

IBM PowerHA SystemMirror

7.1.2 Enterprise Edition

for AIX

Unleashes IBM PowerHA
SystemMirror

Describes new features and
functionalities

Includes implementation
scenarios



Dino Quintero
Venkatesh Balappa
Shawn Bodily
Daniel de Souza Casali
Johann-Georg Dengel
Murali Dhandapani
Machi Hoshino
Jes Kiran
Kunal Langer
Paulo Sergio Lemes Queiroz
Matt Radford
Andrei Socoliuc
Neng Kuan Tu

Redbooks



International Technical Support Organization

**IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for
AIX**

May 2013

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

First Edition (May 2013)

This edition applies to IBM AIX 7.1 TL02 SP01, IBM PowerHA SystemMirror for AIX 7.1 TL02 SP01, for migration IBM PowerHA SystemMirror for AIX 6.1 SP09, IBM Systems Director 6.1.3, IBM PowerHA Director Plug-in 7.1.2.1.

For the 4 nodes/2 sites, 2 nodes and one storage system each site: Nodes: AIX 6.1TL8 SP1, PowerHA 7.1.2 SP1; multipath software - AIX MPIO included in AIX. Storage: DS8800(2141-951); Mcode versions (DS8K4: 7.6.30.160; DS8K6:7.6.31.136).

© Copyright International Business Machines Corporation 2013. 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

Notices	ix
Trademarksx
Preface	xi
The team who wrote this bookxi
Now you can become a published author, too!	xiii
Comments welcome	xiii
Stay connected to IBM Redbooks	xiv
Part 1. Introduction	1
 Chapter 1. Concepts and overview of the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition	3
1.1 High availability and disaster recovery	4
1.1.1 Disaster recovery considerations	5
1.2 Local failover, campus style, and extended distance clusters	8
1.3 Storage replication and the new HyperSwap feature	11
 Chapter 2. Differences between IBM PowerHA SystemMirror Enterprise Edition version 6.1 and version 7.1.2	15
2.1 Architecture changes	16
2.1.1 Cluster Aware AIX	16
2.1.2 ODM updates	17
2.2 New features and functionalities	18
2.2.1 Stretched and linked clusters	19
2.2.2 Split/merge handling overview	21
2.2.3 HyperSwap overview	21
2.2.4 IPv6 support	22
2.2.5 PowerHA Smart Assists	22
2.2.6 IBM Systems Director	24
2.2.7 New or changed cluster administration tools	25
2.2.8 Features at a glance added since PowerHA 6.1	25
2.3 Limitations	26
2.3.1 Deprecated features	26
2.3.2 Restrictions	27
2.3.3 Unsupported changes	27
2.4 Hardware and software prerequisites	27
 Chapter 3. Planning	29
3.1 Infrastructure considerations and support	30
3.1.1 Hostname and node name	30
3.1.2 Network considerations	30
3.1.3 Storage and SAN considerations	40
3.1.4 Cluster repository disk	47
3.1.5 Tie breaker disk	48
3.2 Hardware and software requirements	50
3.2.1 Hardware requirements for the AIX solution	50
3.2.2 Software requirements for PowerHA 7.1.2	50
3.2.3 Capacity on Demand (CoD) support	55

Part 2. Campus style disaster recovery (stretched clusters)	57
Chapter 4. Implementing DS8800 HyperSwap 59	
4.1 Overview of HyperSwap	60
4.2 Traditional disaster recovery	60
4.3 HyperSwap failover	61
4.4 The architecture	64
4.4.1 In-band storage management	64
4.4.2 AIX support for HyperSwap	66
4.4.3 AIX view of HyperSwap disks	68
4.5 HyperSwap functionalities	68
4.6 Hardware and software requirements	69
4.7 Limitations	69
4.8 Considerations of implementation	69
4.9 Test environment of a HyperSwap cluster	69
4.10 Initial disk configuration	72
4.11 Preparation of a HyperSwap-capable disk	73
4.11.1 Selecting disks to be mirrored for HyperSwap disks	73
4.11.2 Establishing the connection path from hdiskA to hdiskB	77
4.11.3 Establishing the connection path from hdiskB to hdiskA	78
4.11.4 Establishing volume pairs in one direction	79
4.11.5 Enabling HyperSwap on all nodes	79
4.12 Initial configuration for the PowerHA cluster	81
4.13 Configuring the HyperSwap disk in PowerHA	83
4.13.1 Creating a volume group with HyperSwap capable disks	84
4.13.2 Defining the storage systems and their site association	86
4.13.3 Setting up the mirror group you want these HyperSwaps to be in	88
4.13.4 Create a resource group with site policy	94
4.13.5 Add a mirror group and a volume group into the resource group	96
4.13.6 Verification and synchronization	98
4.14 Swap demonstration for HyperSwap	98
4.14.1 Planned swap for user mirror group	99
4.14.2 Planned swap for a cluster repository mirror group	103
4.14.3 Planned swap for a system mirror group	107
4.14.4 Unplanned swap for pure write applications	110
4.14.5 Unplanned swap for pure read applications	115
4.14.6 Summary for HyperSwap performance	116
4.15 Oracle standalone database in a PowerHA HyperSwap environment	117
4.15.1 The environment description	117
4.15.2 HyperSwap disk configuration	119
4.15.3 Verifying AIX and the PowerHA cluster environment for the database	121
4.15.4 The Oracle database environment	124
4.15.5 Configure the database in PowerHA using the Oracle Smart Assist	127
4.15.6 Testing the cluster environment using PowerHA HyperSwap	138
Chapter 5. Cross-site LVM mirroring with IBM PowerHA SystemMirror 7.1.2 Standard Edition 151	
5.1 Cross-site LVM mirroring overview	152
5.1.1 Requirements	152
5.1.2 Planning considerations	153
5.2 Testing the environment	155
5.3 Configuring the cross-site LVM cluster	156
5.3.1 Configuring the cluster topology	157
5.3.2 Configuring the cluster resources	158

5.4	Test scenarios.....	169
5.4.1	Test case 1: Both nodes down site failure	169
5.4.2	Test case 2: Rolling node failures for site outage.....	171
5.4.3	Test case 3: Outage of a storage subsystem	172
5.4.4	Test case 4: Rolling disaster.....	176
5.5	Maintaining cross-site LVM.....	179
5.5.1	Test environment overview.....	179
5.5.2	Storage administration scenarios	181
Part 3.	Extended disaster recovery (linked clusters)	193
Chapter 6. Configuring PowerHA SystemMirror Enterprise Edition linked cluster with SVC replication..... 195		
6.1	Overview of SVC management.....	196
6.2	Planning and prerequisites overview	199
6.2.1	Planning	199
6.2.2	Prerequisite overview	201
6.3	Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with SVC remote copy 204	204
6.3.1	Scenario overview.....	204
6.3.2	Environment configuration for PowerHA.....	205
6.3.3	Configuring the cluster basic definitions	212
6.3.4	Configuring SVC replicated resources	213
6.3.5	Configuring resources and resource groups.....	218
6.4	Storage management with SVC replicated resources	221
6.4.1	Removing a disk from a PowerHA environment with SVC remote copy.....	221
6.4.2	Adding a volume to the PowerHA cluster with SVC remote copy.....	223
6.5	Testing the resource group move	226
6.5.1	Resource group move within a site.....	226
6.5.2	Resource group move across sites.....	227
6.5.3	Node failure.....	228
6.5.4	Site failure.....	229
6.6	Testing storage failure on SVC and V7000 mixed environment.....	230
6.6.1	SVC-V7000 mixed environment	230
6.6.2	Local site SAN failure	233
6.6.3	PPRC link failure.....	237
6.7	PowerHA SVC PPRC commands.....	252
6.8	IBM PowerHA SystemMirror and SVC global mirror failover demonstration.....	252
Chapter 7. Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replication 253		
7.1	Introduction to XIV remote mirroring	254
7.2	Planning PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV remote mirroring 257	257
7.2.1	Operational considerations	257
7.2.2	Software requirements	258
7.3	Preliminary steps for XIV storage configuration	258
7.4	Installing the XIV CLI.....	261
7.5	Creating an XIV replicated cluster using SMIT and XIV GUI	262
7.5.1	Our scenario description	263
7.5.2	Configuration of the XIV remote mirroring.....	265
7.5.3	Installation and configuration of the PowerHA SystemMirror software	273
7.5.4	Testing the cluster.....	291
7.6	Creating an XIV replicated cluster via clmgr and XIV CLI.....	302
7.6.1	Implementation overview	302
7.6.2	Configuring XIV remote mirror couplings using xcli	304

7.6.3 Configuring the cluster using clmgr.....	313
7.7 Administrating the cluster with XIV replicated resources	326
7.7.1 Adding volumes to the XIV remote mirror replicated resource	326
7.7.2 Remove volumes from a XIV remote mirror replicated resource	331
7.7.3 Adding a file system to a volume group	334
7.7.4 Change the recovery action	338
7.8 IBM PowerHA SystemMirror demonstration with XIV replication	340
Part 4. System administration, monitoring, maintenance, and management.....	341
Chapter 8. Migrating to PowerHA SystemMirror 7.1.2 Enterprise Edition.....	343
8.1 Migration planning.....	344
8.1.1 Requirements	344
8.1.2 Migration options.....	346
8.1.3 Offline method.....	347
8.1.4 Rolling method	347
8.1.5 Snapshot method	348
8.2 Clmigcheck explained.....	349
8.3 Migration scenarios.....	351
8.3.1 Rolling from PowerHA SystemMirror 6.1 Enterprise Edition.....	351
8.3.2 Snapshot from IBM PowerHA SystemMirror 6.1 Enterprise Edition.....	364
8.3.3 Offline migration from PowerHA 6.1 Enterprise Edition with IPv6 configuration.	373
8.4 Solving migration errors	386
8.4.1 Node name not set to hostname.....	386
8.4.2 Stuck in migration	389
8.4.3 Non-IP network not deleted after migration completed.....	389
8.4.4 Clodmgmt not found.....	392
Chapter 9. PowerHA 7.1.2 for IBM Systems Director plug-in enhancements	393
9.1 Installing the IBM Systems Director environment	394
9.1.1 Installing IBM Systems Director	394
9.1.2 Installing IBM Systems Director plug-in for PowerHA SystemMirror.....	397
9.1.3 Installing PowerHA SystemMirror plug-in fixes.....	400
9.1.4 Installing PowerHA SystemMirror agent fixes.....	407
9.2 Environment configuration	414
9.3 Configuring IBM PowerHA SystemMirror 7.1.2 Enterprise Edition using IBM Systems Director.....	415
9.3.1 Discovering servers.....	415
9.3.2 Creating the cluster.....	416
9.3.3 Verifying and synchronizing the cluster.....	423
9.3.4 Configuring resources.....	424
9.3.5 Configuring replicated mirror groups.....	426
9.3.6 Create a resource group.....	429
9.4 Administering a cluster	437
9.4.1 Bringing cluster services online and offline.....	437
9.4.2 Cluster snapshot.....	440
9.4.3 Moving cluster resource groups	445
Chapter 10. Cluster partition management.....	449
10.1 Cluster partitioning	450
10.2 Methods to avoid cluster partitioning.....	451
10.3 Planning to avoid cluster partitioning	452
10.4 Detailed behavior of cluster partitioning	453
10.4.1 Test environment overview.....	453

10.4.2 Configuring the split and merge policy	455
10.4.3 Split policy: None, merge policy: Majority	456
10.4.4 Split policy: TieBreaker, merge policy: TieBreaker	464
Part 5. Appendices	477
Appendix A. Configuring IBM PowerHA SystemMirror with IPv6.....	479
Configuring IBM PowerHA SystemMirror with IPv6	480
Enabling IPv6 on AIX	480
Configuration tips for PowerHA with IPv6.....	483
CAA command enhancements	491
Migrating steps from IPv4 to IPv6.....	492
Appendix B. DNS change for the IBM Systems Director environment with PoweHA	495
Configuring DNS on an IBM Systems Director server	496
Configuring the application server scripts	496
Testing the application	496
Related publications	499
IBM Redbooks	499
Other publications	499
Online resources	499
Help from IBM	500

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 grant 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites 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.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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 illustrate 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.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

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

AIX®	IBM®	Rational®
DB2®	Jazz™	Redbooks®
developerWorks®	Lotus®	Redbooks (logo)  ®
Domino®	MQSeries®	Storwize®
DS6000™	Orchestrate®	System p®
DS8000®	Power Systems™	System Storage®
Enterprise Storage Server®	POWER6®	SystemMirror®
eServer™	PowerHA®	Tivoli®
GPFS™	PowerVM®	XIV®
HACMP™	POWER®	
HyperSwap®	pSeries®	

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redbooks® publication helps strengthen high availability solutions for IBM Power Systems™ with IBM PowerHA® SystemMirror® Enterprise Edition (hardware, software, and tools) with a well-defined and documented deployment model within an IBM Power Systems environment, offering clients a planned foundation for a dynamic, highly available infrastructure for their enterprise applications.

This book addresses topics to leverage the strengths of IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX on IBM Power Systems to solve client application high availability challenges, and maximize system availability, and management. The book examines the tools, utilities, documentation, and other resources available to help the IBM technical teams provide solutions and support for IBM high availability solutions with IBM PowerHA in an IBM Power Systems environment.

This book is targeted toward technical professionals (consultants, technical support staff, IT Architects, and IT Specialists) responsible for providing high availability solutions and support with IBM Power Systems and IBM PowerHA.

The team who wrote this book

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

Dino Quintero is a complex solutions project leader and IBM Senior Certified IT Specialist with the ITSO in Poughkeepsie, NY. His areas of knowledge include enterprise continuous availability, enterprise systems management, system virtualization, technical computing, and clustering solutions. He is currently an Open Group Distinguished IT Specialist. Dino holds a Master of Computing Information Systems degree and a Bachelor of Science degree in Computer Science from Marist College.

Venkatesh Balappa is a System Software Engineer in IBM India. He is currently working at IBM ISTL, which is the India Systems and Technology Lab, where he is working on GFW (Power firmware), Hardware Management Console (HMC) and PowerHA System Mirror Service Pack Test. He has more than five years of experience in system testing on various IBM POWER® system features such as PowerHA SystemMirror, Dynamic Logical Partitioning (DLPAR), Live Partition Mobility, FSP reset/reload, LDAP, Virtual IO Server and Hardware Management Console (HMC). His area of expertise includes PowerHA SystemMirror, DLPAR, FSP reset/reload, Virtual IO Server and HMC. Venkatesh holds a Bachelor of Engineering degree in Computer Science from VTU Belgaum, India.

Shawn Bodily is a Certified Consulting IT Specialist for Advanced Technical Skills Americas in Dallas, Texas. He has worked for IBM for 14 years and has 16 years of AIX® experience, with 12 years specializing in High Availability Cluster Multi-Processing (HACMP™). He is double Advanced Technical Expert certified. Shawn has written and presented extensively on high availability and storage and has coauthored six other Redbooks publications.

Daniel de Souza Casali is a Lab Services Senior Consultant in Brazil. He has 10 years of experience in the UNIX systems field. Daniel holds an Engineering degree in Physics from the Federal University of São Carlos (UFSCar). His areas of expertise include UNIX, SAN networks, IBM Disk Subsystems and clustering solutions.

Johann-Georg Dengel is a systems engineer and member of the AIX Engineering Team in the Dynamic Platform Services of T-Systems International GmbH, Germany. He has more than 22 years of experience in various IT fields, including operating systems, SAN-engineering and administration, database administration and HADR solutions. Johann-Georg is double advanced Technical Expert certified. This is his first experience co-authoring a Redbooks publication.

Murali Dhandapani is a Certified IT Specialist in Systems Management in IBM India. He is working for the IBM India Software Lab Operations team, where he is a technical lead for IBM Rational® Jazz™ products infrastructure, high availability, and disaster recovery deployment. He has 10 years of experience. His areas of expertise include Linux, AIX, IBM POWER Virtualization, PowerHA SystemMirror, System Management, and Rational tools. Murali has a Master of Computer Science degree. He is an IBM developerWorks® Contributing Author, IBM Certified Specialist in System p® administration and an IBM eServer™ Certified Systems Expert - pSeries® High Availability Cluster Multi-Processing (IBM HACMP).

Machi Hoshino is an IT Specialist for Power Systems and AIX in IBM Japan Systems Engineering, which provides part of the GTS function in Japan. He has four years of experience in AIX and the PowerHA field. He holds a Bachelor in Liberal Arts degree with a major in Information Science from the Christian University. His areas of expertise include Power Systems, AIX, PowerHA SystemMirror for AIX, PowerVM®, GPFS™, and Linux on Power.

Jes Kiran is an Advisory Software Engineer in IBM India. He has more than 11 years of experience and is a lead developer for PowerHA SystemMirror. He is involved with various features of PowerHA. He is an expert in system programming and kernel areas. His areas of expertise include PowerHA, RSCT, CAA, SSP, VIOS, and HMC. Jes holds a Bachelor of Technology degree from Kakatiya University.

Kunal Langer is a Technical Consultant for Power Systems and AIX in STG Lab Based Services in IBM India. He has more than six years of experience in Power Systems. He is a certified technical expert for AIX and PowerHA. His areas of expertise include AIX, PowerHA, PowerVM and IBM System Storage®. He has co-authored previous IBM PowerHA SystemMirror Redbooks publications. Kunal holds a Bachelor of Computer Science degree from VTU, India.

Paulo Sergio Lemes Queiroz is an IT Specialist with IBM Brazil. He has 10 years of experience in UNIX and Linux, ranging from systems design and development to systems support. His areas of expertise include AIX Power Systems, AIX, GPFS, RHCS, KVM, and Linux. Paulo is a Certified Advanced Technical Expert for Power Systems with AIX (CATE) and a Red Hat Certified Engineer (RHCE).

Matt Radford is a UNIX support specialist in the United Kingdom. He has worked in IBM for 15 years and has five years of experience in AIX and High-Availability Cluster Multi-Processing (HACMP). He holds a degree in Information Technology from the University of Glamorgan. Matt has co-authored two other IBM Redbooks publications.

Andrei Socoliuc is a Certified IT Specialist in Systems and Infrastructure, working in IBM Global Technologies Services Romania. He has more than 12 years of experience in IT infrastructure. Andrei holds a Master's degree in Computer Science from the Polytechnical University of Bucharest. He is a Certified Advanced Technical Expert IBM System p and a Certified Tivoli® Storage Manager specialist. He has worked extensively on HACMP and Disaster Recovery projects and is also a coauthor of several HACMP IBM Redbooks publications.

Neng Kuan Tu is a creative AIX consultant. He started his IBM career as an AIX developer in 1989 in Austin, Texas. His innovative programming ideas, including a patent in binary to

decimal conversion, improved AIX performance significantly. He has been involved in AIX technical support to China's telecommunication industry for almost seven years. After successfully supporting IBM's business in China, he joined Huawei Technologies to build Huawei's IT infrastructure for R&D. In 2011, he rejoined IBM and supports IBM worldwide business. His expertise includes performance tuning, problem determination, server virtualization, and high availability solutions. Neng holds a PhD in Physics from the Ohio State University.

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

Octavian Lascu, Richard Conway, David Bennin, Alfred Schwab (editor)
International Technical Support Organization, Poughkeepsie Center

PI Ganesh, Michael Coffey, Patrick Buah, Ravi A. Shankar, Paul Moyer, Tom Weaver, Bob Maher, Ken Fleck, Nick Fernholz, Mark Gurevich, Stephen Tovcimak, Steven Finnes, Gary Lowther, Kam Lee, Bob McNamara
IBM USA

Lakshmipriya Kanduru, Seema K.T., Mamatha Medarametla, Saritha Garipelly
IBM India

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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 books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:
ibm.com/redbooks
- ▶ Send your comments in an email to:
redbooks@us.ibm.com

- ▶ Mail your comments to:
IBM Corporation, International Technical Support Organization
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:
<http://www.redbooks.ibm.com/rss.html>



Part 1

Introduction

In the first part of this IBM Redbooks publication, we provide an overview and concepts of the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX. We also discuss the differences between the current and previous versions of this product. Moreover, we include a planning chapter that shows the prerequisites needed before implementing the high availability solution.

In this part we introduce the following chapters:

- ▶ Chapter 1, “Concepts and overview of the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition” on page 3.
- ▶ Chapter 2, “Differences between IBM PowerHA SystemMirror Enterprise Edition version 6.1 and version 7.1.2” on page 15.
- ▶ Chapter 3, “Planning” on page 29.



Concepts and overview of the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition

In this book, we discuss the new features, differences from previous versions, and give some example scenarios to help you get used to and exploit PowerHA SystemMirror 7.1.2 Enterprise Edition.

In this chapter we cover the following topics:

- ▶ High availability and disaster recovery
- ▶ Local failover, campus style, and extended distance clusters
- ▶ Storage replication and the new HyperSwap feature

1.1 High availability and disaster recovery

Datacenter and services availability are some of the most important topics for IT infrastructure, and each day draws more attention. Since not only natural disasters affect normal operations, but human errors and terrorist acts may affect business continuity and even with fully redundant infrastructure, services are vulnerable to such disasters.

Replication of data between sites is a good way to minimize business disruption since backup restores can take too long to meet business requirements, or equipment may be damaged depending on disaster extent and not available for restoring data. Recovery options typically range in cost from the least expensive having a longer time for recovery to the most expensive providing the shortest recovery time and the closest to having zero data loss. A fully manual failover would normally require many specialists to coordinate and perform all the needed steps to bring the services up to another site, and even with a good disaster recovery plan can take longer than needed by business requirements. High availability software is intended to minimize downtime of services by automating recovery actions when failures are detected on the various elements of the infrastructure.

PowerHA SystemMirror 7.1.2 Standard Edition helps you automate node failures and application events and provide high availability. Even in this version it is possible to have two site clusters using LVM Cross Site mirror as a solution. Testing was done and documented in Chapter 5, “Cross-site LVM mirroring with IBM PowerHA SystemMirror 7.1.2 Standard Edition” on page 151. The PowerHA Enterprise Edition, in addition, will help you automate recovery actions on storage failures for selected storage, controlling storage replication between sites and enabling recoveries for entire site failures, thus ensuring copies are in a consistent state to make the failover, enabling you to build a disaster recovery solution.

If you already have PowerHA SystemMirror 6.1 Enterprise Edition, and have implemented it as part of your disaster recovery solution, refer to Chapter 2, “Differences between IBM PowerHA SystemMirror Enterprise Edition version 6.1 and version 7.1.2” on page 15 for the differences between the versions. Then if you are planning on migrating PowerHA SystemMirror Enterprise Edition, we show three different migration scenarios to get your current version to 7.1.2 (snapshot, offline, and rolling migration). These migration scenarios guide you to getting your existing cluster migrated to the latest level using the option that better suits your business requirements. All the testing scenarios are documented in Chapter 8, “Migrating to PowerHA SystemMirror 7.1.2 Enterprise Edition” on page 343.

For both PowerHA Standard and PowerHA Enterprise Edition, the IBM Systems Director server can be enabled to manage clusters via its integrated GUI just by installing the PowerHA plug-in that has been enhanced to support the disaster recovery enablement features added in PowerHA SystemMirror 7.1.2 Enterprise Edition.

The IBM Systems Director with the PowerHA plug-in gives the ability to:

- ▶ Discover the existing PowerHA SystemMirror clusters.
- ▶ Collect information and a variety of reports about the state and configuration of applications and clusters.
- ▶ Receive live and dynamic status updates for clusters, sites, nodes, and resource groups.
- ▶ Having single sign-on capability which allows full access to all clusters with only one user ID and password.
- ▶ Access and search log files to display a summary page that you can view for the overall status of all known clusters and resource groups.
- ▶ Create clusters and add resource groups with wizards.

- ▶ Apply updates to the PowerHA SystemMirror Agent using the Director Update Manager.

Chapter 9, “PowerHA 7.1.2 for IBM Systems Director plug-in enhancements” on page 393 was developed to discuss how to install and use the IBM Systems Director PowerHA plug-in to manage a cluster.

1.1.1 Disaster recovery considerations

The idea of a fast failover in the event of a problem, or the recovery time objective (RTO), is important but should not be the only area of focus. Ultimately, the consistency of the data and whether the solution meets the recovery point objective (RPO) is what make the design worth the investment. You should not enter a disaster recovery planning session and expect to truly achieve the five nines of availability by solely implementing a clustering solution. Table 1-1 outlines the calculations for the uptime criteria.

Table 1-1 Five nines of availability

Uptime	Uptime	Maximum downtime per year
Five nines	99.999%	5 minutes 35 seconds
Four nines	99.99%	52 minutes 33 seconds
Three nines	99.9%	8 hours 46 minutes
Two nines	99.0%	87 hours 36 minutes
One nine	90.0%	36 days 12 hours

There are some considerations when planning a disaster recovery solution to achieve an accurate RTO. For example, is the time for planned maintenance accounted for? If so, is that time deducted from the maximum downtime figures for the year? While PowerHA has the ability to nondisruptively update the cluster, you have to consider the impact of other service interruptions in the environment such as upgrades involving the applications, the AIX operating system, and the system firmware, which often require the services to go offline a certain amount of time.

The IBM PowerHA SystemMirror 7.1.2 Enterprise Edition solution can provide a valuable proposition for reliably orchestrating the acquisition and release of cluster resources from one site to another. It can also provide quick failover in the event of an outage or natural disaster. Figure 1-1 on page 6 shows the various tiers of disaster recovery solutions and how the PowerHA Enterprise Edition is considered a tier 7 recovery solution. Solutions in the alternate tiers can all be used to back up data and move it to a remote location, but they lack the automation that the PowerHA Enterprise Edition provides. Looking over the recovery time axis, you can see how meeting an RTO of under four hours could be achieved with the implementation of an automated multisite clustering solution.

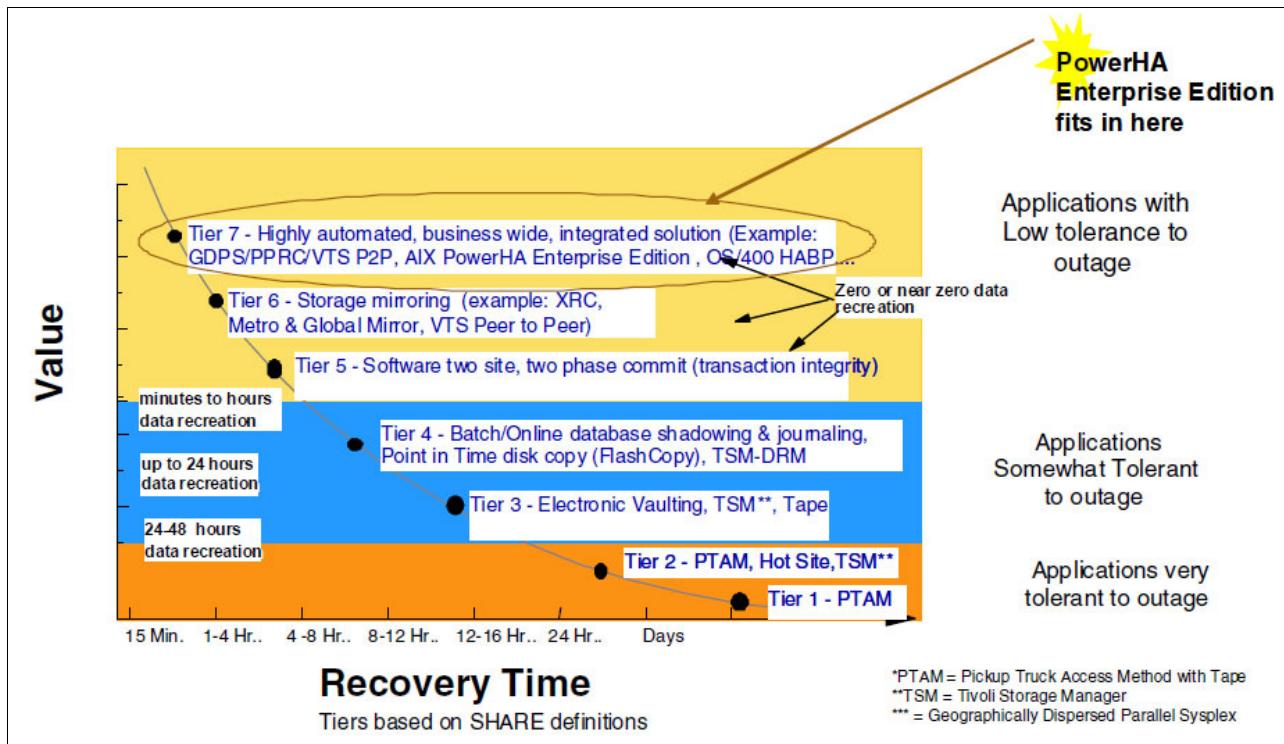


Figure 1-1 Tiers of disaster recovery solutions - IBM PowerHA SystemMirror 7.1.2 Enterprise Edition

Table 1-2 describes the tiers of disaster recovery in more detail, and outlines POWER Systems solutions available in each tier for disaster recovery.

Table 1-2 Disaster recovery tier

DR planning model reference	Power Systems solutions	100% recovery of application data possible?	Automatic detection of site failure?	Facility locations supported
Tier 7 Zero data loss. Recovery up to application restart in minutes	Cross Site LVM GLVM Metro Mirror Global Mirror HyperSwap®	Yes - minus data in transit at time of disaster	Yes - Failover and fallback automated	All PowerHA Standard Edition suffices for a campus-style DR solution
Tier 6 Two-site two-phase commit. Recovery time varies from minutes to hours	Oracle or DB2® log shipping to a remote standby database Oracle or DB2 active data replication to a remote database DB2 HADR solution	No - does not include active log data Yes Yes	No No No	All All All

DR planning model reference	Power Systems solutions	100% recovery of application data possible?	Automatic detection of site failure?	Facility locations supported
Tier 5 Continuous electronic vaulting of backup data between active sites. Active data management at each site is provided.	TSM with copy pool duplexing between sites, and active TSM servers at each site	No Recovery in days or weeks Must restore from backups	No	All
Tier 4 Electronic vaulting of critical backup data to a hot site. The hot site is not activated until a disaster occurs.	TSM with copy pool duplexing to the hot site, and TSM server at active site only	No Recovery in days or weeks	No	N/A
Tier 3 Off-site vaulting with a hot site. Backup data is transported to the hot site manually. The hot site is staffed and equipped but not active until a disaster occurs.	TSM with multiple local storage pools on disk and tape at active site	No Recovery in days or weeks	No	N/A
Tier 2 Off-site vaulting of backup data by courier. A third-party vendor collects the data at regular intervals and stores it in its facility. When a disaster occurs: a) A hot site must be prepared. b) Backup data must be transported.	TSM with multiple local storage pools on disk and tape at active site	No Recovery in days or weeks	No	N/A
Tier 1 Pickup Truck Access Method with tape	Tape-backup based solution	No	No	No

DR planning model reference	Power Systems solutions	100% recovery of application data possible?	Automatic detection of site failure?	Facility locations supported
Tier 0 No disaster recovery plan or protection	Local backup solution may be in place, but no offsite data storage	No DR recovery - site and data lost	N/A	N/A

High availability and disaster recovery is a balance between recovery time requirements and cost. Various external studies are available that cover dollar loss estimates for every bit of downtime experienced as a result of service disruptions and unexpected outages. Thus, key decisions must be made to determine what parts of the business are important and must remain online in order to continue business operations.

Beyond the need for secondary servers, storage, and infrastructure to support the replication bandwidth between two sites, there are items that can easily be overlooked, such as:

- ▶ Where does the staff go in the event of a disaster?
- ▶ What if the technical staff managing the environment is unavailable?
- ▶ Are there facilities to accommodate the remaining staff, including desks, phones, printers, desktop PCs, and so on?
- ▶ Is there a documented disaster recovery plan that can be followed by non-technical staff if necessary?

Chapter 3, “Planning” on page 29 describes the considerations to get a better understanding of the infrastructure needs, and how to plan the implementation of PowerHA to improve availability.

1.2 Local failover, campus style, and extended distance clusters

Clients implementing local PowerHA clusters for high availability often mistake failover functions between a pair of machines within the same server room from the same failover functions between machines located at dispersed sites. Although from a cluster standpoint the graceful release and reacquire functions are effectively the same, in a scenario with remote sites you always have a higher risk of data loss in the event of a hard failure and will likely take longer than a local failover since more steps are needed; for example, storage replication management. Thus, it is often more appropriate to use a local high availability cluster as the building block for a disaster recovery solution where two local nodes are paired within a site and at least one remote node is added into the cluster as part of a second site definition.

If the business requirements of a client state that the service level agreements (SLAs) claim to maintain a highly available environment even after a site failure, the client needs a cluster with four nodes, two in each site. This is the configuration that we used in most of the testing environments discussed in this book.

For two-site failover, two different situations can be used to accomplish replication needs:

- ▶ The sites can be near enough to have shared LUNs in the same Storage Area Network (SAN), as seen in Figure 1-2 on page 9.

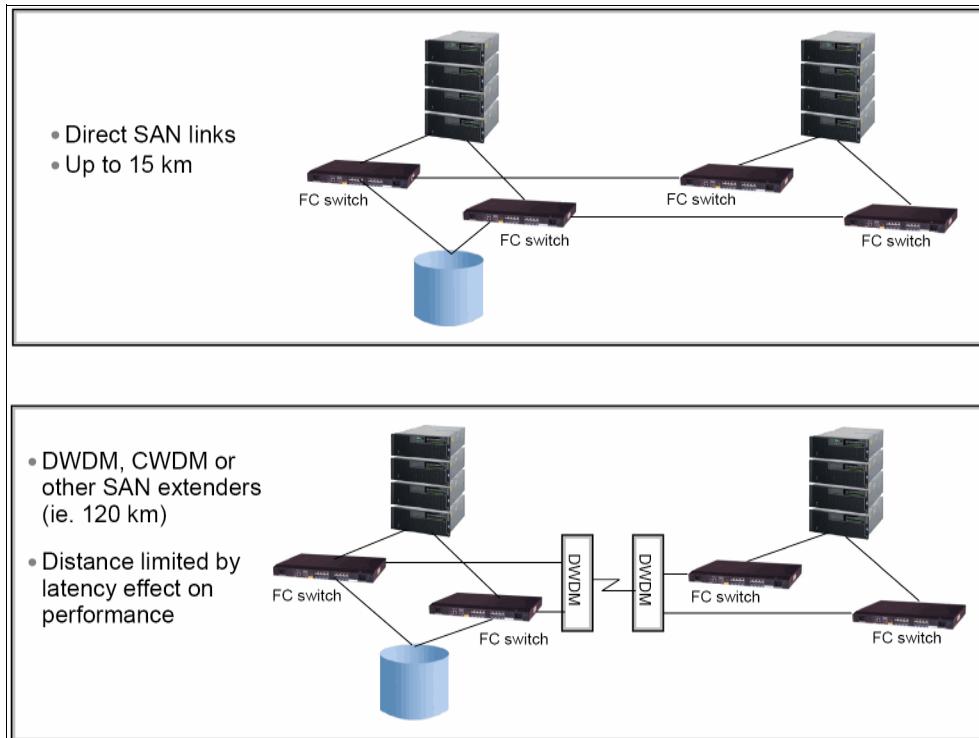


Figure 1-2 Extended SAN between sites

These environments often present a variety of options for configuring the cluster, including a cross-site LVM mirrored configuration, using disk-based metro-mirroring, or scenarios using San Volume Controller (SVC) Vdisk mirroring with a split I/O group between two sites, which can be implemented since the SVC firmware 5.1 release. Being inherently synchronous, all of these solutions experience minimal to zero data loss, very similar to that in a local cluster sharing LUNs from the same storage subsystem. Even asynchronous replication such as GlobalMirror technologies can be used in this kind of stretched SAN structure if performance with synchronous replication is too slow for business requirements. But data replication cannot be guaranteed between sites since the I/O is completed to the application before the replication is complete.

- The infrastructure uses the network to accomplish the storage replication needs, using GLVM for example, and there is no SAN link between the two sites to share the same LUN across all cluster nodes, as shown in Figure 1-3 on page 10.

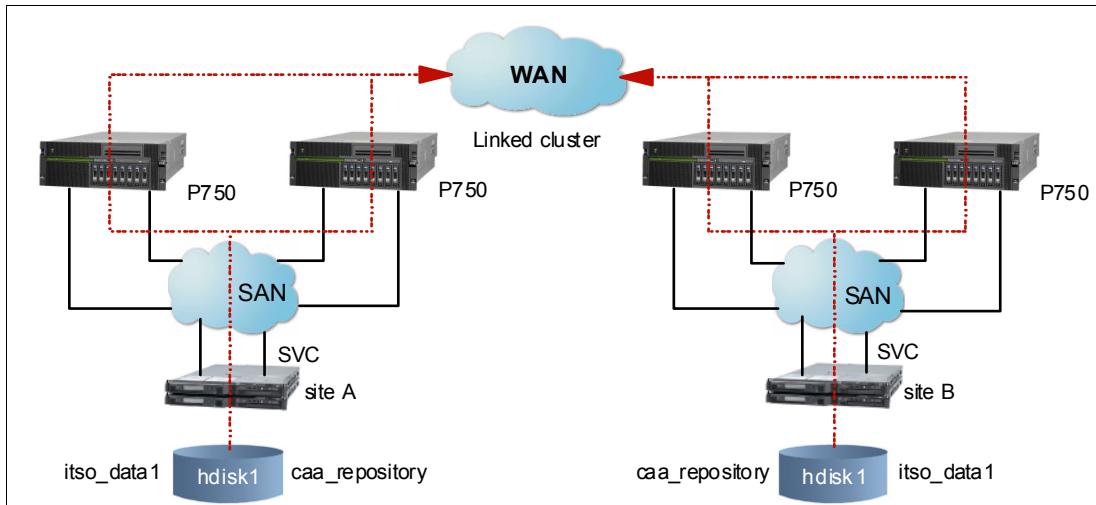


Figure 1-3 One different SAN in each site

Since PowerHA SystemMirror 7.1.2 Enterprise Edition uses the CAA repository disks (more details about repository disks can be found in 2.1.1, “Cluster Aware AIX” on page 16), in the case where the infrastructure is unable to provide the same LUN across all nodes, a new concept was created to accommodate these two different scenarios. When the same repository disk can be shared across all nodes in the cluster, it is a stretched cluster. If different disks are used on each site for the repository disks, the cluster is a linked cluster. See section 2.2.1, “Stretched and linked clusters” on page 19 for more information.

For stretched clusters, the repository disk is used as heartbeat path communication too. On linked clusters, this is not possible between the sites, but PowerHA SystemMirror 7.1.2 Enterprise Edition brings a new feature that you can use to automate decisions in split and merge situations with the possibility to use a SCSI-3 reservation tie breaker disk. For additional information about this topic, refer to 2.2.2, “Split/merge handling overview” on page 21 and Chapter 10, “Cluster partition management” on page 449.

The PowerHA cluster nodes can reside on different machine types and server classes as long as you maintain common AIX and cluster file set levels. This can be very valuable in the scenario where you acquire newer machines and leverage the older hardware to serve as the failover targets. In a campus environment, creating this stretched cluster could serve two roles:

- ▶ Providing high availability
- ▶ Providing a recovery site that contains a current second copy of the data

Historically, the main limiting factor for implementing disaster recovery (DR) solutions has been the cost. Today, as more and more clients reap the benefits of geographically dispersing their data, two of the more apparent inhibitors are:

- ▶ The risks associated with a fully automated clustered solution
- ▶ The associated value-add of wrapping the cluster software around the replication

A multisite clustered environment is designed so that by default if the primary site stops responding, the secondary site terminates the replication relationship and activates the resources at the backup location. As a best practice, plan for multiple redundant heart-beating links between the sites to avoid a separation. If you proceed with the assumption that the cluster design has redundant links between the sites, and in turn minimizes the risk of a *false down*, we continue to discuss the value-add that the PowerHA SystemMirror 7.1.2 Enterprise Edition brings.

Many clients use the various disk and IP replication technologies without leveraging the integration within the PowerHA Enterprise Edition only to find that they are left with more to manage on their own. For example, you can control remote physical volumes (RPVs) in Geographic Logical Volume Manager (GLVM) manually, but even after scripting the commands to activate the disks in the proper order, there are various cluster scenarios where you would need to append extra logic to achieve the desired results. This in part is the reason that the GLVM file set is included in AIX—the concept of try and buy. Once you have identified that the replication technology meets your needs, and you have sized the data links appropriately, you quickly realize that the management is significantly simpler when you exploit the integration within the PowerHA Enterprise Edition.

One of the major benefits that Power HA Enterprise Edition brings is that it has been comprehensively tested with not just the basic failover and fallback scenarios, but also with such things as rg_move and selective failover inherent in cluster mechanisms. In addition, using the PowerHA Enterprise Edition automatically reverses the flow and restarts the replication after the original site is restored. The integrated cluster `c1verify` functions are also intended to help identify and correct any configuration errors. The cluster EVENT logging is appended into the existing PowerHA logs, and the nightly verification checks identify whether any changes have occurred to the configuration. The replicated resource architecture in the Enterprise Edition allows finer control over the status of the resources. Through features such as application monitoring or the pager notification methods, you can receive updates any time that a critical cluster event has occurred.

Enabling full integration can also facilitate the ability to gracefully move resources and the testing of your scripts. By simply moving the resources from one site to the next, you can test the application stop and start scripts and ensure that everything is in working order. Leveraging some of the more granular options within the cluster, such as resource group location dependencies, can also facilitate the de-staging of lower priority test or development resources at the failover location whenever a production site failure has taken place. Using the site-specific dependencies, you could also specify that a set of resource groups always coexists within the same site. Another benefit of using PowerHA Enterprise Edition is the integrated logic to pass instructions to the disk storage subsystems automatically, based on the various events detected by the cluster that are integrated into the code and tested in many different ways.

1.3 Storage replication and the new HyperSwap feature

One of the major concerns when considering a replication solution is whether it maintains the integrity of the data after a hard site failover. The truth is that all of the solutions do a very good job, but they are not bullet-proof. One important note to keep in mind is the concept of garbage in garbage out (GIGO). Database corruption at the source site would be replicated to the target copy. This is the main reason that the DR design does not stop after selecting a replication technology.

Replicating the data only addresses one problem. In a well-designed disaster recovery solution, a backup and recovery plan must also exist. Tape backups, snapshots, and flashcopies are still an integral part of an effective backup and recovery solution. The frequency of these backups at both the primary and remote locations should also be considered for a thorough design.

Tip: An effective backup and recovery strategy should leverage a combination of tape and point-in-time disk copies to protect unexpected data corruption. Restore is very important, and regular restore tests need to be performed to guarantee that the disaster recovery is viable.

IBM PowerHA has been providing clustering and data replication functionality for a number of years and it continues to strive to be the solution of choice for IBM clients running on AIX. The software tightly integrates with a variety of existing disk replication technologies under the IBM portfolio and third-party technologies. Support for additional third-party replication technologies is continually being tested and planned for future software releases or via service flashes.

There are two types of storage replication: synchronous and asynchronous. Synchronous replication only considers the I/O completed after the write is done on both storages. Only synchronous replication can guarantee that 100% of transactions were correctly replicated to the other site, but since this can add a considerable amount of I/O time, the distance between sites must be considered for performance matters. This is the main reason asynchronous replication is used between very distant sites or with I/O sensitive applications.

In synchronous mirroring, both the local and remote copies must be committed to their respective subsystems before the acknowledgement is returned to the application. In contrast, asynchronous transmission mode allows the data replication at the secondary site to be decoupled so that primary site application response time is not impacted. Asynchronous transmission is commonly selected with the exposure that the secondary site's version of the data may be out of sync with the primary site by a few minutes or more. This lag represents data that is unrecoverable in the event of a disaster at the primary site. The remote copy can lag behind in its updates. If a disaster strikes, it might never receive all of the updates that were committed to the original copy.

Although every environment differs, more contention and disk latency is introduced the farther the sites reside from each other. However, there are no hard set considerations dictating whether you need to replicate synchronously or asynchronously. It can be difficult to provide an exact baseline for the distance delineating synchronous versus asynchronous replication.

Some clients are replicating synchronously between sites that are hundreds of miles apart, and the configuration suits them quite well. This is largely due to the fact that their environments are mostly read intensive and writes only occur sporadically a few times a day, so that the impact of the application response time due to write activity is minimal. Hence, factors such as the application read and write tendencies should be considered along with the current system utilization.

If a synchronous replication is suitable for your environment, consider a new feature that has been added to this PowerHA SystemMirror 7.1.2 Enterprise Edition release: HyperSwap. This is a good virtualization layer above replicated storage devices since it enables a fast takeover and continuous availability against storage failures. See 2.2.3, "HyperSwap overview" on page 21 and Chapter 4, "Implementing DS8800 HyperSwap" on page 59 for more information about the new PowerHA SystemMirror 7.1.2 Enterprise Edition HyperSwap feature.

Other replication technology configurations and tests made during this residency with SVC and XIV® storage can be found in Chapter 6, "Configuring PowerHA SystemMirror Enterprise Edition linked cluster with SVC replication" on page 195 and in Chapter 7, "Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replication" on page 253.

You can see that the differences between local high availability (HA) and DR revolve around the distance between the sites and ability, or inability (in this case a linked cluster is required),

to extend the storage area network (SAN). Local failover provides a faster transition onto another machine than a failover going to a geographically dispersed site. A scenario where synchronous replication is faster on two different sites than asynchronous replication is where HyperSwap would mask the failure of one of the storage subsystems to the application and not require a failover in the case where only a storage failure occurred.

In environments requiring the replication of data over greater distances where asynchronous disk-based replication might be a better fit, there is a greater exposure for data loss. There may be a larger delta between the data in the source and target copies. Also, the nature of that kind of setup results in the need of a failover if the primary storage subsystem is to go offline.

For local or stretched clusters, licensing of the IBM PowerHA SystemMirror Standard Edition typically suffices, with the exception of the HyperSwap feature and synchronous or asynchronous disk-level mirroring configurations that could benefit from the additional integrated logic provided with the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition solution. The additional embedded logic would provide automation to the management of the role reversal of the source and target copies in the event of a failover. Local clusters, assuming that virtualized resources are being used, can also benefit from advanced functions such as IBM PowerVM Live Partition Mobility between machines at the same site. This combination of the IBM PowerVM functions and IBM PowerHA SystemMirror clustering is useful for helping to avoid any service interruption for a planned maintenance event while protecting the environment in the event of an unforeseen outage.



Differences between IBM PowerHA SystemMirror Enterprise Edition version 6.1 and version 7.1.2



In this chapter we cover:

- ▶ Architecture changes
 - Cluster Aware AIX
 - ODM updates
- ▶ New features and functionalities
 - Stretched and linked clusters
 - Split/merge handling overview
 - HyperSwap overview
 - IPv6 support
 - PowerHA Smart Assists
 - IBM Systems Director
 - New or changed cluster administration tools
 - Features at a glance added since PowerHA 6.1
- ▶ Limitations
 - Restrictions
 - Unsupported changes
- ▶ Hardware and software prerequisites

2.1 Architecture changes

Earlier versions of PowerHA (version 5.1 to version 6.1) used the reliable scalable clustering technologies (RSCT) topology services for heartbeat communications. With the introduction of PowerHA 7.1 and Cluster Aware AIX (CAA) we use Storage Interconnected Resource Collection (SIRCOL). SIRCOL is the equivalent of a local cluster or site in terms of PowerHA.

2.1.1 Cluster Aware AIX

PowerHA SystemMirror 7.1.2 uses the Cluster Aware AIX (CAA) services to configure, verify and monitor the cluster topology. This is a major reliability improvement because core functions of the cluster services such as topology related services, now run in the kernel space. This makes it much less susceptible to be affected by the workload generated in the user space.

Communication paths

This includes the important process of sending and processing the cluster *heartbeats* by each participant node. Cluster communication is achieved by communicating over multiple redundant paths. The following redundant paths provide a robust clustering foundation that is less prone to cluster partitioning:

- ▶ TCP/IP

PowerHA SystemMirror and Cluster Aware AIX, via multicast, uses all network interfaces that are available for cluster communication. All of these interfaces are discovered by default and used for health management and other cluster communication. You can use the PowerHA SystemMirror management interfaces to remove any interface that you do not want to be used by specifying these interfaces in a *private* network.

- ▶ SAN-based

A redundant high-speed path of communication is established between the hosts by using the storage area network (SAN) fabric that exists in any data center between the hosts. Discovery-based configuration reduces the burden for you to configure these links.

- ▶ Repository disk

Health and other cluster communication is also achieved through the central repository disk.

Of these three, only the SAN-based one is optional for a local site cluster. However, none of these communication paths can be used between *two sites* in a linked cluster. In this case, you have to use unicast communication between the sites. More information can be found in “Linked cluster and multiple site support” on page 19.

Repository disk

This is one of the core components of CAA cluster topology services containing vital information about cluster topology and resources. What is new specifically in PowerHA SystemMirror 7.1.2 Enterprise Edition is defining two repository disks, one for each site, when configuring a linked cluster. The repositories between sites are kept in sync internally by CAA. For more information, see 3.1.4, “Cluster repository disk” on page 47.

When sites sunder or split, and then merge, CAA provides a mechanism to reconcile the two repositories. This can be done either through rebooting all the nodes on the losing side or through APIs implemented exclusively for RSCT. More information on this topic can be found in 10.4.2, “Configuring the split and merge policy” on page 455.

New quorum rule

Although the cluster continues operating if one or more nodes lose access to the repository disk, the affected nodes are considered to be in *degraded mode*. If, in addition, the heartbeat communication is also affected, then there is the potential for the nodes to form an independent cluster (partition) by seeing other nodes register an abnormal failure.

Therefore, PowerHA SystemMirror 7.1.2 Enterprise Edition does not allow a node to operate if it no longer has access to the repository disk *and* also registers an abnormal node down event. This allows a double failure scenario to be tolerated.

Tie breaker disk

The tie breaker disk is a new feature that can be configured in a PowerHA SystemMirror 7.1.2 Enterprise Edition cluster. However, to utilize it, you must use any disk that supports SCSI-3 persistent reservation.

It is an optional feature you can use to prevent a *partitioned* cluster, also known as *split brain*. If specified as an “arbitrator” in the split and merge policy of a cluster, the tie breaker decides which partition of the cluster survives. The one containing a node that succeeds in placing a SCSI-3 persistent reserve on the tie breaker disk wins, and hence survives. The loser is rebooted.

Similar behavior happens while merging the partitioned cluster. The nodes belonging to the partition that is unable to place the SCSI-3 persistent reserve belong to the losing side and will be rebooted.

Refer to 10.2, “Methods to avoid cluster partitioning” on page 451 for a full explanation and examples of configuring split/merge options using a tie breaker disk.

2.1.2 ODM updates

The following ODM classes have been added to PowerHA 7.12 to support new features.

HACMPcluster

The following new stanzas have been added to the class HACMPcluster:

- ▶ remote_HB_factor
- ▶ link_timeout
- ▶ multi_site_lc

These are shown in Example 2-1.

Example 2-1 HACMPcluster new ODM stanzas

```
HACMPcluster:  
    id = 1595970743  
    name = "xdtest"  
    nodename = "hacmp21"  
    sec_level = "Standard"  
    sec_level_msg = ""  
    sec_encryption = ""  
    sec_persistent = ""  
    last_node_ids = "1, 2"  
    highest_node_id = 2  
    last_network_ids = ""  
    highest_network_id = 0  
    last_site_ids = ""
```

```

highest_site_id = 0
handle = 2
cluster_version = 14
reserved1 = 0
reserved2 = 0
wlm_subdir = ""
settling_time = 0
rg_distribution_policy = "node"
noautoverification = 0
clvernodenname = ""
clverhour = 0
clverstartuoptions = 0
node_down_delay = 10
node_timeout = 20
remote_HB_factor = 0
link_timeout = 0
multi_site_lc = 0

```

- ▶ **remote_HB_factor** - Ratio of remote (unicast) heartbeats to local (multicast) heartbeat messages. The default is 10.
- ▶ **link_timeout** - Delay (milliseconds) added to the **node_timeout** + **node_down_delay** before marking the remote node down. The default is 30000.
- ▶ **multi_site_lc** - Definition to define new cluster types:
 - 1 = local cluster
 - 0 = stretched cluster
 - 1 = linked cluster

HACMPsircol

This ODM class has the addition of an extra repository disk stanza because linked clusters require two separate CAA repository disks (one stanza for local and stretched clusters and two stanzas for a linked cluster). Example 2-2 shows the file contents of a linked cluster.

Example 2-2 Linked cluster HACMPsircol

```

HACMPsircol:
    name = "xdtest_sircol"
    id = 0
    uuid = "0"
    ip_address = "228.1.10.10"
    repository = "00ccfe7445cd455c"
    repository = "00ccfe7445Rd421J"
    backup_repository = ""

```

HACMPsite

No changes to this stanza.

2.2 New features and functionalities

The IBM PowerHA SystemMirror 7.12 Enterprise Edition introduces the following features:

- ▶ Stretched and linked clusters

- ▶ Split/merge handling overview
- ▶ HyperSwap overview
- ▶ IPv6 support
- ▶ PowerHA Smart Assists
- ▶ IBM Systems Director

2.2.1 Stretched and linked clusters

PowerHA 7.12 Enterprise Edition introduced two new cluster types: *stretched* and *linked* clusters.

Stretched cluster

The term stretched cluster denotes a cluster that has sites defined in the same geographic location, for example, a campus style cluster. The key aspect about a stretched cluster is that it uses a shared repository disk. This means, *extended* distance sites with IP only connectivity are not possible with this configuration. A stretched cluster can support cross-site LVM mirroring, HyperSwap, and Geographical Logical Volume Mirroring (GLVM). Refer to Chapter 8, “Migrating to PowerHA SystemMirror 7.1.2 Enterprise Edition” on page 343 for an example of a GLVM snapshot migration to a stretched cluster. Refer to Figure 2-1 for an illustration of a typical stretched cluster.

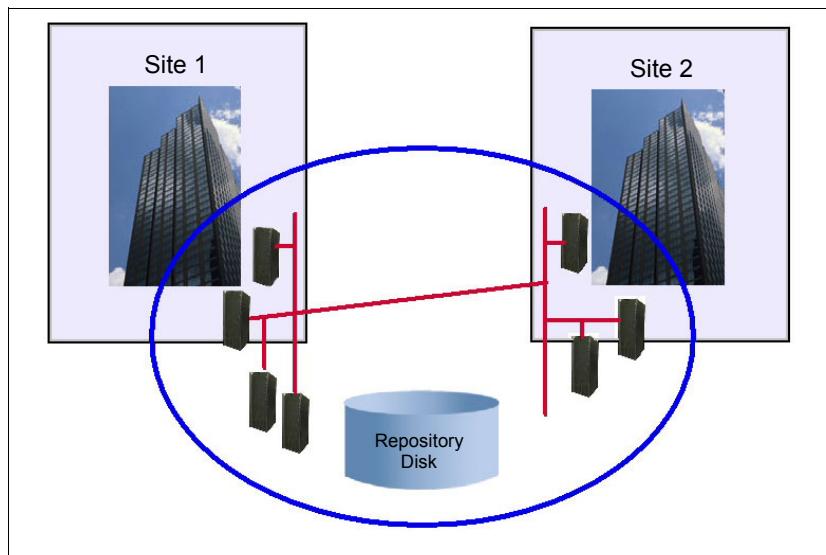


Figure 2-1 Stretched cluster example

A stretched cluster configuration can also be utilized with PowerHA 7.1.2 Standard Edition with the use of LVM Cross Site mirroring. See 5.1, “Cross-site LVM mirroring overview” on page 152. This is similar to what some clients have configured in previous versions.

The stretched cluster configuration is capable of utilizing all three levels of cluster communication provided by 2.1.1, “Cluster Aware AIX” on page 16.

Linked cluster and multiple site support

A linked cluster is expected to be the most common Enterprise Edition configuration (see Figure 2-2 on page 20). It allows the configuration of a traditional extended distance cluster between two sites, for example in London and New York. The key aspect of a linked cluster that makes it different from extended distance clusters in previous versions is the use of SIRCOL in CAA. This means that each site has its own CAA repository disk, which is

replicated automatically between sites by CAA. Linked cluster sites communicate with each other using unicast and not multicast. However, local sites internally still use multicast. Multicast still needs to be enabled in the network at each site.

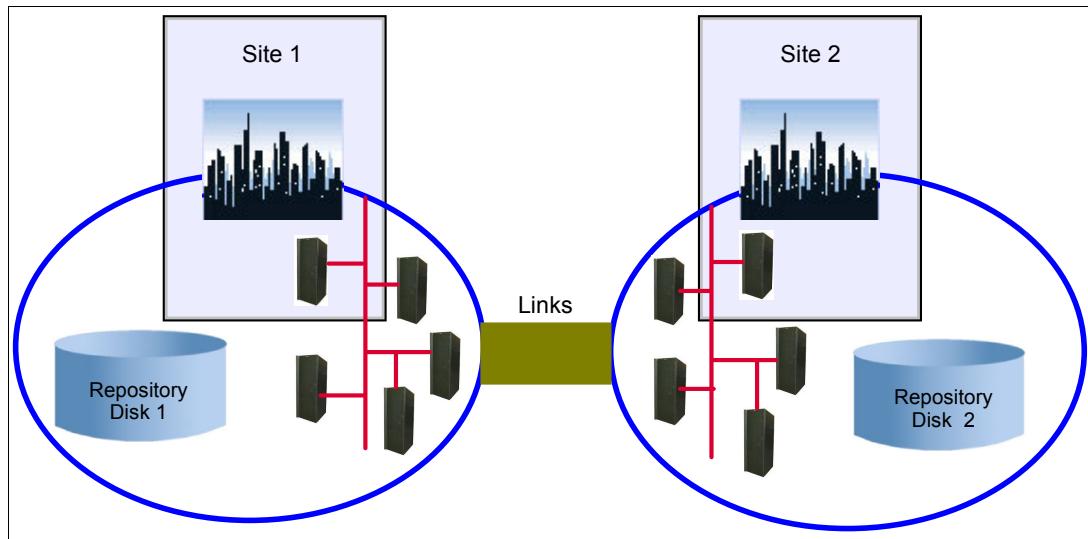


Figure 2-2 *Linked cluster illustration*

All interfaces are defined in this type of configuration as CAA “gateway addresses.” CAA maintains the repository information automatically across sites via unicast address communication. Refer to Table 2-1 for details about stretched and linked clusters.

Table 2-1 *Stretched vs. linked cluster cross reference*

Function	Stretched	Linked
Site communication	Multicast	Unicast
Repository disk	Shared	One local in each site
Cluster communication	- IP network - SAN fabric - Disk (repository)	IP network
Cross site logical volume mirroring	Available	Available
Geographical LVM mirroring (GLVM)	Available	Available
HyperSwap	Available	Available
Concurrent resource group with HyperSwap	Available	Not supported
SVC replicated resources V7000 replicated resources	Available	Available
DS8K replicated resources	Available	Available
XIV replicated resources	Available	Available
EMC SRDF*	Not supported	Available
Hitachi TrueCopy/HUR* HP Continuous Access	Not supported	Available

Note: The initial release of IBM PowerHA SystemMirror 7.1.2 Enterprise Edition had a support limitation for all third party storage replication previously supported with PowerHA 6.1 Enterprise Edition. However this limitation was removed and appropriate levels required can be found at:

<http://www-03.ibm.com/support/techdocs/atstr.nsf/WebIndex/FLASH10822>

2.2.2 Split/merge handling overview

In the current version of PowerHA, there are exactly *two* options to choose from how the cluster will behave in the case that a partitioned or split condition occurs. PowerHA 7.1.2 allows us to configure this while we are defining the cluster.

Cluster split

The default behavior, called *None*, means that each partition becomes an independent cluster. This can be very dangerous because the resources will be activated in both partitions!

The other option is *tie breaker*. The partition that survives is the one that is able to perform a SCSI-3 persistent reserve on the tie breaker disk. *All* nodes in the losing partition will reboot.

Cluster merge

When nodes rejoin the cluster it is called a merge operation. What happens next depends on which split policy is chosen. You have to use *Majority* if the split policy was *None*, and *Tie Breaker* if the split policy was *Tie Breaker*.

For a detailed description of this topic see 10.4.2, “Configuring the split and merge policy” on page 455.

2.2.3 HyperSwap overview

HyperSwap is a function that provides continuous availability against storage errors. It is based upon storage-based synchronous replication. When directed (or upon disk errors), AIX hosts accessing the primary disk subsystem automatically switch over to the backup copy of the data.

HyperSwap technology enables PowerHA SystemMirror to support the following capabilities for you:

- ▶ Enable storage maintenance without any application downtime.
- ▶ Enable migration from old to new storage.
- ▶ No disruption to the dependent applications.
- ▶ Natural extension of disaster recovery configurations.

Thus, HyperSwap helps eliminate primary disk subsystems as the single point of failure to provide the next level of continuous operations support within metro distances.

For complete details and examples of exploiting HyperSwap, see Chapter 4, “Implementing DS8800 HyperSwap” on page 59.

2.2.4 IPv6 support

PowerHA has been supporting IPv6 in previous releases. The following is the history of PowerHA and IPv6 development:

- ▶ PowerHA 5.4.1 or before - No IPv6 support was available.
- ▶ PowerHA 5.5 - Added support for IPv6 services and persistent labels. Since the underlying topology service (RSCT) did not have full native IPv6 support, boot IP labels were delayed.
- ▶ PowerHA 6.1 - Full native IPv6 support with the following limitations:
 - Ethernet and IPAT via aliasing only
 - Cannot change prefix length or netmask of service label via dare
 - Smart Assists not supported
- ▶ PowerHA 7.1 and PowerHA 7.1.1 - Due to CAA implementation, IPv6 boot label supports were again dropped.

PowerHA 7.1.2 re-introduced support for IPv6. This feature is introduced with an update to CAA. The IPv6 support includes the following:

- ▶ Support communications of IPv6 packets through CAA.
- ▶ Multicast IPv6.
- ▶ Gathering and displaying of IPv6 statistics and addresses in use.
- ▶ Support for new cluster and migration from IPv4.
- ▶ All interfaces of TCP/IP sockets support the AF_INET6 socket family.
- ▶ Derivation of the IPv6 address takes place in user space.
- ▶ IPv6 multicast address is derived from the IPv4 multicast address and conversion is handled in the user space.
- ▶ Mping support added for IPv6.
- ▶ During discovery autoconf6 is no longer executed.
- ▶ Smart Assists support for IPv6 label.

Table 2-2 shows a quick overview of IPv6 capability for each version.

Table 2-2 PowerHA and IPv6 support overview

	PowerHA 5.4.1 or before	PowerHA 5.5	PowerHA 6.1	PowerHA 7.1 PowerHA 7.1.1	PowerHA 7.1.2
IPv6 boot IP	Not supported	Not supported	Supported	Not supported	Supported
IPv6 service IP or persistent IP	Not supported	Supported	Supported	Supported	Supported

For more information on implementing IPv6 with PowerHA, see “Configuring IBM PowerHA SystemMirror with IPv6” on page 480.

2.2.5 PowerHA Smart Assists

PowerHA SystemMirror 7.1.2 provides the following Smart Assists for application support:

- ▶ DB2 UDB non-DPF Smart Assist
- ▶ DHCP Smart Assist

- ▶ DNS Smart Assist
- ▶ Lotus® Domino® Smart Assist
- ▶ Filenet P8 Smart Assist
- ▶ IBM HTTP Server Smart Assist
- ▶ SAP MaxDB Smart Assist
- ▶ Oracle Database Smart Assist
- ▶ Oracle Application Server Smart Assist
- ▶ Print Subsystem Smart Assist
- ▶ SAP Smart Assist
- ▶ Tivoli Directory Server Smart Assist
- ▶ TSM Admin Smart Assist
- ▶ TSM Client Smart Assist
- ▶ TSM Server Smart Assist
- ▶ Websphere Smart Assist
- ▶ Websphere MQ Smart Assist

The Smart Assists are provided as part of the licensed product filesets. When you install PowerHA, you need to install the appropriate Smart Assist required to support your application.

Note: Review the Smart Assist documentation carefully for any application-specific requirements.

Table 2-3 Smart Assist application version support

	SystemMirror 7.1.0	SystemMirror 7.1.2
DB2 Enterprise Edition	9.5	9.7
WAS	6.1	6.1
WAS N/D	6.1	6.1
HTTP Server	6.1	6.1
TSM	6.1	6.2
TDS	5.2	6.3
Filenet	4.5.1	4.5.1
Lotus Domino Server	8.5.1	8.5.1
Oracle Database Server	11gR1	11gR1
Oracle Application Server	10gR2	10gR2
SAP	SAP ERP Netweaver 2004s	SAP SCM 7.0 with Netweaver EHP1 for FVT SAP SCM 7.0 with Netweaver EHP2 for SVT
- MaxDB	N/A	7.6
- Oracle	10gR2	10gR2
- DB2	N/A	9.7
MQSeries®	7.0.1.5	7.0.1.5
AIX Print Server	AIX 6.1	AIX 6.1

	SystemMirror 7.1.0	SystemMirror 7.1.2
AIX DHCP	AIX 6.1	AIX 6.1
AIX DNS	AIX 6.1	AIX 6.1

2.2.6 IBM Systems Director

A new plug-in is provided for PowerHA management with IBM Systems Director 6.3. Enhancements include support for:

- ▶ Enterprise Edition
- ▶ Replicated storage
- ▶ Mirror pool support
- ▶ Volume group wizard
- ▶ Federated security wizard
- ▶ Capacity on demand
- ▶ Events management
- ▶ Reports management

PowerHA Enterprise Edition support

Support has now been added for PowerHA SystemMirror 7.12 Enterprise Edition. This includes the cluster configuration wizard allowing the defining of sites and additions to the cluster resource group management view.

Replicated storage support

A new wizard has been provided for replicated storage support called the *Replicated Mirror Group Wizard*. Support is provided for:

- ▶ DS8000® Series, Global Mirror and HyperSwap
- ▶ SAN Volume Controller (SVC) [7.1.2.1]
- ▶ XIV

Mirror pool support

Mirror pools are displayed for the volume group they reside in. There are a number of management options available, including the ability to create a new mirror pool.

Volume group wizard

This new wizard allows for the easy creation of a volume group.

Federated security wizard

This wizard is to provide end-to-end server security configuration such as LDAP server/client, encrypted file systems (EFS), user groups, users, and roles in a single wizard. Configuration of LDAP users, user groups and roles is also possible.

Capacity On Demand support

Support for adding a communication path to the HMC and creating Capacity On Demand object application controller has been added.

Event management

Accessible from the Events and Alerts Tab, this feature allows for the editing, viewing and creation of custom event actions. Users can customize commands, notifications, and recovery scripts of the event. The recovery counter can also be changed.

Reports management

This provides the ability to run and save reports through the Director plug-in. Saved reports are stored on the directory server. See Chapter 9, “PowerHA 7.1.2 for IBM Systems Director plug-in enhancements” on page 393 for more information.

2.2.7 New or changed cluster administration tools

This section covers command updates to CAA to support PowerHA SystemMirror 7.1.2 Enterprise Edition, mainly for site support.

mkcluster	<code>[-S sitename{ [cle_uuid=<UUID>:cle_globid=<id>:cle_prio=<prio>] }]</code> Specifies the name of the local site. If not specified, a default site with the name <i>local</i> is created. Currently, a cluster can support only two sites. To create a second site, use the chcluster command. The following site information may be specified: <code>cle_uuid</code> - The site UUID, which is honored as long as it is unique across the cluster. If not specified, the site UUID is automatically generated. <code>cle_globid</code> - The short ID of site, which must be a unique unsigned number greater than zero. If not specified, the site short ID is automatically generated. The following site attribute may be specified: <code>cle_prio</code> - The priority of a site. A lower value indicates a higher priority. The priority is mainly used in the context of synchronizing the repository metadata.
chcluster	<code>[-S sitename{ [cle_uuid=<UUID>:cle_globid=<id>:cle_prio=<prio>] }]</code> This uses the same attributes as mkcluster. <code>-r + remote_reposdisk</code>
lscuster	<code>-m</code> has been enhanced to include site information. <code>-s</code> enhanced to include <i>remote pkts sent</i> and <i>remote pkts recv</i> . <code>-d</code> shows the disks known to each node along with the repository. Shared disks are site specific, shown by the disks discovered by each node.
ciras dumprepos	<code>-v</code> (undocumented) option added for verbose repos disk content.
mping	Support for AF_INET6 (IPv6) socket family has been added.

2.2.8 Features at a glance added since PowerHA 6.1

Figure 2-3 on page 26 shows a list of features and the corresponding PowerHA levels in which they were added.

PowerHA Feature Name	7.1.0 SE	7.1.1 SE	7.1.2 SE	7.1.2 EE
Rootvg system event protection	✓	✓	✓	✓
User Defined Resources	✓	✓	✓	✓
Start/Stop After Resource Group Dependency	✓	✓	✓	✓
Service IP Distribution Policies with Source IP	✓	✓	✓	✓
Clmgr command line	✓	✓	✓	✓
IBM Director based graphical interface	✓	✓	✓	✓
Adaptive Failover Policy	✓	✓	✓	✓
Clicmd distributed command	✓	✓	✓	✓
SAP Smart Assist	✓	✓	✓	✓
SAP LiveCache Hot Standby Fast Failover (SVC,DS8K)		✓	✓	✓
JFS2 mount guard support		✓	✓	✓
Repository disk resilience and backup		✓	✓	✓
Federated Security (encrypted JFS, LDAP, etc)		✓	✓	✓
Application foreground startup option		✓	✓	✓
Repository disk resilience		✓	✓	✓
Physical volume rename		✓	✓	✓
Private networks		✓	✓	✓
Repository disk backup			✓	✓
IPv6 support			✓	✓
Replicated Resource (GLVM,DS8k,SVC,v7000,etc)				✓
HyperSwap				✓
Split/Merge management options				✓

Figure 2-3 PowerHA version features list

2.3 Limitations

In this section we talk about limitations in PowerHA SystemMirror 7.1.2 Enterprise Edition. We also discuss the restrictions and unsupported changes in the product.

2.3.1 Deprecated features

Starting with PowerHA SystemMirror 7.1, the following features are no longer available:

- ▶ IP address takeover (IPAT) via IP replacement
- ▶ Locally administered address (LAA) for hardware MAC address takeover (HWAT)
- ▶ Heartbeat over IP aliases
- ▶ The following IP network types:
 - ATM
 - FDDI
 - Token ring
- ▶ The following point-to-point (non-IP) network types:
 - RS232
 - TMSCSI
 - TMSSA
 - Disk heartbeat (diskhb)
 - Multinode disk heartbeat (mndhb)

- ▶ Two-node configuration assistant
- ▶ WebSMIT (replaced with the IBM Systems Director plug-in)

Though PowerHA Enterprise Edition was never supported with WebSMIT, the 7.1.2 version is supported with the IBM Systems Director plug-in.

2.3.2 Restrictions

In this section, we list the restrictions in the 7.1.2 version at the time of writing this book. Some of these restrictions might change in the future. Users are advised to contact IBM sales and support to get the latest information. The restrictions are as follows:

- ▶ OEM storage features
- ▶ Network tunables

OEM storage replicated support

At the time of writing, PowerHA SystemMirror 7.1.2 Enterprise Edition does not support EMC SRDF and Hitachi TrueCopy. This restriction is expected to be removed in a future update.

Network tunables

Configure NIC not to use **autonegotiate**, but to run at the desired speed and duplex value. For the switches turn off the following options:

- ▶ Spanning tree algorithm
- ▶ portfast
- ▶ uplinkfast
- ▶ backbonefast

PowerHA SystemMirror 7.1.2 Enterprise Edition ignores network tunables.

2.3.3 Unsupported changes

Here we list some of the unsupported changes in PowerHA SystemMirror cluster:

- ▶ The hostname of a cluster node cannot be changed after the cluster is configured. To change the hostname, you must first remove the Cluster Aware AIX (CAA) cluster definition, update PowerHA SystemMirror and AIX operating system configurations, and then synchronize the changes to recreate the CAA cluster with the new hostname.
- ▶ You cannot change the IP address that corresponds to the hostname of a cluster node after the cluster is configured in Cluster Aware AIX (CAA).
- ▶ In PowerHA SystemMirror 7.1.2 Enterprise Edition, you cannot change *linked* cluster to *local* or *stretched* cluster. There are limitations on changing *local* and *stretched* cluster to a *linked* cluster.

2.4 Hardware and software prerequisites

For information about supported hardware and required AIX levels for PowerHA SystemMirror 7.1.2 Enterprise Edition, see 3.2, “Hardware and software requirements” on page 50.



Planning

In this chapter we provide some guidelines on the planning topic for PowerHA SystemMirror 7.1.2 Enterprise Edition. The planning is an important step during a PowerHA cluster deployment. We focus on several infrastructure and software requirements for deploying a PowerHA SystemMirror 7.1.2 Enterprise Edition cluster. For general planning purposes, refer also to the PowerHA SystemMirror planning publication for the version you are using at:

http://pic.dhe.ibm.com/infocenter/aix/v6r1/topic/com.ibm.aix.powerha.plangd/hac_mpplangd_pdf.pdf

This chapter contains the following topics:

- ▶ Infrastructure considerations and support
 - Network considerations
 - Storage and SAN considerations
 - Cluster repository disk
 - Tie breaker disk
- ▶ Hardware and software requirements
 - Hardware requirements for the AIX solution
 - Software requirements for PowerHA 7.1.2

3.1 Infrastructure considerations and support

In this section, we detail several aspects regarding the infrastructure requirements for PowerHA SystemMirror 7.1.2 Enterprise Edition.

3.1.1 Hostname and node name

Unlike earlier versions, now PowerHA SystemMirror 7.1.1 has strict rules for which interface can be the hostname due to the new CAA layer requirements.

Important:

- ▶ The hostname cannot be an alias in the /etc/hosts file.
- ▶ The name resolution for the hostname must work for both ways, therefore a limited set of characters can be used.
- ▶ The IP address that belongs to the hostname must be reachable on the server, even when PowerHA is down.
- ▶ The hostname cannot be a service address.
- ▶ The hostname cannot be an address located on a network which is defined as private in PowerHA.
- ▶ The hostname, the CAA node name, and the “communication path to a node” must be the same.
- ▶ By default, the PowerHA, nodename, the CAA nodename, and the “communication path to a node” are set to the same name.
- ▶ The hostname and the PowerHA nodename can be different.
- ▶ The hostname cannot be changed after the cluster configuration.

The rules leave the base addresses and the persistent address as candidates for the hostname. You can use the persistent address as the hostname only if you set up the persistent alias manually before you configure the cluster topology.

On the Domain Name System (DNS) server, you can specify any “external” name for the service addresses so that the clients can use this name when they connect to the application that runs in the cluster.

Note: We could not configure a PowerHA 7.1.1 cluster because of a mismatch in the way the hostname was used. The hostname was defined using lowercase characters but in the /etc/hosts file it was written using uppercase. All the default TCP/IP commands worked fine. But the cluster setup failed. We updated the hosts file and then cluster setup worked fine.

3.1.2 Network considerations

The network infrastructure plays a major role in defining the cluster configuration. When using sites several considerations apply: network technologies used for node communication within or between the sites, network bandwidth and latency for the case of replicating the data over TCP/IP with GLVM, IP segments for each site, the firewall, and the DNS configurations. The network configuration also dictates what communication paths are used for heartbeating.

You can find general aspects regarding the networking in an environment using PowerHA Enterprise Edition in *Exploiting IBM PowerHA SystemMirror Enterprise Edition*, SG24-7841.

Cluster multicast IP address and PowerHA site configuration

Cluster monitoring and communication require that multicast communication be used.

Multicast consists of sending messages or information from the source to a group of hosts simultaneously in a single transmission. This kind of communication uses network infrastructure very efficiently because the source sends a packet only once, even if it needs to be delivered to a large number of receivers, and other nodes in the network replicate the packet to reach multiple receivers only when necessary. Cluster multicast communication is implemented at the CAA level. A multicast address is also known as a class D address. Every IP datagram whose destination address starts with 1110 is an IP multicast datagram. The remaining 28 bits identify the multicast group on which the datagram is sent.

Starting with PowerHA SystemMirror 7.1.2 sites can be used in the cluster configuration in both Standard and Enterprise editions. In this version, only two sites can be configured in a cluster. When using sites, the multicast communication depends on the cluster type:

- ▶ *Stretched* clusters use multicast communication between all nodes across all sites. In this case, a single multicast IP address is used in the cluster. CAA is unaware of the PowerHA site definition.
- ▶ *Linked* clusters use multicast communication only within the site and unicast protocol for communication between the cluster nodes across the sites. Unlike multicast, unicast communication involves a one-to-one communication path with one sender and one receiver. This cluster type uses two multicast IP group addresses, one in each site. For the particular case of a linked cluster with two nodes, one at each site, the multicast addresses are defined, but the node-to-node communication is using only unicast.

For PowerHA operation, the network infrastructure must handle the IP multicast traffic properly:

- ▶ Enable multicast traffic on all switches used by the cluster nodes.
- ▶ Check the available multicast traffic IP address allocation.
- ▶ Ensure that the multicast traffic is properly forwarded by the network infrastructure (firewalls, routers) between the cluster nodes, according to your cluster type requirements (stretched or linked).

You can specify the multicast address when you create the cluster, or you can have it generated automatically when you synchronize the initial cluster configuration. The following examples detail the multicast IP address generation by the CAA software for the two types of PowerHA clusters:

- ▶ A stretched cluster. The multicast IP address is generated as in the case of a cluster without sites. CAA generates the multicast address based on a local IP, associated with the hostname of a cluster node, by replacing the first byte of the IP address with 228: The address is generated during the first synchronization, at the time of CAA cluster creation. For example, you have a 3-node cluster with a single IP segment at both sites:

Site1:

```
svca1: 192.168.100.60 (Site 1)  
svca2: 192.168.100.61 (Site 1)
```

Site2:

```
svcb1: 192.168.100.62 (Site 2)
```

The default generated multicast IP is 228.168.100.60.

You can check the generated multicast address, after CAA cluster creation, using **lscuster -c**. See Example 3-1.

Example 3-1 Node IP addresses and multicast IP - stretched cluster

```
root@svca1:/>lscuster -c
Cluster Name: ihs_cluster
Cluster UUID: 73a14dec-42d5-11e2-9986-7a40cdea1803
Number of nodes in cluster = 3
    Cluster ID for node svca1: 1
    Primary IP address for node svca1: 192.168.100.60
    Cluster ID for node svca2: 2
    Primary IP address for node svca2: 192.168.100.61
    Cluster ID for node svcb1: 3
    Primary IP address for node svcb1: 192.168.100.62
Number of disks in cluster = 1
    Disk = hdisk2 UUID = ffe72932-2d01-1307-eace-fc7b141228ed cluster_major
= 0 cluster_minor = 1
Multicast for site LOCAL: IPv4 228.168.100.60 IPv6 ff05::e4a8:643c
```

In our test environment we observed that the IP address used for building the multicast IP is the IP address of the node initializing the CAA cluster configuration, at the time of first synchronization of the PowerHA cluster configuration (node svca1 in our case).

- ▶ A linked cluster. Two multicast IP addresses are generated: one at each site. For Site1 the multicast IP is generated as in the stretched cluster case. The second multicast IP address is generated using the first three bytes of the multicast IP in Site1, adding the last byte of the IP address of one of the nodes in Site2. Site1 and Site2 are associated according to their IDs in the *HACMPsite* ODM class. For example, for a cluster with four nodes, two in each site and one IP segment in each site:

Site1:

```
glvma1:192.168.100.55
glvma2:192.168.100.56
```

Site2:

```
glvmb1:10.10.100.57
glvmb2:10.10.100.58
```

The generated multicast IP addresses are:

Site1: **228.168.100.56**

Site2: **228.168.100.57**

Example 3-2 shows the output of **lscuster -c** command for this cluster.

Example 3-2 Node IP addresses and multicast IP - linked cluster

```
# lscuster -c
Cluster Name: glvma2_cluster
Cluster UUID: 6609a6cc-4512-11e2-9fce-7a40cf0aea03
Number of nodes in cluster = 4
    Cluster ID for node glvma1: 1
    Primary IP address for node glvma1: 192.168.100.55
    Cluster ID for node glvma2: 3
    Primary IP address for node glvma2: 192.168.100.56
    Cluster ID for node glvmb1: 4
    Primary IP address for node glvmb1: 10.10.100.57
```

```

Cluster ID for node glvmb2: 8
Primary IP address for node glvmb2: 10.10.100.58
Number of disks in cluster = 2
Disk = hdisk2 UUID = 565cf144-52bd-b73d-43ea-4375b50d290b cluster_major
= 0 cluster_minor = 1
Disk = UUID = c27d8171-6da8-728a-7ae8-ee9d455b95df cluster_major = 0
cluster_minor = 2
Multicast for site siteA: IPv4 228.168.100.56 IPv6 ff05::e4a8:6438
Multicast for site siteB: IPv4 228.168.100.57 IPv6 ff05::e4a8:6439

```

In our case, we generated the CAA cluster by creating the PowerHA cluster definition on node glvma2 and synchronizing the configuration. We observed that the multicast IP address in the second site is the one associated with the node having the lowest ID in the second site, as indicated by the `lsccluster` command in Example 3-2 on page 32.

You can specify the multicast IP addresses at the time of creating the PowerHA cluster rather than having a generated one. In that case, do not use the following multicast groups:

- 224.0.0.1 This is the all-hosts group. If you ping that group, all multicast-capable hosts on the network should answer because every multicast-capable host must join that group at startup on all its multicast-capable interfaces.
- 224.0.0.2 This is the all-routers group. All multicast routers must join that group on all its multicast-capable interfaces.
- 224.0.0.4 This is the all DVMRP (Distance Vector Multicast Routing Protocol) routers group.
- 224.0.0.5 This is the all OSPF (Open Shortest Path First) routers group.
- 224.0.0.13 This is the all PIM (Protocol Independent Multicast) routers group.

Note: The range 224.0.0.0 to 224.0.0.255 is reserved for local purposes such as administrative and maintenance tasks, and data that they receive is never forwarded by multicast routers. Similarly, the range 239.0.0.0 to 239.255.255.255 is reserved for administrative purposes. These special multicast groups are regularly published in the Assigned Numbers RFC. You can check the RFCs published on the Internet at:

<http://www.ietf.org/rfc.html>

To verify whether nodes in your environment support multicast-based communication, use the `mping` command, which is part of the CAA framework in AIX, and is included in the `bos.cluster.rte` fileset. The `mping` command options are shown in Figure 3-1 on page 34.

```

# mping
mping version 1.1
Must specify exactly one of -r or -s

Usage: mping -r|-s [-a address] [-p port] [-t ttl] [-c|-n pings] [-v]
-r|-s      Receiver or sender. Required argument,
           and are mutually exclusive
-a address Multicast address to listen/send on,
           overrides the default of 227.1.1.1.
           This can accept either an IPv4 or IPv6 address in decimal
notation.
-p port    Multicast port to listen/send on,
           overrides the default of 4098.
-t ttl     Multicast Time-To-Live to send,
           overrides the default of 1.
-n|-c pings The number of pings to send,
           overrides the default of 5.
-6         Sets the default multicast group address to the IPv6 default of
ff05::7F01:0101.
-v         Verbose mode. Declare multiple times to increase verbosity.
-?|-h     This message.

```

Figure 3-1 mping command options

For an example of using the **mping** command we used a stretched cluster with three nodes (nodea1, nodea2, and nodeb1) and the multicast group address 228.100.100.10. We sent five packets from nodea1 and received them on nodea2 and nodeb1, using the multicast IP address. We first started **mping** in the receive mode on nodes nodea2 and nodeb1 using the following command options:

```
mping -r -v -a 228.100.100.10
```

Then we started the sender (nodea1) using the following command options:

```
mping -s -v -c 5 -a 228.100.100.10
```

The whole picture of the test is illustrated in Figure 3-2 on page 35.

```

nodea1>mping -s -v -c 5 -a 228.100.100.10
mping version 1.1
Connecting using IPv4.
mpinging 228.100.100.10/4098 with ttl=1:

32 bytes from 192.168.100.61 seqno=0 ttl=1 time=0.240 ms
32 bytes from 192.168.100.62 seqno=0 ttl=1 time=0.374 ms
32 bytes from 192.168.100.61 seqno=1 ttl=1 time=0.231 ms
32 bytes from 192.168.100.62 seqno=1 ttl=1 time=0.348 ms
32 bytes from 192.168.100.61 seqno=2 ttl=1 time=0.235 ms
32 bytes from 192.168.100.62 seqno=2 ttl=1 time=0.391 ms
32 bytes from 192.168.100.61 seqno=3 ttl=1 time=0.264 ms
32 bytes from 192.168.100.62 seqno=3 ttl=1 time=0.355 ms
32 bytes from 192.168.100.61 seqno=4 ttl=1 time=0.225 ms
32 bytes from 192.168.100.62 seqno=4 ttl=1 time=0.357 ms
Sleeping for 1 second to wait for any additional packets to arrive.

--- 228.100.100.10 mping statistics ---
5 packets transmitted, 10 packets received, 0% packet loss
round-trip min/avg/max = 0.225/0.302/0.391 ms
nodea1>

nodeb1>mping -r -v -a 228.100.100.10
mping version 1.1
Connecting using IPv4.
Listening on 228.100.100.10/4098:

Replies to mping from 192.168.100.60 bytes=32 seqno=0 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=1 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=2 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=3 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=4 ttl=1
nodea2>mping -r -v -a 228.100.100.10
mping version 1.1
Connecting using IPv4.
Listening on 228.100.100.10/4098:

Replies to mping from 192.168.100.60 bytes=32 seqno=0 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=1 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=2 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=3 ttl=1
Replies to mping from 192.168.100.60 bytes=32 seqno=4 ttl=1

```

Figure 3-2 mping example

IPv6 address planning

This section explains the IPv6 concepts and provides details for planning a PowerHA cluster using IPv6 addresses. IPv6 support is available for PowerHA 7.1.2, and later versions.

IPv6 address format

IPv6 increases the IP address size from 32 bits to 128 bits, thereby supporting more levels of addressing hierarchy, a much greater number of addressable nodes, and simpler auto configuration of addresses.

Figure 3-3 shows the basic format for global unicast IPv6 addresses.

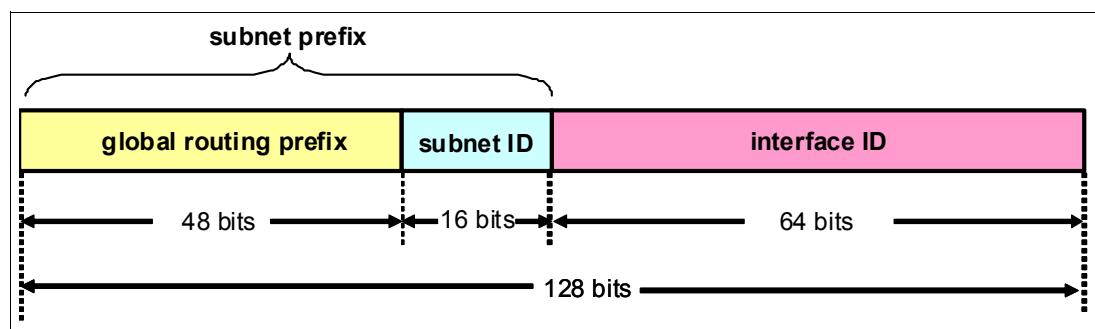


Figure 3-3 IPv6 address format

IPv6 addresses contain three parts:

- ▶ Global Routing Prefix
 - The first 48-bits (in general) are for a global prefix, distributed for global routing.
- ▶ Subnet ID
 - The second 16 bits are freely configurable in a field available to define within sites.
- ▶ Interface ID
 - The last 64 bits for distributing for each network device.

Subnet prefix considerations

The subnet prefix, which corresponds to the subnet mask for IPv4, is a combination of the *global routing prefix* and *subnet ID*. Although you are free to have longer subnet prefixes, in general 64 bits is a suitable length. IPv6 functions such as *Router Advertisement* are designed and assumed to use the 64-bit length subnet prefix. Also, the 16-bit subnet ID field allows 65,536 subnets, which are usually enough for general purposes.

IPv6 address considerations

The three basic IPv6 addresses are the following:

- ▶ Link-local address

Link-local addresses are IP addresses that are configured for the purpose to communicate within the site (that cannot go outside the network router). This term also exists in IPv4. The IP range of the link-local address for IPv4 and IPv6 is:

- 169.254.0.0/16 for IPv4
- fe08::/10 for IPv6

Although this address was optional in IPv4, in IPv6 it is now required. Currently in AIX, these are automatically generated based on the EUI-64 format which uses the network cards MAC address. The logic of the link-local address creation is as follows:

Say you have a network card with a MAC address of 96:D8:A1:5D:5A:0C.

- a. The 7th bit will be flipped and FFEE will be added after the 24th bit making it 94:D8:A1:FF:EE:5D:5A:0C.
- b. The subnet prefix fe08:: will be added, providing you with the link-local address FE08::94D8:A1FF:EE5D:5A0C.

In AIX, the **autoconf6** command is responsible for creating the link-local address.

- ▶ Global unicast address

These are IP addresses configured to communicate outside of the router. The range 2000::/3 is provided for this purpose. The following ones are predefined global unicast addresses:

- 2001:0000::/32 - Teredo address defined in RFC 4380
- 2001:db8::/32 - Provided for document purposes defined in RFC 3849
- 2002::/16 - 6 to 4 address defined in RFC 3056

- ▶ Loopback address

The same term as for IPv4. This uses the following IP address:

- ::1/128

For PowerHA, you can have your boot IPs configured to the link-local address if this suits you. However, for configurations involving sites, it will be more suitable for configuring boot IPs with

global unicast addresses that can communicate with each other. The benefit is that you can have additional heartbeating paths, which helps prevent cluster partitions.

The global unicast address can be configured *manually* and *automatically* as follows:

- ▶ Automatically configured IPv6 global unicast address
 - Stateless IP address
 - Global unicast addresses provided through a Neighbor Discovery Protocol (NDP). Similar to link-local addresses, these IPs will be generated based on the EUI-64 format. In comparison, the subnet prefix will be provided by the network router. The client and the network router must be configured to communicate through the NDP for this address to be configured.
 - Stateful IP address
 - Global unicast addresses provided through an IPv6 DHCP server.
- ▶ Manually configured IPv6 global unicast address
 - The same term as for IPv4 static address.

In general, automatic IPv6 addresses are suggested for unmanaged devices such as client PCs and mobile devices. Manual IPv6 addresses are suggested for managed devices such as servers.

For PowerHA, you are allowed to have either automatic or manual IPv6 addresses. However, take into consideration that automatic IPs will have no guarantee to persist. CAA restricts you to have the hostname labeled to a configured IP address, and also does not allow you to change the IPs when the cluster services are active.

IPv4/IPv6 dual stack environment

When migrating to IPv6, in most cases, it will be suitable to keep your IPv4 networks. An environment using a mix of different IP address families on the same network adapter is called a dual stack environment.

PowerHA allows you to mix different IP address families on the same adapter (for example, IPv6 service label on the network with IPv4 boot, IPv4 persistent label on the network with IPv6 boot). However, the best practice is to use the same family as the underlying network for simplifying planning and maintenance.

Figure 3-4 on page 38 shows an example of this configuration.

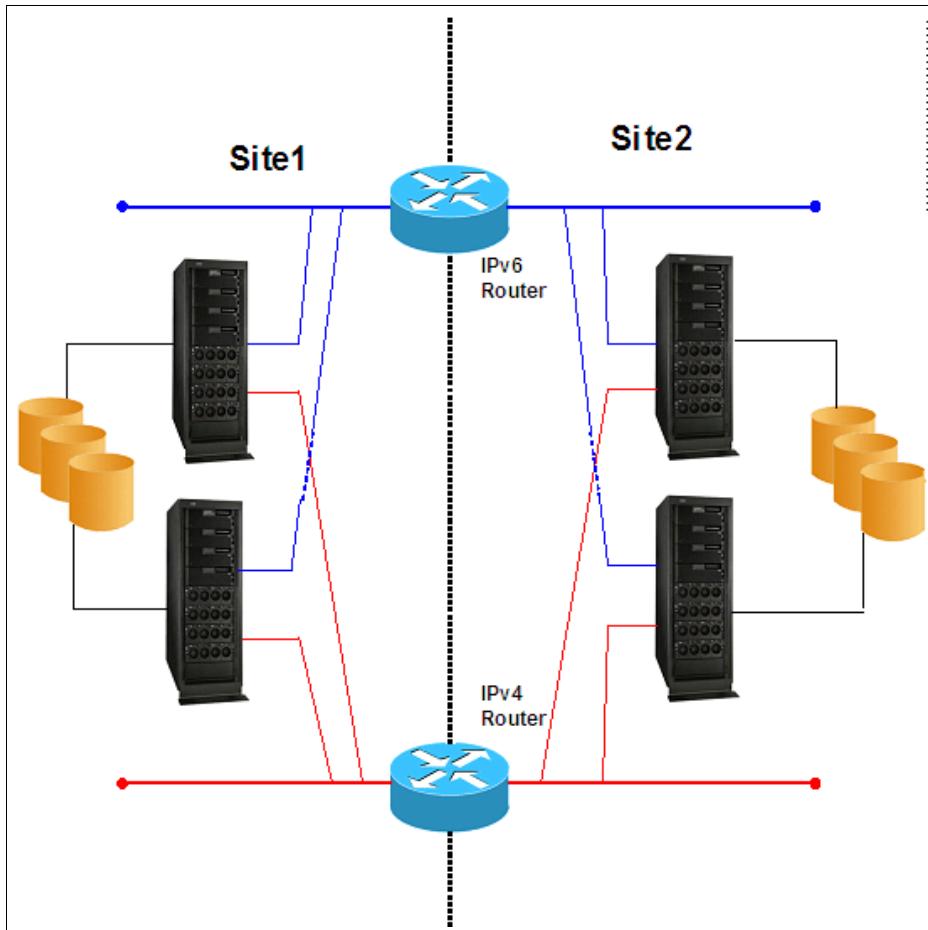


Figure 3-4 IPv6 dual stack environment

Multicast and IPv6

PowerHA SystemMirror 7.1.2 or later supports IP version 6 (IPv6). However, you cannot explicitly specify the IPv6 multicast address. CAA uses an IPv6 multicast address that is derived from the IP version 4 (IPv4) multicast address.

To determine the IPv6 multicast address, a standard prefix of 0xFF05 is combined using the logical OR operator with the hexadecimal equivalent of the IPv4 address. For example, the IPv4 multicast address is 228.8.16.129 or 0xE4081081. The transformation by the logical OR operation with the standard prefix is 0xFF05:: | 0xE4081081. Thus, the resulting IPv6 multicast address is 0xFF05::E408:1081.

The netmon.cf file

The netmon.cf file is an optional configuration file that can be used to complement the normal detection of interface failures based on the available communication paths between the cluster nodes by providing additional targets not part of the cluster itself that can be reached by the cluster nodes.

This file is used by the RSCT services. It was introduced in earlier releases for clusters with single adapter networks. With the introduction of CAA features exploited by PowerHA 7.1, this file is optional for the clusters using real network adapters, since the interface status is determined by CAA. However, in a VIOS environment, failure of the physical network adapter and network components outside the virtualized network might not be detected reliably. To

detect external network failures, you must configure the netmon.cf file with one or more addresses outside of the virtualized network.

As with the virtual environment case, the netmon.cf file can also be used for site configurations to determine a network down condition when the adapters are active in the site, but the link between the sites is down.

In the current implementation of RSCT 3.1.4 and PowerHA 7.1.2, the netmon functionality is supported by the RSCT group services instead of the topology services. Group services will report adapter DOWN if all the following conditions are met simultaneously:

- ▶ !REQD entries are present in netmon.cf for a given adapter
- ▶ Netmon reports adapter down
- ▶ The adapter is reported “isolated” by CAA

You can configure the netmon.cf file in a manner similar to previous releases. In PowerHA clusters, the file has the following path: /usr/es/sbin/cluster/netmon.cf, and it needs to be manually populated on each cluster node. Environments using virtual Ethernet and shared Ethernet adapters on VIOS to access external resources require the use of the netmon.cf file populated with IP addresses outside the machine hosting the logical partition for proper interface status determination. In this case, the specification of the netmon.cf file is the same as in prior releases:

```
!REQD <owner> <target>
<owner>      It is the originating interface name or IP address of that interface.
<target>      It is the target ping address to be used to test the connectivity.
```

You can find more details in the description text of APAR IZ01331, which introduces the netmon functionality for VIOS environments:

<http://www.ibm.com/support/docview.wss?uid=isg1IZ01332>

Additional considerations for the netmon.cf file are:

- ▶ You can add up to 32 lines for one adapter. If there is more than one entry for an adapter, then PowerHA tries to ping them all. As long as at least one target replies, the interface is marked good.
- ▶ Ensure that you add at least one line for each base adapter. It is advised to add a ping target for your persistent addresses.
- ▶ The file can be different on each cluster node.
- ▶ The target IP address should be associated with a critical device in your network, such as a gateway, a firewall system, or a DNS.

Example 3-3 shows a netmon.cf file with multiple IP/host interface examples.

Example 3-3 netmon.cf definition examples

```
#This is an example for monitoring en0 Virtual Ethernet interface on an LPRAR
#using two IP addresses of the DNS systems
!REQD en0 192.168.100.1
!REQD en0 192.168.100.2
```

```
#Example using the IP address of the interface instead of the interface name
#
!REQD 10.10.10.60 10.10.10.254
```

```
#Example using the IP interface names (they need to be resolvable to the actual IP
#addresses !!!)
```

```
!REQD host1_en2 gateway1
```

Note: PowerHA 7.1.2 supports IPv6. In this case, `netmon.cf` can be configured with IPv6 addresses in a similar manner as the IPv4 case. There are no special considerations for the IPv6 case.

3.1.3 Storage and SAN considerations

This section describes the storage and SAN considerations.

SAN-based heartbeat

PowerHA 7.1.2 supports SAN-based heartbeat only within a site. IBM intends to enhance its facility for inter-site heartbeating in upcoming releases of the PowerHA SystemMirror software solution.

The SAN heartbeating infrastructure can be accomplished in several ways:

- ▶ Using real adapters on the cluster nodes and enabling the storage framework capability (`sfwcomm device`) of the HBAs. Currently, FC and SAS technologies are supported. Refer to the following Infocenter page for further details about supported HBAs and the required steps to set up the storage framework communication:
http://pic.dhe.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.cluster_aware/claware_comm_setup.htm
- ▶ In a virtual environment using NPIV or vSCSI with a VIO server, enabling the `sfwcomm` interface requires activating the target mode (the `tme` attribute) on the real adapter in the VIO server and defining a private VLAN (ID 3358) for communication between the partition containing the `sfwcomm` interface and the VIO server. The real adapter on the VIO server needs to be a supported HBA as indicated in the previous reference link. For a practical example of setting up a storage communication framework interface in a virtual environment, refer to Chapter 3.7.13,"Configure SAN heart beating in virtual environment" of *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030.

