4 at Site B. Local failures, such as disk, disk adapter, network adapter, and processor or node failures are handled locally.

Note: HAGEO requires the disks containing geographically mirrored data to be handled by HACMP and must be part of a resource group. We cannot implement HAGEO using logical volumes on non-shared (local) disks.

Data space design

When designing a geographical cluster, you should group your data spaces into three groups:

Critical data	The critical data is expected to be synchronized online and kept ready for use at the backup site while it is used on the primary site
Static data	The static data does not change frequently. This data is, for example, the operating system files, application binaries, static archives, and so on. This data does not need to be synchronized online, but synchronization must be done on a regular basis, such as after every installation of new software, after administrative system changes, and so on.
Temporary data	The temporary data does not need to be synchronized at all. The loss of any temporary data does not lead to any loss to the overall system consistency. Examples of this data group are the /tmp, /scratch or /transfer file systems, or some temporary tablespaces in databases.

The result of the data grouping is a table that gives us information about the size of each group of data. Focus on the critical data, because this data should be mirrored geographically, and the information is one of the basic pieces of information needed for the network planning.

There is a required measurement for the data update behavior. The most important value is the amount of writes and type of writes of the critical data.

Network planning

The HAGEO network design consists of planning the local and the geographic networks. There are four necessary network types to plan:

- ▶ Primary geographic network
- ▶ Backup geographic network
- ▶ Public network
- ▶ Local HACMP networks

Primary geographic network

The primary geographic network, called GeoPrimary, is *the* core network for the geographic mirroring of the data because it has to transfer all updates in a minimal amount of time across long distances. With increasing distance, we need to consider several additional impacts, as compared to local high availability.

Check the following network characteristics for their overall performance impacts:

Bandwidth	The network speed measured in amount of bits, or bytes, per second, is a good measurement tool for large volume streams using large data blocks. It is also a required parameter when calculating the time needed to transfer the changed data across the network. The higher the bandwidth is, the better performance we achieve.
Latency	The delay in transferring the data, called the network latency, plays a significant role in online synchronization of small blocks of data. As each data block must be accepted and confirmed by the remote location, the transfer protocol must wait for this confirmation. The wait time is the time that a data block moves to a remote location plus the time the confirmation comes back. The smaller the latency is the better performance we achieve.
Link stability	During the transmission of data across a network, we expect that the network may lose data blocks or may have drop outs for relatively long intervals. These intervals may then cause synchronization failures between primary and remote locations. HAGEO handles these failures, but the user may believe that the system is not responding for some time.
Price	The price of the link is often an influence on a decision, so it should be considered.

Attention: When speaking of networking terminology, the term “bit per second” (b/s, Kb/s, Mb/s) is used. If speaking of data terminology, the term “byte per second” (B/s, KB/s, MB/s) is used. In several calculations, we assume that 1B (byte) = 8b (bit).

Also, remember that the bandwidth is the burst data that can be transferred through the network and not the real data. During transfers across networks, we should consider protocol headers, frame control, network management communications, and so on.

HAGEO allows us to implement multiple GeoPrimary networks. If multiple networks are defined, HAGEO will apply load balancing to the GeoPrimary networks on a round-robin basis. Multiple GeoPrimary networks will also increase the availability of the GeoPrimary networks as a whole.

There is a very useful tool for planning your GeoPrimary Networks that is installed from the hageo.gmdsizing filesset on the HAGEO install media. Start the tool during the most disk intensive operations on your current system, for

example, peak work hours, or year or quarter end processing. The tool is started by the **gmdsizing** command or by running **smit hageo** and selecting **HAGEO Utilities -> Measure Disk Utilization**.

The **gmdsizing** command is used to estimate network bandwidth requirements for the GeoPrimary networks used to support GeoMirror traffic. It monitors disk utilization over a given period of time and prints a report. This report can then be used to help determine the bandwidth needs. See the README file supplied with the product or the system man pages for the command syntax.

Backup geographic network

The primary geographic network is the core of geographic mirroring. If this network fails, the backup site tries to contact the primary site. If there is no other network implemented, the backup site cannot determine if the primary site failed, the network failed, or the IP layer failed. For this reason, there must be at least one other network, that is, non-IP, implemented that can be used to diagnose if the remote site is down or the primary geographic network failed.

The backup network is not used to carry any traffic; it only provides a way to differentiate site failures from network failures. For this purpose, we can use two designs:

GeoSecondary The GeoSecondary network is a non IP network, for example, a RS232 line, that can carry heartbeat messages. If the sites are distant, the RS232 line may be converted to a different protocol. The GeoSecondary network is sometimes referenced as the Secondary Geographical Network (SGN). SGN is a serial line connection between two nodes at different geographic sites. They are usually routed through an X.25 network or SLIP modem lines. No GMD traffic occurs on these lines. They are only used by the HACMP cluster manager to better determine the nature of the failure (to prevent site isolation)

Dial Back Fail Safe The Dial Back Fail Safe (DBFS) is a dialed line connection through standard modems. Its configuration requires you to have public telephone lines available at each site that are supported by the selected modems. The minimal requirement is to have one modem at each site. The optimal requirement is to have one modem at each node.

The DBFS is a more popular design, because the price of the public telephone lines are much lower than the prices of private telecommunication networks.

Public network

The public network is a network through that the users access their applications that are running on the cluster. This network is usually an Ethernet network, which may be connected between sites into one LAN emulation, or the sites may be interconnected through routers.

If the sites are connected to the same LAN that is used as a public network, we can configure the IP address takeover (IPAT) to the HACMP resource group that handles resources geographically. In this case, adapters in each node should be planned according to the HACMP planning for IPAT. For details, see the *HACMP for AIX 4.4.1 Planning Guide*, SC23-4277.

Note: Most of the planning mistakes are related to the number of standby adapters on nodes at the backup site. If the primary site fails, the nodes at the backup site must takeover all service addresses from the primary site.

If the sites are interconnected through routers, the IPAT cannot be implemented between sites, because the routers cannot handle this situation by default. The dynamic routing protocols, like RIP, OSPF, and so on, are not supported on the HACMP cluster nodes. A solution is to generate SNMP traps to the routers after each failover and do additional programming of the routers, which is beyond this scope of this book. However, IPAT can be implemented only locally at each site and, in case of site failure, the users will connect to the backup site using a different IP address.

The GeoPrimary network cannot be used for client communications, because this network does not support IP address takeover (IPAT), and the network overload of the public network may negatively affect the performance and monitoring of the GeoPrimary network.

Restriction: Organization policies sometimes restrict site interconnections. Check them before you plan the geographic networks.

Local HACMP networks

The local HACMP networks are related to the design of an basic HACMP planning and are not part of the HAGEO planning. An example of these networks could be a serial network between nodes, SP Switch2 networks, and so on. However, these networks are configured into the HACMP and the HAGEO configuration with the same environment. These networks are part of the local high availability referred to “Local high availability” on page 227.

4.2.2 Software requirements

We define the required software versions according to the HAGEO 2.4 requirements.

Operating system

The required level of the operating system is AIX 5L Version 5.1. However, we recommend you install the latest maintenance level. The following list is a list of minimum prerequisite filesets for HAGEO 2.4.0.0:

- ▶ bos.rte 5.1.0.0 (all bos.rte filesets)
- ▶ bos.net.tcp.client 5.1.0.0
- ▶ perl.rte 5.5.0.0
- ▶ bos.adt.lib 5.1.0.0
- ▶ bos.adt.libm 5.1.0.0
- ▶ bos.adt.syscalls 5.1.0.0
- ▶ perfactent.tools 2.2.32.0
- ▶ xIC.rte 3.6.4.0
- ▶ bos.rte.man 5.1.0.0 (if the man pages for HAGEO need to be installed)

The AIX operating system must be installed on every node of the HAGEO cluster. We do not recommend having different maintenance levels on the nodes within the same cluster.

RSCT

The version of RSCT should tightly correspond to the version of the operating system and HACMP. For AIX 5L 5.1 use RSCT Version 2.2.1. If recommended maintenance levels exist, install them, unless otherwise noted. The following list is a list of minimum prerequisite filesets for HAGEO 2.4.0.0:

- ▶ rsct.basic.hacmp 2.2.1.0
- ▶ rsct.compat.basic.rte 2.2.1.0
- ▶ rsct.compat.basic.hacmp 2.2.1.0
- ▶ rsct.compat.clients.hacmp 2.2.1.0
- ▶ rsct.compat.clients.rte 2.2.1.0

The RSCT software must be installed on every node of the HAGEO cluster. The RSCT version and maintenance should be at the same level on every node.

HACMP

HACMP 4.5 is required to use HAGEO 2.4. Check that the correct level of HACMP is installed before HAGEO is installed. The following list is a list of minimum prerequisite filesets for HAGEO 2.4.0.0:

- ▶ cluster.es.server 4.5.0.0
- ▶ cluster.es.client 4.5.0.0

- ▶ cluster.es.cspoc 4.5.0.0

Important: HAGEO 2.4 requires HACMP 4.5 ES (Enhanced Scalability) to be installed. The HACMP HAS (“Classic”) version is not supported.

The HACMP software must be installed on every node of the HAGEO cluster. The HACMP version and maintenance should be at the same level on every node.

4.2.3 Configuration examples

Here we consider the three most popular HAGEO designs:

- ▶ Two nodes at each site
- ▶ Two nodes at the primary site and one node at the backup site
- ▶ One node at each site

Two nodes at each site

This design is the most recommended one and gives the most flexibility. The primary and also the backup site has an full HACMP cluster installed. All component failures are handled locally at each site. Only site failures result in failover across sites. See Figure 4-2 for a diagram of this layout.

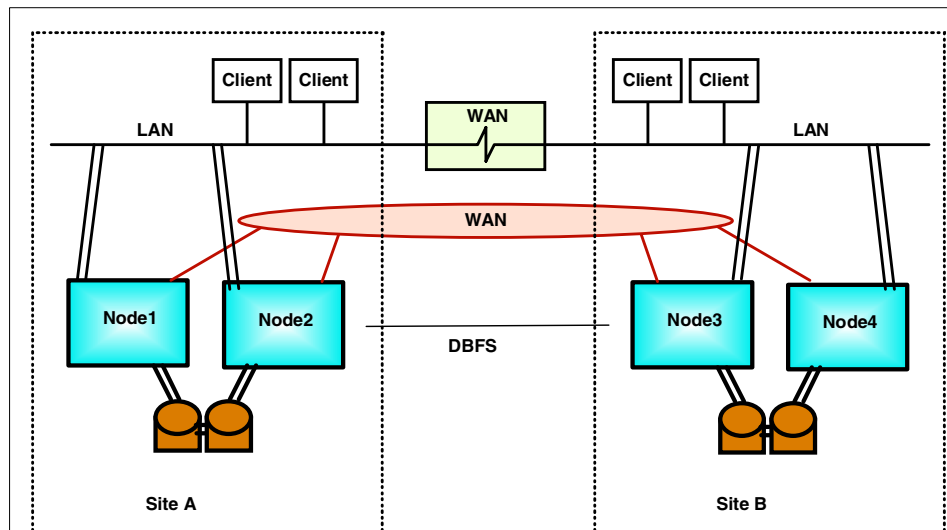


Figure 4-2 Two nodes at each site

Two nodes at the primary site and one node at the backup site

This is not a preferred design, but it is more popular. We handle all component failures locally only at the primary site. The secondary site propagates any node failures to a site failure. Even though the backup site can be configured to fail over anything from the primary, this design is more satisfactory if there are no applications running at the backup site, which avoids unnecessary disruption in case of maintenance on the backup. See Figure 4-3 for a diagram of this layout.