Storage-based replication environments

The PowerHA clusters using storage-based replication can be implemented only with the PowerHA Enterprise Edition version. In this section, we describe the storage-based replication technologies integrated with PowerHA SystemMirror 7.1.2 Enterprise Edition.

Out-of-band versus in-band storage control

The communication between the cluster nodes and the storage system is required in a cluster using storage-based replication for management of the replication pairs under the PowerHA software control. This communication is dependent on the storage technology; it can be performed in two ways:

- ▶ Out-of-band control

The cluster software uses the storage management CLI functions and communicates with the storage in a different communication path from the current I/O traffic, usually on a TCP/IP network.

- ▶ In-band control

The cluster software communicates with the storage system using the same path as the disk I/O path, usually the SAN fiber channel network.

In PowerHA 7.1.2, in-band communication can be used with DS8800 storage systems while out-of-band control can be used with all DS8000 storage models. Figure 3-5 shows a comparative approach between the two communication models.

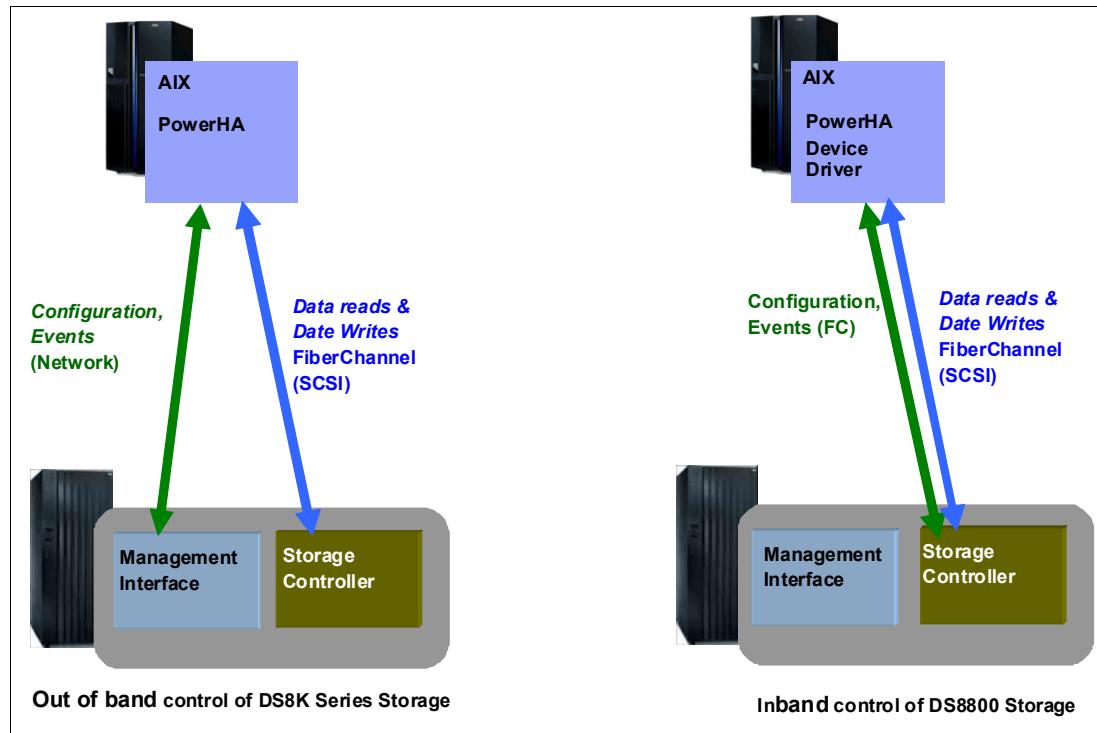


Figure 3-5 Out-of-band versus in-band storage control

Using in-band communication offers multiple benefits over the out-of-band model:

- ▶ Better performance - A TCP network is usually slower than the Fiber Channel infrastructure. Using a storage agent (storage HMC) for the CLI command execution causes higher delays in the command execution and event flow. As a direct result, a reduced failover time is achieved using the in-band communication.
- ▶ Facilitates tighter integration with the host SCSI disk driver.
- ▶ Ease of configuration - In-band control is embedded in the disk driver, so no PowerHA configuration is required for storage communication.
- ▶ Enhanced reliability - Disks on an AIX host can be easily mapped to their corresponding DS8000 volume IDs, resulting in enhanced RAS capability. Due to its tight integration with the host disk driver, it allows for more robust error checking.

In PowerHA 7.1.2, the following storage-based replication technologies are included:

- ▶ Out-of-band:
 - ESS/DS8000/DS6000™ PPRC
 - DS8700/DS8800 Global Mirror
 - XIV
 - SVC
- ▶ In-band
 - DS8800 in-band and HyperSwap

DS8000 Metro and Global Mirror

ESS/DS6000/DS8000 Metro Mirror and DS8000 Global Mirror are technologies supported with PowerHA Enterprise Edition.

ESS/DS replicated resources can use the following replication configurations:

- ▶ For Metro Mirror Replication (formerly PPRC):

- Direct Management (ESS 800)

This is the oldest configuration type used for IBM ESS storage systems. In this configuration, PowerHA directly manages the failover and resynchronization of the PPRC pairs by issuing commands directly to the ESS systems. PowerHA manages the PPRC resources by communicating with the Copy Services Server (CSS) on ESS systems via the ESS CLI.

- DSCLI management (ESS/DS6000/DS8000)

This type of configuration uses the DS CLI interface management for issuing commands to the storage system, via an Enterprise Storage Server® Network Interface (ESSNI) server on either storage controller or storage HMC.

DSCLI-based PowerHA SystemMirror supports more than one storage system per site. You can configure and use more than one DS storage on a single site. Each PPRC-replicated resource still has only one primary and one secondary storage per site, but you can use any of the configured DSs having a single one per site in a PPRC-replicated resource group.

- ▶ Global Mirror

This is an asynchronous replication technology using a Global Copy relationship between two storage systems and a flash copy relationship at the remote storage site, coordinated by a Global Mirror session that ensures consistent point-in-time data generated at the remote site. Currently DS8700 and DS8800 models are supported with PowerHA SystemMirror 7.1.2 using Global Mirror replicated resources.

Note: PowerHA Enterprise Edition with DSCLI Metro Mirror replicated resources supports Virtual I/O environments with vSCSI and NPIV disk attachment. For Global Mirror replicated resources, vSCSI attachment is not supported.

For more details on DS8000 copy services features refer, to the following publication:

IBM System Storage DS8000: Copy Services in Open Environments, SG24-6788-03 at:

<http://www.redbooks.ibm.com/abstracts/sg246788.html>

For DS8700/DS8800 specific information, refer to:

IBM System Storage DS8000: Architecture and Implementation, SG24-8886-02

<http://www.redbooks.ibm.com/abstracts/sg248886.html>

IBM SAN Volume Controller (SVC)

PowerHA Enterprise Edition with SVC-based replication provides a fully automated, highly available disaster recovery management solution by taking advantage of SVC's ability to provide virtual disks derived from varied disk subsystems.

SVC hardware supports only FC protocol for data traffic inside and between sites. It requires a SAN switched environment. FCIP routers can also be used to transport the Fiber Channel (FC) data frames over a TCP/IP network between the sites.

Management of the SVC replicated pairs is performed using *ssh* over a TCP/IP network. Each cluster node needs to have the *openssh* package installed and configured to access the SVCs in both sites.

PowerHA Enterprise Edition using SVC replication supports the following options for data replication between the SVC clusters:

- ▶ Metro Mirror providing synchronous remote copy
Changes are sent to both primary and secondary copies, and the write confirmation is received only after the operations are complete at both sites.
- ▶ Global Mirror providing asynchronous replication
Global Mirror periodically invokes a point-in-time copy at the primary site without impacting the I/O to the source volumes. Global Mirror is generally used for greater distances and complements the synchronous replication facilities. This feature was introduced in SVC Version 4.1.

For PowerHA Enterprise Edition integration, SAN Volume Controller code Version 4.2 or later is required. At the time of writing, the latest supported SVC code version is 6.4.

Note: PowerHA using SVC/V7000 replicated resources can be used in an environment using Virtual I/O resources and supports both vSCSI and NPIV disk attachment.

At this time SVC and Storwize® V7000 storage systems use the same version of code (v6.x). Storwize V7000 replication is also supported with PowerHA Enterprise Edition under the same circumstances as SVC replication. A mixed environment using SVC and V7000 is also supported.

Note: In a mixed SVC/V7000 replication environment, V7000 storage needs to be changed from the default “storage” mode of interoperation with an SVC to the “replication” mode. This can be accomplished only using the CLI by issuing the command:

```
chsystem -layer replication
```

For more details on PowerHA Enterprise Edition with SVC replicated resources and for practical implementation examples refer to Chapter 6, “Configuring PowerHA SystemMirror Enterprise Edition linked cluster with SVC replication” on page 195.

IBM XIV Remote Mirroring

PowerHA Enterprise Edition supports XIV Remote Mirroring technology. The XIV Remote Mirror function of the IBM XIV Storage System enables a real-time copy between two or more storage systems over Fibre Channel or iSCSI links. This function provides a method to protect data from site failures for both synchronous and asynchronous replication.

XIV enables a set of remote mirrors to be grouped into a consistency group. When using synchronous or asynchronous mirroring, the consistency groups handle many remote mirror pairs as a group to make mirrored volumes consistent. Consistency groups simplify the handling of many remote volume pairs because you do not have to manage the remote volume pairs individually.

PowerHA using XIV replicated resources requires the use of consistency groups for both types of replication. Plan to allocate the volumes on the XIV systems in consistency groups according to the application layout. All the volumes of an application should be part of the same consistency group. Remember when creating the consistency groups that the same

name must be used for a set of XIV replicated pairs on both XIV systems for proper integration with PowerHA.

PowerHA with XIV replicated resources can be implemented in an environment using virtual storage resources. Disk volumes in a Virtual I/O server environment can be attached using NPIV. vSCSI disks are not supported.

For more details about PowerHA Enterprise Edition with XIV replicated resources and for practical implementation examples, refer to Chapter 7, “Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replication” on page 253.

DS8800 in-band communication and HyperSwap

PowerHA SystemMirror 7.1.2 Enterprise Edition and later can use in-band communication for management of the replication pairs on a DS8800 storage system. This functionality is enabled by the following components:

- ▶ DS8800 storage system with the appropriate level of firmware
- ▶ AIX MPIO with AIX 61TL8/AIX 71TL2. SDDPCM is not supported
- ▶ PowerHA 7.1.2 Enterprise Edition, or later

The in-band communication and HyperSwap features are configured using common *smit* panels. The following storage-related considerations apply for an environment using DS8000 in-band resources and HyperSwap:

- ▶ Currently only the DS8800 model is supported. The minimum code bundle required is 86.30.49.0.
- ▶ DS8800 can be attached to the cluster nodes using FC, NPIV or FCoE. vSCSI is not supported.
- ▶ HyperSwap is supported only with Metro Mirror replication between the DS8800 storage systems. DS8800 requires the Metro Mirror license to be applied on both storage systems.

Note: Metro/Global Mirror is not supported at this time with PowerHA Enterprise Edition. Also the HyperSwap functionality cannot be used with volumes on a storage, part of the Metro/Global Mirror relationships.

- ▶ The storage systems need to be accessed from all nodes in both sites. Hardware connectivity and SAN zoning configuration needs to allow a cluster node to access the local storage in the local site and the secondary storage in the remote site.
- ▶ SCSI Reservations are not supported with HyperSwap disks. AIX hdisks must be configured with the *reserve_policy* attribute set to *no_reserve*.
- ▶ The host profile for the volumes attached to the cluster nodes must be set to “IBM pSeries - AIX with Powerswap”. You can check this setting using the **lshostconnect dscli** command. See Example 3-4.

Example 3-4 Host type attachment required for HyperSwap

```
dscli> lshostconnect  
Date/Time: December 22, 2012 7:51:44 AM CST IBM DSCLI Version: 6.6.0.305 DS:  
IBM.2107-75TL771
```

Name	ID	WWPN	HostType	Profile
portgrp	volgrpID	ESSIOport		

r9r2m12_fcs0	0008 1000000C951075D -	IBM pSeries - AIX with
Powerswap support	46 V4	all

- Storage level Peer-to-Peer Remote Copy (PPRC) relationships and PPRC paths must be defined before you configure HyperSwap for PowerHA SystemMirror. Plan for the volumes and the logical subsystem (LSS) assignment. The LSS is the first two bytes of the LUNID on a DS8000 storage. You can check the LSS and LUNID of a volume in the storage using the *dscli* CLI tool and the **lsfbvol** command. See Example 3-5.

Example 3-5 lsfbvol output example

```
dscli> lsfbvol
Date/Time: December 22, 2012 6:22:09 AM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771
Name           ID   accstate datastate configstate deviceMTM datatype extpool
cap (2^30B)  cap (10^9B) cap (blocks)
=====
=====
r1r4m27-m38_1  B000 Online  Normal  Normal    2107-900 FB 512  P0
2.0          -     4194304
r1r4m27-m38_2  B001 Online  Normal  Normal    2107-900 FB 512  P0
2.0          -     4194304
r1r4m27-m38_caa B200 Online  Normal  Normal    2107-900 FB 512  P0
2.0          -     4194304
r1r4m27-m38_rvg B401 Online  Normal  Normal    2107-900 FB 512  P0
2.0          -     4194304
.......
```

We suggest allocating different LSSs for the HyperSwap storage resources than other replicated volumes not being part of the HyperSwap. Also take into consideration to allocate distinct LSSs between the application, system and CAA repository disks, because they will be included in different PowerHA mirror groups.

Take into consideration that suspended operations on DS8800 must function on the entire LSS. If a single DS8800 LSS contains PPRC volumes from more than one application and if one of the replication connections breaks, all PPRC paths are removed. If the applications are not managed by PowerHA SystemMirror, the PPRC paths must be manually recreated after the replication connection is reestablished.

Note: The DS8800 in-band communication can be used without HyperSwap enablement for DS8800 Metro Mirror replicated resources. This case is similar to the DSCLI PPRC case, except that the cluster nodes communicate with the storage system for managing the PPRC pairs using the SAN FC infrastructure used for disk I/O, rather than using *dscli* and the IP network connectivity.

For more details about HyperSwap functionality and practical implementation examples, refer to Chapter 4, “Implementing DS8800 HyperSwap” on page 59.

Storage and Firmware requirements

Table 3-1 on page 46 summarizes the particular types of replication technologies currently supported with PowerHA SystemMirror 7.1.2 Enterprise Edition and the storage software related requirements.

Table 3-1 Storage technology replication supported with PowerHA 7.1.2

Storage type	Type of replication environment	Firmware/management software requirements
DS8000/DS6000/ESS	Metro Mirror	ibm2105cli.rte 32.6.100.13(ESS) ibm2105esscli.rte 2.1.0.15(ESS) DSCLI version 5.3.1.236 or later(DS6K/DS8K)
DS8700/DS8800	Global Mirror	Minimum levels for DS8700: Code bundle 75.1.145.0 or later DSCLI version 6.5.1.203 or later
SVC/V7000	Metro/Global	SVC code 4.2, or later (Currently 6.3 is the latest supported version) SVC code 6.2/6.3 (V7000 storage) openssh version 3.6.1 or later (for access to SVC interfaces)
XIV- Gen2/Gen3 models	Sync/Async	System Firmware 10.2.4 (Gen2) System Firmware 11.0.0a (Gen3) XCLI 2.4.4 or later
DS8800	DS8000 in and/ HyperSwap	DS8800 Code bundle 86.30.49.0, or later

We suggest using the latest supported version of firmware and storage management software for integration with PowerHA Enterprise Edition.

Note that IBM intends to support the following storage-based mirroring in upcoming releases of the PowerHA SystemMirror software solution. While management interfaces might be visible for these HADR solutions, their support is still pending qualification:

- ▶ PowerHA SystemMirror Enterprise Edition for EMC Symmetrix Remote Data Facility (SRDF)
- ▶ PowerHA SystemMirror Enterprise Edition for Hitachi TrueCopy and Universal Replicator (HUR)

Firmware levels can be verified using the management tools GUI/CLI for that particular storage type. Table 3-2 provides the CLI subcommands that can be used to get the current firmware level for the currently supported storage with PowerHA SystemMirror 7.1.2 Enterprise Edition.

Table 3-2 Getting the firmware level of the storage system

Storage system	CLI tool	Subcommand
DS6000/DS8000	dscli	ver -l
SVC	ssh	svcinfo lscluster
XIV	xcli	version_get

In case of a DS8000 storage system, dscli can be used for querying the current firmware version of the storage system using the command **ver -l** as shown in Example 3-6.

Example 3-6 DS8800 verifying the code version

```
dscli> lssi
Date/Time: December 13, 2012 8:12:03 PM CST IBM DSCLI Version: 6.6.0.305 DS: -
```

```

Name ID           Storage Unit     Model WWNN          State ESSNet
=====
DS8K4 IBM.2107-75TL771 IBM.2107-75TL770 951   500507630AFFC16B Online Enabled

dscli> ver -l
Date/Time: December 13, 2012 8:14:26 PM CST IBM DSCLI Version: 6.6.0.305 DS: -
DSCLI      6.6.0.305
StorageManager 7.7.3.0.20120215.1
=====Version=====
Storage Image    LMC
=====
IBM.2107-75TL771 7.6.30.160

```

Status of the replication relationships and the MANUAL recovery option

PowerHA Enterprise Edition using storage-based replication provides the possibility for using manual recovery of the storage replicated resources in certain conditions.

When defining the replicated resources, two recovery options are available:

- ▶ AUTO - Involves automatic recovery of the storage-replicated resource during the site failover. The PowerHA software performs the activation of the mirrored copy at the recovery site, and brings up the resource group at the remote site.
- ▶ MANUAL - User action is required at site failover time. The cluster will not automatically bring up the replicated resources and the related resource groups. User intervention is required to manually recover the replicated volumes at the recovery site. This option is taken into consideration only in certain situations, such as in a replication link down case. At the time of site failover, PowerHA software checks the replication status. If the replication is found in a normal state, the cluster performs an automatic failover of the associated resource group in the recovery site.

The status of the replicated pairs determines whether the MANUAL recovery action prevents a failover. The states vary between the replication types. The states shown in Table 3-3 do not failover automatically during a site failover.

Table 3-3 Replication states for the MANUAL option

Replication type	Replication state
DS Metro Mirror	Target-FullDuplex-Source-Unknown Source-Suspended-Source-Unknown Source-Unknown-Target-FullDuplex Source-Unknown-Source-Suspended
SVC Metro/Global Mirror	idling_disconnected consistent_disconnected
XIV Remote Mirror	Unsynchronized RPO lagging

3.1.4 Cluster repository disk

PowerHA SystemMirror uses a shared disk to store Cluster Aware AIX (CAA) cluster configuration information. You must have at least 512 MB and no more than 460 GB of disk space allocated for the cluster repository disk. This feature requires that a dedicated shared disk be available to all nodes that are part of the cluster. This disk cannot be used for application storage or any other purpose.

When planning for a repository disk in case of a multi-site cluster solution:

- ▶ Stretched cluster

Requires and shares only one repository disk. When implementing the cluster configuration with multiple storages in different sites, consider allocating the CAA repository and the backup repositories in different storages across the sites for increasing the availability of the repository disk in case of a storage failure. As example, when using a cross-site LVM mirroring configuration with a storage subsystem in each site, you can allocate the primary disk repository in Site1 and the backup repository on the storage in Site2.

- ▶ Linked clusters

Requires a repository disk to be allocated to each site. If there is no other storage at a site, plan to allocate the backup repository disk on a different set of disks (other arrays) within the same storage for increasing the repository disk availability in case of disk failures.

Considerations for a stretched cluster

This section describes some considerations when implementing a stretched cluster:

- ▶ There is only *one* repository disk in a stretched cluster.
- ▶ Repository disks *cannot* be mirrored using AIX LVM. Hence we strongly advise to have it RAID protected by a redundant and highly available storage configuration.
- ▶ All nodes must have access to the repository disk.
- ▶ In the event the repository disk *fails* or becomes *inaccessible* by one or more nodes, the nodes stay online and the cluster is still able to process events such as node, network or adapter failures, and so on. Upon failure, the cluster *ahaFS event REP_DOWN* occurs. However, no cluster *configuration changes* can be performed in this state. Any attempt to do so will be stopped with an error message.
- ▶ A backup repository disk can be defined in case of a failure. When planning the disks that you want to use as repository disks, you must plan for a backup or replacement disks, which can be used in case the primary repository disk fails. The backup disk must be the same size and type as the primary disk, but could be in a different physical storage. Update your administrative procedures and documentation with the backup disk information. You can also replace a working repository disk with a new one to increase the size or to change to a different storage subsystem. To replace a repository disk, you can use the SMIT interface or PowerHA SystemMirror for IBM Systems Director. The cluster *ahaFS event REP_UP* occurs upon replacement.

Additional considerations for linked clusters

This section provides additional considerations when implementing linked clusters:

- ▶ The nodes *within a site* share a common repository disk with all its characteristics specified earlier.
- ▶ The repositories between sites are kept in sync internally by CAA.

When sites *sunder* and then are *merged*, CAA provides a mechanism to reconcile the two repositories. This can be done either through reboot (of all the nodes on the losing side) or through APIs implemented exclusively for RSCT; see 10.4.2, “Configuring the split and merge policy” on page 455.

3.1.5 Tie breaker disk

The tie breaker disk concept was introduced in PowerHA 7.1.2 as an additional mechanism for preventing cluster partitioning. In case of a long distance cluster where nodes

communicate between the sites using only IP networks, the risk of cluster partitioning by the network links failing can be mitigated by using a tie breaker disk in a tertiary location accessible by both sites. Although it proves to be more efficient in the linked clusters, the tie breaker disk can be used in both stretched and linked cluster types.

Requirements for a tie breaker

The following requirements and restrictions apply for a tie breaker:

- ▶ A disk supporting SCSI-3 persistent reserve. Such a device can include a Fibre Channel, FCoE or i-SCSI attachment supporting the SCSI-3 persistent reservation.
- ▶ Disk device must be accessible by all nodes.
- ▶ The repository disk cannot be used as tie breaker.

Note: There is no specific capacity or bandwidth access required for the tie breaker disk. PowerHA SystemMirror software does not store any configuration data on it or use it for heartbeating. It is only used with SCSI-3 persistent reservation for handling site split/merge events.

Where to locate the tie breaker disk

The tie breaker disk should be located where the sites can access it in any situation including total site failures. Ideally, this should be configured to a different site from where the cluster nodes are being activated.

Figure 3-6 illustrates this design.

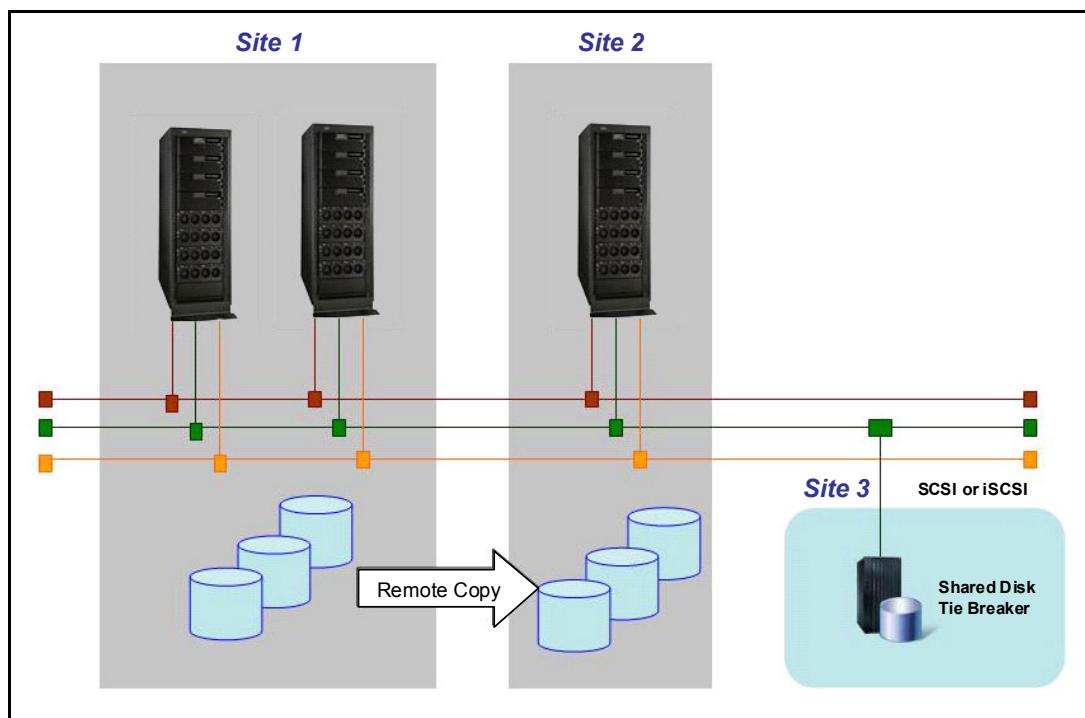


Figure 3-6 Tie breaker disk location

Keep in mind that the AIX nodes must be able to initiate SCSI commands to the tie breaker disk to gain the reservation.

The simplest way to accomplish this is to configure a physical fiber channel access to where the tie breaker disk is located. However, in many cases this is not realistic. Instead, methods that imitate SCSI commands through IP networks should suit users.

The following are examples of technology that can be used for connecting the tie breaker disk:

- ▶ Fiber channel (FC) using direct attachment or FCIP
- ▶ Fibre Channel over Ethernet (FCoE)
- ▶ Internet Small Computer Systems Interface (iSCSI)

Important: The current AIX iSCSI device driver does not support SCSI-3 persistent reservations. IBM PowerHA development is currently working to remove this limitation, but at the time of writing, iSCSI devices could not be used for tie breaker disks.

For further details regarding the tie breaker and site split/merge handling, see Chapter 10, “Cluster partition management” on page 449.

3.2 Hardware and software requirements

This section describes the hardware and software requirements for the IBM PowerHA SystemMirror for AIX solution.

3.2.1 Hardware requirements for the AIX solution

PowerHA SystemMirror 7.1.2 can be installed on any hardware supported by AIX 6.1 TL8 SP2 or AIX 7.1 TL2 SP2. For supported IBM server models, extension cards, storage subsystems, SAN Volume Controller, NAS, Adapters, refer to the PowerHA SystemMirror hardware support matrix at:

<http://www-03.ibm.com/support/techdocs/atstr.nsf/WebIndex/TD105638>

For the list of supported Fibre Channel adapters for SAN heart beating, refer to the “Setting up cluster storage communication” page at:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.cluster_aware/claware_comm_setup.htm

PowerHA supports full system partitions as well as fully virtualized LPARs. You can use a mix of physical and virtual interfaces for cluster communication.

3.2.2 Software requirements for PowerHA 7.1.2

This section provides the software requirements for PowerHA 7.1.2.

Base requirements for PowerHA 7.1.2

PowerHA SystemMirror 7.1.2 requires the following minimum levels of software on the nodes:

- ▶ AIX 6.1 TL8 or AIX 7.1 TL2
- ▶ RSCT 3.1.4

Since the AIX installation might be a new or a migrated one, check also that the fileset *bos.cluster:rte* is at 6.1.8 level for the AIX 6.1 case or at 7.1.2 level for the case of AIX 7.1.

The latest service pack for PowerHA SystemMirror software should be installed for all components. At the time of writing, we are referring to Service Pack 1 as the latest SP, available as of November 2012. We suggest installing the latest service pack (SP) for the AIX operating system. You can download the latest SP for PowerHA SystemMirror and for AIX from the IBM FixCentral page at:

<http://www-933.ibm.com/support/fixcentral/>

The IBM PowerHA SystemMirror Enterprise Edition requires the installation and acceptance of license agreements for both the Standard Edition cluster.license fileset and the Enterprise Edition cluster.xd.license fileset as shown in Table 3-4, in order for the remainder of the filesets to install.

Table 3-4 PowerHA Enterprise Edition - required fileset

Required package	Filesets to install
Enterprise Edition license	cluster.xd.license

The base filesets in the Standard Edition are required to install the Enterprise Edition filesets. The Enterprise package levels must match those of the base runtime level (cluster.es.server.rte). Table 3-5 displays the itemized list of filesets for each of the integrated offerings.

Table 3-5 PowerHA Enterprise Edition - integrated offering solution filesets

Replication type	Fileset to install
ESS-Direct Management PPRC	cluster.es.pprc.rte cluster.es.pprc.cmds cluster.msg.en_US.pprc
ESS/DS6000/DS8000 Metro Mirror (DSCLI PPRC)	cluster.es.spprc.cmds cluster.es.spprc.rte cluster.es.cgpprc.cmds cluster.es.cgpprc.rte cluster.msg.en_US.cgpprc
SAN Volume Controller (SVC)	cluster.es.svcpprc.cmds cluster.es.svcpprc.rte cluster.msg.en_US.svcpprc
XIV, DS8800 in-band and HyperSwap, DS8700/DS8800 Global Mirror	cluster.es.genxd.cmds cluster.es.genxd.rte cluster.msg.en_US.genxd
Geographic Logical Volume Mirroring	cluster.doc.en_US.glvm.pdf cluster.msg.en_US.glvm cluster.xd.glvm <i>glvm.rpv.client (part of AIX base install)</i> <i>glvm.rpv.man.en_US (part of AIX base install)</i> <i>glvm.rpv.msg.en_US (part of AIX base install)</i> <i>glvm.rpv.server (part of AIX base install)</i> <i>glvm.rpv.util (part of AIX base install)</i>
EMC SRDF (see Note)	cluster.es.sr.cmds cluster.es.sr.rte cluster.msg.en_US.sr
Hitachi TrueCopy/Universal Replicator (see Note)	cluster.es.tc.cmds cluster.es.tc.rte cluster.msg.en_US.tc

Note: Although the current package of PowerHA 7.1.2 includes the filesets for Hitachi and EMC replication technologies, IBM intends to provide support in upcoming releases of the PowerHA SystemMirror software solution.

Additional software requirements

The following optional filesets can be installed:

- ▶ devices.common.IBM.storfwk (for SAN-based heartbeating)
- ▶ cas.agent (optional, used for IBM Systems Director plug-in)
- ▶ clic.rte (for secure encryption communication option of clcomd)

The following additional software requirements apply:

- ▶ PowerHA SystemMirror Enterprise Edition and HyperSwap support requires PowerHA SystemMirror 7.1.2 Service Pack 1 with APAR IV27586.
- ▶ IBM Systems Director plug-in for PowerHA SystemMirror 7.1.2 has been certified with IBM Systems Director version 6.3.1.
- ▶ In a virtual environment, use VIOS 2.2.0.1-FP24 SP01 or later.

Visit the IBM FixCentral website for all available service packs for AIX, Virtual I/O server, PowerHA SystemMirror, RSCT and Systems Director, at:

<http://www.ibm.com/support/fixcentral/>

Multipath software requirements

The multipath software is required for redundant access to the storage system using the SAN communication paths. You have to consider the appropriate device driver option for a specific environment.

When using PowerHA SystemMirror 7.1.2 Enterprise Edition 7.1.2 storage-replicated resources, you have the following software multipath options:

- ▶ Native AIX MPIO driver

AIX 6.1 TL8 and 7.1 TL2 have been enhanced to support DS8800 in-band communication and HyperSwap features. The XIV FC storage attachment is also natively supported by the AIX MPIO driver with no additional software needed. This driver supports two options: active/passive (AIX_AAPCM) used for the active/passive controllers such as DS4000/5000 storage models and active/active (AIX_AAPCM), currently supported with DS8000 models and XIV storage systems.

- ▶ Subsystem Device Driver (SDD)

This is the legacy multipath driver for ESS/DS6000/DS8000/SVC. When using SDD an hdisk# is created for each path to a LUN and a pseudo device named vpath# is created for multipath I/O access, which is used for AIX LVM operations.

- ▶ Subsystem Device Driver MPIO (SDDPCM)

This is an extension to AIX MPIO designed for ESS, DS6000, DS8000, and SVC and also for DS4000, DS5000, and DS3950 storage systems. With SDDPCM, a single hdisk# is presented by AIX for use in the LVM operations.

At the time of writing, the support for integrating EMC and Hitachi storage replications with PowerHA 7.1.2 is pending qualification. In such cases, vendor-specific multipath software might be required. For example, EMC storage uses Powerpath software for AIX.

Multiple driver options can apply for a particular environment. As example for an environment using DS8000 dscli metro mirror replicated resources, you can use either SDD or SDDPCM. You must use a consistent multipath driver option across the cluster nodes.

Note: The SDD and SDDPCM multipath drivers cannot coexist on the same server.

The following SDD software levels are required:

- ▶ IBM 2105 Subsystem Device Driver (SDD): ibmSdd_510nchacmp.rte 1.3.3.6 or later
- ▶ IBM SDD FCP Host Attachment Script: devices.fcp.disk.ibm.rte version 1.0.0.12 or later
- ▶ IBM Subsystem Device Driver (SDD): devices.sdd.XX.rte version 1.6.3.0 or later (where XX corresponds to the associated level of AIX); see Table 3-6.

The latest AIX SDD levels available at the time of writing that we suggest to install, and the supported operating system, are shown in Table 3-6.

Table 3-6 SDD and operating system levels

AIX OS level	ESS 800	DS8000	DS6000	SVC
AIX 6.1	1.7.2.1	1.7.2.8	1.7.2.3	1.7.2.5

For AIX 7.1 use AIX MPIO or SDDPCM.

To obtain the latest version of this driver, go to:

<http://www.ibm.com/servers/storage/support/software/sdd/>

The following SDDPCM prerequisite software and microcode levels are required:

- ▶ IBM Multipath Subsystem Device Driver Path Control Module (SDDPCM): devices.sddpcm.XX.rte version 2.2.0.0 or later (where XX corresponds to the associated level of AIX); see Table 3-7.

The host attachment for SDDPCM adds 2105, 2145, 1750, and 2107 device information to allow AIX to properly configure 2105, 2145, 1750, and 2107 hdisks. The device information allows AIX to:

- ▶ Identify the hdisk as a 2105, 2145, 1750, or 2107 hdisk.
- ▶ Set default hdisk attributes such as queue_depth and timeout values.
- ▶ Indicate to the AIX device driver configure method to configure these hdisks as MPIO-capable devices.

The AIX SDDPCM levels available at the time of writing, which we suggest to install, and the supported operating systems, are shown in Table 3-7.

Table 3-7 SDDPCM and operating system levels

AIX OS level	ESS	DS8000	DS6000	SVC
AIX 6.1	2.2.0.4	2.6.3.2	2.4.0.2	2.6.3.2
AIX 7.1	N/A	2.6.3.2	N/A	2.6.3.2

Note: Persistent Reservation with PowerHA SystemMirror 7.1 is not supported. Shared volume groups managed by PowerHA SystemMirror and accessed through SDDPCM must be set in enhanced concurrent mode.

For the latest version of SDDPCM for AIX, refer to:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S4000201>

In the case of DS8800 in-band configuration and HyperSwap, only the AIX PCM device driver is supported. The minimum required AIX level is AIX 6.1 TL8 or AIX 7.1 TL2. The `manage_disk_drivers` command was updated to support DS8000 storage as an option. The default option for DS8000 storage is NO_OVERRIDE, which uses the highest priority ODM mapping. Note that with this option, when SDDPCM is present, it has a higher priority than the AIX PCM (AIX_AAPCM) option.

To set AIX PCM as driver option, run:

```
manage_disk_drivers -d 2107DS8K -o AIX_AAPCM
```

To set NO_OVERRIDE as driver option, run:

```
manage_disk_drivers -d 2107DS8K -o NO_OVERRIDE
```

Note: A reboot is required for applying the changes to the driver option.

Refer also to the System Storage Interoperability Center (SSIC) for verifying the combination between the multipath software level, AIX, and storage firmware, at:

<http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss>

GLVM considerations for PowerHA Enterprise Edition

The following considerations apply for Geographical LVM mirroring:

- ▶ Sites

GLVM for PowerHA SystemMirror Enterprise Edition requires two PowerHA SystemMirror sites. Each PowerHA SystemMirror site must have the same name as the RPV server site name.

- ▶ Enhanced concurrent volume groups

In addition to non-concurrent volume groups, you can have enhanced concurrent mode volume groups configured with RPVs, so that they can serve as geographically mirrored volume groups. You can include such volume groups in both concurrent and non-concurrent resource groups in PowerHA SystemMirror.

Note: Enhanced concurrent volume groups can be accessed concurrently only on nodes within the same site. You cannot access enhanced concurrent mode geographically mirrored volume groups across sites concurrently. Fast disk takeover is not supported for remote disks that are part of a geographically mirrored volume group.

- ▶ Replication networks

In a PowerHA SystemMirror cluster that has sites configured, you can have up to four XD_data networks used for data mirroring. This increases data availability and mirroring performance. For instance, if one of the data mirroring networks fails, the GLVM data mirroring can continue over the redundant networks. Also, you have the flexibility to configure several low bandwidth XD_data networks and take advantage of the aggregate

network bandwidth. Plan the data mirroring networks so that they provide similar network latency and bandwidth. This is because, for load balancing, each RPV client communicates with its corresponding RPV server over more than one IP-based network at the same time (by sending I/O requests across each of the networks in a round-robin order).

- ▶ PowerHA SystemMirror lets you configure site-specific service IP labels, thus you can create a resource group that activates a given IP address on one site and a different IP address on another site. Site-specific IP labels are supported within different subnets, thus allowing you to have subnetting between sites. When using site-specific IP addresses you can also plan for DNS integration. At the time of site failover, a user-defined script can update the DNS entry associated with a unique application service name with the service IP address specific to the active site. See an example in Appendix B, “DNS change for the IBM Systems Director environment with PoweHA” on page 495.
- ▶ Asynchronous mirroring allows the local site to be updated immediately and the remote site to be updated as network bandwidth allows. The data is cached and sent later, as network resources become available. While this can greatly increase application response time, there is some inherent risk of data loss after a site failure due to the nature of asynchronous replication. Asynchronous mirroring requires AIX super-strict mirror pools.

For further details about PowerHA SystemMirror using GLVM, refer to:

http://pic.dhe.ibm.com/infocenter/aix/v6r1/topic/com.ibm.aix.powerha.geolvm/ha_g1vm_kick.htm

Required release of AIX and PowerHA SystemMirror for GLVM

GLVM data mirroring functionality is provided by the following filesets, which are available from the base AIX installation media:

glvm.rpv.client	Remote Physical Volume Client
glvm.rpv.server	Remote Physical Volume Server
glvm.rpv.util	Geographic LVM Utilities

The software requirements for PowerHA 7.1.2 also apply for GLVM. See “Base requirements for PowerHA 7.1.2” on page 50. For implementing GLVM for PowerHA SystemMirror Enterprise Edition we suggest using the latest version of PowerHA SystemMirror, AIX, and RSCT.

Refer also to the release notes file for the PowerHA version you are using, located in: /usr/es/sbin/cluster/release_notes_xd.

3.2.3 Capacity on Demand (CoD) support

PowerHA supports Dynamic LPAR, Capacity on Demand (CoD) processor, and memory resources. You can configure the minimum and desired number of processors, virtual processors, and memory for each application. When an application is activated on a node, PowerHA contacts the HMC to acquire this resource in addition to the current resources of the LPAR.

Table 3-8 on page 56 describes the types of Capacity on Demand (CoD) licenses that are available. It also indicates whether PowerHA supports the use of a particular license.

Table 3-8 Capacity on Demand (CoD) licenses

License Type	Description	PowerHA support	Comments
On/Off	CPU: Allows you to start and stop using processors as needs change. Memory: not allowed.	CPU: Yes Memory: N/A	PowerHA does not manage licenses. The resources remain allocated to an LPAR until PowerHA releases them through a DLPAR operation, or until you release them dynamically outside of PowerHA. If the LPAR node goes down outside of PowerHA, the CoD resources are also released.
Trial	CPU and Memory: The resources are activated for a single period of 30 consecutive days. If your system was ordered with CoD features and they have not yet been activated, you can turn the features on for a one-time trial period. With the trial capability, you can gauge how much capacity you might need in the future, if you decide to permanently activate the resources you need.	CPU: Yes Memory: Yes	PowerHA activates and deactivates trial CoD resources. Note: Once the resources are deactivated, the trial license is used and cannot be reactivated.



Part 2

Campus style disaster recovery (stretched clusters)

In this part we introduce the HyperSwap concept and implementation, and the cross-site LVM mirroring with the IBM PowerHA SystemMirror Standard Edition. The following topics are discussed in this part:

- ▶ Chapter 4, “Implementing DS8800 HyperSwap” on page 59.
- ▶ Chapter 5, “Cross-site LVM mirroring with IBM PowerHA SystemMirror 7.1.2 Standard Edition” on page 151.



Implementing DS8800 HyperSwap

This chapter includes the following topics for implementing HyperSwap with PowerHA SystemMirror 7.1.2 Enterprise Edition:

- ▶ Overview of HyperSwap
- ▶ Traditional disaster recovery
- ▶ HyperSwap failover
- ▶ The architecture
- ▶ HyperSwap functionalities
- ▶ Hardware and software requirements
- ▶ Limitations
- ▶ Considerations of implementation
- ▶ Test environment of a HyperSwap cluster
- ▶ Preparation of a HyperSwap-capable disk
- ▶ Initial configuration for the PowerHA cluster
- ▶ Swap demonstration for HyperSwap
- ▶ Oracle standalone database in a PowerHA HyperSwap environment

4.1 Overview of HyperSwap

HyperSwap is a function that provides continuous availability against storage errors. It is based upon storage-based synchronous replication. When directed (or upon disk errors), an AIX host accessing the primary disk subsystem would transparently switch over to the backup copy of the data such that consumers of the disks (such as middleware) are not affected. HyperSwap technology enables PowerHA SystemMirror to support the following capabilities for you:

- ▶ Eliminates primary disk subsystems as the single point of failure to provide the next level of continuous operations support within metro distances.
- ▶ Enables storage maintenance without any application downtime.
- ▶ Enables migration from old to new storage.
No disruption to the dependent applications.
- ▶ Natural extension of disaster recovery configurations.

4.2 Traditional disaster recovery

Figure 4-1 shows a traditional disaster recovery (DR) configuration. The storage is mirrored across the sites in either synchronous or asynchronous mode. The mirroring can be performed from either the host server via GLVM or via SAN replication. SAN-based replication is usually better for performance.

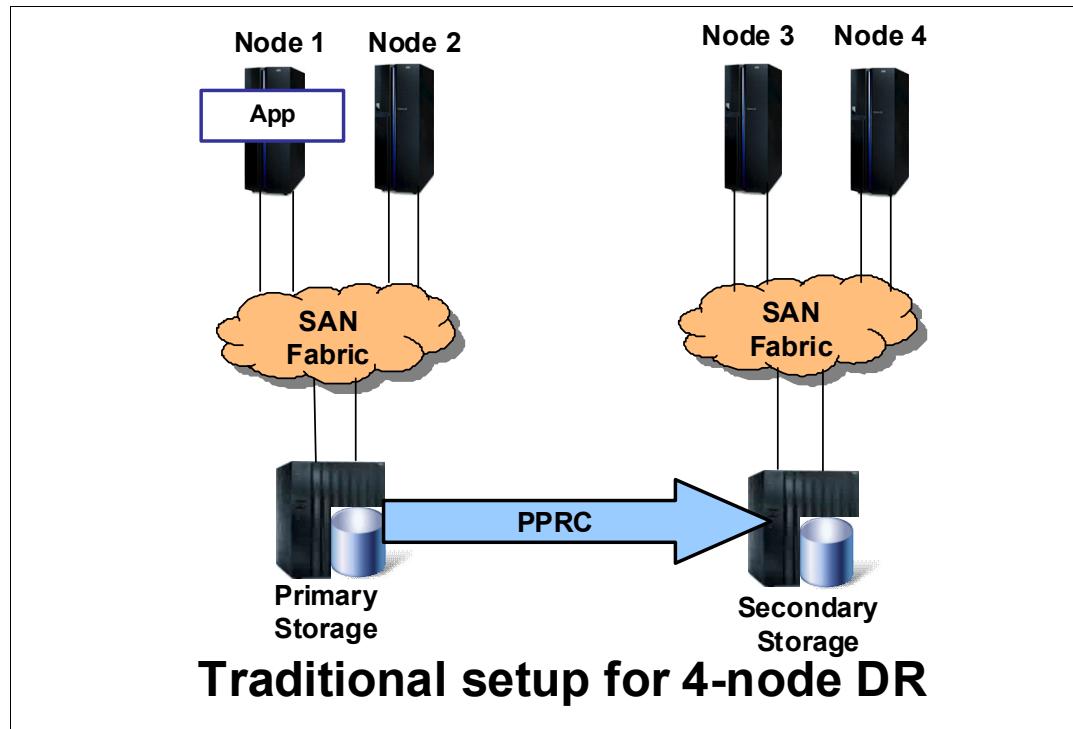


Figure 4-1 Traditional setup for 4-node disaster recovery

Figure 4-2 on page 61 shows the scenario when the primary storage is broken. Because there is no path from the primary node, Node 1, to the secondary storage, the application must move to siteB. Site failover is more complicated than a node failover within a site. It usually takes a few minutes longer for the service to recover.

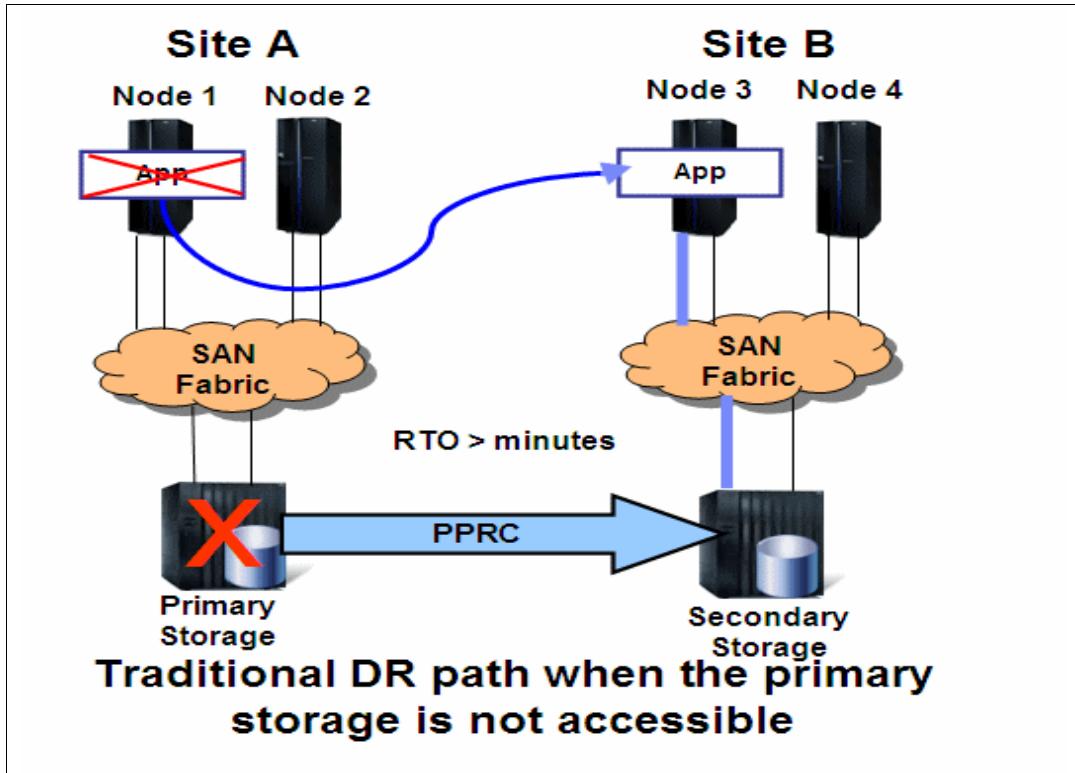


Figure 4-2 Traditional DR path when the primary storage is not accessible

4.3 HyperSwap failover

Figure 4-3 on page 62 shows a typical setup for HyperSwap, where the storage is mirrored by SAN-based utilities. Currently, only DS8800 storage with synchronous mirroring is supported.

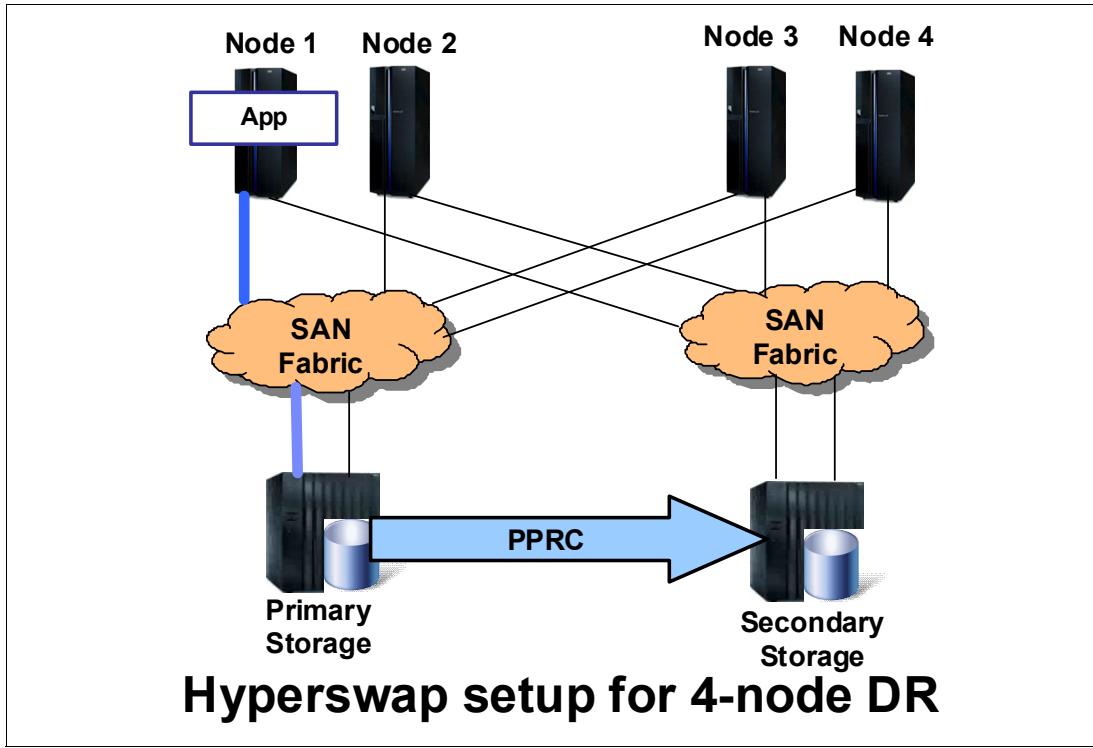


Figure 4-3 HyperSwap setup for 4-node DR

Figure 4-4 on page 63 shows the scenario when the primary storage is broken. Because all nodes in the HyperSwap cluster can access both primary and secondary storage, it does not result in a site failover. When the primary storage fails, AIX detects and reacts to the event by performing a PPRC failover. The application I/Os are transparently redirected to the secondary storage subsystem, thereby allowing the applications to continue running. Note that in this case, errors are detected by the AIX's SCSI disk drivers and a decision is made across multiple hosts to switch over to the secondary storage subsystem. For the duration of the HyperSwap swapping process, I/O is temporarily frozen. Note that during this time, applications would not experience failure, but instead experience non-operations delays. The duration of this I/O delay would take from two seconds to a minute, depending on the number of nodes and disks involved.

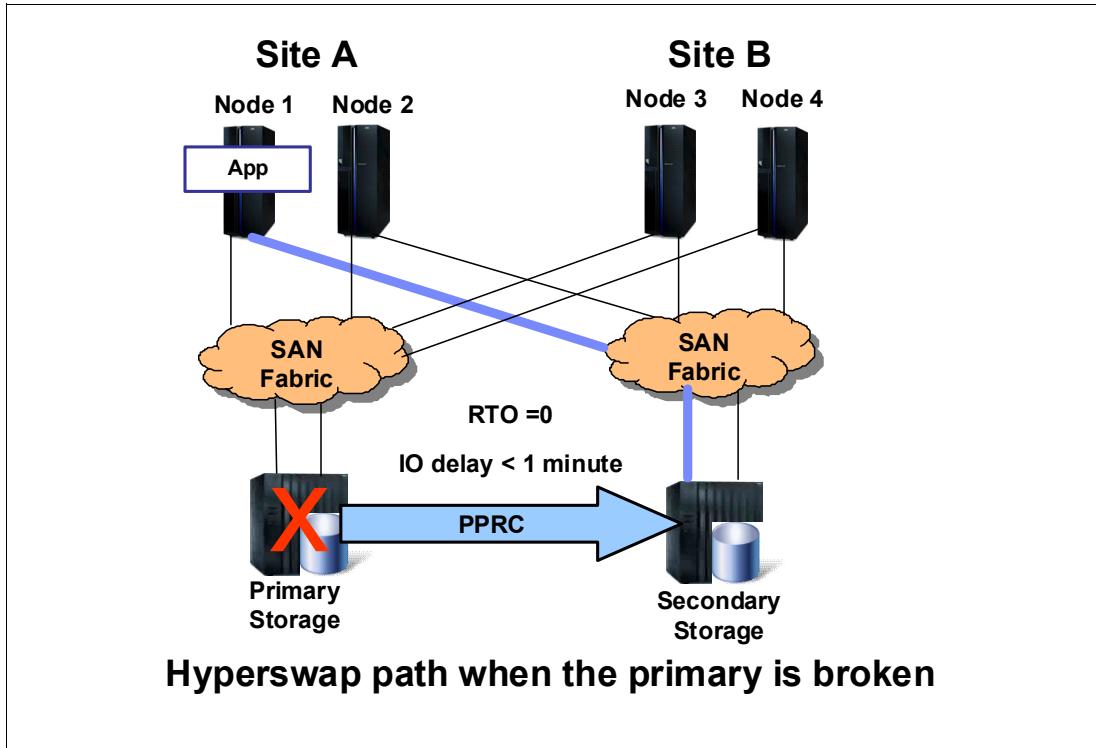


Figure 4-4 HyperSwap path when the primary storage is broken

Other scenarios of failure in the HyperSwap environment use HyperSwap whenever possible to minimize the impact to the application. For instance, Figure 4-5 on page 64 shows that the link between Node 1 and the primary storage is broken. Even though the primary disk is not responding, the application would not fail over to Node 2 as in the regular PowerHA environment. Instead, AIX of Node 1 redirects the I/O to the secondary storage. This process has less impact on the applications than node failover and site failover. In most situations, it is transparent to the application.

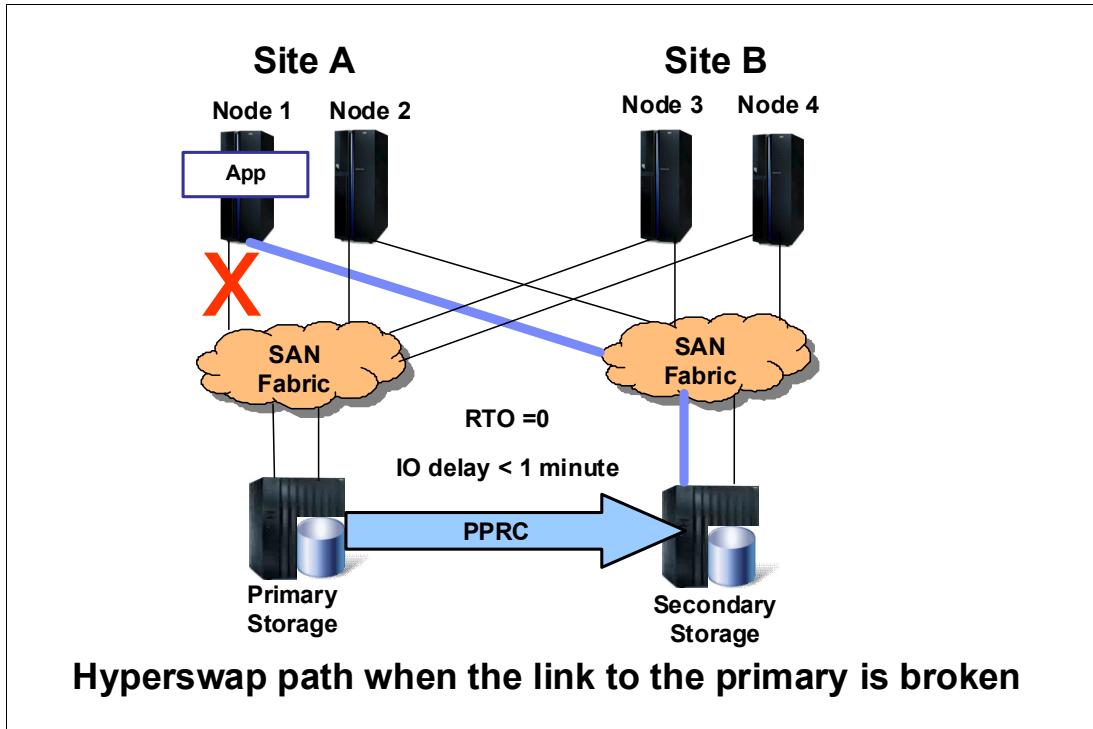


Figure 4-5 HyperSwap path when the link to the primary is broken

4.4 The architecture

In order to support HyperSwap, IBM has made changes in both the storage and AIX. The change in storage is the implementation of in-band communication, which is described in 4.4.1, “In-band storage management” on page 64. The AIX changes are described in 4.4.2, “AIX support for HyperSwap” on page 66.

4.4.1 In-band storage management

In order to support a more reliable, more resilient, and lower latency environment for storage systems capable of supporting HyperSwap, IBM developed in-band storage management to replace out-of-band storage management used in the traditional SAN storage environment. This in-band storage management infrastructure plays an important role, especially in clusters across sites.

Out-of-band and in-band storage management differences

Data path between host server and the storage controller is very critical to the reliability and performance of a storage system. Therefore, storage management usually uses a separate path to place the storage commands for storage actions and messages. This path is usually via a TCP/IP network. This kind of storage management is called out-of-band, as shown in Figure 4-6 on page 65.

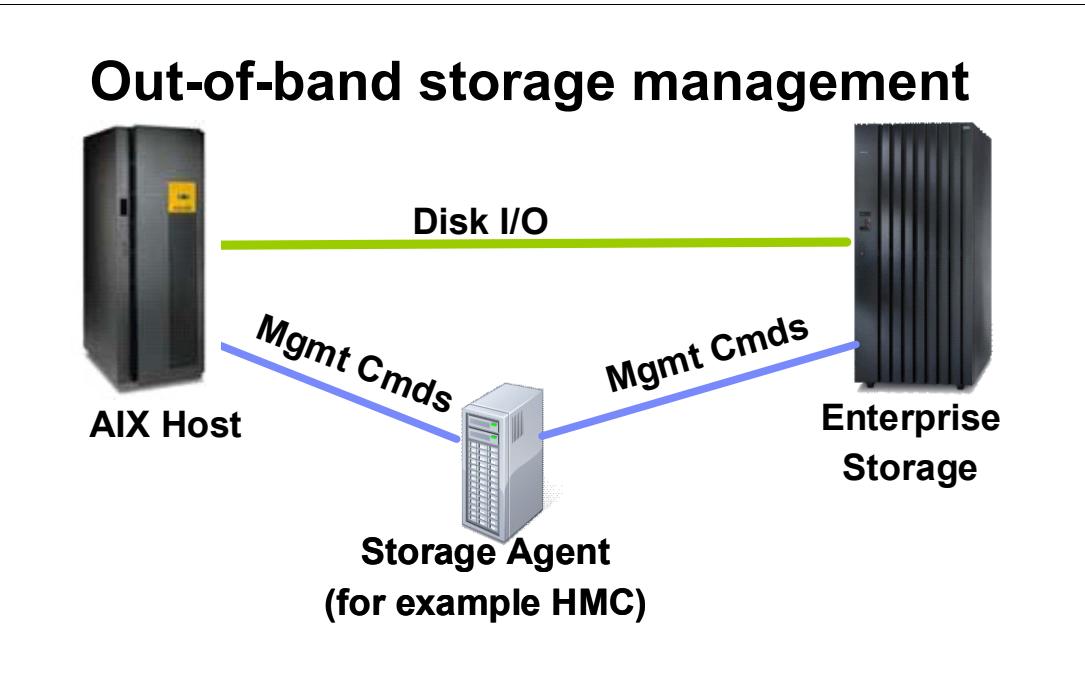


Figure 4-6 Out-of-band storage system

As the storage system grows bigger, out-of-band architecture becomes inadequate for the following reasons:

- ▶ The original consideration of moving the storage management communication out of the data path to eliminate the impact on performance of the critical data throughput. This consideration becomes a lower issue as the bandwidth of the data path grows significantly.
- ▶ As the SAN network spans a longer distance, the reliability and latency of the TCP/IP network becomes an issue.

Therefore, it becomes necessary to replace the TCP/IP network for storage management to support a larger area of storage systems. An in-band communication structure is best suited for this purpose. Figure 4-7 on page 66 shows an in-band storage system.

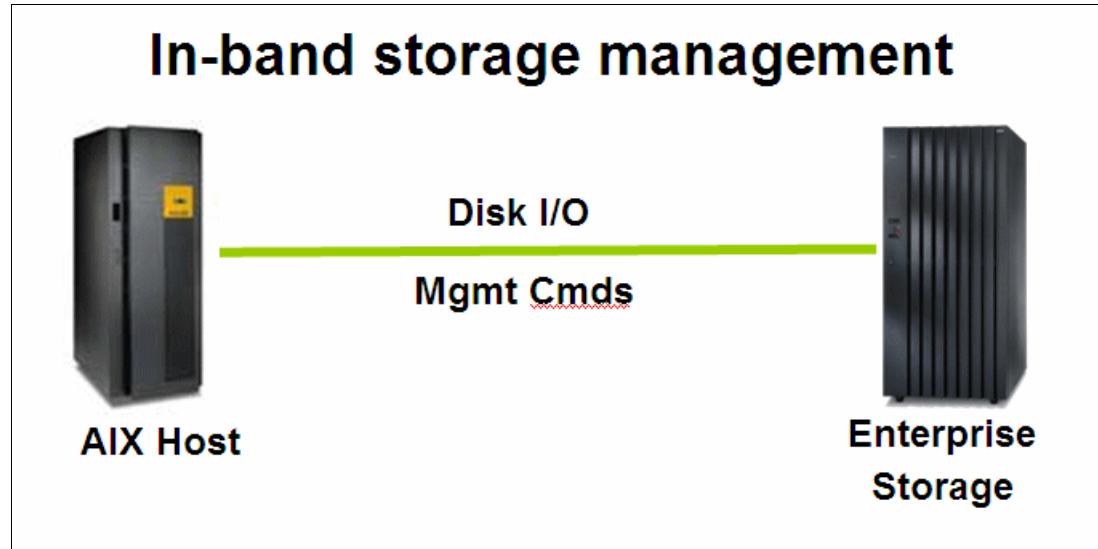


Figure 4-7 In-band storage system

Both data and storage management share the same fiber channel (FC) network. The advantages are:

- ▶ The FC network is usually faster than a TCP network.
- ▶ The HMC, used in the out-of-band structure as an agent, is no longer needed. The communication between host server and storage controller becomes more direct, more reliable, and faster.

4.4.2 AIX support for HyperSwap

Figure 4-8 on page 67 shows the diagram of the components supporting HyperSwap.

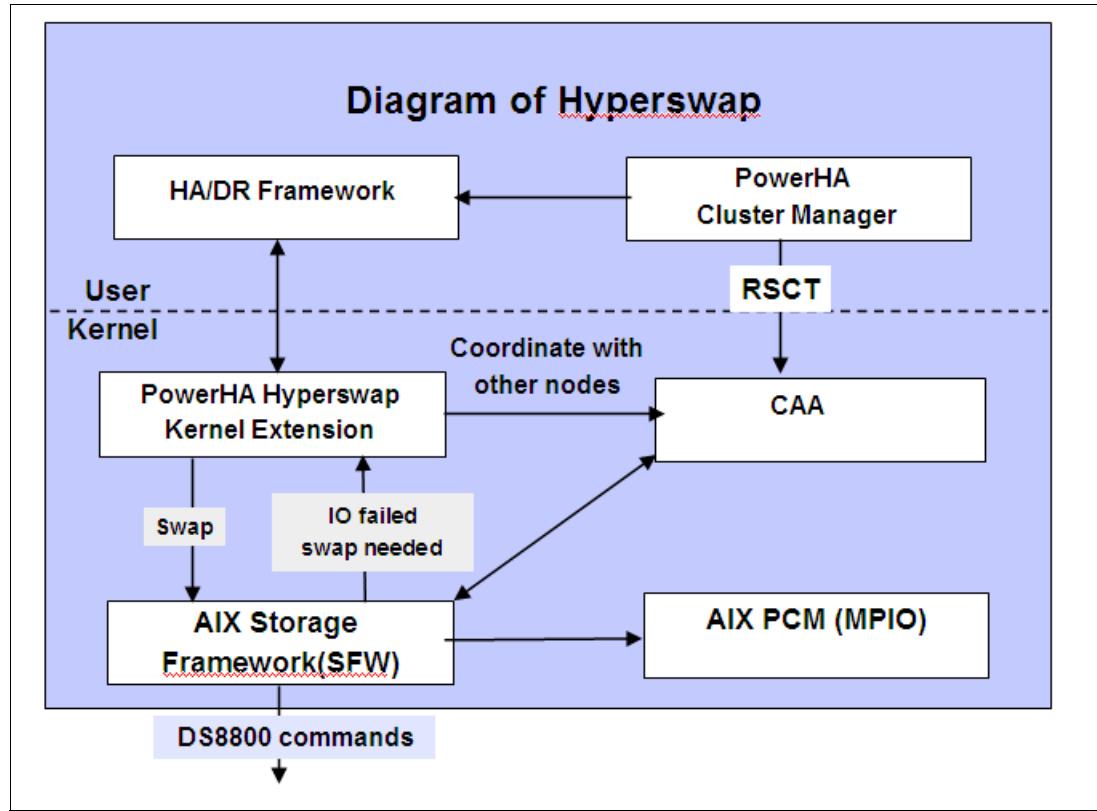


Figure 4-8 Diagram of HyperSwap

HyperSwap-related components are:

- ▶ Cluster Aware AIX (CAA)
 - Orchestrates® clusterwise actions.
- ▶ PowerHA HyperSwap kernel extension
 - Works with CAA to coordinate actions with other nodes.
 - Analyzes the messages from the PowerHA and AIX storage frameworks (SFW) and takes proper actions.
 - Determines the swap action.
- ▶ AIX Storage Framework (SFM)
 - Works as AIX interface to the storage.
 - Works closely with the PowerHA HyperSwap kernel extension.
 - Manages the status of the storage.
 - Informs the PowerHA HyperSwap kernel extension about I/O errors.
 - Gets swap decisions from the PowerHA HyperSwap kernel extension and sends orders to AIX PCM (MPIO).

4.4.3 AIX view of HyperSwap disks

Figure 4-9 shows the AIX view of the HyperSwap disks.

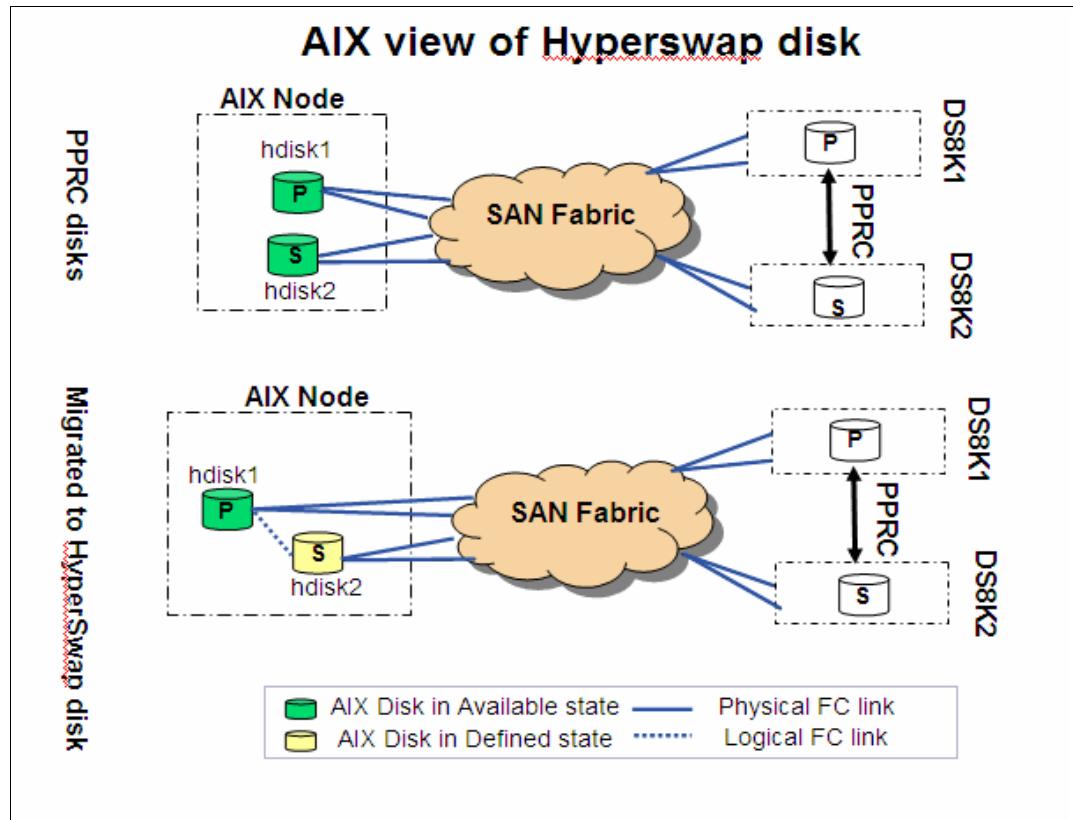


Figure 4-9 AIX view of the HyperSwap disk

Initially, AIX sees PPRC paired disks hdisk1 and hdisk2, one from each of two storage subsystems, DS8K1 and DS8K2. Hdisk1 is in DS8K1 and hdisk2 is in DS8K2. These two disks are in the *available* state. AIX Node has four FC paths to hdisk1 and four FC paths to hdisk2. A new disk attribute, `migrate_disk`, has been implemented for HyperSwap. When one of the PPRC paired disks, say hdisk1, has been configured as `migrate_disk`, its peer paired disk, hdisk2, is changed to the *defined* state. At this point, AIX can see eight paths to hdisk1, which is in the available state. In case AIX Node cannot access the original hdisk1, the disk from DS8K1, the AIX kernel extension will change the path to the disk on DS8K2 while still using hdisk1 in AIX. This is called HyperSwap and is usually transparent to the application.

4.5 HyperSwap functionalities

The following lists the functionalities of HyperSwap on PowerHA SystemMirror 7.1.2 Enterprise Edition:

- ▶ Supports raw disk, volume group, and logical volume.
- ▶ Supports user disks, repository disk and system disk (such as rootvg, paging space disk, and so on).
- ▶ Performs coordinated HyperSwap across multiple nodes.
Useful for applications with concurrent I/O spanned over multiple nodes.
- ▶ Allows planned as well as unplanned HyperSwap operations.
- ▶ Provides consistency semantics for data volumes spanning multiple storage systems.

4.6 Hardware and software requirements

The requirements of AIX and DS8800 microcode are:

- ▶ AIX version: higher than AIX 7.1 TL2 or AIX 6.1 TL8
 - ▶ PowerHA SystemMirror 7.1.2 Enterprise Edition
- If not all filesets of PowerHA SystemMirror 7.1.2 Enterprise Edition are installed, make sure that all HyperSwap-specific filesets are installed, which are:
- cluster.es.genxd.cmds
 - cluster.es.genxd.rte
- ▶ DS8800 with microcode 86.30.49.0 or higher
 - ▶ DS8800 has to be attached with NPIV, FC or FCoE
vSCSI is not supported.

4.7 Limitations

The following limitations apply to the current version of HyperSwap:

- ▶ SCSI reservation is not supported for HyperSwap-enabled configurations.
- ▶ Automatic resynchronization is not supported for HyperSwap-enabled configurations.
 - Users have to manually resume replication after a replication link heals.
 - For DS8800 in-band metro mirror PPRC resources, automatic resynchronization is done through a SystemMirror join cleanup event.
- ▶ Live Partition Mobility (LPM) requires HyperSwap to be disabled for all the affected Mirror Groups.

4.8 Considerations of implementation

Implementation considerations are:

- ▶ I/O freeze operation on DS8K operates on the whole LSS. If a single DS8K LSS contains PPRC volumes from more than one application and if one of the replication links fails, then all PPRC paths get destroyed; and if some of these applications are not managed by PowerHA, then some PPRC paths have to be manually recreated by the client.
- ▶ The PowerHA rediscovery utility has to run after any storage level PPRC configuration change. This includes anytime an add or remove or change of new PPRC paths is performed. Furthermore, HyperSwap functions performed (or automatically triggered) during this time window can cause unexpected or undesired behavior.
- ▶ Disk replication relationships must adhere to a 1-to-1 relationship between the underlying LSS(s).
- ▶ Enabling the HyperSwap for a repository disk would require an alternate disk to be specified.
- ▶ Applications using raw disks are expected to open all the disks up front to enable the HyperSwap capability.
- ▶ HyperSwap does not automatically transfer the SCSI reservations (if any) from the primary to the secondary disks.

4.9 Test environment of a HyperSwap cluster

While our test environment was not ideally configured as a production environment, there is no difference for the purpose of testing functionalities of HyperSwap. The differences between our test environment and a production environment are as follows:

- ▶ All nodes are micropartitions with virtual network and virtual FC port supported by VIOS. The production systems might use physical adapters for both network and FC ports. Moreover, for redundancy considerations, it might use more than one adapter of each kind.
- ▶ The DS8800 DSCLI client is installed in all nodes for convenience during testing. This is not a requirement. It can be installed in any system. In fact, for security reasons, some implementation may choose to manage DS8800 from a system that is accessible only by authorized personnel.
- ▶ There is only one SAN switch. We used the switch zoning capability to separate groups of FC connections. The production system usually contains two or more switches to avoid a single point of failure.

Figure 4-10 shows the environment used for the HyperSwap test in this document.

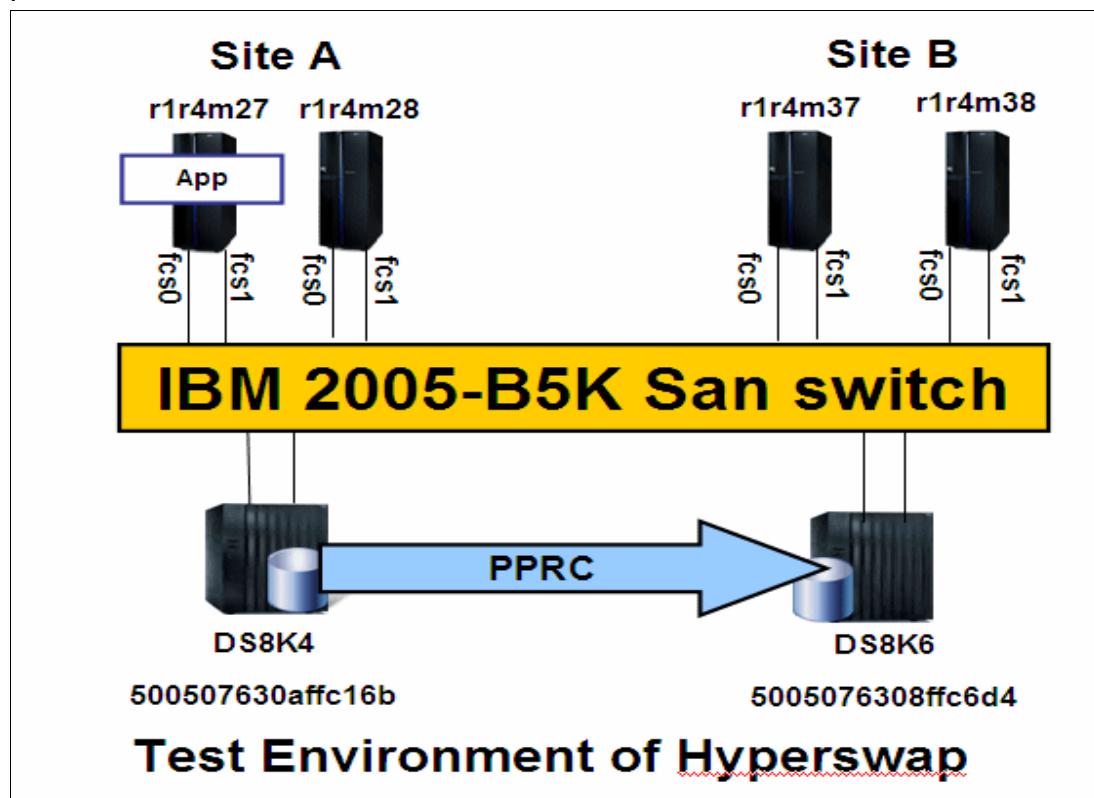


Figure 4-10 HyperSwap test environment

There are four nodes with two FC connections for each of them, two DS8800 storage subsystems and two sites. Nodes r1r4m27, r1r4m28 and storage sys1 are in siteA. Nodes r1r4m37, r1r4m38 and storage sys2 are in siteB. Each node has two virtual fiber channel adapters, fcs0 and fcs1. All nodes are micropartitions hosted on two POWER6® 550 servers.

- ▶ Sites
 - siteA and siteB
 - siteA:
 - Nodes: Two micropartitions, r1r4m27 and r1r4m28
 - Storage: DS8800 DS8K4
 - siteB:

- Nodes: Two micropartitions, r1r4m37 and r1r4m38
 - Storage: DS8800 DS8K6
 - ▶ Nodes
 - r1r4m27 and r1r4m28, r1r4m37 and r1r4m38
 - All nodes are micropartitions of POWER6 550 servers.
 - Operating system: AIX 6.1 TL08.
 - PowerHA: 7.1.2 Enterprise Edition SP1.
 - FC adapters: Two virtual FC adapters, fcs0 and fcs1.
 - Network adapters: Provided by VIOS. The focus of the test is on HyperSwap behavior, which does not involve any traditional failovers such as network adapter failover, node failover, or site failover. Therefore, other than providing the necessary infrastructure for us to access these nodes, the number and types of the network adapters are irrelevant while testing HyperSwap behavior.
 - ▶ SAN switch
 - IBM 2005-B5K SAN switch
 - SAN switch is zoned to simulate multiple switches in functionality.
 - Enhances the high availability for production systems with multiple switches.
 - The connection between each node and each DS8800 subsystem is configured in a separate zone. Because we had two FC connections between each node and SAN switch and two FC connections between each DS8800 and SAN switch, therefore there were four paths between each node and each DS8800.
 - ▶ Storages
 - Two DS8800 storages, DS8K4 and DS8K6.
 - Name ID Storage Unit Model WWNN
 - DS8K4 IBM.2107-75TL771 IBM.2107-75TL770 951 500507630AFFC16B
 - DS8K6 IBM.2107-75LY981 IBM.2107-75LY980 951 5005076308FFC6D4
- These storage system IDs and related information can be obtained with the **lssi** command, as shown in Figure 4-11.

```

From DS8K4
dscli> lssi
Date/Time: December 11, 2012 7:19:18 PM CST IBM DSCLI Version: 6.6.0.305 DS: -
Name ID                   Storage Unit                   Model WWNN                   State ESSNet
=====
DS8K4 IBM.2107-75TL771 IBM.2107-75TL770 951 500507630AFFC16B Online Enabled

From DS8K6
dscli> lssi
Date/Time: December 11, 2012 7:45:59 PM CST IBM DSCLI Version: 6.6.0.305 DS: -
Name ID                   Storage Unit                   Model WWNN                   State ESSNet
=====
ds8k6 IBM.2107-75LY981 IBM.2107-75LY980 951 5005076308FFC6D4 Online Enabled
dscli>

```

Figure 4-11 DS8800 storage system IDs

- Software and firmware version (can be obtained by using the storage command line interface DSCLI as user admin, as shown in Figure 4-12 on page 72).

- DSCLI version 6.6.0.305
- StorageManager version 7.7.3.0.20120215.1

```
dscli> ver -l
Date/Time: December 11, 2012 7:02:55 PM CST IBM DSCLI Version:
6.6.0.305 DS: -
DSCLI          6.6.0.305
StorageManager 7.7.3.0.20120215.1
=====Version=====
Storage Image   LMC
=====
IBM.2107-75TL771 7.6.30.160
dscli>
```

Figure 4-12 DS8800 software and firmware version

- Only one controller modulus per DS8800 was used in the test.
 - Ports for DS8K4: DS8K4_I0100, DS8K4_I0101.
 - Ports for DS8K6: DS8K6_I0330, DS8K6_I0331.
 - Does not affect the functionality test.
 - Should use more than one controller in production for high availability.
- ▶ Repository disk

Our environment was a stretched cluster. Therefore, there was only one repository disk. It resided in one of the DS8800 storages and was shared by all nodes. In a production system, we suggest to put the repository disk on both sites to avoid a site failure, thus causing multiple impacts to the system, or use the HyperSwap capability to support repository mirror groups, which will be demonstrated later in this chapter. HyperSwap also supports a linked cluster as long as the replication between the primary and the secondary storage subsystems is synchronous.

4.10 Initial disk configuration

Before we start, confirm the following:

- ▶ AIX Path Control Module (AIX_AAPCM) driver must be used. Currently IBM Subsystem Device Driver Path Control Module (SDDPCM) or IBM Subsystem Device Driver (SDD) is not supported for HyperSwap functionality. Enter the following command to configure all disks that are part of the storage system to use the AIX_AAPCM driver. A reboot is required after execution.

```
manage_disk_drivers -d device -o AIX_AAPCM
```

- ▶ SCSI reservations are not supported for disks that are used in a HyperSwap mirror group. Verify that no disk reservations are set:

```
devrsrv -c query -l hdisk_name
```

The command returns the following information:

ODM Reservation Policy	:	NO RESERVE
Device Reservation State	:	NO RESERVE

Enter the following command to change the disk reservation policy to no_reserve:

```
chdev -a reserve_policy=no_reserve -l hdisk_number
```

4.11 Preparation of a HyperSwap-capable disk

To create HyperSwap disks, you need to prepare the disk pairs in the storage subsystems and AIX first, then configure them in PowerHA.

This section describes the procedures for preparing HyperSwap disks. An overview of these steps is as follows:

1. Select two disks, one from each storage subsystem to be mirrored for HyperSwap disks. Assume that these two disks are hdiskA (in DS8K4) and hdiskB (in DS8K6).
2. Establish the connection path from hdiskA to hdiskB (from DS8K4, using `mkpprcpath`).
3. Establish the connection path from hdiskB to hdiskA (from DS8K6, using `mkpprcpath`).
4. Establish the volume pair of diskA and diskB in one direction (from DS8K4, using `mkpprc`).
5. Enable HyperSwap for hdiskA on all nodes (from all nodes, using `chdev`).

The following section provides detailed steps to prepare the HyperSwap disks.

Note: In our test environment, the DSCLI client was installed on all AIX nodes in /opt/ibm/dscli/bin. Other implementations may have the DSCLI client in a different location.

4.11.1 Selecting disks to be mirrored for HyperSwap disks

We needed two disks, one from each DS8K storage system. Any DS8K disk accessible by all the nodes in the cluster is eligible for HyperSwap. Disks already used can be used for HyperSwap, too. However, special care needs to be taken to ensure data integrity for the existing disks. For instance, if diskA is already in a volume group containing data, the HyperSwap should be made from diskA, instead of to diskB.

The `lshostvol.sh` command located in /opt/ibm/dscli/bin/ displays the disk attributes, including the containing storage system LSS ID, and so on, as shown in Figure 4-13 on page 74.

```
(0) root @ r1r4m27:(REG) /opt/ibm/dscli/bin
> ./lshostvol.sh
Device Name      Volume ID
-----
hdisk1           IBM.2107-75TL771/1A00
hdisk2           IBM.2107-75TL771/9A00
hdisk3           IBM.2107-75TL771/9A01
hdisk4           IBM.2107-75TL771/9A02
hdisk5           IBM.2107-75TL771/9A03
hdisk6           IBM.2107-75TL771/9A04
hdisk7           IBM.2107-75TL771/B000
-----
hdisk22          IBM.2107-75TL771/BA00
hdisk23          IBM.2107-75TL771/BA01
hdisk24          IBM.2107-75TL771/BA02
hdisk25          IBM.2107-75TL771/BC00
hdisk26          IBM.2107-75TL771/BC01
hdisk27          IBM.2107-75TL771/BC02
-----
hdisk36          IBM.2107-75LY981/5C03
hdisk37          IBM.2107-75LY981/5C04
hdisk56          IBM.2107-75LY981/9A00
hdisk60          IBM.2107-75LY981/9C01
hdisk62          IBM.2107-75LY981/9E00
```

Figure 4-13 Sample output of lshostvol.sh

The Volume ID contains the following:

<vendor_name>.<storage_type>-<serial_number>/<LSS_ID><volume_ID>.

For instance, hdisk25 has the volume ID IBM.2107-75TL771/BC00.

Choose two disks, one from each storage, to make a pprc pair. Figure 4-14 on page 75 shows some examples to examine the status of the disk. This example shows that hdisk24 is already a HyperSwap disk, while hdisk25 and hdisk56 are not configured to be HyperSwap disks yet.

Tips: Because the HyperSwap unit is based on LSS, it is a good idea to follow the following rules to get better performance for HyperSwap:

- ▶ Keep all the HyperSwap disks of the same applications in the same LSS whenever possible.
- ▶ Do not mix HyperSwap disks of different applications in the same LSS whenever possible.

```

(0) root @ r1r4m27:(REG) /opt/ibm/dscli/bin
> lspprc -p hdisk24
path      WWNN          LSS  VOL   path
group id
=====
0(s)      500507630affc16b 0xba 0x02  PRIMARY
1          5005076308ffc6d4 0x98 0x02  SECONDARY

path      path  path      parent  connection
group id  id    status
=====
0     0     Enabled   fscsi0  500507630a08016b,40ba400200000000
0     1     Enabled   fscsi0  500507630a08416b,40ba400200000000
0     2     Enabled   fscsi1  500507630a08016b,40ba400200000000
0     3     Enabled   fscsi1  500507630a08416b,40ba400200000000
1     6     Enabled   fscsi1  50050763081b06d4,4098400200000000
1     7     Enabled   fscsi1  50050763081b46d4,4098400200000000
1     4     Enabled   fscsi0  50050763081b06d4,4098400200000000
1     5     Enabled   fscsi0  50050763081b46d4,4098400200000000

(0) root @ r1r4m27:(REG) /opt/ibm/dscli/bin
> lspprc -p hdisk25
hdisk25 is not a hyperswap disk

(255) root @ r1r4m27:(REG) /opt/ibm/dscli/bin
> lspprc -p hdisk56
hdisk56 is not in Available state

```

Figure 4-14 Sample output of lspprc

To create a PPRC pair, the WWNN for both storages are needed, which can be obtained with the **lssi** command from each storage, as shown in Figure 4-15 on page 76.

```

> /opt/ibm/dscli/dscli -user fvtadmin -passwd my_password -hmc1 9.3.18.145
Date/Time: November 21, 2012 5:23:36 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771

dscli> lssi
Date/Time: November 21, 2012 5:23:39 PM CST IBM DSCLI Version: 6.6.0.305 DS: -
Name ID Storage Unit Model WWNN State ESSNet
=====
DS8K4 IBM.2107-75TL771 IBM.2107-75TL770 951 500507630AFFC16B Online Enabled
dscli>
(130) root @ r1r4m27:(REG) /
> /opt/ibm/dscli/dscli -user fvtadmin -passwd my_password -hmc1 9.3.18.238
Date/Time: November 21, 2012 5:24:30 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75LY981

dscli> lssi
Date/Time: November 21, 2012 5:24:33 PM CST IBM DSCLI Version: 6.6.0.305 DS: -
Name ID Storage Unit Model WWNN State ESSNet
=====
ds8k6 IBM.2107-75LY981 IBM.2107-75LY980 951 5005076308FFC6D4 Online Enabled
dscli>

```

Figure 4-15 Sample output of the lssi command

You also need to know port numbers available to connect this pair of disks. It can be obtained with the **lavailpprcpair** command, as shown in Figure 4-16 on page 77.

```

dscli> lsavailpprcport -remotedev IBM.2107-75LY981 -remotewwnn
5005076308FFC6D4 BC:9A
Date/Time: November 21, 2012 4:46:09 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771
Local Port Attached Port Type
=====
I0030    I0000      FCP
I0030    I0001      FCP
I0030    I0200      FCP
I0030    I0201      FCP
I0031    I0000      FCP
I0031    I0001      FCP
I0031    I0200      FCP
I0031    I0201      FCP
I0032    I0002      FCP
I0032    I0003      FCP
I0033    I0002      FCP
I0033    I0003      FCP
I0100    I0130      FCP
I0100    I0131      FCP
I0100    I0330      FCP
I0100    I0331      FCP
I0101    I0130      FCP
I0101    I0131      FCP
I0101    I0332      FCP
I0101    I0333      FCP
I0102    I0132      FCP
I0102    I0133      FCP
I0102    I0334      FCP      ****
I0102    I0335      FCP
I0103    I0132      FCP
I0103    I0133      FCP
I0103    I0334      FCP
I0103    I0335      FCP
I0236    I0532      FCP
I0236    I0533      FCP
I0237    I0532      FCP
I0237    I0533      FCP
I0306    I0330      FCP
I0306    I0331      FCP
I0306    I0332      FCP
I0306    I0333      FCP
dscli>

```

Figure 4-16 Sample output of the lsavailpprcpair command

We chose ports I0102 and I0334 for this demonstration.

4.11.2 Establishing the connection path from hdiskA to hdiskB

We established the connection path from hdiskA to hdiskB with the DSCLI **mkpprcpath** command. First, we logged in to DS8K4, as shown in Figure 4-17 on page 78.

```
(0) root @ r1r4m27:(REG) /
> /opt/ibm/dscli/dscli -user fvtadmin -passwd my_password -hmc1 9.3.18.145
Date/Time: November 25, 2012 4:24:28 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771

dscli>
```

Figure 4-17 Log in to storageA

On DS8K4, we established the connection path from hdiskA to hdiskB with the **mkpprcpath** command, and checked the status of the path with the **1spprcpath** command, as shown in Figure 4-18. Note that we checked the path before and after **mkpprcpath** to show the change that **mkpprcpath** made.

```
dscli> 1spprcpath BC
Date/Time: November 25, 2012 4:58:35 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771
CMUC00234I 1spprcpath: No Remote Mirror and Copy Path found.
dscli> mkpprcpath -dev IBM.2107-75TL771 -srcLss BC -tgtLss 9A -remotewwnn
50050763081B06D4 I0102:I0334
Date/Time: November 25, 2012 5:06:06 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771
CMUC00149I mkpprcpath: Remote Mirror and Copy path BC:9A successfully
established.
dscli> 1spprcpath BC
Date/Time: November 25, 2012 5:06:59 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771
Src Tgt State SS Port Attached Port Tgt WWNN
=====
BC 9A Success FF9A I0102 I0334 50050763081B06D4
```

Figure 4-18 Establish the connection path from hdiskA to hdiskB

4.11.3 Establishing the connection path from hdiskB to hdiskA

We established the connection path from hdiskB to hdiskA with the DSCLI **mkpprcpath** command. First, we logged in to DS8K6, as shown in Figure 4-19.

```
> /opt/ibm/dscli/dscli -user fvtadmin -passwd my_password -hmc1 9.3.18.238
Date/Time: November 25, 2012 4:37:23 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75LY981

dscli>
```

Figure 4-19 Login to storageB

On DS8K6, we established the connection path from hdiskB to hdiskA with the **mkpprcpath** command and checked the status of the path with the **1spprcpath** command, as shown in Figure 4-20 on page 79. Note that we checked the path before and after **mkpprcpath** to show the change that **mkpprcpath** made.

```

dscli> lspprcpath 9A
Date/Time: November 25, 2012 4:59:05 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75LY981
CMUC00234I lspprcpath: No Remote Mirror and Copy Path found.
dscli> mkpprcpath -dev IBM.2107-75LY981 -srcLss 9A -tgtLss BC -remotewwnn
500507630AFFC16B I0334:I0102
Date/Time: November 25, 2012 5:06:27 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75LY981
CMUC00149I mkpprcpath: Remote Mirror and Copy path 9A:BC successfully
established.
dscli> lspprcpath 9A
Date/Time: November 25, 2012 5:06:49 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75LY981
Src Tgt State SS Port Attached Port Tgt WNN
=====
9A BC Success FFBC I0334 I0102      500507630AFFC16B
dscli>

```

Figure 4-20 Establishing a connection path from hdiskB to hdiskA

4.11.4 Establishing volume pairs in one direction

Now we established the volume pair of diskA and diskB in one direction with the **mkpprc** command, as shown in Figure 4-21, submitted on the storage subsystem DS8K4.

```

dscli> mkpprc -dev IBM.2107-75TL771 -remotedev IBM.2107-75LY981 -mode full
-type mmir BC00:9A00
Date/Time: November 25, 2012 5:08:09 PM CST IBM DSCLI Version: 6.6.0.305 DS:
IBM.2107-75TL771
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship BC00:9A00
successfully created.
dscli>

```

Figure 4-21 Establishing the volume pair of diskA and diskB in one direction

4.11.5 Enabling HyperSwap on all nodes

We enabled HyperSwap capability for this PPRC pair. Before doing this, we checked the status of both disks with the **lspprc** command from one of the nodes, as shown in Figure 4-22 on page 80.

```

(0) root @ r1r4m28: /
> lspprc -p hdisk25
path      WWNN          LSS  VOL   path
group id
=====
0(s)      500507630affc16b 0xbc 0x00  PRIMARY
-1        5005076308ffc6d4 0x9a 0x00

path      path  path      parent  connection
group id  id    status
=====
0         0     Enabled   fscsi0  500507630a08016b,40bc400000000000
0         1     Enabled   fscsi0  500507630a08416b,40bc400000000000
0         2     Enabled   fscsi1  500507630a08016b,40bc400000000000
0         3     Enabled   fscsi1  500507630a08416b,40bc400000000000

(0) root @ r1r4m28: /
> lspprc -p hdisk56
path      WWNN          LSS  VOL   path
group id
=====
0(s)      5005076308ffc6d4 0x9a 0x00  SECONDARY
-1        500507630affc16b 0xbc 0x00

path      path  path      parent  connection
group id  id    status
=====
0         0     Enabled   fscsi0  50050763081b06d4,409a400000000000
0         1     Enabled   fscsi0  50050763081b46d4,409a400000000000
0         2     Enabled   fscsi1  50050763081b06d4,409a400000000000
0         3     Enabled   fscsi1  50050763081b46d4,409a400000000000

(0) root @ r1r4m28: /

```

Figure 4-22 Status of the PPRC pair disk before enabling HyperSwap

Then we made the disk HyperSwap capable with the **chdev** command, as shown in Figure 4-23.

```

(0) root @ r1r4m28: /
> chdev -a san_rep_cfg=migrate_disk -l hdisk25 -U
hdisk25 changed

(0) root @ r1r4m28: /

```

Figure 4-23 Enabling the disk as HyperSwap capable

Then we checked the status again as shown in Figure 4-24 on page 81.

```

(0) root @ r1r4m28: /
> lspprc -p hdisk56
hdisk56 is not in Available state

(22) root @ r1r4m28: /
> lspprc -p hdisk25
path      WWNN          LSS  VOL   path
group id
=====
0(s)      500507630affc16b 0xbc 0x00  PRIMARY
1          5005076308fffc6d4 0x9a 0x00  SECONDARY

path      path  path      parent  connection
group id  id    status
=====
0     0     Enabled   fscsi0  500507630a08016b,40bc400000000000
0     1     Enabled   fscsi0  500507630a08416b,40bc400000000000
0     2     Enabled   fscsi1  500507630a08016b,40bc400000000000
0     3     Enabled   fscsi1  500507630a08416b,40bc400000000000
1     4     Enabled   fscsi0  50050763081b06d4,409a400000000000
1     5     Enabled   fscsi0  50050763081b46d4,409a400000000000
1     6     Enabled   fscsi1  50050763081b06d4,409a400000000000
1     7     Enabled   fscsi1  50050763081b46d4,409a400000000000

(0) root @ r1r4m28: /

```

Figure 4-24 The status of the PPRC pair disk after being HyperSwap-enabled

Note that the secondary disk hdisk56 becomes unavailable. It is changed to the “defined” state. Repeat the same procedures from Figure 4-22 on page 80 to Figure 4-24 for all other nodes.

Tip: Before enabling these HyperSwap-capable disks in 4.12, “Initial configuration for the PowerHA cluster” on page 81, make sure you have prepared enough HyperSwap-capable disks for use. In general, you definitely need to have HyperSwap disks for user data. It is also a good idea to have HyperSwap disks for repository disks to make the cluster more robust. Sometimes you should also keep the system disks on the HyperSwap disks.

4.12 Initial configuration for the PowerHA cluster

This section provides the steps for creating the PowerHA cluster. In our configuration we created the stretched cluster.

First confirm that the necessary filesets, including the genxd filesets, are installed on all nodes, as shown in Example 4-1.

Example 4-1 Confirming that the filesets have been installed

```
# lslpp -l cluster.es.genxd*
Fileset           Level  State       Description
-----
Path: /usr/lib/objrepos
cluster.es.genxd.cmds    7.1.2.1  APPLIED    PowerHA SystemMirror
```

		Enterprise Edition - Generic
cluster.es.genxd.rte	7.1.2.1 APPLIED	XD support - Commands
		PowerHA SystemMirror
		Enterprise Edition - Generic
		XD support - Runtime
		Environment
<hr/>		
Path: /etc/objrepos		
cluster.es.genxd.cmds	7.1.2.0 COMMITTED	PowerHA SystemMirror
		Enterprise Edition - Generic
		XD support - Commands
cluster.es.genxd.rte	7.1.2.0 COMMITTED	PowerHA SystemMirror
		Enterprise Edition - Generic
		XD support - Runtime
		Environment

Next populate the CAA rhosts files (/etc/cluster/rhosts) on all nodes with the IP labels to be used for the communication path, as shown in Example 4-2.

Example 4-2 Configuring the /etc/cluster/rhosts files

```
# cat /etc/cluster/rhosts
r1r4m27
r1r4m28
r1r4m37
r1r4m38
```

After editing the rhost file, restart the clcomd daemon with the **stopsrc -s clcomd** and the **startsrc -s clcomd** commands, respectively. Confirm that the daemon is running, as shown in Example 4-3.

Example 4-3 Confirming the clcomd status

```
# lssrc -s clcomd
Subsystem      Group          PID      Status
  clcomd        caa           5177510  active
```

Configure the cluster through **smitty sysmirror** → **Cluster Nodes and Networks** → **Multi Site Cluster Deployment** → **Setup a Cluster, Nodes and Networks**, which leads to the panel shown in Figure 4-25 on page 83.

Setup Cluster, Sites, Nodes and Networks	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
* Cluster Name	[Entry Fields] [r1r4m27_cluster]
* Site 1 Name	[siteA]
* New Nodes (via selected communication paths)	[r1r4m27 r1r4m28] +
* Site 2 Name	[siteB]
* New Nodes (via selected communication paths)	[r1r4m37 r1r4m38] +
Cluster Type	[Stretched Cluster] +

Figure 4-25 Creating the stretched cluster

After pressing Enter, a discovery process is also executed. It automatically creates the cluster networks and saves a list of all shared disks. The next step is to choose repository disks and multicast addresses to be used by CAA. This can be done through **smitty sysmirror** → **Cluster Nodes and Networks** → **Multi Site Cluster Deployment** → **Define Repository Disk and Cluster IP Address**, which leads to the panel shown in Figure 4-26.

Define Repository Disk and Cluster IP Address	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
* Cluster Name	[Entry Fields] [r1r4m27_cluster]
* Repository Disk	[hdisk21] +
Cluster IP Address	[]

Figure 4-26 Defining repository disks and cluster IP addresses

Finally, verify and synchronize the cluster through **smitty sysmirror** → **Cluster Nodes and Networks** → **Verify and Synchronize Cluster Configuration**. This creates the basic cluster.

4.13 Configuring the HyperSwap disk in PowerHA

In this section we describe the procedures for configuring HyperSwap disks in PowerHA. An overview of the steps follows:

1. Create a volume group with HyperSwap-capable disks previously prepared in “Preparation of a HyperSwap-capable disk” on page 73 for all nodes.
2. Define the storage systems and their site association.
3. Set up the mirror group for the HyperSwap disks.
 - User mirror group
 - Cluster_repository mirror group
 - SystemMirror group

4. Create a resource group with site policy.
5. Add a mirror group and a volume group into the resource group.
6. Verification and synchronization.

The following are the details to configure HyperSwap disks to PowerHA. Details of SMIT path as well as fast path are shown in all procedures. To save effort, always start the SMIT fast path with **smit sysmirror**.

4.13.1 Creating a volume group with HyperSwap capable disks

We previously prepared HyperSwap disks in 4.11, “Preparation of a HyperSwap-capable disk” on page 73. They must be enabled as HyperSwap disks.

Use the fastpath **smit cl_vg** or follow **smitty sysmirror** → **System Management (C-SPOC)** → **Storage** → **Volume Group** → **Create a Volume Group** to create the volume group with a HyperSwap disk as shown in Figure 4-27. Select all the nodes, one by one with the F7 key.

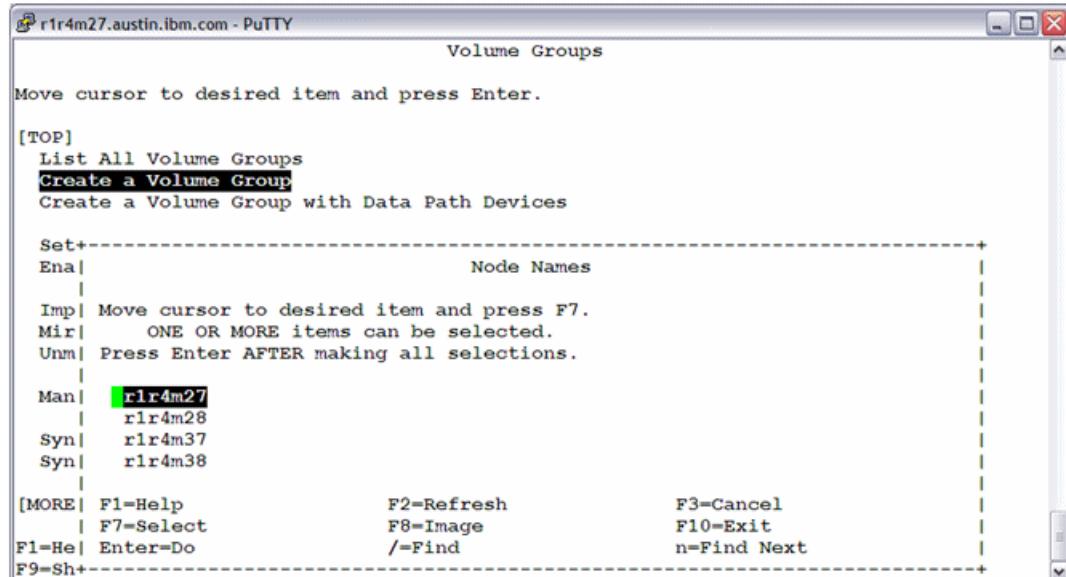


Figure 4-27 Create the volume group with the HyperSwap disk

Afterwards, press Enter and a window with physical volume names appears as shown in Figure 4-28 on page 85. Select the desired disks. In this scenario, we used hdisk25.

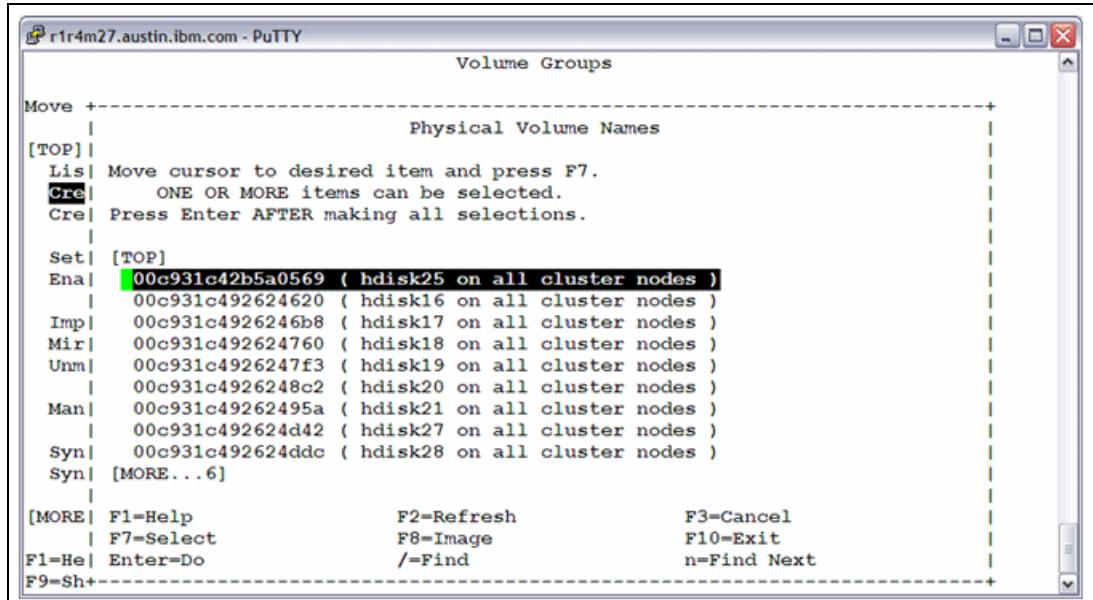


Figure 4-28 Create the volume group with the HyperSwap disk, continued

After pressing Enter, a window with various types of volume groups appears, as shown in Figure 4-29.

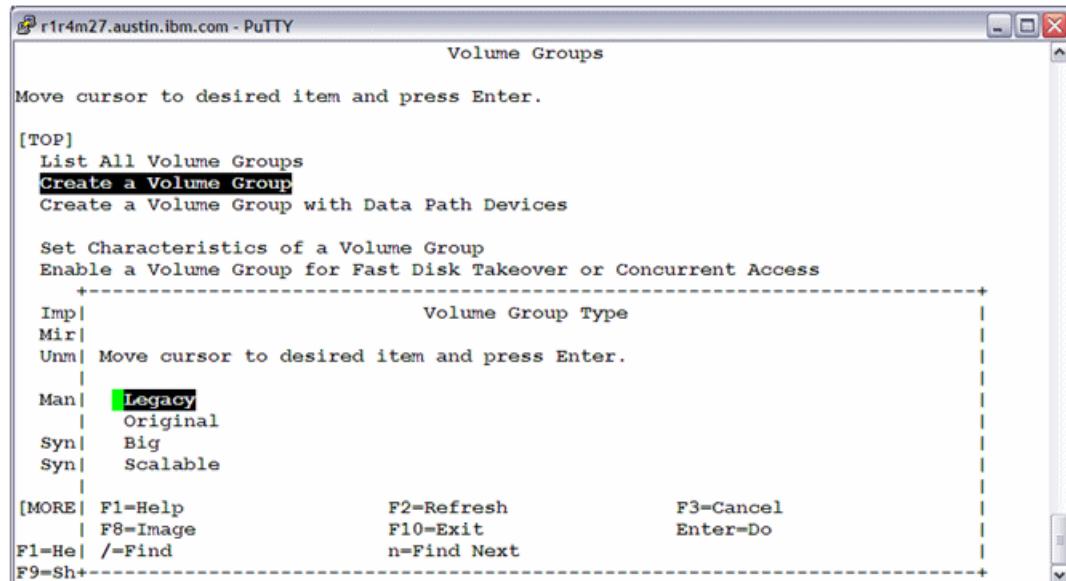


Figure 4-29 Create the volume group with the HyperSwap disk, continued

Select the appropriate type depending on the number of disks you are including in this volume group. We chose Original which supports up to 32 disks. Big supports 128 physical volumes and 512 logical volumes. Scalable supports 1024 physical volumes, 256 logical volumes and 32768 physical partitions.

Next press Enter, and the last menu is displayed, shown in Figure 4-30 on page 86.

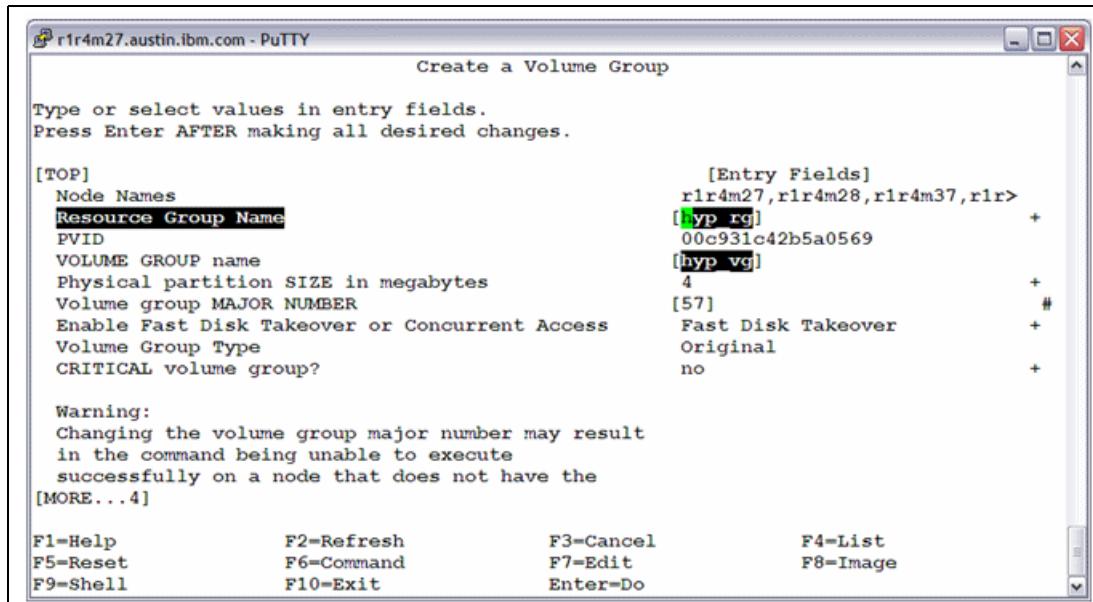


Figure 4-30 Create the volume group with the HyperSwap disk, continued

If a resource group already exists and you want to add a volume group to it, press F4 to generate a pick list and chose the defined resource group. Otherwise, just enter the resource group name in the field. Also enter the volume group name in the field.

After pressing Enter, you see a window processing the creation of the volume group. When successful, you see the output window shown in Figure 4-31.

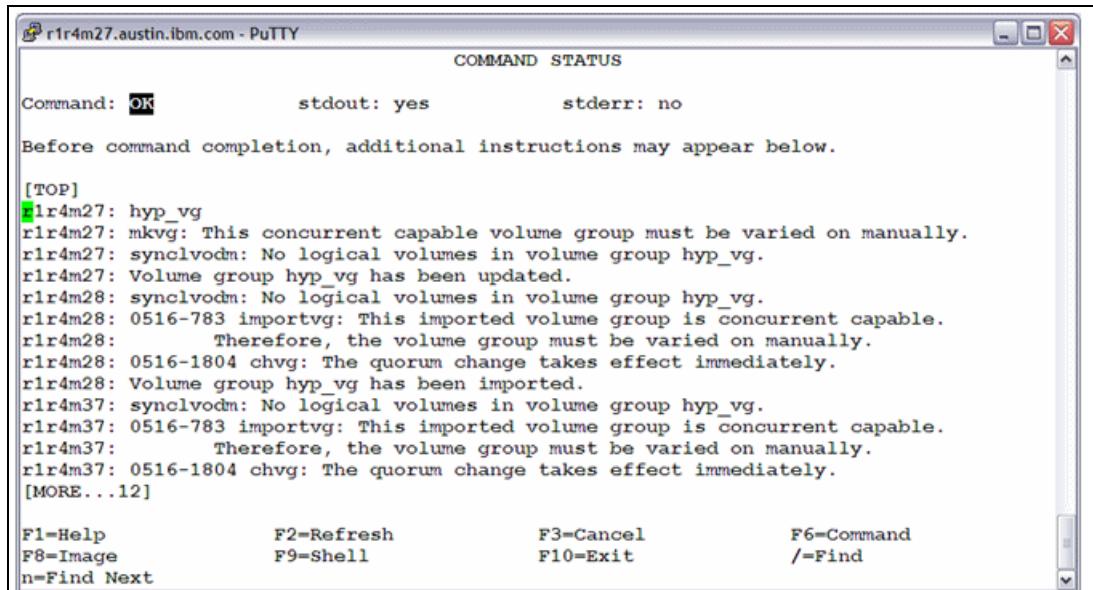


Figure 4-31 Create the volume group with the HyperSwap disk completed

4.13.2 Defining the storage systems and their site association

Use SMIT fastpath `smit cm_add_strg_system` or follow the SMIT path `smitty sysmirror` → **Cluster Applications and Resources** → **Resources** → **Configure DS8000 Metro Mirror**

| (In-Band) Resources → Configure Storage Systems → Add a Storage System, as shown in Figure 4-32.

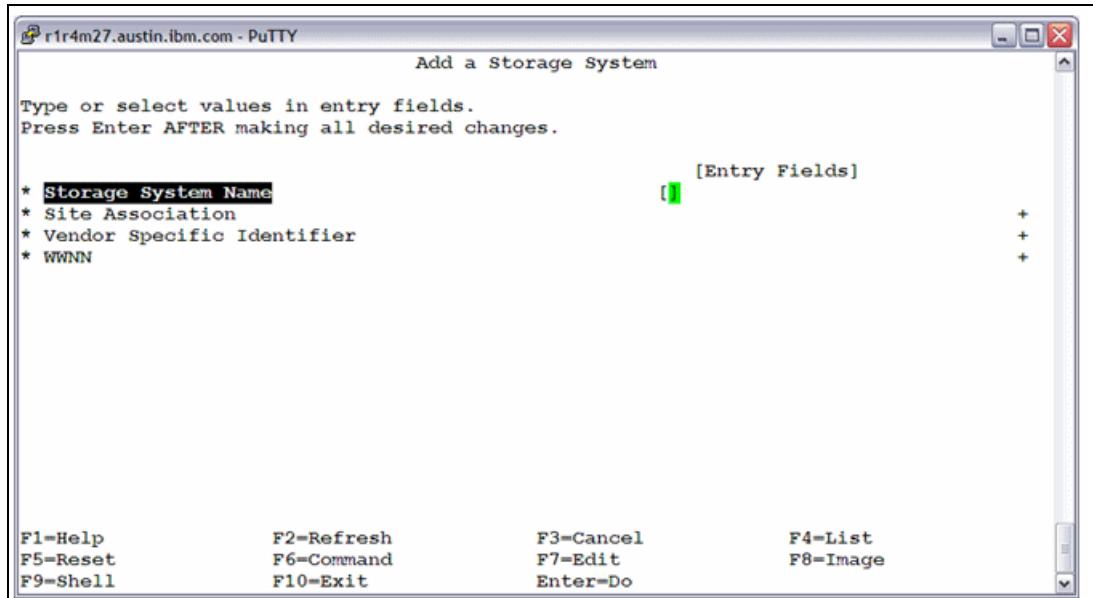


Figure 4-32 Define the storage system

Enter the storage system name. Then move the cursor to *Site Association* and press F4 to get a popup window with available sites, as shown in Figure 4-33.

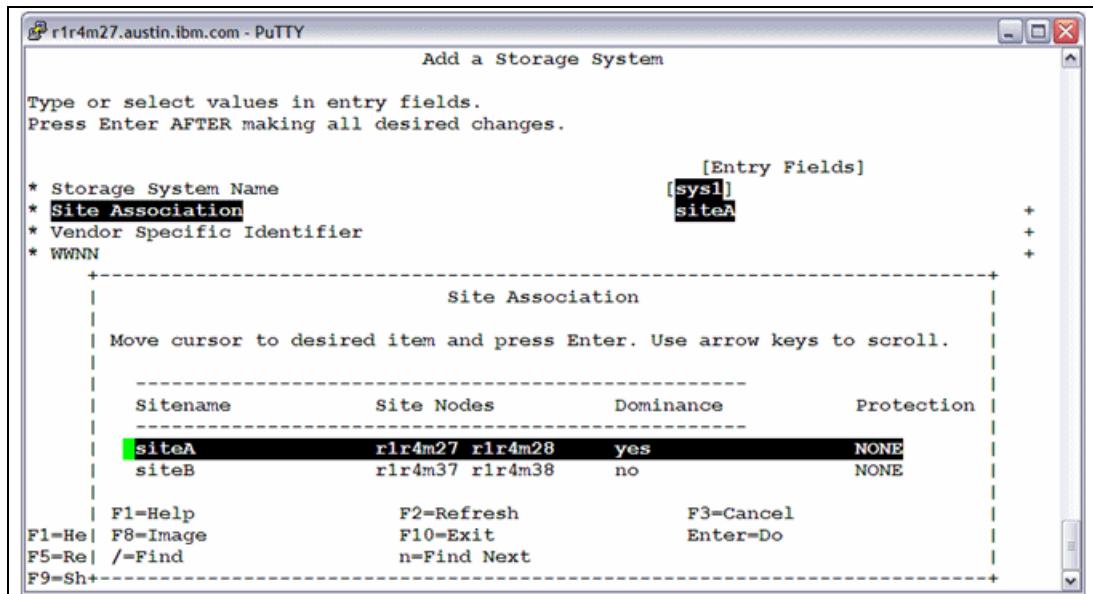


Figure 4-33 Define the storage system, continued

Select the site association for this storage system. Then move the cursor to *Vendor Specific Identifier* and *WWNN*, and press F4 to generate a pick list in a pop up window. Select your choice of Vendor Specific Identifier and WWNN as shown in Figure 4-34 on page 88.

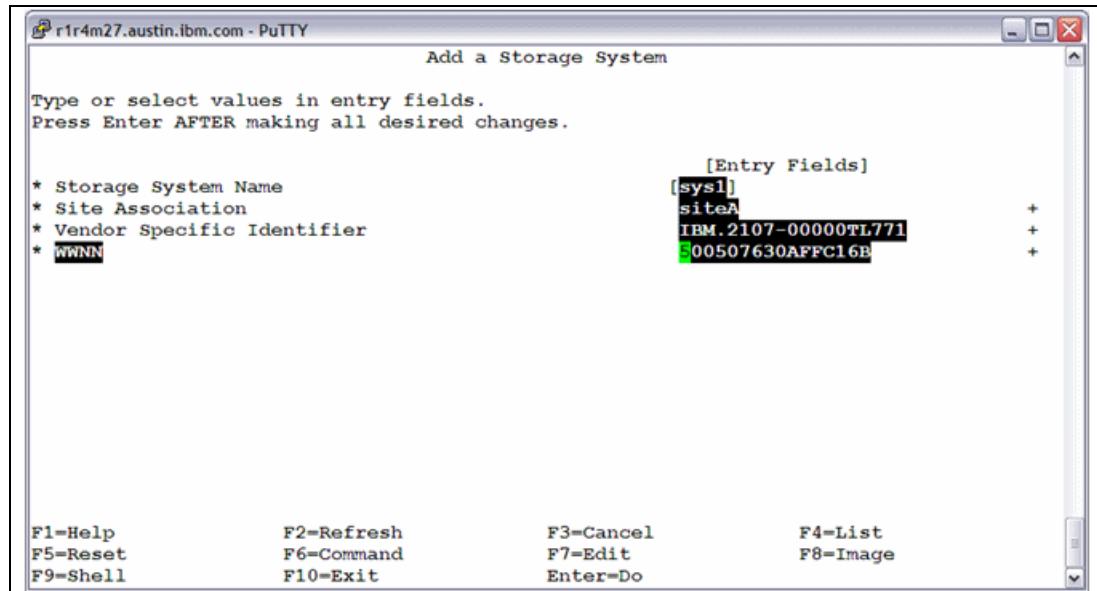


Figure 4-34 Define the storage system, continued

Repeat the previous procedure to add the second storage system, as shown in Figure 4-35.

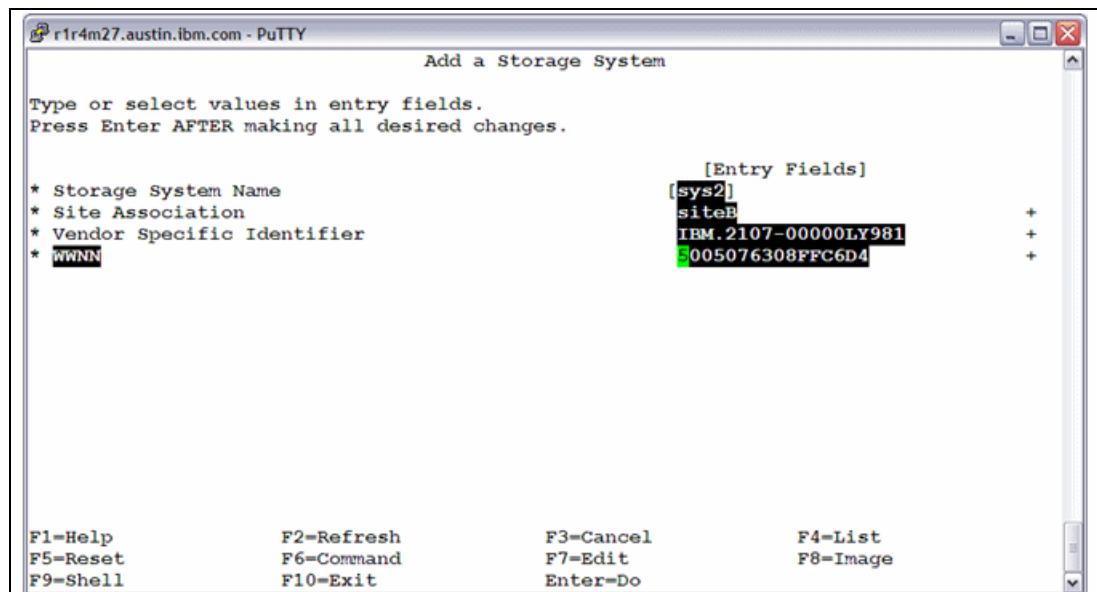


Figure 4-35 Define the storage system, continued

Note: WWNN chosen in this section corresponds to the DS8800 subsystems for our test environment described in 4.9, “Test environment of a HyperSwap cluster” on page 69 and Figure 4-11 on page 71.

4.13.3 Setting up the mirror group you want these HyperSwaps to be in

Use SMIT fast path `smit cm_cfg_mirr_grps` or follow the SMIT path `smitty sysmirror → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror`

(In-Band) Resources → Configure Mirror Groups → Add a Mirror Group as shown in Figure 4-36. Move the cursor to *Add a Mirror Group*, and a window pops up to show the choice of mirror group. There are three mirror groups: user, system, and cluster_repository. User is used for application data. System can be used for rootvg, paging space, and dump devices. Cluster_repository is used for keeping cluster vital information.

User mirror group

This user mirror group is used for application data.

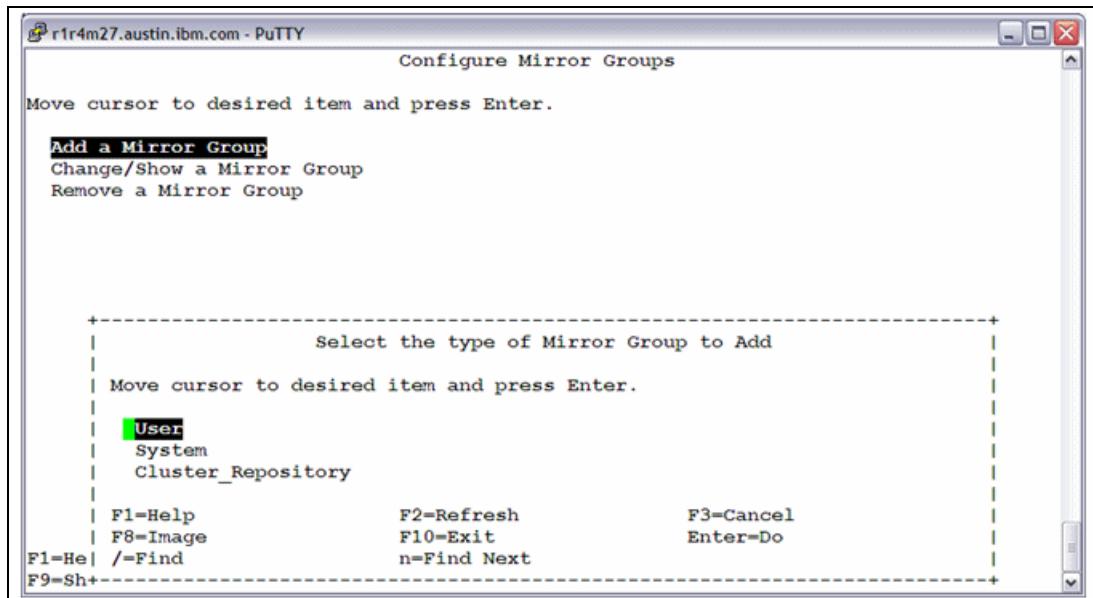


Figure 4-36 Configure user mirror group

The window *Add a User Mirror Group* is shown in Figure 4-37.

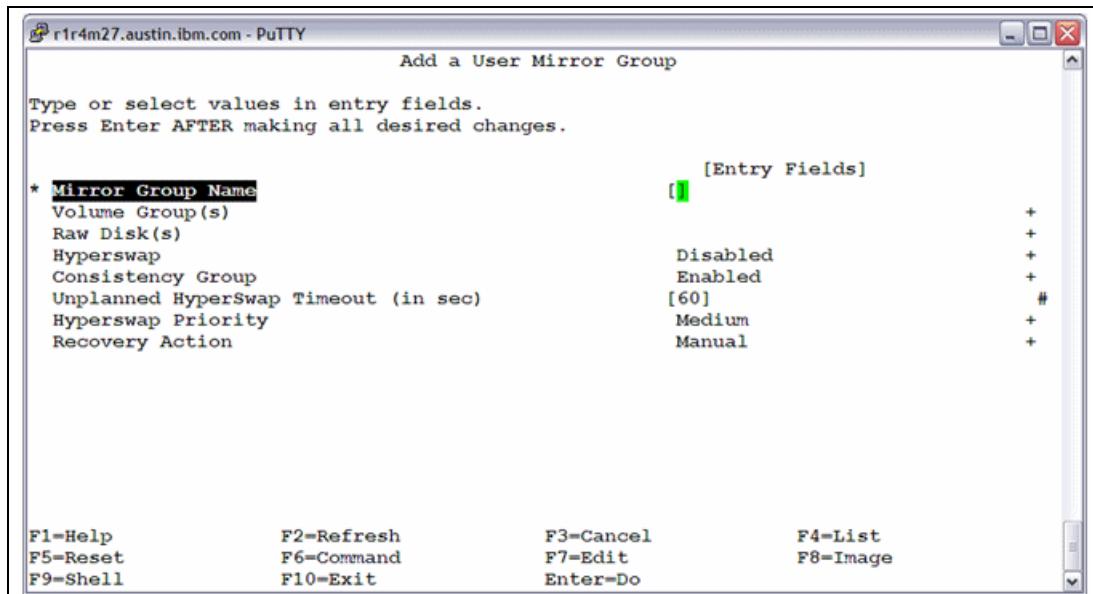


Figure 4-37 Configure user mirror group, continued

Enter the mirror group name, change the HyperSwap field from *Disable* to *Enable*, press F4 in the volume group field to get a pop-up window showing available values, as shown in Figure 4-38.

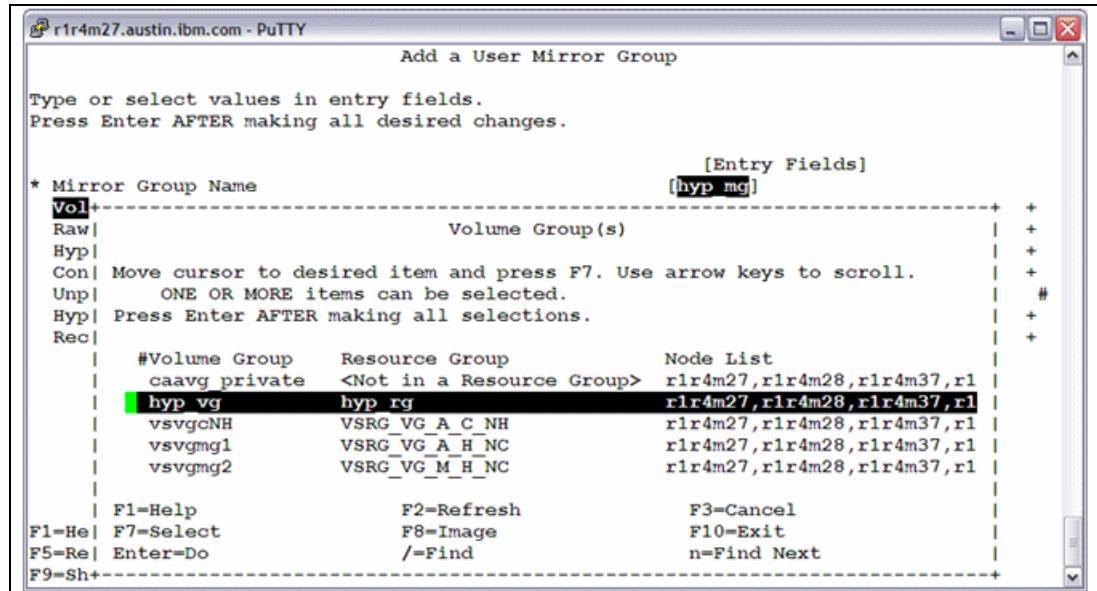


Figure 4-38 Configure user mirror group, continued

Select the volume group **hyp_vg**, which was previously created in Figure 4-30 on page 86. The window then looks as shown in Figure 4-39.

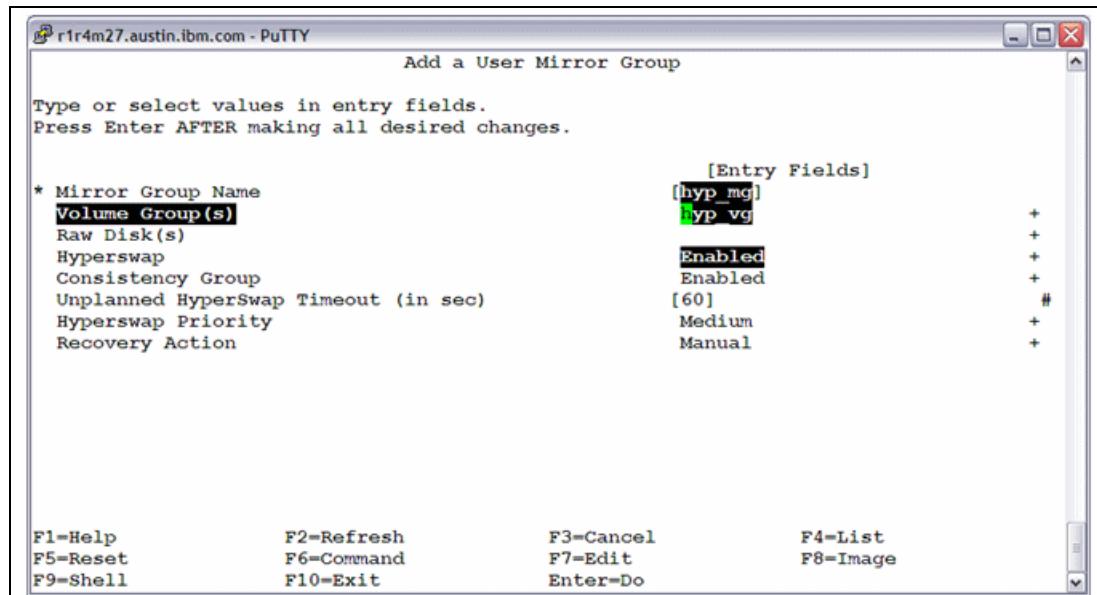


Figure 4-39 Configure user mirror group, continued

Press Enter to add the user mirror group.

Cluster repository mirror group

The setup of the cluster repository mirror group is slightly different from that for a user mirror group. Before configuring the cluster repository mirror group, you need to know which disk is

currently used for the repository disk. The **lspv** command displays the repository disk with the label **caavg_private** and **hdisk21** as shown in Figure 4-40.

(0) root @ r1r4m27:(REG) /			
> lspv			
hdisk0	00c931c40c99bd1f	rootvg	active
hdisk1	00c931c492623e86	oravg	active
hdisk2	00c931c42c5def32	None	
hdisk3	00c931c492623fd0	None	
hdisk4	00c931c492624077	None	
hdisk5	00c931c49262410a	None	
hdisk6	00c931c4926241b2	None	
hdisk7	00c931c492624248	vsvgmg1	concurrent
hdisk8	00c931c492624300	vsvgmg1	concurrent
hdisk9	00c931c49262439f	vsvgmg1	concurrent
hdisk10	00c931c49262540d	None	
hdisk11	00c931c4926254b7	None	
hdisk12	00c931c492625570	None	
hdisk13	00c931c49262444f	hyp_vg	concurrent
hdisk14	00c931c4926244de	hyp_vg	concurrent
hdisk15	00c931c49262458a	hyp_vg	concurrent
hdisk16	00c931c492624620	None	
hdisk17	00c931c4926246b8	None	
hdisk18	00c931c492624760	None	
hdisk19	00c931c4926247f3	None	
hdisk20	00c931c4926248c2	None	
hdisk21	00c931c49262495a	caavg_private	active
hdisk22	00c931c492624a21	vsvgcNH	concurrent
hdisk23	00c931c492624abe	vsvgcNH	concurrent
hdisk24	00c931c492624b5c	vsvgcNH	concurrent
hdisk26	00c931c492624c9a	None	

Figure 4-40 Repository disk configuration for HyperSwap

Instead of choosing user type in Figure 4-36 on page 89, select the cluster_repository mirror group to set up the HyperSwap repository disk and the window as shown in Figure 4-41.

Add cluster Repository Mirror Group	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
[Entry Fields]	
* Mirror Group Name	[]
* Site Name	[]
* Non Hyperswap Disk	[]
* Hyperswap Disk	[]
Hyperswap	Disabled
Consistency Group	Enabled
Unplanned HyperSwap Timeout (in sec)	[60]
Hyperswap Priority	High

Figure 4-41 Adding the cluster repository mirror group

- ▶ Mirror Group Name - Enter a name of your choice.
- ▶ Site Name - Use F4 for the choices, then select both sites in the list.

Note: There are two kinds of clusters in PowerHA: stretched and linked clusters. Stretched clusters require multicast within the cluster and share the only repository disk in the cluster, while linked clusters use unicast to communicate between two sites and each site has its own repository disk. Stretched clusters are more robust than linked clusters but require higher bandwidth for communication. Therefore, it is hard to implement them for two sites that are far apart from each other. They are usually suitable for a metropolitan cluster. Because HyperSwap requires synchronous replication between the primary and secondary storages, which requires higher bandwidth than asynchronous replication that linked clusters usually use, and therefore HyperSwap is primary for stretched clusters and we chose both sites. Additional details can be found in 2.2.1, “Stretched and linked clusters” on page 19.

- ▶ Non HyperSwap disk - Use the original repository disk from Figure 4-40 on page 91.
- ▶ HyperSwap disk - Use one of the HyperSwap disks you have prepared in “Preparation of a HyperSwap-capable disk” on page 73.
- ▶ HyperSwap - Change it from Disabled to Enabled so that the new HyperSwap-capable repository disk is used after cluster verification and synchronization.

Figure 4-42 shows the window for the selections.

Add cluster Repository Mirror Group	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
[Entry Fields]	
* Mirror Group Name	[repos_mg]
* Site Name	siteA siteB +
* Non Hyperswap Disk	hdisk21:9d725962-004d-d729-719d-ae7425ba72c8 +
* Hyperswap Disk	hdisk16:edd406a1-1945-7ea0-2901-a709d6ed116e +
Hyperswap	Enabled +
Consistency Group	Enabled +
Unplanned HyperSwap Timeout (in sec) [60]	#
Hyperswap Priority	High +

Figure 4-42 Adding the cluster repository mirror group, continued

Be very careful in selecting the right Non HyperSwap Disk. If you choose the wrong Non HyperSwap Disk, the system may perform properly without any notice until the HyperSwap repository disk is disabled for any reason. When that situation happens, the wrong Non HyperSwap Disk will be used. The disaster can be no repository disk to use or the data on this wrong Non Hyperwap Disk will be corrupted. It is not clear whether cluster verification is smart enough to find out the mistake in choosing the wrong repository disk.

After cluster verification and synchronization, the repository disk is changed to hdisk16 as shown in Figure 4-43 on page 93.

(0) root @ r1r4m27:(REG) /			
> lspv			
hdisk0	00c931c40c99bd1f	rootvg	active
hdisk1	00c931c492623e86	oravg	active
hdisk2	00c931c42c5def32	None	
hdisk3	00c931c492623fd0	None	
hdisk4	00c931c492624077	None	
hdisk5	00c931c49262410a	None	
hdisk6	00c931c4926241b2	None	
hdisk7	00c931c492624248	vsvgmg1	concurrent
hdisk8	00c931c492624300	vsvgmg1	concurrent
hdisk9	00c931c49262439f	vsvgmg1	concurrent
hdisk10	00c931c49262540d	None	
hdisk11	00c931c4926254b7	None	
hdisk12	00c931c492625570	None	
hdisk13	00c931c49262444f	hyp_vg	concurrent
hdisk14	00c931c4926244de	hyp_vg	concurrent
hdisk15	00c931c49262458a	hyp_vg	concurrent
hdisk16	00c931c492624620	caavg_private	active
hdisk17	00c931c4926246b8	None	
hdisk18	00c931c492624760	None	
hdisk19	00c931c4926247f3	None	
hdisk20	00c931c4926248c2	None	
hdisk21	00c931c49262495a	None	
hdisk22	00c931c492624a21	vsvgcNH	concurrent
hdisk23	00c931c492624abe	vsvgcNH	concurrent
hdisk24	00c931c492624b5c	vsvgcNH	concurrent
hdisk26	00c931c492624c9a	None	

Figure 4-43 Repository disk after HyperSwap configured

System mirror group

The HyperSwap disks used in the user mirror group and cluster repository mirror group are accessible to all nodes in the cluster. However, the system mirror group is used for only one node. The system mirror should not be included in a resource group because it is not managed by PowerHA. Examples of objects for the system mirror group are boot image, system files, paging device, system dump device, system logs, and so on. Usually, all of these objects are included in rootvg. However, they can be in another volume group. Just use the regular AIX tool to change the configuration of these objects to use HyperSwap disks established in 4.11, “Preparation of a HyperSwap-capable disk” on page 73. In particular, the following procedures can be used to configure HyperSwap disk for PowerHA:

1. Configure a HyperSwap-capable disk as described in “Preparation of a HyperSwap-capable disk” on page 73.
2. Clone rootvg with this HyperSwap-capable disk with the following command:
`alt_disk_install -r -C hdisk1`

Note: If the original system disk is a SAN disk that can be configured and is configured as part of a HyperSwap disk pair, then it is not necessary to clone the system disk.

3. Reboot.

Enable the HyperSwap disk with PowerHA as described in Figure 4-44 and Figure 4-45.

Add System Mirror Group

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

[Entry Fields]	
* Mirror Group Name	[]
Volume Group(s)	[] +
Raw Disk(s)	[] +
* Node Name	[] +
* Hyperswap	Disabled [] +
Consistency Group	Enabled [] +
Unplanned HyperSwap Timeout (in sec)	[60] #
Hyperswap Priority	High [] +

Figure 4-44 Adding system mirror group

Name your system mirror group, then select the volume group you want to be HyperSwap mirrored. Remember that this is exclusive for one node. Therefore, you also need to specify which node will use this system mirror group. You also need to enable HyperSwap. The SMIT window for the configuration is shown in Figure 4-45.

Add System Mirror Group

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

[Entry Fields]	
* Mirror Group Name	[sys_mg]
Volume Group(s)	rootvg [] +
Raw Disk(s)	[] +
* Node Name	r1r4m38 [] +
* Hyperswap	Enabled [] +
Consistency Group	Enabled [] +
Unplanned HyperSwap Timeout (in sec)	[60] #
Hyperswap Priority	High [] +

Figure 4-45 Adding a system mirror group, continued

If you dedicated a volume group for paging device, system dump device, or any other system information, you can repeat the procedures described to add them to the system mirror group.

Note: A system mirror group is created for each node.

4.13.4 Create a resource group with site policy

Use SMIT fast path of **smit cm_resource_groups** or follow the smit path **smitty sysmirror** → **Cluster Applications and Resources** → **Resource Groups**. Move the cursor to “Change/Show Nodes and Policies for a Resource Group.” Press Enter. A pick list is generated that displays the choice of resource group, as shown in Figure 4-46.

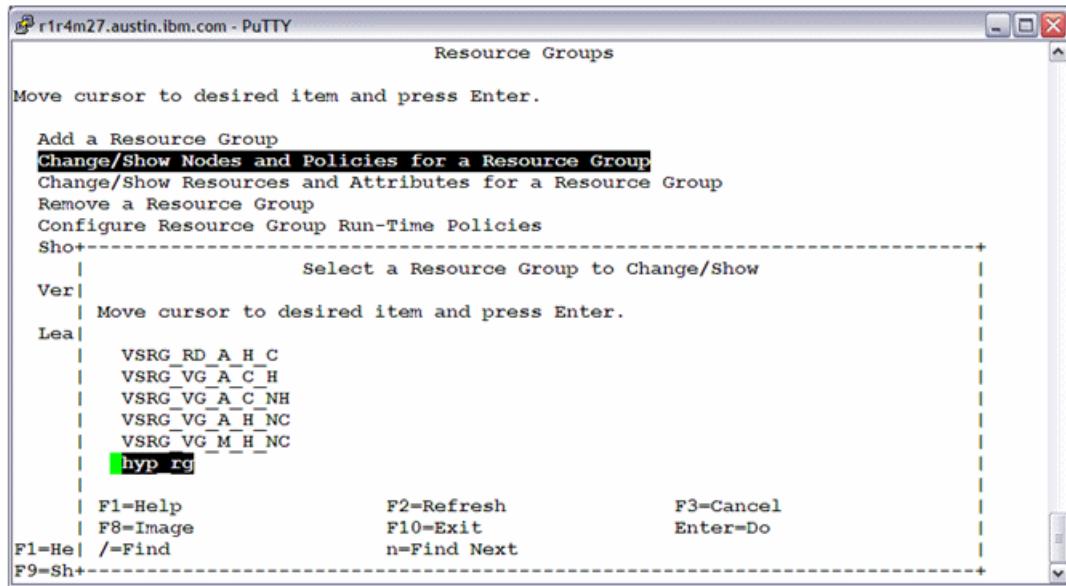


Figure 4-46 Creating the resource group with site policy

If the resource group does not exist, one needs to be created. Otherwise, the window shown in Figure 4-47 displays the configuration of the resource group.

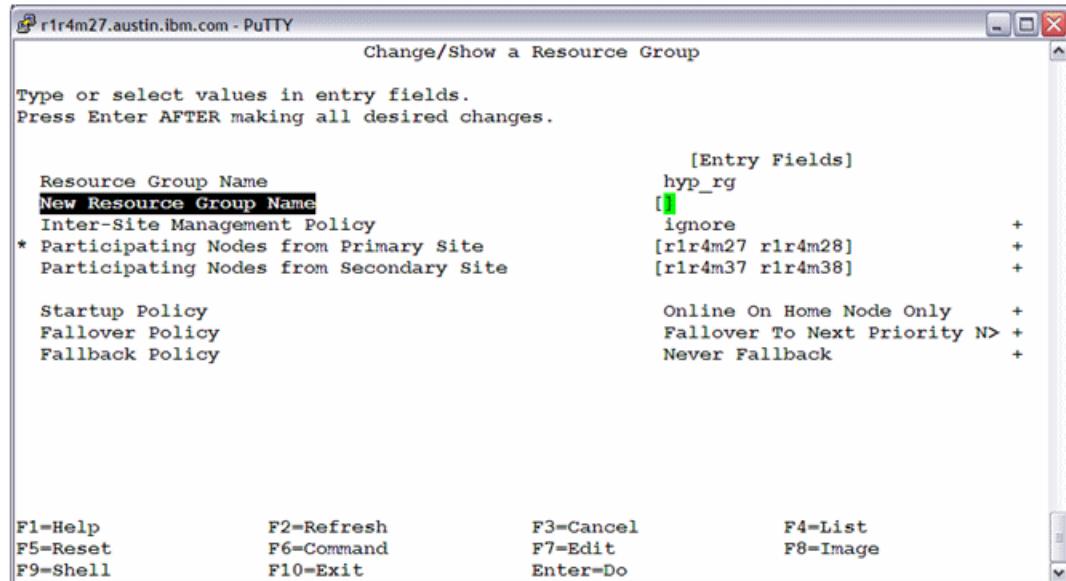


Figure 4-47 Creating the resource group with site policy, continued

Change the Inter-Site Management Policy to Online On Either Site and press Enter to create the resource group shown in Figure 4-48 on page 96.

Important note: For the DS8000 metro mirror resource groups only, the following Inter-Site Management Policy is supported:

- ▶ Prefer Primary Site - In a two-site configuration, replicated resources at startup are on the site with the highest priority, fall over to the other site, and then fall back to the site with the highest priority.
- ▶ Online on Either Site - Replicated resources are on either site at startup, fall over to the other site and remain on that site after failover. This selection simplifies resource group takeover rules, which is helpful if you have a number of resource groups.

For more information, refer to the Planning PowerHA SystemMirror Enterprise Edition for Metro Mirror resource groups at:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/ha_pprc_plan_rg.htm

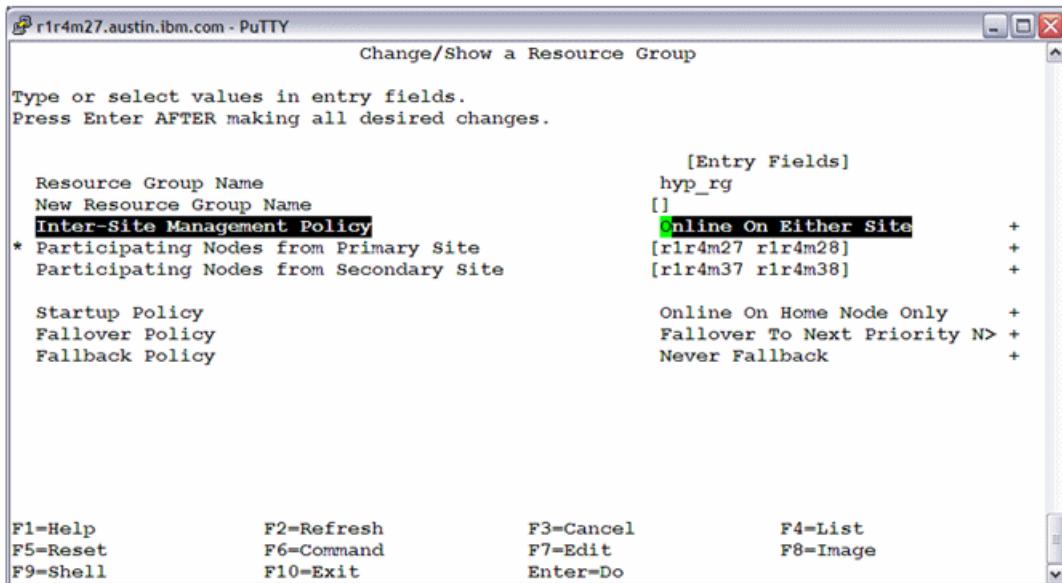


Figure 4-48 Creating the resource group with site policy, continued

4.13.5 Add a mirror group and a volume group into the resource group

Use SMIT fast path `smit cm_resource_groups` or follow the SMIT path `smitty sysmirror` → **Cluster Applications and Resources** → **Resource Groups**. Move the cursor to “Change/Show Resources and Attributes for a Resource Group” and press Enter. A pick list shows the choice of resource groups, as shown in Figure 4-49 on page 97.

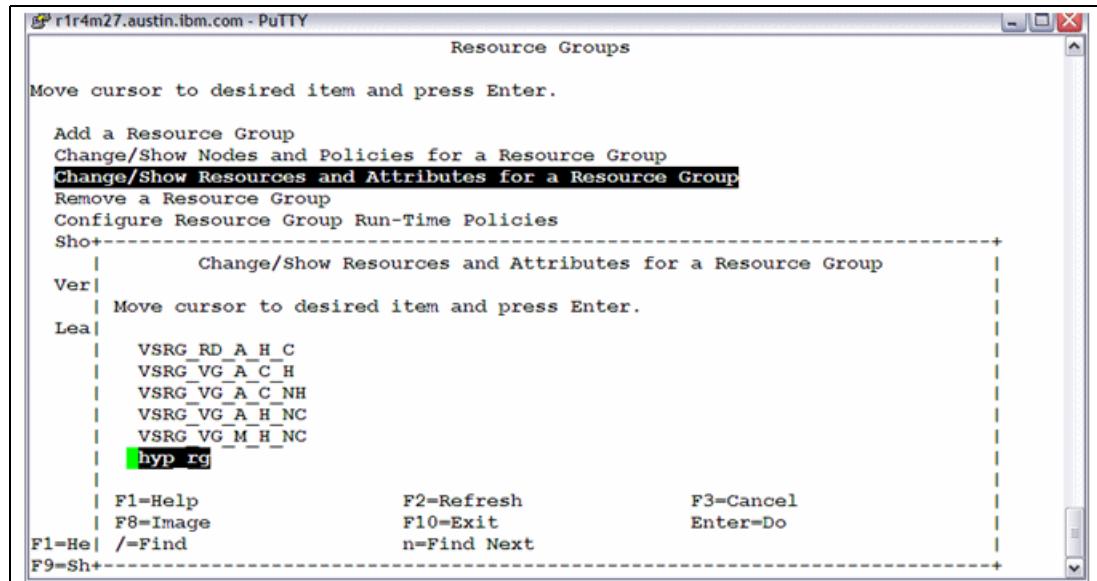


Figure 4-49 Adding the mirror group and volume group into the resource group

Choose **hyp_rg** and then press Enter to see the next action panel. Scroll down to the bottom to the “*DS8000-Metro Mirror (In-band) Resources*” field, as shown in Figure 4-50.

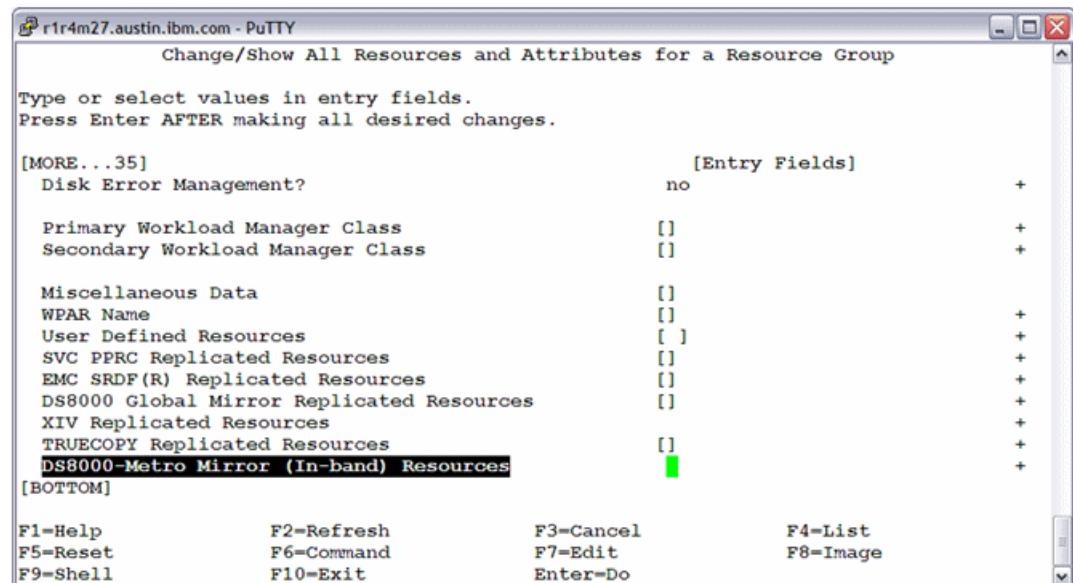


Figure 4-50 Adding the mirror group and volume group into the resource group, continued

Press F4 to get the choice as in Figure 4-51 on page 98. Select the mirror group created in Figure 4-36 on page 89, then press Enter to add the mirror group and then volume group into the resource group.

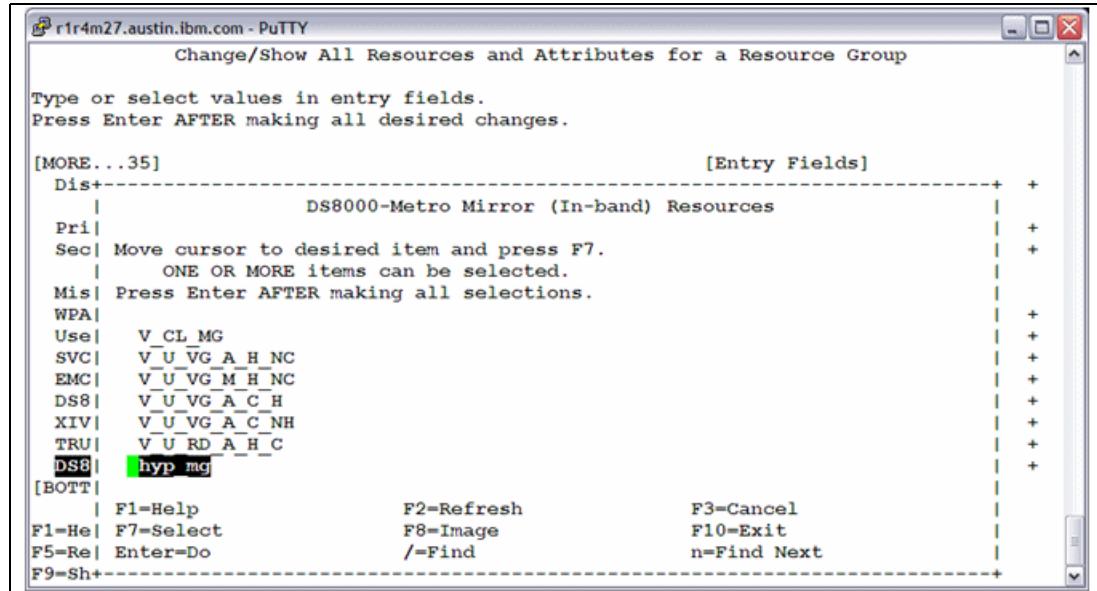


Figure 4-51 Adding the DS8000-Metro Mirror (in-band) resource

4.13.6 Verification and synchronization

Finally, verify whether there is any error in the configuration. If not, do the synchronization with all other nodes. Use SMIT fast path `smit cm_apps_resources` or follow the SMIT path `smitty sysmirror → Cluster Applications and Resources → Verify and Synchronize Cluster Configuration` as shown in Figure 4-52.

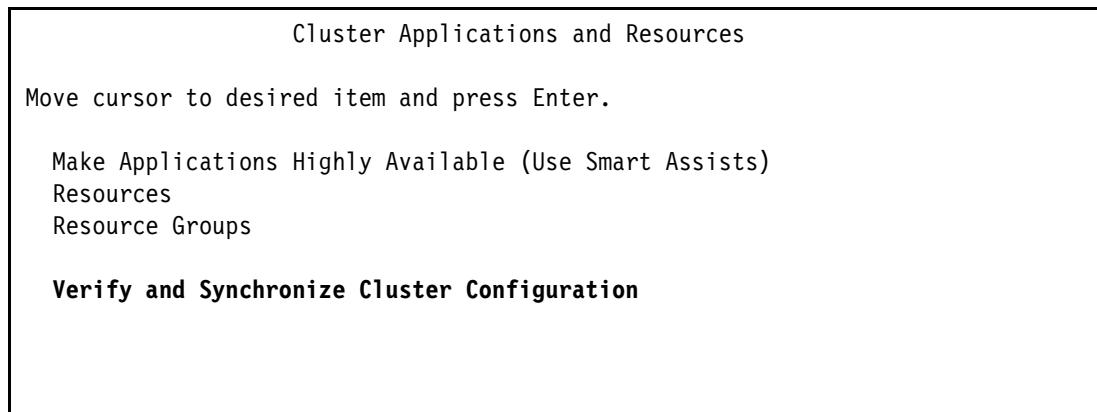


Figure 4-52 Cluster verification and synchronization

Upon completion, if the OK status is displayed, then the configuration is fine.

4.14 Swap demonstration for HyperSwap

Swap can be planned and unplanned. Planned swap can be used for the maintenance of the storage so that there is no interrupt to the application. Unplanned swap can be caused by any hardware or software problems that affect the availability of the active storage. We focused on the impact to the application when the swap occurs.

PowerHA provides the tools for a planned swap. We did the planned swap for all three types of the mirror group.

The unplanned swap is provoked by removing the zone for the connections between the active storage to the active node to force it to make HyperSwap automatically. All the disks on this storage, regardless of the types of mirror group, will be inaccessible after the removing of the zone. Therefore, we did not investigate the unplanned swap per mirror group type as we do for planned swaps. We focused on the impact of unplanned swaps on the applications. The interests to be investigated were:

- ▶ Unplanned swaps for pure write applications
- ▶ Unplanned swaps for pure read applications

We also investigated the tuning that might affect the availability of the applications. Table 4-1 shows the suggested AIX setting for HyperSwap configurations. With the understanding of the impact for various types of transactions, we knew what to expect, the impact, and how we could deal with it.

Table 4-1 Suggested AIX settings for HyperSwap configurations

Tip: Most of these tunables cannot be modified when the device is busy. For best practice, modify the tunables before using the devices. Example of the command to change the tunable:

```
chdev -l myhdisk -a rw_timeout=60
```

4.14.1 Planned swap for user mirror group

We performed a planned swap execution for a user mirror group by using SMIT fast path or smit **cm_user_mirr_gp** or following the SMIT path **smitty sysmirror** → **System Management (C-SPOC)** → **Storage** → **Manage Mirror Groups** → **Manage User Mirror Group(s)**, as shown in Figure 4-53 on page 100.

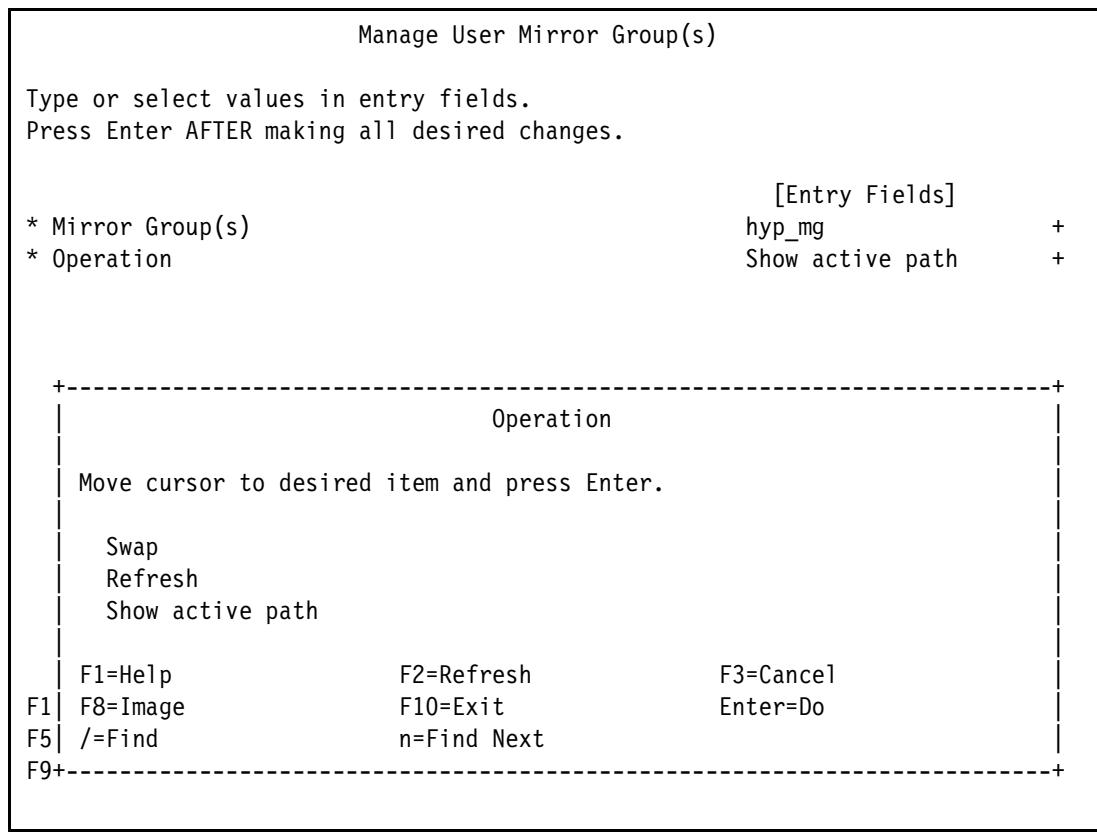


Figure 4-53 Planned swap for user mirror group

Select the mirror group you want to manage. There are three kinds of operations you can manage:

- ▶ Swap: Perform the swap.
- ▶ Refresh: Rediscover the PPRC path characteristics for the specified mirror group.
- ▶ Show active path: Show which half of the PPRC disk pair is actively used.

Before the swap, the status of the HyperSwap disk is seen in Figure 4-54 on page 101. It shows that the active site of this user mirror group is in siteB and the active storage is sys1.

Note: siteB and storage sys1 are defined in 4.13.2, “Defining the storage systems and their site association” on page 86. sys1 and sys2 are storage system names used by PowerHA.

Manage User Mirror Group(s)

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

* Mirror Group(s)
* Operation

[Entry Fields]
hyp_mg +
Show active path +

```
r1r4m27: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE  
r1r4m27: hyp_mg:siteA:siteB:sys1
```

Figure 4-54 Status of user mirror group before planned swap

By selecting **Swap** in Figure 4-53 on page 100, we executed the planned swap for a mirror group as shown in Figure 4-55. SMIT reported the OK status of the swap completion in about four seconds.

Manage User Mirror Group(s)

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

* Mirror Group(s)
* Operation

[Entry Fields]
hyp_mg +
Swap +

Figure 4-55 Planned swap for user mirror group

After the swap, the status of the HyperSwap disk is seen in Figure 4-56. It shows that the active site of this user mirror group changed to siteB and the active storage is changed to sys2.

Manage User Mirror Group(s)

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

* Mirror Group(s)
* Operation

[Entry Fields]
hyp_mg +
Show active path +

```
r1r4m27: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE  
r1r4m27: hyp_mg:siteB:siteA:sys2
```

Figure 4-56 User mirror group status after planned swap

Figure 4-57 shows the **iostat** command output for the HyperSwap disk. It shows about 15% lower throughput for about four seconds (blue part of the output), while the swap is almost unnoticeable for the application.

Disks:	% tm_act	Kbps	tps	Kb_read	Kb_wrt
Thu Dec 6 23:41:30 CST 2012 hdisk14	95.0	207232.0	1612.0	0	207232
Thu Dec 6 23:41:31 CST 2012 hdisk14	98.0	198016.0	1545.0	0	198016
Thu Dec 6 23:41:32 CST 2012 hdisk14	97.0	211456.0	1618.0	0	211456
Thu Dec 6 23:41:33 CST 2012 hdisk14	100.0	205056.0	1602.0	0	205056
Thu Dec 6 23:41:34 CST 2012 hdisk14	92.0	196096.0	1531.0	0	196096
Thu Dec 6 23:41:35 CST 2012 hdisk14	94.0	185472.0	1399.0	0	185472
<i>swap starts here. It takes about 4 seconds for SMIT to report the completion of swap operation.</i>					
Thu Dec 6 23:41:36 CST 2012 hdisk14	93.0	204928.0	1564.0	0	204928
Thu Dec 6 23:41:37 CST 2012 hdisk14	94.0	187904.0	1459.0	0	187904
Thu Dec 6 23:41:38 CST 2012 hdisk14	91.1	165632.0	1294.0	0	165632
Thu Dec 6 23:41:39 CST 2012 hdisk14	86.0	168832.0	1315.0	0	168832
Thu Dec 6 23:41:40 CST 2012 hdisk14	85.0	168192.0	1237.0	0	168192
Thu Dec 6 23:41:41 CST 2012 hdisk14	86.0	206464.0	1561.0	0	206464
Thu Dec 6 23:41:42 CST 2012 hdisk14	87.6	200192.0	1542.0	0	200192
Thu Dec 6 23:41:43 CST 2012 hdisk14	92.0	203136.0	1577.0	0	203136
Thu Dec 6 23:41:44 CST 2012 hdisk14	93.0	208640.0	1624.0	0	208640

Figure 4-57 iostat command output for the planned swap of a user mirror group with high I/O load

Sample collection from Figure 4-57 is performed with a large I/O load. It is over 200 MB/s. For a lower I/O load, the swap is almost transparent to the application even though it still takes the same amount of time (four seconds) to complete the planned swap. Figure 4-58 on page 103 shows this situation with about a 1 MB/s I/O load. I/O throughput is degraded for only one second and by about 10% throughput.

Disks:	% tm_act	Kbps	tps	Kb_read	Kb_wrt
Fri Dec 7 22:06:04 CST 2012	hdisk14	0.0	1152.0	9.0	0 1152
Fri Dec 7 22:06:05 CST 2012	hdisk14	0.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:06 CST 2012	hdisk14	1.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:07 CST 2012	hdisk14	3.0	1280.0	10.0	0 1280
Fri Dec 7 22:06:08 CST 2012	hdisk14	0.0	1152.0	9.0	0 1152
Fri Dec 7 22:06:09 CST 2012	hdisk14	0.0	1152.0	9.0	0 1152
<i>swap starts here. It takes about 4 seconds for SMIT to report the completion of swap operation.</i>					
Fri Dec 7 22:06:10 CST 2012	hdisk14	0.0	896.0	7.0	0 896
Fri Dec 7 22:06:11 CST 2012	hdisk14	1.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:12 CST 2012	hdisk14	4.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:13 CST 2012	hdisk14	0.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:14 CST 2012	hdisk14	1.0	1152.0	9.0	0 1152
Fri Dec 7 22:06:15 CST 2012	hdisk14	1.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:16 CST 2012	hdisk14	3.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:17 CST 2012	hdisk14	1.0	1408.0	11.0	0 1408
Fri Dec 7 22:06:18 CST 2012	hdisk14	0.0	1152.0	9.0	0 1152
Fri Dec 7 22:06:19 CST 2012	hdisk14	0.0	1024.0	8.0	0 1024
Fri Dec 7 22:06:20 CST 2012	hdisk14	1.0	1024.0	8.0	0 1024

Figure 4-58 iostat output for the planned swap of user mirror group with low I/O load

Based on these two tests with different I/O loads, we concluded that:

- ▶ It takes about 4 seconds to complete the swap, regardless of the I/O load.
- ▶ Other than minor interrupt to I/O performance, a planned swap is almost unnoticeable to the application.
- ▶ Higher I/O load can have higher impact to I/O performance around the time of the swap. However, it does not seem to be any issue to the application.

4.14.2 Planned swap for a cluster repository mirror group

You can execute a planned swap execution for a repository disk with SMIT fast path **smit cm_cluser_repos_mirr_gp** or follow the smit path **smitty sysmirror → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage Cluster Repository Mirror Group**, as shown in Figure 4-59 on page 104.

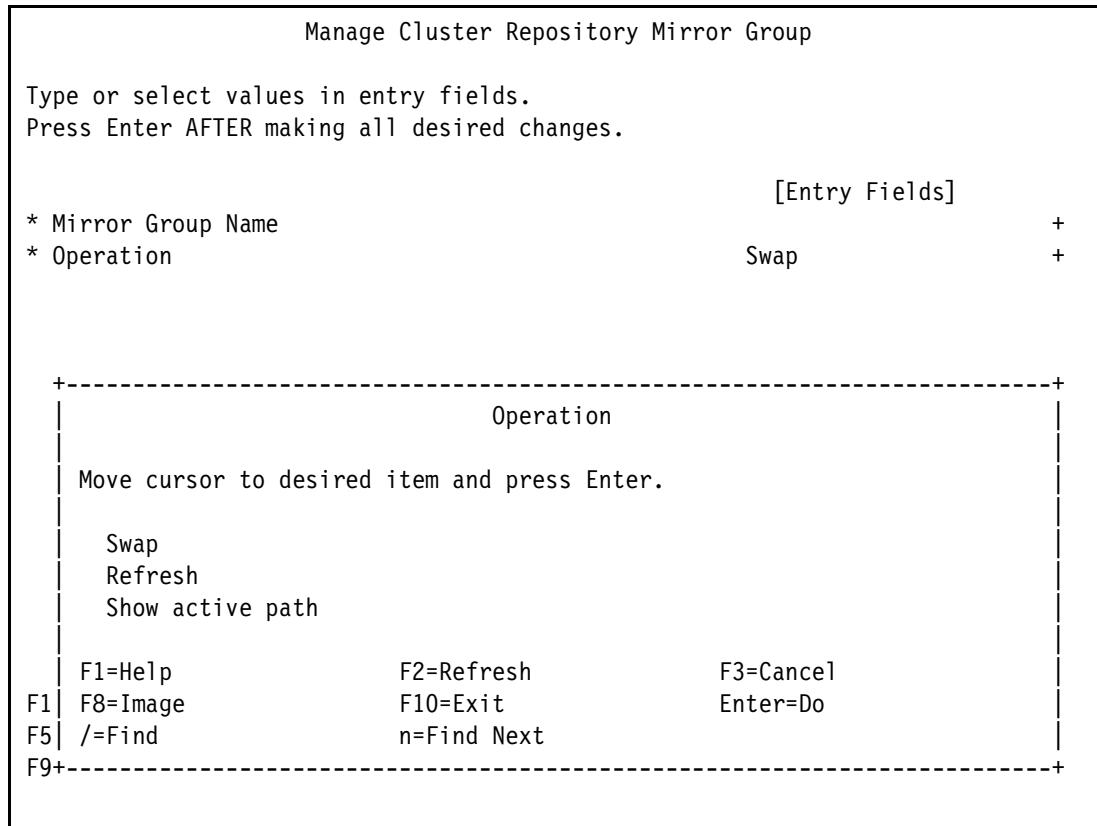


Figure 4-59 Planned swap for a cluster repository mirror group

Select the mirror group you want to manage. There are three kinds of operations you can manage:

- ▶ Swap: Perform the swap.
- ▶ Refresh: Rediscover the PPRC path characteristics for the specified mirror group.
- ▶ Show active path: Show which half of the PPRC disk pair is actively used.

Before the swap, the status of the HyperSwap disk is shown in Figure 4-60 on page 105. It shows that the active site of the cluster repository mirror group is in siteB and the active storage is sys1. Unlike the user mirror group, which is not cluster-wide, the cluster repository disk is cluster-wide. Therefore, it shows the view of the repository disk from each node in the cluster. All nodes need to see the same repository disk.

Manage Cluster Repository Mirror Group

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

* Mirror Group Name * Operation	[Entry Fields] repos_mg + Show active path +
------------------------------------	--

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

```
r1r4m27: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m27: repos_mg:siteA:siteB:sys1
r1r4m28: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m28: repos_mg:siteA:siteB:sys1
r1r4m37: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m37: repos_mg:siteA:siteB:sys1
r1r4m38: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m38: repos_mg:siteA:siteB:sys1
```

Figure 4-60 Status of the cluster repository mirror group before the planned swap

By selecting **Swap** in Figure 4-59 on page 104, we executed the planned swap for the cluster repository mirror group, as shown in Figure 4-61. SMIT reports the OK status of swap completion within one second.

Manage Cluster Repository Mirror Group

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

* Mirror Group Name * Operation	[Entry Fields] repos_mg + Swap +
------------------------------------	--

Figure 4-61 Planned swap for cluster repository mirror group

After the swap, the status of the HyperSwap disk is seen in Figure 4-62 on page 106. It shows that the active site of this repository mirror group is changed to siteB and the active storage is changed to sys2.

```

Manage Cluster Repository Mirror Group

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

[Entry Fields]
* Mirror Group Name      repos_mg      +
* Operation               Show active path      +


COMMAND STATUS

Command: OK          stdout: yes          stderr: no

Before command completion, additional instructions may appear below.

r1r4m27: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m27: repos_mg:siteB:siteA:sys2
r1r4m28: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m28: repos_mg:siteB:siteA:sys2
r1r4m37: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m37: repos_mg:siteB:siteA:sys2
r1r4m38: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m38: repos_mg:siteB:siteA:sys2

```

Figure 4-62 Cluster repository mirror group status after planned swap

Even though there is no user data in a repository disk, it is interesting to know how the swap of a repository disk would affect user data and the application. Figure 4-63 on page 107 shows the **iostat** command output of a user mirror group during the planned swap for the repository disk. The report shows some drop of I/O throughput for one second (blue part of the output). Other than that, the swap is almost unnoticeable. The repository disk and the user I/O load are in the same storage. It is speculated that the effect of the swap would be lower if they were not using the same storage.

```

Disks:          % tm_act     Kbps      tps   Kb_read   Kb_wrt
Thu Dec 6 23:02:38 CST 2012 hdisk14 88.0 161792.0 1264.0 0 161792
Thu Dec 6 23:02:39 CST 2012 hdisk14 86.0 161152.0 1259.0 0 161152
Thu Dec 6 23:02:40 CST 2012 hdisk14 90.0 161792.0 1264.0 0 161792
Thu Dec 6 23:02:42 CST 2012 hdisk14 93.1 56880.0 6147.0 0 56880
Thu Dec 6 23:02:42 CST 2012 hdisk14 96.0 79328.0 2096.0 0 79328
Thu Dec 6 23:02:43 CST 2012 hdisk14 92.0 163584.0 1278.0 0 163584
Thu Dec 6 23:02:44 CST 2012 hdisk14 88.0 161408.0 1261.0 0 161408
Thu Dec 6 23:02:45 CST 2012 hdisk14 89.0 159232.0 1244.0 0 159232
Thu Dec 6 23:02:47 CST 2012 hdisk14 92.1 31348.0 5996.0 0 31348
Thu Dec 6 23:02:47 CST 2012 hdisk14 96.0 94748.0 2168.0 0 94748
Thu Dec 6 23:02:48 CST 2012 hdisk14 82.0 160000.0 1250.0 0 160000
Thu Dec 6 23:02:49 CST 2012 hdisk14 88.0 158720.0 1240.0 0 158720
Thu Dec 6 23:02:50 CST 2012 hdisk14 84.0 156800.0 1225.0 0 156800
swap starts. smit report swap completion within one second.
Thu Dec 6 23:02:52 CST 2012 hdisk14 95.4 17384.0 4180.0 0 17384
Thu Dec 6 23:02:52 CST 2012 hdisk14 90.0 104232.0 3949.0 0 104232
Thu Dec 6 23:02:53 CST 2012 hdisk14 79.0 138496.0 1044.0 0 138496
Thu Dec 6 23:02:54 CST 2012 hdisk14 77.0 138880.0 1085.0 0 138880
Thu Dec 6 23:02:55 CST 2012 hdisk14 83.0 152272.0 1459.0 0 152272
Thu Dec 6 23:02:57 CST 2012 hdisk14 86.4 69184.0 6935.0 0 69184

```

Figure 4-63 iostat output for the planned swap of a cluster repository mirror group

4.14.3 Planned swap for a system mirror group

You can execute a planned swap execution for a system mirror group with the SMIT fast path **smit cm_system_mirr_gp** or follow the SMIT path **smitty sysmirror → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage System Mirror Group**, as shown in Figure 4-64 on page 108.

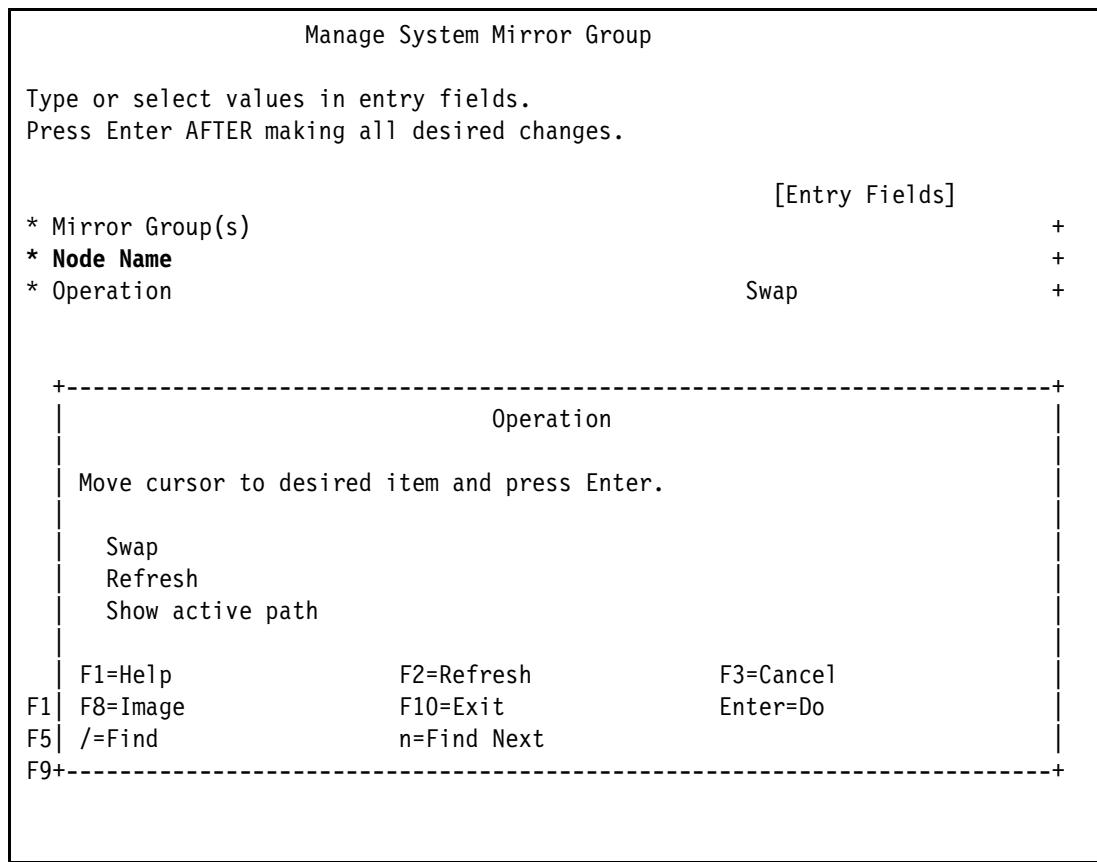


Figure 4-64 Planned swap for a system mirror group

Unlike the menus for both the cluster repository mirror group and the user mirror group, this menu has an additional field, Node Name, to fill. This is because the system mirror group is per node. It is not necessary to get all nodes to have a system mirror group. It is all up to the implementation. One implementation is to have a system mirror group on the primary site and no system user group on the secondary site.

Select the mirror group you want to manage. There are three kinds of operations you can manage:

- ▶ Swap - Perform the swap.
- ▶ Show active path - Show which half of the PPRC disk pair is actively used.
- ▶ Refresh - Rediscover the PPRC path characteristics for the specified mirror group.

The status of the HyperSwap disk before the swap is seen in Figure 4-65 on page 109. It shows that the active site of this user mirror group is in siteB and the active storage is sys1.

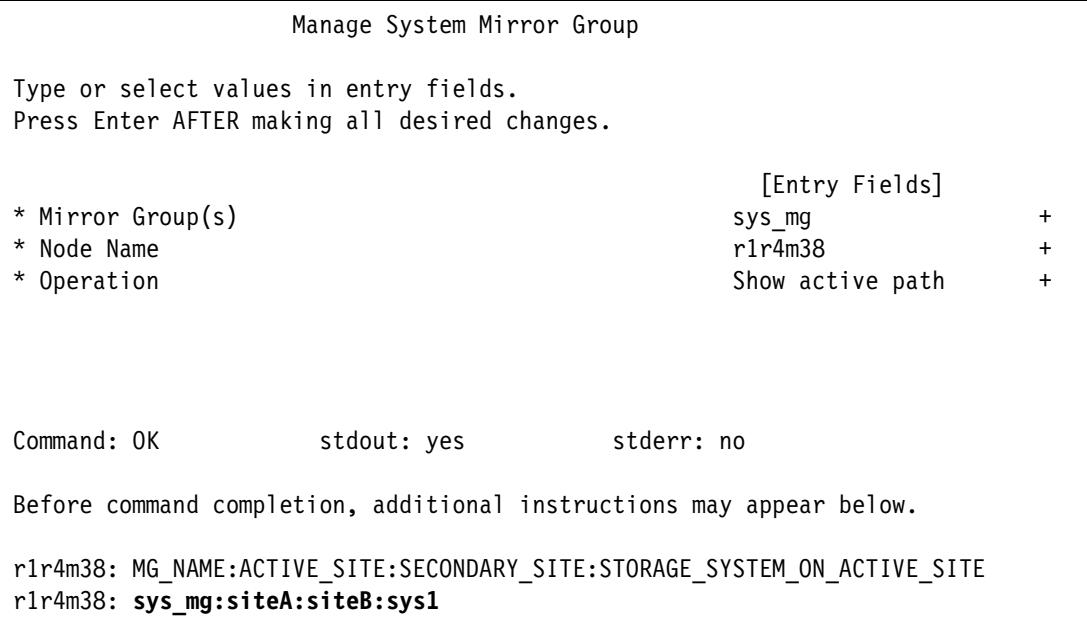


Figure 4-65 System mirror group status before a planned swap

By selecting **Swap** in Figure 4-64 on page 108, execute the planned swap for the system mirror group as shown in Figure 4-66. SMIT reports the OK status of the swap completion in one second.

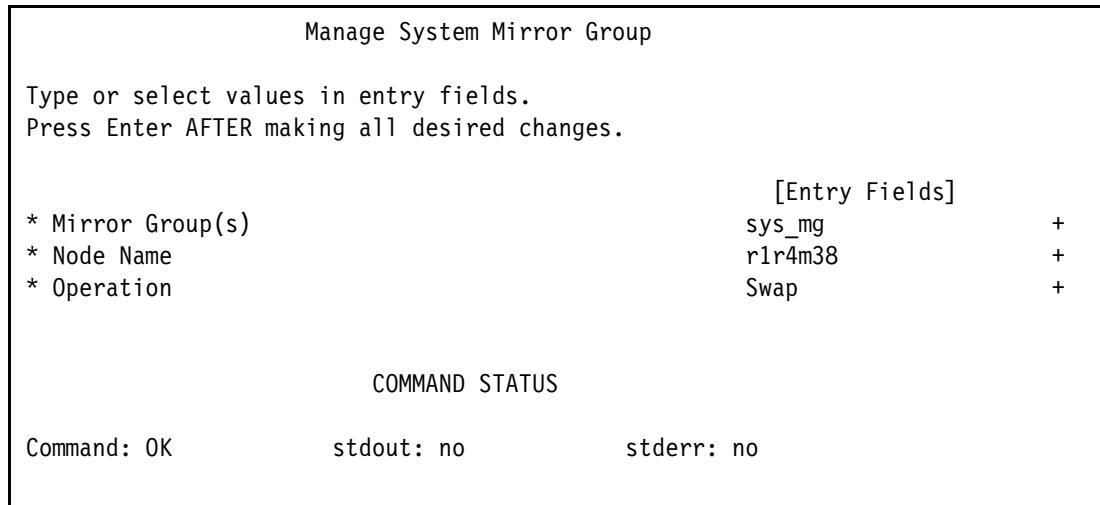


Figure 4-66 Planned swap for a user mirror group

After the swap, the status of the HyperSwap disk is shown in Figure 4-67 on page 110. It shows that the active site of this user mirror group changed to siteB and the active storage is changed to sys2.

```

Manage System Mirror Group

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

[Entry Fields]
* Mirror Group(s)          sys_mg      +
* Node Name                 r1r4m38    +
* Operation                  Show active path +


COMMAND STATUS

Command: OK           stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

r1r4m38: MG_NAME:ACTIVE_SITE:SECONDARY_SITE:STORAGE_SYSTEM_ON_ACTIVE_SITE
r1r4m38: sys_mg:siteB:siteA:sys2

```

Figure 4-67 Status of the system mirror group after a planned swap

It takes less than a couple of seconds to complete the planned swap for the system mirror group. This swap is just like that of a user mirror group. The difference is per node and it is considered a user mirror group with the operating system as the application. Other than this difference, it is just like a user mirror group. Therefore, it is expected that the impact of the planned swap of a system mirror group is just like that of a user mirror group. It is almost unnoticeable.

4.14.4 Unplanned swap for pure write applications

It takes a much longer time to complete an unplanned swap than a planned swap. Therefore, we monitored the impact from the point of view of application and system.

Figure 4-68 on page 111 shows the application view of an unplanned swap with pure write at a load of about 150 KB/s. The storage becomes inaccessible at 21:59:01. The continuity of the listing of time indicates that the application proceeded without knowing that the storage was not accessible.

```
Tue Dec 25 21:58:50 CST 2012
Tue Dec 25 21:58:51 CST 2012
Tue Dec 25 21:58:52 CST 2012
Tue Dec 25 21:58:53 CST 2012
Tue Dec 25 21:58:54 CST 2012
Tue Dec 25 21:58:55 CST 2012
Tue Dec 25 21:58:56 CST 2012
Tue Dec 25 21:58:57 CST 2012
Tue Dec 25 21:58:58 CST 2012
Tue Dec 25 21:58:59 CST 2012
Tue Dec 25 21:59:01 CST 2012
primary storage becomes unaccessible at 21:59:01.
Tue Dec 25 21:59:02 CST 2012
Tue Dec 25 21:59:03 CST 2012
Tue Dec 25 21:59:04 CST 2012
Tue Dec 25 21:59:05 CST 2012
Tue Dec 25 21:59:06 CST 2012
Tue Dec 25 21:59:07 CST 2012
Tue Dec 25 21:59:08 CST 2012
Tue Dec 25 21:59:09 CST 2012
Tue Dec 25 21:59:10 CST 2012
Tue Dec 25 21:59:11 CST 2012
Tue Dec 25 21:59:12 CST 2012
Tue Dec 25 21:59:14 CST 2012
Tue Dec 25 21:59:15 CST 2012
Tue Dec 25 21:59:16 CST 2012
Tue Dec 25 21:59:17 CST 2012
Tue Dec 25 21:59:18 CST 2012
Tue Dec 25 21:59:19 CST 2012
Tue Dec 25 21:59:20 CST 2012
Tue Dec 25 21:59:21 CST 2012
Tue Dec 25 21:59:23 CST 2012
Tue Dec 25 21:59:24 CST 2012
Tue Dec 25 21:59:25 CST 2012
Tue Dec 25 21:59:26 CST 2012
Tue Dec 25 21:59:27 CST 2012
Tue Dec 25 21:59:28 CST 2012
Tue Dec 25 21:59:29 CST 2012
Tue Dec 25 21:59:30 CST 2012
Tue Dec 25 21:59:31 CST 2012
Tue Dec 25 21:59:34 CST 2012
Tue Dec 25 21:59:35 CST 2012
Tue Dec 25 21:59:37 CST 2012
Tue Dec 25 21:59:38 CST 2012
Tue Dec 25 21:59:39 CST 2012
Tue Dec 25 21:59:40 CST 2012
```

Figure 4-68 Application view of an unplanned swap for pure write of about 150 KB/s

Figure 4-69 on page 112 shows the system view of the unplanned swap for a pure write of about 150 KB/s. The AIX tool **iostat** was used to monitor the actual I/O activity around the time of the swap. The system had about 31 seconds of I/O hang. Right after the swap was completed at 21:59:33, the system resumed I/O with a spike to process all queued I/Os. It is

seen that all the accumulated I/O during 31 seconds (I/O hang) was processed in one second.

Disks:	% tm_act	Kbps	tps	Kb_read	Kb_wrt
Tue Dec 25 21:58:52 CST 2012 oradisk2	0.0	256.0	2.0	0	256
Tue Dec 25 21:58:53 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:58:54 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:58:55 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:58:56 CST 2012 oradisk2	0.0	256.0	2.0	0	256
Tue Dec 25 21:58:57 CST 2012 oradisk2	1.0	128.0	1.0	0	128
Tue Dec 25 21:58:58 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:58:59 CST 2012 oradisk2	0.0	256.0	2.0	0	256
Tue Dec 25 21:59:00 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:59:01 CST 2012 oradisk2	0.0	128.0	1.0	0	128
primary storage becomes unaccessible at 21:59:01.					
Tue Dec 25 21:59:02 CST 2012 oradisk2	9.0	0.0	0.0	0	0
Tue Dec 25 21:59:03 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:04 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:05 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:06 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:07 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:08 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:09 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:10 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:11 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:12 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:13 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:14 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:15 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:16 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:17 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:18 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:19 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:20 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:21 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:22 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:23 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:24 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:25 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:26 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:27 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:28 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:29 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:30 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:31 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:32 CST 2012 oradisk2	100.0	0.0	0.0	0	0
Tue Dec 25 21:59:33 CST 2012 oradisk2 49.5 4640.0 36.0 0 4640					
I0 spike to process all I0 queued in the previous 31 second and the current time. 4640/32=145(KB/s) is roughly about the same as the average I0 load in the system.					
Tue Dec 25 21:59:34 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:59:35 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:59:36 CST 2012 oradisk2	1.0	128.0	1.0	0	128
Tue Dec 25 21:59:37 CST 2012 oradisk2	0.0	256.0	2.0	0	256
Tue Dec 25 21:59:38 CST 2012 oradisk2	0.0	128.0	1.0	0	128
Tue Dec 25 21:59:39 CST 2012 oradisk2	0.0	128.0	1.0	0	128

Figure 4-69 System view of an unplanned swap for a pure write of about 150 KB/s

We were also interested in knowing how the swap affects the application if the I/O load is higher. Figure 4-70 shows the same pure write application but with a higher I/O load of 10 MB/s. Unlike the previous example of a lower I/O load in Figure 4-68 on page 111 where the application does not hang, this swap with higher I/O load causes the application to hang 2 or 3 seconds after the storage is inaccessible. However, higher I/O does not affect the time to complete the swap. It still takes about 30 seconds with a lower I/O load.

```
Fri Dec 28 06:27:53 CST 2012
Fri Dec 28 06:27:54 CST 2012
Fri Dec 28 06:27:56 CST 2012
Fri Dec 28 06:27:57 CST 2012
Fri Dec 28 06:27:58 CST 2012
Fri Dec 28 06:27:59 CST 2012
Fri Dec 28 06:28:00 CST 2012
primary storage becomes unaccessible at 06:28:00.
Fri Dec 28 06:28:01 CST 2012
Fri Dec 28 06:28:02 CST 2012
Fri Dec 28 06:28:03 CST 2012
application hang for about 30 seconds.
Fri Dec 28 06:28:34 CST 2012
Fri Dec 28 06:28:35 CST 2012
Fri Dec 28 06:28:36 CST 2012
Fri Dec 28 06:28:37 CST 2012
Fri Dec 28 06:28:38 CST 2012
Fri Dec 28 06:28:39 CST 2012
Fri Dec 28 06:28:41 CST 2012
Fri Dec 28 06:28:42 CST 2012
Fri Dec 28 06:28:43 CST 2012
Fri Dec 28 06:28:44 CST 2012
Fri Dec 28 06:28:45 CST 2012
Fri Dec 28 06:28:46 CST 2012
Fri Dec 28 06:28:47 CST 2012
Fri Dec 28 06:28:48 CST 2012
Fri Dec 28 06:28:49 CST 2012
Fri Dec 28 06:28:50 CST 2012
```

Figure 4-70 Application view of an unplanned swap for a pure write of 10 MB/s

Figure 4-71 on page 114 shows the system view of an unplanned swap for a pure write of about 10 MB/s. The AIX tool **iostat** was used to monitor the actual I/O activity around the time of the swap. The system had about 31 seconds of I/O hang. Right after the swap was completed at 06:28:33, the system resumed I/O with a small I/O spike to process the queued I/O of about 2 to 3 seconds before the application hung.

```

Disks:      % tm_act     Kbps      tps   Kb_read   Kb_wrtm
Fri Dec 28 06:27:55 CST 2012 oradisk2 9.0 10880.0 170.0 0 10880
Fri Dec 28 06:27:56 CST 2012 oradisk2 7.0 10752.0 168.0 0 10752
Fri Dec 28 06:27:57 CST 2012 oradisk2 7.0 10756.0 169.0 4 10752
Fri Dec 28 06:27:58 CST 2012 oradisk2 8.0 10880.0 170.0 0 10880
Fri Dec 28 06:27:59 CST 2012 oradisk2 8.0 10752.0 168.0 0 10752
Fri Dec 28 06:28:00 CST 2012 oradisk2 8.0 10756.0 169.0 4 10752
Fri Dec 28 06:28:01 CST 2012 oradisk2 8.0 10880.0 170.0 0 10880
Fri Dec 28 06:28:02 CST 2012 oradisk2 25.0 0.0 0.0 0 0
Fri Dec 28 06:28:03 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:04 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:05 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:06 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:07 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:08 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:09 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:10 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:11 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:12 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:13 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:14 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:15 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:16 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:17 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:18 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:19 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:20 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:21 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:22 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:23 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:24 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:25 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:26 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:27 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:28 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:29 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:30 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:31 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:32 CST 2012 oradisk2 100.0 0.0 0.0 0 0
Fri Dec 28 06:28:33 CST 2012 oradisk2 53.0 21508.0 155.0 4 21504
Small IO spike to process the queued IO of about 2 to 3 second before the application hang.
Fri Dec 28 06:28:34 CST 2012 oradisk2 9.0 10880.0 170.0 0 10880
Fri Dec 28 06:28:35 CST 2012 oradisk2 6.0 10752.0 168.0 0 10752
Fri Dec 28 06:28:36 CST 2012 oradisk2 9.0 10756.0 169.0 4 10752
Fri Dec 28 06:28:37 CST 2012 oradisk2 9.0 10752.0 168.0 0 10752
Fri Dec 28 06:28:38 CST 2012 oradisk2 6.0 6912.0 108.0 0 6912
Fri Dec 28 06:28:39 CST 2012 oradisk2 3.0 3968.0 62.0 0 3968
Fri Dec 28 06:28:40 CST 2012 oradisk2 8.0 10756.0 169.0 4 10752
Fri Dec 28 06:28:41 CST 2012 oradisk2 9.0 10752.0 168.0 0 10752
Fri Dec 28 06:28:42 CST 2012 oradisk2 8.0 10880.0 170.0 0 10880

```

Figure 4-71 System view of an unplanned swap for a pure write of 10 MB/s

Based on these two tests with different I/O loads, we concluded that:

- ▶ It takes about 30 seconds to complete the swap, regardless of the amount of I/O load.
- ▶ If the system has enough buffer to queue the I/O for 30 seconds, the application does not hang at all.
- ▶ We tried some tuning and found that the `rw_timeout` value of the disk attribute can affect the time to complete the swap. When we set it to 60 seconds with the command `chdev -1 myhdisk - a rw_timeout=60`, the swap time became about 60 seconds. The minimum and default values of `rw_timeout` are 30 seconds. See “Swap demonstration for HyperSwap” on page 98 for information about some HyperSwap-related tunables.