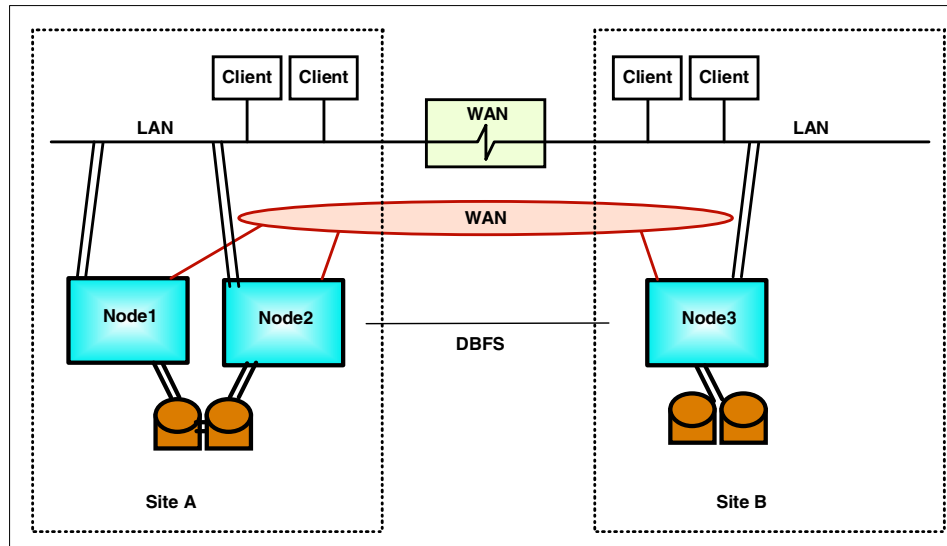


Figure 4-3 Two nodes at the primary site and one node at the secondary site

This configuration looks nice, but we need to remember that if maintenance is being performed on the backup site (for example, a processor upgrade), the node at the backup site (the only one node) will be down and the data updated at the primary site cannot be synchronized to the backup site. When the node at the backup site comes up, it has to synchronize all changes during the site down. However, by using the design described in “Two nodes at each site” on page 233 is implemented, the node at the backup site can first fall over to the second node of this backup site, keeping the site online. We use this design for our example in Section 4.4, “Clustering with HAGEO” on page 246.

One node at each site

This design is a very restricted one. It may handle disk or adapter failures, but node failures are propagated to site failures, because there are no local peer nodes available in this model. See Figure 4-4 on page 235 for a diagram of the layout. We recommend weighing your system requirements appropriately before choosing a design that best fits your needs.

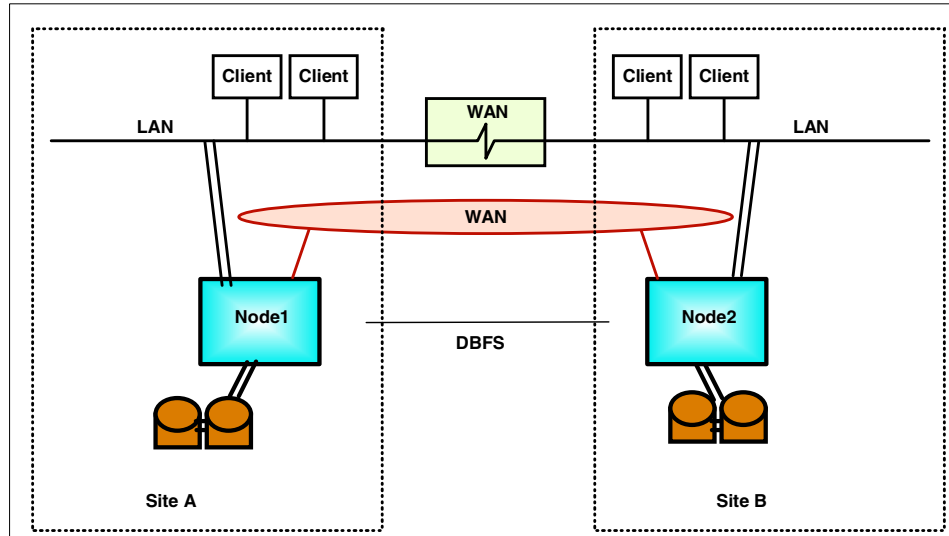


Figure 4-4 One node at each site

4.3 New features of HAGEO 2.4

This section describes the new features of HAGEO 2.4.

4.3.1 Integration with HACMP

HAGEO 2.4 is supported with HACMP/ES 4.5. Configuration with HACMP HAS ("Classic" HACMP) is not supported. Therefore, throughout this section, we always assume HACMP/ES when referring to HACMP.

The main focus of the new version of the HAGEO is on its integration with the HACMP environment. This integration is done in the following way:

- ▶ Configuration of the HAGEO topology is now implemented through HACMP topology configuration only.
- ▶ HACMP monitoring, using RSCT, is used to monitor the GeoPrimary and GeoSecondary networks. If DBFS is configured, HACMP calls the DBFS utilities from HAGEO.
- ▶ Resource groups are configured to reflect both node and site relationships within the same resource group.
- ▶ Events are generated and handled by HACMP.

HAGEO topology configuration

The configuration of the HAGEO topology is integrated into the HACMP configuration procedures. The HAGEO installation adds the GeoPrimary and GeoSecondary network definitions into HACMP. Also, the site definitions are defined through the HACMP menus.

To correctly configure the topology, do the following using the HACMP configuration tools:

1. Run **smit hacmp**.
2. Configure all nodes at both sites (primary and backup) belonging to one cluster.
3. Configure site definitions. The site definitions contain a list of nodes belonging to a site and the characteristics of a site.
4. Configure all networks and adapters, including the GeoPrimary and GeoSecondary geographic networks.
5. Synchronize the HACMP topology.

The HACMP and HAGEO topology configuration is now ready, so HACMP can now monitor the geographic networks.

If you compare the HAGEO configuration menu to the previous versions of HAGEO, there is one item missing in the HAGEO 2.4 initial SMIT menu (SMIT HAGEO) for configuring HAGEO. This menu item is Configure Geographic Topology, which is fully integrated into HACMP. These functions are fully integrated into HACMP.

Note: You no longer import any configurations from HACMP into HAGEO. Configure the geographic topology in the HACMP topology only.

If DBFS is used instead of the GeoSecondary network, than additional configuration should be done, similar to the previous versions.

Node configuration

The node configuration is done directly through HACMP node configuration, as shown in Figure 4-5 on page 237. Add all nodes of the cluster, including all nodes from the primary and backup sites.

Add Cluster Nodes

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

* Node Names

[Entry Fields]
[YOK01 YOK02 SEOUL]

Figure 4-5 Define cluster nodes

Site configuration

Configure site definitions through the `smit hacmp` command, as shown in Figure 4-6. Run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Configure Sites -> Add a Site**.

Add Site

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

* Site Name

[Entry Fields]
[YOKOHAMA]

* Site Nodes

[YOK01 YOK02]

+

* Site Dominance

[yes]

+

* Site Backup Communications Type

[dbfs]

+

Figure 4-6 SMIT HACMP - Add Site

Configure geographical networks and adapters

The GeoPrimary and GeoSecondary networks are configured through HACMP configuration. Both geographical networks are defined as *private* networks. See Figure 4-7 on page 238 for an example of the menu.

To configure the geographical networks and adapters, run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Configure Networks -> Configure IP-based Networks -> Add a Network**.

Add an IP-based Network

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

	[Entry Fields]	
* Network Name	[GEO_NET1]	
* Network Attribute	private	+
Network Type	[Geo_Primary]	+
Subnet(s)	[192.168.6.32/27]	+

Figure 4-7 Configure GeoPrimary network

The GeoPrimary adapters are defined as any standard HACMP adapter. A detailed description on how to configure all networks and network adapters can be found in Section 4.4, “Clustering with HAGEO” on page 246.

Synchronize topology

The HACMP configuration for HAGEO nodes, sites, networks and adapters are synchronized by standard HACMP Topology synchronization. Run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Synchronize Cluster Topology**.

The topology synchronization copies the cluster, site, node, network, and adapter configuration to all defined nodes of the cluster, including both sites.

HACMP monitoring of geographic networks

The HACMP exploits the RSCT monitoring capabilities using the topology services. The topology services are represented by the topsvcs subsystem, which is handled by the System Resource Controller. The topsvcs subsystem starts the hats process, which is the master process for the topology services. The hats process starts a network interface module for every network as a plug-in to the hats process. Example 4-1 shows the hats processes on a cluster node.

Example 4-1 List of hats processes

```
# ps -ef | grep hats
root 16678 6974 0 12:19:50 - 0:07 /usr/sbin/rsct/bin/hatsd -n 4 -o
deadManSwitch
root 19904 16678 0 12:19:55 - 0:05 /usr/sbin/rsct/bin/hats_nim
root 24244 16678 0 12:19:54 - 0:07 /usr/sbin/rsct/bin/hats_nim
```

The network interface module plug-in is an interface that binds to the network interface and monitors its availability by sending packets through the interface. When looking at the output of the `lssrc -ls topsvcs` command, as shown in

Example 4-2, we can see statistics of all defined HACMP networks, including the GeoPrimary network.

Example 4-2 Output of lssrc -ls topsvcs command

```
# lssrc -ls topsvcs
Subsystem      Group      PID      Status
topsvcs        topsvcs    16678    active
Network Name   Indx Defd Mbrs St Adapter ID      Group ID
PUB_NET1_0     [ 0]    2    2  S 192.168.6.131    192.168.6.132
PUB_NET1_0     [ 0] en1      0x2d10af2d    0x2d10b0ca
HB Interval = 1 secs. Sensitivity = 10 missed beats
Missed HBs: Total: 0 Current group: 0
Packets sent   : 7404 ICMP 0 Errors: 0 No mbuf: 0
Packets received: 12522 ICMP 0 Dropped: 0
NIM's PID: 24244
GEO_NET1_0     [ 1]    3    3  S 192.168.6.35     192.168.6.37
GEO_NET1_0     [ 1] en0      0x2d10af2e    0x2d10b665
HB Interval = 2 secs. Sensitivity = 12 missed beats
Missed HBs: Total: 0 Current group: 0
Packets sent   : 4376 ICMP 0 Errors: 0 No mbuf: 0
Packets received: 7920 ICMP 0 Dropped: 0
NIM's PID: 19904
  2 locally connected Clients with PIDs:
haemd( 15770) hagsd( 35682)
  Dead Man Switch Enabled:
    reset interval = 1 seconds
    trip interval = 48 seconds
  Configuration Instance = 118
  Default: HB Interval = 1 secs. Sensitivity = 4 missed beats
  Daemon employs no security
  Segments pinned: Text Data.
  Text segment size: 676 KB. Static data segment size: 344 KB.
  Dynamic data segment size: 4297. Number of outstanding malloc: 139
  User time 2 sec. System time 5 sec.
  Number of page faults: 190. Process swapped out 0 times.
  Number of nodes up: 3. Number of nodes down: 0.
```

The following list contains other subsystems used by HACMP:

topsvcs	Topology services
grpsrvcs	Group services
emsvcs	Event management services
clstrmgrES	Cluster manager
clsmuxpdES	Cluster SNMP interface
clinfoES	Cluster information daemon

Find more information on these subsystems in *RS/6000 SP High Availability Infrastructure*, SG24-4838 and *HACMP Enhanced Scalability*, SG24-2081.

HAGEO resources

The resource groups describing the geographic takeover are defined in a new way. Define the resource group that contains all nodes participating in the takeover. The order of the nodes reflects the order of the takeover. In the previous versions of HACMP and HAGEO, you were required to create three resource groups, two local at each site and one between sites. This can now all be handled by a single resource group.

Note: You no longer define site names in the node list for the resource group.

The site relationship is configured within the resource group by the **Site Relationship** menu. Select Cascading, Rotating, Concurrent, or Ignore, depending on you configuration.

See Figure 4-8 for an example for creating a geographic resource group. Run **smit hacmp** and select **Cluster Configuration -> Cluster Resources -> Define Resource Groups -> Add a Resource Group**.

Add a Resource Group

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

	[Entry Fields]	
* Resource Group Name	[GEO_RG]	
* Node Relationship	cascading	+
* Site Relationship	cascading	+
* Participating Node Names / Default Node Priority	[YOK01 YOK02 SE0UL]	+

Figure 4-8 Creating geographic resource groups

Because of the integrated features, you no longer need to define the GMD devices in Miscellaneous Data field of the Resource Group attributes. Instead, define the GMD devices in the new menu field, GeoMirror Devices, in the HACMP Resource Group definitions. See Figure 4-9 on page 241 for an example of defining GMD devices. Run **smit hacmp** and select **Cluster Configuration -> Cluster Resources -> Change/Show Resources/Attributes for a Resource Group**.

Change/Show Resources/Attributes for a Resource Group

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

[TOP]	[Entry Fields]	
Resource Group Name	GEO_RG	
Node Relationship	cascading	
Site Relationship	cascading	
Participating Node Names / Default Node Priority	YOK01 YOK02 SEOUL	
Dynamic Node Priority	<input type="checkbox"/>	+
Service IP label	<input type="checkbox"/>	+
Filesystems (default is All)	[/fkm]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems/Directories to Export	<input type="checkbox"/>	+
Filesystems/Directories to NFS mount	<input type="checkbox"/>	+
Network For NFS Mount	<input type="checkbox"/>	+
Volume Groups	[fkmvg1]	+
Concurrent Volume groups	<input type="checkbox"/>	+
Raw Disk PVIDs	<input type="checkbox"/>	+
GeoMirror Devices	[fkmgmd1log fkmgmd1v1]	+
Connections Services	<input type="checkbox"/>	+
Fast Connect Services	<input type="checkbox"/>	+
Tape Resources	<input type="checkbox"/>	+
Application Servers	<input type="checkbox"/>	+
Communication Links	<input type="checkbox"/>	+
Primary Workload Manager Class	<input type="checkbox"/>	+
Secondary Workload Manager Class	<input type="checkbox"/>	+
Miscellaneous Data	<input type="checkbox"/>	
Automatically Import Volume Groups	false	+
Inactive Takeover Activated	false	+
Cascading Without Fallback Enabled	false	+
Disk Fencing Activated	false	+
Filesystems mounted before IP configured	false	+
[BOTTOM]		

Figure 4-9 GMD replicated resources

Synchronize resources

The HACMP resource synchronization is required to synchronize the configuration of the resource groups. Run **smit hacmp** and select **Cluster Configuration -> Cluster Resources -> Synchronize Cluster Resources**.

Note: The synchronization of the HACMP resources does not synchronize the configuration of the GMD devices. The GMD devices should be created by the HAGEO utilities in the HAGEO SMIT menu.

Section 4.4, “Clustering with HAGEO” on page 246 gives a complete procedure on configuring an HAGEO cluster.

Configuration by HAGEO

HAGEO still contains the GMD devices and the setup tools to configure the GMDs. It also contains tools necessary to verify, document, diagnose, monitor and maintain the GMD devices. See Figure 4-10 for the smit menu related to the HAGEO. Start the HAGEO menu by typing the **smit hageo** command.

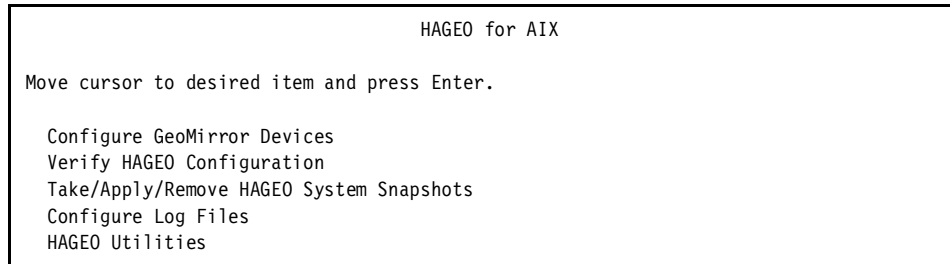


Figure 4-10 SMIT HAGEO menu

A full configuration example is provided in Section 4.4, “Clustering with HAGEO” on page 246.

4.3.2 TCP option for remote mirroring

TCP and UDP are the two most common transport layer communication protocols according to the ISO/OSI model.

UDP Simple protocol that connects the source logical port number and IP address to the destination logical port and IP address. The protocol does not handle any flow control. The application using this protocol has to handle the flow control.

TCP Complex protocol that connects the source logical port number and IP address to the destination logical port and IP address. TCP provides flow control. The difference to the UDP is that TCP provides a complex flow control, network buffering, and error handling, and the applications using TCP do not need to worry about flow control.

HAGEO 2.4 introduces a new option to use the TCP instead of the UDP protocol. The UDP protocol is still supported. The previous versions of HAGEO (2.3 and lower) used only UDP.

The TCP protocol has several benefits:

- Simple programming** TCP has all the features related to the flow control integrated. In UDP, the GeoMirror Device driver had to handle the flow control, which was not efficient.
- Better performance** Typically, the performance of the TCP protocol is better, because it allows network buffering, which improves the data flow.
- Better administration** The administration of the network options are more flexible via the provided network options.

The configuration of the GMD device driver to use TCP protocol is done through the SMIT interface, as shown in Figure 4-11. Run **smit hageo** and select **Configure GeoMirror Devices -> Configure Global GeoMirror Properties**.

Configure Global GeoMirror Properties

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

	[Entry Fields]	
GMDs for HACMP to start in parallel	[4]	#
Network Protocol	[TCP]	+#
Temporal Ordering Policy	[SYSTEM]	+#
Autoset Network Parameters	[Yes]	+#
Send/Receive Space Size (KBytes)	[256]	+#

Figure 4-11 SMIT HAGEO - Configure Global GeoMirror Properties

There are two performance tuning parameters related to the network options that are integrated into the HAGEO configuration when the TCP protocol is chosen:

- Autoset Network Parameters** If Yes, the TCP/IP network options set by HAGEO are used when configuring the GMD devices. If No, the default AIX network options are used.
- Send/Receive Space Size (KBs)** If the Autoset Network Parameters menu item is set to Yes, then this item sets the size of the network buffer used for send and receive buffers.

The configuration is done per cluster level, and the combination of UDP and TCP GMD device drivers are not possible. The configuration is done at one node of the cluster (any node) and synchronized to the other nodes using HAGEO synchronization for Global GeoMirror Properties, as described in “Cluster synchronization” on page 262.

4.3.3 Selection of temporal ordering policies

The temporal ordering policy reflects the requirement of several applications, such as databases, to order the writes to the devices in the order they are called from the application. This ordering may have performance impacts.

The temporal ordering policy controls how writes are done across GMDs. The temporal ordering policy allows you to configure VOLUME_GROUP, SYSTEM or NONE policies by the **geo_set_globals** command or through the SMIT menu, as shown in Figure 4-12. TCP supports all three types (SYSTEM, VOLUME GROUP, and NONE) of Temporal Ordering and UDP supports only two (SYSTEM and VOLUME GROUP).

```

                                Configure Global GeoMirror Properties

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

                                [Entry Fields]
GMDs for HACMP to start in parallel      []                #
Network Protocol                         [TCP]              +#
Temporal Ordering Policy                 [SYSTEM]          +#
Autoset Network Parameters               [Yes]              +#
Send/Receive Space Size (KBytes)         [256]              +#

+-----+
|                                     Temporal Ordering Policy                                     |
| Move cursor to desired item and press Enter.                                                                 |
|                                                                                                                 |
|  VOLUME_GROUP                                                                                                   |
|  SYSTEM                                                                                                       |
|  NONE                                                                                                         |
|                                                                                                                 |
|  Esc+1=Help      Esc+2=Refresh      Esc+3=Cancel                                                            |
Es| Esc+8=Image    Esc+0=Exit          Enter=Do                                                                |
Es| /=Find         n=Find Next                                                |
Es+-----+

```

Figure 4-12 SMIT HAGEO - Temporal Ordering Policy

SYSTEM ordering policy

If the SYSTEM policy is used, writes are performed as they were in previous releases. They are done in the same sequential order across all GMDs in the system, and one write must be completed before another begins. This ensures complete data integrity, but it can impact performance.

Volume_GROUP ordering policy

If the VOLUME_GROUP policy is used, writes are sequentially ordered on a volume group basis. This has less impact on performance, and data integrity should not be impacted in most cases, because typically related resources are organized in the same volume group.

NONE ordering policy

Choose NONE only if you are sure that your system has no dependency on the order in which writes occur on the GMDs in the system. An example of use of this policy is the implementation of independent file systems in HAGEO.

4.3.4 Support for 64-bit kernel environment

The device drivers with 64-bit device support must be aware of the 64-bit addressing methods. They use 64-bit structures, pointers, system calls, and so on. For our purposes, there are two versions of device drivers for the GMD available in HAGEO 2.4. Example 4-3 shows the device drivers stored in archives for both 32- and 64-bit systems. The AIX loader automatically detects the correct version of the device driver and loads only that version.

Example 4-3 Example of device drivers with 32- and 64-bit support

```
# cd /usr/lib/hageo/drivers
# ar -X any -tv gmd
rwxr-xr-x 30007/1      46334 May 21 13:53 2002 gmd32
rwxr-xr-x 30007/1      47822 May 21 13:53 2002 gmd64
# ar -X any -tv gmdpin
rwxr-xr-x 30007/1     108810 May 21 13:53 2002 gmdpin32
rwxr-xr-x 30007/1     114100 May 21 13:53 2002 gmdpin64
```

The gmd and the gmdpin files are kernel loadable libraries for pinned and pageable memory. They are loaded by the cfggmd system method, which needs to be aware of the 64-bit architecture. There is no need to have two versions loaded because of the AIX 32- and 64-bit compatibility for user programs. The cfggmd system method only loads the GMD device drivers in the kernel and exits. There are additional commands to manage the state of the device driver, which include **cfggmd**, **defgmd**, **startgmd**, **ucfggmd**, **chgmd**, **gmddown**, **stopgmd**, and **undefgmd**. See the HAGEO manual pages for the syntax and explanation of the commands using the **man** command. The user does not need to use these commands, as we recommend using the SMIT menus.

The 32- and 64-bit AIX is supported also in combinations of 32- and 64-bit versions on primary and backup sites. For example, if any node at the primary site runs on a 64-bit kernel, then nodes at the backup site may run on a 32 or 64-bit kernel, or vice versa.

4.4 Clustering with HAGEO

In this section, we assume that one remote node (and one remote site) will be added to our existing local cluster, as shown in Figure 4-13, which is configured with two nodes and one resource group.

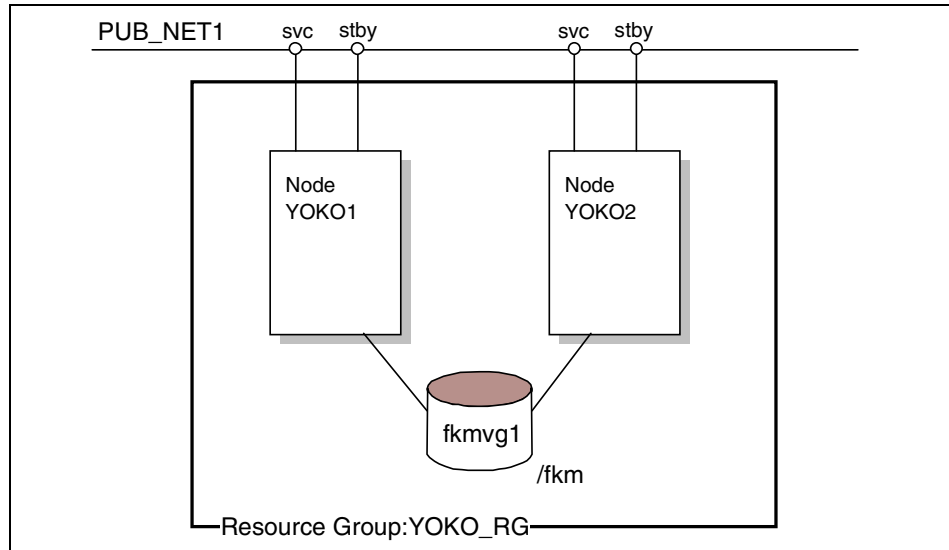


Figure 4-13 Our basic existing cluster

Figure 4-14 on page 247 shows our cluster configuration with a remote site using HAGEO.