4.14.5 Unplanned swap for pure read applications

Unlike the write transaction, which can be buffered in the system so that applications can proceed without waiting for the completion of data being written to the storage, even a small amount of I/O read can hang the application when the storage is not accessible. Figure 4-72 shows the application view of an unplanned swap with pure read. The storage becomes inaccessible at about 11:10:14 as shown in Figure 4-73 on page 116 and the application hangs immediately when the storage is inaccessible. The application stays hung for about 30 seconds until the swap is completed.

```
Tue Jan 1 11:09:45 CST 2013
Tue Jan 1 11:09:49 CST 2013
Tue Jan 1 11:09:53 CST 2013
Tue Jan 1 11:09:56 CST 2013
Tue Jan 1 11:09:59 CST 2013
Tue Jan 1 11:10:02 CST 2013
Tue Jan 1 11:10:06 CST 2013
Tue Jan 1 11:10:09 CST 2013
Tue Jan 1 11:10:12 CST 2013
primary storage becomes unaccessible shortly after 11:10:12
The application hangs for about 35 seconds.
Tue Jan 1 11:10:49 CST 2013
Tue Jan 1 11:11:03 CST 2013
Tue Jan 1 11:11:12 CST 2013
Tue Jan 1 11:11:22 CST 2013
Tue Jan 1 11:11:31 CST 2013
Tue Jan 1 11:11:42 CST 2013
Tue Jan 1 11:11:54 CST 2013
Tue Jan 1 11:12:07 CST 2013
Tue Jan 1 11:12:16 CST 2013
```

Figure 4-72 Application view of an unplanned swap for a pure read

```

Disks:      % tm_act     Kbps      tps   Kb_read   Kb_wrtm
Tue Jan 1 11:10:10 CST 2013 oradisk2 82.0 253960.0 520.0 253960 0
Tue Jan 1 11:10:11 CST 2013 oradisk2 81.0 216576.0 439.0 216576 0
Tue Jan 1 11:10:12 CST 2013 oradisk2 25.0 60924.0 124.0 60924 0
Tue Jan 1 11:10:13 CST 2013 oradisk2 50.0 136708.0 298.0 136708 0
Tue Jan 1 11:10:14 CST 2013 oradisk2 78.0 205316.0 422.0 205312 4
primary storage becomes unaccessible at 11:10:14
System IO hangs for about 30 seconds showing that it takes about 30 seconds to complete the swap.
Tue Jan 1 11:10:15 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:16 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:17 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:18 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:19 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:20 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:21 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:22 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:23 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:24 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:25 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:26 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:27 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:28 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:29 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:30 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:31 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:32 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:33 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:34 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:35 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:36 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:37 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:38 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:39 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:40 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:41 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:42 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:43 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:44 CST 2013 oradisk2 100.0 0.0 0.0 0 0
Tue Jan 1 11:10:45 CST 2013 oradisk2 43.0 74848.0 188.0 74848 0
Tue Jan 1 11:10:46 CST 2013 oradisk2 23.0 48640.0 100.0 48640 0
Tue Jan 1 11:10:47 CST 2013 oradisk2 10.0 21504.0 44.0 21504 0
Tue Jan 1 11:10:48 CST 2013 oradisk2 16.0 37376.0 76.0 37376 0
Tue Jan 1 11:10:49 CST 2013 oradisk2 4.0 7164.0 15.0 7164 0
Tue Jan 1 11:10:50 CST 2013 oradisk2 17.8 43528.0 96.0 43528 0

```

Figure 4-73 System view of an unplanned swap for a pure read

4.14.6 Summary for HyperSwap performance

The swap time of the HyperSwap disk depends on the number of nodes and disks involved and is not related to the I/O load. However, higher I/O load makes the application more

sensitive to the swap. Table 4-2 describes the swap time and how the application feels the swap event in several different scenarios.

Table 4-2 Swap time and its effect to the application for various scenarios

Swap type	Swap time	Transparent to application
Planned user mirror group	4 seconds	Yes if no read and reasonable amount of write, otherwise almost transparent.
Planned system mirror group	< 1 second	Yes.
Planned repository mirror group	< 1 second	Yes.
Unplanned swap for a pure write application	30 seconds (tunable)	Yes if I/O write is not too high.
Unplanned swap for a pure read application	30 seconds (tunable)	No. Application hangs in the entire duration of the swap.

4.15 Oracle standalone database in a PowerHA HyperSwap environment

This section describes an implementation scenario for a PowerHA SystemMirror 7.1.2 Enterprise Edition cluster with Oracle Database version 11gR2 in a standalone configuration. This environment uses a local cluster configuration in a primary site with two nodes, extended with a third node in a secondary location, and a single resource group containing the database application. The database volumes are placed on IBM DS8800 storage systems which are replicated across the sites using DS8000 synchronous replication (Metro Mirror).

In this environment, we configure the cluster DS8000 Metro Mirror (in-band) replicated resources demonstrating the HyperSwap and the disaster recovery capabilities of the environment when using the PowerHA SystemMirror software.

In this section we do not detail all the steps required to configure the HyperSwap disks as they are already described in previous sections (4.10, “Initial disk configuration” on page 72 and in 4.15.2, “HyperSwap disk configuration” on page 119). Also, for the details of installing the Oracle database, creating a database, and the listener configuration refer to the Oracle database documentation library at:

http://www.oracle.com/pls/db112/portal.all_books

4.15.1 The environment description

The environment used for implementing the Oracle database system with the IBM PowerHA Enterprise Edition is based on the hardware configuration used for implementing the PowerHA HyperSwap as described in 4.9, “Test environment of a HyperSwap cluster” on page 69.

The Oracle database cluster environment is a stretched cluster containing:

- ▶ Two sites named siteA and siteB
- ▶ Three nodes - Two nodes in the primary site siteA: r1r4m27 and r1r4m28 and a third node, r1r4m37 in the secondary site siteB. The nodes have AIX 6.1 TL8 SP1 installed and PowerHA 7.1.2 SP1. The Oracle database version used in our environment was 11.2.0.1.
- ▶ Two IBM storage subsystems, model DS8800 named DS8K4 in siteA and DS8K6 in siteB.

Figure 4-74 on page 118 illustrates the environment.

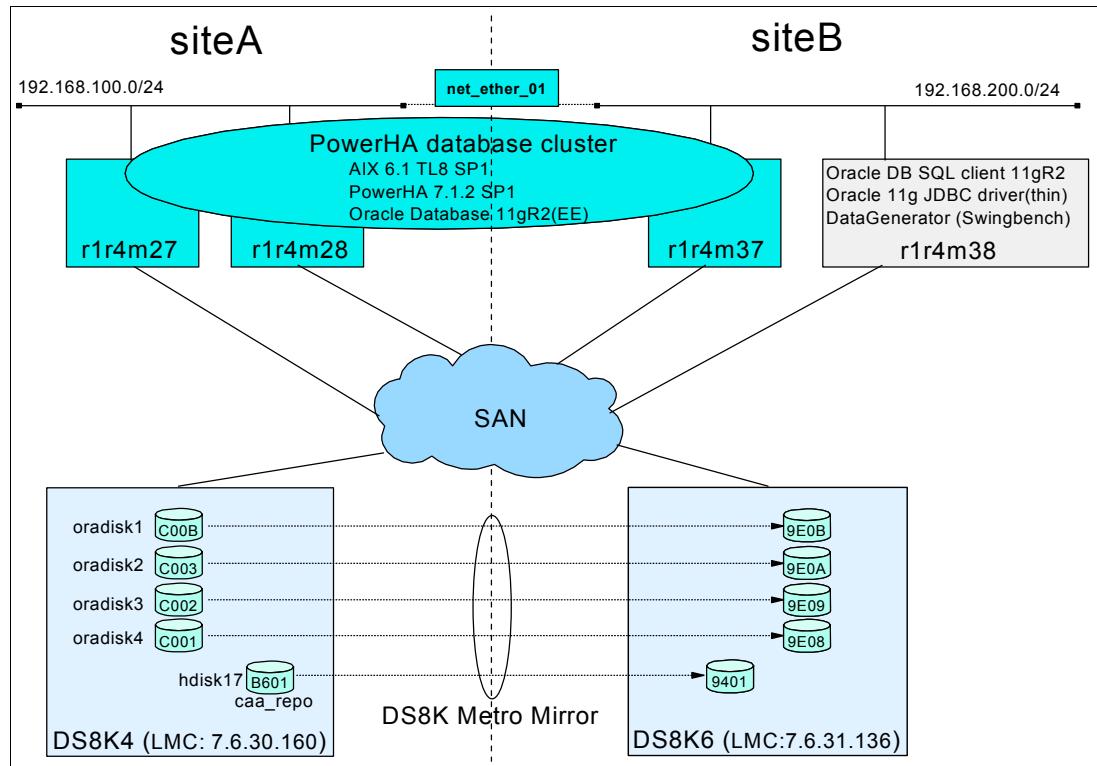


Figure 4-74 The database cluster environment

A fourth node, r1r4m38 in siteB, is used for test purposes. It has the Oracle database client installed and is used to test the connection to the database during the node failover tests. We also used the Swingbench DataGenerator application installed on this node to generate a database load during the HyperSwap tests. For more details about the Swingbench DataGenerator software, refer to:

<http://www.dominicgiles.com/index.html>

The storage environment consists of one DS8800 system in each site named DS8K4 (siteA) and DS8K6 (siteB). On each storage, we defined four volumes (LUNs) allocated for the Oracle database file systems. These volumes use the same Logical Storage Subsystem (LSS). The LSS is the first two octets from the DS8000 volume ID. See Table 4-3 for the details on volume allocation in the DS8800 in our environment.

Table 4-3 DS8800 volumes for Oracle database storage and CAA repository

LUN ID DS8K4	LUN ID DS8K6	LUN capacity	AIX device (hdisk#)	AIX volume group	AIX file systems (jfs2)	Role
C001	9E08	50 GB	oradisk4	oraarchvg	/u04	Archive logs, flash recovery area
C002	9E09	10 GB	oradisk3	oraredovg	/u03	Redo logs
C003	9E0A	20 GB	oradisk2	oradatavg	/u02	Data files
C00B	9E0B	20 GB	oradisk1	oraclevg	/u01	Oracle code

LUN ID DS8K4	LUN ID DS8K6	LUN capacity	AIX device (hdisk#)	AIX volume group	AIX file systems (jfs2)	Role
B601	9401	2 GB	hdisk17	caavg_private	N/A	CAA repository

An additional pair of volumes is used for the Cluster Aware AIX (CAA) repository, which is also configured with the HyperSwap protection. The DS8800 volumes for the CAA repository use a different LSS ID than the Oracle database disks. We used a PowerHA stretched cluster type which contains a single repository disk shared on all cluster nodes. The rootvg disks are allocated on internal storage on each node—they are not part of the HyperSwap set of disks. The entire Oracle database environment, including the Oracle binaries, the database files, and the Oracle user home directory are placed on the DS8800 storage, and on HyperSwap-capable disks shared to all the cluster nodes.

In order to facilitate AIX disk management and to distinguish the AIX database disk devices from the rest of the disks allocated to each node, we changed the names of these devices using the following AIX command:

```
rendev -l hdisk# -n oradisk#
```

The cluster configuration has a TCP/IP network (net_ether_01) used for external client communication with the database services, such as the SQL client and the database load software on node r1r4m38. This cluster network has two different IP subnets, one specific to each site, which are routed between the sites. In our case, we also used two distinct subnets for the base and the service IP labels. The interface IP addresses are summarized in Table 4-4.

Table 4-4 Node network interfaces

Site	Node	IP label	IP Address/mask	PowerHA role
siteA	r1r4m27	r1r4m27.austin.ibm.com	9.3.18.233/24	base address
	r1r4m28	r1r4m28.austin.ibm.com	9.3.18.183/24	base address
	r1r4m27/r1r4m28	orasrva.austin.ibm.com	192.168.100.100/24	service address
siteB	r1r4m37	r1r4m37.austin.ibm.com	9.3.207.248/24	base address
	r1r4m37	orasrvb.austin.ibm.com	192.168.200.100/24	service address

In a real case environment, additional IP networks such as XD_ip or Ether type networks and non-IP networks (SAN-based heartbeat within the sites) should be defined to increase the cluster heartbeat path redundancy.

4.15.2 HyperSwap disk configuration

Both the database volumes on the DS8800 storages and the CAA repository disks in our environment were enabled for HyperSwap. The AIX disk configuration must be performed before including them in the PowerHA cluster.

The AIX MPIO driver is used for the multipath access to the DS8800 storage volumes. Each node uses two FC ports and four paths to access each individual LUN on the storage. The disk path configuration is illustrated in Figure 4-75 on page 120.

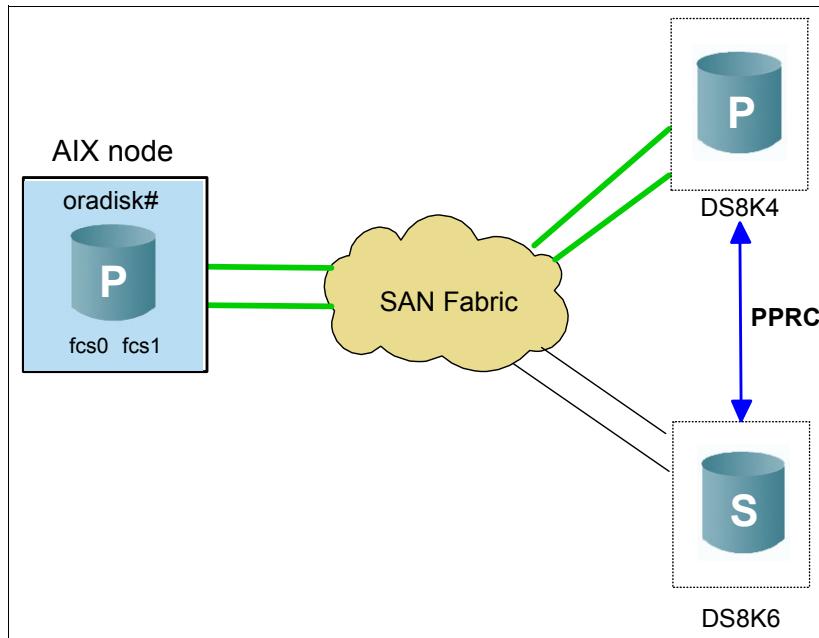


Figure 4-75 Disk path configuration

During normal operation, the active paths for the disk I/O traffic are the paths to the primary storage volumes (marked with P in the figure) on the DS8K4 storage in siteA, while the paths to the secondary volumes (marked with S in the figure) on the storage DS8K6 in siteB are inactive. The AIX HyperSwap disk configuration can be verified using the AIX **1spprc** command shown in Example 4-4.

Example 4-4 *1spprc* output for a HyperSwap configured disk

(0)	root @ r1r4m27:(REG) /				
>	1spprc -p oradisk1				
path	WWNN	LSS	VOL	path	
group id				group status	
=====					
0(s)	500507630affc16b	0xc0	0x0b	PRIMARY	
1	5005076308ffc6d4	0x9e	0x0b	SECONDARY	
path	path	path	parent	connection	
group id	id	status			
=====					
0	0	Enabled	fscsi0	500507630a08016b,40c0400b00000000	
0	1	Enabled	fscsi0	500507630a08416b,40c0400b00000000	
0	2	Enabled	fscsi1	500507630a08016b,40c0400b00000000	
0	3	Enabled	fscsi1	500507630a08416b,40c0400b00000000	
1	4	Enabled	fscsi0	50050763081b06d4,409e400b00000000	
1	5	Enabled	fscsi0	50050763081b46d4,409e400b00000000	
1	6	Enabled	fscsi1	50050763081b06d4,409e400b00000000	
1	7	Enabled	fscsi1	50050763081b46d4,409e400b00000000	

The output of the **1spprc -p** command shows all the paths of the AIX HyperSwap disk to the corresponding LUNs on each storage. The storage identification is based on the World Wide Node Name (WWNN) of the storage system, which can be obtained using the **dscli**

administrative command line interface of the DS8000 storage and the **dscli lssi** command. The DS8800s have the following WWNNs:

DS8K4: 500507630affc16b
DS8K6: 5005076308ffc6d4

For performing the initial disk configuration of the database volumes and the HyperSwap enablement, refer to 4.10, “Initial disk configuration” on page 72 and 4.11, “Preparation of a HyperSwap-capable disk” on page 73.

4.15.3 Verifying AIX and the PowerHA cluster environment for the database

In the following sections, we point out specific AIX and cluster related configurations prepared in our environment. Check the hardware and software system requirements and also the suggested AIX parameters in the Oracle 11g Installation Guide for AIX before installing the database software, at:

http://docs.oracle.com/cd/E11882_01/install.112/e24332/pre_install.htm#BABFDGHJ

Oracle user and group definitions

The users and groups defined for the Oracle database are summarized in Table 4-5. Observe that the home directory of the Oracle database user is placed on the shared disk oradisk1 in the file system /u01 along with the Oracle database code.

Table 4-5 Users and groups for the Oracle components implementation

Oracle component	User (ID)	Primary group (ID)	Other groups	Home directory
Database	oracle(500)	oinstall(500)	dba(501)	/u01/home/oracle

Tip: User and group definitions must be the same on all cluster nodes. You can use PowerHA SystemMirror C-SPOC operations to define the users and groups across all cluster nodes. This facilitates performing the operations from a single node, instead of doing the same operation on each cluster node. The smit fastpath is: **smit c1_usergroup**.

The actual configuration of the AIX user is shown in the output of the **lsuser** and **lsgroup** commands. See Example 4-5.

Example 4-5 Define the Oracle user and groups

```
(0) root @ r1r4m27:(REG) /
> lsuser oracle
oracle id=500 pgrp=oinstall groups=oinstall,dba home=/u01/home/oracle
shell=/usr/bin/ksh login=true su=true rlogin=true daemon=true admin=false
sugroups=ALL admgroups= tpath=nosak ttys=ALL expires=0 auth1=SYSTEM auth2=NONE
umask=22 registry=files SYSTEM=compat logintimes= loginretries=0 pwdwarntime=0
account_locked=false minage=0 maxage=0 maxexpired=-1 minalpha=0 minloweralpha=0
minupperalpha=0 minother=0 mindigit=0 minspecialchar=0 mindiff=0 maxrepeats=8
minlen=0 histexpire=0 histsize=0 pwdchecks= dictionlist=
capabilities=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE default_roles= fsize=-1 cpu=-1
data=-1 stack=-1 core=2097151 rss=-1 nofiles=200000 roles=
```

```
(0) root @ r1r4m27:(REG) /
> lsgroup oinstall
```

```
oinstall id=500 admin=false users=oracle adms=root registry=files  
(0) root @ r1r4m27:(REG) /  
> lsgroup dba  
dba id=501 admin=false users=oracle adms=root registry=files
```

Important: According to the Oracle 11g documentation for AIX, for the operating system to use 16 MB pages or pinned memory when allocating shared memory, the Oracle user ID must have set the following capabilities: CAP_BYPASS_RAC_VMM and CAP_PROPAGATE. Use the following command to change the user capabilities:

```
chuser capabilities=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE <user id>
```

Database volume groups and file systems

The volume group definitions need to be imported on all cluster nodes prior to defining the cluster resources using the Oracle Smart Assist. Due to AIX devices being renamed on each cluster node, the volume group association with the AIX physical volumes is the same on all nodes. See the **lspv** command output in Example 4-6.

Example 4-6 lspv output on a cluster node

```
> lspv | grep oradisk  
oradisk4      00c931c44a365fd5          oraarchvg  
oradisk3      00c931c44a35ac96          oraredovg  
oradisk2      00c931c44a352a9c          oradatavg  
oradisk1      00c931c44a34736b          oraclevg
```

The file systems are defined for each volume group. The logical volume and associated file systems defined are listed in Example 4-7.

Example 4-7 Logical volumes defined in the database volume groups

```
0) root @ r1r4m27:(REG) /  
> lsvg -o | grep ora | lsvg -il  
oraarchvg:  
LV NAME      TYPE    LPs    PPs    PVs   LV STATE    MOUNT POINT  
oraarchloglv jfs2log 1       1       1     closed/syncd N/A  
u041v        jfs2    638    638    1     closed/syncd /u04  
oraredovg:  
LV NAME      TYPE    LPs    PPs    PVs   LV STATE    MOUNT POINT  
oraredologlv jfs2log 1       1       1     closed/syncd N/A  
u031v        jfs2    352    352    1     closed/syncd /u03  
oradatavg:  
LV NAME      TYPE    LPs    PPs    PVs   LV STATE    MOUNT POINT  
u021v        jfs2    620    620    1     closed/syncd /u02  
oradataloglv jfs2log 1       1       1     closed/syncd N/A  
oraclevg:  
LV NAME      TYPE    LPs    PPs    PVs   LV STATE    MOUNT POINT  
oracleloglv  jfs2log 1       1       1     closed/syncd N/A  
u011v        jfs2    544    544    1     closed/syncd /u01
```

For the Smart Assist software to properly discover the instances, databases, and listeners, particularly when multiple instances or versions are used, the directory layout must comply with the Oracle Flexible Architecture (OFA) standard. See more information about OFA at:

http://docs.oracle.com/cd/E11882_01/install.112/e24332/appendix_ofa.htm

In order to comply with the OFA standard, we created the following directory structures inside the /u01 to /u04 file systems, logged in on the node as *oracle* user:

- ▶ Oracle code:

```
mkdir -p /u01/app/oracle  
chmod 755/u01/app/oracle
```

- ▶ Oracle data files

```
mkdir -p /u02/app/oracle/oradata  
chmod 775 /u02/app/oracle/oradata
```

- ▶ Oracle redo log files

```
mkdir -p /u03/app/oracle/oradata/  
chmod 775 /u03/app/oracle/oradata/
```

- ▶ Oracle archive logs

```
mkdir -p /u04/app/oracle/admin  
chmod 775 /u04/app/oracle/admin
```

- ▶ Oracle fast recovery area

```
mkdir -p /u04/app/oracle/flash_recovery_area  
chmod 775 /u04/app/oracle/flash_recovery_area
```

PowerHA cluster topology

Due to disk sharing requirements by the HyperSwap environment, a cluster repository disk can also be shared and can be included in the PowerHA HyperSwap configuration. The stretched cluster is usually the most appropriate configuration in a HyperSwap environment. However, the linked cluster type is also supported. One of the benefits of the stretched clusters is the heartbeat using the cluster repository disk, which in the case of nodes placed on different sites creates an additional line of defense against cluster partitioning besides the regular TCP/IP and SAN heartbeat networks defined.

The cluster topology used for the Oracle database is shown in the **cktopinfo** command output in Example 4-8.

Example 4-8 Cluster topology configuration

```
(0) root @ r1r4m27:(REG) /  
> cktopinfo  
Cluster Name: r1r4m27_cluster  
Cluster Connection Authentication Mode: Standard  
Cluster Message Authentication Mode: None  
Cluster Message Encryption: None  
Use Persistent Labels for Communication: No  
Repository Disk: hdisk17  
Cluster IP Address: 228.3.18.233  
There are 3 node(s) and 1 network(s) defined
```

```
NODE r1r4m27:  
    Network net_ether_01  
        r1r4m27 9.3.18.233
```

```
NODE r1r4m28:  
    Network net_ether_01  
        r1r4m28 9.3.18.183
```

```
NODE r1r4m37:  
    Network net_ether_01  
        r1r4m37 9.3.207.248
```

No resource groups defined

There are no resource groups defined at this time. They are created at the time of running the Oracle Smart Assist.

4.15.4 The Oracle database environment

In this section, we describe how we verified and adjusted the database environment for running the Smart Assist software. We had the database software already installed and a database named ITSOTEST1 had been created on the prepared AIX jfs2 file systems (/u01 to /u04) in directory structures already created in the previous paragraph. An Oracle listener had also been defined with the default name LISTENER and an associated TCP/IP port 1521.

The following steps were performed:

1. Activate the volume groups and mount the Oracle database file systems on a cluster node.
2. Install the Oracle 11gR2 code.
3. Configure a database on the jfs2 file systems using the Oracle Database Configuration Assistant (DBCA) tool.
4. Configure a database listener using the Network Configuration Assistant (NETCA).

For more details about Oracle database installation and configuration, consult the Oracle documentation at:

http://www.oracle.com/pls/db112/portal.portal_db

We performed the following steps to verify and update the database configuration for proper Oracle Smart Assist software discovery of the database instance and listener. Unless specified, the following steps are performed by the Oracle user ID (oracle):

1. Update the oracle user profile with the Oracle environment variable (see Example 4-9).

Example 4-9 Oracle user profile

```
#Oracle environment  
export ORACLE_OWNER=oracle  
export ORACLE_HOME=/u01/app/oracle/product/11.2.0/db112_1  
export ORACLE_SID=ITSOTEST1  
export PATH=$PATH:$ORACLE_HOME/bin
```

2. Check the Oracle listener configuration file located in the \$ORACLE_HOME/network/admin directory. The SID name of the database must be present in the SID_LIST stanza. See the listener.ora file we configured in our environment in Example 4-10.

Example 4-10 Oracle database environment listener configuration

```
$ pwd  
/u01/app/oracle/product/11.2.0/db112_1/network/admin
```

```

$ cat listener.ora
# listener.ora Network Configuration File:
/u01/app/oracle/product/11.2.0/db112_1/network/admin/listener.ora
# Generated by Oracle configuration tools.

LISTENER =
  (DESCRIPTION_LIST =
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = r1r4m27.austin.ibm.com)(PORT = 1521))
      (ADDRESS = (PROTOCOL = IPC)(KEY = EXTPROC1521))
    )
  )

SID_LIST_LISTENER =
  (SID_LIST =
    (SID_DESC =
      (SID_NAME = PLSExtProc)
      (ORACLE_HOME = /u01/app/oracle/product/11.2.0/db112_1)
      (PROGRAM = extproc)
    )
    (SID_DESC =
      (GLOBAL_DBNAME = ITSOTEST1)
      (ORACLE_HOME = /u01/app/oracle/product/11.2.0/db112_1)
      (SID_NAME = ITSOTEST1)
    )
  )

ADR_BASE_LISTENER = /u01/app/oracle

```

Attention: The content of `listener.ora` is important for the Smart Assist to properly detect the listener configuration. The cluster Smart Assist software looks for the `SID_LIST` stanza to be present in the Oracle listener configuration file.

- Verify that the Oracle database startup file (`spfile`) is present in the `$ORACLE_HOME/dbs` directory. Example 4-11 shows the `spfile` for the Oracle instance `ITSOTEST1` in our environment.

Example 4-11 spfile for the instance ITSOTEST1 database

```

$ cd $ORACLE_HOME/dbs
$ pwd
/u01/app/oracle/product/11.2.0/db112_1/dbs
$ ls -l
total 56
-rw-rw---- 1 oracle oinstall 1544 Nov 27 21:05 hc_DBUA0.dat
-rw-rw---- 1 oracle oinstall 1544 Nov 28 23:17 hc_ITSOTEST1.dat
-rw-r--r-- 1 oracle oinstall 2851 May 15 2009 init.ora
-rw-r----- 1 oracle oinstall 24 Nov 27 21:50 1kITSOTEST
-rw-r----- 1 oracle oinstall 24 Nov 27 21:52 1kITSOTEST1
-rw-r----- 1 oracle oinstall 1536 Nov 27 21:53 orapwITSOTEST1
-rw-r----- 1 oracle oinstall 3584 Nov 28 22:57 spfileITSOTEST1.ora

```

4. Check the Oracle database and listener status. The database should be running after completing the DBCA configuration. If the listener file was updated, you need to restart the listener service. You can use the following commands to do that:

```
lsnrctl stop [ <listener_name> ]
lsnrctl start [ <listener_name> ]
```

Example 4-12 provides an example of verifying the database instance status using the sqplus Oracle client and the **lsnrctl** command for checking the listener status.

Example 4-12 Checking the database and listener status

```
$ sqlplus / as sysdba

SQL*Plus: Release 11.2.0.1.0 Production on Wed Nov 28 15:57:56 2012

Copyright (c) 1982, 2009, Oracle. All rights reserved.

Connected to:
Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options

SQL> select * from v$instance;

INSTANCE_NUMBER INSTANCE_NAME
-----
HOST_NAME
-----
VERSION        STARTUP_T STATUS      PAR    THREAD# ARCHIVE LOG_SWITCH_WAIT
-----
LOGINS        SHU DATABASE_STATUS INSTANCE_ROLE    ACTIVE_ST BLO
-----
           1 ITSOTEST1
r1r4m27.austin.ibm.com
11.2.0.1.0      27-NOV-12 OPEN        NO      1 STARTED
ALLOWED        NO  ACTIVE          PRIMARY_INSTANCE NORMAL    NO

SQL> quit
Disconnected from Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 -
64bit Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options

$ lsnrctl status

LSNRCTL for IBM/AIX RISC System/6000: Version 11.2.0.1.0 - Production on
29-NOV-2012 08:53:01

Copyright (c) 1991, 2009, Oracle. All rights reserved.

Connecting to
(DESCRIPTION=(ADDRESS=(PROTOCOL=TCP)(HOST=r1r4m27.austin.ibm.com)(PORT=1521)))
STATUS of the LISTENER
-----
```

```

Alias           LISTENER
Version        TNSLSNR for IBM/AIX RISC System/6000: Version 11.2.0.1.0
- Production
Start Date     29-NOV-2012 00:40:39
Uptime         0 days 8 hr. 12 min. 22 sec
Trace Level    off
Security       ON: Local OS Authentication
SNMP          ON

Listener Parameter File
/u01/app/oracle/product/11.2.0/db112_1/network/admin/listener.ora
Listener Log File
/u01/app/oracle/diag/tnslsnr/r1r4m27/listener/alert/log.xml
Listening Endpoints Summary...
  (DESCRIPTION=(ADDRESS=(PROTOCOL=tcp)(HOST=r1r4m27.austin.ibm.com)(PORT=1521)))
  (DESCRIPTION=(ADDRESS=(PROTOCOL=ipc)(KEY=EXTPROC1521)))

Services Summary...
Service "ITSOTEST1" has 2 instance(s).
  Instance "ITSOTEST1", status UNKNOWN, has 1 handler(s) for this service...
  Instance "ITSOTEST1", status READY, has 1 handler(s) for this service...
Service "ITSOTEST1XDB" has 1 instance(s).
  Instance "ITSOTEST1", status READY, has 1 handler(s) for this service...
Service "PLSExtProc" has 1 instance(s).
  Instance "PLSExtProc", status UNKNOWN, has 1 handler(s) for this service...
The command completed successfully

```

5. As root user, stop the cluster services on the nodes if they are already running. Verify that the node is stopped in the cluster using the following command:

```
#lssrc -ls clstrmgrES | grep state
```

Current state: ST_INIT

4.15.5 Configure the database in PowerHA using the Oracle Smart Assist

In this section we explain, step-by-step, the configuration of the cluster resources and resource groups for the existing Oracle database environment, and also the additional cluster customizations performed to configure the PowerHA HyperSwap.

Configuring the database using the Oracle Smart Assist wizard

For configuring the database resources in the PowerHA SystemMirror cluster, we performed the following steps:

1. Set the following environment variables as root user for Smart Assist discovery:


```
#export ORACLE_HOME=/u01/app/oracle/product/11.2.0/db112_1
#export ORACLE_USER=oracle
```
2. Ran PowerHA SystemMirror Smart Assist for Oracle using the smit menus: **smitty c1sa** → **Add an Application to the PowerHA SystemMirror Configuration**, then selected **Oracle Database Smart Assist**. Refer to Figure 4-76 on page 128.

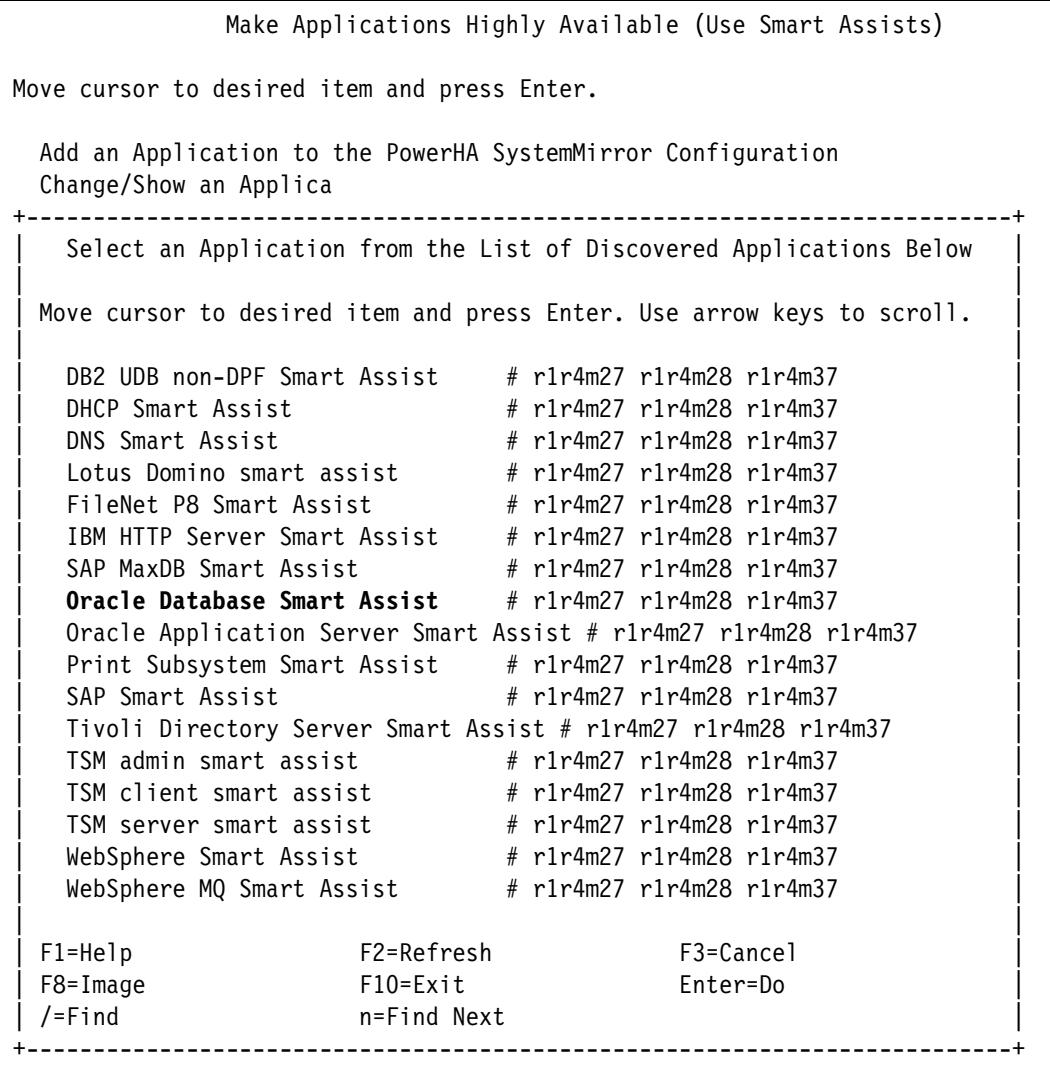


Figure 4-76 Select the Oracle Database smart assist

- In the next menu, select **Automatic Discovery And Configuration** as shown in Figure 4-77.

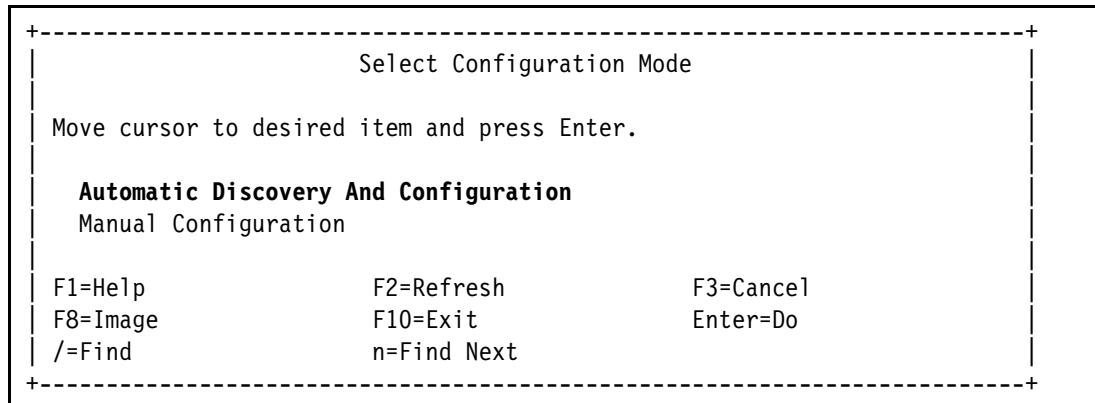


Figure 4-77 Select the detection and configuration methods

4. In the next panel, the Oracle RDBMS option is displayed and the node currently having the database started. See Figure 4-78.

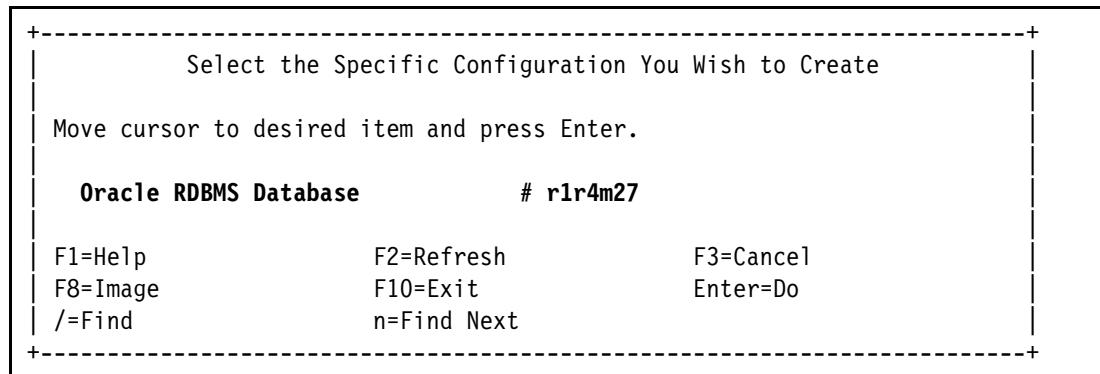


Figure 4-78 Select the specific configuration you wish to create

5. The next panel displays the Oracle version detected and the existing databases (DB SIDs), as shown in Figure 4-79.

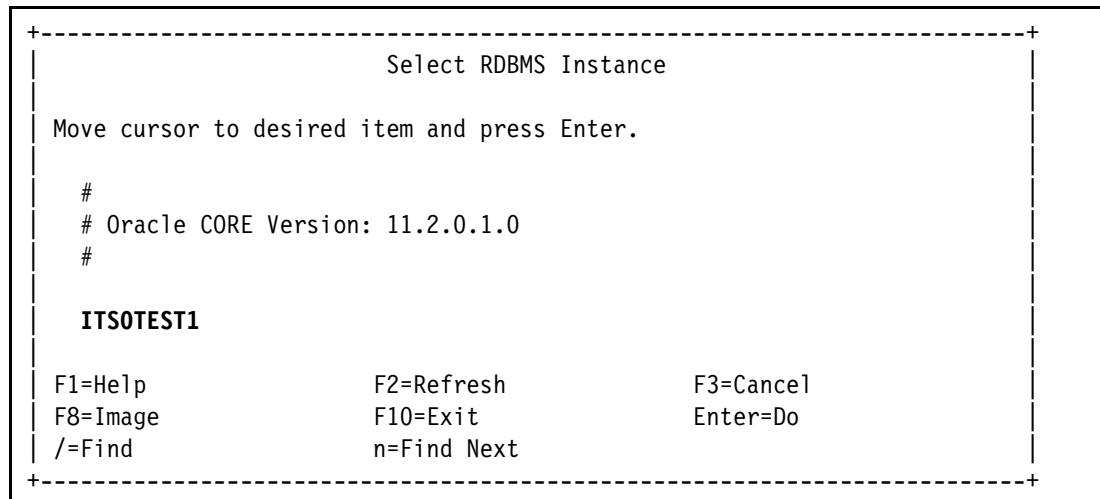


Figure 4-79 Choose the Oracle Database you want to configure in PowerHA

6. In this step we added the Oracle RDBMS instance to the cluster configuration. Customize the cluster resource configuration according to your environment. See our configuration in Figure 4-80 on page 130.

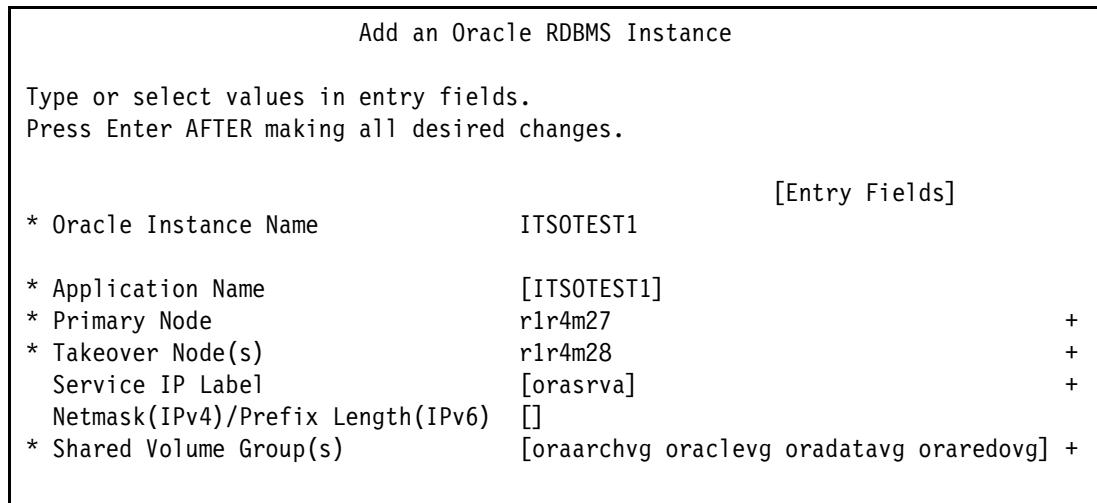


Figure 4-80 Add the Oracle Instance in the PowerHA configuration

NOTE: Observe that the cluster resource configuration generated at this time contains only the nodes and the service IP address for the primary site, siteA. The second site will be added later during the update operation of the resource group.

- After starting the Smart Assist configuration process, the cluster resources and resource groups are configured. See the output of the Smart Assist configuration process in Example 4-13.

Example 4-13 Adding Oracle instance to the cluster configuration

```

instance= listener=ITSOTEST1 LISTENER r1r4m27.austin.ibm.com
instance= listener=ITSOTEST1 LISTENER orasrva.austin.ibm.com
Adding RDBMS database: ITSOTEST1 to the PowerHA SystemMirror configuration.

      Adding Oracle RDBMS Database Listener
      [ IP Label: ITSOTEST1, Listener Name: LISTENER]

      Adding Oracle RDBMS Database Listener
      [ IP Label: ITSOTEST1, Listener Name: LISTENER]

-----
Configuring service IP label: orasrva
Service IP label configuration complete.

-----
Creating PowerHA SystemMirror resource group: ITSOTEST1_CFC_RG_RDBMS to support
RDBMS Database: ITSOTEST1

      Creating PowerHA SystemMirror application server: ITSOTEST1_CFC_AP_RDBMS

      Creating PowerHA SystemMirror application monitor: ITSOTEST1_CFC_AM_RDBMS

      Oracle RDBMS Database version: 11.2.0.1.0

.....

```

8. Run the cluster verification and synchronization process to propagate the changes on all cluster nodes: **smitty hacmp** → **Cluster Applications and Resources** → **Verify and Synchronize Cluster Configuration**.
9. The cluster configuration results can be displayed using the **cltopinfo** command. See Example 4-14.

Example 4-14 cltopinfo output

```
> cltopinfo
Cluster Name: r1r4m27_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: hdisk21
Cluster IP Address: 228.3.18.233
There are 4 node(s) and 1 network(s) defined

NODE r1r4m27:
    Network net_ether_01
        orasrva 192.168.100.100
        r1r4m27 9.3.18.233

NODE r1r4m28:
    Network net_ether_01
        orasrva 192.168.100.100
        r1r4m28 9.3.18.183

NODE r1r4m37:
    Network net_ether_01
        orasrva 192.168.100.100
        r1r4m37 9.3.207.248

Resource Group ITSOTEST1_CFC_RG_RDBMS
    Startup Policy Online On Home Node Only
    Failover Policy Failover To Next Priority Node In The List
    Fallback Policy Never Fallback
    Participating Nodes r1r4m27 r1r4m28
    Service IP Label orasrva
```

10. Stop the Oracle database and the listener services.

Customize the Oracle network configuration

In this section, we updated the following Oracle network configuration files:

- ▶ The **listener.ora** file, by adding the cluster service IP address.
- ▶ The **tnsnames.ora** file, used for the sql client connection to the database server. The file location is on a shared drive: **\$ORACLE_HOME/network/admin/tnsnames.ora**. We also configured this file on node **r1r4m38** used for testing the connection to the database server.

We had to specify the cluster service IP address in the listener configuration file, **listener.ora**. Our extended distance cluster used two IP segments, one specific to each site, and two site-specific service IP addresses: **orasrva** (in siteA) and **orasrvb** (in siteB).

???In order to avoid a listener start error due to a nonexistent IP address in case of activating the resource group in either site, we defined two versions of this file each containing the service IP address used in that site. The location of the `listener.ora` file is on the cluster shared volumes which have the same content for all nodes when the volume groups and file systems get activated with the cluster resource group. A possible way to circumvent this situation is to create a symbolic link of the `listener.ora` to a local directory on each cluster node. For example:

```
$ORACLE_HOME/network/admin/listener.ora ---sym link → /<node  
localdir>/listner.ora
```

We created the symbolic link using the following command:

```
ln -s /ha_oracle/listener.ora $ORACLE_HOME/network/admin/listener.ora
```

Example 4-15 shows the `listener.ora` file for the primary site siteA. Observe in the output of the example, that we replaced the existing hostname address with the service IP address for the site siteA. A similar file is created for the secondary site siteB, using the `orasrvb` service address instead of `orasrva`.

Example 4-15 listener.ora file on the nodes in siteA

```
$ cat listener.ora  
# listener.ora Network Configuration File:  
/u01/app/oracle/product/11.2.0/db112_1/network/admin/listener.ora  
# Generated by Oracle configuration tools.  
  
LISTENER =  
(DESCRIPTION_LIST =  
(DESCRIPTION =  
(ADDRESS = (PROTOCOL = TCP)(HOST = orasrva.austin.ibm.com)(PORT = 1521))  
(ADDRESS = (PROTOCOL = IPC)(KEY = EXTPROC1521))  
)  
)  
  
SID_LIST_LISTENER =  
(SID_LIST =  
(SID_DESC =  
(SID_NAME = PLSExtProc)  
(ORACLE_HOME = /u01/app/oracle/product/11.2.0/db112_1)  
(PROGRAM = extproc)  
)  
(SID_DESC =  
(GLOBAL_DBNAME = ITSOTEST1)  
(ORACLE_HOME = /u01/app/oracle/product/11.2.0/db112_1)  
(SID_NAME = ITSOTEST1)  
)  
)  
ADR_BASE_LISTENER = /u01/app/oracle
```

The `tnsnames` file is used by the Oracle sql clients and can contain multiple IP addresses associated with the cluster nodes. We used a basic failover mechanism with the two site-specific IP addresses. See the `tnsnames.ora` file in our environment in Example 4-16 on page 133. This file is configured on the shared file system `/u01` and on the Oracle SQL client on node `r1r4m38`.

Example 4-16 The tnsnames.ora file

```
ITSOTEST1 =
(DESCRIPTION =
  (ADDRESS = (PROTOCOL = TCP)(HOST = orasrva)(PORT = 1521))
  (ADDRESS = (PROTOCOL = TCP)(HOST = orasrvb)(PORT = 1521))
  (FAILOVER = yes)
  (CONNECT_DATA =
    (SERVER = DEDICATED)
    (SERVICE_NAME = ITSOTEST1)
  )
)
```

Performing the site-specific configuration

Before starting the configuration steps, we unmounted the Oracle database file systems and varied off the volume groups to prepare the shared storage for activation by the cluster services—after the additional configuration operations are performed and the cluster configuration is synchronized.

The following additional steps were performed to extend the current resource group configuration for inter-site failover:

1. We added the third node r1r4m37 to the database resource group in the cluster and configured the inter-site policy. Access the resource group operation using the following smit path:

smitty hacmp → Cluster Applications and Resources → Resource Groups → Change/Show Nodes and Policies for a Resource Group, then we selected the database resource group for the modification action, ITSOTEST1_CFC_RG_RDBMS.

In the next panel, we added the node r1r4m37 from the secondary site, siteB, and let the default inter-site policy set to Prefer Primary Site; see Figure 4-81.

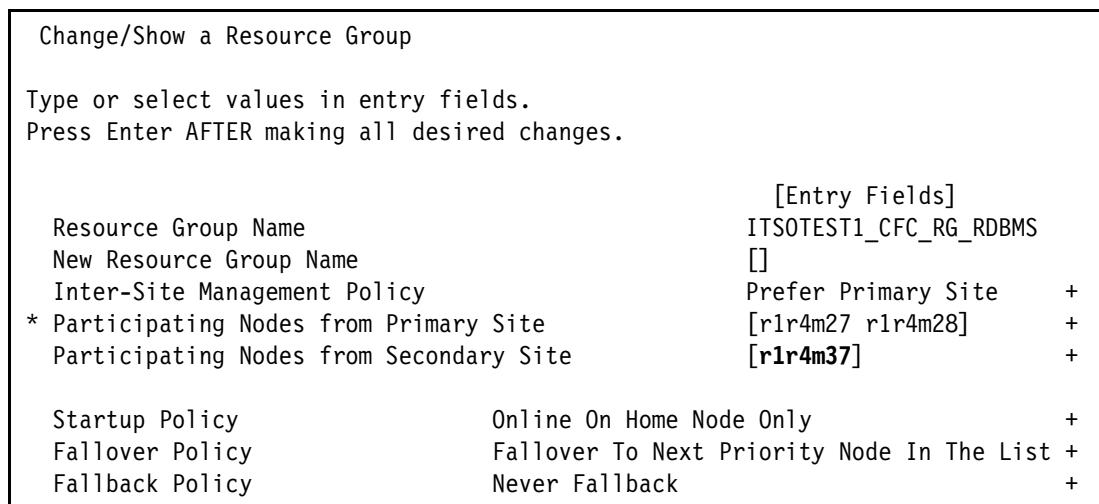


Figure 4-81 Adding the third node to the database resource group

2. In the next step, we configured the site-specific service IP addresses associated with the database resource group. In order to change an already defined service IP resource to a

site-specific configuration, we had to delete it and redefine it back. See Figure 4-82 for how to create a site-specific IP address resource in the cluster configuration:

smitty hacmp → **Cluster Applications and Resources** → **Resources** → **Configure Service IP Labels/Addresses** → **Add a Service IP Label/Address**, then we selected the network for the IP address being configured (net_ether_01 in our case).

Add a Service IP Label/Address configurable on Multiple Nodes (extended)	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
IP Label/Address	[Entry Fields] orasrvb.austin.ibm.com +
Netmask(IPv4)/Prefix Length(IPv6)	[]
* Network Name	net_ether_01
Associated Site	siteB +

Figure 4-82 Adding a service site-specific IP address

3. We changed the database resource group attributes to include the new IP addresses; see Figure 4-83.

smitty hacmp → **Cluster Applications and Resources** → **Resource Groups** → **Change/Show Resources and Attributes for a Resource Group**

Change/Show All Resources and 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	ITSOTEST1_CFC_RG_RDBMS
Inter-site Management Policy	Prefer Primary Site
Participating Nodes from Primary Site	r1r4m27 r1r4m28
Participating Nodes from Secondary Site	r1r4m37
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority Node In The List
Fallback Policy	Never Fallback
Service IP Labels/Addresses	[orasrva orasrvb] +
Application Controller Name	[ITSOTEST1_CFC_AP_RDBMS] +
Volume Groups	[oraarchvg oraclevg oradatavg oraredovg] +
Use forced varyon of volume groups, if necessary	false +
Automatically Import Volume Groups	false +
.....	

Figure 4-83 Adding the site-specific IP address to the resource group

4. We verified and synchronized the cluster configuration to propagate the changes on all cluster nodes.

Configure the HyperSwap resource in the cluster configuration

In order to complete our cluster configuration, we had to configure the DS8000 Metro Mirror (in-band) cluster resources by creating the repository and user mirror group, and associate the user mirror group with the database resource group.

The following steps were performed to define the PowerHA mirror groups:

1. We defined the storage systems in the PowerHA configuration and their association with the sites. See Figure 4-84. Use the following smit path:

```
smitty sysmirror → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Storage Systems → Add a Storage System
```

Add a Storage System	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
[Entry Fields]	
* Storage System Name	[DS8K4]
* Site Association	siteA +
* Vendor Specific Identifier	IBM.2107-00000TL771 +
* WWNN	500507630AFFC16B +

Figure 4-84 Adding a storage system to the cluster configuration

Note: The vendor-specific information is the DS8000 storage image ID, a unique identifier of the DS8000 storage system. This attribute and the WWNN of the storage system can be chosen from the drop-down menu or by using the **dscli lssi** command.

The same operation was performed for defining the secondary storage in siteB, DS8K6.

2. We defined the cluster repository mirror group with:

```
smitty sysmirror → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Mirror Groups → Add a Mirror Group
```

We selected the mirror group type=Cluster_Repository and provided the cluster repository mirror group parameters. See our configuration in Figure 4-85 on page 136.

Add cluster Repository Mirror Group

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

[Entry Fields]	
* Mirror Group Name	[repo_mg]
* Site Name	siteA siteB +
* Non Hyperswap Disk	hdisk3:99afd053-caef-e467-cfb6-270253fccd54 +
* Hyperswap Disk	hdisk17:a1a3aa01-b341-c962-ae99-0303b64684c6+
Hyperswap	Enabled +
Consistency Group	Enabled +
Unplanned HyperSwap Timeout (in sec)	[60] #
Hyperswap Priority	High +

Figure 4-85 Adding the cluster repository mirror group

3. We defined the user mirror group for the database resource group using the following smit path:

smitty sysmirror → Cluster Applications and Resources → Resources → Configure DS8000 Metro Mirror (In-Band) Resources → Configure Mirror Groups → Add a Mirror Group

We selected the mirror group type=User and entered the user mirror group parameters. See our configuration in Figure 4-86.

Add a User Mirror Group

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

[Entry Fields]	
Mirror Group Name	oracle_mg
Volume Group(s)	oraarchvg oraclevg oradatavg oraredovg +
Raw Disk(s)	[] +
Hyperswap	Enabled +
Consistency Group	Enabled +
Unplanned HyperSwap Timeout (in sec)	[60] #
Hyperswap Priority	High +
Recovery Action	Manual +

Figure 4-86 Adding the user mirror group

Observe that we did not specify the volumes associated with the database application, but the volume groups. The cluster automatically associates the volume groups with the oradisk# devices in AIX and the storage LUN IDs.

4. We then added the mirror group in the resource group of the Oracle database. We accessed the SMIT menu for Change>Show the resource group attributes, and added the user mirror group previously defined in the existing resource group configuration. See Figure 4-87 on page 137.

Change/Show All Resources and 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	ITSOTEST1_CFC_RG_RDBMS	
Inter-site Management Policy	Prefer Primary Site	
Participating Nodes from Primary Site	r1r4m27 r1r4m28	
Participating Nodes from Secondary Site	r1r4m37	
Startup Policy	Online On Home Node	
Only		
Fallover Policy	Fallover To Next	
Priority Node In The List		
Fallback Policy	Never Fallback	
Service IP Labels/Addresses	[orasrvA orasrvB]	+
Application Controller Name	[ITSOTEST1_CFC_AP_RDBMS]	+
Volume Groups	[oraarchvg oraclevg oradatavg oraredovg]	+
.....		
Miscellaneous Data	[]	
WPAR Name	[]	+
User Defined Resources	[]	+
SVC PPRC Replicated Resources	[]	+
EMC SRDF(R) Replicated Resources	[]	+
DS8000 Global Mirror Replicated Resources	[]	+
XIV Replicated Resources		+
TRUECOPY Replicated Resources	[]	+
DS8000-Metro Mirror (In-band) Resources	oracle_mg	+

Figure 4-87 Adding the oracle_mg to the database resource group configuration

- We verified and synchronized the cluster configuration to propagate the changes to all cluster nodes.
- We started the cluster services on all nodes using **smitty clstart** and verified the resource group status using the **c1RGinfo** command. See Example 4-17.

Example 4-17 Resource group status after starting the cluster services

```
> c1RGinfo
-----
Group Name      State          Node
-----
ITSOTEST1_CFC_ ONLINE        r1r4m27@siteA
                  OFFLINE       r1r4m28@siteA
                  ONLINE SECONDARY r1r4m37@siteB
```

4.15.6 Testing the cluster environment using PowerHA HyperSwap

The followings tests were performed in our Oracle environment:

- ▶ Planned swap
- ▶ Unplanned swap
- ▶ Site failover

During the HyperSwap tests, we generated a database load using Swingbench DataGenerator program. Using this tool, we generated multiple insert operations, which in turn determined multiple write I/O operations on the HyperSwap disks in the DS8800 storage systems.

All tests assumed as a starting point that the cluster services were started on all nodes and that the Oracle database resource group (ITSOTEST1_CFC_RG_RDBMS) was active on the node r1r4m27. The disk I/O was performed to the primary storage DS8K4 from siteA. See the initial cluster resource group status in Example 4-18.

Example 4-18 Cluster resource group status

```
> c1RGinfo -p
```

```
Cluster Name: r1r4m27_cluster
```

```
Resource Group Name: ITSOTEST1_CFC_RG_RDBMS
```

```
Site settings:
```

```
The following site temporarily has the highest priority for this group:  
siteA, user-requested rg_move performed on Tue Jan 15 07:16:13 2013
```

Node	Primary State	Secondary State
r1r4m27@siteA	ONLINE	OFFLINE
r1r4m28@siteA	OFFLINE	OFFLINE
r1r4m37@siteB	OFFLINE	ONLINE SECONDARY

Planned swap

In this test case, we moved the database workload from the primary storage to the secondary storage using the smit C-SPOC menus. This case is useful for various planned storage interruptions such as storage maintenance or upgrade operations with possible outage exposure on the primary storage. The swap operation is transparent to the database application activity so it runs without any interruptions.

We started the database load using the Datagenerator tool, and also monitored the database disks using I/O stat. The tool runs multiple insert operations, which in turn generate multiple I/O write operations on the disks. Figure 4-88 on page 139 shows the panel of the DataGenerator software tool while performing the insert operations into the database.

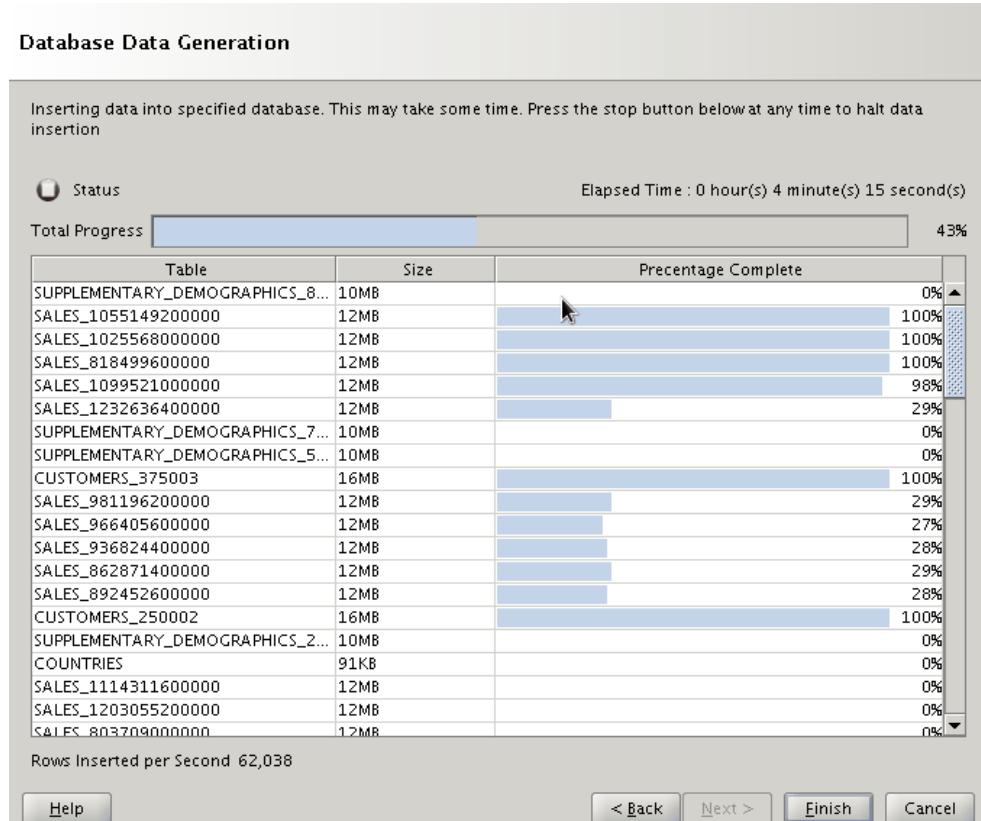


Figure 4-88 Inserting data into the database

We generated a planned swap for all mirror groups associated with the cluster volumes from siteA where the Oracle database resource group was active using the smit CSPOC menus:

smitty hacmp → System Management (C-SPOC) → Storage → Manage Mirror Groups → Manage Mirror Groups by Site. See Figure 4-89.

The screenshot shows the "Manage Mirror Groups by Site" menu. It prompts the user to type or select values in entry fields and press Enter after making changes. The menu lists several options with asterisks (*): Site Name, Include System Mirror Group(s), Include Cluster Repository MG, and Operation. To the right, there is a section titled "[Entry Fields]" containing four entries: siteA (with value yes), yes (with value yes), yes (with value yes), and Swap (with value +). The "Swap" entry is highlighted with a red box.

[Entry Fields]	
siteA	+
yes	+
yes	+
Swap	+

Figure 4-89 Swapping the access to the secondary storage in siteB

During the planned swap, we did not notice any disruption of I/O throughput and the entire swap operation performed on all mirror groups (cluster repository and Oracle database disks) was transparent. See the output of the **iostat** command run on the redo file disk during the swap test as shown in Example 4-19 on page 140.

Example 4-19 iostat output on the database disks

oradisk3	8.0	3456.0	226.0	0	3456
oradisk3	3.0	3296.0	202.0	0	3296
oradisk3	5.0	3324.0	211.0	0	3324
oradisk3	7.0	3120.0	201.0	0	3120
oradisk3	2.0	3032.0	183.0	0	3032
oradisk3	5.0	2940.0	182.0	0	2940
oradisk3	8.0	1928.0	119.0	0	1928
oradisk3	7.0	2344.0	147.0	0	2344
oradisk3	4.0	2696.0	170.0	0	2696
oradisk3	2.0	2348.0	141.0	0	2348
oradisk3	5.0	2640.0	159.0	0	2640

We changed the I/O access back to the primary storage DS8K4 using the same smit C-SPOC operations as for swapping the access to storage DS8K6 specifying *siteB* in the Manage Mirror Groups by Site menu. See Figure 4-90.

smitty cl_lvm → Manage Mirror Groups → Manage Mirror Groups by Site

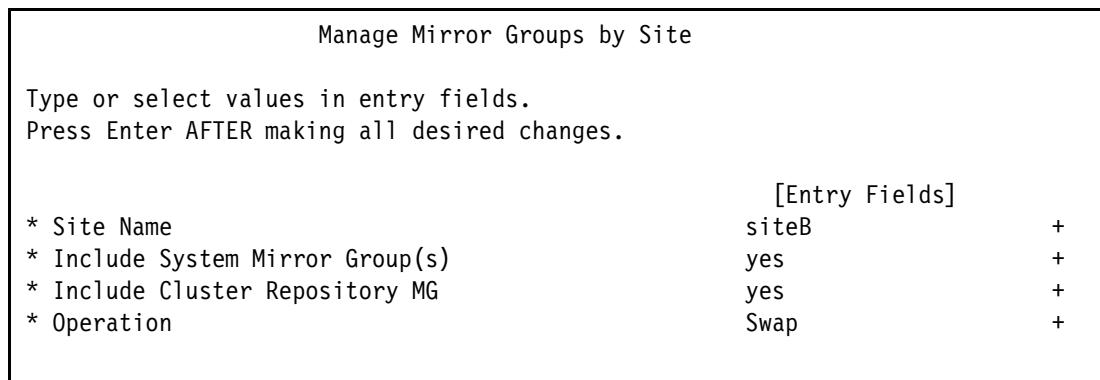


Figure 4-90 Falling back the I/O access to the primary storage

Example 4-20 shows another output of the **iostat** command during the second swap operation while running an additional I/O load on the oradisk2 (/u02) generated using the **dd** command.

Example 4-20 iostat output during the swap test

oradisk2	50.0	110304.0	577.0	0	110304
oradisk2	51.0	114444.0	380.0	0	114444
oradisk2	48.0	113668.0	370.0	0	113668
oradisk2	48.0	102648.0	525.0	0	102648
oradisk2	42.0	87912.0	216.0	0	87912
oradisk2	49.0	123432.0	379.0	16	123416
oradisk2	50.0	118280.0	868.0	0	118280
oradisk2	44.0	98420.0	301.0	0	98420
oradisk2	49.0	87024.0	197.0	0	87024
oradisk2	47.0	77056.0	164.0	0	77056
oradisk2	55.0	109956.0	241.0	0	109956
oradisk2	62.4	109240.0	453.0	0	109240
oradisk2	75.0	174984.0	397.0	0	174984

oradisk2	49.0	122752.0	273.0	0	122752
oradisk2	61.0	133456.0	324.0	96	133360

Unplanned swap

An unplanned swap operation is an interruption in accessing the SAN storage disks caused by various unplanned events such as a total storage failure or SAN access failure to the storage subsystem. In such cases, the PowerHA software automatically switches the I/O access to the secondary storage subsystem. While disk I/O operations are running, such an event causes a temporary hang of the disk traffic during failure detection and switches to the secondary storage. After the disk access is switched, the I/O traffic is resumed allowing the applications to continue their processing. This operation is normally transparent to the database system without causing an I/O error or failing the database transactions.

Before performing the test, we generated an I/O load on the database using the DataGenerator software in a manner similar to the planned swap test.

In our environment, we generated an unplanned swap by disrupting the I/O traffic between the node running the Oracle database, r1r4m27, and the primary storage in siteA, DS8K4. The operation was performed on the SAN switch by removing the zone containing the node r1r4m27 and the storage DS8K4 from the active configuration. See the actual commands issued on the Brocade FC switch in Example 4-21.

Example 4-21 Removing the zone access of node r1r4m27 to the storage DS8K4

```
HACMP_STK6-1:admin> cfgRemove hacmp_stk6_cfg,r1r4m27_ds8k4
HACMP_STK6-1:admin> cfgEnable hacmp_stk6_cfg
You are about to enable a new zoning configuration.
This action will replace the old zoning configuration with the
current configuration selected. If the update includes changes
to one or more traffic isolation zones, the update may result in
localized disruption to traffic on ports associated with
the traffic isolation zone changes
Do you want to enable 'hacmp_stk6_cfg' configuration (yes, y, no, n): [no] y
zone config "hacmp_stk6_cfg" is in effect
Updating flash ...
```

While the access to the primary storage failed, the insert operations were temporarily suspended. You can see the panel of the DataGenerator software while the I/O traffic is suspended in Figure 4-91 on page 142.

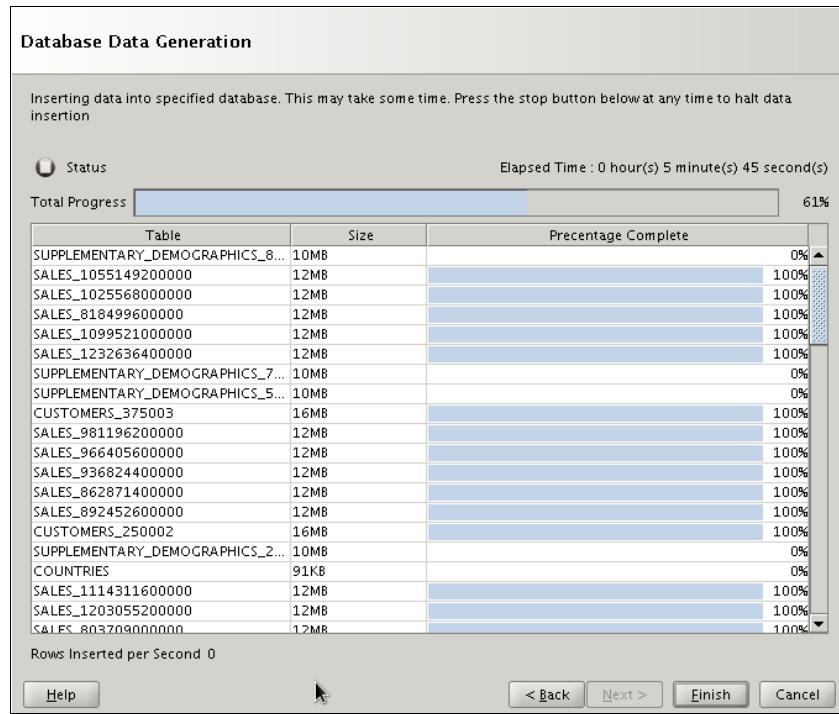


Figure 4-91 Database load process during the swap operation

After the access was swapped to the secondary storage, the I/O traffic resumed and the database load process continued. You can see a panel output of the DataGenerator load process in Figure 4-92.

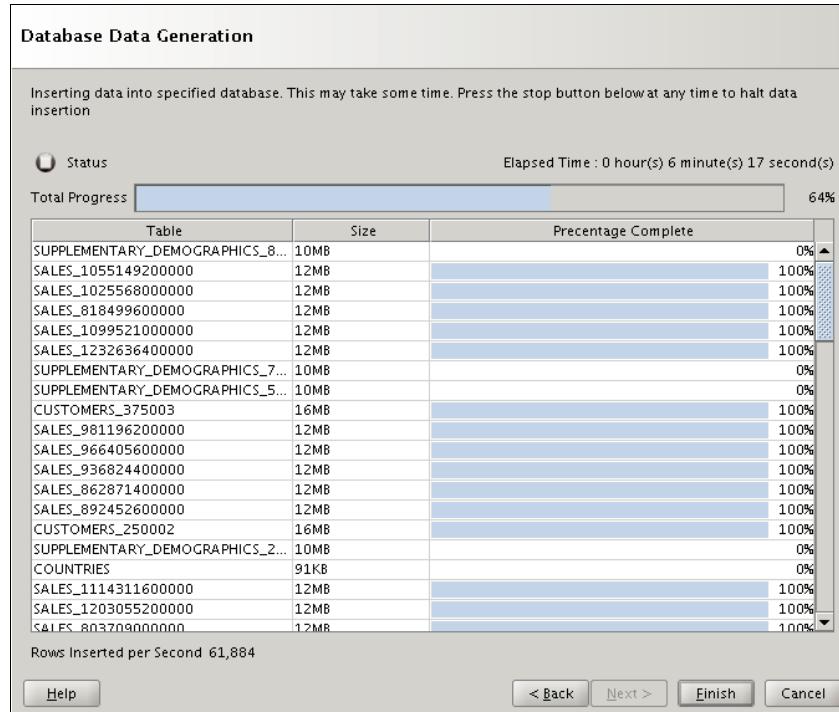


Figure 4-92 Database load traffic resumed after the swap operation to the secondary storage

We verified the status of the disk paths and the successful switch to the secondary storage. See the `1spprc` output for a database disk in Example 4-22. The paths to the primary storage are marked "Failed".

Example 4-22 Verifying the switch to the secondary storage for a database volume

```

> lspprc -p oradisk3
path          WWNN                  LSS    VOL    path
group id
group id      group status
=====
0            500507630affc16b  0xc0   0x02   SECONDARY
1(s)         5005076308ffc6d4  0x9e   0x09   PRIMARY

path      path      path      parent      connection
group id  id       status
group id
=====
0        0        Failed    fscsi0    500507630a08016b,40c0400200000000
0        1        Failed    fscsi0    500507630a08416b,40c0400200000000
0        2        Failed    fscsi1    500507630a08016b,40c0400200000000
0        3        Failed    fscsi1    500507630a08416b,40c0400200000000
1        4        Enabled   fscsi0    50050763081b06d4,409e4009000000000
1        5        Enabled   fscsi0    50050763081b46d4,409e4009000000000
1        6        Enabled   fscsi1    50050763081b06d4,409e4009000000000
1        7        Enabled   fscsi1    50050763081b46d4,409e4009000000000

```

The **iostat** command output during the unplanned HyperSwap operation is shown in Example 4-23.

Example 4-23 iostat output during the swap operation

oradisk3	100.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	92.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	100.0	0.0	0.0	0	0
oradisk3	69.3	0.0	0.0	0	0
oradisk3	1.0	384.0	10.0	0	384
oradisk3	0.0	2176.0	86.0	0	2176
oradisk3	0.0	2744.0	110.0	0	2744

The total swap time, including the detection and the actual switch of the access to the secondary storage, was about 30 seconds.

We reactivated the access to the primary storage DS8K4 on the FC switch by adding back the zone between the node r1r4m27 and the storage DS8K4 in siteA, as shown in Example 4-24.

Example 4-24 Adding the zone access of node r1r4m27 to the storage DS8K4

```
HACMP_STK6-1:admin> cfgAdd hacmp_stk6_cfg,r1r4m27_ds8k4
HACMP_STK6-1:admin> cfgEnable hacmp_stk6_cfg
You are about to enable a new zoning configuration.
This action will replace the old zoning configuration with the
current configuration selected. If the update includes changes
to one or more traffic isolation zones, the update may result in
localized disruption to traffic on ports associated with
the traffic isolation zone changes
Do you want to enable 'hacmp_stk6_cfg' configuration (yes, y, no, n): [no] y
zone config "hacmp_stk6_cfg" is in effect
Updating flash ...
```

After reactivating the access to the primary storage, the paths are automatically reactivated. The access is maintained to the storage DS8K6 in siteB as primary, while the paths to the storage in siteA are reintegrated back as secondary. See the **lspprc** output command on oradisk3 in Example 4-25.

Example 4-25 Verifying the paths status for a database disk

```
> lspprc -p oradisk3
path      WWNN          LSS   VOL    path
group id
=====
0         500507630affc16b 0xc0 0x02    SECONDARY
1(s)     5005076308ffc6d4 0x9e 0x09    PRIMARY

path      path  path        parent connection
group id  id   status
=====
0       0     Enabled     fscsi0  500507630a08016b,40c0400200000000
0       1     Enabled     fscsi0  500507630a08416b,40c0400200000000
```

0	2	Enabled	fscsi1	500507630a08016b,40c0400200000000
0	3	Enabled	fscsi1	500507630a08416b,40c0400200000000
1	4	Enabled	fscsi0	50050763081b06d4,409e400900000000
1	5	Enabled	fscsi0	50050763081b46d4,409e400900000000
1	6	Enabled	fscsi1	50050763081b06d4,409e400900000000
1	7	Enabled	fscsi1	50050763081b46d4,409e400900000000

At this time, in order to restore the access of the node to the local storage DS8K4 in siteA, we performed a C-SPOC swap operation using the smit menus shown in Figure 4-90 on page 140.

Site failover

This test is performed as an unattended stop of the nodes on the primary site, siteA, and demonstrates the node's high availability capabilities in our HyperSwap environment. This is not a Hyperswap test, but a node failover for an environment using DS8000 Metro Mirror replication and in-band communication. We performed this test in two stages: we failed the primary node r1r4m27 on siteA, which triggered a node failover to node r1r4m28 on the same site, and in the second stage we failed the node r1r4m28 for a site failover to the secondary siteB.

For testing the connection to the database, we used an Oracle client installed on the node r1r4m38, placed outside the database resource group's associated nodes. The Oracle client configuration uses a connection to the cluster service IP labels, the site-specific labels (orasrva on siteA, and orasrvb on siteB) defined in the tnsnames.ora file on node r1r4m38. See the tnsname.ora file listed in Example 4-26.

Example 4-26 The tnsnames.ora file on node r1r4m38

```
$ cat tnsnames.ora
# tnsnames.ora Network Configuration File:
/u01/app/oracle/product/11.2.0/db112_1/network/admin/tnsnames.ora
# Generated by Oracle configuration tools.

ITSOTEST1 =
(DESCRIPTION =
  (ADDRESS = (PROTOCOL = TCP)(HOST = orasrva)(PORT = 1521))
  (ADDRESS = (PROTOCOL = TCP)(HOST = orasrvb)(PORT = 1521))
  (FAILOVER = yes)
  (CONNECT_DATA =
    (SERVER = DEDICATED)
    (SERVICE_NAME = ITSOTEST1)
  )
)
```

The following steps were performed:

1. We halted the node r1r4m27 currently running the Oracle database: **halt -q**.

The resource group ITSOTEST1_CFC_RG_RDBMS was failed over to the second node r1r4m28, inside siteA. The resource group status after failover is shown in the output of Example 4-27.

Example 4-27 Resource group status after primary node r1r4m27 node failure

```
> clRGinfo -p
```

Cluster Name: r1r4m27_cluster

Resource Group Name: ITSOTEST1_CFC_RG_RDBMS

Site settings:

The following site temporarily has the highest priority for this group:
siteA, user-requested rg_move performed on Tue Jan 15 07:16:13 2013

Node	Primary State	Secondary State
r1r4m27@siteA	OFFLINE	OFFLINE
r1r4m28@siteA	ONLINE	OFFLINE
r1r4m37@siteB	OFFLINE	ONLINE SECONDARY

For verifying the connection to the database, after the database failover to node r1r4m28, we used the external Oracle client on node r1r4m38. We verified the database connection using an sql client connection to the database server. See the **sqlplus** command output run on node r1r4m38 in Example 4-28.

Example 4-28 Verifying the connection to the database

```
$ sqlplus system/xxxxxx@ITSOTEST1
```

```
SQL*Plus: Release 11.2.0.1.0 Production on Thu Jan 17 15:39:03 2013
```

```
Copyright (c) 1982, 2009, Oracle. All rights reserved.
```

Connected to:

```
Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options
SQL> select INSTANCE_NAME,HOST_NAME,VERSION,STATUS,DATABASE_STATUS from
v$instance;
```

INSTANCE_NAME

HOST_NAME

VERSION STATUS DATABASE_STATUS

```
-----
```

ITSOTEST1		
r1r4m28.austin.ibm.com		
11.2.0.1.0	OPEN	ACTIVE

2. We generated a site failover by running the following operations at the same time:
 - Halted the node r1r4m28, which had the Oracle database resource group active using the command **halt -q**.
 - Removed the access to the storage DS8K4 in siteA. For this operation, we removed the zone access on the FC switch for all three nodes associated with the Oracle database resource group to the storage subsystem DS8K4 on siteA. See an example of how to remove a zone in a Brocade switch B5K in Example 4-21 on page 141.

These events caused a second failover of the Oracle database resource group to the secondary siteB on the volumes of the storage DS8K6 in siteB. We verified the resource

group status after the site failover on the node r1r4m37 in siteB by running the **c1RGinfo** command shown in Example 4-29.

Example 4-29 Resource group status after site failover

```
> c1RGinfo -p

Cluster Name: r1r4m27_cluster

Resource Group Name: ITSOTEST1_CFC_RG_RDBMS
Site settings:
The following site temporarily has the highest priority for this group:
siteA, user-requested rg_move performed on Tue Jan 15 07:15:59 2013

Node           Primary State   Secondary State
-----
r1r4m27@siteA          OFFLINE      OFFLINE
r1r4m28@siteA          OFFLINE      OFFLINE
r1r4m37@siteB        ONLINE       OFFLINE
```

The database was running using the volumes in the storage in siteB. We verified the disk path status using the **1spprc** command. See Example 4-30.

Example 4-30 Disk path status for an Oracle database disk

```
> 1spprc -p oradisk3
path      WWNN          LSS  VOL  path
group id
group status
=====
0(s)    5005076308ffc6d4  0x9e  0x09  PRIMARY
1        500507630affc16b  0xc0  0x02  SECONDARY

path      path  path      parent  connection
group id  id   status
=====
0     0   Enabled   fscsi0  50050763081b86d4,409e400900000000
0     1   Enabled   fscsi0  50050763081bc6d4,409e400900000000
0     2   Enabled   fscsi1  50050763081b86d4,409e400900000000
0     3   Enabled   fscsi1  50050763081bc6d4,409e400900000000
1     4   Failed    fscsi0  500507630a08016b,40c0400200000000
1     5   Failed    fscsi0  500507630a08816b,40c0400200000000
1     6   Failed    fscsi1  500507630a08016b,40c0400200000000
1     7   Failed    fscsi1  500507630a08816b,40c0400200000000
```

In the above example, we observed that the paths to the DS8K4 storage were in the Failed state. The node r1r4m37 was, at this time, using the volumes on the storage DS8K6 on siteB.

The connection to the database was checked using the sql client on node r1r4m38. See the output of the sql connection test in Example 4-31.

Example 4-31 Testing the database connection

```

$ sqlplus system/oracle@ITSOTEST1

SQL*Plus: Release 11.2.0.1.0 Production on Thu Jan 17 18:19:29 2013

Copyright (c) 1982, 2009, Oracle. All rights reserved.

Connected to:
Oracle Database 11g Enterprise Edition Release 11.2.0.1.0 - 64bit Production
With the Partitioning, OLAP, Data Mining and Real Application Testing options

SQL> select INSTANCE_NAME,HOST_NAME,VERSION,STATUS,DATABASE_STATUS from
v$instance;

INSTANCE_NAME
-----
HOST_NAME
-----
VERSION      STATUS      DATABASE_STATUS
-----
ITSOTEST1
r1r4m37.austin.ibm.com
11.2.0.1.0    OPEN        ACTIVE

```

In order to fallback the resource group to its primary state, after the nodes and the storage access to the storage in the primary site, siteA, are up again, we reintegrated the nodes r1r4m27 and r1r4m28 back into the cluster.

Before starting the cluster services on the nodes in siteA, we added back the zones for the storage access on the FC switch. See the path status of the database disks on the cluster node r1r4m37 in Example 4-32.

Example 4-32 Status of the disk paths on a node after restoring access to the storage DS8K4

```

(0) root @ r1r4m37: /
> lspprc -p oradisk3
path      WWNN          LSS  VOL   path
group id
=====
0(s)      5005076308ffc6d4 0x9e 0x09  PRIMARY
1          500507630afffc16b 0xc0 0x02  SECONDARY

path      path  path      parent  connection
group id  id   status
=====
0     0     Enabled   fscsi0  50050763081b86d4,409e400900000000
0     1     Enabled   fscsi0  50050763081bc6d4,409e400900000000
0     2     Enabled   fscsi1  50050763081b86d4,409e400900000000
0     3     Enabled   fscsi1  50050763081bc6d4,409e400900000000
1     4     Enabled   fscsi0  500507630a08016b,40c0400200000000
1     5     Enabled   fscsi0  500507630a08816b,40c0400200000000
1     6     Enabled   fscsi1  500507630a08016b,40c0400200000000

```

1	7	Enabled	fscsi1	500507630a08816b,40c0400200000000
---	---	---------	--------	-----------------------------------

We started the cluster services on the siteA nodes using the normal smit C-SPOC operation: **smitty clstart**. Because of the resource group inter-site policy set to Prefer Primary Site, the resource group automatically fell back to the primary siteA, on the highest priority available node, r1r4m27. See the **c1RGinfo** command output in Example 4-33 that shows the cluster status after the entire fallback operation had completed.

Example 4-33 Cluster resource group status after resource group fallback to the primary site siteA

```
(0) root @ r1r4m37: /  
> c1RGinfo -p
```

Cluster Name: r1r4m27_cluster

Resource Group Name: **ITSOTEST1_CFC_RG_RDBMS**

Site settings:

The following site temporarily has the highest priority for this group:
siteA, user-requested rg_move performed on Tue Jan 15 07:15:59 2013

Node	Primary State	Secondary State
r1r4m27@siteA	ONLINE	OFFLINE
r1r4m28@siteA	OFFLINE	OFFLINE
r1r4m37@siteB	OFFLINE	ONLINE SECONDARY

After the Oracle database resource group had fallen back to the primary siteA, the active paths were maintained to the storage DS8K6 in the secondary siteB (see Example 4-33). In order to activate the paths to the primary storage DS8K4, we performed a new swap operation for all the disks (database and repository LUNs), using smit C-SPOC menus and selecting the **Manage Mirror Groups by Site** operation, selecting as source volumes the volumes associated with siteB. The access was restored back to the primary storage. You can see the path status for one of the Oracle database disks in Example 4-34.

Example 4-34 Swapping the primary I/O access of the cluster nodes to the storage DS8K4 in siteA

```
> lspprc -p oradisk4  
path      WWNN          LSS  VOL   path  
group id                           group status  
=====  
0(s)      500507630afffc16b  0xc0  0x01  PRIMARY  
1          5005076308ffc6d4  0x9e  0x08  SECONDARY  
  
path      path  path        parent  connection  
group id  id    status  
=====  
0       0     Enabled    fscsi0  500507630a08016b,40c0400100000000  
0       1     Enabled    fscsi0  500507630a08416b,40c0400100000000  
0       2     Enabled    fscsi1  500507630a08016b,40c0400100000000  
0       3     Enabled    fscsi1  500507630a08416b,40c0400100000000  
1       4     Enabled    fscsi0  50050763081b06d4,409e400800000000  
1       5     Enabled    fscsi0  50050763081b46d4,409e400800000000
```

1	6	Enabled	fscsi1	50050763081b06d4,409e400800000000
1	7	Enabled	fscsi1	50050763081b46d4,409e400800000000



Cross-site LVM mirroring with IBM PowerHA SystemMirror 7.1.2 Standard Edition

In this chapter we provide details about how to set up a cross-site Logical Volume Manager (LVM) mirroring cluster using IBM AIX 7.1 and PowerHA SystemMirror 7.1.2 Standard Edition. This is a common scenario for a campus-style disaster recovery (DR) solutions.

We cover the following topics:

- ▶ Cross-site LVM mirroring overview
 - Requirements
 - Planning considerations
- ▶ Testing the environment
 - Configuring the cluster topology
 - Configuring the cluster resources
- ▶ Configuring the cross-site LVM cluster
- ▶ Test scenarios
 - Test case 1: Both nodes down site failure
 - Test case 2: Rolling node failures for site outage
 - Test case 3: Outage of a storage subsystem
 - Test case 4: Rolling disaster

5.1 Cross-site LVM mirroring overview

The main difference between local clusters and cluster solutions with cross-site mirroring is as follows:

- ▶ In local clusters, all nodes and storage subsystems are located in the same location.
- ▶ With cross-site mirrored clusters, nodes and storage subsystems reside on different sites.
- ▶ Each site has at least one cluster node and one storage subsystem with all necessary IP and SAN connectivity, similar to a local cluster.
- ▶ Use *ignore* for the resource group inter-site management policy.

The increased availability of metropolitan area networks (MAN) in recent years has made this solution more feasible and popular.

This solution offers automation of AIX LVM mirroring within SAN disk subsystems between different sites. It also provides automatic LVM mirroring synchronization and disk device activation when, after a disk or site failure, a node or disk becomes available.

Each node in a cross-site LVM cluster accesses all storage subsystems. The data availability is ensured through the LVM mirroring between the volumes residing on different storage subsystems on different sites.

In case of a complete site failure, PowerHA performs a takeover of the resources to the secondary site according to the cluster policy configuration. It activates all defined volume groups from the surviving mirrored copy. In case one storage subsystem fails, I/O may experience a temporary delay but it continues to access data from the active mirroring copy on the surviving disk subsystem.

PowerHA drives automatic LVM mirroring synchronization, and after the failed site joins the cluster, it automatically fixes removed and missing physical volumes (PV states removed and missing) and synchronizes data. However, automatic synchronization is not possible for all cases. But C-SPOC can be used to synchronize the data from the surviving mirrors to stale mirrors after a disk or site failure as needed.

5.1.1 Requirements

The following requirements must be met to assure data integrity and appropriate PowerHA reaction in case of site or disk subsystem failure:

- ▶ The force varyon attribute for the resource group must be set to *true*.
- ▶ The logical volumes allocation policy must be set to superstrict (this ensures that LV copies are allocated on different volumes, and the primary and secondary copy of each LP is allocated on disks located in different sites).
- ▶ The LV mirrored copies must be allocated on separate volumes that reside on different disk subsystems (on different sites).
- ▶ Mirror pools should be used to help define the disks at each site.

When adding additional storage space, for example increasing the size of the mirrored file system, it is necessary to assure that the new logical partitions will be allocated on different volumes and different disk subsystems according to the requirements above. For this task, it is required to increase the logical volume first with the appropriate volume selections, then increase the file system, preferably by using C-SPOC.

5.1.2 Planning considerations

Here we describe some considerations regarding SAN setup, fiber channel connections and the LAN environment. The considerations and limitations are based on the technologies and protocols used for cross-site mirroring cluster implementation.

The SAN network can be expanded beyond the original site, by way of advanced technology. Here is an example of what kind of technology could be used for expansion. This list is not exhaustive:

- ▶ FCIP router
- ▶ Wave division multiplexing (WDM) devices. This technology includes:
 - CWDM stands for Coarse Wavelength Division Multiplexing, which is the less expensive component of the WDM technology.
 - DWDM stands for Dense Wave length Division Multiplexing.

Of course, the infrastructure should be resilient to failures by providing redundant components. Also, the SAN interconnection must provide sufficient bandwidth to allow for adequate performance for the synchronous-based mirroring that LVM provides. Distance between sites is important for latency that also can result in performance issues. Consult the whitepaper *Understanding the Performance Implications of Cross-Site Mirroring with AIX's Logical Volume Manager*, which is available at:

<http://tinyurl.com/xsitevmpref>

Tips for faster disk failure detection

AIX disk storage subsystem failure detection can be accelerated by judiciously changing attributes of hdisk and interfaces. Most changes recommended here are generally applicable, although as noted in Table 5-1, some changes are applicable only to certain disk storage subsystems. Always consult with the storage vendor for additional recommendations.

Table 5-1 Device type, device name, device attributes and settings

Device type	Device name	Device attribute	Setting
Virtual SCSI	vscsi#	vscsi_err_recov	fast_fail
FC SCSI I/O	fscsi#	dyntrk	yes
		fc_err_recover	fast_fail
Disk	hdisk#	algorithm	round_robin
		hcheck_interval	30
		timeout_policy	fail_path
DS8000 specific		FC3_REC	true

Tip: Details on MPIO best practices, including new algorithm attribute in AIX 6.1.9/7.1.3 of *shortest_queue* can be found at:

<http://www.ibm.com/developerworks/aix/library/au-aix-mpio/index.html>

It is important to note that the AIX disk storage subsystem failure detection is prolonged in a redundant virtual SCSI environment. Virtual Fibre Channel (NPIV) is a much better choice than virtual SCSI to minimize the time required for AIX to detect and recover from a disk storage subsystem failure in a virtualized environment.

The number of paths to each hdisk/LUN affects the time required to detect and recover from a disk storage subsystem failure. The more paths the longer recovery will take. We generally recommend no more than two paths per server port through which a given LUN can be accessed. Two server ports per LUN are generally sufficient unless I/O bandwidth requirements are extremely high.

When a disk storage subsystem fails, AIX must detect failure of each LUN presented by the subsystem. The more LUNs, the more potential application write I/O stalls while AIX detects a LUN failure. The number of LUNs is dictated by the LUN size and the disk storage capacity required. Since many I/O requests can be driven simultaneously to a single hdisk/LUN (limited only by the hdisk's queue_depth attribute), we generally recommend fewer larger LUNs as compared to more smaller LUNs.

But too few LUNs can sometimes lead to a performance bottleneck. The AIX hdisk device driver is single threaded, which limits the number of IOPS which AIX can drive to a single hdisk. It is therefore inadvisable to drive more than 5,000 IOPS to a single hdisk.

Note: In most cases IOPS to a single hdisk/LUN will be constrained by disk storage subsystem performance well before the hdisk IOPS limit is reached, but the limit can easily be exceeded on LUNs residing on FlashSystem or on solid-state disk drives.

If the anticipated workload on a volume group is such that the hdisk IOPS limit might be exceeded, create the volume group on more smaller hdisks/LUNs.

If an anticipated workload is such that it tends to drive I/O to only one file or file system at a time (especially when I/O requests are sequential), then to help AIX process failures of multiple LUNs simultaneously, stripe logical volumes (including those underlying file systems) across hdisks/LUNs using the -S flag on the `mk1v` command.

Note: It will not help to configure logical volumes with physical partition striping by specifying -e x on the `mk1v` command. If physical partition striping is used with such a workload, AIX might still process a disk storage subsystem failure one LUN at a time.

Pros and cons

The following is a list of reasons why implementing cross-site LVM mirroring may be desirable:

- ▶ It is inexpensive. LVM mirroring is a no charge feature of AIX.
- ▶ Storage agnostic. This means it does not require special storage to utilize.
- ▶ Easy to implement. Most AIX admins are knowledgeable about LVM mirroring.
- ▶ Unlike most storage replication offerings, both copies have full host read/write access.

The following is a list of reasons why implementing cross-site LVM mirroring may *not* be desirable:

- ▶ Specific to AIX, no heterogeneous operating system support.
- ▶ Synchronous replication only.
- ▶ Cannot copy *raw* disks.
- ▶ Potential system performance implications because twice the write I/Os are generated.
- ▶ No way to define primary (source) or secondary (target) copies. Though you can specify preferred read option for storage like flash.
- ▶ AIX I/O hang times in the event of a copy loss may still not prevent application failure as experienced in “Test case 3: Outage of a storage subsystem” on page 172.

- ▶ Quorum is usually disabled, and forced varyon of the volume group enabled. This can lead to data inconsistencies as discovered in “Test case 4: Rolling disaster” on page 176.
- ▶ Like most data replication, it too is good at copying bad data. This means it does not eliminate the need backup and backout procedures.

Important: Even with redundant components, most environments, such as cross-site LVM mirroring, can only handle any one failure. That failure needs to be corrected before another failure occurs. A series of rolling failures may still result in an outage, or worse, data corruption.

5.2 Testing the environment

In this environment we had two sites, *athens* and *rome*. Each site contained two nodes and a single storage subsystem. Site *athens* nodes were *xlvma1* and *xlvma2*. Site *rome* nodes were *xlvmb1* and *xlvmb2*. An overview of our environment is shown in Figure 5-1. Although the site names suggest a significant distance between them, they are located in the same data center.

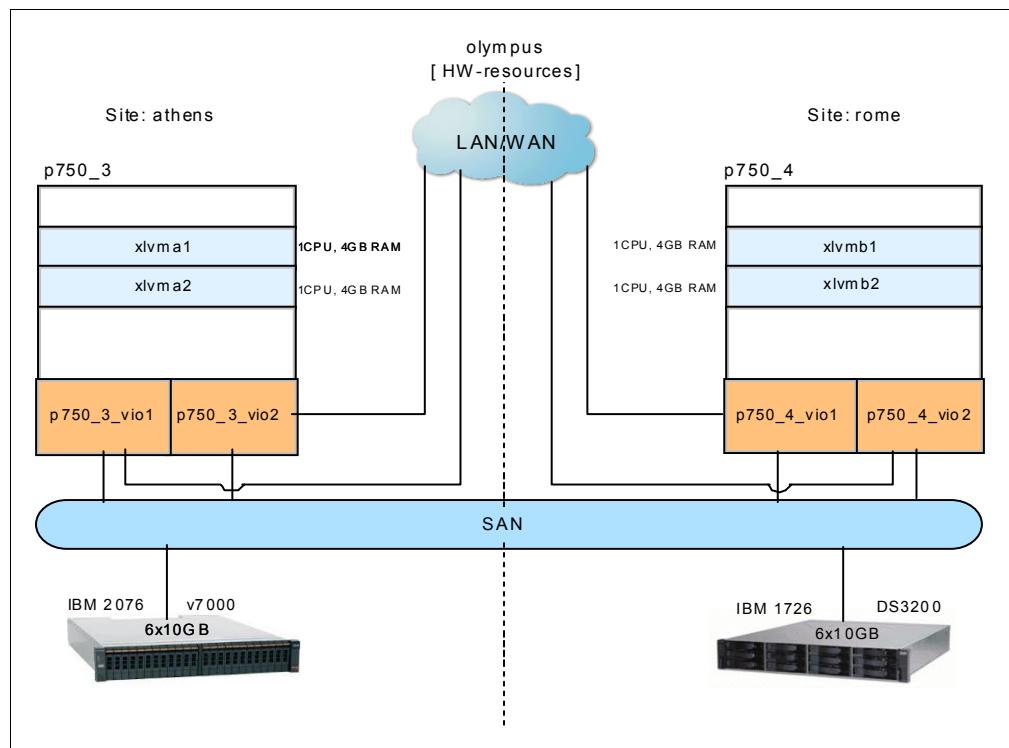


Figure 5-1 Cross-site LVM test cluster overview

Nodes

All nodes are logical partitions (LPAR) on an IBM 750 POWER Systems server. Each has its own *rootvg* located on the external storage, as shown in Figure 5-2 on page 156. Each LPAR contains one processing unit and 4 GB RAM.

SAN access

LPAR access to the external SAN resources is provided by a redundant configuration of two Virtual I/O Servers on the IBM p750 server. Thus, four Virtual I/O Servers are connected to both storage subsystems via two FC adapters.

We used N_Port ID Virtualization (NPIV) for the FC connections and Virtual Ethernet adapters for the LAN connections. From the LPAR perspective, each virtual FC client adapter was attached to both storage subsystems, with connectivity provided by appropriate *zoning* in the SAN switches.

Network access

We connected all four nodes to two different network segments, one dedicated for administration purposes named *admin_net*, and the other for application and client access named *service_net*, as shown in Figure 5-2.

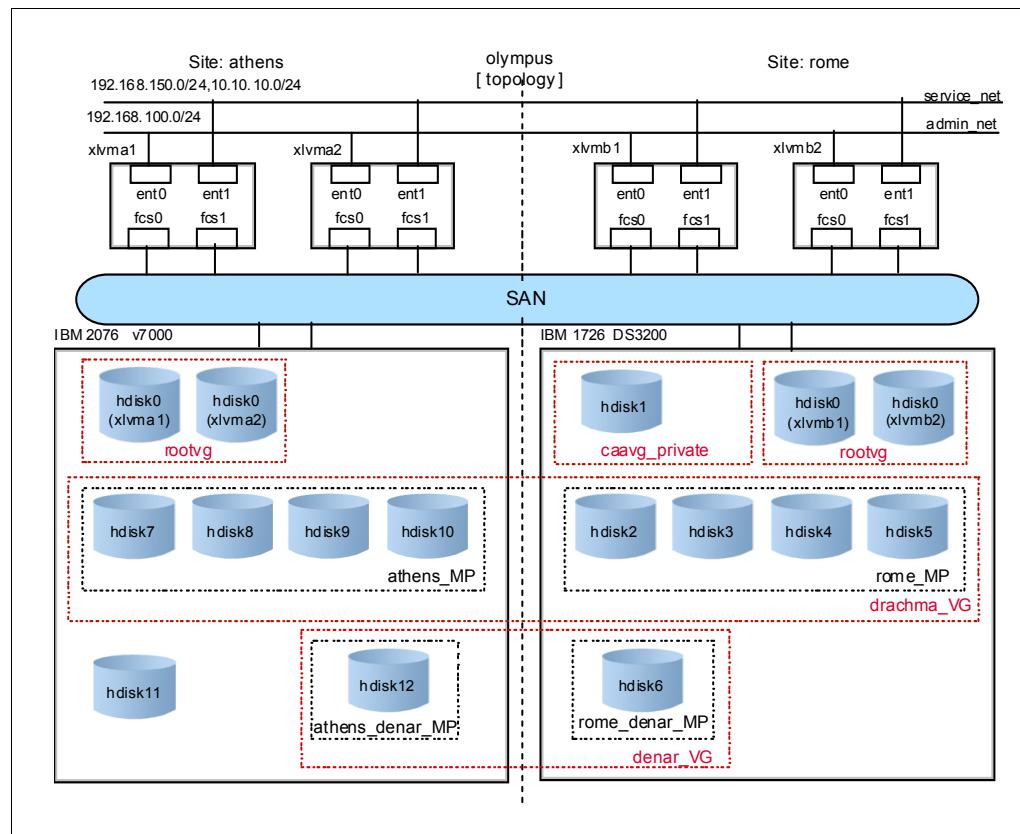


Figure 5-2 Test cluster details

Storage resources

On both storage subsystems we configured a pool of six LUNs each as shared disks among the participating nodes, dedicated to the cluster repository (*hdisk1*) and two Volume Groups to be used (*hdisk2*, *hdisk3*, ..., *hdisk10*, *hdisk12*). The LUN corresponding to the name *hdisk11* was not assigned to any Volume Group, as shown in Figure 5-2.