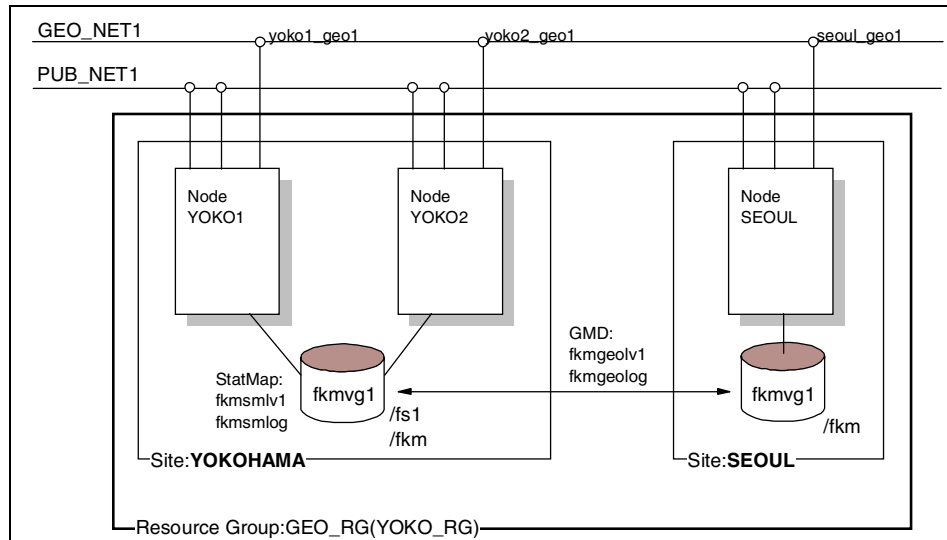


Figure 4-14 Sample cluster with a remote site

The following procedures describe configuring HAGEO 2.4 (using SMIT HACMP) for configuring a remote site and node.

4.4.1 Configure geographic topology

In order to add a remote site (and one node) to our cluster, we need to add one node, one network, and two site definitions to our topology.

1. Add a remote node definition.

We have to add one node as a remote node, but we can add this node the same way we add a local node using the SMIT HACMP menu, as shown in Figure 4-15.

Run **smit hacmp** and select **Cluster Configuration -> Cluster Topology -> Configure Nodes -> Add Cluster Nodes**.

Add Cluster Nodes

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

* Node Names

[Entry Fields]

[SEOUL]

Figure 4-15 Adding a remote node

2. Add a geographic network.

If we have already installed HAGEO, we can choose Geo_Primary for our network type to configure an additional network to our topology (see Figure 4-16).

Run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Configure Adapter -> Configure IP-based Interfaces / IP Labels -> Add Initial Interfaces**.

Add an IP-based Network

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

	[Entry Fields]	
* Network Name	[GEO_NET1]	
* Network Attribute	public	+
Network Type	[Geo_Primary]	+
Subnet(s)	[192.168.6.32/27]	+

Figure 4-16 Add an IP-based Network

Note: If you configure more than one Geo_Primary network, the networks will be able to send/receive the data with load balancing using the round robin technique.

We can then define additional adapters that are used for a geographic network on each node, as shown in Figure 4-17.

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[seoul_geol]	+
* Network Type	[Geo_Primary]	+
Network Name	[GEO_NET1]	+
* Network Attribute	[public]	+
* Interface Function	[service]	+
Interface IP Address	[]	
* Node Name	[SEOUL]	+
Netmask	[]	+

Figure 4-17 Add an additional network adapter

As shown in Figure 4-14 on page 247, in order to enable an IP Address Takeover (IPAT) between two sites, we define one or more boot address (or standby address) for IPAT on the remote node “SEOUL“ to the PUB_NET1.

Figure 4-18 shows a SMIT panel where we define the boot address on remote node SEOUL.

Add an Initial Interface

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

	[Entry Fields]	
* IP Label	[pub_seoul_boot]	+
Network Type	ether	
Network Name	PUB_NET1	
* Interface Function	[boot]	+
Interface IP Address	[]	
* Node Name	[SEOUL]	+
Netmask	[]	+

Figure 4-18 Add an boot adapter for IPAT between sites

Note: IP address takeover is not supported on a Geo_Primary network, but we can take over the IP address by using another network_type definition on a separate network (that is, PUB_NET1).

If we have a serial line network, we can add this connection to HACMP definitions, as shown in Figure 4-19, in order to avoid site isolation.

Run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Configure Adapter -> Configure Adapters on Non IP-based networks -> Add an Adapter**.

Add a Non IP-based Adapter

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

	[Entry Fields]	
* Adapter Label	[yokoi_tty]	
Network Type	rs232	
Network Name	GEO_NET2	
* Device Name	[/dev/tty1]	
* Node Name	[YOK01]	+

Figure 4-19 Add a Non IP-based Adapter

3. Add sites.

We need two site definitions. One is a remote site known as YOKOHAMA, as shown in Figure 4-20 on page 250, and the other is a local site known as SEOUL, as shown in Figure 4-21 on page 250. The site definition is also required for our local existing cluster.

Run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Configure Sites -> Add Site**.

The Dominance field defines which site will be halted when site isolation occurs. Site isolation happens when all the geographic networks are down, but at least one node at each site is still up. To prevent data divergence, the non-dominant site is halted.

To avoid site isolation, you can select the type of backup communication to use as a secondary geographic network. Select DBFS if you are using a telephone line. Select SGN if you have a serial line connection.

Add Site

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

[Entry Fields]

* Site Name	[YOKOHAMA]	
* Site Nodes	[YOK01 YOK02]	+
* Site Dominance	[yes]	+
* Site Backup Communications Type	[none]	+

Figure 4-20 Adding a local site

Add Site

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

[Entry Fields]

* Site Name	[SEOUL]	
* Site Nodes	[SEOUL]	+
* Site Dominance	[no]	+
* Site Backup Communications Type	[none]	+

Figure 4-21 Adding a remote site

4. Synchronize Cluster Topology.

Now we need to synchronize our resource group the traditional way. Run `smit hacmp` and select **Cluster Configuration -> Cluster Topology -> Synchronize Cluster Topology**.

4.4.2 Configure GeoMirror devices

If you plan to mirror a file system across the geographic network, you need to set up a geo-mirrored file system using GeoMirror Devices (GMD).

You need to create GMDs that will mirror the local data volume to the remote data volume. After you complete the procedure in this section, you will have the following items on the local and remote nodes:

- ▶ The /fkm file system
- ▶ The fkmgmdlv1 and fkmgmdlog GeoMirror devices
- ▶ The fkmlv1 and fkmlog logical volumes for the GeoMirror devices
- ▶ The fkmsmlv1 and fkmsmlog state maps for the GeoMirror devices
- ▶ Equivalent state maps and logical volumes on the remote nodes

To create GMDs that will mirror the local volume to the remote one, do the following:

1. Creating logical volumes.

After defining or using an existing volume group, create logical volumes for the file system, for example, fkmlv1 and fkmlog. You can use the **mk1v** command or SMIT to do this:

```
mk1v -y'fkmlog' -t'jfslog' -c'2' fkmvg1 1
mk1v -y'fkmlv1' -c'2' fkmvg1 32
```

You can create logical volumes for the state map devices for the GeoMirror file system device and for the GeoMirror file system log device, for example, /dev/rfkmsmlv1 and /dev/rfkmsmlog. The state map devices must be raw devices.

Use the following **mk1v** commands:

```
mk1v -y'fkmsmlv1' -c'2' fkmvg1 1
mk1v -y'fkmsmlog' -c'2' fkmvg1 1
```

2. Configuring GeoMirror devices.

To configure fkmgmdlv1 and fkmgmdlog as GeoMirror devices on the local node, run **smi t hageo** and select **Configure GeoMirror Devices -> Configure a GeoMirror Device -> Add a GeoMirror Device**, as shown in Figure 4-22 on page 252.

```

                                Add a GeoMirror Device

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

[TOP]                                [Entry Fields]
Device Name                          fkmgmdlog
* Minor Device Number                 [200] #
* State Map Logical Volume            [/dev/rfkmsmlog]
* Local Logical Volume                [/dev/rfkmllog]
* Device Mode                         mwc +
* Device Role                         none +
High Water Mark                       [128] #
Sync Concurrency Rate                 [32] +#
* Remote Node, LV, and Statemap       [SE0UL@/dev/rfkmllog@/dev/rfkmsmlog]
Remote Node, LV, and Statemap         []
Remote Node, LV, and Statemap         []
Remote Node, LV, and Statemap         []
Local Peer and State Map Device       [Y0K02@/dev/rfkmsmlog]
Local Peer and State Map Device       []
Local Peer and State Map Device       []
[BOTTOM]
```

Figure 4-22 Add a GeoMirror Device

Repeat the previous steps to create the GeoMirror file system log device, changing the names where appropriate to use fkmlog instead of fkm1v1.

3. Synchronize GeoMirror Devices.

When you finished configuring GMDs, you need to synchronize them across the cluster. Run **smit hageo** and select **Configure GeoMirror Devices -> Configure a GeoMirror Device -> Synchronize GeoMirror devices**, as shown in Figure 4-23.

```

                                Synchronize GeoMirror Devices

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

* GMD(s) to synchronize                [Entry Fields]
* Overwrite Existing GMD Definitions    [synchronize_all_gmds] +
                                         [yes] +
```

Figure 4-23 Synchronize GeoMirror Devices

4. Starting the GMDs.

You have to varyon all the volume groups that contain logical volumes to be used by the GeoMirror devices. Then start GeoMessage by running **smit hageo** and selecting **HAGEO Utilities -> GeoMessage Utilities -> Start GeoMessage**, as shown in Figure 4-24 on page 253.

```
GeoMessage Utilities

Move cursor to desired item and press Enter.

Show GeoMessage Statistics
Start GeoMessage
Stop GeoMessage
```

Figure 4-24 GeoMessage Utilities

Next, you can start the GeoMirror devices by running `smit hageo` and selecting **Configure GeoMirror Devices -> GeoMirror Utilities -> Start ALL GeoMirror Devices**, as shown in Figure 4-25.

```
Start ALL GeoMirror Devices

Move cursor to desired item and press Enter.

Start ALL GeoMirror Devices (Stopped -> Available)
Configure ALL GeoMirror Devices (Defined -> Stopped)
```

Figure 4-25 Start ALL GeoMirror Devices

Note: Do all of the procedures referenced in this step on each node in the mirror. Starting the device on a local node does not start the device on a remote node.

5. Create a file system.

On each node, add a definition of the file system by editing the `/etc/filesystems` file. Add an entry for the GeoMirror device file system, `/fkm` in this example, with the following stanzas:

```
/fkm:
dev = /dev/fkmgmdlvl
vfs = jfs
log = /dev/fkmgmdllog
mount = false
check = false
options = rw
account = false
```

Note: The device and log both point to the newly created GeoMirror devices and not the logical volumes.

On each node, make the directory `/fkm` a mount point:

```
mkdir /fkm
```

On the local node, put the fkm file system on the GeoMirror device:

```
mkfs -l /fkm /dev/fkm
```

6. Changing a Resource group definition.

Because we have two sites, we have to change our resource group definitions. We do not need additional resource groups, even if we add another site (see Figure 4-14 on page 247).

Optionally, you can also add an additional resource group that is separated from your existing resource group, as shown in Figure 4-26, but we do not discuss this configuration in this document.

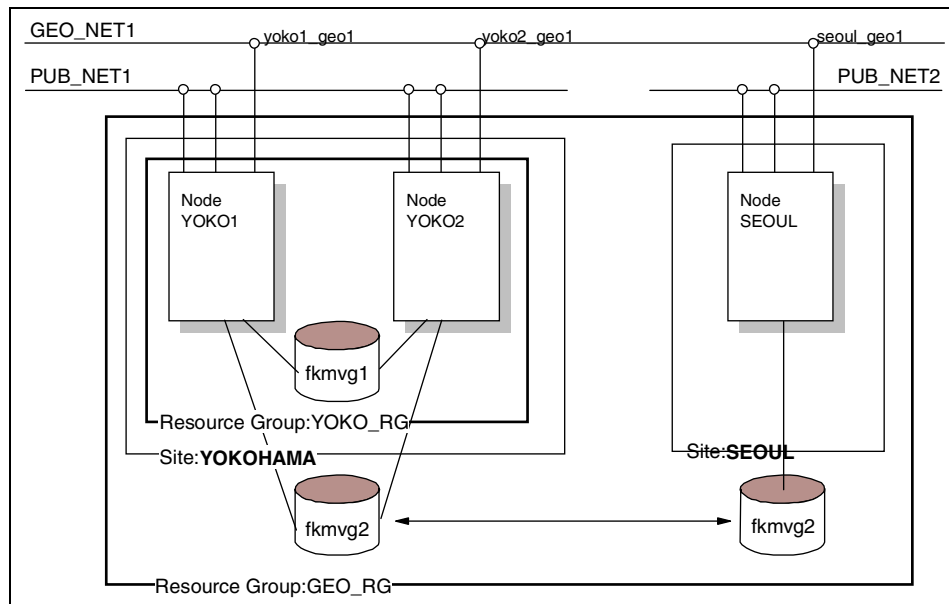


Figure 4-26 Another resource group configuration

We only have to add a remote node and change the site relationship, as shown in Figure 4-27 on page 255. The resource group name can also be changed if desired.

The site relationship is Ignore by default. You can choose Cascading, Rotating, or Concurrent.

Ignore (default)

Use this option if sites are not being used in the cluster.

Cascading

Select this option if you want replicated resources to be taken over by multiple sites in a prioritized manner. When a site fails, the active site with the highest priority acquires the resource. When the failed site

rejoins, the site with the highest priority acquires the resource.

Rotating

Select this option if you want replicated resources to be acquired by any site in the resource chain. When a site fails, the resource will be acquired by the highest priority standby site. When the failed node rejoins, the resource remains with its new owner.

Concurrent

Select this option if you want replicated resources to be accessible from all sites.

Change/Show a Resource Group

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

	[Entry Fields]	
Resource Group Name	YOKO_RG	
New Resource Group Name	[GEO_RG]	
Node Relationship	cascading	+
Site Relationship	cascading	+
Participating Node Names / Default Node Priority	[YOKO1 YOKO2 SEOUL]	+
Total Time To Process Events For This Group	180 Seconds	

Figure 4-27 Add a remote node to our resource group

7. Adding GMDs to a resource group.

In order to define these GMDs into our resource group, run `smit hacmp` and select **Cluster Configuration -> Cluster Resources -> Change/Show Resources/Attributes for a Resource Group**, as shown in Figure 4-28.

Change/Show Resources/Attributes for a Resource Group

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

[MORE...6]	[Entry Fields]	
Service IP label	<input type="checkbox"/>	+
Filesystems (default is All)	[/fkm]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems/Directories to Export	<input type="checkbox"/>	+
Filesystems/Directories to NFS mount	<input type="checkbox"/>	+
Network For NFS Mount	<input type="checkbox"/>	+
Volume Groups	<input type="checkbox"/>	+
Concurrent Volume groups	<input type="checkbox"/>	+
Raw Disk PVIDs	<input type="checkbox"/>	+
GeoMirror Devices	[fkmgmd1log fkmgmd1v1]	+
Connections Services	<input type="checkbox"/>	+
Fast Connect Services	<input type="checkbox"/>	+
Tape Resources	<input type="checkbox"/>	+
[MORE...11]		

Figure 4-28 Add GMD replicated resources

8. Synchronizing a resource group.

Now, we need to synchronize our resource group the traditional way. Run **smi t hacmp** and select **Cluster Configuration -> Cluster Resources -> Synchronize Cluster Resources**.

4.4.3 Managing the Geo Cluster

This section describes starting and stopping the Geo Cluster.

Starting and stopping geo-mirroring using HACMP

After your HAGEO software has been integrated with HACMP and the cluster has been completely and successfully configured with appropriate event scripts, you need only start and stop geo-mirroring using HACMP commands for routine administration.

Stopping HACMP on the node automatically stops the applications that use the GeoMirror devices, unmounts geo-mirrored file systems, stops the GeoMirror devices, stops the GeoMessage subsystem, and varies off the volume groups.

Restarting HACMP automatically restarts everything in the proper sequence and synchronizes the GeoMirror devices.

At times when you need to have more direct control, you must start and stop the individual HAGEO components in a specific order on each node in order to start and stop geo-mirroring.

Starting GeoMirror devices manually

1. Varyon all the volume groups that contain logical volumes to be used by the GeoMirror devices you are about to start.
2. Start the GeoMessage subsystem. In order for the device driver to be loaded, the GeoMessage subsystem, which is a Kernel Remote Procedure Call (KRPC) extension, must be loaded. Run **smi t hageo** and select **HAGEO Utilities -> GeoMessage Utilities -> Start GeoMessage**, as shown in Figure 4-24 on page 253.
3. Start each GeoMirror device. There are several stages to starting a GeoMirror device:
 - a. Change the state of the device from defined to stopped. When the first GeoMirror device is brought into the stopped state, the GeoMirror device driver is loaded. Run **smi t hageo** and select **Configure GeoMirror Devices -> GeoMirror Utilities -> Start ALL GeoMirror Devices**, as shown in Figure 4-25 on page 253.

- b. If starting the first part of a mirror, mark the remote mirror as down. When a GMD becomes available, it attempts to begin geo-mirroring. Before geo-mirroring can start, the GMD needs to be in the available state at both sites. If one becomes available while the other is not available, you have to wait for the geo-mirroring to time out before you can continue. Marking the remote GMD as down on the local node eliminates this unnecessary delay.
 - c. Change the device state to available. When the second of a GMD pair becomes available, it notifies the other nodes in the mirror. At that point, geo-mirroring begins.
4. Mount any file systems that reside on the GeoMirror devices.
 5. Start any applications that use GeoMirror devices.

You have to determine the order for starting the nodes, taking the following items into consideration:

- ▶ The local and remote parts of a GeoMirror device cannot be started simultaneously. Always start the local and remote nodes in the reverse order from which they were stopped. For example, if gmd1 was stopped first on node B and second on node A, start it first on node A and second on node B.
- ▶ For a specific GMD, wait for it to start completely, that is, when it is in the available state on either the local or remote side. Then start the GMD on the other node.
- ▶ You can start different GMDs simultaneously on different nodes. For example, you can start gmd1 on node A and gmd2 on node B at the same time.
- ▶ Start any node with staleness markings before the other. You can use the **gmd_show_state** command to determine if there are any staleness markings on either the local or remote node.

Stopping GeoMirror devices

You must stop any application that uses a GeoMirror device before stopping the GeoMirror devices themselves. You cannot stop a GeoMirror device if it is in use. Stop the GeoMessage subsystem last.

The procedure to stop geo-mirroring involves completing the following steps on each node:

1. Stop any applications that use a GeoMirror device.
2. Unmount any file systems that are mounted on GeoMirror devices.
3. Stop each GeoMirror device. Run **smit hageo** and select **Configure GeoMirror Devices -> GeoMirror Utilities -> Stop a GeoMirror Devices**.
4. Stop the GeoMessage subsystem. Run **smit hageo** and select **HAGEO Utilities -> GeoMessage Utilities -> Stop GeoMessage**.

5. Varyoff the volume groups for the GeoMirror devices.

4.4.4 Performance considerations

There are many factors that can affect HAGEO performance, such as network latency, AIX tunables, GMD mirroring modes, GMD tunables, disks, and applications. Refer to the “Understanding Latency” and “Planning for Performance” sections in Chapter 2, “Planning a HAGEO Cluster”, in *IBM HAGEO for AIX: Planning and Administration Guide*, SC23-1886 to learn more about these subjects.

4.4.5 Migration considerations

This section describes things to consider if you are migrating from previous versions of HAGEO.

Migration of resource group definitions

Previous HACMP releases do not support configuring non-concurrent disk resources at two sites using a single resource group. The workaround for this limitation is to create two or three HACMP resource groups for what could be logically considered to be one resource group, as shown in Figure 4-29 on page 259. There is an utility that can migrate the old style of resource groups to new style of resource group. The utility is `geo_migrate_rg`; use the `man` command for more details.

HACMP/ES 4.5 supports the disk resources that can be replicated at two sites, and it provides the capability to define all of the resources for a single HAGEO workload in a single HACMP resource group, as mentioned in Section 4.4.2, “Configure GeoMirror devices” on page 250. You are strongly encouraged to migrate your HAGEO-related resource groups to this new format.

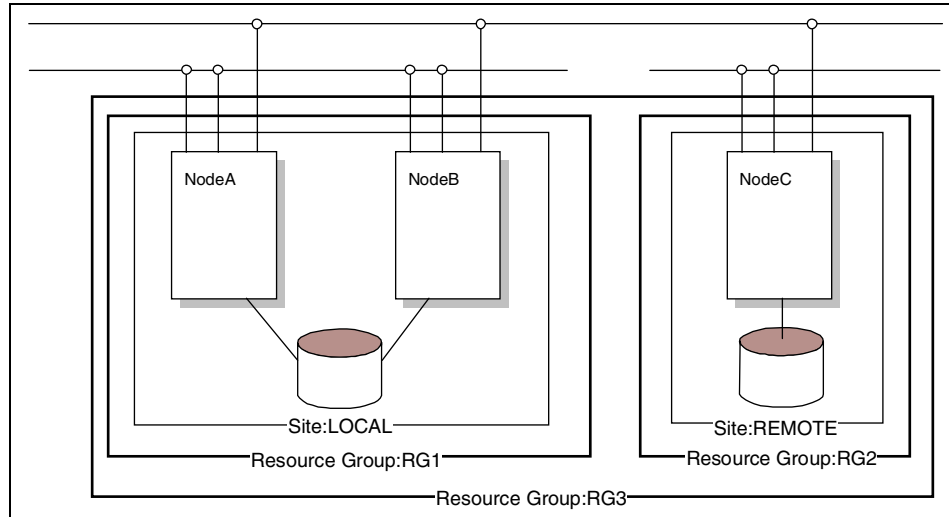


Figure 4-29 Resource group definition on previous HAGEO configuration

Removal of obsolete pre- and post-event definitions

Previous HAGEO releases provided HACMP pre- and post-event scripts to supplement the event processing that is done by HACMP. HACMP/ES 4.5 has an additional functionality that eliminates the need for these pre- and post-event scripts by using site-aware resource groups when HAGEO resources are defined.

After you have migrated all of your cluster nodes to HAGEO 2.4, and all of your HAGEO-related resource groups into site-aware resource groups, you should remove these pre- and post event script definitions from HACMP. Failure to remove these definitions will not harm your cluster. However, once they are not needed, execution of these scripts will unnecessarily slow down HACMP event processing. You can remove these pre- and post-event script definitions from HACMP by running the following command:

```
/usr/sbin/gmd/scripts/configure_events -d
```

You only need to run this command from one cluster node, as it will automatically synchronize its changes to the rest of the nodes in the cluster.

4.4.6 Troubleshooting

There are several aspects we have to consider when troubleshooting an HAGEO cluster. These aspects may arise from mistake of the configuration, network problems, and software errors.

We point to the following areas, because they should be checked first:

- ▶ Cluster planning
- ▶ Log files and status reports
- ▶ Cluster synchronization

Cluster planning

Always have the cluster diagram and cluster planning sheets easily available. HACMP and HAGEO provide snapshot tools to document the configuration.

The HACMP snapshot can be generated by the `clnsnapshot` utility or through SMIT (see Example 4-4).

Example 4-4 HACMP clnsnapshot

```
/usr/es/sbin/cluster/utilities/clnsnapshot -c -i -n'clsnap1' -d 'Cluster Snapshot 1'
clnsnapshot: Creating file /usr/es/sbin/cluster/snapshots/clsnap1.odm...
clnsnapshot: Creating file /usr/es/sbin/cluster/snapshots/clsnap1.info...
clnsnapshot: Executing clnsnapshotinfo command on node: SE0UL...
clnsnapshot: Executing clnsnapshotinfo command on node: YOK01...
clnsnapshot: Executing clnsnapshotinfo command on node: YOK02...
clnsnapshot: Succeeded creating Cluster Snapshot: clsnap1.
```

The HACMP snapshot contains all information related to HACMP, like topology, resources, network configuration, disk layout, and so on. It does not contain the GMD configuration of the HAGEO.