5.3 Configuring the cross-site LVM cluster

This section describes how to configure a cross-site LVM cluster.

5.3.1 Configuring the cluster topology

We created a cluster and named it *olympus*. We used the SMIT menus shown in Figure 5-3. We entered the **smitty sysmirror** command and then selected **Cluster Nodes and Networks → Multi Site Cluster Deployment → Setup a Cluster, Sites, Nodes and Networks**.

Setup Cluster, Sites, Nodes and Networks			
Type or select values in entry fields. Press Enter AFTER making all desired changes.			
[Entry Fields]			
* Cluster Name	[olympus]		
* Site 1 Name	[athens]		
* New Nodes (via selected communication paths)	[xlvma1 xlvma2] +		
* Site 2 Name	[rome]		
* New Nodes (via selected communication paths)	[xlvmb1 xlvmb2] +		
Cluster Type	[Stretched Cluster] +		
F1=Help	F2=Refresh	F3=Cancel	F4=List
F5=Reset	F6=Command	F7>Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Figure 5-3 Define sites, nodes, and cluster type

As required by PowerHA version 7, we had to choose a *repository disk* as shown in Figure 5-4. We executed **smitty sysmirror** and then select **Cluster Nodes and Networks → Multi Site Cluster Deployment → Setup a Cluster, Sites, Nodes and Networks → Define Repository Disk and Cluster IP Address**.

Define Repository and Cluster IP Address	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
[Entry Fields]	
* Cluster Name	olympus
* Repository Disk	[hdisk1] +
Cluster IP Address	[]

Figure 5-4 Define repository and cluster multicast address

In order for *hdisk1* to be used, or chosen from the F4 pick list, the disk must have a *PVID* assigned to it and known to every node in the cluster.

If no specific multicast is entered for the *cluster IP address*, PowerHA creates a valid one automatically. This is done by replacing the first octet in the IP address of the first node defined, *xlvma1*, with 228. In our case this resulted in a multicast address of 228.168.100.65.

We defined two networks, one for administration purposes named *admin_net*, and another for application and client access named *service_net*. We then added the interfaces *en0* from each node to the *admin_net* and the interfaces *en1* from each node to *service_net*. The resulting topology is shown in Figure 5-5.

```
root@x1vma1:/>/usr/es/sbin/cluster/utilities/cltopinfo -w
Network admin_net
  NODE x1vma1:
    x1vma1 192.168.100.65
  NODE x1vma2:
    x1vma2 192.168.100.66
  NODE x1vmb1:
    x1vmb1 192.168.100.67
  NODE x1vmb2:
    x1vmb2 192.168.100.68

Network service_net
  NODE x1vma1:
    x1vma1_boot      192.168.150.65
  NODE x1vma2:
    x1vma2_boot      192.168.150.66
  NODE x1vmb1:
    x1vmb1_boot      192.168.150.67
  NODE x1vmb2:
    x1vmb2_boot      192.168.150.68
```

Figure 5-5 Networks configured

5.3.2 Configuring the cluster resources

In our example the cluster has already been created with nodes and networks.

Add site specific service IP(s)

Next, we defined the service IP labels used within the cluster. These labels must be resolvable locally by being included in /etc/hosts. Since we had a resource group, *drachma*, accessible on the two sites via different IP addresses, the service IP labels to be used within *drachma*, *hermes* and *artemis* were configured accordingly, as shown in Figure 5-6 on page 159 and Figure 5-7 on page 159. We executed **smitty sysmirror** command and then select **Cluster Applications and Resources → Resources → Configure Service IP Labels/Addresses → Add a Service IP Label/Address** and then chose the desired network *service_net*.

Add a Service IP Label/Address configurable on Multiple Nodes (extended)

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

<ul style="list-style-type: none"> * IP Label/Address Netmask(IPv4)/Prefix Length(IPv6) * Network Name Associated Site 	[Entry Fields] artemis + [255.255.255.0] service_net rome +
--	---

Figure 5-6 Adding site-specific service address for rome

Add a Service IP Label/Address configurable on Multiple Nodes (extended)

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

<ul style="list-style-type: none"> * IP Label/Address Netmask(IPv4)/Prefix Length(IPv6) * Network Name Associated Site 	[Entry Fields] hermes + [] service_net athens +
--	---

Figure 5-7 Adding site-specific service address for site athens

The service IP labels belonging to the resource group *denar* are not site-specific, so we configured them as shown in Figure 5-8.

Add a Service IP Label/Address configurable on Multiple Nodes (extended)

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

<ul style="list-style-type: none"> * IP Label/Address Netmask(IPv4)/Prefix Length(IPv6) * Network Name Associated Site 	[Entry Fields] diana + [] service_net ignore +
--	--

Figure 5-8 Adding a service address to stay within a site

Specify service IP distribution preference

It was our preference to always have the service IP label and address be configured on the *service_net*. This means they were bounded to the Ethernet adapter *en1*. We configured this by using the **smitty sysmirror** command and then selected **Cluster Applications and Resources** → **Resources** → **Configure Service IP Labels/Addresses** → **Configure Service IP Labels/Address Distribution Preference**, as shown in Figure 5-9 on page 160.

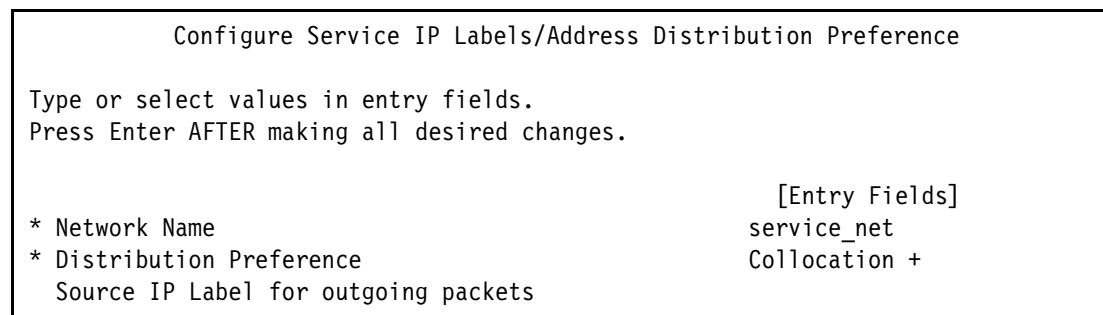


Figure 5-9 Service IP address distribution preference

Add application controllers

We now defined the application controllers by executing the **smitty sysmirror** command and then selected **Cluster Applications and Resources** → **Resources** → **Configure User Applications (Scripts and Monitors)** → **Application Controller Scripts** → **Add Application Controller Scripts**, as shown in Figure 5-10. We then added two resource groups named *drachma* and *denar*.

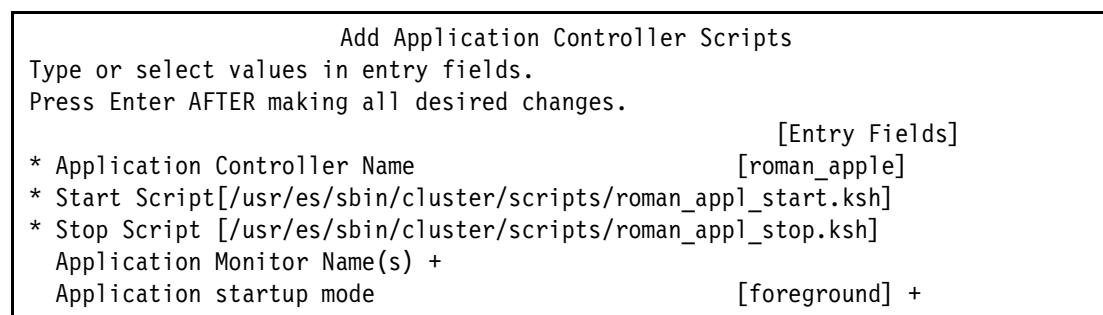


Figure 5-10 Application controller scripts

Add resource group(s)

Resource group *drachma* was designed to failover across sites (Figure 5-11), whereas *denar* ran only within site *rome* as shown in Figure 5-10 and Figure 5-12 on page 161, respectively. To achieve this we executed the **smitty sysmirror** command and then selected **Cluster Applications and Resources** → **Resource Groups** → **Add a Resource Group (extended)**.

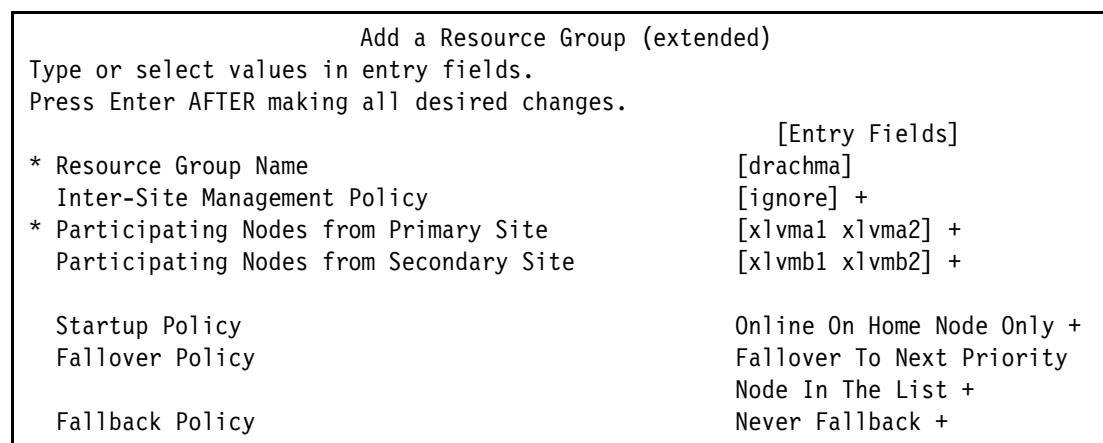


Figure 5-11 Add resource group for either site

Attention: When configuring cross-site LVM mirroring, the *inter-site management policy* for the resource group should be set to *ignore*.

```
Add a Resource Group (extended)
Type or select values in entry fields.
Press Enter AFTER making all desired changes.

* Resource Group Name [Entry Fields]
  Inter-Site Management Policy [denar]
* Participating Nodes from Primary Site [ignore] +
  Participating Nodes from Secondary Site [xlvmb1 xlvmb2] +
[] +

Startup Policy [Online On Home Node Only +]
Fallback Policy [Failover To Next Priority]
                [Node In The List +]
                [Never Fallback +]

Fallback Policy
```

Figure 5-12 Adding a resource group for preferred primary site

Create volume group(s)

Next, we created two volume groups, *drachma_VG* and *denar_VG*, and assigned them to their respective resource groups, as shown in Figure 5-13 and Figure 5-15 on page 162. In this case, we used C-SPOC via the **smitty sysmirror** command and then selected **System Management (C-SPOC) → Storage → Volume Groups → Volume Groups → Create a Volume Group**. We chose the hdisks and the *scalable* volume group type.

```
Create a Scalable Volume Group
Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP] [Entry Fields]
  Node Names xlvmb2,xlvma1,xlvma2
  Resource Group Name [drachma] +
  PVID 00f6f5d00624aaa5 00f6f5d00624ab46 >
  VOLUME GROUP name [drachma_VG]
  Physical partition SIZE in megabytes 16 +
  Volume group MAJOR NUMBER [36] #
  Enable Fast Disk Takeover or Concurrent Access Fast Disk Takeover +
  Volume Group Type Scalable
  CRITICAL volume group? no +
  Maximum Physical Partitions in units of 1024 32 +
  Maximum Number of Logical Volumes 256 +
  Enable Strict Mirror Pools superstrict +
  Mirror Pool name []
  Warning:
    Changing the volume group major number may result
    in the command being unable to execute
    successfully on a node that does not have the
    major number currently available. Please check
  [MORE...3]
```

Figure 5-13 Creating the scalable shared volume group *drachma_VG*

It is normal to see the messages as shown in Figure 5-14. Though not specifically stated, these are generally warnings and not errors.

```
COMMAND STATUS

Command: OK          stdout: yes          stderr: no

Before command completion, additional instructions may appear below.

xlvma1: ReserveOfGreece
xlvma1: mkvg: This concurrent capable volume group must be varied on manually.
xlvma1: synclovdm: No logical volumes in volume group drachma_VG.
xlvma1: Volume group ReserveOfGreece has been updated.
xlvmb2: synclovdm: No logical volumes in volume group drachma_VG.
xlvmb2: 0516-783 importvg: This imported volume group is concurrent capable.
xlvmb2: Therefore, the volume group must be varied on manually.
xlvmb2: 0516-1804 chvg: The quorum change takes effect immediately.
xlvmb2: Volume group ReserveOfGreece has been imported.
xlvma2: synclovdm: No logical volumes in volume group drachma_VG.
xlvma2: 0516-783 importvg: This imported volume group is concurrent capable.
xlvma2: Therefore, the volume group must be varied on manually.
xlvma2: 0516-1804 chvg: The quorum change takes effect immediately.
xlvma2: Volume group ReserveOfGreece has been imported.
xlvmb1: synclovdm: No logical volumes in volume group drachma_VG.
xlvmb1: 0516-783 importvg: This imported volume group is concurrent capable.
xlvmb1: Therefore, the volume group must be varied on manually.
xlvmb1: 0516-1804 chvg: The quorum change takes effect immediately.
xlvmb1: Volume group ReserveOfGreece has been imported.
cl_mkvg: The HACMP configuration has been changed - Volume Group drachma_VG has been added. The configuration must be synchronized to make this change effective across the cluster
cl_mkvg: Discovering Volume Group Configuration...
```

Figure 5-14 Volume group creation messages

Create a Scalable Volume Group		
Type or select values in entry fields.		
Press Enter AFTER making all desired changes.		
[TOP] Node Names Resource Group Name PVID 00f6f5d00624aela VOLUME GROUP name Physical partition SIZE in megabytes Volume group MAJOR NUMBER Enable Fast Disk Takeover or Concurrent Access Volume Group Type CRITICAL volume group? Maximum Physical Partitions in units of 1024 Maximum Number of Logical Volumes Enable Strict Mirror Pools Mirror Pool name	[Entry Fields] xlvmb1,xlvmb2 [denar] + 00f6f5d00624ac28 [denar_VG] 16 + [37] Fast Disk Takeover + Scalable no + 32 + 256 + yes + []	
Warning: Changing the volume group major number may result in the command being unable to execute successfully on a node that does not have the major number currently available. Please check		

Figure 5-15 Creating scalable shared volume group drachma_VG

Define mirror pools

In order to provide data strictly mirrored between the two sites, we defined two *mirror pools* for each Volume Group. Again, we used **smitty sysmirror** and then selected **System Management (C-SPOC) → Storage → Volume Groups → Manage Mirror Pools for Volume Groups → Add Disks to a Mirror Pool** as shown in Figure 5-16 and Figure 5-17.

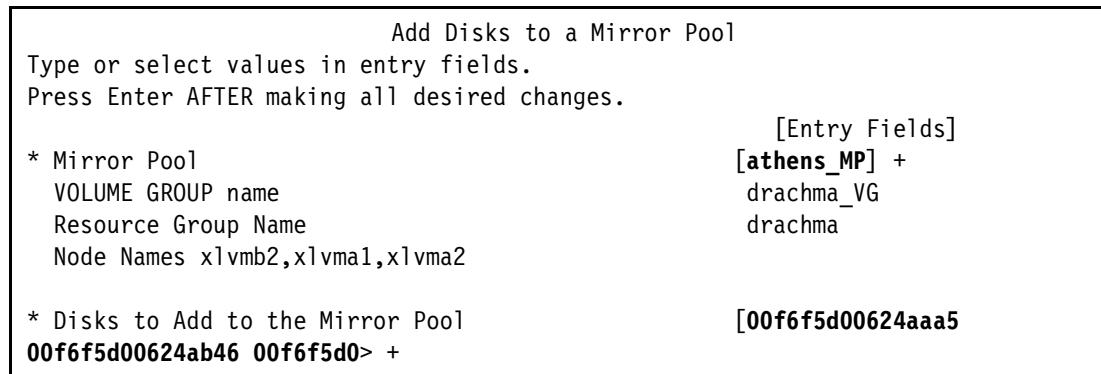


Figure 5-16 Mirror pool for drachma_VG

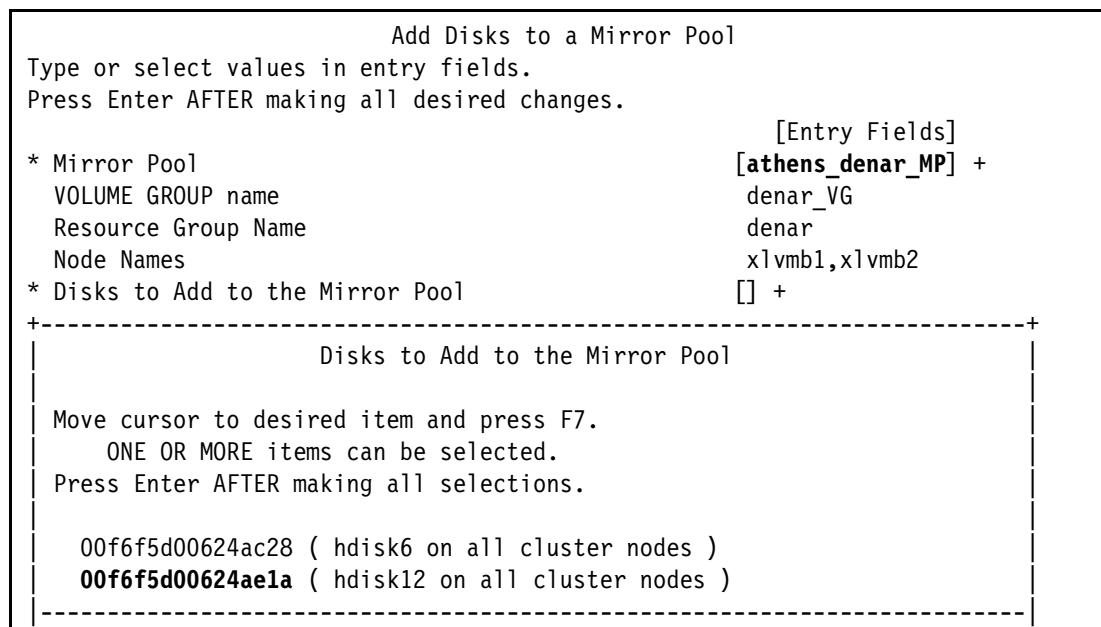


Figure 5-17 Mirror pool for denar_VG

Add logical volume(s)

Now we added logical volumes and file systems on the volume groups. Be aware of keeping the strictness in the mirror policy as shown in Figure 5-18 on page 165. We used **smitty sysmirror** and then selected **System Management (C-SPOC) → Storage → Logical Volumes → Add a Logical Volume**. Next we selected the desired volume group and the physical volumes to contain the new logical volume.

Add resources into resource group

Finally, we added the resources to the two resource groups as shown in Figure 5-20 on page 166. We executed **smitty sysmirror** and then selected **Cluster Applications and Resources → Resource Groups → Change/Show All Resources and Attributes for a**

Custom Resource Group and selected there the desired resource group, in our example *drachma*.

Synchronize cluster

At this point the cluster configuration was complete. We now needed to verify and synchronize the cluster to propagate the information to all the other nodes in the cluster. We used the **smitty sysmirror** command and then selected **Cluster Applications and Resources → Verify and Synchronize Cluster Configuration**.

During our first verification and synchronization of the cluster topology and resources some errors were logged by the verification tools invoked by SMIT as shown in Figure 5-21 on page 167. However, it did actually complete successfully. All subsequent actions performed after changing the cluster configuration no longer produced these errors.

An overview of our cluster topology and resources was retrieved with the PowerHA commands **c1topinfo** and **c1RGinfore**. Both of these commands are located in /usr/es/sbin/cluster/utilities. See Figure 5-22 on page 168 and Figure 5-23 on page 169.

Add a Logical Volume	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
[TOP]	[Entry Fields]
Resource Group Name	drachma
VOLUME GROUP name	drachma_VG
Node List x1vmb2,x1vma1,x1vma2	
Reference node	x1vma1
* Number of LOGICAL PARTITIONS	[20] #
PHYSICAL VOLUME names	hdisk2 hdisk3 hdisk4 hdisk5 hdisk7 hdisk8 >
Logical volume NAME	[drachma_LV1]
Logical volume TYPE	[jfs2] +
POSITION on physical volume	outer_middle +
RANGE of physical volumes	minimum +
MAXIMUM NUMBER of PHYSICAL VOLUMES	[]
to use for allocation	
Number of COPIES of each logical	
partition	2 +
Mirror Write Consistency?	active +
Allocate each logical partition copy	yes +
on a SEPARATE physical volume?	
RELOCATE the logical volume during reorganization?	yes +
Logical volume LABEL	[]
MAXIMUM NUMBER of LOGICAL PARTITIONS	[512] #
Enable BAD BLOCK relocation?	yes +
SCHEDULING POLICY for reading/writing	parallel +
logical partition copies	
Enable WRITE VERIFY?	no +
File containing ALLOCATION MAP	[]
Stripe Size?	[Not Striped] +
Serialize I/O?	no +
Make first block available for applications?	no +
Mirror Pool for First Copy	athens_MP +
Mirror Pool for Second Copy	rome_MP +
Mirror Pool for Third Copy + ...	

Figure 5-18 Creating the logical volumes

As an alternative to **c1stat** to obtain the status of the cluster, you can also use the command sequence shown in Figure 5-19.

```
root@x1vma1:/var/adm/ras>c1cmd lssrc -ls c1strmgrES|egrep 'NODE |Current state'
NODE x1vmb2
Current state: ST_INIT
NODE x1vmb1
Current state: ST_STABLE
NODE x1vma2
Current state: ST_INIT
NODE x1vma1
Current state: ST_STABLE
```

Figure 5-19 c1cmd command showing the cluster status

Change/Show All Resources and Attributes for a Custom Resource Group
Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[TOP]

Resource Group Name	[Entry Fields] drachma
Inter-site Management Policy	Online On Either Site
Participating Nodes from Primary Site	x1vma1 x1vma2
Participating Nodes from Secondary Site	x1vmb2
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority
Node In The List	Never Fallback
Fallback Policy	[artemis hermes] +
Service IP Labels/Addresses	[greek_apple] +
Application Controllers	[drachma_VG] +
Volume Groups	true +
Use forced varyon of volume groups, if necessary	false +
Automatically Import Volume Groups	
Allow varyon with missing data updates? (Asynchronous GLVM Mirroring Only)	true +
Default choice for data divergence recovery (Asynchronous GLVM Mirroring Only)	ignore
Filesystems (empty is ALL for VGs specified)	[] +
Filesystems Consistency Check	fsck +
Filesystems Recovery Method	sequential +
PPRC Replicated Resources	[] +
Filesystems mounted before IP configured	true + +
Filesystems/Directories to Export (NFSv2/3)	[] +
Filesystems/Directories to Export (NFSv4)	[] +
Stable Storage Path (NFSv4)	[]

Figure 5-20 Adding resources to the resource group

```
COMMAND STATUS
Command: OK          stdout: yes        stderr: no
Before command completion, additional instructions may appear below.
[TOP]
Timer object autoclverify already exists
0519-003 libodm: The specified object identifier did not refer to
a valid object. The object identifier must be an integer greater than
zero.
ERROR: failed to set athens as the secondary site.
0519-003 libodm: The specified object identifier did not refer to
a valid object. The object identifier must be an integer greater than
zero.
ERROR: failed to set rome as the dominant site.
ERROR: failed to properly set the site dominance.Verification to be performed
on the following:
    Cluster Topology
    Cluster Resources
Verification will interactively correct verification errors.
Retrieving data from available cluster nodes. This could take a few minutes.
Start data collection on node xlvmal
    Start data collection on node xlvmal2
    Start data collection on node xlvmbl
    Start data collection on node xlvmb2
    Collector on node xlvmbl completed
    Collector on node xlvmb2 completed
    Collector on node xlvmal completed
[MORE...90]
```

Figure 5-21 Errors for the initial verification and synchronization

```

root@xlvma1:/var/adm/ras>/usr/es/sbin/cluster/utilities/cltopinfo
Cluster Name: olympus
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: hdisk1
Cluster IP Address: 228.168.100.65
There are 4 node(s) and 2 network(s) defined
NODE xlvma1:
    Network admin_net
        xlvma1 192.168.100.65
    Network service_net
        artemis 10.10.10.68
        hermes 10.10.10.67
        diana 10.10.10.66
        merkury 10.10.10.65
        xlvma1_boot 192.168.150.65
NODE xlvma2:
    Network admin_net
        xlvma2 192.168.100.66
    Network service_net
        artemis 10.10.10.68
        hermes 10.10.10.67
        diana 10.10.10.66
        merkury 10.10.10.65
        xlvma2_boot 192.168.150.66
NODE xlvmb1:
    Network admin_net
        xlvmb1 192.168.100.67
    Network service_net
        artemis 10.10.10.68
        hermes 10.10.10.67
        diana 10.10.10.66
        merkury 10.10.10.65
        xlvmb1_boot 192.168.150.67
NODE xlvmb2:
    Network admin_net
        xlvmb2 192.168.100.68
    Network service_net
        artemis 10.10.10.68
        hermes 10.10.10.67
        diana 10.10.10.66
        merkury 10.10.10.65
        xlvmb2_boot 192.168.150.68
        ....

```

Figure 5-22 Cluster topology from the cltopinfo command

```

root@xlvma1:/usr/tmp>c1RGinfo -v
Cluster Name: olympus
Resource Group Name: drachma
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Fallback Policy: Never Fallback
Site Policy: Online On Either Site
Node Primary State Secondary State
-----
xlvma1@athens ONLINE OFFLINE
xlvma2@athens OFFLINE OFFLINE
xlvmb2@rome OFFLINE OFFLINE

Resource Group Name: denar
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Fallback Policy: Never Fallback
Site Policy: ignore
Node Group State
-----
xlvmb1 ONLINE
xlvmb2 OFFLINE

root@xlvma1:/usr/tmp>c1RGinfo -v drachma
Cluster Name: olympus
Resource Group Name: drachma
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Fallback Policy: Never Fallback
Site Policy: Online On Either Site
Node Primary State Secondary State
-----
xlvma1@athens ONLINE OFFLINE
xlvma2@athens OFFLINE OFFLINE
xlvmb2@rome OFFLINE OFFLINE

```

Figure 5-23 Resource group information from the c1RGinfo command

5.4 Test scenarios

With the cluster configuration completed, we now cover some test scenarios. We also provide expectations, results, and additional observations about the way of improving the high availability.

We focused on the site-specific related testing of the cluster and not individual components such as a single node, a single network adapter, or a network segment.

5.4.1 Test case 1: Both nodes down site failure

We begin with all cluster nodes active and in the STABLE state. The location and status of the resource groups are shown in Figure 5-24 on page 170.

We provoked a site failure by simultaneously halting both nodes, *xlvma1* and *xlvma2*, as shown in Figure 5-25. The expectation is that node *xlvmb2* takes over resource group *drachma*. After a short time, we noticed everything worked as expected. The new state of the cluster is shown in Figure 5-26 on page 171.

root@xlvma1:/usr/tmp>clRGinfo -r		
Group Name	Group State	Node
drachma	ONLINE	xlvma1@athens
	OFFLINE	xlvma2@athens
	OFFLINE	xlvmb2@rome
denar	OFFLINE	xlvmb1
	ONLINE	xlvmb2

Figure 5-24 Resource group status before site failure

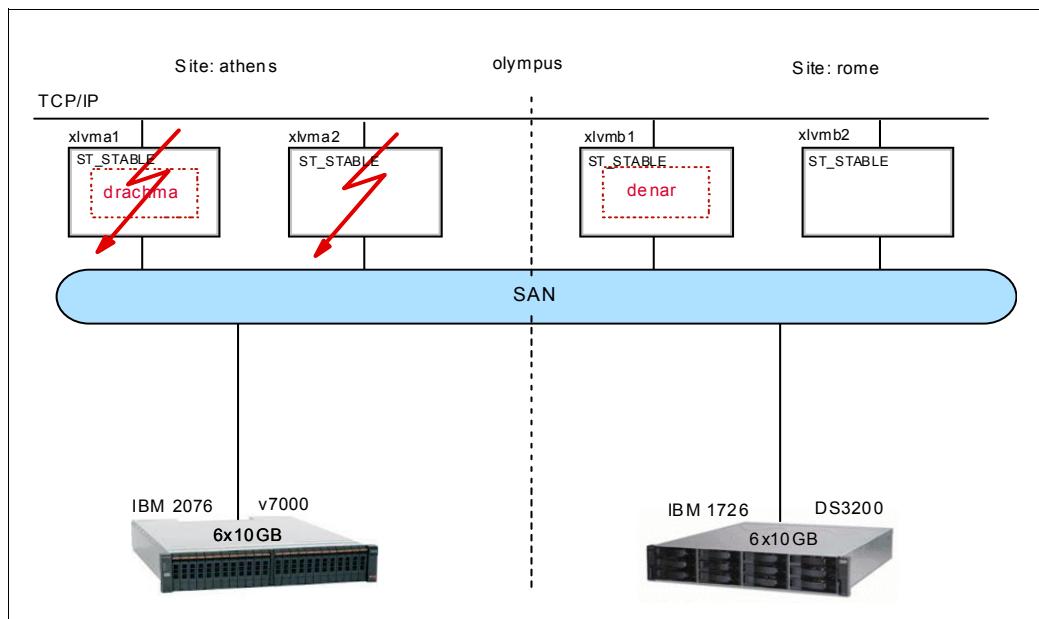


Figure 5-25 Site failure invoked

```

root@xlvma2:/>clRGinfo -v
Cluster Name: olympus
Resource Group Name: drachma
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Fallback Policy: Never Fallback
Site Policy: Online On Either Site
Node Primary State Secondary State
-----
xlvma1@athens OFFLINE OFFLINE
xlvma2@athens OFFLINE OFFLINE
xlvmb2@rome ONLINE OFFLINE

Resource Group Name: denar
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Fallback Policy: Never Fallback
Site Policy: ignore
Node Group State
-----
xlvmb1 ONLINE
xlvmb2 OFFLINE

```

Figure 5-26 Resource group state after site failure

5.4.2 Test case 2: Rolling node failures for site outage

We begin again with all cluster nodes and resource groups active, and the cluster in the STABLE state as shown in Figure 5-24 on page 170.

We provoked a site failover by rebooting the nodes *xlvma2* and *xlvma1* one at a time by executing `/usr/sbin/reboot -q`. We performed this on *xlmva2* first, then waited 3 to 4 seconds and submitted the same command on *xlvma1*, as shown in Figure 5-27.

We had the same expectation as in Test case 1. Namely, that node *xlvmb2* takes over resource group *drachma*, while *denar* stays unaffected on *xlvmb1*.

Indeed, after a few minutes we noticed the new state of the cluster shown in Figure 5-28 on page 172. To get the current cluster status we could use the command shown in Figure 5-27.

```

root@xlvma1:/>clcmd lssrc -ls clstrmgrES|egrep 'NODE |Current state'
NODE xlvmb2
Current state: ST_STABLE
NODE xlvmb1
Current state: ST_STABLE
NODE xlvma2
Current state: ST_INIT
NODE xlvma1
Current state: ST_INIT

```

Figure 5-27 Cluster status after site failure

5.4.3 Test case 3: Outage of a storage subsystem

In this test case, we wanted to determine how the outage of the entire mirror pools associated with one of the two sites is reflected in the cluster, as well as the impact on the resources active on node *xlvma1*. This initial state of the cluster resources is shown in Figure 5-28.

For this purpose, we generated an I/O-load in the volume group *drachma_VG* on node *xlvma1* by using the AIX command **/usr/bin/dd** in an endless loop:

```
while true
do
dd if=/dev/zero of=/drachma2/tst.1 bs=1k count=1024000
dd if=/dev/zero of=/drachma3/tst.1 bs=1k count=1024000
done
```

We also recorded the performance behavior of the LPAR using the tool **topasrec**, taking measured values every ten seconds.

This produces a large amount of disk writes to the file systems /drachma2 and /drachma3 and hence to all the corresponding logical volumes. Since these logical volumes are mirrored, having one copy on the remote site *rome*, within the other storage subsystem, IBM-1726 (DS3200) that will be unaffected by our generated outage, we expected that the applications running on *xlvma1* would continue to work *without disruption*.

We simulated the outage of the storage subsystem IBM-2076 (V7000) by instantly removing all the LUNs mapped to the LPAR *xlvma1*, as shown in Figure 5-29 on page 173.

```
root@xlvma1:/>c1RGinfo
```

Group Name	Group State	Node
drachma	ONLINE	xlvma1@athens
	OFFLINE	xlvma2@athens
	OFFLINE	xlvmb2@rome
denar	ONLINE	xlvmb1
	OFFLINE	xlvmb2

Figure 5-28 Resource group status prior to storage subsystem loss

The effects and the impact showed up immediately on *xlvma1*. The AIX error log is “flooded” with error messages, as shown in Figure 5-30 on page 173. The cluster services stay active and stable with no cluster events logged in the */var/hacmp/adm/cluster.log*. All the resource groups remain online on their respective nodes.

The only log entries showing that the Cluster Event Manager took notice of storage subsystem loss are found in the AIX error log */var/adm/ras/errlog*, shown in Figure 5-31 on page 174.

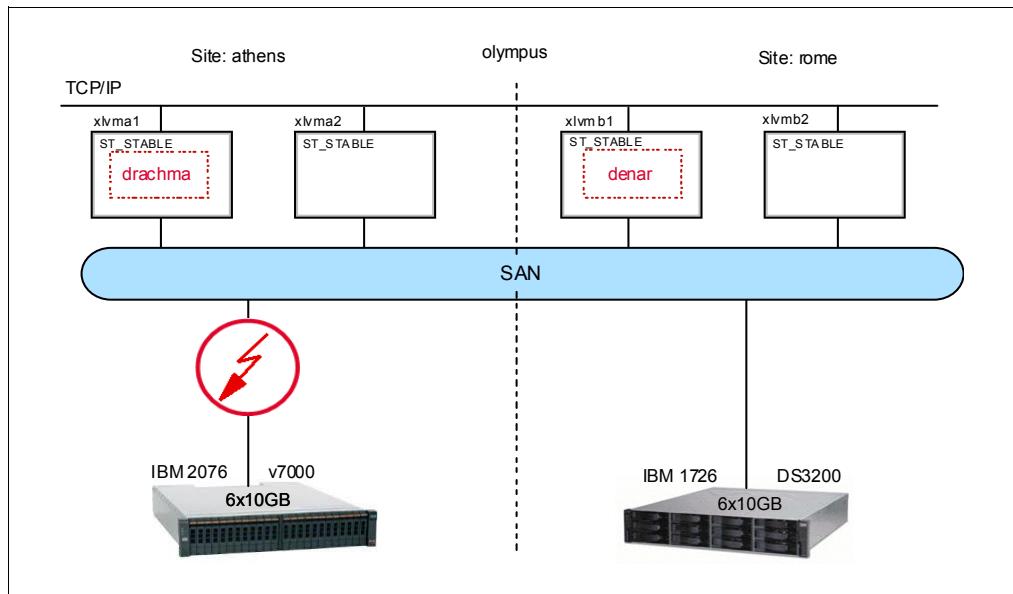


Figure 5-29 Simulating storage loss

```

...
B6267342 1122145512 P H hdisk7      DISK OPERATION ERROR
B6267342 1122145512 P H hdisk7      DISK OPERATION ERROR
DE3B8540 1122145512 P H hdisk9      PATH HAS FAILED
DE3B8540 1122145512 P H hdisk9      PATH HAS FAILED
DE3B8540 1122145512 P H hdisk9      PATH HAS FAILED
DE3B8540 1122145312 P H hdisk7      PATH HAS FAILED
DE3B8540 1122145312 P H hdisk7      PATH HAS FAILED
DE3B8540 1122145312 P H hdisk7      PATH HAS FAILED
...

```

Figure 5-30 Error report summary entries upon storage loss

t

```
...
-----
LABEL:      OPMG
IDENTIFIER: AA8AB241
Date/Time:   Thu Nov 22 14:57:16 2012
Sequence Number: 7737
Machine Id:  00F70C994C00
Node Id:    xlvma1
Class:      0
Type:       TEMP
WPAR:       Global
Resource Name: clevmgrd
Description
OPERATOR NOTIFICATION
User Causes
ERRLOGGER COMMAND
Recommended Actions
    REVIEW DETAILED DATA
Detail Data
MESSAGE FROM ERRLOGGER COMMAND
INFORMATION: Invoked process_rawdisk_event with PID 8126616 for selective failover. +++
-----
LABEL:      OPMG
IDENTIFIER: AA8AB241
Date/Time:   Thu Nov 22 14:57:16 2012
Sequence Number: 7736
Machine Id:  00F70C994C00
Node Id:    xlvma1
Class:      0
Type:       TEMP
WPAR:       Global
Resource Name: clevmgrd
Description
OPERATOR NOTIFICATION
User Causes
ERRLOGGER COMMAND
Recommended Actions
    REVIEW DETAILED DATA
Detail Data
MESSAGE FROM ERRLOGGER COMMAND
Error: Node 0xF300EBA341711E290C37A40C3E22102 has lost access to disk hdisk9. +++
-----
```

Figure 5-31 Error report details after storage loss

During this time, we observe a degradation of the I/O performance. It appears to have taken up to six minutes until the inaccessible volumes are declared *missing* as shown in Figure 5-32.

```
root@xlvma1:/>errpt|grep -i missing
F7DDA124  1122145912 U H LVDD          PHYSICAL VOLUME DECLARED MISSING
F7DDA124  1122145912 U H LVDD          PHYSICAL VOLUME DECLARED MISSING
F7DDA124  1122145712 U H LVDD          PHYSICAL VOLUME DECLARED MISSING
F7DDA124  1122145512 U H LVDD          PHYSICAL VOLUME DECLARED MISSING
F7DDA124  1119164112 U H LVDD          PHYSICAL VOLUME DECLARED MISSING
```

Figure 5-32 Physical volumes missing

After the physical volumes are declared missing, the I/O performance returns, meaning that the values are close to the values expected for unmirrored disk resources as shown in Figure 5-33 on page 175.

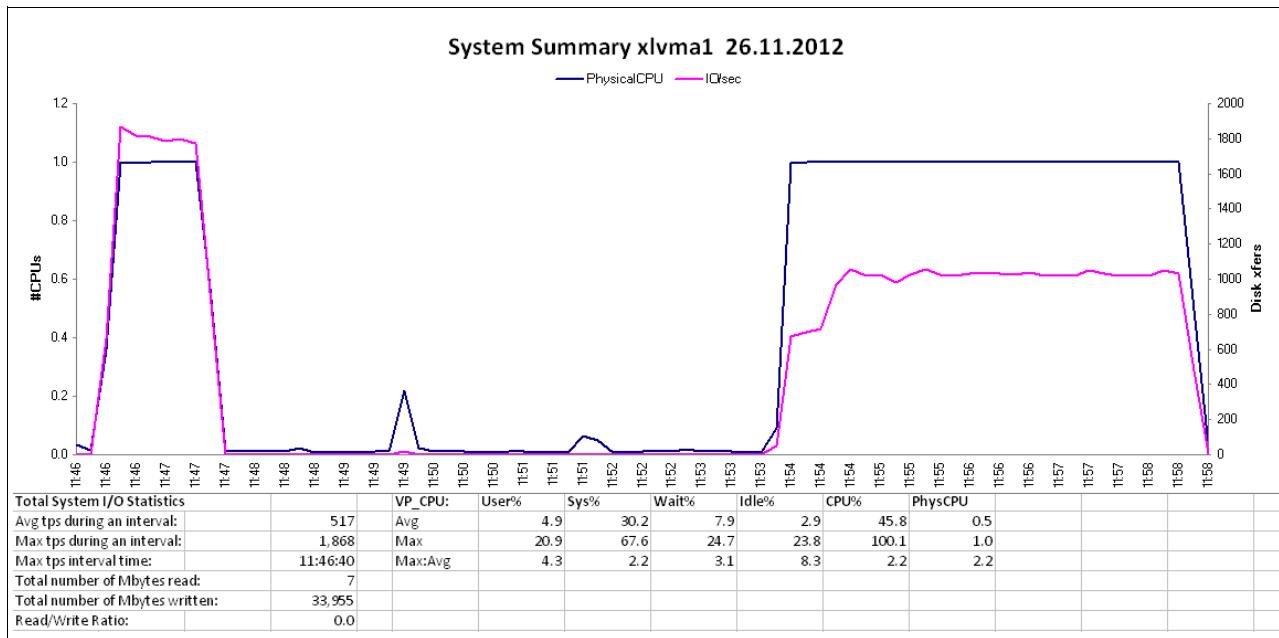


Figure 5-33 Disk I/O performance decay with storage subsystem failure

Gradually, all the physical partitions (PPs) on the missing hdisks are marked as STALE. This is also reflected on logical volumes themselves, as shown in Figure 5-34.

root@xlvm1:/>lsvg -l drachma_VG							
drachma_VG:							
LV NAME	TYPE	LPs	PPs	PVs	LV STATE	MOUNT POINT	
drachma_LV1	jfs2	20	40	2	open/stale	/drachma1	
drachma_LV2	jfs2	104	208	2	open/stale	/drachma2	
drachma_LV3	jfs2	124	248	2	open/stale	/drachma3	

Figure 5-34 Stale partitions

Though the resources were always available, there was a transition phase of about six minutes when the I/O-performance decreased considerably. This coincided with numerous LVDD write errors in the error report. This negatively impacts most applications because they may be unable to survive such an extended delay.

We retested this scenario with different values for the disk parameters. First we changed the MPIO parameter *algorithm* from *fail_over* (default value) to *round_robin*. No notable changes in the behavior could be seen.

In the tests above, six minutes were required to mark three LUNs as stale (two minutes per LUN).

In conclusion, when recovering from a disk storage subsystem failure, I/O to each LUN will stall for more than a minute. Please note that AIX can process failures of multiple LUNs simultaneously, but in the tests above, AIX recovered from simultaneous failure of three LUNs one at a time. To reduce the impact on applications and users of I/O stalls while AIX recovers from a disk storage subsystem failure:

- ▶ Minimize the number of LUNs. A few large LUNs will cause far fewer stalls than many small ones.

- If an anticipated workload is such that it tends to drive I/O to only one file or file system at a time (especially when I/O requests are sequential as was our testing), then to help AIX process failures of multiple LUNs simultaneously, stripe logical volumes (including those underlying file systems) across LUNs using the **-S** flag on the **mklv** command.

Note: It will not help to configure logical volumes with physical partition striping by specifying **-e x** on the **mklv** command. If physical partition striping is used with such a workload, AIX might still process a disk storage subsystem failure one LUN at a time.

5.4.4 Test case 4: Rolling disaster

The initial state of the cluster is similar to the ones in the previous test cases as shown in Figure 5-24 on page 170. However, there is an exception: cluster services are not running on node *xlvma2*!

During an I/O workload generated with the same procedure used in 5.4.3, “Test case 3: Outage of a storage subsystem” on page 172, we provoked a “Rolling disaster.”

In the first step, we cut the SAN link between site *athens* and site *rome*. This meant that the cluster nodes *xlvma1* and *xlvma2* were unable to access the storage resources from site *rome*. *xlvmb1* and *xlvmb2* will lose connection to their storage resources on site *athens*, as shown in Figure 5-35.

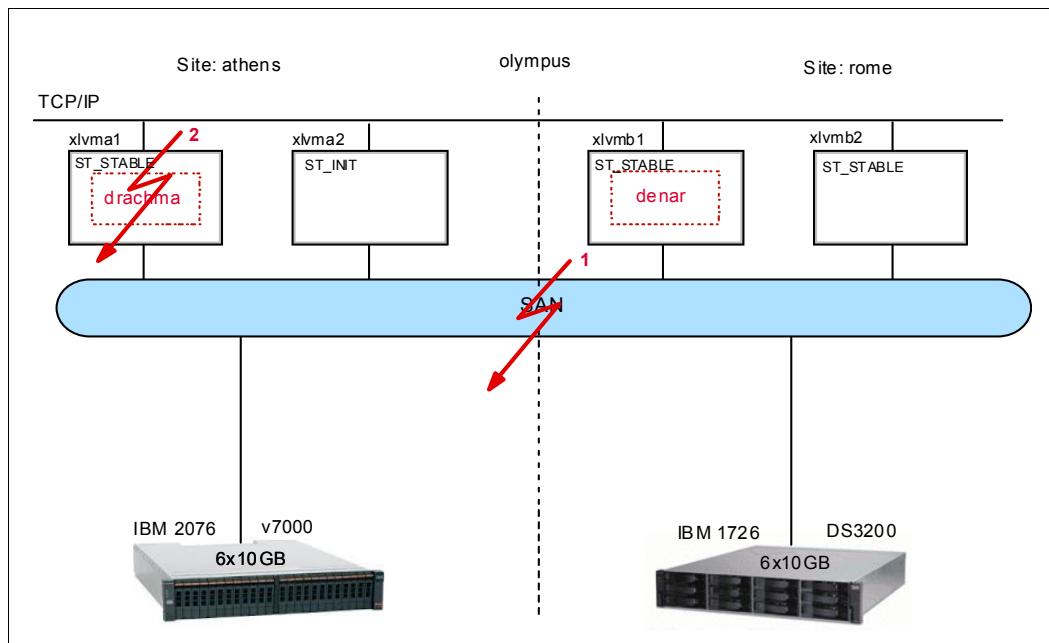


Figure 5-35 Rolling storage and site failure

During this time the data on site *rome* was not updated by applications running on site *athens*. This meant, from the application perspective, there was a considerable *data inconsistency* between the data copy on site *athens* and the one on site *rome*.

In the second step, we generated a site failure on site *athens* by crashing node *xlvma1*, also shown in Figure 5-35.

The last node remaining available for resource group *drachma*, which is *xlvmb2*, noticed the site failure and started a failover of the resource group, as shown in Figure 5-36 on page 177.

```

Nov 22 16:22:47 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT START: site_down athens
Nov 22 16:22:47 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT START: site_down_remote athens
Nov 22 16:22:47 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: site_down_remote athens 0
Nov 22 16:22:47 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: site_down athens 0
Nov 22 16:22:47 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT START: node_down xlvmal
Nov 22 16:22:47 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: node_down xlvmal 0
...
Nov 22 16:23:07 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: acquire_takeover_addr 0
...
Nov 22 16:23:08 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT START: start_server greek_apple
Nov 22 16:23:08 xlvmb2 user:notice PowerHA SystemMirror for AIX: Warning: syncvg can take considerable amount of time
, depending on data size and network bandwidth.
Nov 22 16:23:08 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: start_server greek_apple 0
Nov 22 16:23:09 xlvmb2 user:notice PowerHA SystemMirror for AIX: EVENT COMPLETED: rg_move_complete xlvmbl 2 0
Nov 22 16:23:09 xlvmb2 user:notice PowerHA SystemMirror for AIX: syncvg did not complete successfully
Nov 22 16:23:09 xlvmb2 user:notice PowerHA SystemMirror for AIX: Warning: syncvg can take considerable amount of time
, depending on data size and network bandwidth.
Nov 22 16:23:09 xlvmb2 user:notice PowerHA SystemMirror for AIX: syncvg did not complete successfully
Nov 22 16:23:09 xlvmb2 user:notice PowerHA SystemMirror for AIX: Warning: syncvg can take considerable amount of time
, depending on data size and network bandwidth.
Nov 22 16:23:10 xlvmb2 user:notice PowerHA SystemMirror for AIX: syncvg did not complete successfully

```

Figure 5-36 Unable to synchronize the volume group warning on failover

Although there were warnings logged in /var/hacmp/adm/cluster.log regarding the impossibility to synchronize the volume group *drachma_VG*, the resource group *drachma*, containing this volume group was started on xlvmb2, resulting in making the IP addresses and the applications available, as shown in Figure 5-37 on page 178. However, it was brought online using stale data, though the LVM on the remote side had no idea the data was stale.

```

root@xlvmb2:/>clRGinfo
-----
Group Name      Group State          Node
-----
drachma        OFFLINE             xlvma1@athens
                OFFLINE             xlvma2@athens
                ONLINE              xlvmb2@rome
denar          ONLINE              xlvmb1
                OFFLINE             xlvmb2

root@xlvmb2:/>netstat -i
Name  Mtu   Network     Address          Ipkts Ierrs    Opkts Oerrs    Coll
en0   1500  link#2    ee.af.3.c1.be.2  264860  0       43915  0       0
en0   1500  192.168.100 xlvmb2           264860  0       43915  0       0
en1   1500  link#3    ee.af.3.c1.be.3  237444  0       13166  0       0
en1   1500  10.10.10   artemis          237444  0       13166  0       0
en1   1500  192.168.150 xlvmb2_boot     237444  0       13166  0       0
lo0   16896 link#1           localhost      303686  0       303686  0       0
lo0   16896 127      localhost      303686  0       303686  0       0
lo0   16896 loopback        localhost      303686  0       303686  0       0
root@xlvmb2:/>lsvg -l drachma_VG
drachma_VG:
LV NAME      TYPE    LPs    PPs    PVs  LV STATE    MOUNT POINT
drachma_LV1 jfs2    20     40     2    open/stale   /drachma1
drachma_LV2 jfs2    104    208    2    open/stale   /drachma2
drachma_LV3 jfs2    124    248    2    open/stale   /drachma3

root@xlvmb2:/>mount | grep drachma
/dev/drachma_LV1 /drachma1          jfs2  Nov 22 16:23 rw,log=INLINE
/dev/drachma_LV2 /drachma2          jfs2  Nov 22 16:23 rw,log=INLINE
/dev/drachma_LV3 /drachma3          jfs2  Nov 22 16:23 rw,log=INLINE

```

Figure 5-37 Active volume group in the stale state after failover

When we removed the SAN connections between sites, xlvma1 started to experience I/O failures. As LVM attempted to update its stale partition information, it discovered that half the disks were missing, and marked them so. However, since quorum was off, the volume group stayed on line. **gsclvmd** on xlvma1 attempted to communicate the missing disks to its peer on the other node, xlvmb2.

At this point, we ran into a special interaction built into LVM **gsclvmd** as it was first developed. **gsclvmd** on xlvmb2 finds that it cannot see any of the disks that its peer can see, which meant that LVM had no way to unify VGDA/VGSA information across the cluster. This forced an LVM_SA_QUORCLOSE error and *drachma_VG* is brought offline on xlvmb2. This was by design. Unfortunately, it meant that xlvmb2 had absolutely no chance to become aware of the stale data on *drachma_VG*.

After a while, we forced a failure of xlvma1. The resource group *drachma* was moved to xlvmb2, and PowerHA attempted to bring the volume group *drachma_VG* online. Because force varyon was specified for the VG activation, and because there was a complete mirror set available (one physical partition for each logical partition), the volume group did activate. The application was started on old data.

We could not prevent the activation of *drachma_VG*, for example, with a pre-event script, because the information about the stale PP was not stored in any VGDA/VGSA accessible to

xlvmb2. Even when forcing an activation of *drachma_VG* in concurrent PASSIVE mode, both the **1svg** and **readvgda** commands could find no stale PP or PV.

In conclusion, the options currently available to PowerHA SystemMirror 7.1.2 Enterprise Edition and LVM will not prevent data inconsistencies in the case of a “Rolling disaster” as just described. However, as previously noted in this chapter, most solutions are designed to handle any *one* failure.

5.5 Maintaining cross-site LVM

It is just as important to properly maintain the configuration as it is to set it up properly from the start. In general, you should use C-SPOC to maintain the storage space allocation in your environment. However, it is very important to know exactly how to do it correctly because parts of C-SPOC are not completely cross-site aware.

The most common tasks to perform are:

- ▶ Creating a new volume group
- ▶ Adding volumes into an existing volume group
- ▶ Adding new logical volumes
- ▶ Adding additional space to an existing logical volume
- ▶ Adding a file system
- ▶ Increasing the size of a file system

Note: None of these procedures require a cluster synchronization because the resources technically have not changed. Also C-SPOC updates the information on the other cluster nodes. The only time a synchronization would be required is if adding or remove a volume group and an associated resource group change was needed.

5.5.1 Test environment overview

We have two nodes, *jessica* and *cassidy*. One assigned to each of two sites, *Dallas* and *FortWorth*, respectively. There are four shared disks total, two from each site and storage subsystem. Initially only two disks, hdisk4 and hdisk6, are part of the cross site LVM mirrored volume group *xsitevg*. All four of our disk, hdisk4-hdisk7, definitions match across the two systems. There are two mirror pools defined, *dallasmp* and *fortworthmp*. Dallasmp currently only contains hdisk4, and fortworthmp only contains hdisk6. An overview of cluster configuration is shown in Example 5-1. Also a diagram of our starting test configuration is shown in Figure 5-38 on page 181.

Example 5-1 Test cluster initial configuration

```
[jessica:root] / # cltopinfo |pg
Cluster Name: xsitelvmcluster
Cluster Type: Stretched
Heartbeat Type: Unicast
Repository Disk: hdisk1 (00f6f5d015a4310b)
Cluster Nodes:
    Site 1 (Dallas):
        jessica
    Site 2 (FortWorth):
        cassidy
```

There are 2 node(s) and 2 network(s) defined

```

NODE cassidy:
    Network net_ether_01
        cassidy_xd      192.168.150.52
    Network net_ether_010
        cassidy 192.168.100.52

NODE jessica:
    Network net_ether_01
        jessica_xd      192.168.150.51
    Network net_ether_010
        jessica 192.168.100.51
[jessica:root] / # clshowres
Resource Group Name           xsitelvmRG
Participating Node Name(s)    jessica cassidy
Startup Policy                 Online On Home Node Only
Failover Policy                Fallover To Next Priority Node
In The List
Fallback Policy               Never Fallback
Site Relationship              ignore
Dynamic Node Priority
Service IP Label               dallasserv ftwserv
Volume Groups                  xsitevg
Use forced varyon for volume groups, if necessary   true
Application Servers           dummyapp

[jessica:root] / # lsmmp -A xsitevg
VOLUME GROUP:     xsitevg          Mirror Pool Super Strict: yes

MIRROR POOL:      dallasmp         Mirroring Mode:           SYNC
MIRROR POOL:      fortworthmp      Mirroring Mode:           SYNC

[jessica:root] / # lsvg -P xsitevg
Physical Volume  Mirror Pool
hdisk4            dallasmp
hdisk6            fortworthmp

```

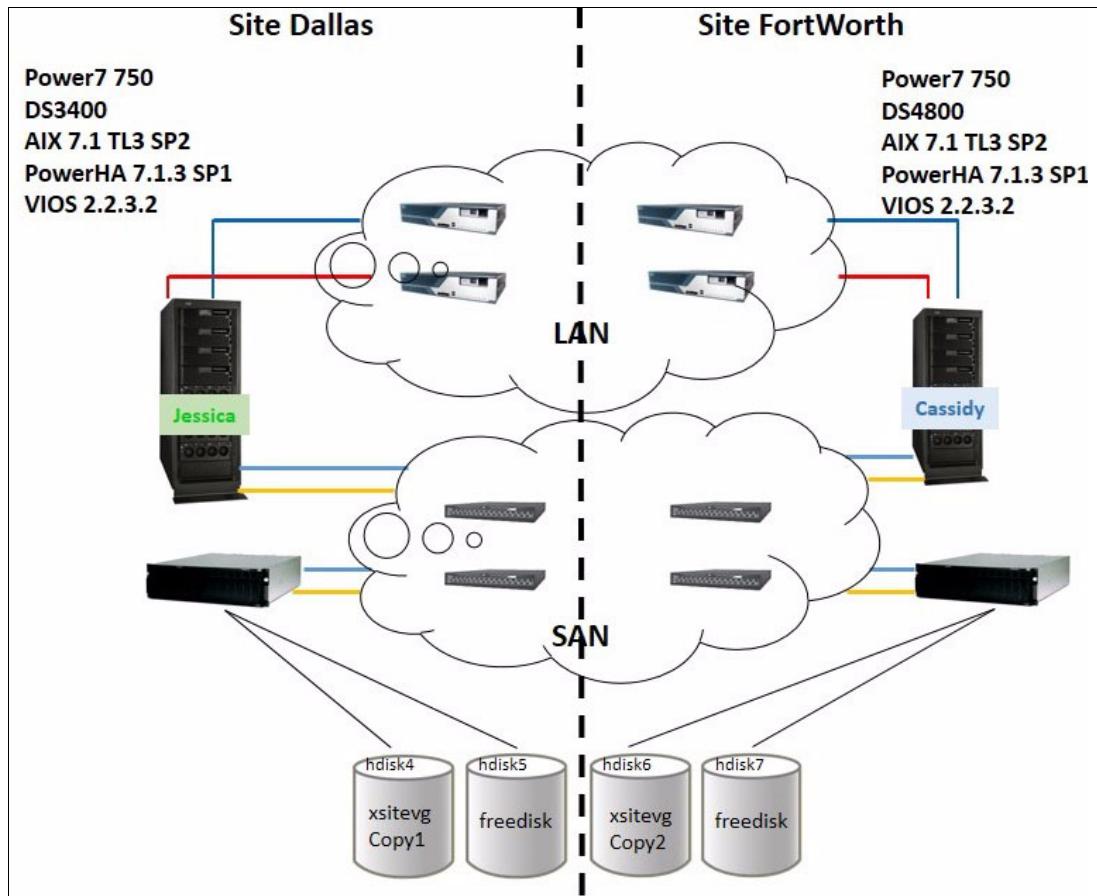


Figure 5-38 Cross-site LVM test environment overview

5.5.2 Storage administration scenarios

This section provides storage administration details.

Creating a new volume group

When creating a new volume group, refer to the initial configuration steps in “Create volume group(s)” on page 161.

Adding volumes into an existing volume group

When adding disks into an existing volume group, it is important to add them in pairs, one for each site. Then also choose to add the disks into their respective mirror pools. Also check that the PVID of the disks are known to each system.

In our scenario, we will add one disk at a time because we can also add them to their respective mirror pool at the same time.

To add volumes/disks into an existing volume group, run:

- ▶ **smitty cl_vgsc** → **Add a Volume to a Volume Group**, press Enter, select the appropriate resource group and volume group and press Enter.

You will then be presented with a list of volumes to choose from. Remember to choose one from each site as shown in Figure 5-39 on page 182.

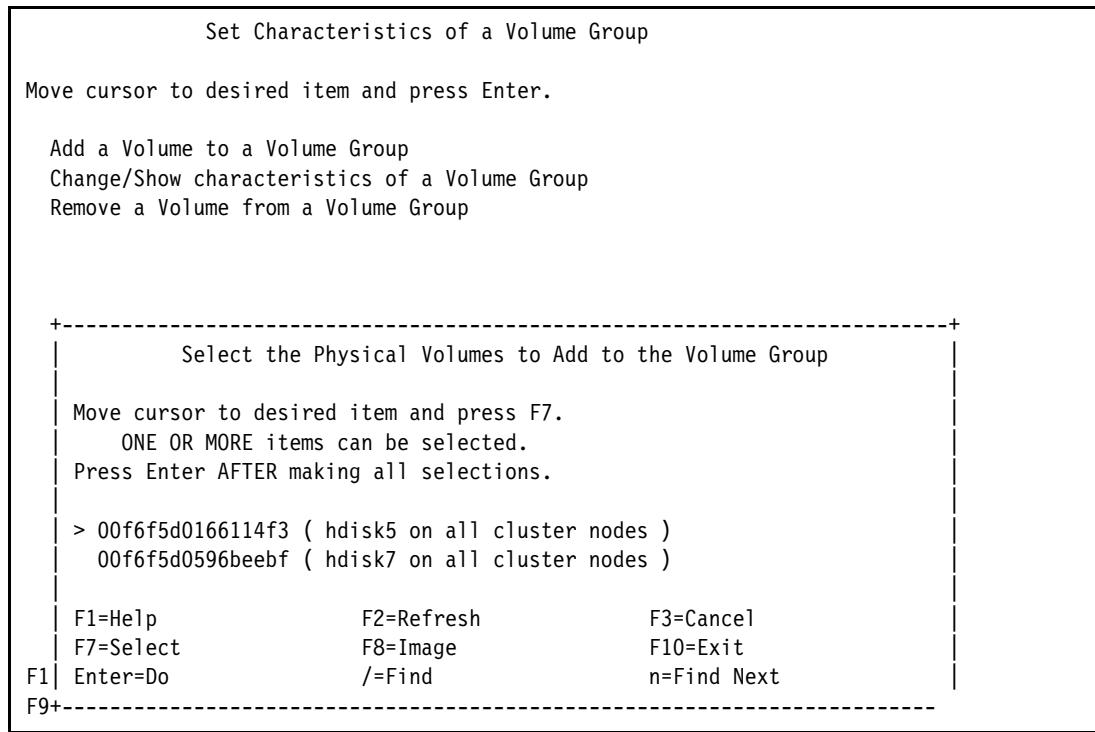


Figure 5-39 Choose disk(s) to add to volume group

Upon choosing the appropriate disk, verify the fields in the final menu as shown in Figure 5-40 and press Enter to complete the addition.

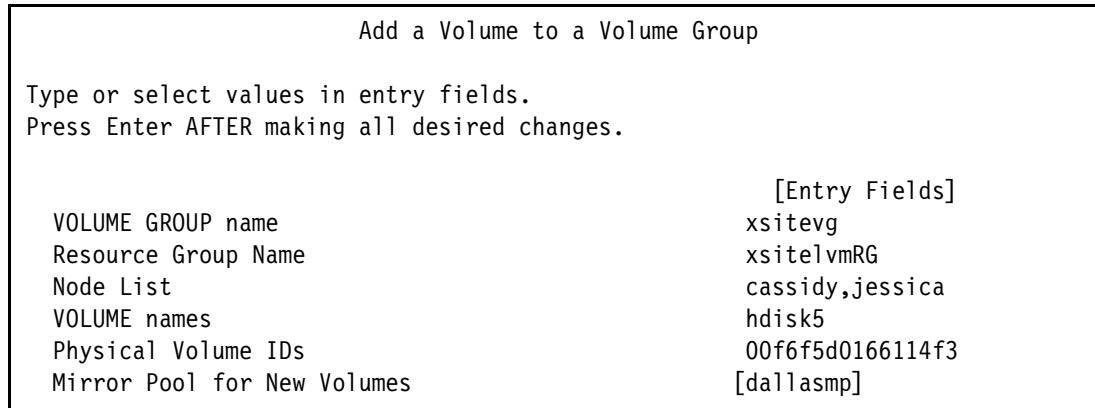


Figure 5-40 Add a volume to volume group specifying mirror pool

In our scenario, we repeat this process choosing hdisk7 and the appropriate mirror pool of fortworthmp. Once completed our volume group and mirror pool definitions are as shown in Example 5-2.

Example 5-2 New volume group and mirror pool definitions

[jessica:root] / # lspv grep xsitevg			
hdisk4	00f6f5d0166106fa	xsitevg	concurrent
hdisk5	00f6f5d0166114f3	xsitevg	concurrent
hdisk6	00f6f5d029906df4	xsitevg	concurrent
hdisk7	00f6f5d0596beebf	xsitevg	concurrent

```
[jessica:root] / # lsvg -P xsitevg
Physical Volume   Mirror Pool
hdisk4           dallasmp
hdisk6           fortworthmp
hdisk5           dallasmp
hdisk7           fortworthmp
```

Adding new logical volumes

When adding a new logical volume, it is important to create the logical volume copies on separate disks, and those disks are specific to each site. It is also crucial to set the allocation policy to *superstrict* as shown in Figure 5-42 on page 184.

To add a new logical volume, run:

- ▶ **smitty cl_lv** → **Add a Logical Volume**, press Enter, select the appropriate resource group and volume group, and press Enter.

Then choose the appropriate disks, one from each site, using **F7** as shown in Figure 5-41.

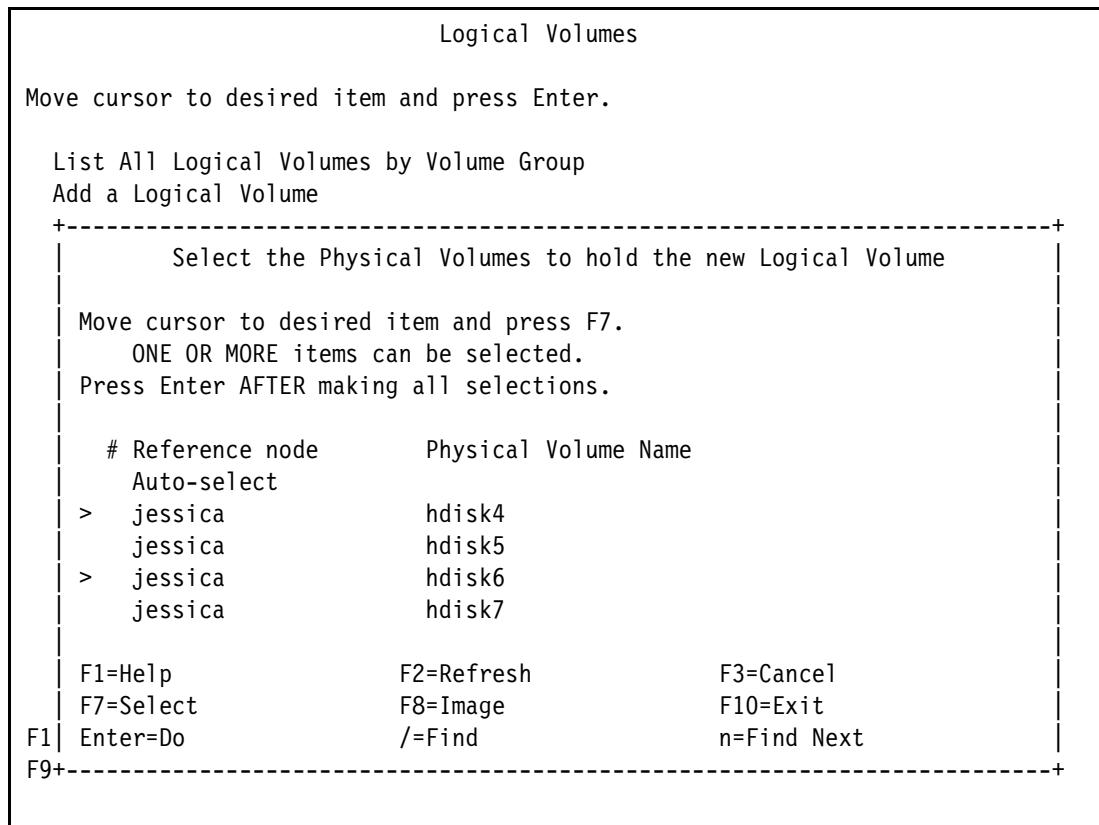


Figure 5-41 Select physical volumes to add logical volumes

Important: Do *not* use the Auto-select option.

Upon pressing Enter, you will be presented with the final menu to create the new shared logical volume see Figure 5-42. Choose the appropriate unique name, type, and size of the logical volume. Keep the *RANGE of physical volumes* set to minimum. Also specify two copies and set the *Allocate each logical partition copy on a SEPARATE physical volume?* to superstrict. Although the disks are already members of existing mirror pools, it is recommended to specify the mirror pools for each copy. Repeat as needed for each logical volume that needs to be created.

Add a Logical Volume		
Type or select values in entry fields. Press Enter AFTER making all desired changes.		
[TOP]	[Entry Fields]	
Resource Group Name	xsitevmRG	
VOLUME GROUP name	xsitevg	
Node List	cassidy,jessica	
Reference node	jessica	
* Number of LOGICAL PARTITIONS	[25] #	
PHYSICAL VOLUME names	hdisk4 hdisk6	
Logical volume NAME	[xsitev1]	
Logical volume TYPE	[jfs2] +	
POSITION on physical volume	outer_middle	+ (highlighted)
RANGE of physical volumes	minimum	+ (highlighted)
MAXIMUM NUMBER of PHYSICAL VOLUMES to use for allocation	[] #	
Number of COPIES of each logical	2 +	
Mirror Write Consistency?	active +	
Allocate each logical partition copy on a SEPARATE physical volume?	superstrict +	
RELOCATE the logical volume during reorganization?	yes +	
Logical volume LABEL	[]	
MAXIMUM NUMBER of LOGICAL PARTITIONS	[512] #	
Enable BAD BLOCK relocation?	yes +	
SCHEDULING POLICY for writing logical partition copies	parallel +	
Enable WRITE VERIFY?	no +	
File containing ALLOCATION MAP	[] /	
Stripe Size?	[Not Striped] +	
Serialize I/O?	no +	
Make first block available for applications?	no +	
Mirror Pool for First Copy	dallasmp +	
Mirror Pool for Second Copy	fortworthmp +	
Mirror Pool for Third Copy	+ (highlighted)	
User ID	+ (highlighted)	
Group ID	+ (highlighted)	
Permissions	[] X	

Figure 5-42 Add a shared logical volume with superstrict allocation policy

You can verify that the copies have been created correctly by viewing the mapping of the partitions for each logical volume. Use the `lslv -m lvgname` command as shown in Figure 5-43.

```
[jessica:root] / # lslv -m xsitelv1
xsite1v1:N/A
LP   PP1  PV1          PP2  PV2          PP3  PV3
0001 0193 hdisk4      0193 hdisk6
0002 0194 hdisk4      0194 hdisk6
0003 0195 hdisk4      0195 hdisk6
0004 0196 hdisk4      0196 hdisk6
0005 0197 hdisk4      0197 hdisk6
0006 0198 hdisk4      0198 hdisk6
0007 0199 hdisk4      0199 hdisk6
0008 0200 hdisk4      0200 hdisk6
0009 0201 hdisk4      0201 hdisk6
0010 0202 hdisk4      0202 hdisk6
0011 0203 hdisk4      0203 hdisk6
0012 0204 hdisk4      0204 hdisk6
0013 0205 hdisk4      0205 hdisk6
0014 0206 hdisk4      0206 hdisk6
0015 0207 hdisk4      0207 hdisk6
0016 0208 hdisk4      0208 hdisk6
0017 0209 hdisk4      0209 hdisk6
0018 0210 hdisk4      0210 hdisk6
0019 0211 hdisk4      0211 hdisk6
0020 0212 hdisk4      0212 hdisk6
0021 0213 hdisk4      0213 hdisk6
0022 0214 hdisk4      0214 hdisk6
0023 0215 hdisk4      0215 hdisk6
0024 0216 hdisk4      0216 hdisk6
0025 0217 hdisk4      0217 hdisk6
```

Figure 5-43 Mirrored logical volume partition mapping

Adding additional space to an existing logical volume

Similar to creating a logical volume, it is important to allocate additional space properly to maintain the mirrored copies at each site. To add more space, run:

1. **smitty c1_lvsc** → **Increase the Size of a Shared Logical Volume** → and press Enter.
2. Choose the appropriate volume group and resource group from the pop-up list as shown in Figure 5-44 on page 186. Then choose the proper logical volume the pop-up list displayed as shown in Figure 5-45 on page 186. You will then be presented with a list of disks that belong to the same volume group as the logical volume previously chosen. The list will be very similar to the one seen when creating a new logical volume as shown in Figure 5-41 on page 183. Choose the appropriate disks with **F7** and press Enter.

Important: Do *not* use the Auto-select option.

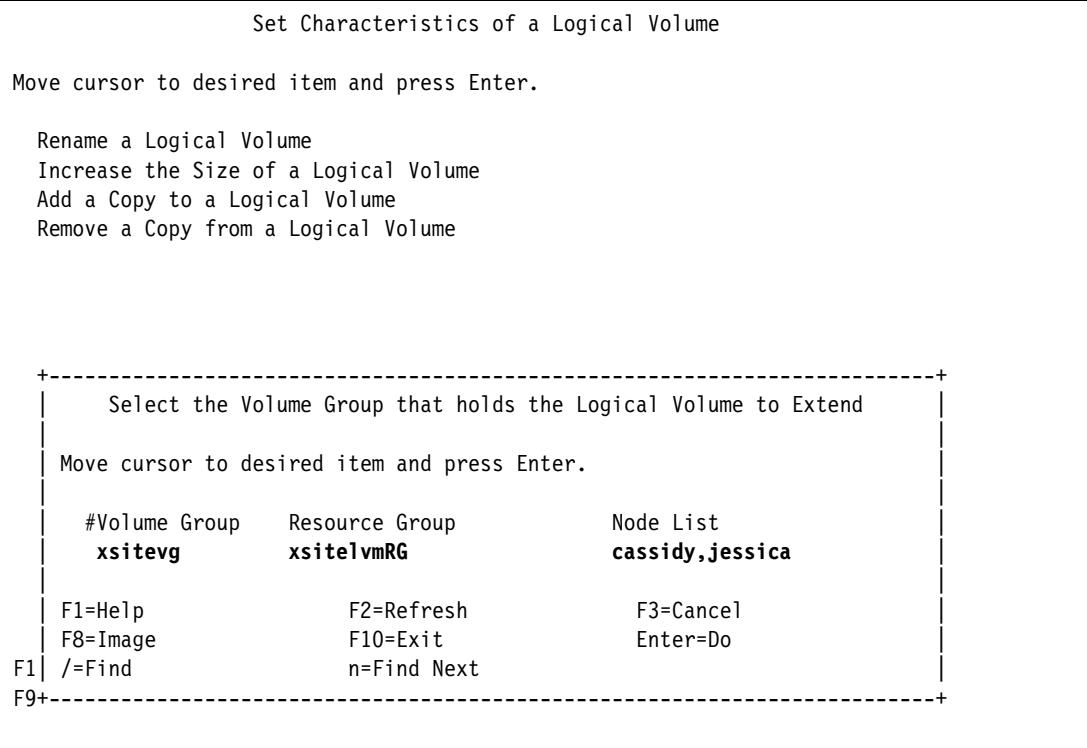


Figure 5-44 Shared volume group pop-up list

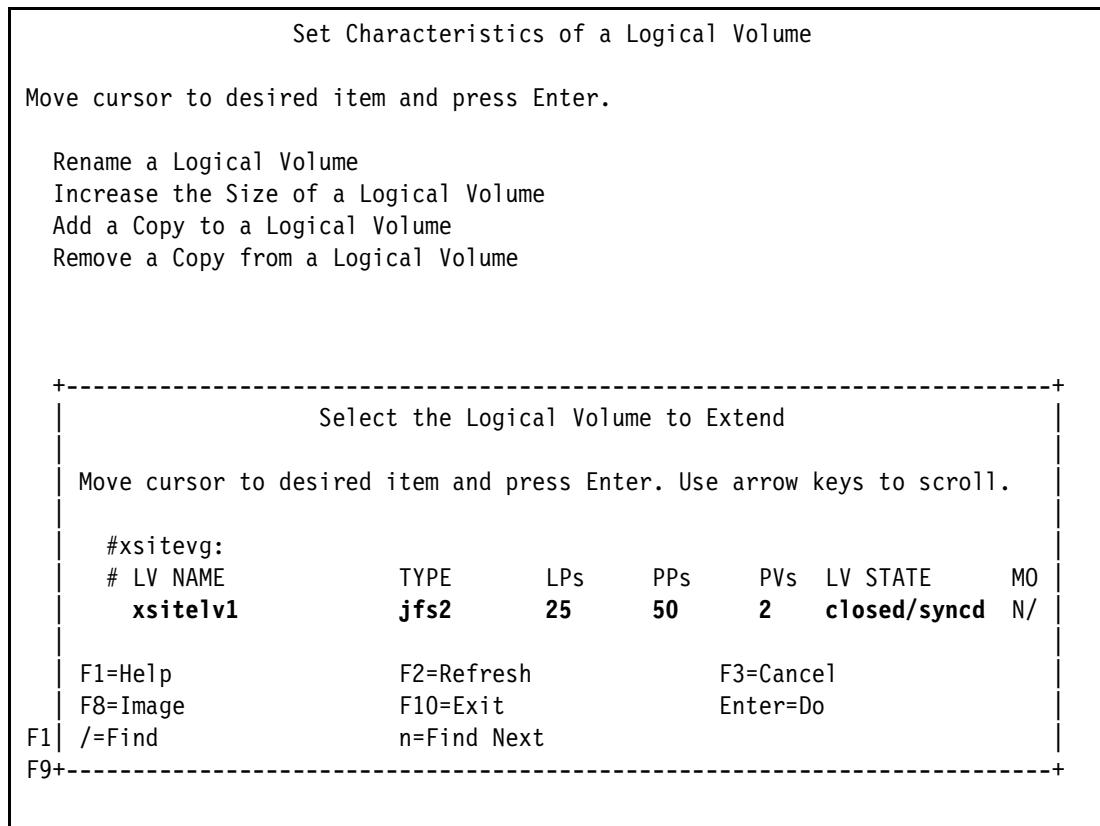


Figure 5-45 Logical volume pop-up list selection

3. After selecting the target disks, you will be presented with the final *Increase the Size of a Shared Logical Volume* menu as shown in Figure 5-46.

Be sure to keep the *RANGE of physical volumes* set to *minimum* and set the *Allocate each logical partition copy on a SEPARATE physical volume?* to *superstrict* as it should already be set properly, if the logical volume was originally created correctly.

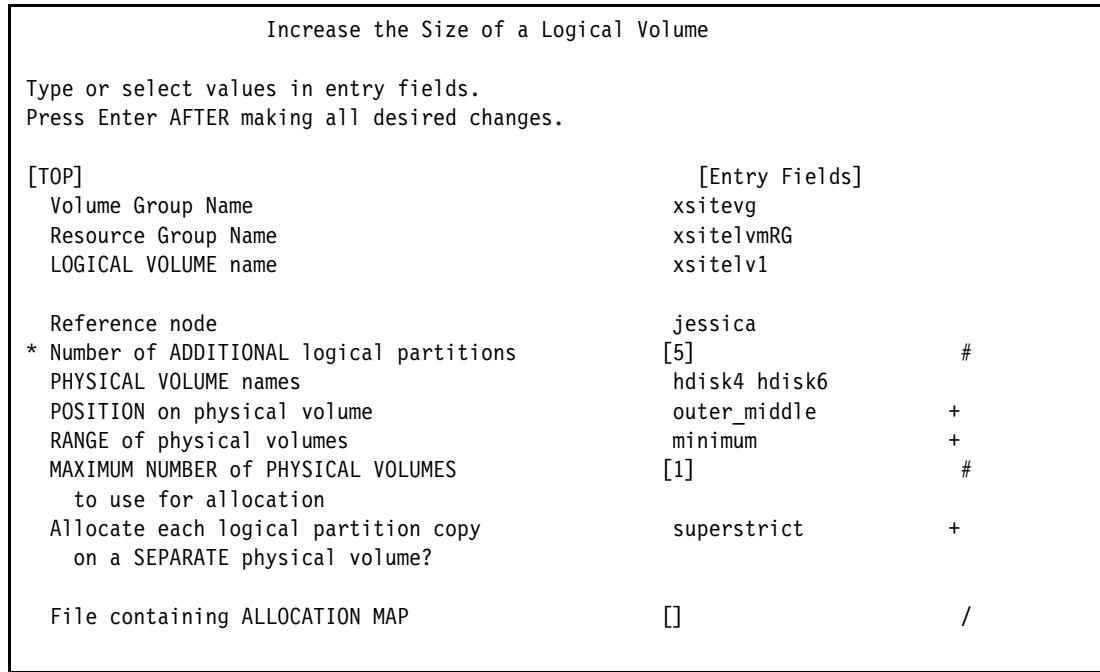


Figure 5-46 Increase the size of a shared logical volume

4. After adding additional space, verify that the partition mapping is correct by running the **lslv -m lvname** command again. Shown in bold are the five new partitions just added to the logical volume.

```
[jessica:root] / # lslv -m xsitev1
xsitev1:N/A
LP   PP1  PV1          PP2  PV2          PP3  PV3
0001 0193 hdisk4      0193 hdisk6
0002 0194 hdisk4      0194 hdisk6
0003 0195 hdisk4      0195 hdisk6
0004 0196 hdisk4      0196 hdisk6
0005 0197 hdisk4      0197 hdisk6
0006 0198 hdisk4      0198 hdisk6
0007 0199 hdisk4      0199 hdisk6
0008 0200 hdisk4      0200 hdisk6
0009 0201 hdisk4      0201 hdisk6
0010 0202 hdisk4      0202 hdisk6
0011 0203 hdisk4      0203 hdisk6
0012 0204 hdisk4      0204 hdisk6
0013 0205 hdisk4      0205 hdisk6
0014 0206 hdisk4      0206 hdisk6
0015 0207 hdisk4      0207 hdisk6
0016 0208 hdisk4      0208 hdisk6
0017 0209 hdisk4      0209 hdisk6
0018 0210 hdisk4      0210 hdisk6
0019 0211 hdisk4      0211 hdisk6
```

0020	0212 hdisk4	0212 hdisk6
0021	0213 hdisk4	0213 hdisk6
0022	0214 hdisk4	0214 hdisk6
0023	0215 hdisk4	0215 hdisk6
0024	0216 hdisk4	0216 hdisk6
0025	0217 hdisk4	0217 hdisk6
0026	0218 hdisk4	0218 hdisk6
0027	0219 hdisk4	0219 hdisk6
0028	0220 hdisk4	0220 hdisk6
0029	0221 hdisk4	0221 hdisk6
0030	0222 hdisk4	0222 hdisk6

Adding a file system

C-SPOC can be used to add a file system in a cross-site LVM mirrored configuration. The key here is to *always* create the logical volume first to ensure proper mirroring. Then add the file system on the previously defined logical volume. Also make sure an existing JFSlog is in place or use inline logs. Though C-SPOC can create a JFSlog for you, it may not have a preferred unique name.

Important: Always add a file system by creating the logical volume first, then create the file system on a previously defined logical volume. The reason for this is: If you allow the creation of the logical volume at the time of file system creation, the logical volume mirroring *may* not be created properly.

To add a new JFS2 file system, run:

1. **smitty c1_fs → Add a File System**
2. Choose the volume group from the pick list and press Enter.
3. Choose the file system type from the pick list and press Enter.

4. Then choose the appropriate logical volume from the picklist as shown in Figure 5-47.

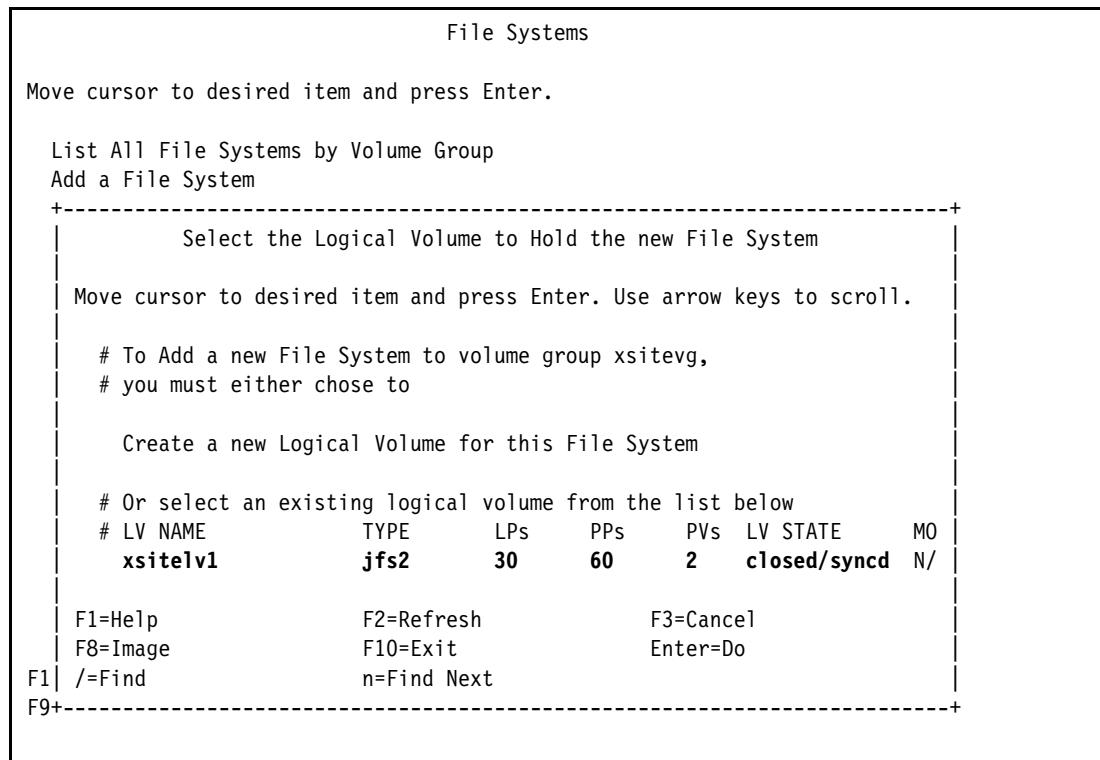


Figure 5-47 Add a file system to a previously defined cross-site logical volume

5. Then proceed to fill out the final menu fields appropriately as needed, as shown
 Example 5-3. Press Enter twice to complete the creation.

Example 5-3 Final Add a JFS2 SMIT menu

Add an Enhanced Journaled File System on a Previously Defined Logical Volume

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

[TOP]	[Entry Fields]
Resource Group	xsitevmRG
* Node Names	cassidy,jessica
Logical Volume name	xsitev1
Volume Group	xsitevg
* MOUNT POINT	[/xsitefs] /
PERMISSIONS	read/write +
Mount OPTIONS	[] +
Block Size (bytes)	4096 +
Inline Log?	no +
Inline Log size (MBytes)	[] #
Logical Volume for Log	[] +
Extended Attribute Format	Version 1 +
ENABLE Quota Management?	no +
Enable EFS?	no +
Esc+1=Help	Esc+2=Refresh
Esc+5=Reset	F6=Command
	Esc+3=Cancel
	F7>Edit
	Esc+4=List
	F8=Image

6. Repeat this step as needed for each logical volume previously created that requires a file system.

Once created, the file system will automatically be mounted on the node that is currently hosting the resource group that contains the volume group the file system belongs to.

Increasing the size of a file system

C-SPOC can be used to increase the size of a file system in a cross-site LVM mirrored configuration. The key here is to *always* increase the size of the logical volume first to ensure proper mirroring. See “Adding additional space to an existing logical volume” on page 185 for more information. Then increase the size of the file system on the previously defined logical volume.

Important: Always add more space to a file system, by adding more space to the logical volume first. Never add the additional space to the JFS first when using cross-site LVM mirroring as it *may not* maintain the mirroring properly.

To add additional space to a JFS2 file system, run:

1. **smitty c1_fs → Change / Show Characteristics of a File System.**
2. Then choose the appropriate file system, volume group, and resource group.
3. Then fill in the rest of the fields appropriately as shown in Figure 5-48 on page 191.

Change/Show Characteristics of a Enhanced Journaled File System

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

[TOP]	[Entry Fields]		
Volume group name	xsitevg		
Resource Group Name	xsiteLvmRG		
* Node Names	cassidy,jessica		
* File system name	/xsitesefs		
NEW mount point	[/xsitesefs]		
SIZE of file system			/
Unit Size	512bytes	+	
Number of Units	[2097152]	#	
Mount GROUP	[]		
Mount AUTOMATICALLY at system restart?	no	+	
PERMISSIONS	read/write	+	
Mount OPTIONS	[]		
Start Disk Accounting?	no	+	
Block Size (bytes)	4096		
Inline Log?	no		
Inline Log size (MBytes)	[0]	#	
Extended Attribute Format	[v1]		
ENABLE Quota Management?	no	+	
Allow Small Inode Extents?	[yes]	+	
Logical Volume for Log	xsiteLoglv	+	
Encrypted File System	no		
Esc+1=Help	Esc+2=Refresh	Esc+3=Cancel	Esc+4=List
Esc+5=Reset	F6=Command	F7>Edit	F8=Image
F9=Shell	F10=Exit	Enter=Do	

Figure 5-48 Increase the size of Shared Enhanced Journaled File System

4. Ensure that the size of the file system matches the size of the logical volume. If you are unsure, you can use the **lsfs -q mountpoint** command as shown in the following example:

```
# lsfs -q /xsitesefs
Name           Nodename   Mount Pt          VFS  Size    Options     Auto
Accounting
/dev/xsiteLvm1 --         /xsitesefs      jfs2  1966080  rw      no  no
(1v size: 2097152, fs size: 1966080, block size: 4096, sparse files: yes, inline
log: no, inline log size: 0, EAformat: v1, Quota: no, DMAPI: no, VIX: yes, EFS:
no, ISNAPSHOT: no, MAXEXT: 0, MountGuard: yes)
```


Extended disaster recovery (linked clusters)

This part provides information about disaster recovery techniques with linked clusters.

This part contains the following chapters:

- ▶ Chapter 6, “Configuring PowerHA SystemMirror Enterprise Edition linked cluster with SVC replication” on page 195.
- ▶ Chapter 7, “Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replication” on page 253.



Configuring PowerHA SystemMirror Enterprise Edition linked cluster with SVC replication

The main difference between PowerHA SystemMirror Enterprise Edition version 6.1 and 7.1.2 is the way that the cluster behaves when having a repository disk at each site. This is called a *linked* cluster. For more information about linked clusters, see 2.2.1, “Stretched and linked clusters” on page 19.

This chapter shows how to configure an SVC linked cluster along with how to test PowerHA SystemMirror Enterprise Edition functionality.

This chapter covers the following topics:

- ▶ Overview of SVC management
- ▶ Planning and prerequisites overview
- ▶ Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with SVC remote copy
- ▶ Storage management with SVC replicated resources
- ▶ Testing the resource group move
- ▶ Testing storage failure on SVC and V7000 mixed environment

6.1 Overview of SVC management

The IBM SAN Volume Controller (SVC) is a virtualization appliance solution that maps virtualized volumes, visible to hosts and applications, to the physical volumes on storage devices. The SVC is a SAN block aggregation appliance that is designed for attachment to a variety of host computer systems.

There are three major approaches in use today for the implementation of block-level aggregation:

- ▶ Network-based: Appliance

The device is a SAN appliance that sits in the data path, and all I/O flows through the device. This kind of implementation is also referred to as *symmetric virtualization* or *in-band*. The device is both a target and an initiator. It is the target of I/O requests from the host perspective and the initiator of I/O requests from the storage perspective. The redirection is performed by issuing new I/O requests to the storage.

- ▶ Switch-based: Split-path

The device is usually an intelligent SAN switch that intercepts I/O requests on the fabric and redirects the frames to the correct storage location. The actual I/O requests are themselves redirected. This kind of implementation is also referred to as *asymmetric virtualization* or *out-of-band*. Data and the control data path are separated, and a specific (preferably highly available and disaster tolerant) controller outside of the switch holds the metainformation and the configuration to manage the split data paths.

- ▶ Controller-based

The device is a storage controller that provides an internal switch for external storage attachment. In this approach, the storage controller intercepts and redirects I/O requests to the external storage as it does for internal storage.

It is an in-band implementation that minimizes dependency on unique hardware and software by copying the storage functions expected in a SAN environment from the storage subsystems and managing storage resources. In addition, it provides advanced copy services for data migration and business continuity similar to the PPRC function in ESS and DS8000 storage subsystems. Since the copy services operate on the virtual volumes, simpler replication configurations can be created using the SAN Volume Controller.

Figure 6-1 on page 197 shows a conceptual diagram of numerous storage systems attached to an SVC. It shows a pair of hosts that are connected to a SAN fabric. In implementations that have high availability requirements, the SAN fabric cloud represents a redundant SAN. A redundant SAN is composed of a fault-tolerant system of two or more counterpart SANs, providing alternate paths for each SAN-attached device.

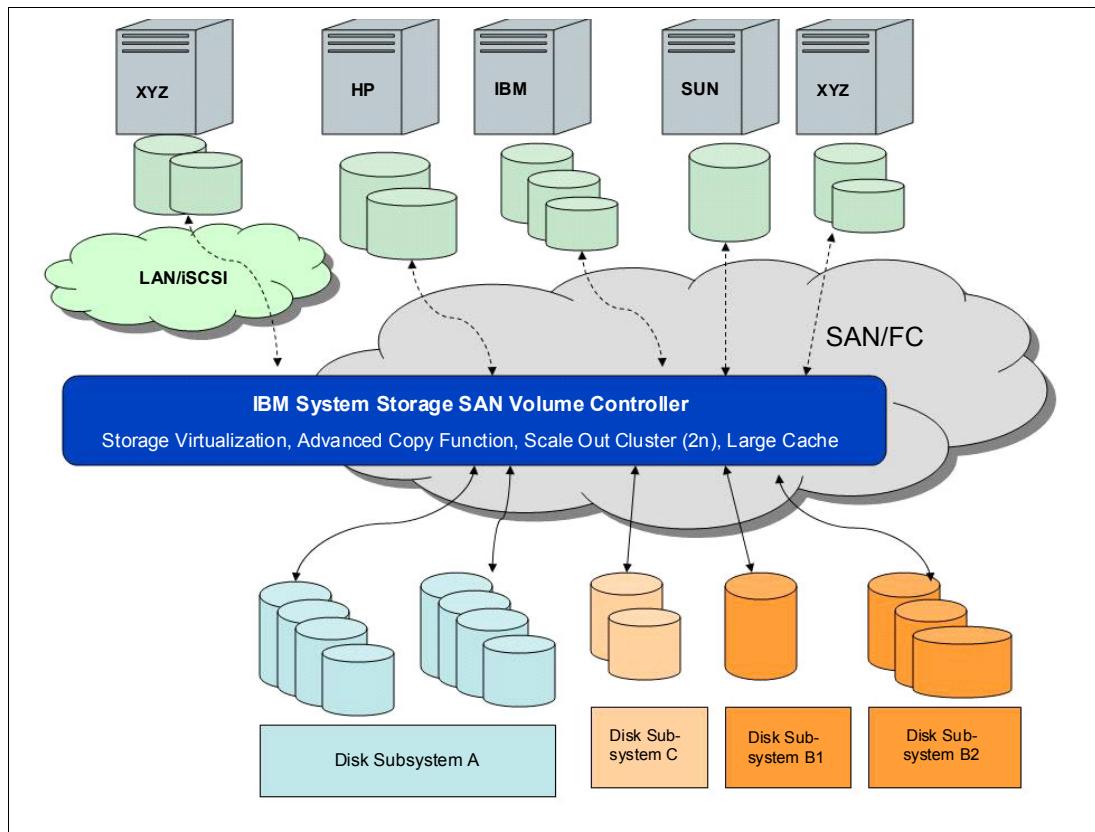


Figure 6-1 SVC conceptual overview

A cluster of SVC nodes is connected to the same fabric and presents VDisks to the hosts. These VDisks are created from MDisks that are presented by the RAID controllers. Two zones are shown in the fabric:

- ▶ A host zone, in which the hosts can see and address the SVC nodes.
- ▶ A storage zone, where SVC nodes can see and address the MDisks and logical unit numbers (LUNs) presented by RAID controllers. All the data transfer happens through the SVC nodes.

PowerHA SystemMirror Enterprise Edition for Metro Mirror with SVC Management enhances the Peer-to-Peer Remote Copy (PPRCs) ability to provide a fully automated, highly available disaster recovery (HADR) management solution by taking advantage of SVC's ability to provide virtual disks derived from varied disk subsystems. PowerHA SystemMirror's interface is designed so that once the basic SVC environment is configured, SVC-PPRC relationships are automatically created; no additional access is needed.

The integration of PowerHA SystemMirror and SVC-PPRC provides:

- ▶ PowerHA SystemMirror management of SVC PPRC for automatic failover and reintegration of SVC PPRC-protected virtual disks (VDisks) between sites.
- ▶ Support for user-defined policy-based resource groups.
- ▶ Support for the following inter-site management policies for resource groups:
 - Prefer primary site
 - Online on either site
- ▶ Support for Subsystem Device Drivers (SDD).

- ▶ Support for cluster verification and synchronization.
- ▶ Automatic failover and reintegration of server nodes attached to SVC-provided virtual disk pairs within sites.
- ▶ SVC management to switch the SVC-PPRC relationships over so that the backup site can take control of the PowerHA SystemMirror-managed resource groups from the primary site in case of a site-failure.
- ▶ SVC-PPRC Command Line Interface (CLI) or GUI to manually manage SVC-PPRC consistency groups and relationships.
- ▶ Limited support for C-SPOC.

It is important to understand the terms *master* and *auxiliary* and how they are used in PowerHA SystemMirror Enterprise Edition Metro Mirror for an SVC environment. In general, master and auxiliary refer to SVC virtual disks that are on either end of an SVC PPRC link. *Primary* and *Secondary* refer to the PowerHA sites that host the resource groups that manage SVC PPRC replicated resources that contain those SVC-PPRC links.

The terms master and auxiliary also refer to the SVC clusters themselves. The master SVC cluster is connected to the PowerHA SystemMirror production or primary site, and the auxiliary SVC cluster is connected to the backup or recovery site.

For more information about IBM SAN Volume Controller, refer to *Implementing the IBM System Storage SAN Volume Controller*, SG24-6423, which can be found at:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg246423.pdf>

SVC PPRC states

For managing SVC PPRC replicated resources with PowerHA, you should understand each PPRC state.

The following PPRC volume states are possible, for either Consistency Groups (CG) or PPRC relationships:

- ▶ inconsistent_stopped

In this state, the master is accessible for read and write I/O but the auxiliary is not accessible for either. A copy process needs to be started to make the auxiliary consistent.

- ▶ inconsistent_copying

In this state, the master is accessible for read and write I/O, but the auxiliary is not accessible for either. This state is entered after a **Start** command is issued to an *InconsistentStopped* relationship or Consistency Group. It is also entered when a **Forced Start** is issued to an *Idling* or *ConsistentStopped* relationship or Consistency Group. A background copy process runs which copies data from the master to the auxiliary virtual disk.

- ▶ consistent_stopped

In this state, the auxiliary contains a consistent image but it might be out-of-date with respect to the master.

- ▶ consistent_synchronized

In this state, the master VDisk is accessible for read and write I/O. The auxiliary VDisk is accessible for read-only I/O. Writes that are sent to the master VDisk are sent to both master and auxiliary VDisks. Either good completion must be received for both writes, or

the write must be failed to the host, or a state transition out of consistent_synchronized must take place before a write is completed to the host.