The HAGEO snapshot tool is used to back up the HAGEO configuration. This snapshot is a dump of the HAGEO objects from the ODM. This snapshot file is a text file, but may be difficult to read. We recommend that you double check the planning sheets and update them when changes are done. The snapshot may be used by service personnel to debug a problem.

Create the snapshot by running the **geo_snapshot** command, as shown in Example 4-5.

Example 4-5 HAGEO snapshot

```
# /usr/sbin/gmd/geo_snapshot -t -f snapshot1
Taking Geo snapshot...
```

Log files and status reports

The status of the cluster can be checked by verifying that the user can access the data. This check is the first check visible to the user.

The processes of the HAGEO cluster can be checked by the **ps** or the **lssrc** commands. Check to see if the cluster processes are running. The following subsystems should be running on each cluster node:

topsvcs	This is the topology services monitoring subsystem. It must run on every node where the cluster is active. The topsvcs subsystem starts the hats and hats_nim processes. Check the topsvcs subsystem by running the lssrc -ls topsvcs command, which gives a detailed diagnostic output.
grpsvcs	This is the group services group synchronization protocol. It must run on every node where the cluster is active. The grpsvcs subsystem starts the hags process. Check it by running the lssrc -ls grpsvcs command, which gives a detailed diagnostic output.
emsvcs	This is the event management subsystem that is used for event generation when a monitored condition occurs. It should run on every cluster node. Check it by running the lssrc -ls emsvcs command, which produces a detailed diagnostic output.
emaixos	This is a monitoring interface used to supply data from the operating system into the emsvcs subsystem. This subsystem is started by the emsvcs, if needed.
krpcmond	This is the GeoMessage control subsystem. It should run on every node where active GMDs are. It does not need to run on cluster nodes where GMDs are not active.
gmdmond	This is the GeoMirror control subsystem. It should run on every node where active GMDs are. It does not need to run on cluster nodes where GMDs are not active.
clstrmgrES	This is the cluster manager (the main cluster process). It starts the clstrmgr process, which has the main cluster logic. It should run on every cluster node. Check it by running the lssrc -ls clstrmgrES command, which gives a detailed diagnostic output. An even better diagnostic procedure is to check the clstrmgr log file (the /tmp/clstrmgr.debug file).
clsmuxpdES	This is the SNMP interface to the cluster manager. It has to run on every cluster node. Check it by running the ps -ef grep clsmuxpd command.
clinfoES	This is an user interface to the cluster manager detected events. We recommend that this subsystem be active,

because it initializes, for example, the client interface used by the clstat monitoring tool.

cllockdES

This subsystem is used only by the Oracle Parallel Server for lock management between database instances when accessing shared data in concurrent volume groups.

The next step is to check the log files:

/tmp/hacmp.out	The first place to check. It contains detailed tracing of the executed event scripts.
/tmp/clstrmgr.debug	Contains the clstrmgr tracing. It is a log file aimed at service personnel. However, it may give you some information.
/var/adm/clavan.log	Log of ordered events. The log file is easy to read.
/var/ha/log	Directory containing the tracing of the RSCT subsystems (topsvcs, grpsvcs, and emsvcs). The log files are aimed at service personnel.
/usr/es/adm/cluster.log	Log file generated by the BSD SYSLOG concerning HACMP.
/var/adm/hageo/kerngeo.log	Log generated by the BSD SYSLOG concerning HAGEO GeoMessage.
/tmp/krpc.out	GeoMessage log file.

Cluster synchronization

The synchronization of the HAGEO cluster has changed in HAGEO 2.4. The synchronization is performed by HACMP and HAGEO utilities in five steps, where each step can be performed separately:

1. HACMP topology synchronization
2. HACMP resource synchronization
3. HAGEO synchronization of Global GeoMirror properties
4. HAGEO Synchronization of GeoMirror device configuration
5. HAGEO Synchronization of GMD device content

HACMP topology synchronization

The HACMP topology synchronization first verifies the correct configuration of the networks and adapters and gives information about possible errors or warnings. If there are no errors detected, the synchronization procedure performs the synchronization.

Note: Do not ignore any error messages during the verification. Check the warning messages as they may point to a possible configuration mistake.

Start the HACMP topology synchronization by running the `clldare -t` command or use SMIT. This procedure is a standard HACMP procedure.

HACMP resource synchronization

The procedure of the HACMP resource synchronization is very similar to the HACMP topology synchronization. It can be performed when the cluster is using the DARE of HACMP or when the cluster is stopped. If you are troubleshooting a problem, we recommend you stop the cluster processes, reboot the system, and do not start HACMP after reboot.

The HACMP resource synchronization first verifies the correct configuration and, if there are no errors detected, it performs the synchronization. Similarly with the topology synchronization, do not ignore the error messages and check the warning messages.

Note: You cannot add, rename, or remove a site, node, network, or adapter in the geographic topology while GeoMessage is active on any nodes, because HAGEO does not support HACMP DARE capabilities.

HAGEO synchronization of Global GeoMirror properties

The configuration of the HAGEO Global GeoMirror properties should be synchronized by HAGEO utilities. Configurations related to the communication protocol (TCP and UDP) and to the configuration of temporary ordering policies (NONE, SYSTEM, and VOLUME_GROUP) are synchronized by this procedure.

Start the synchronization of the Global GeoMirror properties by running the `geo_sync_config` command, as shown in Example 4-6.

Example 4-6 geo_sync_config command

```
YOK01 # /usr/sbin/hageo/krpc/geo_sync_config
Synchronizing ODM class GE0globals to node SE0UL ...
... Succeeded.
Synchronizing ODM class GE0globals to node YOK02 ...
... Succeeded.
```

HAGEO synchronization of GeoMirror device configuration

The configuration of the GeoMirror devices (GMDs) is synchronized by the `sync_gmds` command. This command synchronizes the definition (not the

content) of the GMDs to other nodes. Example 4-7 shows the usage of this command.

Example 4-7 sync_gmds command

```
/usr/sbin/hageo/gmd/sync_gmds -l synchronize_all_gmds -d
Synchronizing GMD fkgmdlog definition to node SEOUL ...
Replacing GMD fkgmdlog definition on node SEOUL.
... Succeeded.

Synchronizing GMD fkgmdlog definition to node YOK02 ...
Replacing GMD fkgmdlog definition on node YOK02.
... Succeeded.

Synchronizing GMD fkgmdlv1 definition to node SEOUL ...
Replacing GMD fkgmdlv1 definition on node SEOUL.
... Succeeded.

Synchronizing GMD fkgmdlv1 definition to node YOK02 ...
Replacing GMD fkgmdlv1 definition on node YOK02.
... Succeeded.
```

Note: The `sync_gmd` command synchronizes only the definitions of the GMDs. It does not synchronizes the content of the GMDs.

Synchronize GMD content

The synchronization of the content of the GMDs requires special care, because a mistake may result in a loss of data. Pay attention when selecting the site with the valid data.

We recommend this procedure and assume that HACMP is stopped:

1. Select the site with valid data. Define the other site as having invalid data (see Example 4-8).

Example 4-8 Select valid data

Cluster layout:

Site Yokohama (primary site): YOK01, YOK02

Site Seoul (backup site): SEOUL

Yokohama has the valid data. (Seoul data is invalid.)

The synchronization will be done from Yokohama to Seoul

Site Seoul data will be overwritten by data from Yokohama.

2. Check that the GMD devices are stopped on all nodes using the **lsdev -Cc geo_mirror** command, as shown in Example 4-9. All devices should be in the Defined state.

Example 4-9 Check the GMD devices

```
# lsdev -Cc geo_mirror
fkmgmdlog Defined Geographic Mirror Device
fkmgmdlvl Defined Geographic Mirror Device
```

3. Go to node YOKO1. The Yokohama site was selected in Example 4-8 on page 264 as the site with valid data and YOKO1 is the primary node of the Yokohama site. Activate the volume group where the valid data resides and check it as shown in Example 4-10 using the **lsvg** and **varyonvg** commands.

Example 4-10 Activate volume group with data

```
# lsvg
rootvg
fkmvg1
# lsvg -o
rootvg
# varyonvg fkmvg1
# lsvg -o
rootvg
fkmvg1
# lsvg -l fkmvg1
fkmvg1:
LV NAME          TYPE      LPs   PPs   PVs  LV STATE      MOUNT POINT
fkmlog            jfslog    1     2     2   closed/syncd  N/A
fkm1v1            jfs       32    64     2   closed/syncd  N/A
fkmsmlv1          jfs       1     2     2   closed/syncd  N/A
fkmsmlog          jfs       1     2     2   closed/syncd  N/A
```

4. On node YOKO1, start the GeoMessage or check that it is running (see Example 4-11). Use the **cfgkrpc -ci** command to start the GeoMessage and the **krpcstat** command to check that it is running and responding.

Example 4-11 Start GeoMessage on YOKO1

```
# /usr/sbin/hageo/krpc/cfgkrpc -ci
cfgkrpc: loadit: extension=/usr/sbin/krpc/krpctcp kmid=30046516
Starting krpcmond
# /usr/sbin/krpc/krpcstat
```

Machines			KRPC Statistics				

	rpcs	KB	rp	rpcs	fragd	frags	% frag
SEOUL	0	0.0	0		0	0	0.00

	rpcs	KB	rp	rpcs	fragd	frags	% frag
	0	0.0	0		0	0	0.00
	rpcs	KB	rp	rpcs	fragd	frags	% frag
	0	0.0	0		0	0	0.00

Networks	KRPC Statistics					
	snd (snd/s)	rcv	0 KB (0 KB/s)		I KB	
GEO_NET1	0 (0.00)	0	0.0 (0.00)		0.0	
	resends	timeouts	failures	dups	other	
	0	0		0	0	0

- On YOKO1, configure the GMD devices using the command shown in Example 4-12.

Example 4-12 Configure GMD devices on YOKO1

```
# /etc/methods/cfggmd -l fkgmgmdlog
cfggmd: gmd, kmid = 30087536 loaded.
# /etc/methods/cfggmd -l fkgmgmdlvl
# lsdev -Cc geo_mirror
fkgmgmdlog Stopped Geographic Mirror Device
fkgmgmdlvl Stopped Geographic Mirror Device
```

- On YOKO1, inform the GeoMessage, that all remote nodes are down. Use the **gmddown** command, as shown in Example 4-13.

Example 4-13 Inform the GeoMessage on YOKO1 that SEOUL node is down

```
/etc/methods/gmddown -l fkgmgmdlog SEOUL
/etc/methods/gmddown -l fkgmgmdlvl SEOUL
```

- On YOKO1, start the GMDs using the **startgmd** command, as shown in Example 4-14.

Example 4-14 Start the GMDs on YOKO1

```
# /etc/methods/startgmd -l fkgmgmdlog
startgmd: No remote peers currently available for fkgmgmdlog.
startgmd: fkgmgmdlog started successfully.
# /etc/methods/startgmd -l fkgmgmdlvl
startgmd: No remote peers currently available for fkgmgmdlvl.
startgmd: fkgmgmdlvl started successfully.
# lsdev -Cc geo_mirror
fkgmgmdlog Available Geographic Mirror Device
fkgmgmdlvl Available Geographic Mirror Device
```

- On YOKO1, update the state map for the GMDs and make the state map “dirty” using the **gmddirty** command, and check its state with the

gmd_show_state command, as shown in Example 4-15. The “dirty” state map means, that the local data is valid, and the remote site has invalid (“dirty”) data. The state map value of 0xf represents a dirty state map.

Example 4-15 Dirty the statemaps

```
# /usr/sbin/hageo/gmd/gmddirty -l fkmgmdlog
gmd_update_state: Got fkmgmdlog state map on YOKO1.
# /usr/sbin/hageo/gmd/gmddirty -l fkmgmdlog
gmd_update_state: Got fkmgmdlog state map on YOKO1.
YOKO1 # /usr/sbin/hageo/gmd/gmd_show_state -l fkmgmdlog
gmd_show_state: Got fkmgmdlog state map on YOKO1.
Point of View: Node YOKO1
-----
Point of View GMD List:
-----
Name: fkmgmdlog
Status: AVAILABLE
gmd_show_state: Statemap is not clean.
State Map:
  Cell      Value
  -----
0    0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf 0xf
*
1048576
```

At this time, site YOKO1 is ready. Site YOKO2 does not have any active GMDs, because the GMDs are on node YOKO1.

- 9. Go to node SEOUL at site Seoul, which has invalid data. We want to delete this invalid data and copy the new valid data from the primary site (Yokohama) from its active node (YOKO1).

On node SEOUL, activate the volume group where the data resides and check it, as shown in Example 4-16, using the **lsvg** and **varyonvg** commands.

Example 4-16 Activate volume group on SEOUL

```
# lsvg
rootvg
fkmvg1
# lsvg -o
rootvg
# varyonvg fkmvg1
# lsvg -o
rootvg
fkmvg1
# lsvg -l fkmvg1
fkmvg1:
LV NAME                TYPE      LPs    PPs    PVs   LV STATE      MOUNT POINT
```

fkmllog	jfslog	1	2	2	closed/syncd	N/A
fkmlv1	jfs	32	64	2	closed/syncd	N/A
fkmsmlv1	jfs	1	2	2	closed/syncd	N/A
fkmsmlog	jfs	1	2	2	closed/syncd	N/A

10.On node SEOUL, start the GeoMessage or check that it is running (see Example 4-17). Use the **cfgkrpc -ci** command to start the GeoMessage and the **krpcstat** command to check that it is running and responding.

Example 4-17 Start GeoMessage on SEOUL

```
# /usr/sbin/hageo/krpc/cfgkrpc -ci
cfgkrpc: loadit: extension=/usr/sbin/krpc/krpctcp kmid=30046516
Starting krpcmond
# /usr/sbin/krpc/krpcstat
```

Machines	KRPC Statistics						

SEOUL	rp	KB	rp	rp	fragd	frags	% frag
	0	0.0	0		0	0	0.00
	rp	KB	rp	rp	fragd	frags	% frag
	0	0.0	0		0	0	0.00
	rp	KB	rp	rp	fragd	frags	% frag
	0	0.0	0		0	0	0.00
Networks	KRPC Statistics						

GEO_NET1	snd	(snd/s)	rcv	0 KB (0 KB/s)		I KB	
	0	(0.00)	0	0.0 (0.00)		0.0	
	resends	timeouts	failures	dups		other	
	0	0		0		0	
							0

11.On node SEOUL, configure the GMD devices using the command shown in Example 4-18.

Example 4-18 Configure GMDs on SEOUL

# /etc/methods/cfggmd -l fkmgmdlog	
cfggmd: gmd, kmid = 30087536 loaded.	
# /etc/methods/cfggmd -l fkmgmdlvl	
# lsdev -Cc geo_mirror	
fkmgmdlog	Stopped Geographic Mirror Device
fkmgmdlvl	Stopped Geographic Mirror Device

12.On node SEOUL, inform the GeoMessage that YOKO2 node at the remote site does not have active GMDs (YOKO2 is the standby node). If you do not do this step, you may get a system hang (see Example 4-19 on page 269). Do not run this command for YOKO1 node, because this node is the active node with our valid data.

Example 4-19 Inform the GeoMessage that YOKO2 is down

```
/etc/methods/gmddown -l fkgmgmdlog YOKO2
/etc/methods/gmddown -l fkgmgmdlvl YOKO2
```

13. On node SEOUL, check and “clear” the state maps, if necessary. Clear the state map for all GMDs you want to synchronize. Follow Example 4-20 and check the state maps using the **gmd_show_state** command. Write down the dirty regions, and then clean the state map using the **gmdclean** command.

Example 4-20 Clean the state maps on SEOUL

```
# /usr/sbin/hageo/gmd/gmd_show_state -l fkgmgmdlog
gmd_show_state: Got fkgmgmdlog state map on SEOUL.
Point of View: Node SEOUL
-----
Point of View GMD List:
-----
Name: fkgmgmdlog
Status: STOPPED
gmd_show_state: Statemap is not clean.
State Map:
  Cell      Value
  -----
      0  0xf 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0xf 0x0
     16  0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0
*
1048576
# /usr/sbin/hageo/gmd/gmd_show_state -l fkgmgmdlvl
...
# /usr/sbin/hageo/gmd/gmdclean -l fkgmgmdlog
gmd_update_state: Got fkgmgmdlog state map on SEOUL.
# /usr/sbin/hageo/gmd/gmdclean -l fkgmgmdlvl
gmd_update_state: Got fkgmgmdlvl state map on SEOUL.
# /usr/sbin/hageo/gmd/gmd_show_state -l fkgmgmdlog
gmd_show_state: Got fkgmgmdlog state map on SEOUL.

Point of View: Node SEOUL
-----

Point of View GMD List:
-----

Name: fkgmgmdlog
Status: STOPPED

State Map:

  Cell      Value
```

```

-----
0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0
*
1048576
# /usr/sbin/hageo/gmd/gmd_show_state -l fkmgmdlog
...

```

14. On node SEOUL, start each GMD using the **startgmd** command. The synchronization time may vary, depending on the amount of data that needs to be copied through the network. It may take a long time, because all the data needs to be copied through the network (see Example 4-21). During the synchronization, check the activity using the **iostat** command.

Example 4-21 Start GMDs on SEOUL

```

SEOUL # lsdev -Cc geo_mirror
fkmgmdlog Stopped Geographic Mirror Device
fkmgmdlvl Stopped Geographic Mirror Device
SEOUL # /etc/methods/startgmd -l fkmgmdlog
startgmd: fkmgmdlog started successfully.
SEOUL # /etc/methods/startgmd -l fkmgmdlvl
startgmd: fkmgmdlvl started successfully.
SEOUL # lsdev -Cc geo_mirror
fkmgmdlog Available Geographic Mirror Device
fkmgmdlvl Available Geographic Mirror Device

```

When the GMDs finish starting, the data is synchronized.

15. HAGEO expects that HACMP will start the GeoMessage and GMDs. Return the GMDs, GeoMessage, and volume groups to the original state, as HACMP expects it.

On every node of the cluster where we activated the resources (SEOUL and YOKO1), deactivate the resources, as shown in Example 4-22, using the **stopgmd**, **ucfggmd**, **cfgkrpc** and **varyoffvg** commands. The **-A** parameter means *all* GMDs. It is also possible to simply reboot the nodes.

Example 4-22 Deactivate resources

```

SEOUL # /etc/methods/stopgmd -A
SEOUL # /etc/methods/ucfggmd -A
ucfggmd: gmd, kmid = 30624368 unloaded.
SEOUL # /usr/sbin/hageo/krpc/cfgkrpc -u
initkrpctcp: kmid= 30440916 unloaded
SEOUL # varyoffvg fkmvg1

```

16. Start the HACMP cluster in the standard way, as shown in Section 4.4.3, “Managing the Geo Cluster” on page 256.

4.4.7 Maintenance considerations

You might have parts of your HACMP cluster integrated with HAGEO, while other parts might be separate. For example, you might have local networks, adapters, and applications used by the nodes at a site that are not included in the HAGEO configuration. Changes to those components do not affect HAGEO.

On the other hand, all the HAGEO components must also be included in the HACMP configuration. You must be aware of the application servers, resources, and resource groups that are relevant to the HAGEO configuration, even though they are not part of the GeoMirror or geographic topology configuration process. They form part of the geographic mirroring process and the disaster recovery process.

The application server must be directed to write to a GeoMirror device so that the data will be mirrored across the geography. The Cluster Manager must know about all the HAGEO networks, nodes, and adapters in order to monitor them. The volume groups of the GeoMirror devices must be included in properly configured resource groups so that the Cluster Manager handles them correctly, in case of node or site failure.

After making any change to the environment, be sure to review and test the configuration.

Changing geographic topology

Additional nodes, networks, and adapters, or changes to them, must also be defined to HACMP.

If you need to change geographic topology, you can change it through the HACMP topology menus.

For more information on how to change the HACMP cluster configuration, see Chapter 24 of *HACMP for AIX Enhanced Scalability Installation and Administration Guide*, SC23-4306.

Changing GeoMirror devices

To change the configuration of a GeoMirror device, do the following:

1. Stop the applications that use the GeoMirror device you want to change.
2. Stop and unconfigure the device. This puts the device in the defined state.

Note: A GeoMirror device must be in the defined state for changes to be made to the definition. This is not possible with the device in the Stopped or Available state.

3. Change the attributes of the GeoMirror device. Make any required changes on one node in the configuration.
 - a. Changing the attributes of a GeoMirror device
 - i. Click on **Configure a GeoMirror Device** -> **Change/Show a GeoMirror Device** and press Enter. The Select Device Name menu is displayed, as shown in Figure 4-22 on page 252.
 - ii. The Device Name field is initialized with the selection you made. You can change any of the other entry fields.
 - iii. Click on the Synchronize GeoMirror Devices option and press Enter to distribute these changes to the remote and local peer in the GeoMirror device.
 - b. Changing the size of a GeoMirror file system device

Unlike a regular file system, you can use the **geo_chfs** command to resize a GeoMirror file system device.
4. Synchronize the GeoMirror definitions to the other nodes.
5. Verify that the new configuration is properly written on all affected nodes.
6. Start the device. This reconfigures and starts the device.

Abbreviations and acronyms

AAAT	Application Availability Analysis Tool	RSCT	IBM RS/6000 Cluster Technology
ARP	Address Resolution Protocol	SNA	Systems Network Architecture
CS/AIX	Communications Server for AIX	SSA	Serial Storage Architecture
CWS	Control Workstation	TMR	Tivoli Management Region
DARE	Dynamic Reconfiguration	UID	User Identification
DASD	Direct Access Storage Device	WAN	Wide Area Network
DLC	Data Link Control	WLM	Workload Manager
GID	Group Identification		
GPFS	General Parallel File System		
HA/ES	IBM High Availability Cluster Multi-Processing for AIX Enhanced Scalability		
HACMP	IBM High Availability Cluster Multi-Processing for AIX		
HACMP/ES	IBM High Availability Cluster Multi-Processing for AIX Enhanced Scalability		
HACWS	Highly Available Control Workstation		
HAS	IBM High Availability Cluster Multi-Processing for AIX Classic		
HATivoli	High Availability Tivoli		
HAWAN			
HWAT	Hardware Address Takeover		
IBM	International Business Machines Corporation		
IPAT	IP Address Takeover		
ITSO	International Technical Support Organization		
LAN	Local Area Network		
ODM	Object Database Manager		
OLPW	Online Planning Worksheets		

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 276.

- ▶ *AIX 5L Workload Manager (WLM)*, SG24-5977
- ▶ *Disaster Recovery Using HAGEO and GeoRM*, SG24-2018
- ▶ *HACMP Enhanced Scalability*, SG24-2081
- ▶ *IBM @server pSeries 690 System Handbook*, SG24-7040
- ▶ *RS/6000 SP/Cluster: New Enhancements in PSSP 3.4*, SG24-6604
- ▶ *RS/6000 SP High Availability Infrastructure*, SG24-4838

Other resources

These publications are also relevant as further information sources:

- ▶ *Configuring p690 in an @server Cluster 1600*, REDP0187
- ▶ *HACWS: An HACMP Application for IBM @server Cluster 1600*, REDP0303
- ▶ *Configuring Highly Available p690 Clusters using HACMP 4.5*, REDP0218
- ▶ *HACMP for AIX 4.5 Administration Guide*, SC23-4279
- ▶ *HACMP for AIX 4.5 Enhanced Scalability Installation & Administration Guide*, SC23-4306
- ▶ *HACMP for AIX 4.5 Installation Guide*, SG23-4278
- ▶ *HACMP for AIX 4.5 Planning Guide*, SC23-4277
- ▶ *IBM HAGEO for AIX: Planning and Administration Guide*, SC23-1886
- ▶ *PSSP for AIX: Installation and Migration Guide*, GA22-7347
- ▶ *RS/6000 & pSeries PCI Adapter Placement Reference*, SA38-0538
- ▶ *RS/6000 SP: Planning, Vol. 1, Hardware and Physical Environment*, GA22-7280

- ▶ *RSCT: Event Management Programming Guide and Reference*, SA22-7354

Referenced Web sites

These Web sites are also relevant as further information sources:

- ▶ *AIX 5L Version 5.1 System Management Guide: Operating System and Devices*
<http://www-1.ibm.com/servers/aix/library/>
- ▶ AIX Support Web site
<http://techsupport.services.ibm.com/>
- ▶ ESS official support Web site
<http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/2105>
- ▶ *IBM @server pSeries 690 Availability Best Practices* whitepaper
http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/p690_avai1.html
- ▶ *PSSP Read this First*
http://www.ibm.com/servers/eserver/pseries/library/sp_books/pssp.html

How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

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.