- ▶ idling

Both master and auxiliary disks are operating in the master role. Consequently, both are accessible for write I/O. In this state, the relationship or Consistency Group will accept a *Start* command. Remote Copy maintains a record of regions on each disk which have received write I/O while Idling. This is used to determine what areas need to be copied following a *Start* command.

- ▶ idling_disconnected

The Virtual Disks in this half of the relationship or Consistency Group are all in the master role and accept read or write I/O. No configuration activity is possible (except for deletes or stops) until the relationship becomes connected again. At that point, the relationship transitions to a *Connected* state.

- ▶ inconsistent_disconnected

The Virtual Disks in this half of the relationship or Consistency Group are all in the auxiliary role and do not accept read or write I/O. No configuration activity except for deletes is permitted until the relationship becomes connected again.

- ▶ consistent_disconnected

The VDisks in this half of the relationship or Consistency Group are all in the auxiliary role and accept read I/O but not write I/O. This state is entered from ConsistentSynchronized or ConsistentStopped when the auxiliary side of a relationship becomes disconnected.

- ▶ empty

This state only applies to Consistency Groups. It is the state of a Consistency Group that has no relationships and hence no other state information to show. It is entered when a Consistency Group is first created. It is exited when the first relationship is added to the Consistency Group, at which point the state of the relationship becomes the state of the Consistency Group.

6.2 Planning and prerequisites overview

In this section, we discuss planning and prerequisites for installing and configuring PowerHA 7.1.2 Enterprise Editions for SVC.

6.2.1 Planning

Before configuring PowerHA 7.1.2 Enterprise Edition for SVC, check the following:

- ▶ The PowerHA SystemMirror sites have been planned.
- ▶ Basic SVC and SVC PPRC support has been completely configured. Refer to the SVC documentation about how to install and configure SVC and SVC PPRC support.
- ▶ SVC clusters and native PPRC support on these SVC clusters has already been configured.

To plan for SVC-managed PPRC replicated resources in a PowerHA SystemMirror cluster, identify the following:

- ▶ The SVC clusters to be used at each site.

- ▶ SVC relationships.
- ▶ Volume groups associated with SVC virtual disks (VDisks) to be used in the relationships.
- ▶ SVC consistency groups.
- ▶ The relationships to be used in the consistency groups.
The term Consistency Group Name can be, in this instance, interchanged with the term PPRC Replicated Resource Name. In the context of SVC, they are the same.
- ▶ Plan which resource groups will contain the SVC-managed PPRC replicated resources.

Limitations for PowerHA SystemMirror Enterprise Edition for SVC

PowerHA SystemMirror 7.1.2 Enterprise Edition utilizing SVC PPRC has the following restrictions:

- ▶ SVC Host Naming Convention
Although SVC Host Name Aliases are arbitrary, for PowerHA SystemMirror Enterprise Edition SVC PPRC, they must match the node names you define for the PowerHA SystemMirror Enterprise Edition sites. This ensures that the `ssh` commands used to execute SVC tasks are completed on the correct SVC nodes.
- ▶ SSH must be installed and configured
PowerHA SystemMirror uses commands to communicate with the SVC PPRC cluster that require `ssh`. Therefore, some version of `ssh` must be installed and configured on all SVC PPRC cluster nodes.
- ▶ Volume Groups
Resource Groups to be managed by PowerHA SystemMirror cannot contain volume groups with both SVC PPRC-protected and non SVC-PPRC-protected disks.
For example:
 - VALID: RG1 contains VG1 and VG2, both PPRC-protected disks.
 - INVALID: RG2 contains VG3 and VG4, VG3 is PPRC-protected, and VG4 is not.
 - INVALID: RG3 contains VG5, which includes both PPRC-protected and non-protected disks within the same volume group.
- ▶ SVC cluster names and PowerHA SystemMirror Host aliases
The host aliases on your SVC clusters must match the node names (short names) used in PowerHA SystemMirror to define each cluster node.
- ▶ C-SPOC operations
You cannot use C-SPOC for the following LVM operations to configure nodes at the remote site (that contain the target volumes):
 - Creating or extending a volume group
 - Operations that require nodes at the target site to write to the target volumes (for example, changing file system size, changing mount point, adding LVM mirrors) cause an error message in C-SPOC. This includes functions such as changing file system size, changing mount points, and adding LVM mirrors. However, nodes on the same site as the source volumes can successfully perform these tasks. The changes are subsequently propagated to the other site via lazy update.

For C-SPOC operations to work on all other LVM operations, we suggest performing all C-SPOC operations when the cluster is active on all PowerHA SystemMirror nodes and the underlying SVC consistency groups are in a `consistent_synchronized` state.

Reference: For complete details and corresponding planning, see *Storage-based high availability and disaster recovery for PowerHA SystemMirror Enterprise Edition* at:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/ha_pprc_main.htm

6.2.2 Prerequisite overview

In this section, we discuss the prerequisites to install and configure PowerHA Enterprise Edition for SVC.

Software requirements

The following list describes the minimal required software for implementing PowerHA SystemMirror 7.1.2 Enterprise Edition with SVC:

- ▶ cluster.xd.license
- ▶ cluster.es.svcpprc.cmds
- ▶ cluster.es.svcpprc.rte
- ▶ cluster.msg.en_US.svcpprc
(and other appropriate language message sets)

Because each type of PPRC management has different prerequisites, information about installing the particular filesets for specific support types (for example, cluster.es.pprc filesets) is deferred to the section specific to that PPRC management type.

The following software and microcode levels are required:

- ▶ openssh version 3.6.1 or later (for access to SVC interfaces)

When running SVC version 4.x:

- ▶ Storage microcode/LIC versions as per SVC support requirements
- ▶ Subsystem Device Driver (SDD) v 1.6.3.0 or higher
- ▶ IBM Host attachment scripts:
devices.fcp.disk.ibm.rte 1.0.0.9 or later
- ▶ ibm2105.rte 32.6.100.25 or later (as specified by SVC support)

When using Subsystem Device Driver Path Control Module (SDDPCM):

- ▶ v 2.2.0.0 or higher
- ▶ IBM host attachment scripts:
devices.fcp.disk.ibm.mpio.rte 1.0.0.10 or later

Note: For information on the device driver, consult the System Storage Multipath Subsystem Device Driver User's Guide, GC52-1309-07 at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S7000303&aid=1>

When using Virtual I/O Server 1.5.1.x:

- ▶ v 2.2.0.0 or higher
- ▶ IBM host attachment scripts:

devices.fcp.disk.ibm.mpio.rte 1.0.0.10 or later

Note: At the time of writing this IBM Redbooks publication, SVC v6.4 is the most recent version that the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition supports.

IBM PowerHA SystemMirror for AIX supports IBM System Storage SVC Software V6.4:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10795>

SSH connection configuration

Configure ssh for each node that will communicate to the SVC clusters. Public or private key pairs must be created on the nodes and distributed to the SVC clusters in order for PowerHA SystemMirror Enterprise Edition Metro Mirror for SVC to function. Also see the prerequisite list for fileset name and version in “Software requirements” on page 201.

The SVC PPRC replicated methods used with PowerHA SystemMirror Enterprise Edition rely heavily on remote commands executed using ssh. It's required, for example, to remotely execute a command via `ssh` on all cluster nodes that are part of a resource group that contains a PPRC replicated resource. Private and public key pairs must be installed on both the nodes that will access the SVC clusters, and the SVC clusters themselves in order for the SVC PPRC commands to work correctly.

Make sure the host aliases on your SVC clusters match the node names (short names) used in PowerHA SystemMirror. To get a list of the current host to VDisks maps on the master and auxiliary SVC clusters, use the SVC CLI commands:

```
ssh admin@<ip_for_MASTER_SVC_Cluster> svcinfo lshostvdiskmap | more  
ssh admin@<ip_for_AUXILIARY_SVC_Cluster> svcinfo lshostvdiskmap | more
```

Sample outputs are shown in Example 6-1.

Example 6-1 List of vDisks mapping through ssh

```
# ssh admin@SiteA svcinfo lshostvdiskmap | more  
id name          SCSI_id vdisk_id vdisk_name           vdisk_UID  
12 SiteA_itsoal 1      50      itso_data1          60050768018205B6500000000000000004E  
12 SiteA_itsoal 2      51      itso_data2          60050768018205B6500000000000000004F  
...  
  
# ssh admin@SiteB svcinfo lshostvdiskmap | more  
id name          SCSI_id vdisk_id vdisk_name           vdisk_UID  
10 SiteB_itsob1 0      42      itsob1_rootvg       600507680180863B8000000000000000036  
10 SiteB_itsob1 1      45      itso_data1          600507680180863B800000000000000003A  
10 SiteB_itsob1 2      46      itso_data2          ...
```

Check the host names listed against your PowerHA SystemMirror node names. If they differ, then refer to the SVC documentation about how to change the names to match, either via the SVC CLI or GUI interface.

Additional details for configuring ssh access can be found in “Configuring ssh access to the SVC” on page 208.

Identifying VDisks of SVC to hdisks from AIX clients

Before configuring the SVC PPRC replicated resources, you need to identify the corresponding Vdisk on the SVC to each hdisk device on the AIX client nodes.

The following procedure shows how to determine which devices correspond to each other.

An SVC VDisk has a unique_id (UID) on the SVC, which is also part of the disk device definition in AIX. You can find this information from the SVC master console under **Volumes** → **All Volumes** (Figure 6-2).

Name	Status	Capacity	Storage Pool	UID
itso_data1	Online	40.0 GB	SiteA_pool	60050768018205B65000000000000004E
itso_data2	Online	40.0 GB	SiteA_pool	60050768018205B65000000000000004F
itso_data3	Online	40.0 GB	SiteA_pool	60050768018205B650000000000000050
itso_data4	Online	40.0 GB	SiteA_pool	60050768018205B650000000000000051
itso_data5	Online	40.0 GB	SiteA_pool	60050768018205B650000000000000052
itso_data6	Online	40.0 GB	SiteA_pool	60050768018205B650000000000000053
itso_data7	Online	40.0 GB	SiteA_pool	60050768018205B650000000000000054
itso_data8	Online	40.0 GB	SiteA_pool	60050768018205B650000000000000055

Figure 6-2 VDisk maps from SVC master console

You can also check each UID via the command line. Assuming that ssh access from the client to the SVC has been configured, run:

```
ssh admin@<ip_for_SVC_Cluster> svcinfo lshostvdiskmap | more
```

You can also grep on the host alias name to narrow the list (Example 6-2).

Example 6-2 UID via command line

```
# ssh admin@SiteA svcinfo lshostvdiskmap | grep itso_data2
12 SiteA_itsoa1      2      51      itso_data2      60050768018205B65000000000000004F
13 SiteA_itsoa2      2      51      itso_data2      60050768018205B65000000000000004F
```

You need to collect this information from each VDisk to be used for the PPRC relationship from the SVC cluster in both sites.

On the AIX clients, the UID is in the ODM. You can get this using **lsattr -El hdiskX -a unique_id**. Example 6-3 shows the VDisk UID in bold.

Example 6-3 UID from the lsattr command

```
# lsattr -El hdisk2 -a unique_id
unique_id 3321360050768018205B650000000000004F04214503IBMfcp Device Unique Identification False
```

Also, APAR IV02884 introduced the **-u** flag for the **lspv** command to show the UID of all the disks. Example 6-4 shows the VDisk UID in bold.

Example 6-4 UID from lspv command

```
# lspv -u
hdisk2      00f623c5fbfe9053      siteavg      concurrent
3321360050768018205B650000000000004F04214503IBMfcp      37602ed0-e455-58c4-1443-e212916a7b4e
```

From Example 6-3 or Example 6-4, hdisk2 matches the VDisk UID of itso_data2 in Example 6-2. Repeat the command to match, record and create proper replicated relationships. As the nodes from site siteA share the same disks, you only need to get the VDisk UID from one of these nodes. You can do the same on a node from site siteB.

6.3 Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with SVC remote copy

In this section, we discuss the steps to create a cluster with SVC remote copy.

6.3.1 Scenario overview

This section provides an overview of the SVC cluster configuration scenario.

SVC cluster environment

In this scenario, we had a cluster with two sites. Each site consisted of two nodes and one SVC cluster. Since we did not have one SAN spanning both sites, we had to use a linked cluster. As required for linked clusters, each site had its own CAA repository disk.

The link between the sites of the cluster assured that the CAA repository disks were synced and within the site all the nodes used the same repository disk for maintaining the CAA configuration, as shown in Figure 6-3.

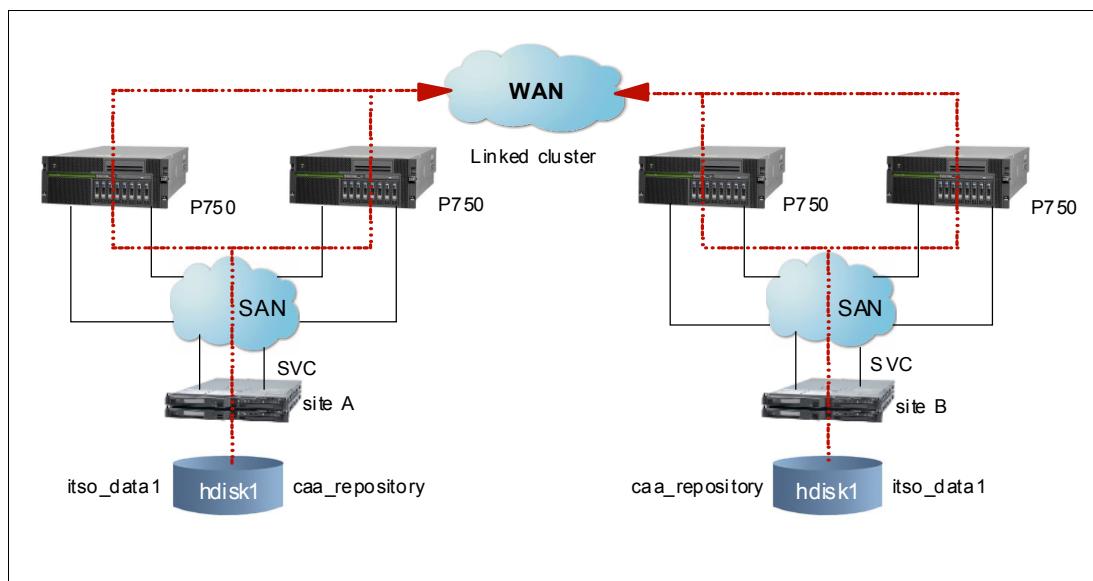


Figure 6-3 Scenario with linked cluster

The site configuration was as follows:

- ▶ siteA:
 - Node itsoa1
 - Node itsoa2
 - SVC SiteA
- ▶ siteB:
 - Node itsob1
 - Node itsob2
 - SVC SiteB

On each node we had nine disks:

- ▶ hdisk0 → rootvg
- ▶ hdisk1 → CAA repository disk

- ▶ hdisk2-hdisk4 → SiteA MetroMirror
- ▶ hdisk5-hdisk8 → SiteB Global Mirror

The final disk configuration is shown in Figure 6-4.

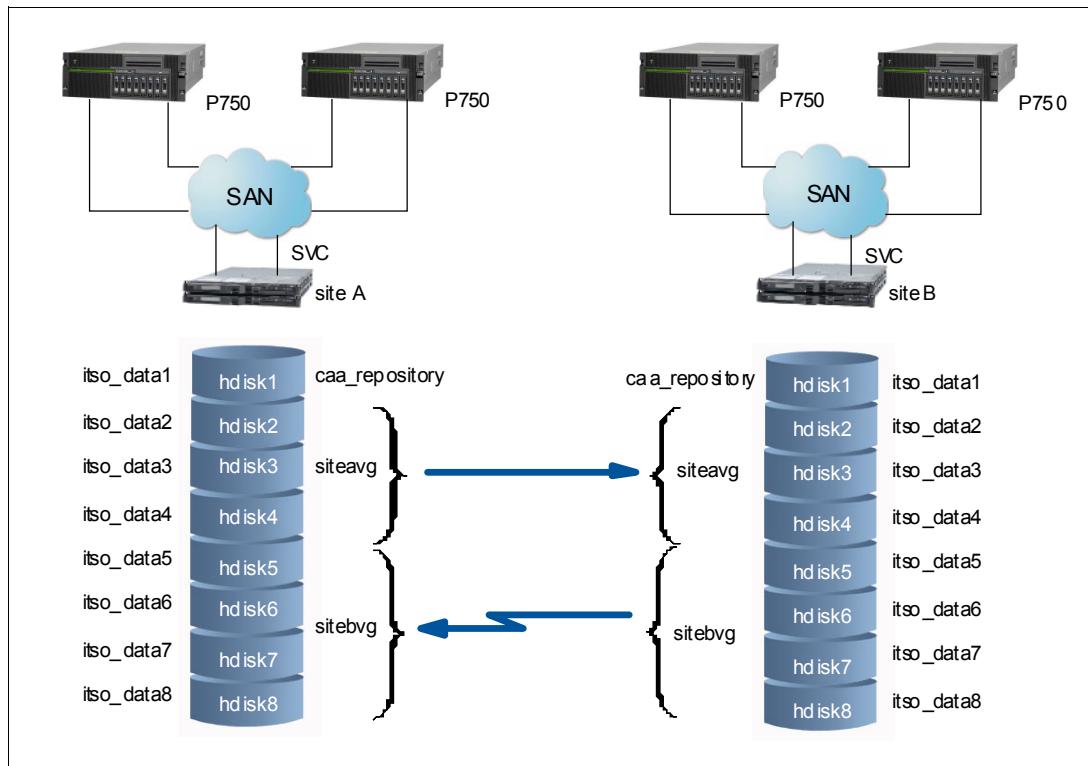


Figure 6-4 Environment disk configuration

6.3.2 Environment configuration for PowerHA

This section provides the prerequisite steps before configuring PowerHA with the following steps:

1. Configure name resolutions.
2. Configure volume groups on one node.
3. Configure SSH access to the SVC.
4. Copy and import volume groups to other nodes.

Configuring name resolutions

First you need to configure your /etc/hosts files to contain all the IP addresses that are used on the cluster configuration, as shown in Example 6-5.

Example 6-5 /etc/hosts

```
# Communication Network

129.40.100.77 itsoa1
129.40.100.79 itsoa2
129.40.100.91 itsob1
129.40.100.92 itsob2

# Boot
```

```
192.168.150.1 itsoa1boot  
192.168.150.2 itsoa2boot  
192.168.160.1 itsob1boot  
192.168.160.2 itsob2boot
```

```
# Service  
192.168.150.3 itsoasrva  
192.168.160.3 itsobsrva  
192.168.150.4 itsoasrvb  
192.168.160.4 itsobsrvb  
  
# SVC  
129.40.100.70 SiteA  
129.40.100.80 SiteB
```

The important aspects for name resolutions are as follows:

- ▶ CAA requires the configured IP addresses to correspond to the IP labels returned by the **/usr/bin/hostname** command.
Check with the **host hostname** command that this is correctly configured.
- ▶ When using SVC replicated resources, they must match the node names you later define for the PowerHA SystemMirror Enterprise Edition sites.

Then, on all nodes, populate the CAA rhosts files (*/etc/cluster/rhosts*) with the IP labels to be used for the communication path, as shown in Example 6-6. After configuring it, restart the clcomd daemon with the **stopsrc** and **startsrc** commands, respectively.

Example 6-6 /etc/cluster/rhosts

```
itsoa1  
itsoa2  
itsob1  
itsob2
```

Configure the volume groups on one node

The next step is to make sure that all disks and volume groups are defined on all nodes. This requires you to create a volume group on one site, then a temporary remote copy relationship, and finally synchronize it so that the volume groups appear on all nodes.

As required in PowerHA 7.1 or later, the volume groups are enhanced concurrent capable. On node itsoa1, one VG is named *siteavg* containing one JFS2 file system, */siteafs*, on the *sitealv* logical volume. On node itsob1, there is a *sitebvg* volume group that contains one JFS2 file system, */sitebfs*, on the *siteblv* logical volume. Example 6-7 shows how our initial disk configuration looked on all nodes. The remote copies are not yet defined on the SVC and the volume groups are defined only on the home node for each site.

Example 6-7 Disk configuration on all nodes

```
root@itsoa1:/>lspv  
hdisk0      00f623c5983627df          rootvg      active  
hdisk1      00f623c5ff66a93a          None        active  
hdisk2      00f623c5fbfe9053         siteavg     active  
hdisk3      00f623c5fbfe9128         siteavg     active
```

```

hdisk4      00f623c5fbfe920e      siteavg      active
hdisk5      none                   None
hdisk6      none                   None
hdisk7      none                   None
hdisk8      none                   None

root@itsoa2:/>lspv
hdisk0      00f623c59859eaa0      rootvg       active
hdisk1      00f623c5ff66a93a      None
hdisk2      00f623c5fbfe9053      None
hdisk3      00f623c5fbfe9128      None
hdisk4      00f623c5fbfe920e      None
hdisk5      none                   None
hdisk6      none                   None
hdisk7      none                   None
hdisk8      none                   None

root@itsob1:/>lspv
hdisk0      00f67bdd9873ac65      rootvg       active
hdisk1      00f67bddff66e38d      None
hdisk2      none                   None
hdisk3      none                   None
hdisk4      none                   None
hdisk5      00f623c5fbfe92cc      sitebvg     active
hdisk6      00f623c5fbfe93a5      sitebvg     active
hdisk7      00f623c5fbfe9440      sitebvg     active
hdisk8      00f623c5fbfe94e6      sitebvg     active

root@itsob2:/>lspv
hhdisk0     00f67bdd9873ac65      rootvg       active
hdisk1      00f67bddff66e38d      None
hdisk2      none                   None
hdisk3      none                   None
hdisk4      none                   None
hdisk5      00f623c5fbfe92cc      None
hdisk6      00f623c5fbfe93a5      None
hdisk7      00f623c5fbfe9440      None
hdisk8      00f623c5fbfe94e6      None

```

We now varied off and exported the VGs, as shown in Example 6-8. We created temporary remote copies on the SVC for siteavg and sitebvg to get the VG information in sync.

Example 6-8 Varyoff and exporting the volume groups

```

root@itsoa1:/>varyoffvg siteavg
root@itsoa1:/>
root@itsoa1:/>exportvg siteavg

root@itsob1:/>varyoffvg sitebvg
root@itsob1:/>
root@itsob1:/>exportvg sitebvg

```

Configuring ssh access to the SVC

To access SVC from PowerHA, we configured the rsa key connection between the root user from all nodes and both SVCs. Using the `ssh-keygen -t rsa` command, we created a pair of private and public keys, as shown in Example 6-9.

Example 6-9 Creating the rsa ssh key pairs

```
ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (//.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in //.ssh/id_rsa.
Your public key has been saved in //.ssh/id_rsa.pub.
The key fingerprint is:
90:57:eb:0c:6e:de:5f:5d:89:ea:74:c8:4a:b8:79:47 root@lbstds2
The key's randomart image is:
+--[ RSA 2048]----+
|          .         |
|       . . .        |
|      o o .        |
|     + + . .        |
|    S o . . .      |
|   o...Eo ..        |
|  ...o= ... .      |
| +.+o..           |
| o.....           |
+-----+
```

Copy both keys in `/.ssh/` on all nodes with the same permission as the initially generated one.

Then you need to pass the public key to both SVC clusters and create a new user. Log in as the administrator user (superuser, for example) go to **User Management** → **New User** and create a user with the public key you just generated, as shown in Figure 6-5 on page 209.



Figure 6-5 Adding a user with the ssh public key

Execute this for both SVC clusters. After the key is exchanged, you should be able to log in to the SVC admin account without password. This allows you to create the temporary relationship via the SVC CLI.

Note: As shown in Figure 6-5, any usernames are allowed. This procedure allows the non-prompt login for the *admin* user on the SVC.

Copying and importing volume groups to other nodes

Although PowerHA automatically creates and maintains SVC copy groups later, for the initial configuration you are required to create a temporary copy group for synchronizing the volume groups to other nodes.

Tips: See also setting up the volume group and file system for PowerHA SystemMirror management at:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/ha_pprc_svc_vg_hacmp.htm

In our case, we made copies of the volumes hdisk2-hdisk4 from siteA to siteB, and hdisk5-hdisk8 from siteB to siteA. We then started the copying procedure as seen in Example 6-10 on 210. The syntax for the command follows:

```
mkrcrelationship -master <master vdisk name> -aux <Auxiliary vdisk name>
-cluster <remote site SVC cluster name> -name <relationship name>
```

The master and auxiliary Vdisk name should be the name given to the LUN (vdisk) on the SVC cluster as shown in 6.1, “Overview of SVC management” on page 196. These names are also used on the PowerHA configuration.

Important: Make sure that all logical volumes (including loglv) and file systems have unique names on each volume group. Otherwise the volume group (VG) cannot be imported after the remote copy finishes.

Example 6-10 Creating and starting the relationship between sites

```
root@itsoa1:/>ssh admin@SiteA
Last login: Thu Nov 15 16:03:27 2012 from 129.40.100.77
SiteA:admin>mkrcrelationship -master itso_data2 -aux itso_data2 -cluster SiteB -name itso_temp1
RC Relationship, id [51], successfully created
SiteA:admin>mkrcrelationship -master itso_data3 -aux itso_data3 -cluster SiteB -name itso_temp2
RC Relationship, id [52], successfully created
SiteA:admin>mkrcrelationship -master itso_data4 -aux itso_data4 -cluster SiteB -name itso_temp3
RC Relationship, id [53], successfully created
SiteA:admin>startrcrelationship itso_temp1
SiteA:admin>startrcrelationship itso_temp2
SiteA:admin>startrcrelationship itso_temp3
SiteA:admin>>exit
exit
Connection to SiteA closed.

root@itsoa1:/>
root@itsoa1:/>ssh admin@SiteB
Last login: Thu Nov 15 13:53:14 2012 from 129.40.100.77
SiteB:admin>mkrcrelationship -master itso_data5 -aux itso_data5 -cluster SiteA -name itso_temp4
RC Relationship, id [49], successfully created
SiteB:admin>mkrcrelationship -master itso_data6 -aux itso_data6 -cluster SiteA -name itso_temp5
RC Relationship, id [50], successfully created
SiteB:admin>mkrcrelationship -master itso_data7 -aux itso_data7 -cluster SiteA -name itso_temp6
RC Relationship, id [51], successfully created
SiteB:admin>mkrcrelationship -master itso_data8 -aux itso_data8 -cluster SiteA -name itso_temp7
RC Relationship, id [52], successfully created
SiteB:admin>startrcrelationship itso_temp4
SiteB:admin>startrcrelationship itso_temp5
SiteB:admin>startrcrelationship itso_temp6
SiteB:admin>startrcrelationship itso_temp7
SiteB:admin>>exit
exit
Connection to SiteB closed.
root@itsoa1:/>
```

The process shown in Example 6-10 starts the remote copy. It takes some time to synchronize all the physical volumes between sites. It is possible to verify the status of the copy with the **lsrcrelationship <relationship name>** command. As seen in Example 6-11, you need to wait until the status is in *consistent_synchronized* mode in all remote copy relationships.

Example 6-11 Checking relationship status

```
SiteA:admin>lsrcrelationship itso_temp1
id 51
name itso_temp1
master_cluster_id 0000020060816D94
master_cluster_name SiteA
```

```
master_vdisk_id 51
master_vdisk_name itso_data2
aux_cluster_id 0000020060218EE0
aux_cluster_name SiteB
aux_vdisk_id 46
aux_vdisk_name itso_data2
primary master
consistency_group_id
consistency_group_name
state inconsistent_copying
bg_copy_priority 50
progress 26
freeze_time
status online
sync
copy_type metro
SiteA:admin>
```

```
SiteA:admin>lsrcrelationship itso_temp1
id 51
name itso_temp1
master_cluster_id 0000020060816D94
master_cluster_name SiteA
master_vdisk_id 51
master_vdisk_name itso_data2
aux_cluster_id 0000020060218EE0
aux_cluster_name SiteB
aux_vdisk_id 46
aux_vdisk_name itso_data2
primary master
consistency_group_id
consistency_group_name
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

After all disks are in sync, stop the remote copy relationship and delete it as shown in Example 6-12.

Example 6-12 Removing the remote copy relationship

```
SiteA:admin>stoprcrelationship itso_temp1
SiteA:admin>stoprcrelationship itso_temp2
SiteA:admin>stoprcrelationship itso_temp3
SiteA:admin>stoprcrelationship itso_temp4
SiteA:admin>stoprcrelationship itso_temp5
SiteA:admin>stoprcrelationship itso_temp6
SiteA:admin>stoprcrelationship itso_temp7
```

```
SiteA:admin>rmmrcrelationship itso_temp1
SiteA:admin>rmmrcrelationship itso_temp2
SiteA:admin>rmmrcrelationship itso_temp3
```

```
SiteA:admin>rmrcrelationship itso_temp4
SiteA:admin>rmrcrelationship itso_temp5
SiteA:admin>rmrcrelationship itso_temp6
SiteA:admin>rmrcrelationship itso_temp7
SiteA:admin>
```

You can import the VG again on all nodes. First remove all remote copied hdisks from the ODM so you can get the synchronized PVID, then run the **cfgmgr** command and then import the volume groups as seen in Example 6-13. You need to repeat the procedure on all nodes.

Example 6-13 Importing the synchronized volume groups

```
root@itsob1:/>rmdev -dl hdisk2
hdisk2 deleted
root@itsob1:/>rmdev -dl hdisk3
hdisk3 deleted
root@itsob1:/>rmdev -dl hdisk4
hdisk4 deleted
root@itsob1:/>rmdev -dl hdisk5
hdisk5 deleted
root@itsob1:/>rmdev -dl hdisk6
hdisk6 deleted
root@itsob1:/>rmdev -dl hdisk7
hdisk7 deleted
root@itsob1:/>rmdev -dl hdisk8
hdisk8 deleted
root@itsob1:/>cfgmgr
root@itsob1:/>lspv
hdisk0      00f67bdd9871d20c          rootvg      active
hdisk1      00f67bddff66e38d          None
hdisk2      00f623c5fbfe9053          None
hdisk3      00f623c5fbfe9128          None
hdisk4      00f623c5fbfe920e          None
hdisk5      00f623c5fbfe92cc          None
hdisk6      00f623c5fbfe93a5          None
hdisk7      00f623c5fbfe9440          None
hdisk8      00f623c5fbfe94e6          None
root@itsob1:/>importvg -V 50 -y siteavg hdisk2
siteavg
0516-783 importvg: This imported volume group is concurrent capable.
Therefore, the volume group must be varied on manually.
root@itsob1:/>importvg -V 51 -y sitebvg hdisk5
sitebvg
0516-783 importvg: This imported volume group is concurrent capable.
Therefore, the volume group must be varied on manually.
root@itsob1:/>
```

6.3.3 Configuring the cluster basic definitions

Now you can start to define all cluster configurations. There are a few changes in the SMIT menu in order to create the linked cluster. To start configuring the cluster, issue **smitty sysmirror** and then select **Cluster Nodes and Networks → Multi Site Cluster Deployment**. This submenu is the starting point for all the other menus in this subsection.

Choose **Setup a Cluster, Nodes and Networks**, which leads to the panel shown in Figure 6-6.

Setup Cluster, Sites, Nodes and Networks	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
* Cluster Name	[Entry Fields] [itso_cluster]
* Site 1 Name	[itsoa]
* New Nodes (via selected communication paths)	[itsoa1 itsoa2] +
* Site 2 Name	[itsob]
* New Nodes (via selected communication paths)	[itsob1 itsob2] +
Cluster Type	[Linked Cluster] +

Figure 6-6 Creating a new linked cluster

After pressing Enter, a discovery process is also executed. The process automatically creates the cluster networks and saves a list of all shared disks. The next step is to choose the repository disks and multicast address to be used by CAA.

Choose **Define Repository Disk and Cluster IP Address**, which displays the SMIT panel as shown in Figure 6-7. We selected hdisk1, which was planned to be the CAA repository disk.

Multi Site with Linked Clusters Configuration	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
* Site Name	[Entry Fields] itsoa
* Repository Disk	[00f623c5ff66a93a]] +
Site Multicast Address	[Default]
* Site Name	itsob
* Repository Disk	[00f67bddff66e38d]] +
Site Multicast Address	[Default]

Figure 6-7 Configuring CAA repository disks and multicast addresses

The minimum requirements to configure the cluster have now been met. Now verify and synchronize the cluster. On the main PowerHA SMIT panel select **Cluster Nodes and Networks** → **Verify and Synchronize Cluster Configuration**.

6.3.4 Configuring SVC replicated resources

Now we configured the SVC replicated resources between sites, issued **smitty sysmirror** on the command prompt and then selected **Cluster Applications and Resources** →

Resources → Configure SVC PPRC-Replicated Resources. We used this as our starting point for this section of SMIT menus.

The following steps are required to configure SVC replicated resources.

1. Add an SVC cluster.
2. Add an SVC PPRC relationship.
3. Add an SVC PPRC replicated resource.

Add an SVC cluster

To add the SVC cluster, go to **SVC Clusters Definition to PowerHA SystemMirror → Add an SVC Cluster**. You can see the SMIT panel for the SVC siteA in Figure 6-8. Remember that the *SVC Cluster Name* must match the name resolutions in the /etc/hosts file and the actual SVC node name.

Add an SVC Cluster	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
* SVC Cluster Name	[Entry Fields]
* SVC Cluster Role	[SiteA]
* PowerHA SystemMirror site	[Master] +
* SVC Cluster IP Address	itsoa +
SVC Cluster Second IP Address	[129.40.100.70]
* Remote SVC Partner	[]
	[SiteB]

Figure 6-8 Defining the SVC siteA cluster on the PowerHA configuration

The field values are as follows:

- ▶ SVC cluster name
Enter the same name used by SVC. This name cannot be more than 20 alphanumeric characters and underscores.
- ▶ SVC cluster role
Select master or auxiliary.
- ▶ PowerHA SystemMirror site
Select the PowerHA SystemMirror site associated with the SVC cluster.
- ▶ SVC cluster IP address
IP address of this cluster.
- ▶ Remote SVC partner
Name of the SVC cluster hosting vDisks from the other side of the SVC PPRC link.

We repeated the previous procedure for the SVC cluster siteB as shown in Figure 6-9 on page 215.

Add an SVC Cluster																							
Type or select values in entry fields.																							
Press Enter AFTER making all desired changes.																							
<table border="0"> <tr> <td>* SVC Cluster Name</td> <td colspan="2">[Entry Fields]</td> </tr> <tr> <td>* SVC Cluster Role</td> <td>[SiteB]</td> <td>+</td> </tr> <tr> <td>* PowerHA SystemMirror site</td> <td>[Master]</td> <td>+</td> </tr> <tr> <td>* SVC Cluster IP Address</td> <td>itsoa</td> <td>+</td> </tr> <tr> <td>SVC Cluster Second IP Address</td> <td>[129.40.100.70]</td> <td></td> </tr> <tr> <td>* Remote SVC Partner</td> <td>[]</td> <td></td> </tr> <tr> <td></td> <td>[SiteB]</td> <td></td> </tr> </table>			* SVC Cluster Name	[Entry Fields]		* SVC Cluster Role	[SiteB]	+	* PowerHA SystemMirror site	[Master]	+	* SVC Cluster IP Address	itsoa	+	SVC Cluster Second IP Address	[129.40.100.70]		* Remote SVC Partner	[]			[SiteB]	
* SVC Cluster Name	[Entry Fields]																						
* SVC Cluster Role	[SiteB]	+																					
* PowerHA SystemMirror site	[Master]	+																					
* SVC Cluster IP Address	itsoa	+																					
SVC Cluster Second IP Address	[129.40.100.70]																						
* Remote SVC Partner	[]																						
	[SiteB]																						

Figure 6-9 Defining SVC siteB cluster on PowerHA configuration

The `/usr/es/sbin/cluster/svcpprc/cmds/c11ssvc` command can be used to confirm your settings. Figure 6-10 shows an example.

```
>./c11ssvc -a
#SVCNAME ROLE SITENAME IPADDR IPADDR2 RPARTNER
SiteA Master itsoa 129.40.100.70 SiteB
SiteB Master itsob 129.40.100.80 SiteA
```

Figure 6-10 c11ssvc command output

Adding the SVC PPRC relationship

To configure the remote copy across sites, go to **SVC PPRC Relationships Definition** → **Add an SVC PPRC Relationship** and include all the VDisk relationships across the sites as shown in Figure 6-11.

Add an SVC PPRC Relationship														
Type or select values in entry fields.														
Press Enter AFTER making all desired changes.														
<table border="0"> <tr> <td>* Relationship Name</td> <td colspan="2">[Entry Fields]</td> </tr> <tr> <td>* Master VDisk Info</td> <td>[itso_rel1]</td> <td></td> </tr> <tr> <td>* Auxiliary VDisk Info</td> <td>[itso_data2@SiteA]</td> <td></td> </tr> <tr> <td></td> <td>[itso_data2@SiteB]</td> <td></td> </tr> </table>			* Relationship Name	[Entry Fields]		* Master VDisk Info	[itso_rel1]		* Auxiliary VDisk Info	[itso_data2@SiteA]			[itso_data2@SiteB]	
* Relationship Name	[Entry Fields]													
* Master VDisk Info	[itso_rel1]													
* Auxiliary VDisk Info	[itso_data2@SiteA]													
	[itso_data2@SiteB]													

Figure 6-11 Defining VDisk relationships

Enter field values as follows:

- ▶ Relationship name

The name used by both SVC and PowerHA SystemMirror for configuration of SVC PPRC relationships. Use no more than 20 alphanumeric characters and underscores.

► Master VDisk info

The master and auxiliary VDisk names use the format vdisk_name@svc_cluster_name. The master VDisk is the disk that resides at the primary site for the resource group that contains the SVC PPRC relationship.

► Auxiliary VDisk info

The auxiliary VDisk is the disk at the backup site for the resource group that contains the SVC PPRC relationship.

Repeat this procedure for all disks that have remote copies. If your environment has reverse copies like ours, remember to change the SVC master cluster.

The **/usr/es/sbin/cluster/svcpprc/cmds/c1lsrelationship** command can be used to confirm your settings, as shown in Figure 6-12.

```
>./c1lsrelationship -a
relationship_name MasterVdisk_info AuxiliaryVdisk_info
itso_rel1      itso_data2@SiteA itso_data2@SiteB
itso_rel2      itso_data3@SiteA itso_data3@SiteB
itso_rel3      itso_data4@SiteA itso_data4@SiteB
itso_rel4      itso_data5@SiteB itso_data5@SiteA
itso_rel5      itso_data6@SiteB itso_data6@SiteA
itso_rel6      itso_data7@SiteB itso_data7@SiteA
itso_rel7      itso_data8@SiteB itso_data8@SiteA
```

Figure 6-12 C1lsrelationship command output

Adding the SVC PPRC replicated resource

Now we configured the relationships on the Consistency Groups. To do so we selected **SVC PPRC-Replicated Resource Configuration** → **Add an SVC PPRC Resource** and provided all required information, as shown in Figure 6-13. Repeat the process for the other copy, as seen in Figure 6-14 on page 217. In our case, for the copy SiteB → SiteA we selected Global mirror for testing purposes.

Add an SVC PPRC Resource	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
[Entry Fields]	
* SVC PPRC Consistency Group Name	[itso_sitea]
* Master SVC Cluster Name	[SiteA] +
* Auxiliary SVC Cluster Name	[SiteB] +
* List of Relationships	[itso_rel1 itso_rel2 itso_rel3] +
* Copy Type	[METRO] +
* PowerHA SystemMirror Recovery Action	[MANUAL] +

Figure 6-13 Adding VDisks in consistency groups SiteA → SiteB

Add an SVC PPRC Resource	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
[Entry Fields]	
* SVC PPRC Consistency Group Name	[itso_siteb]
* Master SVC Cluster Name	[SiteB]
* Auxiliary SVC Cluster Name	[SiteA]
* List of Relationships	[itso_rel14 itso_rel15 itso_rel16 itso_rel17]
* Copy Type	[GLOBAL]
* PowerHA SystemMirror Recovery Action	[MANUAL]

Figure 6-14 Adding VDisks on consistency groups SiteB → SiteA

Enter field values as follows:

- ▶ SVC PPRC consistency group name

The name to be used by SVC and also in the resource group configuration. Use no more than 20 alphanumeric characters and underscores.

- ▶ Master SVC cluster name

Name of the master cluster is the SVC cluster connected to the PowerHA SystemMirror primary site.

- ▶ Auxiliary SVC cluster name

Name of the SVC cluster connected to the PowerHA SystemMirror Backup/Recovery site.

- ▶ List of relationships

List of names of the SVC PPRC relationships

- ▶ Copy type

Global mirror processing provides a long-distance remote copy solution across two sites using asynchronous technology. Metro mirror functions offer a synchronous remote copy option that constantly updates a secondary copy of a volume to match changes made to a source volume. Global or metro.

- ▶ PowerHA SystemMirror recovery action

The PPRC eXtended distance recovery policy to indicate the action to be taken by PowerHA SystemMirror in case of a site failover for PPRC XD Type VolumePairs.

MANUAL: Manual intervention required or AUTOMATED: No manual intervention required.

Important: If you specify manual, this does not indicate that a manual intervention is required for all failover scenarios. There are some conditions, such as cluster partition, in which doing an automatic failover from one site to another can cause potential data divergence and integrity issues. If PowerHA SystemMirror detects the potential for such a case, and if the recovery action associated with the mirror group is set to manual, PowerHA SystemMirror does not execute an automatic failover.

The PowerHA SystemMirror Recovery Action is an important aspect, as explained in 6.6.3, “PPRC link failure” on page 237.

The `/usr/es/sbin/cluster/svcpprc/cmds/c1lssvcpprc` command can be used to confirm your settings, as shown in Figure 6-15.

```
> ./c1lssvcpprc -a
svcpprc_consistencygrp MasterCluster    AuxiliaryCluster relationships   CopyType
RecoveryAction
itso_sitea      SiteA           SiteB           itso_rel1 itso_rel2 itso_rel3 METRO
MANUAL
itso_siteb      SiteB           SiteA           itso_rel4 itso_rel5 itso_rel6 itso_rel7
GLOBAL          MANUAL
```

Figure 6-15 *C1lssvcpprc command output*

6.3.5 Configuring resources and resource groups

First we configured site-specific service IPs for both sites by executing `smitty sysmirror` and selecting **Cluster Applications and Resources** → **Resources** → **Configure Service IP Labels/Addresses** → **Add a Service IP Label/Address**. Then we selected the correct network to use the service IP. In our case, we chose `net_ether_02`. You can see an example in Figure 6-16. Repeat the process for each site-specific service IP address.

Add a Service IP Label/Address configurable on Multiple Nodes (extended)

Type or select values in entry fields.

Press Enter AFTER making all desired changes.

* IP Label/Address

+

[Entry Fields]

itsoasrva

* Network Name

Associated Site

[]

net_ether_02

itsoa

+

Figure 6-16 *Configuring the service IP address*

Then create a resource group, open PowerHA SMIT menu and choose **Cluster Applications and Resources** → **Resource Groups** → **Add a Resource Group** and fill in the information required, as shown in Figure 6-17 on page 219.

Add a Resource Group (extended)		
Type or select values in entry fields. Press Enter AFTER making all desired changes.		
[Entry Fields]		
* Resource Group Name	[sitearg]	
Inter-Site Management Policy	[Prefer Primary Site]	+
* Participating Nodes from Primary Site	[itsoa1 itsoa2]	+
Participating Nodes from Secondary Site	[itsob1 itsob2]	+
Startup Policy	Online On Home Node Only	+
Fallover Policy	Fallover To Next Priority Node In The List	+
Fallback Policy	Never Fallback	+

Figure 6-17 Creating a resource group

The inter-site management policy field values are as follows:

- ▶ Ignore (You cannot choose this for SVC replicated resources.)
The resource groups do not have ONLINE SECONDARY instances. Use this option if you use cross-site LVM mirroring.
- ▶ Prefer Primary Site
The primary instance of the resource group is brought ONLINE on the primary site at startup, the secondary instance is started on the other site. The primary instance falls back when the primary site rejoins.
- ▶ Online on Either Site
During startup, the primary instance of the resource group is brought ONLINE on the first node that meets the node policy criteria (either site). The secondary instance is started on the other site. The primary instance does not fall back when the original site rejoins.
- ▶ Online on Both Sites (You cannot choose this for SVC replicated resources.)
During startup, the resource group (node policy must be defined as ONLINE on All Available Nodes) is brought ONLINE on both sites. There is no failover policy or fallback policy. The resource group moves to another site, if there are no nodes or conditions that the resource group can be brought or kept ONLINE on the site. The site that owns the active resource group is called the primary site.

Important: Remember SVC replicated resources support for the following Inter-Site Management Policies for resource groups:

- ▶ Prefer Primary Site
- ▶ Online on Either Site

Repeat the procedure for each resource group as needed. In our case we had two: *sitearg* and *sitebrg*.

Run **smitty sysmirror** and choose **Cluster Applications and Resources → Resource Groups → Change>Show Resources and Attributes for a Resource Group**. We configured for Site A the *itsoasrva* and *itsobsrva* service IP addresses, *siteavg*, and *itso_sitea*

SVC consistency group as *sitearg* resource group. We did the same for Site B and configured *itsoasrvb* and *itsobsrvb* service IP address, *sitebvg*, and *itso_sitea* SVC consistency group as a *sitebvg* resource group, as shown in Figure 6-18.

Change/Show All Resources and 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	sitearg	
Inter-site Management Policy	Prefer Primary Site	
Participating Nodes from Primary Site	itsoal itsoa2	
Participating Nodes from Secondary Site	itsob1 itsob2	
Startup Policy	Online On Home Node 0>	
Fallover Policy	Fallover To Next Prio>	
Fallback Policy	Never Fallback	
Service IP Labels/Addresses	[itsoasrva itsobsrva] +	
Application Controller Name	[] +	
Volume Groups	[siteavg] +	
.		
.		
.		
SVC PPRC Replicated Resources	[itso_sitea] +	
DS8000 Global Mirror Replicated Resources	[] +	
XIV Replicated Resources	+ +	
DS8000-Metro Mirror (In-band) Resources	+ +	
[BOTTOM]		

Figure 6-18 Resource group *sitearg* resource configuration

Then we synchronized the cluster one last time to update the configuration on all nodes. We executed **smitty sysmirror** → **Cluster Applications and Resources** → **Verify and Synchronize Cluster Configuration**.

After this step, we were ready to start Cluster Services. Execute **smitty sysmirror** → **System Management (C-SPOC)** → **PowerHA SystemMirror Services** → **Start Cluster Services**, choose all nodes to start the cluster services and change the Start Cluster Information Daemon to *true*, as shown in Figure 6-19 on page 221.

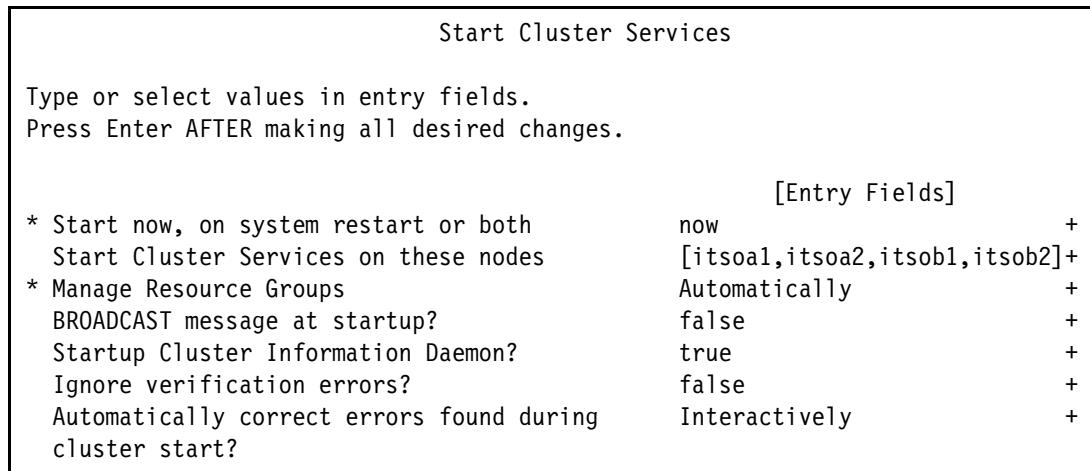


Figure 6-19 Starting Cluster Services

PowerHA starts the remote copy persistent on the SVC clusters and the RG is running on each site. You can see this by viewing the output of the `c1RGinfo` command located in `/usr/es/sbin/cluster/utilities` as shown in Figure 6-20.

root@itsoa1:/>/usr/es/sbin/cluster/utilities/c1RGinfo			
Group	Name	State	Node
sitearg		ONLINE	itsoa1@itsoa
		OFFLINE	itsoa2@itsoa
		ONLINE SECONDARY	itsob1@itsob
		OFFLINE	itsob2@itsob
sitebrg		ONLINE	itsob1@itsob
		OFFLINE	itsob2@itsob
		ONLINE SECONDARY	itsoa1@itsoa
		OFFLINE	itsoa2@itsoa

Figure 6-20 c1RGinfo output

6.4 Storage management with SVC replicated resources

In this section, we show the logical volume management using SVC with global and metro mirror. First we removed one disk from sitebvg. Then we reused the same disk on siteavg. Afterwards we created an LV called `sitealv2`, and a file system called `/siteafs2`.

6.4.1 Removing a disk from a PowerHA environment with SVC remote copy

To accomplish this we removed the physical volume (hdisk) from the sitebvg volume group using CSPOC. From the command line we issued `smitty sysmirror` → **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Set Characteristics of a Volume Group** → **Remove a Volume from a Volume Group**. In our case we chose sitebvg, hdisk8, and all the nodes to remove it from, as shown in Figure 6-21 on page 222.

Remove a Volume from a Volume Group	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
VOLUME GROUP name	[Entry Fields]
Resource Group Name	sitebvg
Node List	sitebrg
VOLUME names	itsoa1,itsoa2,itsob1,itsob2
Reference node	hdisk8
FORCE deallocation of all partitions on	itsob1
	no
	+ []

Figure 6-21 Removing the volume group

Now we removed the disk from the SVC relationship definition of the SVC consistency group configuration. To do that go to the PowerHA main SMIT panel and select **Cluster Applications and Resources** → **Resources** → **Configure SVC PPRC-Replicated Resources** → **SVC PPRC Relationships Definition** → **Remove an SVC PPRC Relationship**. We selected itso_rel7, which was configured to hdisk8 as shown in Figure 6-22.

Select Relationship Name to Remove		
Move cursor to desired item and press Enter.		
itso_rel1		
itso_rel2		
itso_rel3		
itso_rel4		
itso_rel5		
itso_rel6		
itso_rel7		
F1=Help	F2=Refresh	F3=Cancel
F8=Image	F10=Exit	Enter=Do
F1 /=Find	n=Find Next	
F9+-----+		

Figure 6-22 Choosing itso_rel7

Now you just need to remove the definition from the Consistency group by executing **smitty sysmirror** → **Cluster Applications and Resources** → **Resources** → **Configure SVC PPRC-Replicated Resources** → **SVC PPRC-Replicated Resource Configuration** → **Change / Show an SVC PPRC Resource**. Just leave the needed remaining existent relationship for that consistency group (Figure 6-23 on page 223).

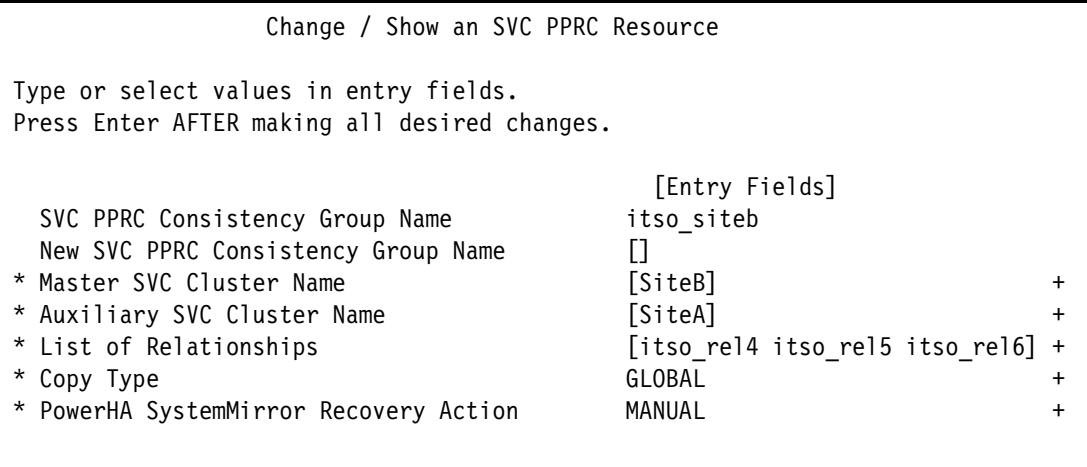


Figure 6-23 Redefining the consistency group

Now verify and synchronize the cluster. PowerHA does not remove the relationship from the SVC cluster. You need to do this manually. So connect to the SVC and issue the `rmrcrelationship` command, as seen in Figure 6-24.

```
SiteA:admin>rmrcrelationship itso_rel7
SiteA:admin>
```

Figure 6-24 Removing the relationship from the SVC cluster

Now we added the volume to SiteA.

6.4.2 Adding a volume to the PowerHA cluster with SVC remote copy

We cleaned the PVID of the disk that we just removed from sitebvg to make the test more meaningful. The PVID was 00f623c5fbfe94e6 as you can see in Example 6-7 on 206, and now it is changed on itsoa1 to 00f623c50b349b68, as seen in Example 6-14.

Example 6-14 Disks on itsoa1

```
root@itsoa1:/>lspv
hdisk0      00f623c5983627df          rootvg      active
hdisk1      00f623c5ff66a93a          caavg_private active
hdisk2      00f623c5fbfe9053          siteavg     concurrent
hdisk3      00f623c5fbfe9128          siteavg     concurrent
hdisk4      00f623c5fbfe920e          siteavg     concurrent
hdisk5      00f623c5fbfe92cc          sitebvg     concurrent
hdisk6      00f623c5fbfe93a5          sitebvg     concurrent
hdisk7      00f623c5fbfe9440          sitebvg     concurrent
hdisk8      00f623c50b349b68          None        
```

Then we created the relationship and synced the remote copy between the volumes on the SVC cluster. After the copy was synced, we added the relationship to the consistency group, as shown in Figure 6-25 on page 224.

```

SiteA:admin>mkrcrelationship -master itso_data8 -aux itso_data8 -cluster SiteB -name itso_rel7
RC Relationship, id [57], successfully created
SiteA:admin>startrcrelationship itso_rel7
SiteA:admin>lsrcrelationship itso_rel7
id 57
name itso_rel7
master_cluster_id 0000020060816D94
master_cluster_name SiteA
master_vdisk_id 57
master_vdisk_name itso_data8
aux_cluster_id 0000020060218EE0
aux_cluster_name SiteB
aux_vdisk_id 52
aux_vdisk_name itso_data8
primary master
consistency_group_id
consistency_group_name
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
SiteA:admin>chrcrelationship -consistgrp itso_sitea itso_rel7
SiteA:admin>

```

Figure 6-25 Creating the consistency group

Then we removed the disk and executed **cfgmgr** to get the new PVID shown in Figure 6-26. If you are adding a new PV, this would be the time to discover the new disk on all other nodes.

```

root@itsob2:/>rmdev -dl hdisk8
hdisk8 deleted
root@itsob2:/>cfgmgr
root@itsob2:/>lspv
hdisk0      00f67bdd9873ac65          rootvg      active
hdisk1      00f67bddff66e38d        caavg_private  active
hdisk2      00f623c5fbfe9053        siteavg      concurrent
hdisk3      00f623c5fbfe9128        siteavg      concurrent
hdisk4      00f623c5fbfe920e        siteavg      concurrent
hdisk5      00f623c5fbfe92cc        sitebvg      concurrent
hdisk6      00f623c5fbfe93a5        sitebvg      concurrent
hdisk7      00f623c5fbfe9440        sitebvg      concurrent
hdisk8      00f623c50b349b68        None         
```

Figure 6-26 Configuring the new PVID

Now we discovered the new PVID from the nodes selecting **Cluster Nodes and Networks → Discover Network Interfaces and Disks**.

Next we added the new PV on the siteavg volume group using C-SPOC by executing **smitty sysmirror** → **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Set**

Characteristics of a Volume Group → Add a Volume from a Volume Group, as shown in Figure 6-27.

Add a Volume to a Volume Group	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
VOLUME GROUP name	[Entry Fields]
Resource Group Name	siteavg
Node List	sitearg
VOLUME names	itsoa1,itsoa2,itsob1,itsob2
Physical Volume IDs	hdisk8 00f623c50b349b68

Figure 6-27 Adding the disk to siteavg

We added the new SVC relationship on the *itso_sitea* consistency group and synchronized the cluster to make the new PVID appear on Site B. Execute **smitty sysmirror** → **Cluster Applications and Resources** → **Resources** → **Configure SVC PPRC-Replicated Resources** → **SVC PPRC Relationships Definition** → **Add an SVC PPRC** and provide the required information, as we did in Figure 6-28.

Add an SVC PPRC Relationship	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
* Relationship Name	[Entry Fields]
* Master VDisk Info	[itso_re17]
* Auxiliary VDisk Info	[itso_data8@SiteA] [itso_data8@SiteB]

Figure 6-28 Adding the PPRC relationship

Then we added it to the *itso_sitea* consistency group via the PowerHA SMIT menu **Cluster Applications and Resources** → **Resources** → **Configure SVC PPRC-Replicated Resources** → **SVC PPRC-Replicated Resource Configuration** → **Change / Show an SVC PPRC Resource**, as shown in Figure 6-29 on page 226.

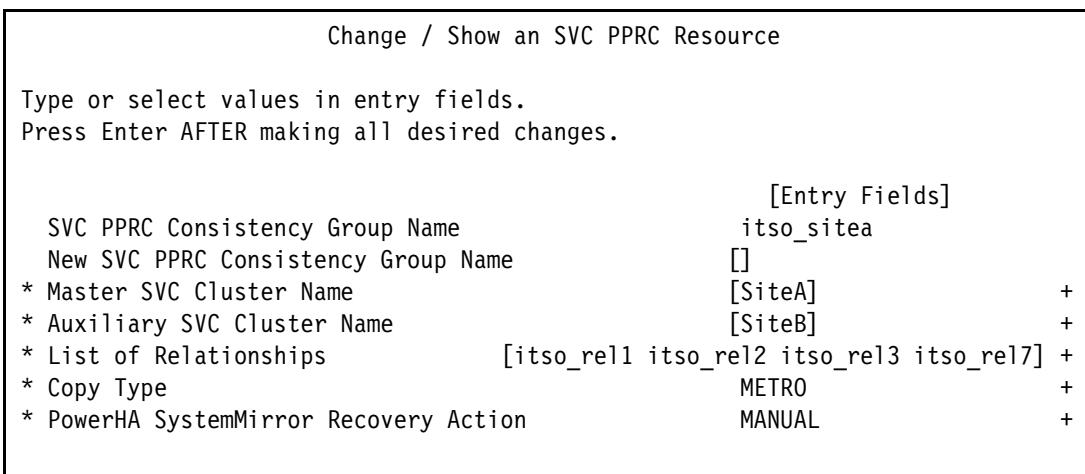


Figure 6-29 Adding the relationship

Now we synchronized the cluster to make all the nodes aware of the changes. To do that just select **Cluster Nodes and Networks** → **Verify and Synchronize Cluster Configuration**.

6.5 Testing the resource group move

We first tested the resource group move from the following:

- ▶ Manual resource group move within the site
- ▶ Manual resource group move between the sites
- ▶ Forced resource group move within the site
- ▶ Forced resource group move between the sites

6.5.1 Resource group move within a site

The initial status of the cluster is shown in Figure 6-30.

root@itsoa1:/>/usr/es/sbin/cluster/utilities/c1RGinfo		
Group Name	State	Node
sitearg	ONLINE	itsoa1@itsoa
	OFFLINE	itsoa2@itsoa
	ONLINE SECONDARY	itsob1@itsob
	OFFLINE	itsob2@itsob
sitebrg	ONLINE	itsob1@itsob
	OFFLINE	itsob2@itsob
	ONLINE SECONDARY	itsoa1@itsoa
	OFFLINE	itsoa2@itsoa

Figure 6-30 Initial cluster status

To start the resource group move, execute `smitty sysmirror` → **System Management (C-SPOC)** → **Resource Group and Applications** → **Move Resource Groups to Another Node**. We selected `sitebrg`, which was online on node `itsob1`, and chose the `itsob2` node as destination node, as shown in Figure 6-31.

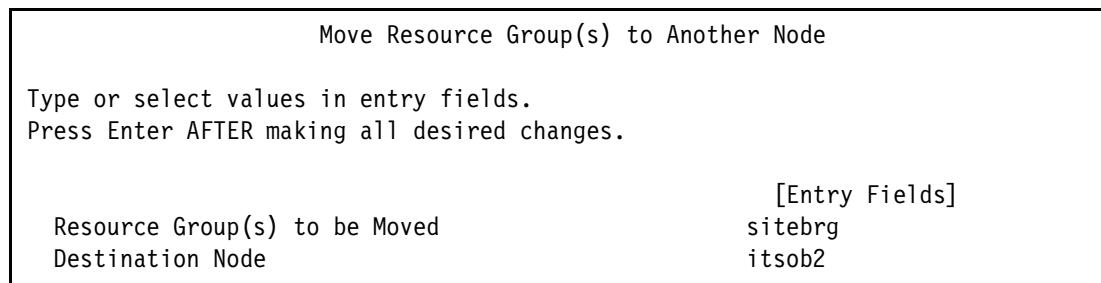


Figure 6-31 Moving a resource group to another node

As shown in Figure 6-32, the resource moved from `itsob1` to `itsob2`.

Group Name	State	Node
sitearg	ONLINE	itsoa1@itsoa
	OFFLINE	itsoa2@itsoa
	ONLINE SECONDARY	itsob1@itsob
	OFFLINE	itsob2@itsob
sitebrg	OFFLINE	itsob1@itsob
	ONLINE	itsob2@itsob
	ONLINE SECONDARY	itsoa1@itsoa
	OFFLINE	itsoa2@itsoa

Figure 6-32 In-site RG move

6.5.2 Resource group move across sites

Now we tested the failover from one site to the other site. Select **System Management (C-SPOC)** → **Resource Group and Applications** → **Move Resource Groups to Another Site** from the PowerHA main menu and choose the `sitebrg` resource group and the `itsoa` site as the destination site, as shown in Figure 6-33.

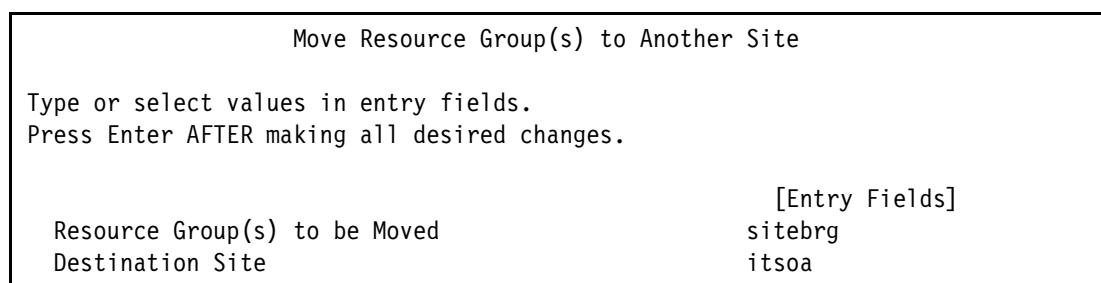


Figure 6-33 Moving the resource group to another site

After a while you can see the resource group on Site A, as shown in Figure 6-34 on page 228.

```

root@itsoa1:/>/usr/es/sbin/cluster/utilities/c1RGinfo
-----
Group Name      State          Node
-----
sitearg        ONLINE         itsoa1@itsoa
                OFFLINE        itsoa2@itsoa
                ONLINE SECONDARY itsoa1@itsob
                OFFLINE        itsoa2@itsob

sitebrg        ONLINE SECONDARY itsoa1@itsob
                OFFLINE        itsoa2@itsob
                ONLINE         itsoa1@itsoa
                OFFLINE        itsoa2@itsoa

```

Figure 6-34 RG state

If you look at the SVC relationship, you see that the direction of the copy has changed as shown in Figure 6-34. The primary site is now the auxiliary one (Figure 6-35).

```

IBM_2145:SiteA:admin>lsrcrelationship itso_rel5
id 55
name itso_rel5
master_cluster_id 0000020060218EE0
master_cluster_name SiteB
master_vdisk_id 50
master_vdisk_name itso_data6
aux_cluster_id 0000020060816D94
aux_cluster_name SiteA
aux_vdisk_id 55
aux_vdisk_name itso_data6
primary aux
consistency_group_id 4
consistency_group_name itso_siteb
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type global
IBM_2145:SiteA:admin>

```

Figure 6-35 Relationship changed direction

6.5.3 Node failure

We executed **halt -q** on node itsoa1 to show the automatic resource group failover. The results of the **halt** command and the resource information (**/usr/es/sbin/cluster/utilities/c1RGinfo**) a minute later on itsoa2 can be seen in Figure 6-36 on page 229.

```

root@itsoa1:/>sync ; sync ; sync ; halt -q
....Halt completed....
```

```

root@itsoa2:/>/usr/es/sbin/cluster/utilities/cLRGinfo
-----
Group Name      State          Node
-----
sitearg        OFFLINE        itsoa1@itsoa
                ONLINE         itsoa2@itsoa
                ONLINE SECONDARY  itsob1@itsob
                OFFLINE        itsob2@itsob

sitebrg        ONLINE         itsob1@itsob
                OFFLINE        itsob2@itsob
                OFFLINE        itsoa1@itsoa
                ONLINE SECONDARY  itsoa2@itsoa

```

Figure 6-36 Resource group moving to secondary on the same site

6.5.4 Site failure

We continued on from the previous section where we already had one node down. We now executed `halt -q` on the `itsoa2` node to simulate a site failure. This resulted in the resource group failing over to Site B, as shown in Figure 6-37.

```

root@itsoa2:/>sync ; sync ; sync ; halt -q
....Halt completed....
```

```

root@itsob2:/>/usr/es/sbin/cluster/utilities/cLRGinfo
-----
Group Name      State          Node
-----
sitearg        OFFLINE        itsoa1@itsoa
                OFFLINE        itsoa2@itsoa
                ONLINE         itsob1@itsob
                OFFLINE        itsob2@itsob

sitebrg        ONLINE         itsob1@itsob
                OFFLINE        itsob2@itsob
                OFFLINE        itsoa1@itsoa
                OFFLINE        itsoa2@itsoa

```

Figure 6-37 Resource group sitearg moved to Site B

We checked the SVC relationship and saw that the copy process was reversed, as shown in Figure 6-38 on page 230. This was because the SVCs were still available at both sites.

```
root@itsob2:/>ssh admin@siteA
Last login: Mon Nov 19 16:21:17 2012 from 129.40.100.92
SiteA:admin>lsrcrelationship itso_re11
id 51
name itso_re11
master_cluster_id 0000020060816D94
master_cluster_name SiteA
master_vdisk_id 51
master_vdisk_name itso_data2
aux_cluster_id 0000020060218EE0
aux_cluster_name SiteB
aux_vdisk_id 46
aux_vdisk_name itso_data2
primary aux
consistency_group_id 3
consistency_group_name itso_sitea
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
SiteA:admin>
```

Figure 6-38 Remote copy reversed

NOTE: As shown in Figure 6-29 on page 226 the **PowerHA SystemMirror Recovery Action** for *sitearg* was defined as *manual*. However, in this scenario we saw that *sitearg* had been acquired automatically to *siteB* when the site fail occurred. This is the expected behavior because the PPRC state was in *consistent_synchronized* during the failover. Discussion about when *manual* or *auto* are affected is covered in 6.6, “Testing storage failure on SVC and V7000 mixed environment” on page 230.

6.6 Testing storage failure on SVC and V7000 mixed environment

In this section we continue testing with storage failures. For comparison, SVC and V7000 mixed environments were used for this environment.

6.6.1 SVC-V7000 mixed environment

This test environment is used in 6.5.4, “Site failure” on page 229 and in “PowerHA 7.1.2 Systems Director Plug-in enhancements” on page 265. The cluster is designed to be a stretched cluster with three different storages, one SVC on Site A, one V7000 on Site B and a DS3K on a third site that can connect to both Site A and Site B. This DS3K only provides the CAA repository disk that should be available to all nodes in a stretched cluster.

The IBM Systems Director server and a DNS server are in the test environment, as shown in Figure 6-39 on page 231.

Replication between sites in a mixed environment

This configuration is presented to show its feasibility and provide test functions in our environment.

V7000 replication: In order to make a V7000 able to replicate with SVC, the layer parameter of the system has to be changed from “storage” (the default configuration) to “replication”. This can be accomplished with the **chsystem -layer replication** command at the V7000 command line prompt.

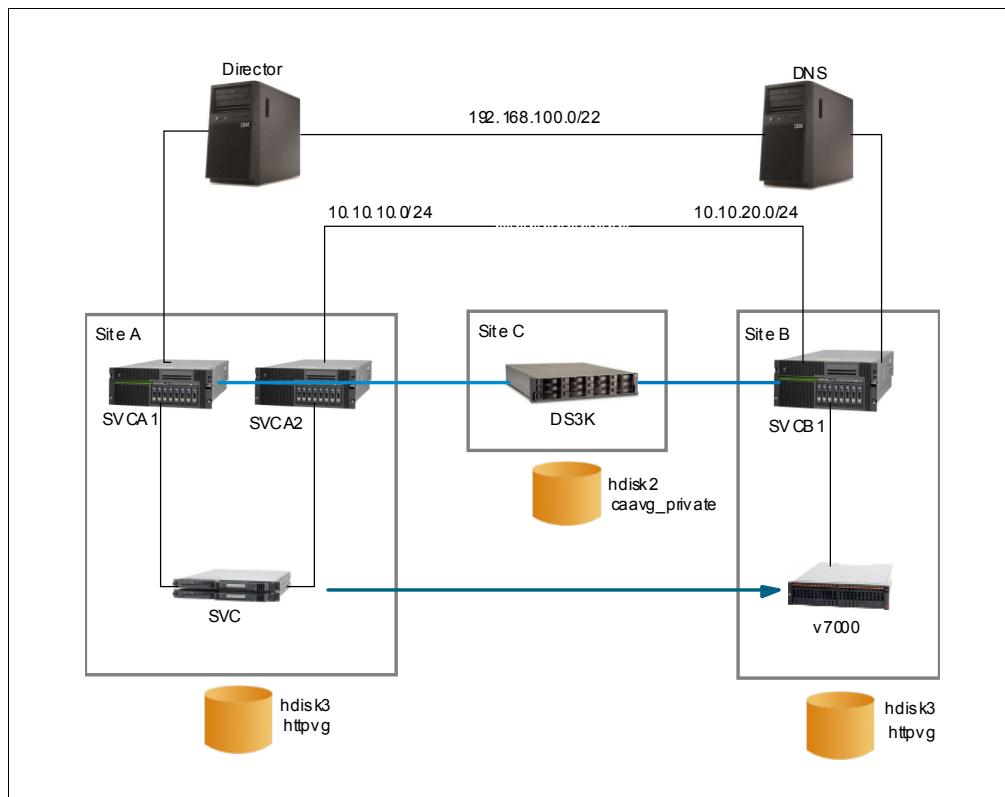


Figure 6-39 Director and SVC/V7000 test environment

The environment is a three-node stretched cluster with two nodes, svca1 and svca2, on siteA and one node, svcb1, on siteB. Two networks, net_ether_01 and net_ether_10. We also had a single resource group, ihs_app_rg with svcb1 configured as the primary site. Example 6-15 shows the cluster topology via the **cltopinfo** command.

Example 6-15 *cltopinfo* output of the *ihs_cluster*

```
#cltopinfo
Cluster Name: ihs_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: hdisk2
Cluster IP Address: 228.168.100.60
There are 3 node(s) and 2 network(s) defined
NODE svca1:
    Network net_ether_01
        svcb_srv      10.10.20.10
```

```

          svca_srv      10.10.10.10
          svca1_pub    10.10.10.11
        Network net_ether_010
          svca1    192.168.100.60
NODE svca2:
        Network net_ether_01
          svcb_srv      10.10.20.10
          svca_srv      10.10.10.10
          svca2_pub    10.10.10.12
        Network net_ether_010
          svca2    192.168.100.61
NODE svcb1:
        Network net_ether_01
          svcb_srv      10.10.20.10
          svca_srv      10.10.10.10
          svcb1_pub    10.10.20.21
        Network net_ether_010
          svcb1    192.168.100.62

Resource Group ihs_app_rg
  Startup Policy   Online On Home Node Only
  Fallback Policy  Fallback To Next Priority Node In The List
  Fallback Policy  Fallback To Higher Priority Node In The List
  Participating Nodes    svcb1 svca1 svca2
  Service IP Label      svca_srv
  Service IP Label      svcb_srv

```

The resource group contains two site-specific single service labels, svca_srv for siteA, and svcb_srv for siteB. A single volume group, httpvg, and an SVC PPRC replicated resource called itso_siteab_grp were also configured. Figure 6-40 shows the **clshowres** output for an outline of the resource group configuration.

# clshowres	
Resource Group Name	ihs_app_rg
Participating Node Name(s)	svcb1 svca1 svca2
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority
	Node In The List
Fallback Policy	Fallback To Higher Priority
	Node In The List
Site Relationship	Prefer Primary Site
Service IP Label	svca_srv svcb_srv
Filesystems	ALL
Filesystems Consistency Check	fsck
Filesystems Recovery Method	sequential
Volume Groups	httpvg
Use forced varyon for volume groups, if necessary	false
Disk Error Management?	no
SVC PPRC Replicated Resources	itso_siteab_grp

Figure 6-40 Clshowres output of ihs_app_rg

When the cluster is activated, the resource group is automatically acquired at svcb1, as shown in Example 6-16 on 233.

Example 6-16 c1RGinfo output of the initial cluster activation

Group Name	Group State	Node
ihs_app_rg	ONLINE	svcb1@SiteB
	ONLINE SECONDARY	svca1@SiteA
	OFFLINE	svca2@SiteA

POWT3V7000, which is the V7000, is configured as the Master role, and ITSO_SVC_CF8, which is the SVC, is configured as the auxiliary role, as shown in Example 6-17.

Example 6-17 V7000 as master and SVC as auxiliary

```
root@svcb1:/usr/es/sbin/cluster/svcpprc/cmds>./cl1ssvc -a
#SVCNAME ROLE SITENAME IPADDR IPADDR2 RPARTNER
POWT3V7000 Master SiteB 9.12.5.6 ITSO_SVC_CF8
ITSO_SVC_CF8 Auxiliary SiteA 9.12.5.67 POWT3V7000
```

6.6.2 Local site SAN failure

We first tested by bringing down the SAN at SiteB. Figure 6-41 illustrates the test.

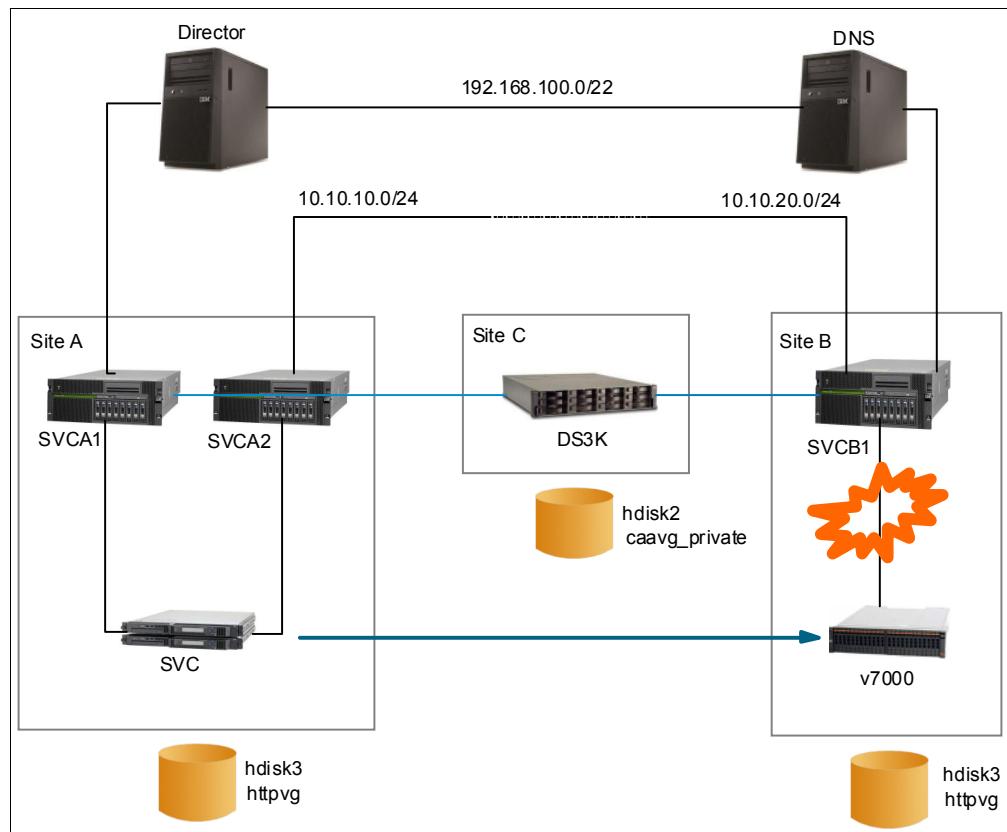


Figure 6-41 Disabling the SAN at SiteB

In this case, the Recover Action is defined as Manual as shown in Example 6-18 on 234.

Example 6-18 SVC PPRC Replicated Resources set to manual

```
root@svca1:/usr/es/sbin/cluster/svcpprc/cmds>./cl1ssvcpprc -a
svcpprc_consistencygrp MasterCluster    AuxiliaryCluster relationships   CopyType      RecoveryAction
itso_siteab_grp          POWT3V7000     ITSO_SVC_CF8       itso_siteab_rel METRO        MANUAL
```

Although the expected result is a site failure, the errlog must log a LVM_SA_QUORCLOSE label for selective failover to initiate. This requires an LVM I/O to indicate a quorum lost.

Tips: For details of selected failover for storage loss, refer to:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.plangd/ha_plan_loss_quorom.htm

In our test scenario, we first created a continuos I/O by writing in the shared volume as shown in Example 6-19.

Example 6-19 Continues I/O creation

```
root@svcb1:/usr/es/sbin/cluster/svcpprc/utils>while true
> do
> cat /etc/hosts > /httpapp/testfile
> sleep 1
> done &
[1] 17498206
```

To disable the SAN at SiteB, we removed the zones on the SAN switch, as shown in Example 6-20.

Example 6-20 Zone removed

```
admin> cfgremove "B32CFG_0_ITSO","svcb1_fcs0_DS3200"
admin> cfgremove "B32CFG_0_ITSO","svcb1_fcs1_DS3200"
admin> cfgenable "B32CFG_0_ITSO"
You are about to enable a new zoning configuration.
This action will replace the old zoning configuration with the
current configuration selected. If the update includes changes
to one or more traffic isolation zones, the update may result in
localized disruption to traffic on ports associated with
the traffic isolation zone changes
Do you want to enable 'B32CFG_0_ITSO' configuration (yes, y, no, n): [no] y
zone config "B32CFG_0_ITSO" is in effect
Updating flash ...
```

Shortly after the zones were disabled, the errlog entries indicating the storage loss were logged, as shown in Example 6-21. We observed that the LVM_SA_QUORCLOSE was logged.

Example 6-21 errlog during the SAN failure

```
root@svcb1:/>errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
5A7598C3 1212225612 I 0 fscsi1 Additional FC SCSI Protocol Driver Infor
5A7598C3 1212225612 I 0 fscsi0 Additional FC SCSI Protocol Driver Infor
AEA055D0 1212225612 I S livedump Live dump complete
AEA055D0 1212225612 I S livedump Live dump complete
CAD234BE 1212225612 U H LVDD QUORUM LOST, VOLUME GROUP CLOSING
```

52715FA5	1212225612 U H LVDD	FAILED TO WRITE VOLUME GROUP STATUS AREA
E86653C3	1212225612 P H LVDD	I/O ERROR DETECTED BY LVM
CAD234BE	1212225612 U H LVDD	QUORUM LOST, VOLUME GROUP CLOSING
F7DDA124	1212225612 U H LVDD	PHYSICAL VOLUME DECLARED MISSING
52715FA5	1212225612 U H LVDD	FAILED TO WRITE VOLUME GROUP STATUS AREA
E86653C3	1212225612 P H LVDD	I/O ERROR DETECTED BY LVM
C62E1EB7	1212225612 P H hdisk3	DISK OPERATION ERROR
C62E1EB7	1212225612 P H hdisk3	DISK OPERATION ERROR
CB4A951F	1212225612 I S SRC	SOFTWARE PROGRAM ERROR
AA8AB241	1212225612 T O clevmgrd	OPERATOR NOTIFICATION
AA8AB241	1212225612 T O clevmgrd	OPERATOR NOTIFICATION
B6DB68E0	1212225612 I O SYSJ2	FILE SYSTEM RECOVERY REQUIRED
78ABDDEB	1212225612 I O SYSJ2	META-DATA I/O ERROR
C1348779	1212225612 I O SYSJ2	LOG I/O ERROR
E86653C3	1212225612 P H LVDD	I/O ERROR DETECTED BY LVM
C62E1EB7	1212225612 P H hdisk3	DISK OPERATION ERROR
EA88F829	1212225612 I O SYSJ2	USER DATA I/O ERROR
E86653C3	1212225612 P H LVDD	I/O ERROR DETECTED BY LVM
C62E1EB7	1212225612 P H hdisk3	DISK OPERATION ERROR
DE3B8540	1212225612 P H hdisk3	PATH HAS FAILED
D5676F6F	1212225612 T H fscsi0	ATTACHED SCSI TARGET DEVICE ERROR
D5676F6F	1212225612 T H fscsi1	ATTACHED SCSI TARGET DEVICE ERROR

```
root@svcb1:/>errpt -a
```

LABEL: LVM_SA_QUORCLOSE
 IDENTIFIER: CAD234BE

Date/Time: Wed Dec 12 22:56:39 2012
 Sequence Number: 55815
 Machine Id: 00F6F5D04C00
 Node Id: svcb1
 Class: H
 Type: UNKN
 WPAR: Global
 Resource Name: LVDD
 Resource Class: NONE
 Resource Type: NONE
 Location:

Description
 QUORUM LOST, VOLUME GROUP CLOSING

Probable Causes
 PHYSICAL VOLUME UNAVAILABLE

Detail Data
 MAJOR/MINOR DEVICE NUMBER
 8000 0032 0000 0000
 QUORUM COUNT
 2
 ACTIVE COUNT
 65535
 SENSE DATA

```
0000 0000 0000 0A77 00F7 0C99 0000 4C00 0000 013B 2185 BE30 00F7 0C99 2185 BDE3
```

From hacmp.out we observed a selective failover upon volume group access loss, as shown in Example 6-22.

Example 6-22 Hacmp.out during failover

HACMP Event Preamble

Enqueued rg_move release event for resource group 'ihs_app_rg'.

Reason for recovery of Primary instance of Resource group 'ihs_app_rg' from TEMP_ERROR state on node 'svcb1' was 'Volume group failure'.

Enqueued rg_move secondary acquire event for resource group 'ihs_app_rg'.

Enqueued rg_move acquire event for resource group 'ihs_app_rg'.

Enqueued rg_move secondary release event for resource group 'ihs_app_rg'.

Cluster Resource State Change Complete Event has been enqueued.

Shortly afterward, the resource groups were activated at SiteA, as shown in Example 6-23.

Example 6-23 Resource group movement

```
root@svcb1:/>/usr/es/sbin/cluster/utilities/c1RGinfo
```

Group Name	Group State	Node
ihs_app_rg	ONLINE SECONDARY	svcb1@siteB
	ONLINE	svca1@siteA
	OFFLINE	svca2@siteA

The SVC replication was switched to aux with the connected_synchronized state, as shown in Example 6-24.

Example 6-24 PPRC state after SAN failure

```
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary aux
consistency_group_id 0
consistency_group_name itso_siteab_grp
state consistent_synchronized
```

```

bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name
-----
root@svcb1:/>ssh admin@ITSO_SVC_CF8
IBM_2145:ITSO_SVC_CF8:admin>lsrcrelationship itso_siteab_rel
id 27
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary aux
consistency_group_id 5
consistency_group_name itso_siteab_grp
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name

```

In summary, although Recovery Action is set to Manual, in the event of a SAN failure on the local sites, the failover occurs automatically.

6.6.3 PPRC link failure

The final test is done by cutting the PPRC links. In this scenario, we tested the behavior during an inconsistent state of the PPRC.

Creating the inconsistent state

We created the inconsistent state by the following steps:

1. Cut the PPRC connection between the SVC and V7000.

- Updated the disks on the SVC node.

Figure 6-42 illustrates the overview of this test.

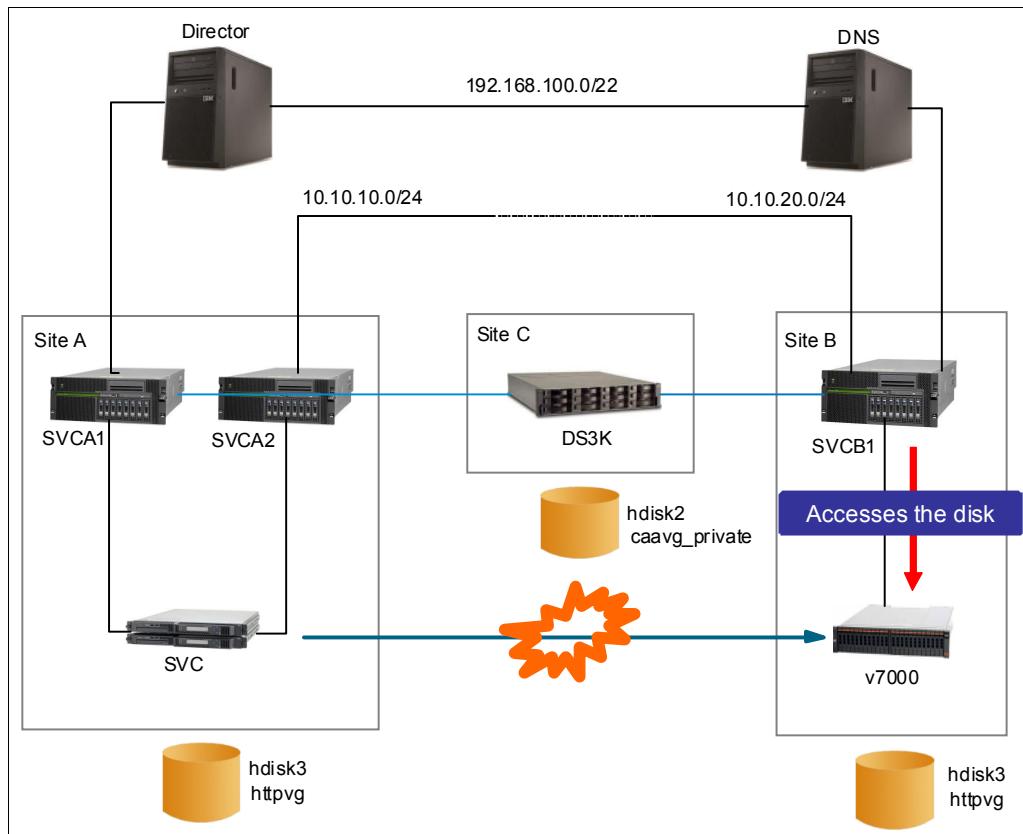


Figure 6-42 PPRC link failure

To cut the PPRC links, we removed the zones for the PPRC connections, as shown in Example 6-25.

Example 6-25 Removing the zones for the PPRC connection

```
admin> cfgremove "B32CFG_0_ITSO", "SVC_CF8_V7K_ITSO"
admin> cfgenable "B32CFG_0_ITSO"
You are about to enable a new zoning configuration.
This action will replace the old zoning configuration with the
current configuration selected. If the update includes changes
to one or more traffic isolation zones, the update may result in
localized disruption to traffic on ports associated with
the traffic isolation zone changes
Do you want to enable 'B32CFG_0_ITSO' configuration (yes, y, no, n): [no] y
zone config "B32CFG_0_ITSO" is in effect
Updating flash ...
```

Almost immediately, POWT3V7000, which is the master, reported the idling_disconnected state, as shown in Example 6-26.

Example 6-26 Copy state of POWT3V7000

```
root@svcb1:/>ssh admin@POWT3V7000
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
```

```
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary master
consistency_group_id 0
consistency_group_name itso_siteab_grp
state idling_disconnected
bg_copy_priority 50
progress
freeze_time
status
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name
```

ITSO_SVC_CF8, as the auxiliary role, reports the consistent_disconnected state as shown in Example 6-27. In this state, the VDisks in this half of the consistency group are all operating in the secondary role and can accept read I/O operations, but not write I/O operations.

Example 6-27 Copy state of ITSO_SVC_CF8

```
root@svcb1:/>ssh admin@ITSO_SVC_CF8
IBM_2145:ITSO_SVC_CF8:admin>lsrcrelationship itso_siteab_rel
id 27
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary master
consistency_group_id 5
consistency_group_name itso_siteab_grp
state consistent_disconnected
bg_copy_priority 50
progress
freeze_time 2012/12/04/17/31/08
status
sync
copy_type metro
cycle_period_seconds 300
```

```
cycling_mode none  
master_change_vdisk_id  
master_change_vdisk_name  
aux_change_vdisk_id  
aux_change_vdisk_name
```

TIPS: For a complete description of each status of the SVC, refer to Troubleshooting PowerHA SystemMirror Enterprise Edition for Metro Mirror for SVC at:

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/ha_pprc_trouble_svc.htm

We then updated the disk at siteB with the **dd** command, as shown in Example 6-28.

Example 6-28 Updating the disks at siteB

```
root@svcb1:/>lsvg -o  
httpvg  
caavg_private  
rootvg  
root@svcb1:/>lsvgfs httpvg  
/httpapp  
root@svcb1:/>dd if=/dev/zero of=/httpapp/test_file bs=1m count=10  
10+0 records in  
10+0 records out  
root@svcb1:/>ls -l /httpapp/test_file  
-rw-r--r--    1 root      system   10485760 Dec  4 20:49 /httpapp/test_file
```

In this state, we observed the behavior when the Recovery Action was defined as manual or auto.

Failing sites with automatic recovery

First, we tested the Recovery Action with auto for the SVC PPRC replicated resources, as shown in Example 6-29.

Example 6-29 SVC PPRC replicated resources set to auto

```
/usr/es/sbin/cluster/svcpprc/cmds>./cl1ssvcpprc -a  
svcpprc_consistencygrp MasterCluster AuxiliaryCluster relationships CopyType RecoveryAction  
itso_siteab_grp      POWT3V7000     ITSO_SVC_CF8    itso_siteab_rel METRO      AUTO
```

We took down the primary site with the **halt** command, as shown in Example 6-30.

Example 6-30 Halting svcb

```
root@svcb1:/>sync ; sync ; sync ; halt -q  
....Halt completed....
```

Shortly after, the site failure occurred and the resource group was acquired at the secondary site, as shown in Example 6-31.

Example 6-31 Resource failover to svca1

```
root@svca1:/>/usr/es/sbin/cluster/utilities/clRGinfo
```

Group Name	Group State	Node
------------	-------------	------

ihs_app_rg	OFFLINE	svcb1@SiteB
	ONLINE	svca1@SiteA
	OFFLINE	svca2@SiteA

On node svca1 the replicated disks are varied on and mounted as shown in Example 6-32. This output also shows that the file created in Example 6-28 on 240 does *not* appear since svca1 is now accessing the outdated disk data contained on the disk prior to the PPRC link failure.

Example 6-32 Disk access of svca1

```
root@svca1:/>lsvg -o
httpvg
caavg_private
rootvg
root@svca1:/>lsvgfs httpvg
/httpapp
root@svca1:/>mount | grep httpapp
/dev/applv      /httpapp          jfs2   Dec 04 21:01
rw,log=/dev/apploglv
root@svca1:/>ls -l /httpapp/test_file
/httpapp/test_file not found
```

During the failover, POWT3V7000 reports the idling_disconnected state as shown in Example 6-33. The state has not changed from Example 6-26 on 238. This means that read and write access are still allowed at siteA.

Example 6-33 Copy state of POWT3V7000 after failover

```
root@svca1:/>ssh admin@POWT3V7000
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary master
consistency_group_id 0
consistency_group_name itso_siteab_grp
state idling_disconnected
bg_copy_priority 50
progress
freeze_time
status
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
```

```
aux_change_vdisk_id  
aux_change_vdisk_name  
IBM_2076:POWT3V7000:admin>exit  
exit  
Connection to POWT3V7000 closed.
```

ITSO_SVC_CF8 now reports an idling_disconnected state as shown in Example 6-34, which now permits write access from nodes in siteB.

Example 6-34 Copy state of ITSO_SVC_CF8 after failover

```
root@svca1:/>ssh admin@ITSO_SVC_CF8  
IBM_2145:ITSO_SVC_CF8:admin>lsrcrelationship itso_siteab_rel  
id 27  
name itso_siteab_rel  
master_cluster_id 00000200A0205BAF  
master_cluster_name POWT3V7000  
master_vdisk_id 55  
master_vdisk_name itso_siteb  
aux_cluster_id 0000020060C17848  
aux_cluster_name ITSO_SVC_CF8  
aux_vdisk_id 27  
aux_vdisk_name itso_sitea  
primary  
consistency_group_id 5  
consistency_group_name itso_siteab_grp  
state idling_disconnected  
bg_copy_priority 50  
progress  
freeze_time  
status  
sync  
copy_type metro  
cycle_period_seconds 300  
cycling_mode none  
master_change_vdisk_id  
master_change_vdisk_name  
aux_change_vdisk_id  
aux_change_vdisk_name
```

In summary, in case of total site failure and having set the Recovery actions to auto, PowerHA automatically acquires the resource group regardless of the state of the data copy. This minimizes recovery time after an outage. However, you must be aware that applications and services may use inconsistent or outdated data copy. Also, in case of cluster partition, both primary and secondary sites are still able to access their own copies and the data will diverge. This may lead to serious data corruption.

When configuring the Recovery Action with auto, you are prompted to be aware of this possibility, as shown in Figure 6-43 on page 243. For these reasons we do not suggest setting the PowerHA Recovery actions to auto.

Recovery Action is currently set to "AUTO" and corresponds to the recovery that would be normally expected on a site failure. However, it must be recognized that PowerHA cannot distinguish between a catastrophic failure of the entire primary site, and a situation where all the links between sites are severed - a partitioned cluster. If the latter case happens, and automatic takeover has been chosen, then both the primary and backup sites will attempt to run their own instances of the application server, with their local copies of the data. In such a situation, the copies of the data at the primary and backup sites will soon diverge - they will no longer be exact copies of each other. Correcting this data divergence can be expected to be a difficult, expensive and time consuming operation, with no guarantee of success.

A partitioned cluster cannot be prevented, but it can be made unlikely by having multiple independent heart beat paths between sites. Care should be given to ensuring that the separate heart beat paths do not travel through common conduits or routers, or any such common physical or logical component whose loss could cause all heart beat paths to fail. (It is IBM's intention to develop and make available technologies in future releases of PowerHA SystemMirror Enterprise Edition that will indeed preclude a partitioned cluster.) Additionally, if the disk subsystem supports it, there should be Copy Services Servers at each site.

Enter "y" to continue or "n" to abort [n]: Y

Figure 6-43 Prompts to set recovery action to normal

Failing sites with manual recovery

Now we tested Recovery Action with manual for the SVC PPRC replicated resources as shown in Example 6-35.

Example 6-35 SVC PPRC replicated resources set to manual

					RecoveryAction
svcpprc_consistencygrp	MasterCluster	AuxiliaryCluster	relationships	CopyType	MANUAL
itso_siteab_grp	POWT3V7000	ITSO_SVC_CF8	itso_siteab_rel	METRO	

We took down the primary site with the **halt** command shown in Example 6-36.

Example 6-36 Halting svcb1

```
root@svcb1:/>sync ; sync ; sync ; halt -q
....Halt completed....
```

Shortly after, the site failure occurred. However, resource groups are not acquired in any of the nodes. The **C1RGinfo** command with the **-p** flag shows Primary State as ERROR on both of the nodes at siteA, as shown in Example 6-37.

Example 6-37 C1RGinfo after failover

```
> c1RGinfo -p

Cluster Name: ihs_cluster

Resource Group Name: ihs_app_rg
Node           Primary State   Secondary State
-----
```

svcb1@SiteB	OFFLINE	OFFLINE
svca1@SiteA	ERROR	ONLINE SECONDARY
svca2@SiteA	ERROR	OFFLINE

During the failover, POWT3V7000 reports the idling_disconnected state as shown in Example 6-33 on 241. The state has not changed from Example 6-38.

Example 6-38 Copy state of POWT3V7000 after failover

```
root@svca1:>ssh admin@POWT3V7000
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary master
consistency_group_id 0
consistency_group_name itso_siteab_grp
state idling_disconnected
bg_copy_priority 50
progress
freeze_time
status
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name
```

Compared to ITSO_SVC_CF8, which remains in consistent_disconnected state as shown in Example 6-39, nodes at siteB are not able to access their disks.

Example 6-39 Copy state of ITSO_SVC_CF8 after failover

```
root@svca1:/usr/es/sbin/cluster/svcpprc/cmds>ssh admin@ITSO_SVC_CF8
IBM_2145:ITSO_SVC_CF8:admin>lsrcrelationship itso_siteab_rel
id 27
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary master
```

```
consistency_group_id 5
consistency_group_name itso_siteab_grp
state consistent_disconnected
bg_copy_priority 50
progress
freeze_time 2012/12/04/18/54/02
status
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name
IBM_2145:ITSO_SVC_CF8:admin>
```

On node svca1 in the /var/hacmp/log/hacmp.out file, a list of suggested steps is shown; see Example 6-40.

Example 6-40 hacmp.out file

RECOMMENDED USER ACTIONS:

We are at this stage because both HACMP and SVC links are DOWN (scenario (b)). It is the responsibility of the user to check
if the production site is active or down

STEP 1: Verify if HACMP Production site nodes are UP or DOWN
STEP 2: Verify if Production site SVC is UP or DOWN

Case 1) Production site SVC and HACMP nodes are both DOWN.

STEP 3: On the SVC GUI, check the states of the SVC consistency groups
If the consistency groups at the remote site are in "consistent_disconnected" state,
Select the consistency groups and run the "stoprcconsistgrp -access" command
against them to enable I/O to the backup VDisks.

```
ssh -n admin@9.12.5.6 svctask stoprcconsistgrp -access itso_siteab_grp
```

NOTE: If the consistency groups are in any other state please consult the
IBM Storage Systems SVC documentation for what further instructions are needed

STEP 4: Wait until the consistency groups state is "idling_disconnected"

STEP 5: Using smitty hacmp select the node you want the RG to be online at.
smitty hacmp -> System Management (C-SPOC) -> HACMP Resource Group and Application Management ->
Bring a Resource Group Online
for the node where you want the RG to come online
Once this completes the RG should be online on the selected site.

Case 2) If Production site SVC cluster and HACMP nodes are both UP.

STEP 3: On the SVC GUI, check the states of the SVC consistency groups
If the consistency groups at the remote site are in "consistent_disconnected" state,
Select the consistency groups and run the "stoprcconsistgrp -access" command.

```
ssh -n admin@9.12.5.6 svctask stoprcconsistgrp -access itso_siteab_grp
```

NOTE: If the consistency groups are in any other state, please consult the IBM Storage Systems SVC documentation for what further actions are needed

Wait until the consistency groups state is "idling_disconnected" you can check them with the following command.

```
ssh -n admin@9.12.5.6 svcinfo lsrrcconsistgrp -delim : itso_siteab_grp  
ssh -n admin@9.12.5.67 svcinfo lsrrcconsistgrp -delim : itso_siteab_grp
```

STEP 4: Check and re-connect all physical links (HACMP and SVC links).

On the SVC GUI, check the states of the SVC consistency groups

If the consistency groups are in "idling" state,
determine which HACMP site that the RG should be coming online. Once you do that run
`/usr/es/sbin/cluster/svcpprc/cmds/cllssvc -ah`

so if you want the RG's online on the secondary site you would pick 9.12.5.67 which is the auxiliary

So next run this command for all the Consistency groups so we restart them in the correct direction. This example is to use the aux site.. however if you wanted to use the master site you should change -primary aux to be -primary master.

```
ssh -n admin@9.12.5.6 svctask startrcconsistgrp -force -primary [ aux | master ]  
itso_siteab_grp
```

If the cluster links are not up we will get the error:

CMMVC5975E The operation was not performed because the cluster partnership is not connected.

NOTE: If the consistency groups are in any other state please consult the IBM Storage Systems SVC documentation for further instructions.

STEP 5: Wait until the consistency groups state is "consistent_synchronized" or "inconsistent_copying".

Run the following commands and check the consistency group state.

```
ssh -n admin@9.12.5.6 svcinfo lsrrcconsistgrp -delim : itso_siteab_grp  
ssh -n admin@9.12.5.67 svcinfo lsrrcconsistgrp -delim : itso_siteab_grp
```

Sample output from one of the above commands:

```
id:255  
name:FVT_CG1  
master_cluster_id:00000200648056E6  
master_cluster_name:HACMPSVC1  
aux_cluster_id:0000020061401C7A  
aux_cluster_name:HACMPSVC2  
primary:aux  
state:consistent_synchronized  
relationship_count:1  
freeze_time:  
status:online  
sync:  
copy_type:global  
RC_rel_id:98  
RC_rel_name:FVT_REL1
```

Note:

Be sure before you fix the connection from the primary HACMP to secondary HACMP site nodes that you make sure only

one site is running HACMP with the resource groups online. If they both have the resources when the network connection is repaired one of the 2 sites will be halted.

STEP 6:

Using smitty hacmp select the node you want the RG to be online at.

smitty hacmp -> System Management (C-SPOC) -> HACMP Resource Group and Application Management ->

Bring a Resource Group Online

for the node where you want the RG to come online

Once this completes the RG should be online on the selected site.

END RECOMMENDED USER ACTIONS:

To recover from this situation, first we stopped the remote copy to make the disks at ITSO_SVC_CF8 accessible. This can be achieved with the **stoprcconsistgrp** command as shown in Example 6-41. This changes the copy state to idling_disconnected.

Example 6-41 Stoprcconsistgrp command

```
root@svcal1:/usr/es/sbin/cluster/svcpprc>ssh -n admin@9.12.5.67 svctask  
stoprcconsistgrp -access itso_siteab_grp  
root@svcal1:/usr/es/sbin/cluster/svcpprc>ssh admin@ITSO_SVC_CF8  
Last login: Tue Dec  4 23:16:04 2012 from 9.12.5.7  
IBM_2145:ITSO_SVC_CF8:admin>lsrcrelationship itso_siteab_rel  
id 27  
name itso_siteab_rel  
master_cluster_id 00000200A0205BAF  
master_cluster_name POWT3V7000  
master_vdisk_id 55  
master_vdisk_name itso_siteb  
aux_cluster_id 0000020060C17848  
aux_cluster_name ITSO_SVC_CF8  
aux_vdisk_id 27  
aux_vdisk_name itso_sitea  
primary  
consistency_group_id 5  
consistency_group_name itso_siteab_grp  
state idling_disconnected  
bg_copy_priority 50  
progress  
freeze_time  
status  
sync  
copy_type metro  
cycle_period_seconds 300  
cycling_mode none  
master_change_vdisk_id  
master_change_vdisk_name  
aux_change_vdisk_id  
aux_change_vdisk_name  
IBM_2145:ITSO_SVC_CF8:admin>
```

We then activated the resource group (Figure 6-44) by executing **smitty sysmirror** → **System Management (C-SPOC)** → **Resource Group and Application** → **Bring a Resource Group Online**.

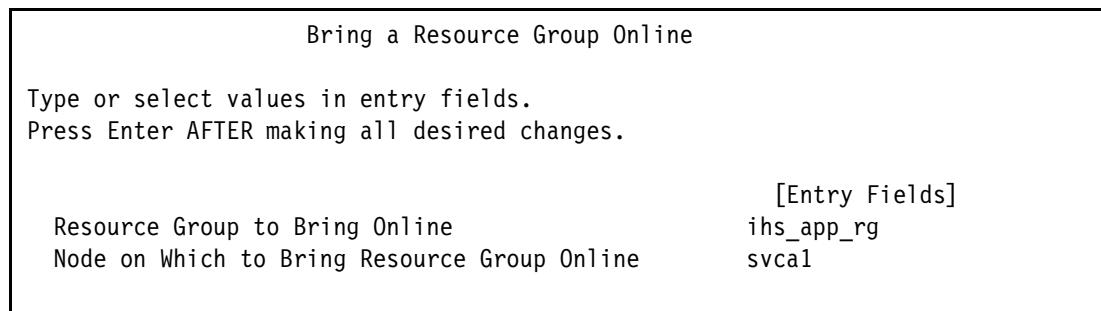


Figure 6-44 Activating the resource group

The resource group was now activated, as shown in Example 6-42.

Example 6-42 CIRGinfo output after activating the cluster

```
# clRGinfo -p
```

Cluster Name:	ihs_cluster	
Resource Group Name:	ihs_app_rg	
Primary instance(s):		
The following node temporarily has the highest priority for this instance: svca1, user-requested rg_move performed on Wed Dec 5 02:16:15 2012		
Node	Primary State	Secondary State
svcb1@SiteB	OFFLINE	OFFLINE
svca1@SiteA	ONLINE	OFFLINE
svca2@SiteA	OFFLINE	OFFLINE

In summary, in case of total site failure and having set the Recovery actions to manual, the resource group is *not* acquired if the copy is in a *not synchronized* state. User interaction is required to recover the resource group. Although this may lead to more down time, this minimizes the risk for applications accessing inconsistent data.

Recovering from a PPRC link failure

This section covers the steps after the PPRC link failure. In both **manual** or **auto** recovery scenarios first, we recovered the physical PPRC links. In our scenario we re-added the zones previously removed in Example 6-25 on 238. Shortly after the recovery of the physical PPRC link, the copy state became **idle**, as shown in Example 6-43.

Example 6-43 Copy state after recovering the physical PPR connection

```
root@svca1:/usr/es/sbin/cluster/svcpprc>ssh admin@ITSO_SVC_CF8
Last login: Tue Dec 4 23:16:50 2012 from 9.12.5.7
IBM_2145:ITSO_SVC_CF8:admin>lsrcrelationship itso_siteab_rel
id 27
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
```

```

master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary
consistency_group_id 5
consistency_group_name itso_siteab_grp
state idling
bg_copy_priority 50
progress
freeze_time
status
sync out_of_sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name

root@svcal:/usr/es/sbin/cluster/svcpprc>ssh admin@POWT3V7000
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary
consistency_group_id 0
consistency_group_name itso_siteab_grp
state idling
bg_copy_priority 50
progress
freeze_time
status
sync out_of_sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name

```

We re-enabled the remote copy with the **startrcconsistgrp** command, as shown in Example 6-44 on 250. Remember to set the primary copy with the **-primary** flag. In our

scenario, because the disk access was done from the secondary site, we chose **aux**. Wait until the copy state becomes **consistent_synchronized**.

Example 6-44 Re-enabling the remote copy

```
IBM_2076:POWT3V7000:admin>startrcconsistgrp -force -primary aux itso_siteab_grp
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary aux
consistency_group_id 0
consistency_group_name itso_siteab_grp
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name
```

To move the resource group back to the siteA, first ensure that the cluster is stable on all nodes by executing **1ssrc -ls clstrmgrES**, as shown in Example 6-45.

Example 6-45 Confirming the current cluster state

```
root@svca1:/httpapp>clcmd 1ssrc -ls clstrmgrES | grep state
Current state: ST_STABLE
Current state: ST_STABLE
Current state: ST_STABLE
root@svca1:/httpapp>c1RGinfo
-----
Group Name      Group State          Node
-----
ihs_app_rg     ONLINE SECONDARY    svcb1@SiteB
                ONLINE             svca1@SiteA
                OFFLINE            svca2@SiteA
```

Move the resource group with **smitty sysmirror** → **System Management (C-SPOC)** → **Resource Groups and Applications** → **Move Resource Groups to Another Site**. Choose the resource group to move, and the site, as shown in Figure 6-45 on page 251.

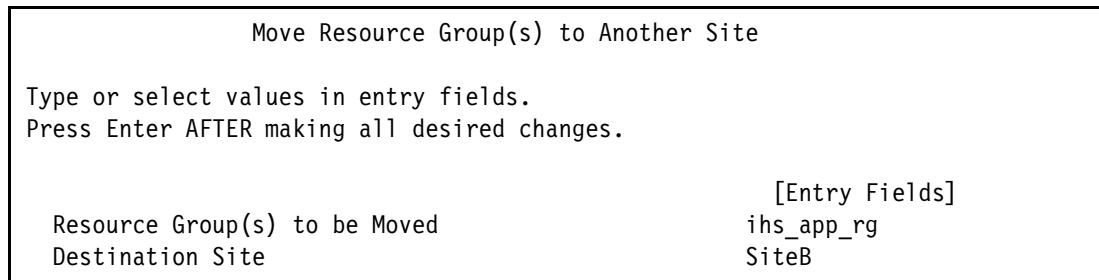


Figure 6-45 Moving the resource group

Upon completion, the resource group is acquired and the copy process direction is reversed, as shown in Example 6-46.

Example 6-46 Resource group state after moving to the sites

```
root@svca2:/var/hacmp/log>c1RGinfo
-----
Group Name      Group State          Node
-----
ihs_app_rg     ONLINE               svcb1@SiteB
                ONLINE SECONDARY      svca1@SiteA
                OFFLINE              svca2@SiteA

root@svca2:/var/hacmp/log>ssh admin@POWT3V7000
IBM_2076:POWT3V7000:admin>lsrcrelationship itso_siteab_rel
id 55
name itso_siteab_rel
master_cluster_id 00000200A0205BAF
master_cluster_name POWT3V7000
master_vdisk_id 55
master_vdisk_name itso_siteb
aux_cluster_id 0000020060C17848
aux_cluster_name ITSO_SVC_CF8
aux_vdisk_id 27
aux_vdisk_name itso_sitea
primary master
consistency_group_id 0
consistency_group_name itso_siteab_grp
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
cycle_period_seconds 300
cycling_mode none
master_change_vdisk_id
master_change_vdisk_name
aux_change_vdisk_id
aux_change_vdisk_name
```

6.7 PowerHA SVC PPRC commands

The following are specific commands for administrating and managing SVC PPRC:

- ▶ **c1ssvc** - List SVC cluster information.

http://pic.dhe.ibm.com/infocenter/aix/v7r1/index.jsp?topic=%2Fcom.ibm.aix.powerha_pprc%2Fha_pprc_trouble_svc.htm

- ▶ **c1ssvcpprc** - List information about all SVC PPRC resources or a specific SVC PPRC resource.

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.hacmp.pprc/ha_pprc_c1ssvcpprc.htm

- ▶ **c1srelationship** - List information about all SVC PPRC relationships or a specific PPRC relationship.

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.hacmp.pprc/ha_pprc_c1srelationship.htm

- ▶ **c1_verify_svccpprc_config** - Verifies the SVC definition in the PowerHA SystemMirror configuration.

http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.hacmp.pprc/ha_pprc_c1verifysvccpprcconfig.htm

6.8 IBM PowerHA SystemMirror and SVC global mirror failover demonstration

The following is a demonstration utilizing IBM PowerHA SystemMirror and SVC global mirror for the replication:

<http://www-03.ibm.com/support/techdocs/atstr.nsf/WebIndex/PRS4594>



Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replication

In this chapter we provide details and practical scenarios on how to configure PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replication. We include the following topics for implementing XIV replication with PowerHA SystemMirror Enterprise Edition (EE) 7.1.2:

- ▶ Introduction to XIV remote mirroring
- ▶ Planning PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV remote mirroring
- ▶ Preliminary steps for XIV storage configuration
- ▶ Installing the XIV CLI
- ▶ Creating an XIV replicated cluster using SMIT and XIV GUI
- ▶ Creating an XIV replicated cluster via clmgr and XIV CLI
- ▶ Administrating the cluster with XIV replicated resources
- ▶ IBM PowerHA SystemMirror demonstration with XIV replication

7.1 Introduction to XIV remote mirroring

The remote mirroring function of the XIV storage system provides a real-time copy between two or more XIV storage systems supported over Fibre Channel (FC) or iSCSI links. This feature provides a method to protect data from storage and site failures.

XIV remote mirroring has two options for replicating the data between two storage systems:

- Synchronous** Write operations are completed on both copies (local and remote sites) before they are considered to be complete. This type of remote mirroring is normally used for short distances to minimize the effect of I/O delays inherent to the distance to the remote site.
- Asynchronous** Consistent sets of data are copied to the remote location at specified intervals and host I/O operations are complete after writing to the primary storage. This is typically used for long distances between sites.

For a detailed and practical understanding of the XIV remote mirroring features, refer to *IBM XIV Storage System: Copy Services and Migration*, SG24-7759-02.

XIV remote mirroring terminology

While explaining various configuration details about XIV remote mirroring, we used a couple of terms and concepts specific to XIV systems. A number of terms, meanings, and usages with regards to XIV and synchronous remote mirroring are as follows:

Local site	This site contains the XIV storage along with the production servers and applications using the XIV storage system.
Remote site	This site holds the mirror copy of the data from the primary storage on another XIV storage system. The remote site is capable of becoming the active production site with consistent data available in the event of a failure at the local site.
Primary storage	It is the XIV system designated under normal conditions to serve hosts and have its data replicated to a secondary XIV for disaster recovery purposes. It is the storage associated with the local site.
Secondary storage	It is the XIV system designated under normal conditions to act as the mirror (backup) for the primary, and that could be set to replace the primary if the primary fails. It is the storage associated with the remote site.
Consistency group	A consistency group (CG) is a set of related volumes on the same XIV storage system that are treated as a single consistent unit. Consistency groups are supported in remote mirroring.
Coupling	This is the pairing of volumes or consistency groups (CGs) to form a mirror relationship between the source of the replication (master) and the target (slave).
Peer	This is one side of a coupling. It can either be a volume or a consistency group, but the peers forming a coupling must be of the same type.
Role	Each peer has a role within the coupling and this can be: <i>Master</i> : A role that indicates that the peer serves host requests and acts as the source for replication. While a peer has the master role, read/write operations are allowed on the storage volume for the associated hosts. Changing a peer's role to master from slave may be

warranted after a disruption of the current master's service either due to a disaster or to planned service maintenance.

Slave: A role that indicates that the peer does not serve host requests and acts as the target for replication. When the role of the peer is slave, the volume will have a lock on it and only read operations are allowed for the hosts attaching the peer volume. Changing a peer's role to slave from master may be warranted after the peer is recovered from a site or system or link failure or disruption that led to the promotion of the other peer from slave to master. Changing roles can also be done in preparation for supporting a planned service maintenance.

Recovery Point Objective

As a general term RPO refers to a maximal point-in-time difference between the image of data in the sites. This term is specific to an asynchronous replication relationship and shows how far behind the secondary image of data is compared with the source data. For a synchronous replication RPO is always zero.

In XIV RPO is a setting applicable to asynchronous mirroring that represents an objective set by the user implying the maximal currency difference considered acceptable between the mirror peers (the actual difference between mirror peers can be shorter or longer than the RPO set). The XIV system then reports the effective RPO and compares it to the required RPO. The status of the mirroring peers will then be reflected to the user in relationship with the user RPO setting. Connectivity, bandwidth, and distance between the XIV system that contains the production volume and the XIV system that contains the replicated copy directly impact the RPO. More connectivity, greater bandwidth, and less distance typically enable a lower RPO.

Sync job

This applies to async mirroring only. It denotes a synchronization procedure run by the master at specified user-configured intervals corresponding to the asynchronous mirroring definition or upon manual execution of a dedicated XCLI command (the related command is `mirror_create_snapshot`). The resulting job is dubbed snapshot mirror sync job or ad-hoc sync job, or manual sync job in contrast with a scheduled sync job. The sync job entails synchronization of data updates recorded on the master since the creation time of the most recent snapshot that was successfully synchronized.

Asynchronous schedule interval

This applies to asynchronous mirroring only. It represents, per given coupling, how often the master automatically runs a new sync job. For example, if the pertinent mirroring configuration parameter specifies a 60-minute interval, then during a period of one day, 24 sync jobs will be created.

Mirroring status

The status of a mirror is affected by a number of factors such as the links between the XIV systems or the initialization state.

Link status

The link status reflects the connection from the master to the slave volume or consistency group (CG). A link has a direction (from local site to remote or vice versa). A failed link or a

failed secondary system both result in a link error status. The link state is one of the factors determining the mirror operational status. Link states are as follows:

- ▶ OK - Link is up and functioning
- ▶ Error - Link is down

If there are several links (at least two) in one direction and one link fails, this usually does not affect mirroring as long as the bandwidth of the remaining link is high enough to keep up with the data traffic.

Mirror operational status

Mirror operational status is defined as either *operational* or *non_operational*. Mirroring is operational if all of the following conditions apply:

- ▶ The activation state is active.
- ▶ The link is UP.
- ▶ The peers have different roles (master or slave).
- ▶ The mirror is active.

Mirroring is non_operational if any of the following conditions apply:

- ▶ The mirror is inactive.
- ▶ The link is in an error state or deactivated (link down).

Synchronous mirroring states

The synchronization status reflects the consistency of the data between the master and slave volumes. Because the purpose of the remote mirroring feature is to ensure that the slave volumes are an identical copy of the master volumes, this status indicates whether this objective is currently being achieved.

The following states or statuses are possible for the synchronous mirroring:

- ▶ Initializing

The first step in remote mirroring is to create a copy of all the data from the master volume or CG to the slave volume or CG. During this initial copy phase, the status remains initializing.

- ▶ Synchronized (master volume or CG only)/consistent (slave volume or CG only)

This status indicates that all data that has been written to the master volume or CG has also been written to the slave volume or CG. Ideally, the master and slave volumes or CGs must always be synchronized. However, this does not always indicate that the two volumes are absolutely identical in case of a disaster because there are situations when there might be a limited amount of data that was written to one volume, but that was not yet written to its peer volume. This means that the write operations have not yet been acknowledged. These are also known as *pending writes* or *data in flight*.

- ▶ Unsynchronized (master volume only) or inconsistent (slave volume only)

After a volume or CG has completed the initializing stage and achieved the synchronized status, it can become unsynchronized (master) or inconsistent (slave). This occurs when it is not known whether all the data that has been written to the master volume has also been written to the slave volume. This status can occur in the following cases:

- The communications link is down and as a result certain data might have been written to the master volume, but was not yet written to the slave volume.
- Secondary XIV system is down. This is similar to communication link errors because in this state, the primary XIV system is updated, whereas the secondary is not.

- Remote mirroring is deactivated. As a result, certain data might have been written to the master volume and not to the secondary volume. The XIV system keeps track of the partitions that have been modified on the master volumes and when the link is operational again or the remote mirroring is reactivated. These changed partitions can be sent to the remote XIV system and applied to the slave volumes there.

Asynchronous mirroring states

The mirror states can be one of the following:

- ▶ Inactive

The synchronization process is disabled. It is possible to delete a mirror.

- ▶ Initializing

The initial copy is not done yet. Synchronization does not start until the initialization completes. When initialization is complete, the synchronization process is enabled. It is possible to run sync jobs and copy data between master and slave. The possible synchronization states are:

- RPO_OK - Synchronization has completed within the specified sync job interval time (RPO).
- RPO_Lagging - Synchronization has completed but took longer than the specified interval time (RPO).

7.2 Planning PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV remote mirroring

Before implementing the PowerHA SystemMirror software with XIV replicated resources, verify the prerequisite hardware (including microcode versions) and software are in place.

7.2.1 Operational considerations

The following restrictions and limitations apply in a PowerHA environment using XIV replicated resources:

- ▶ PowerHA SystemMirror 7.1.2 Enterprise Edition with XIV replicated resources can include XIV storage systems Gen2 (model A14) with minimum code level 10.2.4 and Gen3 (114) with minimum code level 11.0.0. At the time of writing this book, PowerHA 7.1.2 Enterprise Edition supports XIV replicated resources with XIV remote mirroring between systems of the same generation.
- ▶ Hosts can be attached to the XIV using FC links through physical HBAs, N-Port ID Virtualized (NPIV) adapters or iSCSI attachments. Native vSCSI devices are *not* supported. iSCSI multipathing is also not supported.
- ▶ The XIV attachment on FC currently has two types of drivers for AIX: the AIX native MPIO driver (AIX_AAPCM) and the non-MPIO driver usually used for third-party multipath solutions. PowerHA SystemMirror software always uses AIX MPIO. You can check your current driver in AIX using the command `manage_disk_drivers -1`.
- ▶ The XIV configuration cannot be changed by using dynamic automatic reconfiguration (DARE), although the XIV remote mirror replicated resources or mirror groups can be included and excluded from a PowerHA SystemMirror resource group by using DARE.
- ▶ PowerHA SystemMirror does not handle Simple Network Management Protocol (SNMP) events in the context of the XIV remote mirror function. If there is a failure in the XIV links,

in some cases this can put the corresponding remote mirror in a failed state and PowerHA SystemMirror might not be notified. To fix this problem, you must correct the links and then restart the corresponding resource groups on the appropriate nodes.

7.2.2 Software requirements

Use the following software versions to integrate XIV remote mirror with PowerHA SystemMirror 7.1.2 Enterprise Edition:

- ▶ PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX, or later. For this version, a minimum level of AIX 6.1 TL8 or AIX 7.1 TL2 with RSCT version 3.1.4 is required.
- ▶ IBM XIV microcode bundle 10.2.4, or later.
- ▶ IBM XIV command-line interface (XCLI) for AIX 2.4.4, or later, on each PowerHA SystemMirror node.

The XCLI must be installed on all cluster nodes; it provides a mechanism for managing the XIV replication relationships.

You can download the XCLI and the release notes from the following site:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S4000951>

It is not mandatory to install the AIX host attachment kit because now AIX natively supports XIV using ODM changes. However, we strongly suggest installing it. The kit provides support and access to the latest XIV utilities such as `xiv_diag`. The output of these XIV utilities is required for IBM support when opening an XIV-related service call. You can download the XIV AIX attachment kit from:

[http://www.ibm.com/support/entry/portal/Downloads/Hardware/System_Storage/Disk_systems/Enterprise_Servers/XIV_Storage_System_\(2810,_2812\)](http://www.ibm.com/support/entry/portal/Downloads/Hardware/System_Storage/Disk_systems/Enterprise_Servers/XIV_Storage_System_(2810,_2812))

7.3 Preliminary steps for XIV storage configuration

An important step for preparing the XIV systems to be configured with PowerHA SystemMirror software is the XIV remote mirroring configuration. In this section we highlight the most important steps for configuring the XIV remote mirroring.

Before configuring the PowerHA software, consider the following steps for setting up the XIV remote mirroring:

1. Define the XIV mirroring target.

To connect two XIV systems for remote mirroring, each system must be defined to be a mirroring target of the other. An XIV mirroring target is an XIV system with volumes that receive data copied through XIV remote mirroring. To define an XIV mirroring target for an XIV system, name the target and specify whether Fibre Channel or iSCSI protocol is used to copy the data.

As part of the target system definition, set the maximum initialization and synchronization rates. With the XIV system you can specify a user-specific maximum rate (in MBps) for remote mirroring coupling initialization, and a different user-specific maximum rate for resynchronization. The initialization rate and resynchronization rate are specified for each mirroring target. The maximum initialization rate must be less than or equal to the maximum synchronization job rate (asynchronous mirroring only), which must be less than or equal to the maximum resynchronization rate. See an example of defining a target storage named XIV pPETXIV4 to a current XIV system in Figure 7-1 on page 259.

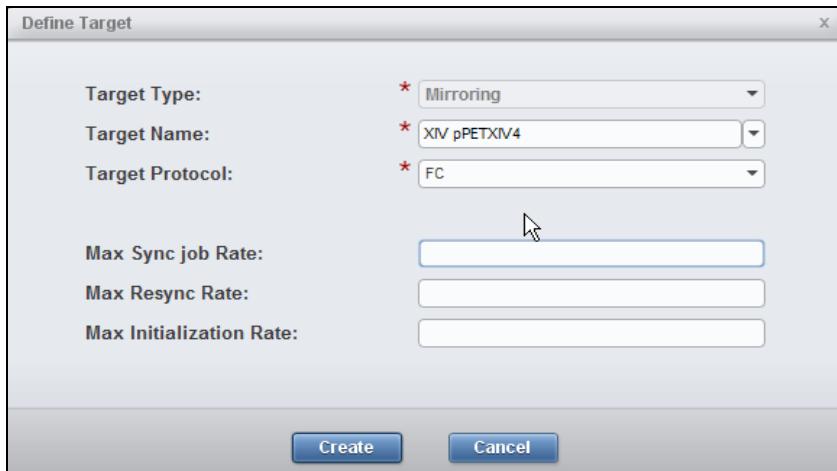


Figure 7-1 Defining a target XIV to a current system

The default settings of the initialization and synchronization rates are:

- Maximum initialization rate: 100 MBps
- Maximum synchronization job: 300 MBps
- Maximum resynchronization rate: 300 MBps

2. Connect the XIV mirroring ports.

After you define the remote mirroring targets, make one-to-one connections between ports on each XIV system. The XIV systems support remote mirroring connectivity for FC or iSCSI ports.

For the XIV's Fibre Channel (FC) ports, connections are unidirectional, from an initiator port (for example Interface Module Port 4 is configured as a Fibre Channel initiator by default) on the source XIV system to a target port (typically Interface Module Port 2) on the target XIV system. Use a minimum of four connections (two connections in each direction, from ports in two different modules, using a total of eight ports) to provide availability protection. See an example in Figure 7-2 on page 260.

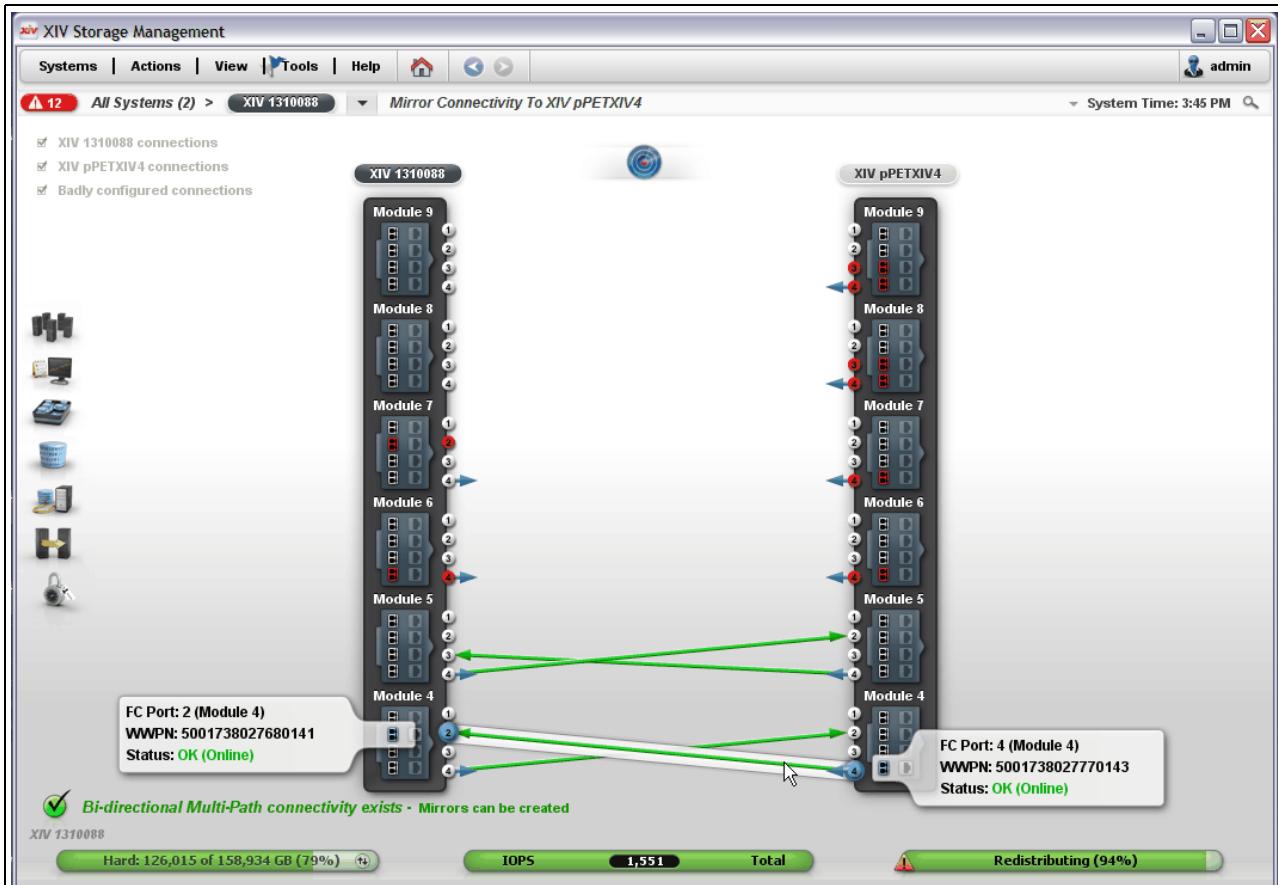


Figure 7-2 XIV remote mirroring - port connections

For iSCSI ports, connections are bidirectional. Use a minimum of two connections (with each of these ports in a different module) using a total of four ports to provide availability protection.

3. Define the XIV volume mirror coupling and peers.

After the mirroring targets are defined, a coupling or mirror might be defined. This creates a mirroring relationship between two peers. For each data volume that participates in the remote mirror on the production site, identify a volume that serves as its remote mirror peer. This volume must reside on an XIV storage unit at the recovery site. For a practical example refer to the scenario sections (7.5.1, “Our scenario description” on page 263).

4. Activate XIV mirror couplings.

When an XIV mirror coupling is activated, all existing data on the master is copied to the slave. This process is referred to as initialization. XIV remote mirroring copies volume identification information (that is, physical volume ID or PVID) and any actual data on the volumes. Space that has not been used is not copied.

5. Create and add the volume mirror couplings to consistency group mirror coupling.

When an XIV consistency group mirror coupling is created, the volumes included in the consistency group must not contain any application data. This prevents data movement and greatly improves the initialization time. After a volume mirror coupling has completed the initialization, the master volume can be added to a mirrored consistency group in the same storage pool. With each mirroring type there are certain additional constraints, such as the same role, target, and schedule. The slave volume is automatically added to the

consistency group on the remote XIV system. One or more additional mirrored volumes can be added to a mirrored consistency group at a later time in the same way.

7.4 Installing the XIV CLI

Perform the following steps in order to install the XCLI software on the PowerHA SystemMirror nodes:

1. Download the latest version of the XIV CLI package. You can use the following IBM sites to download the software for AIX platforms:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S4000951>
<ftp://ftp.software.ibm.com/storage/XIV/GUI/>

Also check the release notes for the version you download for details regarding the platform requirements.

Note: The installation package contains both the GUI and the command line interface (CLI) software. However, only the CLI component is used by PowerHA SystemMirror software when using XIV replicated resources.

2. Install the downloaded package. For the XCLI package version 2.4.4, you need to unpack the bundled package in the installation directory as shown in Example 7-1. Unpacking the installation kit automatically creates the XIVGUI subdirectory in the installation folder.

Example 7-1 Unpacking the XCLI package in the target directory

```
# mkdir /opt/xiv
# cd /opt/xiv
# gzip -dc ./xivcli-2.4.4-build3-aix.tar.gz | tar -xf-
# ls -l
total 8
drwxr-xr-x    6 root      system          4096 Nov 18 16:24 XIVGUI
```

Important: PowerHA SystemMirror software expects that the XCLI executable is located in the /opt/xiv/XIVGUI directory. The verification script applied for XIV replicated resources fails if the XCLI command is not found in the indicated location.

3. Download and install the following additional packages from the AIX Toolbox set:

- readline
- libgcc
- libstdc++

You can select the appropriate packages from the following location:

<http://www-03.ibm.com/systems/power/software/aix/linux/toolbox/alpha.html>

Install the rpm packages on the nodes, as shown in Example 7-2.

Example 7-2 Installing additional libraries

```
#cd /inst/xiv
#rpm -ivh ./readline-4.3-2.aix5.1.ppc.rpm
readline               #####
#rpm -ivh ./libgcc-4.2.0-3.aix6.1.ppc.rpm
libgcc                #####
```

```
#rpm -ivh ./libstdcplusplus-4.2.0-3.aix6.1.ppc.rpm  
libstdc++ #####
```

4. Create the symbolic links for the libgcc and libstdc++ libraries installed in the previous step, as shown in Example 7-3.

Example 7-3 Creating the symbolic links

```
ln -s /opt/freeware/lib/gcc/powerpc-ibm-aix6.1.0.0/4.2.0/libgcc_s.a /usr/lib  
ln -s /opt/freeware/lib/gcc/powerpc-ibm-aix6.1.0.0/4.2.0/libstdc++.a /usr/lib
```

5. Optionally, set the PATH environment variable to point to the XCLI command location:

```
export PATH=$PATH:/opt/xiv/XIVGUI
```

6. Test the XCLI connectivity to the storage subsystem. In Example 7-4, we connected to the storage system pPETXIV4, specifying a management IP address of the XIV.

Example 7-4 Test the node connectivity to the XIV storage system

```
[c581stg10] [/inst/xiv]> xcli -u admin -p xxxxxxxx -m 9.114.63.163 config_get  
Name Value  
dns_primary XIV pPETXIV4  
dns_secondary  
system_name XIV pPETXIV4  
snmp_location Unknown  
snmp_contact Unknown  
snmp_community XIV  
snmp_trap_community XIV  
system_id 10103  
machine_type 2810  
machine_model 114  
machine_serial_number 1310103  
email_sender_address  
email_reply_to_address  
email_subject_format {severity}: {description}  
iscsi_name iqn.2005-10.com.xivstorage:010103  
ntp_server  
support_center_port_type Management  
isns_server ?  
ipv6_state enabled  
ipsec_state disabled  
ipsec_track_tunnels no  
impending_power_loss_detection_method UPS
```

7.5 Creating an XIV replicated cluster using SMIT and XIV GUI

This section provides a practical example for creating a new PowerHA SystemMirror cluster using two nodes and two XIV systems, one at each site. In this section, we detail the XIV remote mirror configuration steps using the graphical management tool (XIVGUI). We also create and test a PowerHA SystemMirror cluster with both synchronous and asynchronous remote mirroring configurations.

7.5.1 Our scenario description

For our implementation example of PowerHA SystemMirror 7.1.2 with XIV Remote Mirroring, we used a sample scenario containing two sites (sitea and siteb). There is also one node at each site. Node c581stg10 in sitea and node c581stg11 in siteb. Nodes are logical partitions (LPARs) defined on two IBM Power Systems 795 servers. They both have AIX 6.1 TL8 SP1 installed.

We used two XIV storage system Gen3 of model 2810-114 in both sites. Figure 7-3 shows an overview of our environment.

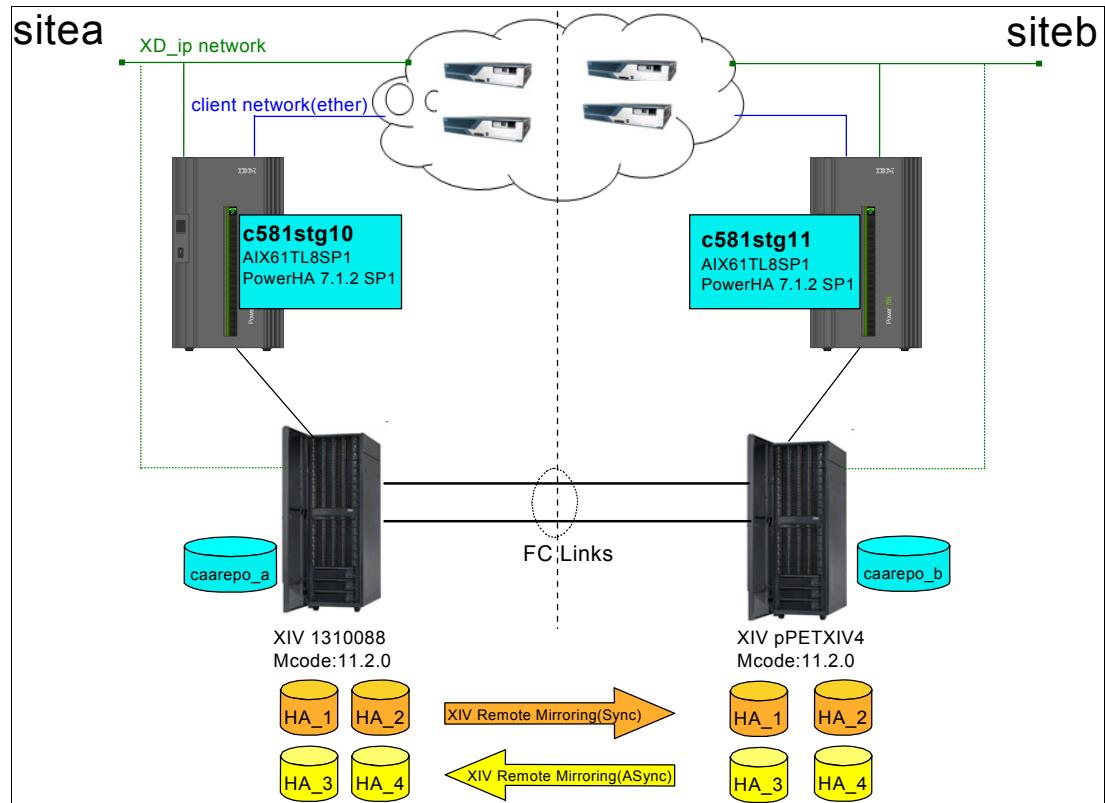


Figure 7-3 An overview of the cluster environment

There are two types of networks defined for the extended distance cluster:

- ▶ **ether** - A client access network where we defined service IP addresses, which can be acquired on all cluster nodes.
- ▶ **XD_ip** - A heartbeat IP network used as an alternate connection between sites. The XIV storages in both sites are connected to this network.

The node communication interfaces are virtual Ethernet, connected to the external network through the Shared Ethernet Adapter (SEA) defined on the Virtual I/O Server (VIOS). We used SEA failover in a dual-VIOS configuration for creating redundant communication paths on each server hosting our LPARs to the external network.

A single IP segment was used for each of the cluster networks. In your environment, you might use different IP segments on both networks, specific to each site. In that case site-specific service IP addresses can also be used.

The communication interfaces used in our cluster are shown in Table 7-1 on page 264.

Table 7-1 Cluster node interfaces

Site	Node	IP label	IP address/mask	AIX interface	PowerHA network	PowerHA net type
sitea	c581stg10	c581stg10	9.114.28.10/26	en0	net_xdip_01	XD_ip
sitea	c581stg10	c581stg10p	10.10.100.10/24	en2	net_ether_01	ether
siteb	c581stg11	c581stg11	9.114.28.11/26	en0	net_xdip_01	XD_ip
siteb	c581stg11	c581stg11p	10.10.100.11/24	en1	net_ether_01	ether
sitea/ siteb	c581stg10/ c581stg11	c581stga	10.10.100.12/24		net_ether_01	ether
sitea/ siteb	c581stg10/ c581stg11	c581stgb	10.10.100.13/24		net_ether_01	ether

In each site we defined volumes mapped to the host in the site. For testing purposes, we defined both synchronous and asynchronous remote mirroring relations between volumes of the XIV storages in the same cluster. Figure 7-4 details the volume configuration on each site and the replication relationships.

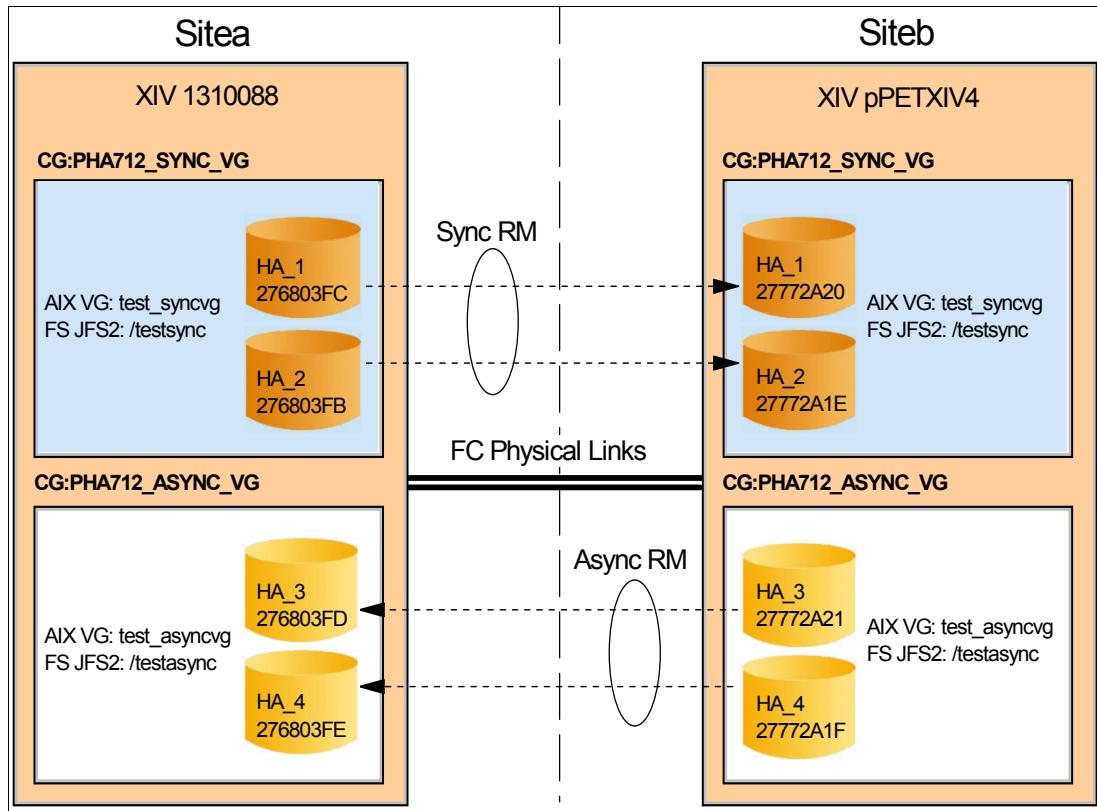


Figure 7-4 XIV remote mirroring pairs

In our case a *linked* cluster configuration was defined. Along with the replicated disk volumes, we also used a volume at each site for the cluster repository disk allocated on the XIV system. Peer volumes from the XIV systems were grouped in consistency groups. The XIV remote mirror configuration includes two consistency groups: one with a synchronous replication and a second with an asynchronous replication.

Table 7-2 details the volumes defined on each XIV storage system, AIX disk devices, and the volume groups created for our test case.

Table 7-2 XIV volumes and AIX logical configuration

Storage system	Volume name	Consist. Group (CG)	Node	AIX device	AIX VG
XIV 1310088	caarepo_a	N/A	c581stg10	caarepo_a	caavg_private
XIV 1310088	HA_1	PHA712_SYNC(CG	c581stg10	HA_1	test_syncvg
XIV 1310088	HA_2	PHA712_SYNC(CG	c581stg10	HA_2	
XIV 1310088	HA_3	PHA712_ASYNC(CG	c581stg10	HA_3	test_asyncvg
XIV 1310088	HA_4	PHA712_ASYNC(CG	c581stg10	HA_4	
XIV pPETXIV4	caarepo_b	N/A	c581stg11	caarepo_b	caavg_private
XIV pPETXIV4	HA_1	PHA712_SYNC(CG	c581stg11	HA_1	test_syncvg
XIV pPETXIV4	HA_2	PHA712_SYNC(CG	c581stg11	HA_2	
XIV pPETXIV4	HA_3	PHA712_ASYNC(CG	c581stg11	HA_3	test_asyncvg
XIV pPETXIV4	HA_4	PHA712_ASYNC(CG	c581stg11	HA_4	

We used the same name for the AIX hdisk device as the volume name in the XIV storage system. Starting with AIX 6.1 TL06 and in AIX 7.1 the following command can be used to rename an AIX device, for example an hdisk device: `rendev -1 hdisk# -n <new_name>`.

7.5.2 Configuration of the XIV remote mirroring

For our test environment, the volumes were created on each XIV system and attached to the nodes in each site. The following operations explain how to create the consistency groups and the XIV remote mirror relationships. For our configuration we used XIV GUI installed previously on a management station. The status of the volumes before creating the remote mirroring pairs is illustrated in Figure 7-5 and Figure 7-6 on page 266.

Name / System	Type	Cluster	LUN ▲
Standalone Hosts			
c581stg10	default		
caarepo_a	vol		1
HA_1	vol		2
HA_2	vol		3
HA_3	vol		4
HA_4	vol		5

Figure 7-5 Volume assignment on XIV systems in Sitea

Name / System	Type	Cluster	LUN ▲
Standalone Hosts			
c567stg07_npiv	default		
c581stg07	default		
c568b01	default		
c581stg1	default		
caarepo_b	vol		1
HA_1	vol		2
HA_2	vol		3
HA_3	vol		4
HA_4	vol		5

Figure 7-6 Volume assignment on XIV systems in Siteb

For more information regarding the configuration and administration of the XIV storage system Gen3, refer to:

http://publib.boulder.ibm.com/infocenter/ibmxiv/r2/topic/com.ibm.help.xivgen3.doc/xiv_gen3homepage.html

Creating the mirroring coupling

We created a remote mirroring coupling between two peers, each placed in an XIV system. The peer of a replication relationship (coupling) can be either an individual volume or a consistency group. We started creating the remote mirroring relationships for the volumes (LUNs) for two types of mirroring:

- ▶ Synchronous remote mirroring, from XIV storage in SiteA (XIV_1310088) to XIV storage in SiteB (XIV_pPETXIV4). In this case, volumes in XIV SiteA have the *master* role and the volumes from XIV in SiteB have the *slave* role.
- ▶ Asynchronous remote mirroring, from XIV storage in SiteB (XIV_pPETXIV4) to XIV storage in SiteA (XIV_1310088). In this case, the volumes from XIV storage in SiteA have the *slave* role, while the volumes from XIV SiteB have the *master* role.

Note: When working with remote mirroring operations, ensure that the time settings and the time zone are consistent across both storage systems. A good practice is to use an NTP server for both XIV systems.

You can check the time settings on an XIV system using the XCLI command `time_list`. See Example 7-5 for an output of this command on our XIV systems.

Example 7-5 Verifying the time and the time zone settings

```
[c581stg10] [/]> xcli -u admin -p adminadmin -m 9.114.63.166 time_list;xcli -u
admin -p adminadmin -m 9.114.63.163 time_list
Time      Date      Time Zone  Daylight Saving Time
04:28:33  2012-11-19  UTC      no
Time      Date      Time Zone  Daylight Saving Time
04:28:36  2012-11-19  UTC      no
```

Creating the synchronous pairs

We created a synchronous remote mirroring relationship for the volumes assigned to this type of replication in our scenario: HA_1 and HA_2. For creating a relationship using XIV GUI, select the source system (in our case XIV 1310088) and then select the Mirroring View (**View → Remote → Mirroring**). Click **Create Mirror** as indicated in Figure 7-7 on page 267.

	Name	RPO	Status	Remote Volume	Remote System
Mirrored Volumes					
	cmm211_datastore		Synchronized	cmm211_datastore	XIV pPETXIV4
	cmm70_datastore		Synchronized	cmm70_datastore	XIV pPETXIV4
	cmm70_raw_VM		Synchronized	cmm70_raw_VM	XIV pPETXIV4
	cmm70_rhel57_os		Synchronized	cmm70_rhel57_os	XIV pPETXIV4
	cmm71_datastore		Synchronized	cmm71_datastore	XIV pPETXIV4
	cmm71_datastore2		Synchronized	cmm71_datastore2	XIV pPETXIV4
	FT_Vol		Synchronized	FT_Vol	XIV pPETXIV4
	FT_Vol Test		Unsynchronized	FT_Vol Test	VII nDFTVMA

Figure 7-7 Mirror view storage SiteA

In the next panel select the volumes on each storage side that form the mirroring coupling. For our scenario we defined two pairs for volumes HA_1 and HA_2 in each site. See Figure 7-8 for creating the synchronous remote mirroring pair between HA_1 volumes. A similar procedure applies for HA_2.

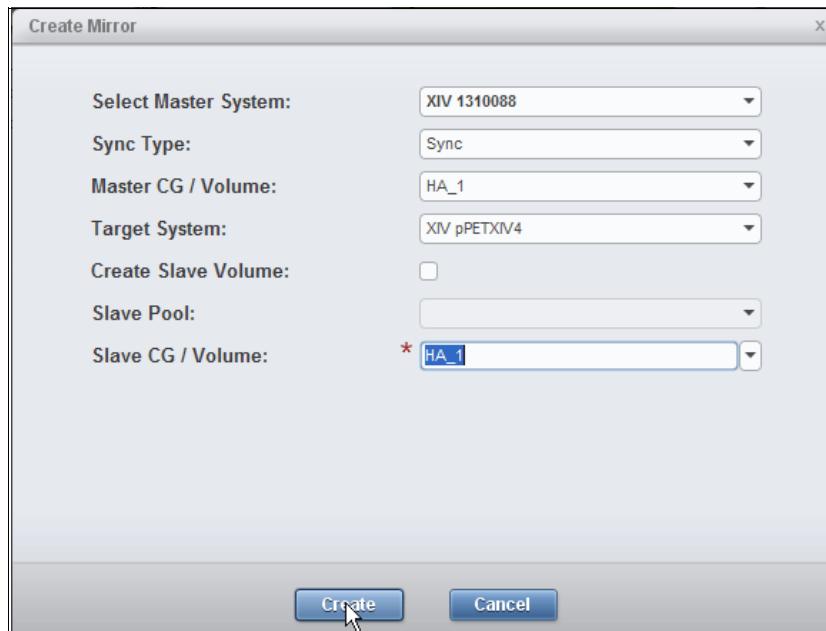


Figure 7-8 Creating a synchronous mirror pair

When creating the synchronous mirroring pairs you need to specify the following:

- ▶ Master system - Name of the source XIV system for the current relationship (in our case XIV in SiteA)
- ▶ Select type - Sync
- ▶ Master CG/volume - The name of the volume in XIV SiteA (HA_1)
- ▶ Target system - Name of the target XIV system for the current relationship (in our case XIV in SiteB)
- ▶ Slave CG/volume - The name of the corresponding volume in XIV SiteB storage (HA_1).

Note: The slave volume must be unlocked and created as formatted (default when it is created from scratch), which also means no associated snapshots. Once mirroring is active, resizing the source volume automatically resizes the target volume to match.

We did *not* check **Create Slave Volume** because the volumes were already defined in each XIV system. If you check this box and specify the remote pool of the slave volume, it creates the target volume on the remote storage before creating the mirroring pair.

Check the status of the pairs in the mirroring view as shown in Figure 7-9.

Name	RPO	Status	Remote Volume	Remote System
Mirrored Volume				
HA_1		Inactive	HA_1	XIV pPETXIV4
HA_2		Inactive	HA_2	XIV pPETXIV4

Figure 7-9 Status of the remote mirror pairs

When the remote mirror pairs get created, the status of the replication is *Inactive*.

Creating the asynchronous pairs

The asynchronous pairs were created in our scenario for HA_3 and HA_4 volumes with master roles for volumes on XIV SiteA (pPETXIV4) and slave role for volumes on XIV SiteB (1310088). To create an asynchronous pair using the XIV GUI, select the source storage and then the mirroring view as in the synchronous case, then click **Create Mirror**. In the following window specify the mirror relationship parameters as shown in Figure 7-10.

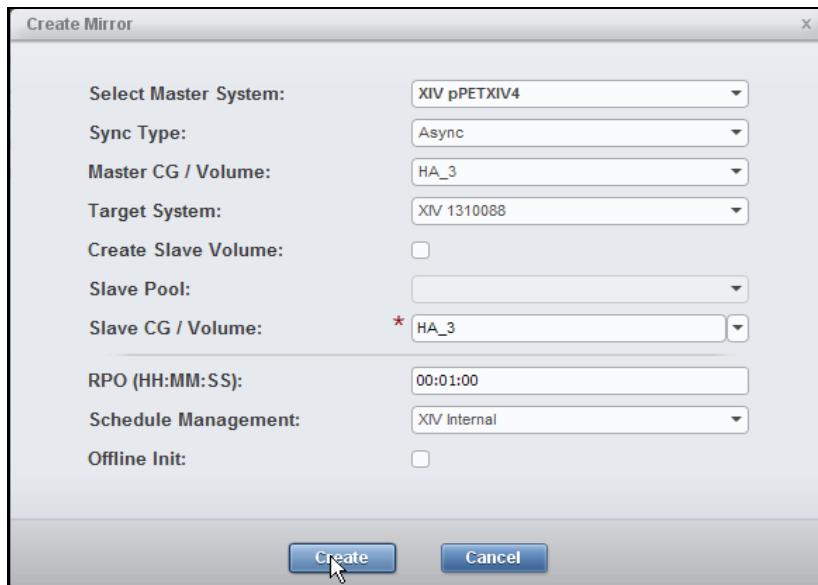


Figure 7-10 Creating an asynchronous mirror pair

The following parameters apply to an async relationship:

- ▶ Master system - Name of the source XIV system for the current relationship (in our case XIV in SiteA)
- ▶ Select type - Async
- ▶ Master CG/volume - The name of the volume in XIV SiteA (HA_3)

- ▶ Target system - Name of the target XIV system for the current relationship (in our case XIV in SiteB)
- ▶ Slave CG/volume - The name of the corresponding volume in XIV SiteB storage (HA_3)

Note: The slave volume must be unlocked and created as formatted (default when it is created from scratch), which also means no associated snapshots. Once mirroring is active, resizing the source volume automatically resizes the target volume to match.

- ▶ RPO - The recovery point objective time designation is the maximum time interval at which the mirrored volume or CG is less current, or lags behind, the master volume. Once the specified interval is reached, a consistent copy of the volume or CG should be available.
- ▶ Schedule management - Set the schedule management field to the XIV internal to create automatic synchronization using scheduled sync jobs. The external option is to specify that no sync jobs will run for this mirror and the interval will be set to Never. At this point you need to run an ad-hoc mirror snapshot to initiate a sync job.
- ▶ Offline init - Only available for selection if the create slave option is not selected. This engages the trucking feature of the XIV storage system that enables initialization of the remote mirror slave peer without requiring the contents of the local master peer to be replicated over an inter-site mirroring link.

You can check the mirroring status using the XIV GUI in the mirroring view as shown in Figure 7-11.

⊖	Name ▲	RPO	Status	Remote Volume	Remote System
Mirrored Volumes					
	HA_1	S	Inactive	HA_1	XIV 1310088
	HA_2	S	Inactive	HA_2	XIV 1310088
	HA_3	M 00:01:00	Inactive	HA_3	XIV 1310088
	HA_4	M 00:01:00	Inactive	HA_4	XIV 1310088

Figure 7-11 Remote mirroring pairs status

Activate the remote mirror relations

For activating the remote mirror replication using XIV GUI, select the source storage for the replication pairs you want to activate, then select the mirror pairs and right-click. Select the **Activate** operation as shown in Figure 7-12 on page 270.

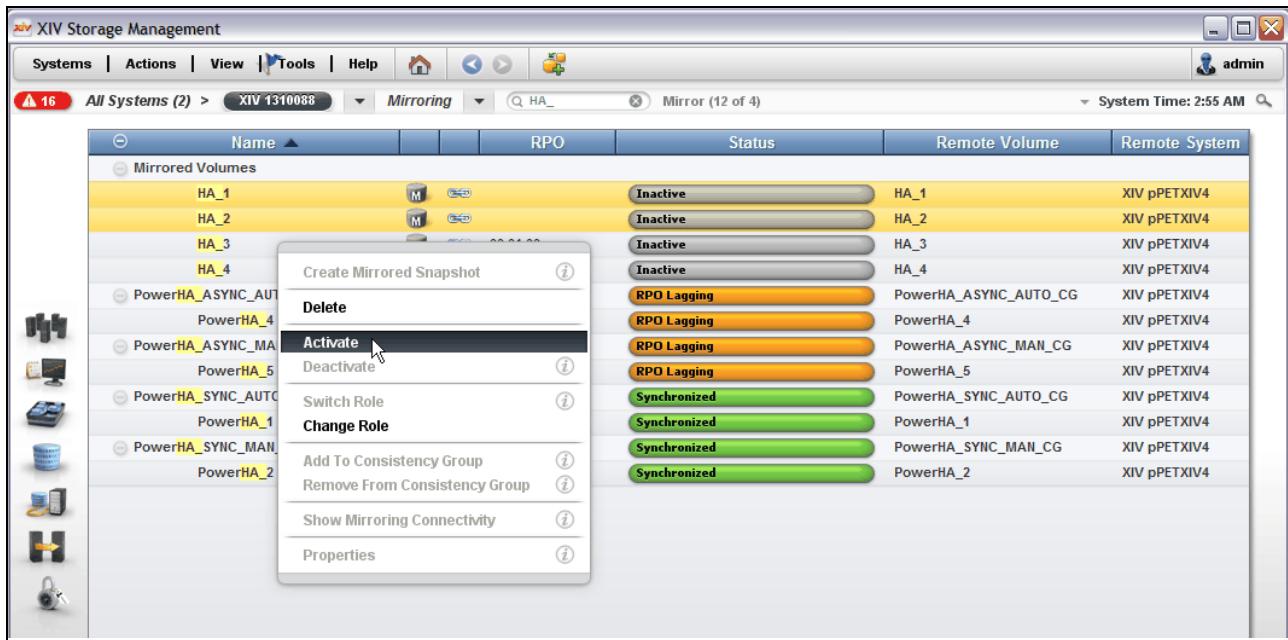


Figure 7-12 Activating the synchronous pairs

The same procedure applies for both types of mirror pairs (sync and async). You can activate the mirroring only by running the operation from the storage containing the peers with the master role. By activating the mirroring pairs, a replication process starts and runs until the volume pairs are synchronized. The initial synchronization process might take some time, depending on the volume capacities and available replication bandwidth. At the end, verify the status of the remote mirroring using the GUI.

Figure 7-13 shows the status of both asynchronous and synchronous pairs defined so far from the XIV storage SiteA. Observe the role and the pair status for both cases:

- ▶ The sync pairs HA_1 and HA_2 have the master role on XIV SiteA (XIV 1310088). The normal status of the mirroring coupling is *Synchronized*.
- ▶ The async pairs HA_3 and HA_4 have a slave role on XIV SiteA. The normal status of the mirroring coupling is *RPO OK*.

Name	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes				
HA_1		Synchronized	HA_1	XIV pPETXIV4
HA_2		Synchronized	HA_2	XIV pPETXIV4
HA_3	00:01:00	RPO OK	HA_3	XIV pPETXIV4
HA_4	00:01:00	RPO OK	HA_4	XIV pPETXIV4

Figure 7-13 Remote mirroring status - SiteA

Figure 7-14 on page 271 shows the status of the remote mirroring pairs from SiteB (XIV pPETXIV4). Observe again the roles of the volumes and the status field for both mirroring cases:

- ▶ The sync pairs HA_1 and HA_2 have here the slave role. The status is *Consistent*.
- ▶ The async pairs HA_3 and HA_4 have the master role on XIV SiteB. The status of the mirroring pairs is *RPO OK*.

Name	RPO	Status	Remote Volume	Remote ...
Mirrored Volumes				
HA_1	S	Consistent	HA_1	XIV 1310088
HA_2	S	Consistent	HA_2	XIV 1310088
HA_3	M	00:01:00	HA_3	XIV 1310088
HA_4	M	00:01:00	HA_4	XIV 1310088

Figure 7-14 Remote mirroring status - SiteB

Add the mirroring pairs to the consistency group

In this step, we add the defined pairs to the consistency groups at both sites:

- ▶ The sync pairs HA_1 and HA_2 are added to consistency group PHA712_SYNC(CG).
- ▶ The async pairs HA_3 and HA_4 are added to consistency group PHA712_ASYNC(CG).

The operation must be performed on both XIV systems. The following are the steps for defining a consistency group (CG) and adding the volumes to the consistency group. We used as example PHA712_SYNC(CG). A similar procedure applies for the second CG.

1. Create a CG on each storage subsystem. We used the same name for both XIV systems. For creating a CG using the XIV GUI, select the storage system in the main panel. Go to **View → Volumes → Consistency Groups**. Click **Create Consistency Group**. In the next window provide the CG parameters, as shown in Figure 7-15.

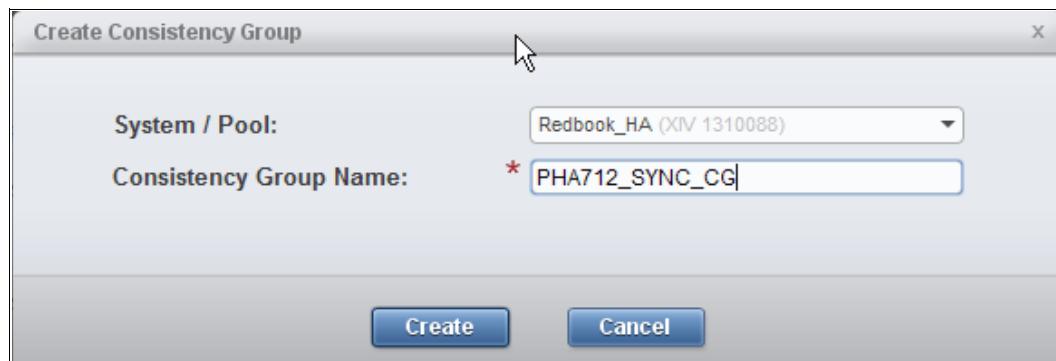


Figure 7-15 Creating an empty CG

The following parameters apply when defining a consistency group:

System/Pool This is the name of the pool containing the volumes that will be included in the CG. In our case, this CG contained HA_1 and HA_2, which were part of the XIV pool named Redbook_HA.

Consistency group name

Provides the name for the consistency group. When you define the consistency group for using with PowerHA SystemMirror XIV replicated resources, always use the same name on both storages.

Apply the same operation for the corresponding CG in the secondary site, using the same name.

2. Create a mirror pair using CGs, following the procedures described in “Creating the mirroring coupling” on page 266. See Figure 7-16 on page 272 for an example of how we created the sync mirror pair between the CG PHA712_SYNC(CG) in both XIV systems.

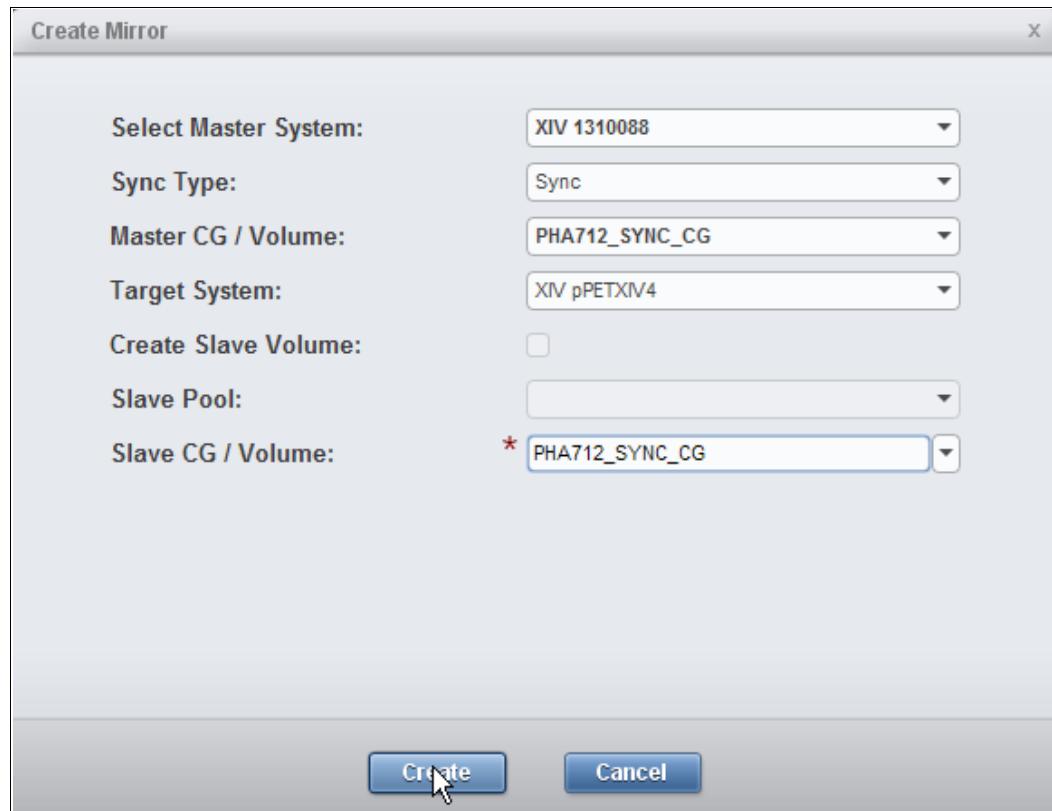


Figure 7-16 Creating the remote mirror pair between the consistency groups

3. Activate the empty CG: In the Mirroring view, right-click the CG name and select **Activate**. The CG is displayed in the *Synchronized* state. See Figure 7-17.

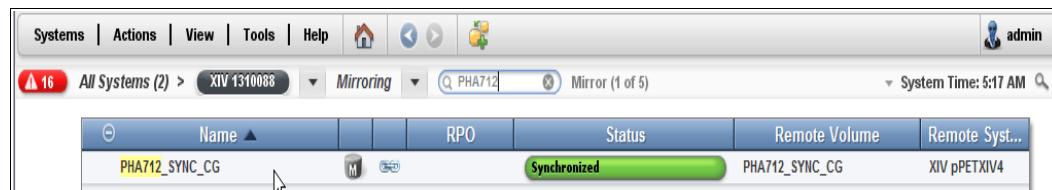


Figure 7-17 CG remote mirroring status

4. Add the volume mirrored pairs to the consistency group, as shown in Figure 7-18.



Figure 7-18 Adding a volume pair to the consistency group(1)

Specify the CG name in the next window, as shown in Figure 7-19.



Figure 7-19 Adding a volume pair to the consistency group(2)

Notes: Volumes are added one by one to the consistency group. When adding a volume in the CG from the primary site, it automatically adds the corresponding peer from the secondary in the target CG.

5. Check the status of the remote mirroring pairs in both sites. See Figure 7-20 for SiteA and Figure 7-21 for SiteB.

All Systems (2) > XIV 1310088	Mirroring	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes					
PHA712_SYNC(CG)	M	S	Synchronized	PHA712_SYNC(CG)	XIV pPETXIV4
HA_1	M	S	Synchronized	HA_1	XIV pPETXIV4
HA_2	M	S	Synchronized	HA_2	XIV pPETXIV4

Figure 7-20 Remote mirror status - SiteA

All Systems (2) > XIV pPETXIV4	Mirroring	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes					
PHA712_SYNC(CG)	s	✓	Consistent	PHA712_SYNC(CG)	XIV 1310088
HA_1	s	✓	Consistent	HA_1	XIV 1310088
HA_2	s	✓	Consistent	HA_2	XIV 1310088

Figure 7-21 Remote mirror status - SiteB

7.5.3 Installation and configuration of the PowerHA SystemMirror software

This section provides details about PowerHA SystemMirror configuration with XIV replicated resources. In our scenario, we used the XIV remote mirror configuration already created in the previous steps and created a resource group for each type of mirroring (sync and async). We now explain the steps to create the replicated resources and the resource group, as well as possible options during the cluster configuration.

Installation of the PowerHA SystemMirror software

Before installing the PowerHA SystemMirror software, verify the software prerequisites as described in 7.2.2, “Software requirements” on page 258. Install the PowerHA software from the installation media using the `installp` command, or use `smitty install_latest` fastpath. Make sure you include the following filesets for enabling the Enterprise Edition features and the XIV replicated resources support:

- ▶ `cluster.xd.base`
- ▶ `cluster.xd.license`
- ▶ `cluster.es.genxd.rte`
- ▶ `cluster.es.genxd.cmds`
- ▶ `cluster.msg.en_US.genxd`

The cluster filesets `genxd` provide the HADR Storage Framework capabilities of the PowerHA software. This includes XIV remote mirroring, DS8000 global mirror and DS8000 metro mirror with in-band communication.

It is always considered a best practice to apply the latest service pack (SP) available. You can verify the latest available SP for PowerHA SystemMirror software at the IBM FixCentral page:

<http://www-933.ibm.com/support/fixcentral/>

See Example 7-6 for the list of filesets installed in our environment (all cluster nodes).

Example 7-6 Cluster fileset list

Fileset	Level	State	Type	Description (Uninstaller)
<hr/>				
<code>cluster.adt.es.client.include</code>	7.1.2.0	C	F	PowerHA SystemMirror Client Include Files
<code>cluster.adt.es.client.samples.clinfo</code>	7.1.2.0	C	F	PowerHA SystemMirror Client CLINFO Samples
<code>cluster.adt.es.client.samples.clstat</code>	7.1.2.0	C	F	PowerHA SystemMirror Client Clstat Samples
<code>cluster.adt.es.client.samples.libcl</code>	7.1.2.0	C	F	PowerHA SystemMirror Client LIBCL Samples
<code>cluster.adt.es.java.demo.monitor</code>	7.1.2.0	C	F	Web Based Monitor Demo
<code>cluster.es.client.clcomd</code>	7.1.2.0	C	F	Cluster Communication Infrastructure
<code>cluster.es.client.lib</code>	7.1.2.1	C	F	PowerHA SystemMirror Client Libraries
<code>cluster.es.client.rte</code>	7.1.2.1	C	F	PowerHA SystemMirror Client Runtime
<code>cluster.es.client.utils</code>	7.1.2.1	C	F	PowerHA SystemMirror Client Utilities
<code>cluster.es.client.wsm</code>	7.1.2.0	C	F	Web based Smit
<code>cluster.es.cspoc.cmds</code>	7.1.2.1	C	F	CSPOC Commands
<code>cluster.es.cspoc.dsh</code>	7.1.2.0	C	F	CSPOC dsh
<code>cluster.es.cspoc.rte</code>	7.1.2.1	C	F	CSPOC Runtime Commands
<code>cluster.es.genxd.cmds</code>	7.1.2.1	C	F	PowerHA SystemMirror Enterprise Edition - Generic

cluster.es.genxd.rte	7.1.2.1	C	F	XD support - Commands PowerHA SystemMirror Enterprise Edition - Generic XD support - Runtime Environment
cluster.es.migcheck	7.1.2.0	C	F	PowerHA SystemMirror Migration support
cluster.es.server.cfgast	7.1.2.0	C	F	Two-Node Configuration Assistant
cluster.es.server.diag	7.1.2.1	C	F	Server Diags
cluster.es.server.events	7.1.2.1	C	F	Server Events
cluster.es.server.rte	7.1.2.1	C	F	Base Server Runtime
cluster.es.server.testtool	7.1.2.0	C	F	Cluster Test Tool
cluster.es.server.utils	7.1.2.1	C	F	Server Utilities
cluster.es.worksheets	7.1.2.0	C	F	Online Planning Worksheets
cluster.license	7.1.2.0	C	F	PowerHA SystemMirror Electronic License
cluster.man.en_US.es.data	7.1.2.1	C	F	Man Pages - U.S. English
cluster.msg.en_US.es.client	7.1.2.0	C	F	PowerHA SystemMirror Client Messages - U.S. English
cluster.msg.en_US.es.server	7.1.2.0	C	F	Recovery Driver Messages - U.S. English
cluster.msg.en_US.genxd	7.1.2.0	C	F	PowerHA SystemMirror Enterprise Edition - Generic XD support - Messages - U.S. English IBM-850
cluster.xd.base	7.1.2.0	C	F	PowerHA SystemMirror Enterprise Edition - Base Support.
cluster.xd.license	7.1.2.0	C	F	PowerHA SystemMirror Enterprise Edition License Agreement Files

Defining the cluster topology

In this section we explain the topology configuration of our cluster according to the scenario description provided in 7.5.1, “Our scenario description” on page 263. We point out important menu panels when defining the multisite configuration. For creating a basic cluster topology, refer also to the PowerHA SystemMirror manuals at:

<http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.navigation/powerha.htm>

We created a linked cluster that included a CAA repository disk local to each site. For easily matching the LUNs defined on the XIV systems with the AIX hdisk devices, we renamed the hdisk devices to the names of the LUNs on the XIV. You can rename an hdisk device in AIX using the command `rendev -l hdisk# -n <newname>` as shown in Example 7-7.

Example 7-7 Renaming an hdisk device in AIX

```
root@c581stg11:/>lspv
hdisk0      00f61a3bc9943dfe          rootvg      active
hdisk1      00f61a3b1b0aea54          None       -
hdisk2      00f61a3bfbe8054d          test_syncvg
```

```

hdisk3      00f61a3bfbe804a5      test_syncvg
hdisk4      00f61a3bfbea9c23      test_asyncvg
hdisk5      00f61a3bfbea9bd8      test_asyncvg

root@c581stg11:/>rendev -l hdisk1 -n caarepo_b
root@c581stg11:/>lspv
hdisk0      00f61a3bc9943dfe      rootvg      active
caarepo_b  00f61a3b1b0aea54      None
hdisk2      00f61a3bfbe8054d      test_syncvg
hdisk3      00f61a3bfbe804a5      test_syncvg
hdisk4      00f61a3bfbea9c23      test_asyncvg
hdisk5      00f61a3bfbea9bd8      test_asyncvg

```

We defined a linked cluster configuration by using SMIT and executing **smitty sysmirror** → **Cluster Nodes and Networks** → **Multi Site Cluster Deployment** → **Setup Cluster, Sites, Nodes and Networks** as shown in Figure 7-22.

Setup Cluster, Sites, Nodes and Networks	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
* Cluster Name	[Entry Fields] [xiv_cluster]
* Site 1 Name	[sitea]
* New Nodes (via selected communication paths)	[c581stg10] +
* Site 2 Name	[siteb]
* New Nodes (via selected communication paths)	[c581stg11] +
Cluster Type	[Linked Cluster] +

Figure 7-22 Defining the linked cluster xiv_cluster

We specified the names for the cluster and the sites in the SMIT panel. The names of the nodes in our case were their hostnames. The hostname must have an associated IP address. All node interfaces (including the hostname interfaces) must be added to /etc/cluster/rhosts prior to running the initial cluster configuration.

Note: The PowerHA SystemMirror software automatically detects the node interfaces and adds them to networks according to their IP subnet definition and VLAN membership.

By default, the cluster software associates the interfaces with PowerHA networks of type *ether*. In order to change the default names and the type of a network to *XD_ip*, we used SMIT menus by executing **smitty sysmirror** → **Cluster Nodes and Networks** → **Manage Networks and Network Interfaces** → **Networks** → **Change>Show a Network**.

In the next step, we defined the repository disks for the sites. We defined both of them in a single SMIT panel by executing **smitty sysmirror** → **Cluster Nodes and Networks** → **Multi Site Cluster Deployment** → **Define Repository Disk and Cluster IP Address**. We selected the candidate disks for the CAA repository in each site and kept the default IP address allocation for multicasting in each site. See Figure 7-23 on page 277.

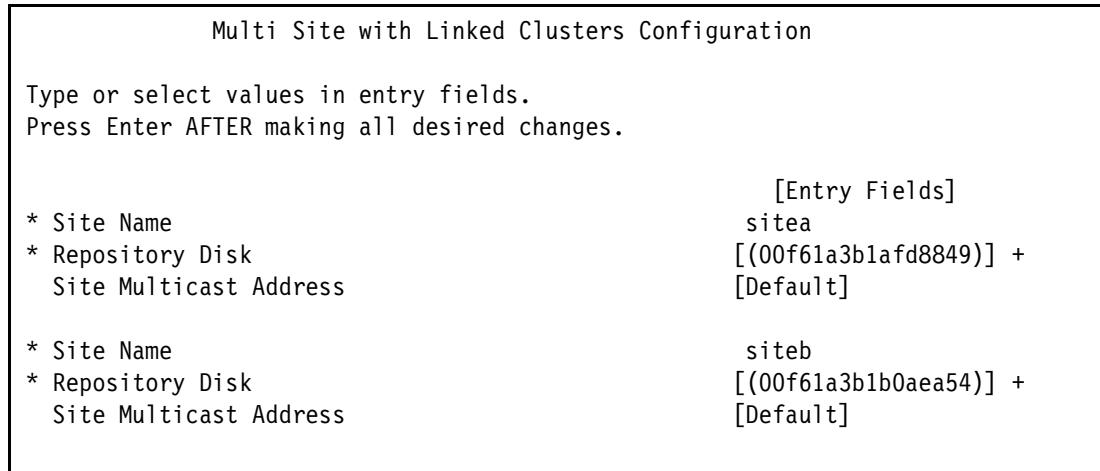


Figure 7-23 Defining the repository disks and the multicast IP addresses

Note: When selecting the CAA repository disk, a PVID is automatically assigned for that disk if it has not a PVID defined already. The SMIT panel gets populated with the PVIDs of the candidate repository disks, making the cluster configuration independent of hdisk device numbers or names.

In the next step, we performed the cluster verification and synchronization by executing **smitty systemmirror** → **Cluster Nodes and Networks** → **Verify and Synchronize Cluster Configuration**.

The configuration we defined so far is shown in the **cltopinfo** command output in Example 7-8.

Example 7-8 Cluster configuration - cltopinfo

```
[c581stg10] [/]> cltopinfo
Cluster Name: xiv_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disks: Site 1: caarepo_a, Site 2:
Cluster IP Addresses: Cluster: , Site 1: 228.114.28.10, Site 2: 228.114.28.11
There are 2 node(s) and 2 network(s) defined

NODE c581stg10:
    Network net_xdip_01
        c581stg10      9.114.28.10
    Network net_ether_01
        c581stg10p    10.10.100.10

NODE c581stg11:
    Network net_xdip_01
        c581stg11      9.114.28.11
    Network net_ether_01
        c581stg11p    10.10.100.11
```

No resource groups defined

Configuring PowerHA XIV replicated resources and resource groups

This section describes the steps performed to create the PowerHA XIV replicated resources and their corresponding resource groups. We first created the resource group associated with the synchronous pairs and later we dynamically added the resource group associated with the async pairs.

The volume group definition and the file systems were already defined on the primary node. We started by exemplifying how you can import the already defined volume group from the primary site, to the secondary site.

Importing the volume group definition to the secondary site

This step is required the first time you import the volume group definition created in the primary site, to the secondary site. In our case, we imported the volume group *test_syncvg* defined on c581stg10 in SiteA to the secondary node c581stg11 in SiteB. The replication services between the disks HA_1 and HA_2 containing the volume group were active at this time. In Example 7-9, we describe the volume group configuration we had on the primary node c581stg10.

Example 7-9 Volume group details in SiteA

```
[c581stg10] [/]> lsvg -l test_syncvg
test_syncvg:
LV NAME      TYPE    LPs    PPs    PVs   LV STATE    MOUNT POINT
tstsncloglv  jfs2log  1      1      1     closed/syncd  N/A
tstsncnv     jfs2      1000   1000   2     closed/syncd  /testsync

[c581stg10] [/]> lsvg -p test_syncvg
test_syncvg:
PV_NAME      PV STATE    TOTAL PPs    FREE PPs    FREE DISTRIBUTION
HA_1          active      512         23        00..00..00..00..23
HA_2          active      512         0         00..00..00..00..00
```

Before importing the volume group to the secondary site, match the disk part of the volume group across the sites. We used the *lspv* command and checked the PVID of the disks in both sites. See a comparative output in Figure 7-24 for matching the disks of the *test_syncvg* volume group in both sites.

[c581stg10] [/]> lspv grep test_syncvg	root@c581stg11:/>lspv
HA_1 00f61a3b1bb46046 test_syncvg active	hdisk0 00f61a3bc9943dfe rootvg active
HA_2 00f61a3b1bb46135 test_syncvg active	caarepo_b 00f61a3b1b0aea54 caavg_private active

Figure 7-24 Matching the disks of the volume group *test_syncvg* on two nodes using PVIDs

If the PVIDs are not found because the disks are newly attached to the host in the secondary site, you need to delete the devices using the *rmdev -dl hdisk#* command and rediscover them using the *cfgmgr* command. Once you have the PVIDs shown in the *lspv* output on the secondary site, vary off the volume group in the primary site to avoid any modifications and import the volume group to the secondary site as shown in Example 7-10 on page 279.

Example 7-10 Importing the test_syncvg to the second site

```
root@c581stg11:/>importvg -y test_syncvg -n -V50 HA_1  
test_syncvg  
root@c581stg11:/>
```

The **-V** flag shown in Example 7-10 is optional. We used it for keeping the VG major number the same in both sites. You can import the volume group without varying it on (using the **-n** flag during the import operation) while keeping the remote mirror relationships active. This can be done with XIV remote mirroring because the slave volume is accessible for read-only operations.

Creating the XIV replicated resources

Having the remote mirroring pairs already defined and activated, and the volumes included in the consistency groups as explained in 7.5.2, “Configuration of the XIV remote mirroring” on page 265, we could now define the PowerHA XIV replicated resources. In the following steps, we explain how we created the XIV remote mirror resources in PowerHA SystemMirror for the synchronous mirrored pairs included in the consistency group PHA712_SYNC(CG).

1. Define the storage agents for the XIV systems. PowerHA SystemMirror uses the **xc1i** command installed on all cluster nodes with the parameters indicated in the storage agent definition to manage the remote mirroring pairs.

You can configure the Storage Agents for XIV in PowerHA SystemMirror using SMIT:

smitty sysmirror → **Cluster Applications and Resources** → **Resources** → **Configure XIV Remote Mirror Resources** → **Configure Storage Agents** → **Add a Storage Agent**.

Figure 7-25 shows the configuration for SiteA XIV.

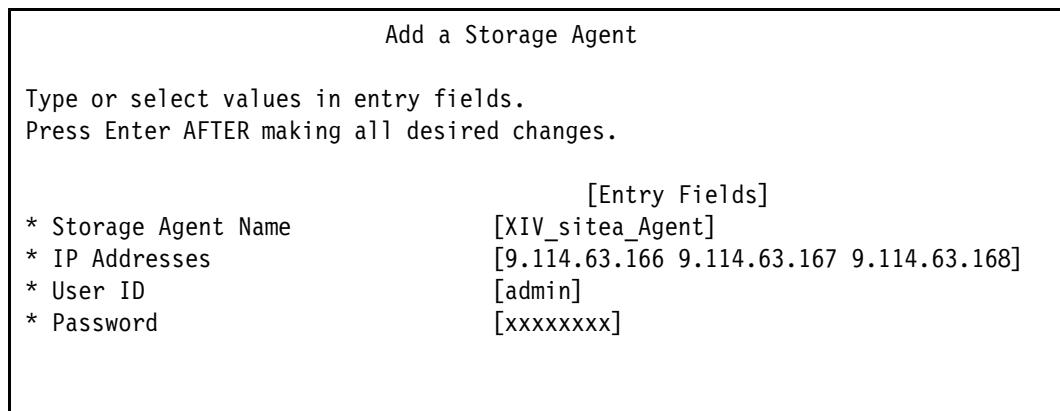


Figure 7-25 Configuring the storage agent for the XIV system in SiteA

Next we defined the storage agent for the XIV system in SiteB as shown in Figure 7-26 on page 280.

Add a Storage Agent	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
[Entry Fields]	
* Storage Agent Name	[XIV_siteb_Agent]
* IP Addresses	[9.114.63.163 9.114.63.164 9.114.63.165]
* User ID	[admin]
* Password	[xxxxxxxx]

Figure 7-26 Configuring the storage agent for the XIV system in SiteB

The following parameters must be specified when defining the XIV storage agent in PowerHA SystemMirror:

Storage agent name This is a user-provided name for the storage agent.

IP addresses Up to four IP addresses can be specified. PowerHA SystemMirror tries to use one of the specified addresses to communicate with the storage system. They are the management IP addresses of the XIV system.

User ID This is the user name required for accessing the XIV system using XCLI.

Password This is the password associated with the previous user ID required for accessing the XIV system using xcli.

2. Define the storage systems in PowerHA SystemMirror. We did that by using SMIT:

smitty sysmirror → **Cluster Applications and Resources** → **Resources** → **Configure XIV Remote Mirror Resources** → **Configure Storage Systems** → **Add a Storage System**. See Figure 7-27 on page 281 for the storage definition associated with the XIV system in SiteA.

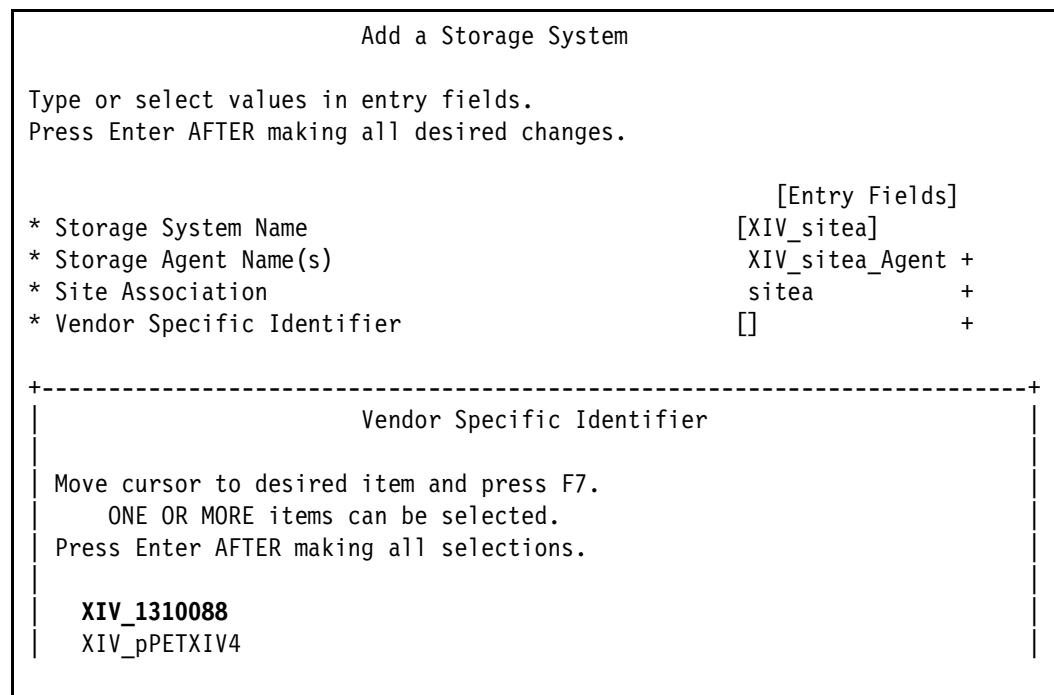


Figure 7-27 Configuring the storage system in SiteA

Note: Except for the storage system name parameter, all other parameters can be provided by using the F4 key to access a drop-down list and selecting the appropriate values.

We next defined the second storage system for SiteB as shown in Figure 7-28.

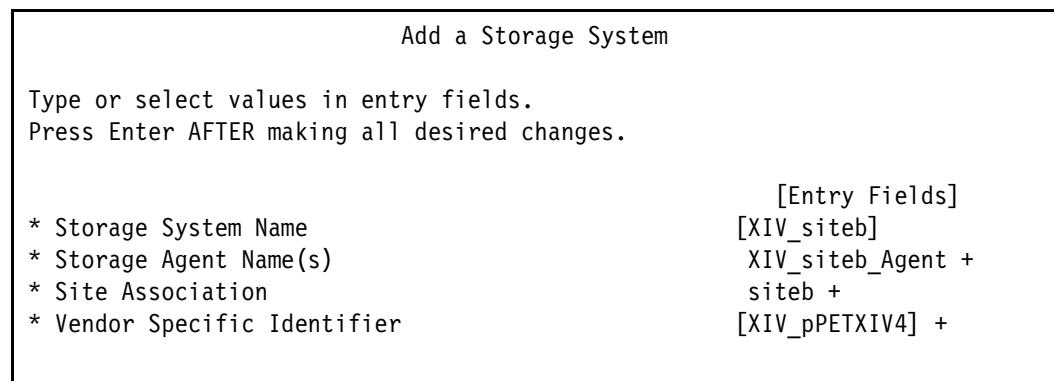


Figure 7-28 Configuring the storage system in SiteB

The following parameters must be provided when configuring an XIV storage system in PowerHA SystemMirror:

Storage system name This is a user-provided arbitrary name for the storage system.

Storage agent name This is the storage agent associated with the storage system you define.

Site association This is one of the sites already defined in PowerHA SystemMirror that is associated with the XIV storage system you define.

Vendor Specific Identifier

This field is associated with the XIV system name defined on the XIV storage system.

You can get the system name of the storage complex with the **xcli** command **config_get** as shown in Example 7-11.

Example 7-11 Output of the xcli command config_get on XIV SiteA

Name	Value
dns_primary	9.12.16.2
dns_secondary	9.12.18.2
system_name	XIV 1310088
snmp_location	Unknown
snmp_contact	Unknown
snmp_community	XIV
snmp_trap_community	XIV
system_id	10088
machine_type	2810
machine_model	114
machine_serial_number	1310088
email_sender_address	
email_reply_to_address	{severity}: {description}
email_subject_format	iqn.2005-10.com.xivstorage:010088
iscsi_name	9.56.248.20
ntp_server	
support_center_port_type	Management
isns_server	?
ipv6_state	disabled
ipsec_state	disabled
ipsec_track_tunnels	no
impending_power_loss_detection_method	UPS

3. Configure the mirror groups. We created the mirror group associated with the consistency group PHA712_SYNC(CG). This CG is already defined on the XIV systems and contains synchronous remote mirroring pairs. For creating a mirror group in PowerHA SystemMirror, use SMIT:

smitty sysmirror → Cluster Applications and Resources → Resources → Configure XIV Remote Mirror Resources → Configure Mirror Groups → Add a Mirror Group.

See Figure 7-29 on page 283 for how we defined the mirror group associated with the consistency group PHA712_SYNC(CG).

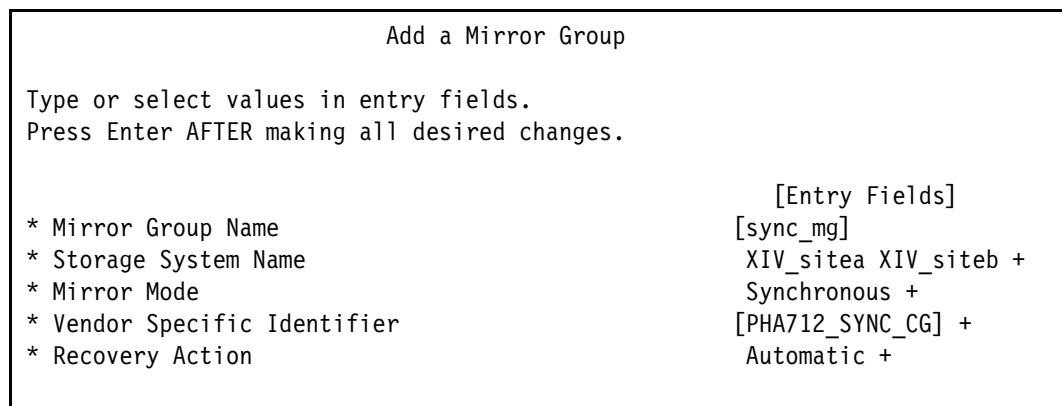


Figure 7-29 Creating the XIV mirror group

Note: Except for the mirror group name, all the parameters can be selected from a drop-down list using the F4 key.

The following parameters must be provided when creating the XIV mirror group:

Mirror group name A user-provided name. It does not have to match a particular XIV field. It can have 1-64 characters.

Storage system name

Select all storage systems containing the remote mirrored consistency groups. In our case, we had only two XIVs and one CG defined with the same name on both XIVs.

Mirror mode

Either specify synchronous or asynchronous depending on your consistency group remote mirroring type. In our case, we used synchronous mode for the consistency group PHA712_SYNC_CG.

Vendor specific identifier

Specify in this field the consistency groups participating in the remote mirroring relationship. In our case, we had a single consistency group of synchronous type defined with the same name on both XIVs.

Recovery action

It can be either manual or automatic. This parameter determines how PowerHA SystemMirror software reacts in case of a site failure:

Automatic - The cluster automatically attempts to acquire the resource group associated with this mirror group in case of a failure of the primary site for the resource group.

Manual - In case of a site failover, the cluster provides an action plan in the log file (/var/hacmp/log/hacmp.out) and does not acquire the resource group on the secondary site. User intervention is required to manually failover the XIV remote mirroring relationships and activate the resource groups in the secondary site.

Note: When the recovery action is set to *manual*, a total site failover will cause the cluster not to activate the resource groups in the secondary location. Cluster checks the status of the mirroring pairs at the time of site failover. While using C-SPOC to perform a resource group move to the other site, or systems being down in the primary site, but the replication consistency groups reflect a consistent state, the cluster software attempts to acquire the resource group in the secondary site.

The following considerations also apply:

- ▶ In the Vendor Specific Identifier you need to provide a name for the consistency group defined on the XIV systems and not a name of a particular volume (LUN) defined on the storage.
- ▶ There is a one-to-one relationship between the mirror group and the consistency group on the storage. You cannot include multiple consistency groups from one or multiple XIV systems in the same mirror group.
- ▶ It is necessary to have the same consistency group name on both XIVs associated with the same remote mirroring pairs.

Creating a resource group with XIV remote mirroring resources

The XIV mirror groups need to be included in a resource group to provide site failover capabilities of the applications using the volumes in the remote mirroring relationships.

In the following steps, we define a resource group associated with the volume group test_syncvg created on the volumes HA_1 and HA_2 using synchronous mirroring. The volumes are part of the XIV storage consistency group PHA712_SYNC(CG associated with the previously defined PowerHA SystemMirror XIV mirror group sync_mg. We set the following characteristics for the resource group associated with this mirror group:

- ▶ Inter-Site Management Policy: Online On Either Site

In this case, resources may be acquired by any site in its resource chain. When a site fails, the resource will be acquired by the highest priority standby site. When the failed site rejoins, the resource remains with its new owner.

- ▶ Intra-Site Management Policy

- Startup Policy: Online On Home Node Only

The resource group will be brought up online *only* on its home node (the first node in the participating node list) during the resource group startup. This requires the highest priority node to be available. In our case we had only one node in each site.

- Fallover Policy: Fallover To Next Priority Node In The List

In the case of failover, the resource group that is online on only one node at a time follows the default node priority order specified in the resource group's nodelist. In our case we had only one node in each site.

- Fallback Policy

A resource group falls back when a higher priority node joins the cluster. If you select this option, then you can use the delayed fallback timer that you previously specified in the Configure Resource Group Run-time Policies SMIT menu. If you do not configure the delayed fallback policy, the resource group falls back immediately when a higher priority node joins the cluster. In our case we had only one node in each site.

We created the resource group using SMIT:

| smitty sysmirror → Cluster Applications and Resources → Resource Groups → Add a Resource Group.

See Figure 7-30 on page 285 for our resource group named sync_rg with the characteristics previously explained.

Add a Resource Group (extended)

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

	[Entry Fields]
* Resource Group Name	[sync_rg]
Inter-Site Management Policy	[Online On Either Site] +
* Participating Nodes from Primary Site	[c581stg10] +
Participating Nodes from Secondary Site	[c581stg11] +
Startup Policy	Online On Home Node Only +
Failover Policy	Failover To Next Priority Node In The List +
Fallback Policy	Fallback To Higher Priority Node In The List +

Figure 7-30 Creating a resource group for the XIV mirror group

Next we defined the resources associated with sync_rg resource groups:

- ▶ Service IP address: c581stgsa

This is an IP address associated with cluster network net_ether_02 (10.10.100.0/24). This address has been defined as a service IP address without an associated site. In case the node in SiteA fails, this IP address will be acquired in the secondary site, SiteB. Refer to PowerHA manuals about how to define a service IP address resource.

- ▶ Volume group: test_syncvg

This is the volume group defined on AIX disks HA_1 and HA_2 in both sites. The volumes are part of the consistency group PHA712_SYNC(CG) defined on each XIV system.

- ▶ XIV replicated resources

Specify the name of the XIV mirror groups associated with the current resource group. In our case we selected sync_mg.

Prior to adding the resources, you can run a discovery process for network and disks, so that the cluster discovers any additional disks and volume groups you might have defined: **smitty sysmirror** → **Cluster Nodes and Networks** → **Discover Network Interfaces and Disks**.

For adding the resources to the resource group, you can use SMIT: **smitty sysmirror** → **Cluster Applications and Resources** → **Resource Groups** → **Change>Show**

Resources and Attributes for a Resource Group. See Figure 7-31 on page 286 for the fields we changed when adding the resources to the sync_rg resource group.

Change/Show All Resources and 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	sync_rg
Inter-site Management Policy	Online On Either Site
Participating Nodes from Primary Site	c581stg10
Participating Nodes from Secondary Site	c581stg11
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority
Fallback Policy	Node In The List
Fallback Timer Policy (empty is immediate)	Fallback To Higher PriorityNodeInTheList
Service IP Labels/Addresses	[c581stgsa] +
Application Controller Name	[] +
Volume Groups	[test_syncvg] +
Use forced varyon of volume groups, if necessary	false +
Automatically Import Volume Groups	false +
Filesystems (empty is ALL for VGs specified)	[] +
Filesystems Consistency Check	fsck +
Filesystems Recovery Method	sequential +
Filesystems mounted before IP configured	false +
Filesystems/Directories to Export (NFSv2/3)	[] +
Filesystems/Directories to NFS Mount	[] +
Network For NFS Mount	[] +
Tape Resources	[] +
Raw Disk PVIDs	[] +
Raw Disk UUIDs/hdisks	[] +
Disk Error Management?	no +
Primary Workload Manager Class	[] +
Secondary Workload Manager Class	[] +
Miscellaneous Data	[] +
WPAR Name	[] +
User Defined Resources	[] +
DS8000 Global Mirror Replicated Resources	[] +
XIV Replicated Resources	sync_mg +
DS8000-Metro Mirror (In-band) Resources	+ +

Figure 7-31 Adding the resources to the resource group sync_rg

After completing the resource group definition, verify and synchronize the cluster definition on the nodes. You can use **smitty cm_apps_resources** → **Verify and Synchronize Cluster Configuration**.

Start the cluster services

After defining the resource group sync_rg, we started the cluster services on all cluster nodes using **smitty clstart** as shown in Figure 7-32 on page 287.

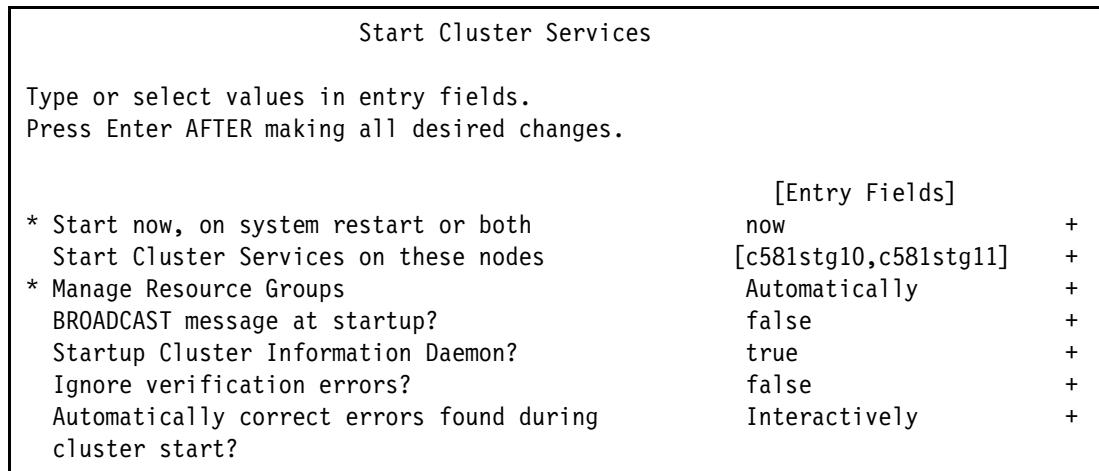


Figure 7-32 Starting the cluster services on the cluster nodes

We verify the resource group sync_rg with the primary node c581stg10 acquired in SiteA. After starting the cluster services and acquiring the resource group on node c581stg10, the cluster resource group status is ONLINE for the node in SiteA and ONLINE SECONDARY for the node in SiteB (c581stg11), as shown in Example 7-12.

Example 7-12 Resource group status

```
root@c581stg11:/>clRGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: sync_rg

Node	Primary State	Secondary State
c581stg10@sitea	ONLINE	OFFLINE
c581stg11@siteb	OFFLINE	ONLINE SECONDARY

Adding the second resource group

We dynamically added a second resource group in the cluster configuration for the XIV asynchronous remote mirroring pairs. This resource group has SiteB as the primary site.