Index

Symbols

#1245 160
/spdata 165
/usr/sbin/hacws/hacws_verify 213

Numerics

64-bit clinfo 5

A

ACQUIRE 19
acquire_service_addr 164
Active control workstation 159
adapter failure 219
adapter failures 204
Adapters 76
AIX I/O pacing 69
AIX Workload Manager 47
aliasing 167
Application Availability Analysis tool 28
application failover 65
Application server 2, 162
application server 187
Application servers 77
ARP 42
attached servers 159
authentication 67

B

Backup control workstation 159
Backup geographic network 228
Bandwidth 229
Basic cluster definitions 64
Boot IP labels 188
boot/service 69

C

Cascading 42, 49, 254
cl_highest_free_mem 12
cl_highest_idle_cpu 12
cl_lowest_disk_busy 12
clavan.log 28

clcommlinkd 36
clcycle 26
Clinfo 24
cllsif 40
clsmuxpd 10
clsmuxpdES 261
clstat 24
clstat.cgi 24
clstrmgrES 261
Cluster 1600 64, 168
cluster event 208
Cluster ID 75
cluster ID 187
cluster manager 21, 168
Cluster network 69
cluster node 64
Cluster nodes 64
Cluster resources 227
Cluster security 67
Cluster topology 227
Cluster with p690 servers 161
clverify 54
Command
 /spdata/sys1/hacws/rc.syspar_aliases -add 163
 /spdata/sys1/hacws/rc.syspar_aliases -delete 163
 /usr/bin/lppchk 182
 /usr/kerberos/etc/kdb_edit 177
 /usr/lpp/ssp/kerberos/bin/kadmin 181
 /usr/lpp/ssp/kerberos/bin/ksrvutil add 178
 /usr/lpp/ssp/kerberos/bin/ksrvutil list 181
 /usr/sbin/hacws/spcw_apps -di 163
 /usr/sbin/hacws/spcw_apps -ua 163
chauthent 176
chdev 82
cl_configure_persistent_address 34
cl_opsconfig 5, 7
clcycle 26
clevsummary 26
clfindres 27, 215
clinfo 10
clsnapshot 260
clstat 24

- clstat.cgi 24
- clstop 221
- clverify 47
- geo_snapshot 260
- geo_sync_config 263
- gmd_show_state 257, 267
- gmdsizing 230
- ifconfig 34
- install_cw 175
- k4list 180
- lsattr -El ssar 202
- lsauthent 176
- lsdev 82
- lsdev -C | grep tmssa 202
- lsdev -Cc geo_mirror 265
- push-kprop 180
- SDRGetObjects Frame 215
- setup_server 163
- startgmd 266
- sync_gmds 263
- topsvcs 238
- varyon 256
- x25status 36
- command
 - gmdclean 269
- Commnad
 - cl_resgrp 219
- Concurrent 255
- concurrent 62
- Concurrent Mode 5, 10
- Concurrent Resource Manage 85
- configuration 72
- control workstation
 - active 162
 - inactive 162
- Create 7
- crontab 165
- Custom cluster events 162
- Custom definitions 227
- CWS 168

D

- DARE 13, 48
- Data availability 160
- Data space design 226
- DBFS 230, 250
- DCE 222
- Dedicated 70

- Design 63
- Dial Back Fail Safe 230
- disaster 223
- Discover network topology 188
- Dominance 250
- drawer 74
- Dynamic node priority 5, 10
- Dynamic Reconfiguration 13

E

- EEH 62
- emsvcs 239
- enhancements
 - administrative 4
 - application support 4
 - device support 4
 - network 4
 - usability 4
- error daemon 21
- errornotify 20
- ESS 71
- event scripts 70
- Event Summaries 27
- extended error handling 62

F

- Failover 2
- Fallover 2
- fast events 22
- FC 62
- Fiber channel 44
- File
 - /dev/rmt0 47
 - /etc/inittab 34, 210
 - /etc/krb.conf 178
 - /etc/krb-servtab 180
 - /etc/rc.backup_cw_alias 210
 - /etc/rc.sp script 163
 - /tftpboot/tuning.cust 210
 - /usr/es/sbin/cluster/etc/harc.net 34
 - /usr/es/sbin/cluster/samples/worksheets 5
 - /usr/es/sbin/cluster/etc 26
 - /usr/sbin/cluster/netmon.cf 204
 - /usr/sbin/hacws/hacws_verify 213
 - cl_event_summaries 26
 - clevsum.txt 27
 - clstrmgr.debug 13
 - cluster.conf 7

- hacmp.out 15
- ipaliase.conf 33
- Fileset
 - hageo.gmdsizing 229
 - ssp.hacws 161, 164
- Filesystems 14
- firmware 65
- frame supervisor card 158
- Full system partition 62
- full system partition mode 64

G

- geographic mirroring 228
- geographic network 248
- geographic topology 247
- GeoMirror Device 240
- GeoMirror device 250
- GeoPrimary 228, 231
- GeoSecondary 230
- gmd 245
- GMD device 240
- gmdsizing 230
- GPFS 57
- group services 68

H

- HACMP xix, 1, 64, 223
- HACMP Classic 14
- HACMP monitoring 69
- HACMP/ES 2
- HACWS 157, 168
- HACWS Support 159
- hacws_post_event 163–164
- hacws_pre_event 163–164
- HAES 15
- HAGEO xix, 223
 - 64-bit device support 245
 - Configuration examples 233
 - maintenance 271
 - Migration 258
 - Software requirements 232
- hardmon 159, 163
- Hardware Address Takeover 42, 88
- Hardware Management Console 61
- HAS 2
- HATivoli 59
- heartbeat 69
- High Availability 1

- High Availability Cluster Mult-Processing 1
- High Availability Control Workstation 157
- High Availability Geographic Cluster for AIX xix, 223
- highly available cluster 62
- HMC 61, 161, 165, 168, 192
- HMC failure 68
- HMC network 69
- Hot Standby 159
- HTML 24
- HWAT 88

I

- I/O pacing 210
- IBM C for AIX 161
- IBM Enterprise Storage Server 71
- IBM eServer Cluster 1600 3
- IBM eServer pSeries Cluster 1600 158
- IBM RS/6000 Cluster Technology 2
- Inactive control workstation 159
- Install 182
- installation 72
- IP Address Takeover 88
- IP address takeover 231
- IP aliases 70
- IP aliasing 32, 39, 69
- IP labels 183
- IPAT 32, 88, 231

J

- JOB_TYPE 19

K

- Kerberos 67, 166
- krb-srvtab 181

L

- lab environment 72
- Latency 229
- license 4
- Link stability 229
- Local HACMP networks 228
- local high availability 227
- Log file 260
- logical partitioning 61
- Logical Volume Manager 70
- LPAR 61, 64

LVM 70
LVM_SA_QUORCLOSE 20–21

M

Managing Geo cluster 256
Migration 16
monitor 219

N

network 226
Network Adapter failures 20
Network design 166
Network Discovery 32, 38
Network Install Manager 72
Network planning 226
Network tuning 176
network_down 69
network_type 249
new features 4
Node 2
Node configuration 236
node_down 12
NODE_PRIORITY_POLICY 13
node_up_complete 164
Nodes 75
NONE ordering 245
non-IP network 201
NVRAM 62

O

ODM Class 20
OLPW 5
Online planning worksheets 5
Operating system failure 64

P

p690 159
p690 attached servers 159
p690 server 63
pager notification 5, 8
parallel 15
PCI host bridge 62
Persistent 188
persistent alias 167, 199
Persistent IP label 33, 188
Persistent node IP alias 32
Planning 158, 225

point-to-point 224
post_event 164
post-events 211
pre_event 164
pre-defined 10
Prerequisites 3, 84
primary 227
Primary control workstation 159
primary control workstation 168
Primary geographic network 228
private network 161, 168, 237
processes 176
profile 74
PSSP 70, 161
PTF 175
Public network 228
public network 69
push-kprop 180

R

RAS 63
Redbooks Web site 276
 Contact us xvii
release_service_addr 164
Requirements 3
requirements 161
Resource Group 2, 19
Resource group 162
resource group 187, 235
Resource group parallel processing 5
Resource groups 77
restrictions 158, 222
rg_move 12
Rotating 255
rotating 165, 168
Rotating or Concurrent 49
rotating resource 158
RS/6000 Cluster Technology 85
RS-232 control lines 174
RSCT 2, 71, 232
RSCT Integration 68

S

Scenario1 87
scenarios 72
SCSI 45
SCSI reservation 70
SDR 159, 162

- serial connection 160
- Serial network 69
- serial network 81, 203
- servers 61
- Service IP address 162
- Service IP labels 188
- SGN 250
- Shared 70
- single point of failure 79, 158
- Site configuration 237
- Site planning 226
- slow events 22
- SP attached servers 161
- SP services 183
- SP Switch 70
- SP Switch network 67, 70
- SP Switch2 70
- SPLAN 168
- SPOF 79
- SSA 62
- standby 69
- Standby IP labels 188
- Starting GMD 252
- State maps 251
- Supported hardware 3
- swap_adapter 33, 41
- symmetric 160
- Synchronize topology 238
- sysctld 163
- syspar 163
- syspar_ctrl 163
- System management 67
- system management network 70
- SYSTEM ordering policy 244

T

- tape 45
- target mode 202
- target mode SCSI 69, 160
- target mode SSA 69, 160
- TCP 242
- TCP/IP 224
- Tivoli 56
- tmssa 79
- Topology 3, 187
- topology 200
- Topology Services 2
- topology services 68

- Troubleshooting 259
- tty 159
- Tune 175

U

- UDP 242
- user defined 10

V

- VisualAge C++ 161
- volume group 77
- volume group loss 5
- Volume_GROUP ordering 245
- VSM utility 58

W

- WAN 226
- WAN communication link 32
- Web link
 - http
 - //192.168.6.35/cgi-bin/clstat.cgi 24
- Web server 24
- WLM 47
- WLM class 53

X

- X25 Communication link failure 20

Y

- Y cable 175
- Y-cable 160



Configuring Highly Available Clusters Using HACMP 4.5



Redbooks

Configuring Highly Available Clusters Using HACMP 4.5

Examples including p690 configuration

Configuring HACWS for Cluster 1600

Explains HACMP and HAGEO integration

The objective of this IBM Redbook is to provide how-to technical information about configuring highly available clusters using HACMP Version 4.5, with a special focus on IBM *@server* pSeries 690 model 681 servers. It describes, in detail, the installation, customization, and configuration procedures of the new LPAR supported servers with HACMP for high availability. As an case study application for HACMP 4.5, this redbook investigates and explains the procedures for configuring an Highly Available Control Workstation (HACWS) in an Cluster 1600 configuration. HAGEO has been integrated to exploit the HACMP 4.5 features, and a special focus on HAGEO 2.4 is described in Chapter 4 of this redbook. Prior knowledge of HACMP will be useful for easy understanding of this document. This document provides various example scenarios and configurations and demonstrates high availability clustering using HACMP 4.5.

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