We had the XIV remote mirroring pairs defined and activated with the master volumes in XIV from SiteB and the slave volumes in XIV SiteA. The AIX disk devices associated with the async pairs were in both sites named HA_3 and HA_4. They were part of the consistency group PHA712_ASYNC(CG) defined on both XIV storage systems. The actual status of the consistency group is shown in Figure 7-33.

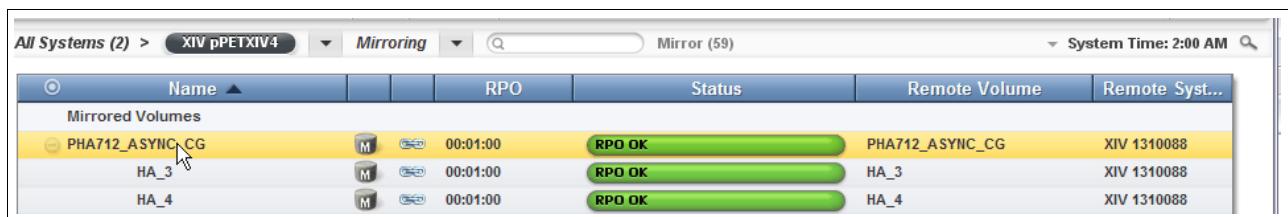


Figure 7-33 Status of the asynchronous consistency group on storage XIV from SiteB

A volume group was defined on disks HA_3 and HA_4 on node c581stg11 in SiteB. See Example 7-13 for the details of the volume group test_asyncvg.

Example 7-13 Volume group logical and physical volumes

```
root@c581stg11:/>lsvg -l test_asyncvg
test_asyncvg:
LV NAME      TYPE     LPs    PPs    PVs   LV STATE    MOUNT POINT
testasyncloglv jfs2log  1      1      1    closed/syncd  N/A
testasynclv    jfs2     1000   1000   2    closed/syncd  /testasync

root@c581stg11:/>lsvg -p test_asyncvg
test_asyncvg:
PV_NAME      PV STATE    TOTAL PPs   FREE PPs   FREE DISTRIBUTION
HA_3          active      512         23        00..00..00..00..23
HA_4          active      512         0         00..00..00..00..00
```

We imported the volume group defined in SiteB on the node in SiteA prior to configuring the cluster resource group. Verify the PVIDs of the volumes in both sites in order to select the appropriate disks. See more details in “Importing the volume group definition to the secondary site” on page 278.

Having the XIV storage agents and XIV storage systems defined at the time of configuring the XIV synchronous replicated resource, we added the second mirror group in the cluster configuration for the asynchronous consistency group PHA712_ASYNC(CG), as shown in Figure 7-34.

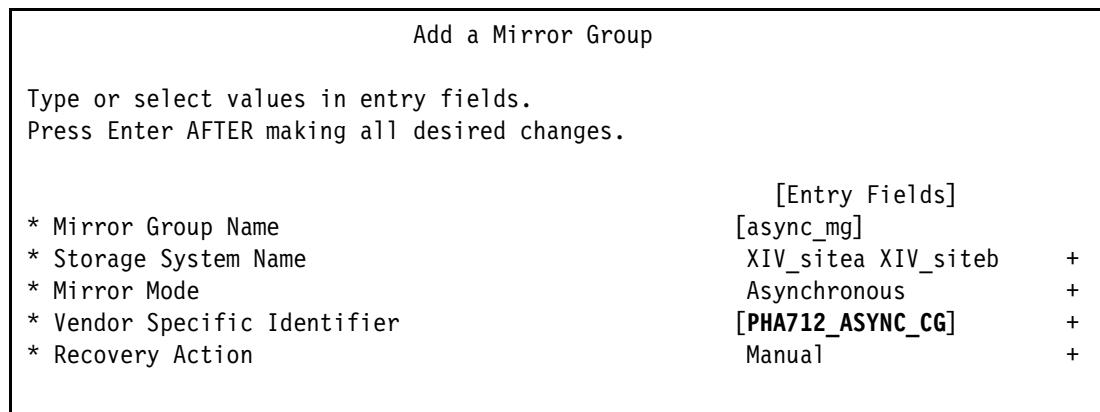


Figure 7-34 Creating the mirror group for the async pairs

We set recovery action for this XIV resource to *manual*, so that a manual intervention is required in the case of a failure of the primary site.

Next we defined a new resource group with the name *async_rg* with inter-site policy “Prefer Primary Site” and the default intra-site policies. Figure 7-35 on page 289 shows how we defined *async_rg* in our environment.

Add a Resource Group (extended)

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

	[Entry Fields]
* Resource Group Name	[async_rg]
Inter-Site Management Policy	[Prefer Primary Site] +
* Participating Nodes from Primary Site	[c581stg11] +
Participating Nodes from Secondary Site	[c581stg10] +
Startup Policy	Online On Home Node Only
Failover Policy	Failover To Next Priority Node In The List
Fallback Policy	Fallback To Higher Priority Node In The List

Figure 7-35 Adding the resource group async_rg

The following resources were added to the resource group `async_rg`, shown in Figure 7-36 on page 290:

- ▶ Service IP address, which can be acquired on all nodes: `c581stgsb`
- ▶ The volume group `test_asyncvg` with all included file systems
- ▶ XIV replicated resource: `async_mp`

Change/Show All Resources and 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	async_rg	
Inter-site Management Policy	Prefer Primary Site	
Participating Nodes from Primary Site	c581stg11	
Participating Nodes from Secondary Site	c581stg10	
Startup Policy	Online On Home Node Only	
Fallover Policy	Fallover To Next Priority Node	
Fallback Policy	In The List	
Fallback Timer Policy (empty is immediate)	Fallback To Higher Priority Node	
	In The List	
	[]	+
Service IP Labels/Addresses	[c581stgsb]	
Application Controller Name	[]	+
Volume Groups	[test_asyncvg]	
Use forced varyon of volume groups, if necessary	false	+
Automatically Import Volume Groups	false	+
Filesystems (empty is ALL for VGs specified)	[]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems mounted before IP configured	false	+
Filesystems/Directories to Export (NFSv2/3)	[]	+
Filesystems/Directories to NFS Mount	[]	+
Network For NFS Mount	[]	+
Tape Resources	[]	+
Raw Disk PVIDs	[]	+
Raw Disk UUIDs/hdisks	[]	+
Disk Error Management?	no	+
Primary Workload Manager Class	[]	+
Secondary Workload Manager Class	[]	+
Miscellaneous Data	[]	
WPAR Name	[]	+
User Defined Resources	[]	+
DS8000 Global Mirror Replicated Resources	[]	+
XIV Replicated Resources	async_mg	+
DS8000-Metro Mirror (In-band) Resources		+

Figure 7-36 Adding resources to resource group `async_rg`

Lastly, verify and synchronize the cluster configuration. At the end of the process, since the cluster is active, it activates the resource `async_rg` on the primary node `c581stg11`. Verify the cluster resource group status using the `c1RGinfo` command as shown in Example 7-14.

Example 7-14 Resource group status

```
root@c581stg11:/>clRGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: sync_rg

Node	Primary State	Secondary State
c581stg10@sitea	ONLINE	OFFLINE
c581stg11@siteb	OFFLINE	ONLINE SECONDARY

Resource Group Name: async_rg

Node	Primary State	Secondary State
c581stg11@siteb	ONLINE	OFFLINE
c581stg10@sitea	OFFLINE	ONLINE SECONDARY

7.5.4 Testing the cluster

In a real case scenario, multiple failures can occur at the hardware or software level, on the infrastructure, or in the application software. The local failures such as adapter or node or disk failures are subject to a local high availability within a site. PowerHA Standard Edition is the clustering solution for high availability for a data center. In this section, we focus on test scenarios specific for a cluster with XIV replicated resources and a multisite configuration.

We used our environment with both sync and async XIV replicated resources within the same cluster to better emphasize the resource group states in correlation with the XIV remote mirror pair states during the failures.

For testing the cross-site cluster configuration, we simulated the following events:

- ▶ Planned site failover: Graceful site failover
- ▶ Unplanned site outage
- ▶ Loss of storage access in a site

The base starting point for our test scenarios was the cluster state shown in Example 7-15.

Example 7-15 Base starting point of the resource group status for our tests

```
root@c581stg11:/>clRGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: sync_rg

Node	Primary State	Secondary State
c581stg10@sitea	ONLINE	OFFLINE
c581stg11@siteb	OFFLINE	ONLINE SECONDARY

Resource Group Name: async_rg

Node	Primary State	Secondary State
c581stg11@siteb	ONLINE	OFFLINE
c581stg10@sitea	OFFLINE	ONLINE SECONDARY

For this resource group state, we also had the state of the XIV mirroring relationships on the XIV systems shown in Figure 7-37 and Figure 7-38.

Name	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes				
PHA712_ASYNC_CG	S 00:01:00	RPO OK	PHA712_ASYNC_CG	XIV pPETXIV4
PHA712_SYNC_CG	M 00:01:00	Synchronized	PHA712_SYNC_CG	XIV pPETXIV4

Figure 7-37 Status of the consistency groups on XIV SiteA

Name	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes				
PHA712_ASYNC_CG	M 00:01:00	RPO OK	PHA712_ASYNC_CG	XIV 1310088
PHA712_SYNC_CG	S 00:01:00	Consistent	PHA712_SYNC_CG	XIV 1310088

Figure 7-38 Status of the consistency groups on XIV SiteB

Site graceful failover

We performed a graceful failover of a site by moving the resource group sync_rg from primary site (SiteA) to the secondary site (SiteB) using C-SPOC. We also used a second example with async_rg, which has the primary node in SiteB by stopping the cluster services on the node c581stg11 in SiteB with the resource group movement option.

The graceful failover test is useful because it tests both the release and the acquisition of the resource groups, including application stop and start scripts.

During a resource group move from a primary site to a secondary one, the following operations take place:

- ▶ Release the primary online instance of the resource group in the primary site. This operation performs the following actions within the site holding the resource group:
 - Stops the application services and the associated monitors.
 - Detaches the service IP addresses.
 - Unmounts the file systems.
 - Varies off the volume groups.
- ▶ Release the secondary online instance of the resource group at the secondary site.
- ▶ Acquire the resource group in the secondary online state in the primary site of the resource group.
- ▶ Acquire the resource group in the online primary state at the secondary site. When acquiring the online primary state, the cluster performs the following operations:
 - Varies on the volumes groups.
 - Mounts the file systems.
 - Acquires the service IP address.
 - Starts the application services and monitors.

We perform the resource group movement using SMIT by executing **smitty sysmirror** → **System Management (C-SPOC)** → **Resource Group and Applications** → **Move Resource Groups to Another Site**, then selecting the resource group sync_rg associated with the online state and with the primary node c581stg10, and selecting next the destination site, SiteB in our case. See the SMIT menu shown in Figure 7-39 on page 293.

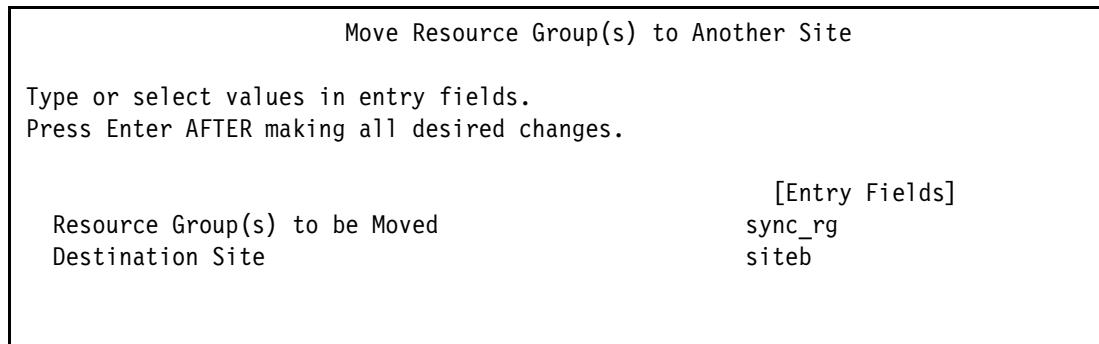


Figure 7-39 Move the resource group to the other site

Alternatively, you can use the **c1RGmove** command to perform the resource group move operation:

```
/usr/es/sbin/cluster/utilities/c1RGmove -s 'false' -x -i -g 'sync_rg' -n 'siteb'
```

The cluster swaps the roles for online primary and secondary between the two nodes. The final state of the resource group is shown in Example 7-16.

Example 7-16 Cluster resource group status after the move operation

```
root@c581stg11:/>c1RGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: **sync_rg**

Node	Primary State	Secondary State
c581stg10@sitea	OFFLINE	ONLINE SECONDARY
c581stg11@siteb	ONLINE	OFFLINE

Resource Group Name: **async_rg**

Node	Primary State	Secondary State
c581stg11@siteb	ONLINE	OFFLINE
c581stg10@sitea	OFFLINE	ONLINE SECONDARY

At the XIV storage level, we observed that the sync relationship between the CGs had been swapped. The master role was moved to the volumes in SiteB, and the slave role was acquired by the volumes in SiteA, as shown in Figure 7-40 on page 294.

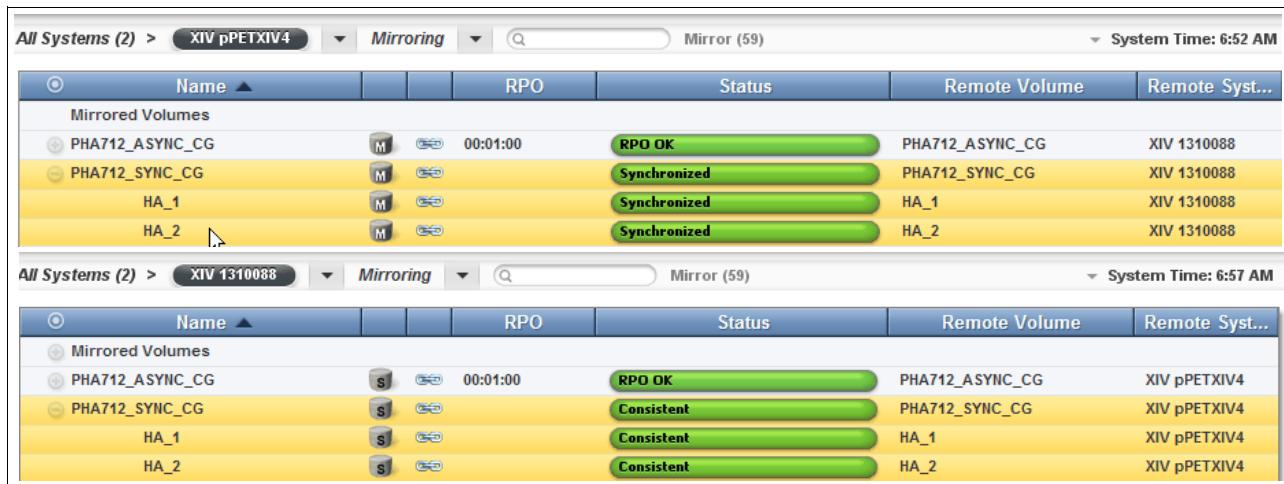


Figure 7-40 XIV pair status after resource group move (view from both XIVs)

You can move back the resource group to the primary site (SiteA) by using the same operation either from SMIT menus or with the **c1RGmove** command. See Example 7-17 for how we moved the sync_rg resource group from SiteB to SiteA using **c1RGmove**.

Example 7-17 Resource group sync_rg move back to SiteA

```
[c581stg10] [/]> c1RGmove -s 'false' -x -i -g 'sync_rg' -n 'sitea'
Attempting to move group sync_rg to site sitea.
```

Waiting for the cluster to process the resource group movement request....

Waiting for the cluster to stabilize.....

Resource group sync_rg is online on site sitea.

Cluster Name: xiv_cluster

Resource Group Name: sync_rg			
Node	Primary State	Secondary State	
c581stg10@sitea	ONLINE	OFFLINE	
c581stg11@siteb	OFFLINE	ONLINE SECONDARY	

Resource Group Name: async_rg			
Node	Primary State	Secondary State	
c581stg11@siteb	ONLINE	OFFLINE	
c581stg10@sitea	OFFLINE	ONLINE SECONDARY	

For a second scenario case, we stopped the cluster services on the node in SiteB using the option to move the resource group shown in Figure 7-41 on page 295.

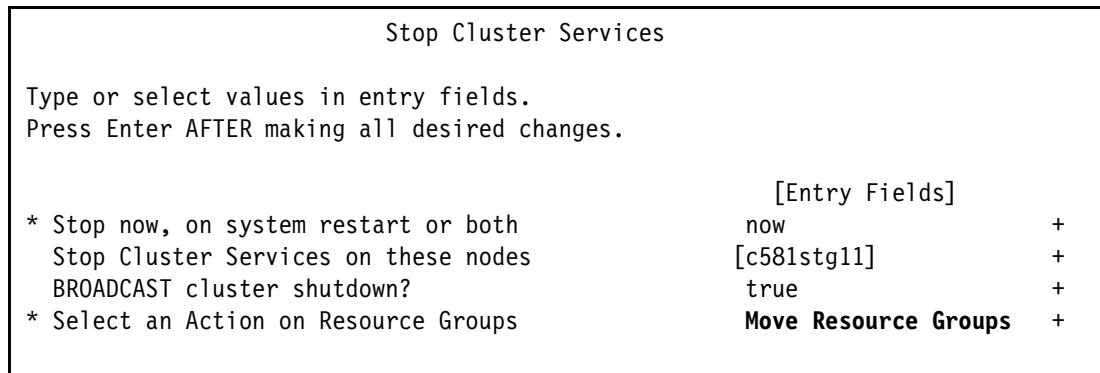


Figure 7-41 Stopping the cluster services with the option of resource group moving

The node c581stg11 currently holding the resource group async_rg stops the cluster services, and the resource group is moved to the other site. The final resource group status is shown in Example 7-18.

Example 7-18 Resource group status after stopping the node in SiteB

```
[c581stg10] [/]> clRGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: sync_rg

Node	Primary State	Secondary State
c581stg10@sitea	ONLINE	OFFLINE
c581stg11@siteb	OFFLINE	OFFLINE

Resource Group Name: async_rg

Node	Primary State	Secondary State
c581stg11@siteb	OFFLINE	OFFLINE
c581stg10@sitea	ONLINE	OFFLINE

Observe that async_rg was acquired on the secondary site of this resource group even if the recovery action is set to MANUAL. This is the normal behavior while moving the resource groups across the sites. The manual intervention is required only in case of a total site failure, including the storage in the failed site. In this case, the XIV consistency group PHA712_ASYNC_CG swaps out the roles, having now the master peers in SiteA and the slave peers in SiteB as shown in Figure 7-42 on page 296.

Name	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes				
PHA712_ASYNC_CG	00:01:00	RPO OK	PHA712_ASYNC_CG	XIV pPETXIV4
HA_3	00:01:00	RPO OK	HA_3	XIV pPETXIV4
HA_4	00:01:00	RPO OK	HA_4	XIV pPETXIV4

Name	RPO	Status	Remote Volume	Remote Syst...
Mirrored Volumes				
PHA712_ASYNC_CG	00:01:00	RPO OK	PHA712_ASYNC_CG	XIV 1310088
HA_3	00:01:00	RPO OK	HA_3	XIV 1310088
HA_4	00:01:00	RPO OK	HA_4	XIV 1310088
PHA712_SYNC_CG		Consistent	PHA712_SYNC_CG	XIV 1310088

Figure 7-42 XIV remote mirroring status (from both sides)

At the time of node reintegration in the cluster, the resource group `async_rg` is automatically acquired on the primary site (SiteB). This is because the inter-site policy for this resource group is set to Prefer Primary Site. We started the cluster services on node `c581stg11` using `smitty clstart`.

After node reintegration in the cluster and resource group movement to SiteB, the resource group status is back to the original state. See the `clRGinfo` state at the beginning of our test scenarios in Example 7-15 on page 291.

Site failover test

During this test we simulated an unplanned interruption in one of the sites by halting the node in that site. Having a single node in our configuration, a site failover occurred and the expected result was the resource group being activated in the second site.

For our test case, we halted node `c581stg10` in SiteA with the command `halt -q`. The result was that the resource group `sync_rg` was acquired on SiteB. There were no online secondary instances since there were no remaining active nodes in SiteA to assume this role. See the cluster resource group status after halting the node in SiteA in Example 7-19.

Example 7-19 Cluster resource group status after SiteA failure

```
root@c581stg11:/> clRGinfo -p
```

Cluster Name: `xiv_cluster`

Resource Group Name: `sync_rg`

Node	Primary State	Secondary State
<code>c581stg10@sitea</code>	OFFLINE	OFFLINE
<code>c581stg11@siteb</code>	ONLINE	OFFLINE

Resource Group Name: `async_rg`

Node	Primary State	Secondary State
<code>c581stg11@siteb</code>	ONLINE	OFFLINE
<code>c581stg10@sitea</code>	OFFLINE	OFFLINE

The status of the XIV mirroring pairs is illustrated in Figure 7-43 on page 297.

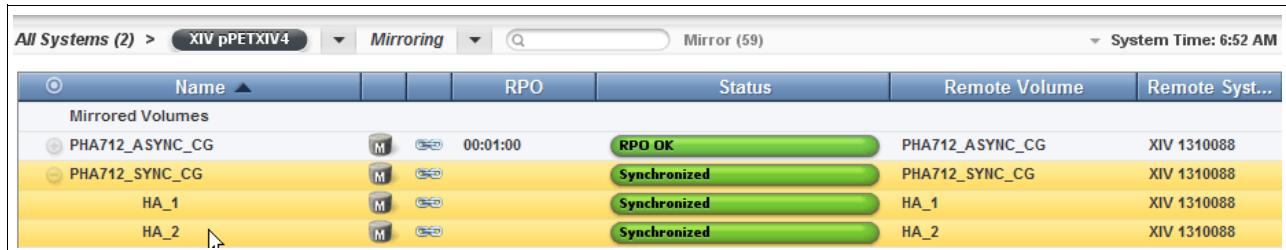


Figure 7-43 Status of the XIV remote mirror pairs after nodes in a site have failed (SiteB view)

Note: The status of the mirroring relationships is *Synchronized* for the synchronous pairs because we simulated an unplanned node failure in SiteA while the storage there was still active. If the storage were also down (total site failure), the mirror peers now having the master role in the secondary site, after the failover, would get the status *Unsynchronized*.

Node reintegration in a cluster

After the node c581stg10 was active again, we reintegrated it to the cluster by starting the cluster services. Because the inter-site policy for the resource group sync_vg having the primary site in SiteA and the secondary in SiteB is Online on Either Site after node reintegration, the resource group status does not change. Only the online secondary instances for both resource groups sync_rg and async_rg will be acquired in SiteA as shown in Example 7-20.

Example 7-20 Node c581stg10 reintegration in cluster

Cluster Name: xiv_cluster

Resource Group Name: **sync_rg**

Node	Primary State	Secondary State
c581stg10@sitea	OFFLINE	ONLINE SECONDARY
c581stg11@siteb	ONLINE	OFFLINE

Resource Group Name: **async_rg**

Node	Primary State	Secondary State
c581stg11@siteb	ONLINE	OFFLINE
c581stg10@sitea	OFFLINE	ONLINE SECONDARY

For bringing back the resource group sync_rg to the primary site SiteA, we needed to move the resource group to the other site. See this operation in Example 7-17 on page 294.

Storage access failure in a site

Loss of access to the storage in one site can be caused by multiple factors. This includes Fibre Channel or iSCSI communication switches and gateways or host HBAs, disk failures, or storage failure. We mitigated them by providing redundancy on each level—multiple HBAs on a host, multiple communication devices (FC switches, Ethernet and gateways), redundant access to storage devices, volume redundancy (such as RAID array devices). However, a disk subsystem failure can result in a serious outage to a production environment. PowerHA SystemMirror using storage replicated resources reacts to such a failure when the volume group part of the resource group becomes unavailable. In that case, a node and/or site failure may occur.

A failure of a host accessing a volume group included in a replicated resource will determine a failover of the resource group on another host available in the chain for acquiring the resource. In case of a total storage failure, this event determines a site failover and the resource group is activated at another site.

To simulate the storage access failure in our environment, we unmapped all the XIV volumes in sitea to node c581stg10. As the result of volume group test_syncvg not being accessible on node c581stg10, the associated resource group sync_rg will failover to SiteB.

The cluster resource group status is shown in Example 7-21.

Example 7-21 Cluster resource status after storage failure in SiteA

```
root@c581stg11:/>c1RGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: sync_rg

Node	Primary State	Secondary State
c581stg10@sitea	OFFLINE	ONLINE SECONDARY
c581stg11@siteb	ONLINE	OFFLINE

Resource Group Name: async_rg

Node	Primary State	Secondary State
c581stg11@siteb	ONLINE	OFFLINE
c581stg10@sitea	OFFLINE	ONLINE SECONDARY

You can observe that sync_rg is now active in SiteB while the ONLINE SECONDARY status is acquired in SiteA for the resource group. This is because node c581stg10 is still up and the network communications are active with the node in SiteB.

Figure 7-44 shows the status of the pairs on the XIV system in SiteB, after failover of the resource group sync_rg in SiteB.

Name	RPO	Status	Remote Volume	Remote Syst...
PHA712_ASYNC_CG	00:01:00	RPO OK	PHA712_ASYNC_CG	XIV 1310088
PHA712_SYNC_CG		Synchronized	PHA712_SYNC_CG	XIV 1310088
HA_1		Synchronized	HA_1	XIV 1310088
HA_2		Synchronized	HA_2	XIV 1310088

Figure 7-44 XIV mirroring status after resource group sync_rg failover to SiteA

Repository disk considerations

In our environment, as a side effect of losing the access to the storage in SiteA, the repository disk *caarepo_a* was no longer available. The cluster service recurrently logs an error message in the log file *hacmp.out* as shown in Example 7-22 on page 299.

Example 7-22 Log file hacmp.out: Error message for the repository disk not being accessible

```
...
ERROR: rep_disk_notify : Fri Nov 23 10:44:04 CST 2012 : Node c581stg10 on Cluster
xiv_cluster has lost access to repository disk caarepo_a. Please recover from this
error or replace the repository
disk using smitty.
```

The status of the CAA repository disk can also be verified with the **lscluster -d** command in AIX as shown in Example 7-23.

Example 7-23 lscluster command output

```
[c581stg10] [/var/hacmp/log]> lscluster -d
Storage Interface Query

Cluster Name: xiv_cluster
Cluster UUID: 27af17f0-32b0-11e2-b3e9-00145e77b4d4
Number of nodes reporting = 2
Number of nodes expected = 2

Node c581stg10
Node UUID = 27b0bb50-32b0-11e2-b3e9-00145e77b4d4
Number of disks discovered = 1
    caarepo_a:
        State : DOWN
        uDid : 2611200173800276803FA072810XIV03IBMfcP
        uUid : 7311d812-619d-0cc2-033f-bdb51dd9af61
        Site uUid : 27a630b8-32b0-11e2-b3e9-00145e77b4d4
        Type : REPDISK

Node c581stg11
Node UUID = 3129da86-32b0-11e2-b4bd-00145e77b4d4
Number of disks discovered = 1
    caarepo_b:
        State : UP
        uDid :
        uUid : 09868313-07bc-80d4-9198-805377d2071c
        Site uUid : 311fe06c-32b0-11e2-b4bd-00145e77b4d4
        Type : REPDISK
```

Note: For a linked cluster, while the repository disk access is lost in a single site, we observed that the cluster verification and synchronization operations can be performed for limited operations. Particularly, operations not related to the cluster topology changes, such as changing the attributes of a mirror group are allowed.

You can recover from the repository disk access error either by restoring the disk access communication (if possible) or by replacing the repository disk. If a repository disk backup has not yet been defined, you can perform the following actions, assuming there is a new disk available on all nodes in the site for the CAA repository:

1. Add a new repository disk by executing **smitty sysmirror** → **Cluster Nodes and Networks** → **Manage Repository Disks** → **Add a Repository Disk**. See an example in Figure 7-45 on page 300.

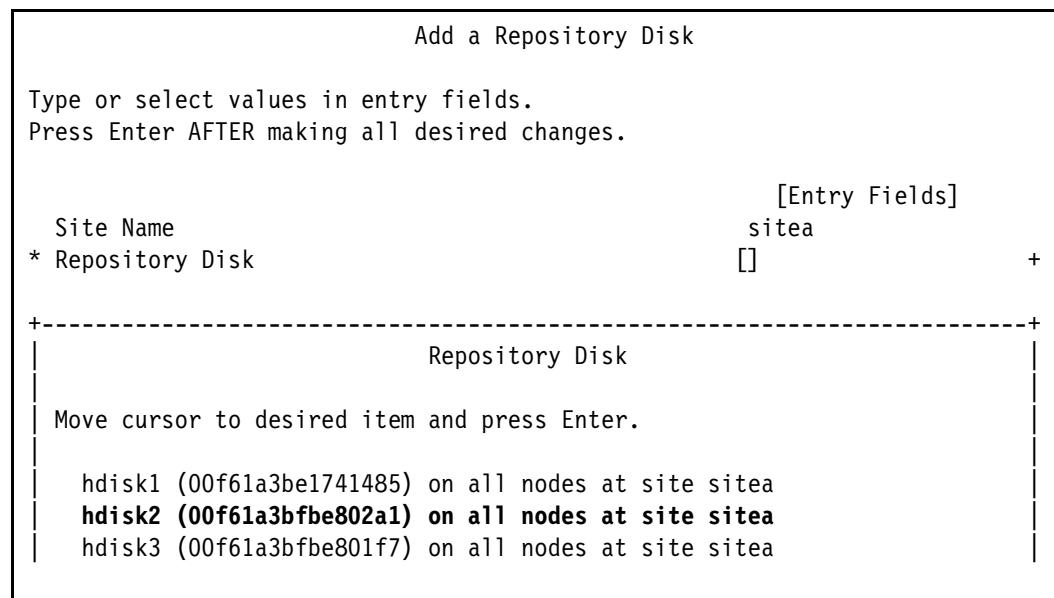


Figure 7-45 Adding a new repository disk

2. Replace the current disk with the newly added one by executing **smitty sysmirror** → **Problem Determination Tools** → **Replace the Primary Repository Disk**, then select the site and then select the PVID of the new candidate disk, as shown in Figure 7-46.

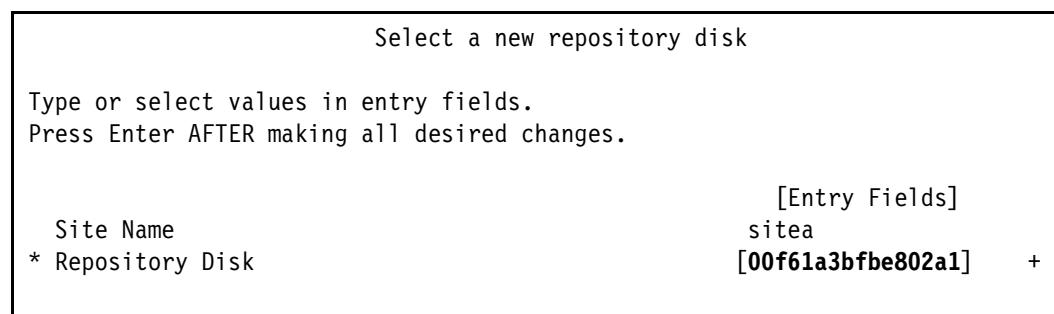


Figure 7-46 Replacing the primary repository disk

3. Run the cluster verification and synchronization task. You can use SMIT menus by executing **smitty sysmirror** → **Cluster Nodes and Networks** → **Verify and Synchronize Cluster Configuration**.

Restoring the storage access

For restoring the disk access of node c581stg10 in SiteA, we reattached the volumes assigned to node c581stg10 to the XIV system in SiteA, in their original configuration. In our environment, by restoring the disk access to all the volumes previously defined, the access to the CAA repository disk was restored. When the repository disk was available again, a message was logged in the hacmp.out file as shown in Example 7-24.

Example 7-24 hacmp.out - Repository disk available

```
.....  
rep_disk_notify: Fri Nov 23 10:58:50 CST 2012 : Access to repository disk has been  
restored on Node c581stg10
```

In our test environment, when remapping back all the disks to the host c581stg10 in their original configuration, we used the same LUN ID as before unmapping, so that their definitions in ODM remained consistent with their new mappings.

We verified that the disks were accessible for read operations on node c581stg10 with the **lquerypv** command, as shown in Example 7-25.

Example 7-25 Checking the read access for a disk

```
[c581stg10] [/var/hacmp/log]> lquerypv -h /dev/HA_2
00000000 C9C2D4C1 00000000 00000000 00000000 |.....
00000010 00000000 00000000 00000000 00000000 |.....
00000020 00000000 00000000 00000000 00000000 |.....
00000030 00000000 00000000 00000000 00000000 |.....
00000040 00000000 00000000 00000000 00000000 |.....
00000050 00000000 00000000 00000000 00000000 |.....
00000060 00000000 00000000 00000000 00000000 |.....
00000070 00000000 00000000 00000000 00000000 |.....
00000080 00F61A3B 1BB46135 00000000 00000000 |...;..a5...
00000090 00000000 00000000 00000000 00000000 |.....
000000A0 00000000 00000000 00000000 00000000 |.....
000000B0 00000000 00000000 00000000 00000000 |.....
000000C0 00000000 00000000 00000000 00000000 |.....
000000D0 00000000 00000000 00000000 00000000 |.....
000000E0 00000000 00000000 00000000 00000000 |.....
000000F0 00000000 00000000 00000000 00000000 |.....
```

After the disks were available again in SiteA, we checked the status of the XIV remote mirroring pairs in SiteA to be active. In case of the storage being down or a failure of the replication links between the XIV systems, the remote mirroring pairs need to be synchronized. We checked the status of the mirroring pairs at both sides using the XIV GUI as shown in Figure 7-47.

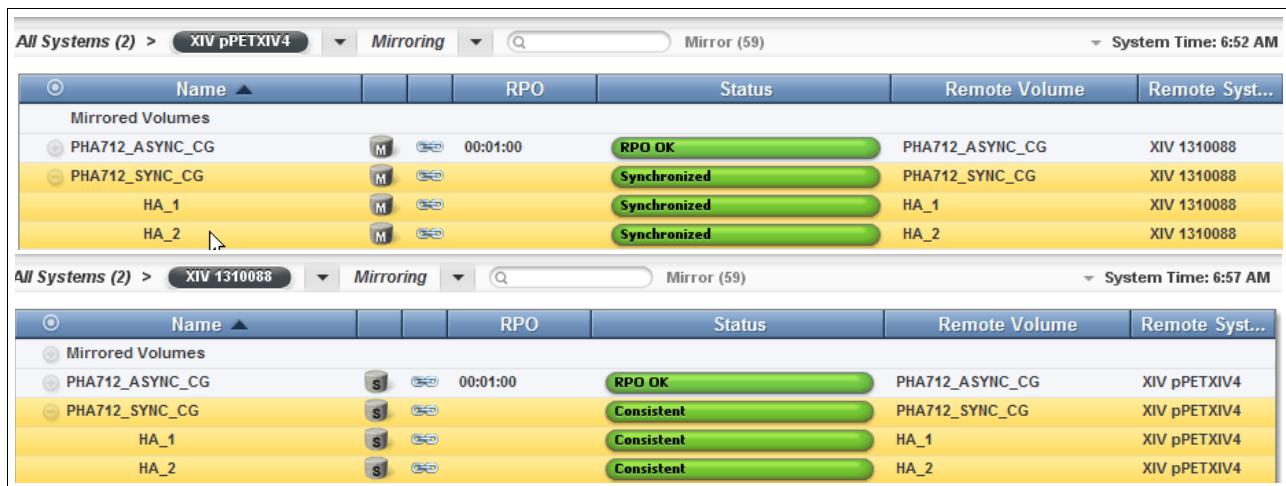


Figure 7-47 XIV remote mirroring status (XIV SiteB, XIV SiteA)

At this time we moved the sync_rg resource group back to the primary site SiteA using the **c1RGmove** command. See this operation described in Example 7-17 on page 294.

Attention: When testing the storage recovery scenario, an error was observed for mounting back the file system from the failed-over resource group (sync_rg). The error was related to APAR IV30226, which had not been released for AIX 6.1 at the time of testing:

<http://www-01.ibm.com/support/docview.wss?uid=isg1IV30226>

The workaround is to stop the cluster services on the affected node, reboot it and reintegrate it back into the cluster.

7.6 Creating an XIV replicated cluster via `clmgr` and XIV CLI

In this section, we show how to use `xcli` to configure remote mirror couplings and `clmgr` to configure the PowerHA cluster.

7.6.1 Implementation overview

We used a simple example to implement PowerHA SystemMirror 7.1.2 linked cluster with XIV remote mirroring. In this configuration, we used two nodes and one XIV storage at each site. We used two different models of XIV storage system Gen3. We had a model 2810-114 at Site1 and model 2810-214 at Site2. An overview of our PowerHA SystemMirror 7.1.2 linked cluster with XIV remote mirror replicated resources is shown in Figure 7-48 on page 303.

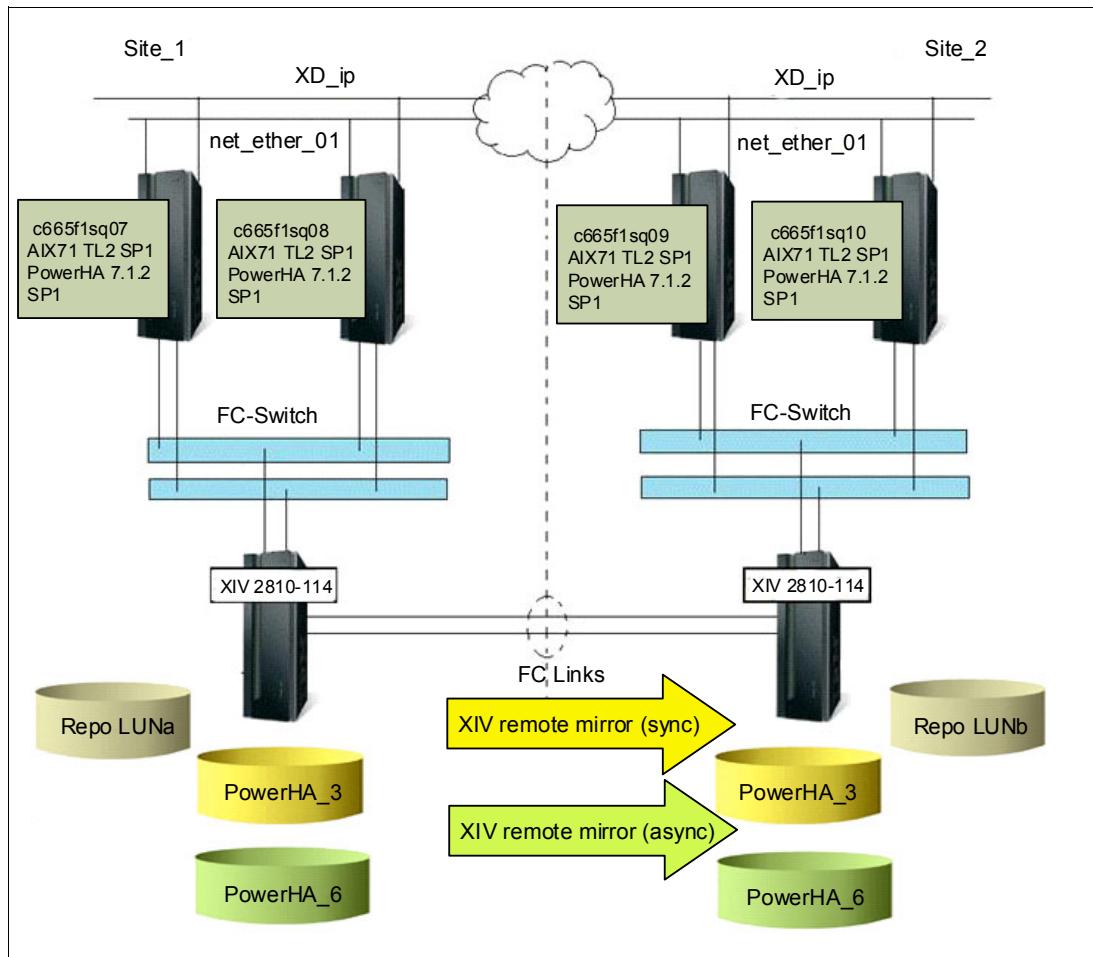


Figure 7-48 PowerHA linked cluster with XIV remote mirror replicated resources

Both the XIV storage and cluster nodes are connected to a dual FC switch in each site. The physical LUN “repo LUNa” from XIV 2810-114 at Site1 is shared across nodes c665f1sq07 and c665f1sq08. The physical LUN “repo LUNb” from XIV 2810-214 at Site2 is shared across nodes c665f1sq09 and c665f1sq10. These two LUNs are used as repository disks in their respective sites for our linked cluster configuration.

The XIV storages communicate with each other via FC links. FC links provide XIV remote mirror couplings. In our scenario, PowerHA_3 LUN on each storage mirrored synchronously and PowerHA_6 LUN on each storage mirrored asynchronously. Refer to 7.6.2, “Configuring XIV remote mirror couplings using xcli” on page 304 for configuring XIV remote mirror couplings.

We defined two networks in our configuration:

- ▶ *net_ether_01* was the client access network of type *ether*. We also defined service IP addresses that could be acquired on all cluster nodes.
- ▶ *XD_ip* was a heartbeat network of type *XD_ip*. It is used as an additional communication path between sites.

The communication interfaces used in this scenario are shown in Example 7-26.

Example 7-26 Network interfaces

NODE c665f1sq07:

```

Network net_XD_ip_01
    c665f1sq07      9.114.135.71
Network net_ether_01
    f1sq08_ensvc   30.30.30.66
    f1sq07_ensvc   30.30.30.65
    f1sq07_enboot  30.30.30.1
    f1sq07_enstby1 30.30.30.129

NODE c665f1sq08:
    Network net_XD_ip_01
        c665f1sq08      9.114.135.72
    Network net_ether_01
        f1sq08_ensvc   30.30.30.66
        f1sq07_ensvc   30.30.30.65
        f1sq08_enstby1 30.30.30.130
        f1sq08_enboot  30.30.30.2

NODE c665f1sq09:
    Network net_XD_ip_01
        c665f1sq09      9.114.135.73
    Network net_ether_01
        f1sq08_ensvc   30.30.30.66
        f1sq07_ensvc   30.30.30.65
        f1sq09_enboot  30.30.30.3
        f1sq09_enstby1 30.30.30.131

NODE c665f1sq10:
    Network net_XD_ip_01
        c665f1sq10      9.114.135.74
    Network net_ether_01
        f1sq08_ensvc   30.30.30.66
        f1sq07_ensvc   30.30.30.65
        f1sq10_enstby1 30.30.30.132
        f1sq10_enboot  30.30.30.4

```

7.6.2 Configuring XIV remote mirror couplings using xcli

In this section, we configure XIV remote mirror couplings using the `xcli` command. We assumed that XIV LUNs were already mapped to the cluster nodes. XIV storage systems details used in our configuration are shown in Example 7-27.

Example 7-27 XIV storage system details

```

# xcli -u admin -p adminadmin -m 9.114.63.166 config_get name=machine_type
Name          Value
machine_type 2810

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.166 config_get name=machine_model
Name          Value
machine_model 114

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 config_get name=machine_type
Name          Value

```

```
machine_type 2810
```

```
(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 config_get name=machine_model
Name          Value
machine_model 214
```

For more information about **xcli** commands, refer to:

http://publib.boulder.ibm.com/infocenter/ibmxiv/r2/topic/com.ibm.help.xivgen3.doc/docs/xiv_11.1.x_xcli.pdf

XIV storage system 2810-114 is connected to nodes c665f1sq07 and c665f1sq08 at Site1. XIV storage system 2810-214 is connected to nodes c665f1sq09 and c665f1sq10 at Site2. The host list mapping from **xcli** is shown in Example 7-28.

Example 7-28 Host lists

```
# xcli -u admin -p adminadmin -m 9.114.63.166 host_list | grep c655
c655f1sq07 default 10000000C94F8D88,10000000C9424EC3 PowerHA none
c655f1sq08 default 10000000C950245B,10000000C940B790 PowerHA none

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 host_list | grep c655
c655f1sq09 default 10000000C945D3A3,10000000C94255FC PowerHA none
c655f1sq10 default 10000000C95021D0,10000000C9425C7D PowerHA none
```

PowerHA_3 and PowerHA_6 volumes from the XIV 2810-114 storage system at Site1 are mapped to c665f1sq07 and c665f1sq08 hosts in the same site. Another pair of volumes called PowerHA_3 and PowerHA_6 from the XIV 2810-214 storage system at Site2 are mapped to c665f1sq09 and c665f1sq10 hosts in the same site. Example 7-29 shows volume mapping on each host.

Example 7-29 Volume mapping to hosts

```
# xcli -u admin -p adminadmin -m 9.114.63.166 mapping_list host=c655f1sq07 | grep
PowerHA_[36]
1    PowerHA_3   17           2299           no
4    PowerHA_6   17           2314           no

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.166 mapping_list host=c655f1sq08 | grep
PowerHA_[36]
1    PowerHA_3   17           2299           no
4    PowerHA_6   17           2314           no

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 mapping_list host=c655f1sq09 | grep
PowerHA_[36]
1    PowerHA_3   17           143            yes
4    PowerHA_6   17           146            yes

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 mapping_list host=c655f1sq10 | grep
PowerHA_[36]
1    PowerHA_3   17           143            yes
```

To allocate space for these volumes, storage pools must already be defined. In our case, we used the storage pool named PowerHA_Pool, as shown in Example 7-30.

Example 7-30 Storage pool used to define volumes

```
# xcli -u admin -p adminadmin -m 9.114.63.166 vol_list | grep PowerHA_[36]
PowerHA_3 17          PowerHA_Pool    admin      0
PowerHA_6 17          PowerHA_Pool    admin      0

(0) root @ c665f1sq07: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 vol_list | grep PowerHA_[36]
PowerHA_3 17          PowerHA_Pool    xiv_development 0
PowerHA_6 17          PowerHA_Pool    xiv_development 0
```

You can check storage pool details with the **pool_list** command shown in Example 7-31.

Example 7-31 PowerHA_Pool storage pool details

```
# xcli -u admin -p adminadmin -m 9.114.63.166 pool_list pool=PowerHA_Pool
Name      Size(GB) SoftVols(GB) SnapSize(GB) SoftEmpty(GB) HardSize(GB) HardVols(GB) Locked HardSnaps(GB)
PowerHA_Pool 240     172        51           17          240         34       no      0
Hard Empty (GB)
206

# xcli -u admin -p adminadmin -m 9.114.63.163 pool_list pool=PowerHA_Pool
Name      Size(GB) SoftVols(GB) SnapSize(GB) SoftEmpty(GB) HardSize(GB) HardVols(GB) Locked HardSnaps(GB)
PPowerHA_Pool 240     172        51           17          240         34       no      0
Hard Empty (GB)
206
```

We created both synchronous and asynchronous XIV remote mirror couplings for our test scenario. PowerHA_3 volumes on both sites are used for synchronous remote mirror coupling, and PowerHA_6 volumes on both sites are used for asynchronous remote mirror coupling.

Synchronous mirror coupling

We now created a consistency group named redbook_sync_cg using the xcli-based command **cg_create**. This is executed on both master and slave XIV storages. We refer to Site1 storage as *master* storage and Site2 storage as *slave* storage. Example 7-32 shows the consistency group creation.

Example 7-32 Creating a consistency group on master and slave storage

```
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 cg_create
cg=redbook_sync_cg pool=PowerHA_Pool
Command executed successfully.

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 cg_create
cg=redbook_sync_cg pool=PowerHA_Pool
Command executed successfully.
```

The syntax for **cg_create** is as follows:

```
cg_create cg=<consistency group> pool=<storage pool name>
```

Verify that the consistency group was created on both master and slave via the **cg_list** command shown in Example 7-33.

Example 7-33 Checking the consistency group on both master and slave

```
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 cg_list | grep  
redbook_sync_cg  
redbook_sync_cg PowerHA_Pool  
  
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 cg_list | grep  
redbook_sync_cg  
redbook_sync_cg PowerHA_Pool
```

Create a mirroring pair between the master and the slave consistency group redbook_sync_cg. Run the **mirror_create** command from the master storage. The syntax for **mirror_create** is as follows:

```
mirror_create < vol=VolName slave_vol=SlaveVolumeName  
[ create_slave=<Yes|No> [ remote_pool=RemotePoolName ] ]  
[ init_type=<online|offline> ] > | <cg=CgName slave_cg=SlaveCgName>  
[ type=<SYNC_BEST EFFORT|ASYNC_INTERVAL> ]  
target=TargetName [ rpo=rpo [ remote_rpo=rpo ]  
schedule=Schedule remote_schedule=Schedule ]
```

The **mirror_create** command utilizes the target XIV storage system name. You can find the XIV storage system names with the **config_get** command, as shown in Example 7-34. The XIV 1310088 system is master and the XIV pPETXIV4 system is slave.

Example 7-34 XIV storage system names

```
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 config_get  
name=system_name  
Name Value  
system_name XIV 1310088  
  
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 config_get  
name=system_name  
Name Value  
system_name XIV pPETXIV4
```

By default, the command creates a synchronous mirroring pair between the consistency groups. We created a mirroring pair between the master and slave consistency group redbook_sync_cg, as shown in Example 7-35.

Example 7-35 Creating synchronous mirroring pair group between consistency groups

```
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_create  
target="XIV pPETXIV4" cg=redbook_sync_cg slave_cg=redbook_sync_cg  
Command executed successfully.  
(0) root @ c665f1sq07: /
```

Verify that a synchronous mirroring pair was created by using the **mirror_list** command. Also verify that the status shows *synchronized* on master storage and *consistent* on slave storage. Since we only defined, and not activated, the mirroring pair, the mirror active status shows no as seen in Example 7-36.

Example 7-36 Checking consistency group mirroring status

```
(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list cg=redbook_sync_cg
Name           Mirror Type      Mirror Object   Role    Remote System   Remote Peer      Active
redbook_sync_cg sync_best_effort  CG            Master XIV pPETXIV4  redbook_sync_cg  no
Status         Link Up
Synchronized   yes

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list cg=redbook_sync_cg
Name           Mirror Type      Mirror Object   Role    Remote System   Remote Peer      Active
redbook_sync_cg sync_best_effort  CG            Slave   XIV 1310088  redbook_sync_cg  no
Status         Link Up
Consistent     yes
```

We now activate the consistency group redbook_sync_cg using the **mirror_activate** command shown in Example 7-37.

Example 7-37 Activating a synchronous mirroring pair consistency group

```
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_activate
cg=redbook_sync_cg
Command executed successfully.
```

Verify that the consistency group synchronous remote mirroring pair is now activated. The *Active* column shows yes, as shown in Example 7-38.

Example 7-38 Checking mirror active status

```
(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list cg=redbook_sync_cg
Name           Mirror Type      Mirror Object   Role    Remote System   Remote Peer      Active
redbook_sync_cg sync_best_effort  CG            Master XIV pPETXIV4  redbook_sync_cg  yes
Status         Link Up
Synchronized   yes

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list cg=redbook_sync_cg
Name           Mirror Type      Mirror Object   Role    Remote System   Remote Peer      Active
redbook_sync_cg sync_best_effort  CG            Slave   XIV 1310088  redbook_sync_cg  yes
Status         Link Up
Consistent     yes
(0) root @ c665f1sq07: /
```

We used the PowerHA_3 volume on both storages for synchronous remote mirror coupling. Create a synchronous mirroring pair between the PowerHA_3 volume on both master and slave storages, as shown in Example 7-39.

Example 7-39 Creating a synchronous mirroring pair between volumes

```
## xcli -u admin -p adminadmin -m 9.114.63.166 mirror_create target="XIV pPETXIV4"
vol=PowerHA_3 slave_vol=PowerHA_3
```

Command executed successfully.

Verify that the synchronous mirroring pair was created and the mirror status shows initializing using the **mirror_list** command, as shown in Example 7-40.

Example 7-40 Checking mirroring pair status

```
# xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list vol=PowerHA_3
Name      Mirror Type      Mirror Object    Role   Remote System   Remote Peer   Active   Status
PowerHA_3  sync_best_effort  Volume        Master  XIV pPETXIV4  PowerHA_3    no       Initializing
Link Up
yes
# xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list vol=PowerHA_3
Name      Mirror Type      Mirror Object    Role   Remote System   Remote Peer   Active   Status
PowerHA_3  sync_best_effort  Volume        Slave   XIV 1310088  PowerHA_3    no       Initializing
Link Up
yes
```

Now activate the PowerHA_3 volume synchronous remote mirroring pair using the **mirror_activate** command shown in Example 7-41.

Example 7-41 Activating the volume mirroring pair

```
# xcli -u admin -p adminadmin -m 9.114.63.166 mirror_activate vol=PowerHA_3
Command executed successfully.
```

Check the PowerHA_3 volume synchronous remote mirroring pair active status using the **mirror_list** command shown in Example 7-42.

Example 7-42 Mirror coupling active status

```
# xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list vol=PowerHA_3
Name      Mirror Type      Mirror Object    Role   Remote System   Remote Peer   Active   Status
PowerHA_3  sync_best_effort  Volume        Master  XIV pPETXIV4  PowerHA_3    yes     Synchronized
Link Up
yes

# xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list vol=PowerHA_3
Name      Mirror Type      Mirror Object    Role   Remote System   Remote Peer   Active   Status
PowerHA_3  sync_best_effort  Volume        Slave   XIV 1310088  PowerHA_3    yes     Consistent
Link Up
yes
```

Now add the volume PowerHA_3 to the consistency group redbook_sync_cg using the **cg_add_vol** command shown in Example 7-43.

Example 7-43 Adding volume PowerHA_3 to consistency group redbook_sync_cg

```
# xcli -u admin -p adminadmin -m 9.114.63.166 vol_list | grep PowerHA_3
PowerHA_3 17 PowerHA_Pool admin 0
# xcli -u admin -p adminadmin -m 9.114.63.166 cg_add_vol cg=redbook_sync_cg
vol=PowerHA_3
Command executed successfully.
```

```
# xcli -u admin -p adminadmin -m 9.114.63.166 vol_list | grep PowerHA_3
PowerHA_3 17 redbook_sync_cg          PowerHA_Pool admin      0
```

Check the *hdisk* names mapped to the PowerHA_3 XIV volume on all cluster nodes using the **clxd_list_xiv_luns** command shown in Example 7-44. This command is located in the `/usr/es/sbin/cluster/xd_generic/xd_xiv_rm` directory. Execute this command on all cluster nodes to find *hdisk* names.

Example 7-44 Volume mapping

```
(0) root @ c665f1sq07: /usr/es/sbin/cluster/xd_generic/xd_xiv_rm
# clxd_list_xiv_luns admin adminadmin 9.114.63.166
Warning: There is no cluster found.
cllsclstr: No cluster defined.
cllsclstr: Error reading configuration.
Warning: There is no cluster found.
cllsnode: Error reading configuration.
HDISKS S/N          XIV VOLUME ID      XIV VOLUME NAME VOLUME GROUP
===== ======      ====== ======      ====== ====== ======
hdisk23 276808F9    00173800276808F9  PowerHA_1 xiv_sync_autovg
hdisk22 276808FA    00173800276808FA  PowerHA_2 xiv_sync_manvg
hdisk21 276808FB    00173800276808FB  PowerHA_3 None
hdisk26 27680908    0017380027680908  PowerHA_4 xiv_autovg
hdisk25 27680909    0017380027680909  PowerHA_5 xiv_manvg
hdisk24 2768090A    001738002768090A  PowerHA_6 Non
hdisk2  27681F88     0017380027681F88  PowerHA_7 None
```

The *hdisks* have different PVIDs on both sites because they were not in a remote mirror relationship earlier. After creating the remote mirror relationship, we needed to remove the *hdisk* using the **rmdev** command and running the **cfgmgr** command on all nodes, as shown in Example 7-45, to flash the same PVID on both sites.

Example 7-45 Removing the hdisk and running cfgmgr to flash the same PVID on both sites

```
(0) root @ c665f1sq08: /
# rmdev -dl hdisk21
hdisk21 deleted

(0) root @ c665f1sq08: /
# cfgmgr

# dsh lspv | grep hdisk21
c665f1sq07.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 None
c665f1sq08.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 None
c665f1sq10.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 None
c665f1sq09.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 None
```

Ideally, to simplify administration, the *hdisk* names should be the same on all cluster nodes. If they are not, then change the *hdisk* name with the **rendev** command on all cluster nodes, as shown in Example 7-46.

Example 7-46 Renaming a physical disk name

```
(0) root @ c665f1sq08: /
# rendev -l hdisk15 -n hdisk21
hdisk21
```

Asynchronous mirror coupling

In this section, we create an asynchronous remote mirroring pair between the PowerHA_6 volume on both master and slave storage systems. Before defining the mirroring pair between volumes, we created an asynchronous remote mirror consistency group, redbook_async_cg, on both storage systems, as shown in Example 7-47.

Example 7-47 Asynchronous consistency group creation

```
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 cg_create  
cg=redbook_async_cg pool=PowerHA_Pool  
Command executed successfully.  
  
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 cg_create  
cg=redbook_async_cg pool=PowerHA_Pool  
Command executed successfully.
```

When creating an asynchronous remote mirroring pair, the additional parameters of *mirror_type*, *recovery point objective* (rpo) and *schedule* are required. You also have to specify the mirror type as ASYNC_INTERVAL. Asynchronous mirroring is based on a schedule-driven replication. The system also offers a predefined schedule object with a non-user-configurable interval of 20 seconds, named *min_interval*. Use the system-provided schedule object while creating the asynchronous mirroring pair. The recovery point objective time designation is the maximum time interval at which the mirrored volume or CG lags behind the master volume. Once the specified interval is reached, a consistent copy of the volume or CG should be available. Example 7-48 shows the consistency group asynchronous remote mirroring pair.

Example 7-48 Creating asynchronous remote mirror coupling between consistency groups

```
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_create  
target="XIV pPETXIV4" cg=redbook_async_cg slave_cg=redbook_async_cg  
type=ASYNC_INTERVAL rpo=60 remote_rpo=60 schedule=min_interval  
remote_schedule=min_interval  
Command executed successfully.
```

Ensure that the consistency group asynchronous mirroring pair was created and the status shows RPO OK, as shown in Example 7-49.

Example 7-49 Checking asynchronous mirror coupling status

```
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list cg=redbook_async_cg  
Name          Mirror Type    Mirror Object   Role   Remote System  Remote Peer   Active  
redbook_async_cg  async_interval  CG           Master XIV pPETXIV4  redbook_async_cg  no  
Status        Link Up  
RPO OK       yes
```

We now activated the consistency group redbook_async_cg remote mirroring pair using the **mirror_activate** command, and checked the active status as shown in Example 7-50.

Example 7-50 Activating asynchronous consistency group and checking the active status

```
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_activate cg=redbook_async_cg  
Command executed successfully.  
  
(0) root @ c665f1sq07: /  
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list cg=redbook_async_cg
```

```

Name      Mirror Type   Mirror Object  Role  Remote System  Remote Peer    Active
redbook_async_cg  async_interval  CG          Master XIV pPETXIV4  redbook_async_cg  yes
Status    Link Up
RPO OK   yes

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list cg=redbook_async_cg
Name      Mirror Type   Mirror Object  Role  Remote System  Remote Peer    Active
redbook_async_cg  async_interval  CG          Slave  XIV 1310088  redbook_async_cg  yes
Status    Link Up
RPO OK   yes

```

We used the volume PowerHA_6 XIV on both master and slave XIV storage systems to create the asynchronous remote mirroring pair. Parameters such as mirror type, recovery point objective (rpo) and schedule object are mandatory for asynchronous mirroring pair creation. We used the system-defined schedule object min_interval and recovery point objective (rpo) 60 seconds, as shown in Example 7-51.

Example 7-51 Creating an asynchronous mirroring pair between volumes

```

# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_create
target="XIV pPETXIV4" vol=PowerHA_6 slave_vol=PowerHA_6 type=ASYNC_INTERVAL rpo=60
remote_rpo=60 schedule=min_interval remote_schedule=min_interval
Command executed successfully.

```

Next we activated the mirroring pair between the PowerHA_6 XIV volumes, and checked the mirror active status, as shown in Example 7-52.

Example 7-52 Activating asynchronous mirror coupling and checking the active status

```

# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_activate vol=PowerHA_6
Command executed successfully.

```

```

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list vol=PowerHA_6
Name      Mirror Type   Mirror Object  Role  Remote System  Remote Peer    Status  Link Up
PowerHA_6  async_interval  Volume       Master XIV pPETXIV4  PowerHA_6     yes     RPO OK  yes

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list vol=PowerHA_6
Name      Mirror Type   Mirror Object  Role  Remote System  Remote Peer    Status  Link Up
PowerHA_6  async_interval  Volume       Slave  XIV 1310088  PowerHA_6     yes     RPO OK  yes

```

We now added the PowerHA_6 mirrored XIV volume pair to the asynchronous remote mirror consistency group redbook_async_cg, as shown in Example 7-53.

Example 7-53 Adding a volume to the consistency group and deactivating the consistency group

```

# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 cg_add_vol
cg=redbook_async_cg vol=PowerHA_6
Command executed successfully.
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 vol_list | grep
PowerHA_[36]
PowerHA_3 17                      redbook_sync_cg        PowerHA_Pool admin 0
PowerHA_6 17                      redbook_async_cg      PowerHA_Pool admin 0
last-replicated-PowerHA_6 17 PowerHA_6    redbook_async_cg PowerHA_Pool

```

Check the *hdisk* names mapped to the PowerHA_6 XIV volume on all cluster nodes using the **c1xd_list_xiv_luns** command as shown in Example 7-44 on page 310. Since the *hdisk* has different PVIDs on both sites, remove the *hdisk* and run the **cfgmgr** command to expose the same PVID on all cluster nodes as shown in Example 7-45 on page 310. Make sure that *hdisk* names are the same on all cluster nodes. If they are different, then rename the *hdisk* name as shown in Example 7-46 on page 310.

7.6.3 Configuring the cluster using clmgr

We now create the PowerHA SystemMirror linked cluster with XIV remote mirror replicated resources using the **clmgr** command, which provides a consistent and reliable interface for performing PowerHA SystemMirror cluster operations via terminal or script. All **clmgr** operations are logged in the *clutils.log* file, including the command that was executed, its start/stop time, and what user initiated the command. The basic format of the **clmgr** is as follows:

```
clmgr <ACTION> <CLASS> [<NAME>] [<ATTRIBUTES...>]
```

Parameters for the **clmgr** command and their descriptions are:

ACTION	This specifies the operation to be performed on a particular CLASS. It is a verb describing a user action. There are four basic ACTIONS that can be performed on almost all the supported CLASSES. They are add, query, modify and delete.
CLASS	The object type on which ACTION will be performed. For example, cluster, network, repository, site and so on.
NAME	Any name given to an object of type CLASS.
ATTRIBUTES	This parameter is optional. Attributes are always defined with ATTR=VALUE pairs. Attribute pairs are used to specify configuration settings. These are specific to the ACTION+CLASS combination.

For more information about the **clmgr** command, consult the man page.

Defining cluster topology

We configured the cluster topology according to the scenario described in 7.6.1, “Implementation overview” on page 302. Also refer to the PowerHA SystemMirror 7.1.2 publications for creating cluster topology at the following website:

<http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.navigation/powerha.htm>

We defined a linked cluster configuration with the **clmgr** command shown in Example 7-54.

Example 7-54 Defining the cluster name using clmgr

```
# clmgr add cluster xiv_cluster NODES=c665f1sq07,c665f1sq08,c665f1sq09,c665f1sq10
TYPE=LC 2>&1 create_cluster.log
Cluster Name: xiv_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disk: None
Cluster IP Address:
There are 4 node(s) and 2 network(s) defined
```

```

NODE c665f1sq07:
    Network net_ether_01
        f1sq07_enboot 30.30.30.1
        f1sq07_enstby1 30.30.30.129
    Network net_ether_02
        c665f1sq07      9.114.135.71

NODE c665f1sq08:
    Network net_ether_01
        f1sq08_enstby1 30.30.30.130
        f1sq08_enboot 30.30.30.2
    Network net_ether_02
        c665f1sq08      9.114.135.72

NODE c665f1sq09:
    Network net_ether_01
        f1sq09_enboot 30.30.30.3
        f1sq09_enstby1 30.30.30.131
    Network net_ether_02
        c665f1sq09      9.114.135.73

NODE c665f1sq10:
    Network net_ether_01
        f1sq10_enboot 30.30.30.4
        f1sq10_enstby1 30.30.30.132
    Network net_ether_02
        c665f1sq10      9.114.135.74

No resource groups defined
Initializing..
Gathering cluster information, which may take a few minutes...

```

In our example, we specified the `xiv_cluster` as a cluster name, linked cluster (LC) as type of cluster. For stretched clusters specify SC, and for non-site clusters specify NSC as cluster type. If you are defining a non-site (NSC) or stretched (SC) cluster then you can also specify the `CLUSTER_IP` attribute value. `CLUSTER_IP` is optional. If you do not specify it, then cluster software automatically assigns an IP for multicasting.

Note: When using sites, the cluster type cannot be modified once the cluster is defined and synchronized.

We also specified node names that are the hostnames of the nodes. Hostnames must have an associated IP address. All hostnames, or hostname IP addresses of the nodes, must be defined in the `/etc/cluster/rhosts` file prior to initial cluster synchronization. PowerHA SystemMirror automatically detects, through discovery, interfaces on the nodes and adds them to the network according to their subnet. By default, PowerHA SystemMirror assigns the interfaces with PowerHA networks of type `ether`. We modified the network `net_ether_02` to `XD_ip` as shown in Example 7-55.

Example 7-55 Changing network type

```
# clmgr modify network net_ether_02 NAME=net_xd_ip1 TYPE=XD_ip
```

```
Network type of Network net_ether_01 changed to XD_ip.
Warning: Network type of interface/label [c665f1sq07] changed to [XD_ip].
```

```
Warning: Network type of interface/label [c665f1sq08] changed to [XD_ip].  
Warning: Network type of interface/label [c665f1sq09] changed to [XD_ip].  
Warning: Network type of interface/label [c665f1sq10] changed to [XD_ip].
```

Define sites and repository disks

We created a linked cluster named xiv_cluster with four nodes. We added sites and assigned nodes to each site defined earlier, as shown in Example 7-56.

Example 7-56 Adding sites to cluster

```
(0) root @ c665f1sq07: /  
# clmgr add site sitea NODES=c665f1sq07,c665f1sq08  
  
claddsite: Node c665f1sq07 has been added to site sitea  
claddsite: Node c665f1sq08 has been added to site sitea  
  
(0) root @ c665f1sq07: /  
# clmgr add site siteb NODES=c665f1sq09,c665f1sq10  
  
claddsite: Node c665f1sq09 has been added to site siteb  
claddsite: Node c665f1sq10 has been added to site siteb
```

The syntax for adding a site is as follows:

```
clmgr add site <sitename> NODES=<node>[,<node#2>,...] [SITE_IP=<multicast_address>]
```

In our example, we created two sites, sitea with nodes c665f1sq07, c665f1sq08, and siteb with nodes c665f1sq09 and c665f1sq10. SITE_IP is optional. If you do not specify it, one is automatically assigned for multicasting.

NOTE: SITE_IP must be used only in a linked cluster (LC).

We used disks local to each site for a repository disk as required when creating a linked cluster, as shown in Example 7-57.

Example 7-57 Adding repository disks to linked cluster

```
(0) root @ c665f1sq07: /  
# clmgr add repository hdisk2 SITE=sitea NODE=c665f1sq07  
  
(0) root @ c665f1sq07: /  
# clmgr add repository hdisk3 SITE=siteb NODE=c665f1sq09
```

The syntax for adding a repository disk is as follows:

```
clmgr add repository <disk>[,<backup_disk#2>,...] [SITE=<site_label>]  
[NODE=<reference_node>]
```

In our example, we added hdisk2 to sitea and referenced the c665f1sq07 node. We also added hdisk3 to siteb and referenced the c665f1sq09 node. The reference node is a node from which the physical disk name is used. Physical disk details are shown in Figure 7-49 on page 316.

```
(0) root @ c665f1sq07: /
# dsh lspv | grep 00c9354cd95befbd
c665f1sq07.ppd.pok.ibm.com: hdisk2          00c9354cd95befbd None
c665f1sq08.ppd.pok.ibm.com: hdisk13           00c9354cd95befbd None

(0) root @ c665f1sq09: /
# dsh lspv | grep 00c9354cd95d61e2
c665f1sq09.ppd.pok.ibm.com: hdisk3          00c9354cd95d61e2 None
c665f1sq10.ppd.pok.ibm.com: hdisk14           00c9354cd95d61e2 None
```

Figure 7-49 Physical details used for repository disk

Configuring the XIV remote mirror couplings

We created synchronous and asynchronous XIV remote mirroring pairs at the storage level in 7.6.2, “Configuring XIV remote mirror couplings using xcli” on page 304. We used the same mirroring pairs to configure XIV remote mirror couplings at the PowerHA SystemMirror level. There are three basic steps to configure a XIV remote mirror replicated resource at the PowerHA SystemMirror level. They are:

- ▶ Defining storage agents
- ▶ Defining storage systems
- ▶ Defining storage mirror groups

First define the XIV storage agents as shown in Example 7-58.

Example 7-58 Defining XIV remote mirror coupling

```
(0) root @ c665f1sq07: /
# clmgr add storage_agent SA_1310088 TYPE=xiv_rm ADDRESSES=9.114.63.166 USER=admin
PASSWORD=adminadmin

(0) root @ c665f1sq07: /
# clmgr add storage_agent SA_pPETXIV4 TYPE=xiv_rm ADDRESSES=9.114.63.163
USER=admin PASSWORD=adminadmin
```

The syntax for adding a storage agent is as follows:

```
clmgr add storage_agent <agent_name> TYPE={ds8k_gm|xiv_rm}
ADDRESSES=<IP>[<IP#2>,...] USER=<user_id> ] PASSWORD=<password> ]
ATTRIBUTES=<NAME>@<VALUE>[,<NAME#2>@<VALUE#2>,...] ]
```

In our example, we added two agents: SA_1310088 and SA_pPETXIV4, one each for master and slave storage. We specified the type of storage as xiv_rm (XIV remote mirror). We also passed the storage system IPs along with user name and password.

Now we added the storage systems. The syntax for adding a storage system is as follows:

```
clmgr add storage_system <storage_system_name> TYPE={ds8k_gm|xiv_rm} SITE=<site>
AGENTS=<agent>[,<agent#2>,...] VENDOR_ID=<identifier> [
WWNN=<world_wide_node_name> ] [ATTRIBUTES=<NAME>@<VALUE>[,<NAME#2>@<VALUE#2>,...] ]
```

The command takes the user-specified storage system name, type of storage mirroring, the site associated with the storage system, agents associated with the storage system, vendor ID of the storage system, world wide node name and attributes. World wide node name and attributes are optional. The vendor ID associated with the storage agent is mandatory. We queried for vendor IDs of the storage agents defined already as shown in Example 7-59.

Example 7-59 Querying vendor ID

```
(0) root @ c665f1sq07: /  
# clmgr -a VENDOR_ID query storage_system TYPE=xiv_rm  
XIV_1310088  
XIV_pPETXIV4
```

The storage system with ID XIV_1310088 is associated to the SA_1310088 storage agent at sitea. The storage system with ID XIV_pPETXIV4 is associated to the SA_pPETXIV4 storage agent at siteb. We defined the storage system as shown in Example 7-60.

Example 7-60 Adding the XIV storage system

```
(0) root @ c665f1sq07: /  
# clmgr add storage_system SS_1310088 TYPE=xiv_rm SITE=sitea AGENTS=SA_1310088  
VENDOR_ID=XIV_1310088  
  
(0) root @ c665f1sq07: /  
# clmgr add storage_system SS_pPETXIV4 TYPE=xiv_rm SITE=siteb AGENTS=SA_pPETXIV4  
VENDOR_ID=XIV_pPETXIV4
```

The final step is to add the storage mirror group. The syntax for adding the storage mirror group is as follows:

```
clmgr add mirror_group <mirror_group_name> TYPE={ds8k_gm|xiv_rm} MODE={sync|async}  
RECOVERY={auto>manual} [ STORAGE_SYSTEMS=<storage_system>[,<ss#2>,...] ] [  
CONSISTENT={yes|no} ] [ VENDOR_ID=<vendor_specific_identifier> ] [  
ATTRIBUTES=<NAME>@<VALUE>[,<NAME#2>@<VALUE#2>,...] ]
```

The command takes user-specified mirror group name, type of storage mirror, mirroring mode, recovery action method, storage systems defined earlier, vendor ID, and attributes. We used vendor IDs as XIV remote mirror replicated resources defined in “Synchronous mirror coupling” on page 306 and “Asynchronous mirror coupling” on page 311. We created synchronous and asynchronous mirror groups for testing purposes, as shown in Example 7-61.

Example 7-61 Adding XIV remote mirror group

```
(0) root @ c665f1sq07: /  
# clmgr add mirror_group sync_mirror TYPE=xiv_rm MODE=sync RECOVERY=auto  
STORAGE_SYSTEMS=SS_1310088,SS_pPETXIV4 VENDOR_ID=redbook_sync_cg  
  
(0) root @ c665f1sq07: /  
# clmgr add mirror_group async_mirror TYPE=xiv_rm MODE=async RECOVERY=auto  
STORAGE_SYSTEMS=SS_1310088,SS_pPETXIV4 VENDOR_ID=redbook_async_cg
```

We created the synchronous mirror group sync_mirror associated with the redbook_sync_cg consistency group. Also, we created an asynchronous mirror group called async_mirror associated with the redbook_async_cg consistency group. Now we created volume groups at the AIX level called redbook_sync_vg and async_vg, respectively. Then we imported them to all cluster nodes, as shown in Figure 7-50 on page 318.

Attention: You can import the volume group without varying it on (using the **-n** flag during the import operation) while keeping the remote mirror relationships active. This is possible with XIV remote mirroring because the slave volume is accessible for read-only operations.

```

# dsh lspv | grep hdisk21
c665f1sq07.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 redbook_sync_vg
c665f1sq08.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 redbook_sync_vg
c665f1sq10.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 redbook_sync_vg
c665f1sq09.ppd.pok.ibm.com: hdisk21      00c9354c47833c82 redbook_sync_vg

(0) root @ c665f1sq07: /
# dsh lspv | grep hdisk24
c665f1sq07.ppd.pok.ibm.com: hdisk24      00c9354c756cf9d6 async_vg
c665f1sq08.ppd.pok.ibm.com: hdisk24      00c9354c756cf9d6 async_vg
c665f1sq10.ppd.pok.ibm.com: hdisk24      00c9354c756cf9d6 async_vg
c665f1sq09.ppd.pok.ibm.com: hdisk24      00c9354c756cf9d6 async_vg

```

Figure 7-50 Volume groups

Configuring the service IP and resource group

We configured the service IPs for each resource group. In our example, we created two resource groups, one for each mirror group defined earlier. So we defined two service IPs with network type *ether* as shown in Example 7-62.

Example 7-62 Defining service IPs

```

(0) root @ c665f1sq07: /
# clmgr add service_ip 30.30.30.65 NETWORK=net_ether_01

(0) root @ c665f1sq07: /
# clmgr add service_ip 30.30.30.66 NETWORK=net_ether_01

```

Important: There must be a one-to-one association between the mirror group and the resource group.

Now we configured resource groups. The syntax for adding a resource group is shown in Figure 7-51 on page 319.

```

clmgr add resource_group <resource_group>[,<rg#2>,...]
    NODES=<node_P1>[,<node_P2>,...]
    [ SECONDARYNODES=<node_S1>[,<node_S2>,...]
    [ SITE_POLICY={ignore|primary|either|both}
    [ STARTUP={OHN|OFAN|OAAN|OUDP}
    [ FALLOVER={FNPN|FUDNP|BO}
    [ FALLBACK={NFB|FBHPN}
    [ NODE_PRIORITY_POLICY={default|mem|cpu|
        disk|least|most}
    [ NODE_PRIORITY_POLICY_SCRIPT=</path/to/script>
    [ NODE_PRIORITY_POLICY_TIMEOUT=###
    [ SERVICE_LABEL=service_ip#1[,service_ip#2,...]
    [ APPLICATIONS=appctr#1[,appctr#2,...]
    [ SHARED_TAPE_RESOURCES=<TAPE>[,<TAPE#2>,...]
    [ VOLUME_GROUP=<VG>[,<VG#2>,...]
    [ FORCED_VARYON={true|false}
    [ VG_AUTO_IMPORT={true|false}
    [ FILESYSTEM=/file_system#1[,/file_system#2,...]
    [ DISK=<raw_disk>[,<raw_disk#2>,...]
    [ FS_BEFORE_IPADDR={true|false}
    [ WPAR_NAME="wpar_name"
    [ EXPORT_FILESYSTEM=/expfs#1[,/expfs#2,...]
    [ EXPORT_FILESYSTEM_V4=/expfs#1[,/expfs#2,...]
    [ STABLE_STORAGE_PATH="/fs3"
    [ NFS_NETWORK="nfs_network"
    [ MOUNT_FILESYSTEM=/nfs_fs1;/expfs1,/nfs_fs2;,...
    [ MIRROR_GROUP=<replicated_resource>
    [ FALLBACK_AT=<FALLBACK_TIMER>

```

Figure 7-51 Add resource group command syntax

Refer to the **clmgr** man page for the resource group policies. We added the resource group sync_rg with the following policies and attributes:

Site_policy	Prefer Primary Site
Startup policy	OHN (Online on Home Node Only)
Fallover policy	FNPN (Failover to Next Priority Node)
Fallback policy	FBHPN (Fall Back to Higher Priority Node)
Service IP Label	f1sq07_ensvc
Volume group	redbook_sync_vg
Filesystem	/redbook_sync_fs
Mirror group	sync_mirror

We also added another resource group, async_rg, with the following policies and attributes:

Site policy	Either (Online on Either Site)
Startup policy	OUDP (Online Using Node Distribution Policy)
Fallover policy	FNPN (Failover to Next Priority Node)
Fallback policy	NFB (Never Fallback)
Service IP Label	f1sq08_ensvc
Volume group	async_vg
Filesystem	/async_fs
Mirror group	async_mirror

Creating both resource groups is shown in Example 7-63. We also specified volume groups associated with mirror groups.

Example 7-63 Adding resource group to cluster

```
# clmgr add resource_group sync_rg NODES=c665f1sq07,c665f1sq08  
SECONDARYNODES=c665f1sq09,c665f1sq10 SITE_POLICY=primary STARTUP=OHN FALLOVER=FNPN  
FALLBACK=FBHPN SERVICE_LABEL=f1sq07_ensvc VOLUME_GROUP=redbook_sync_vg  
FILESYSTEM=/redbook_sync_fs MIRROR_GROUP=sync_mirror
```

Auto Discover/Import of Volume Groups was set to true.
Gathering cluster information, which may take a few minutes.

```
(0) root @ c665f1sq07: /  
# clmgr add resource_group async_rg NODES=c665f1sq08,c665f1sq07  
SECONDARYNODES=c665f1sq10,c665f1sq09 SITE_POLICY=either STARTUP=OUDP FALLOVER=FNPN  
FALLBACK=NFB SERVICE_LABEL=f1sq08_ensvc VOLUME_GROUP=async_vg FILESYSTEM=/async_fs  
MIRROR_GROUP=async_mirror
```

Auto Discover/Import of Volume Groups was set to true.
Gathering cluster information, which may take a few minutes.

Now verify and synchronize the cluster configuration as shown in Example 7-64.

Example 7-64 Sync and verify the cluster

```
(0) root @ c665f1sq07: /  
# clmgr verify cluster CHANGES_ONLY=no SYNC=yes  
Saving existing /var/hacmp/clverify/ver_mping/ver_mping.log to  
/var/hacmp/clverify/ver_mping/ver_mping.log.bak  
Verifying clcom communication, please be patient.
```

Verify the cluster configuration topology using the **cltopinfo** command shown in Example 7-65.

Example 7-65 Verifying the cluster configuration

```
# cltopinfo  
Cluster Name: xiv_cluster  
Cluster Connection Authentication Mode: Standard  
Cluster Message Authentication Mode: None  
Cluster Message Encryption: None  
Use Persistent Labels for Communication: No  
Repository Disks: Site 1: hdisk2, Site 2:  
Cluster IP Addresses: Cluster: , Site 1: 228.114.135.71, Site 2: 228.114.135.72  
There are 4 node(s) and 2 network(s) defined  
  
NODE c665f1sq07:  
    Network net_ether_01  
        f1sq08_ensvc      30.30.30.66  
        f1sq07_ensvc      30.30.30.65  
        f1sq07_enstby1    30.30.30.129  
        f1sq07_enboot     30.30.30.1  
    Network net_xd_ip1  
        c665f1sq07       9.114.135.71  
  
NODE c665f1sq08:
```

```

Network net_ether_01
    f1sq08_ensvc      30.30.30.66
    f1sq07_ensvc      30.30.30.65
    f1sq08_enboot     30.30.30.2
    f1sq08_enstby1    30.30.30.130
Network net_xd_ip1
    c665f1sq08        9.114.135.72

NODE c665f1sq09:
    Network net_ether_01
        f1sq08_ensvc      30.30.30.66
        f1sq07_ensvc      30.30.30.65
        f1sq09_enstby1    30.30.30.131
        f1sq09_enboot     30.30.30.3
    Network net_xd_ip1
        c665f1sq09        9.114.135.73

NODE c665f1sq10:
    Network net_ether_01
        f1sq08_ensvc      30.30.30.66
        f1sq07_ensvc      30.30.30.65
        f1sq10_enstby1    30.30.30.132
        f1sq10_enboot     30.30.30.4
    Network net_xd_ip1
        c665f1sq10        9.114.135.74

Resource Group sync_rg
    Startup Policy   Online On Home Node Only
    Fallback Policy  Fallback To Next Priority Node In The List
    Fallback Policy  Fallback To Higher Priority Node In The List
    Participating Nodes  c665f1sq07 c665f1sq08 c665f1sq09 c665f1sq10
    Service IP Label  f1sq07_ensvc

Resource Group async_rg
    Startup Policy   Online Using Distribution Policy
    Fallback Policy  Fallback To Next Priority Node In The List
    Fallback Policy  Never Fallback
    Participating Nodes  c665f1sq08 c665f1sq07 c665f1sq10 c665f1sq09
    Service IP Label  f1sq08_ensvc

```

Starting cluster services using SMIT

Now start cluster services on the primary nodes, first using the **smit clstart** panel as shown in Figure 7-52 on page 322. Move the cursor and press F4 to select nodes. Set the “Startup Cluster Information Daemon” field to true. Press Enter to continue.

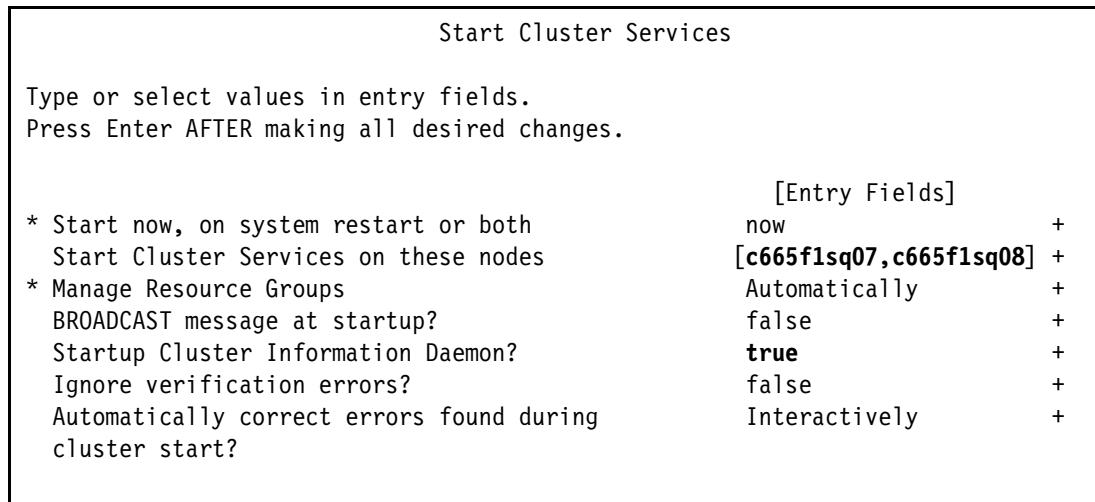


Figure 7-52 Start cluster services on primary nodes

Wait for cluster services to start and the cluster state to become stable. The resource groups will come online on the primary site nodes as shown in Example 7-66.

Example 7-66 Resource group status on primary nodes

# clRGinfo -p			
Cluster Name: xiv_cluster			
Resource Group Name: sync_rg			
Node	Primary State	Secondary State	
c665f1sq07@sitea	ONLINE	OFFLINE	
c665f1sq08@sitea	OFFLINE	OFFLINE	
c665f1sq09@siteb	OFFLINE	OFFLINE	
c665f1sq10@siteb	OFFLINE	OFFLINE	
Resource Group Name: async_rg			
Node	Primary State	Secondary State	
c665f1sq08@sitea	OFFLINE	OFFLINE	
c665f1sq07@sitea	ONLINE	OFFLINE	
c665f1sq10@siteb	OFFLINE	OFFLINE	
c665f1sq09@siteb	OFFLINE	OFFLINE	

Now start cluster services on the secondary nodes using the **smit clstart** panel as shown in Figure 7-52.

Then wait until the nodes join the cluster and the cluster state stabilizes. The resource groups become *online secondary* on secondary site nodes, as shown in Example 7-67.

Example 7-67 Resource group status on secondary nodes

```
# clRGinfo -p
```

Cluster Name: xiv_cluster

Resource Group Name: sync_rg		
Node	Primary State	Secondary State
c665f1sq07@sitea	ONLINE	OFFLINE
c665f1sq08@sitea	OFFLINE	OFFLINE
c665f1sq09@siteb	OFFLINE	ONLINE SECONDARY
c665f1sq10@siteb	OFFLINE	OFFLINE

Resource Group Name: async_rg		
Node	Primary State	Secondary State
c665f1sq08@sitea	OFFLINE	OFFLINE
c665f1sq07@sitea	ONLINE	OFFLINE
c665f1sq10@siteb	OFFLINE	OFFLINE
c665f1sq09@siteb	OFFLINE	ONLINE SECONDARY

Starting cluster services using clmgr

You can also start cluster services using the **clmgr** command. The command syntax for starting cluster services is as follows:

```
clmgr online cluster [WHEN={now|restart|both}] [MANAGE={auto|manual}]
[BROADCAST={false|true}] [CLINFO={false|true|consistent}] [FORCE={false|true}]
[FIX={no|yes|interactively}] [TIMEOUT=<seconds_to_wait_for_completion>]
```

This command starts cluster services on all nodes in the cluster. If you want to start cluster services on a single or more than one node using the **clmgr** command, then use the command syntax as follows:

```
clmgr online node [<node>[,<node#2>,...]] [WHEN={now|restart|both}]
[MANAGE={auto|manual}] [BROADCAST={false|true}] [CLINFO={false|true|consistent}]
[FORCE={false|true}] [FIX={no|yes|interactively}]
[TIMEOUT=<seconds_to_wait_for_completion>]
```

You need to specify the node names on which you want to start the cluster services. If you want to start cluster services on all nodes at a particular site, use the command syntax as follows:

```
clmgr online site <sitename> [WHEN={now|restart|both}] [MANAGE={auto|manual}]
[BROADCAST={false|true}] [CLINFO={false|true|consistent}] [FORCE={false|true}]
[FIX={no|yes|interactively}] [TIMEOUT=<seconds_to_wait_for_completion>]
```

You need to mention the site name where you want to start cluster services on all nodes.

The command takes various parameters in all the above cases. A brief description of these parameters follows:

WHEN	This field indicates when you want to start cluster services. Specify <i>now</i> , if you want to start cluster services immediately. Specify <i>restart</i> , if you want to start cluster services when the node reboots. Otherwise specify <i>both</i> .
MANAGE	This field indicates resource group management. If you set <i>auto</i> , cluster software brings the resource group online automatically. If you set <i>manual</i> , you have to bring the resource group online manually.
BROADCAST	If you set this to <i>true</i> , then it will broadcast a cluster start service message to all nodes. It uses the wall command to broadcast the message.
CLINFO	If you want to start the cluster information daemon, set this field to <i>true</i> or <i>consistent</i> .

FORCE	Cluster services will not start if there are any verification errors. Set this field to <i>true</i> to ignore verification errors and force the cluster startup services.
FIX	If you want to correct errors found during verification automatically, set to <i>true</i> . Set to <i>interactively</i> , to have verification prompt you to correct resulting verification errors. Otherwise set to <i>false</i> .
TIMEOUT	This is an optional field. Specify values in seconds to wait for the completion.

We started cluster services on each site for demonstration purposes. Starting cluster services at site sitea is shown in Example 7-68.

Example 7-68 Starting cluster services on all nodes at site sitea

```
(0) root @ c665f1sq07: /
# clmgr online site sitea WHEN=now MANAGE=auto BROADCAST=true CLINFO=true
FORCE=false FIX=interactively 2>&1 start_cluster.log

== INFO >>
Verifying HACMP support for GENXD.
== INFO >>
Resource Group 'sync_rg' contains GENXD Replicated Resource 'sync_mirror'.
Verifying its configuration.

== INFO >>
Checking for Storage Agent connectivity for GENXD configuration .

== INFO >>
Checking Node 'c665f1sq07' for connectivity with all GENXD Storage Agents.

== INFO >>
Checking Node 'c665f1sq08' for connectivity with all GENXD Storage Agents.
.....
.....
Starting Cluster Services on node: c665f1sq07
This may take a few minutes. Please wait...
c665f1sq07: Dec 13 2012 03:05:46 Starting execution of
/usr/es/sbin/cluster/etc/rc.cluster
c665f1sq07: with parameters: -boot NOW -A -b -i interactively start_cluster.log -P
cl_rc_cluster
c665f1sq07:
c665f1sq07: Dec 13 2012 03:05:46 Checking for srcmstr active...
c665f1sq07: Dec 13 2012 03:05:46 complete.
c665f1sq07: Dec 13 2012 03:05:47
c665f1sq07: /usr/es/sbin/cluster/utilities/clstart: called with flags -m -G -B -A
c665f1sq07:
c665f1sq07:      Dec 13 2012 03:08:51
c665f1sq07: Completed execution of /usr/es/sbin/cluster/etc/rc.cluster
c665f1sq07: with parameters: -boot NOW -A -b -i interactively start_cluster.log -P
cl_rc_cluster.
c665f1sq07: Exit status = 0
c665f1sq07:
```

```

Starting Cluster Services on node: c665f1sq08
This may take a few minutes. Please wait...
c665f1sq08: Dec 13 2012 03:07:27 Starting execution of
/usr/es/sbin/cluster/etc/rc.cluster
c665f1sq08: with parameters: -boot NOW -A -b -i interactively start_cluster.log -P
cl_rc_cluster
c665f1sq08:
c665f1sq08: Dec 13 2012 03:07:28 Checking for srcmstr active...
c665f1sq08: Dec 13 2012 03:07:28 complete.
c665f1sq08: Dec 13 2012 03:07:28
c665f1sq08: /usr/es/sbin/cluster/utilities/clstart: called with flags -m -G -B -A
c665f1sq08:
c665f1sq08: Dec 13 2012 03:10:26
c665f1sq08: Completed execution of /usr/es/sbin/cluster/etc/rc.cluster
c665f1sq08: with parameters: -boot NOW -A -b -i interactively start_cluster.log -P
cl_rc_cluster.
c665f1sq08: Exit status = 0
c665f1sq08:

```

In Example 7-68 on page 324, we started cluster services on all nodes at sitea. We automatically set the correct error field to *interactively* so that we knew the kind of errors in our cluster configuration. We set the clinfo daemon to true. We did not force the cluster startup. If there are any errors that are not resolved by cluster software interactively, then start cluster services. Cluster services starts on all nodes at sitea. Wait till all the nodes join the cluster and the cluster state becomes stable. Since we set the MANAGE field to auto, the resource groups came online on nodes at sitea as shown in Example 7-69.

Example 7-69 Resource group status on nodes at sitea

```

(0) root @ c665f1sq07: /
# c1RGinfo -p

Cluster Name: xiv_cluster

Resource Group Name: sync_rg
Node Primary State Secondary State
-----
c665f1sq07@sitea ONLINE OFFLINE
c665f1sq08@sitea OFFLINE OFFLINE
c665f1sq09@siteb OFFLINE OFFLINE
c665f1sq10@siteb OFFLINE OFFLINE

Resource Group Name: async_rg
Node Primary State Secondary State
-----
c665f1sq08@sitea OFFLINE OFFLINE
c665f1sq07@sitea ONLINE OFFLINE
c665f1sq10@siteb OFFLINE OFFLINE
c665f1sq09@siteb OFFLINE OFFLINE
(0) root @ c665f1sq07: /

```

Similarly, start cluster services at site siteb as shown in Example 7-70 on page 326. Wait till nodes join the cluster and the cluster state becomes stable. The resource groups become online secondary on the secondary site nodes as shown in Example 7-70 on page 326.

Example 7-70 Resource group status after starting cluster services at siteb

```
# c1RGinfo -p

Cluster Name: xiv_cluster

Resource Group Name: sync_rg
Node Primary State Secondary State
-----
c665f1sq07@sitea ONLINE OFFLINE
c665f1sq08@sitea OFFLINE OFFLINE
c665f1sq09@siteb OFFLINE ONLINE SECONDARY
c665f1sq10@siteb OFFLINE OFFLINE

Resource Group Name: async_rg
Node Primary State Secondary State
-----
c665f1sq08@sitea OFFLINE OFFLINE
c665f1sq07@sitea ONLINE OFFLINE
c665f1sq10@siteb OFFLINE OFFLINE
c665f1sq09@siteb OFFLINE ONLINE SECONDARY
```

7.7 Administrating the cluster with XIV replicated resources

The XIV configuration cannot be changed by using dynamic automatic reconfiguration (DARE). The XIV remote mirror replicated resource or mirror group cannot be removed from the resource group using DARE. But you can add multiple mirror groups to an existing resource group using DARE. Make sure that the new mirror group has the same *mirror mode* and *recovery action* as that of an existing mirror group in a resource group.

We can perform the following operations on PowerHA SystemMirror with XIV replicated resources using DARE.

- ▶ Add/Remove volumes to a XIV remote mirror replicated resource or mirror group.
- ▶ Add file systems to a volume group associated with a XIV remote mirror replicated resource or mirror group.
- ▶ Change the recovery action.
- ▶ Add a XIV remote mirror replicated resource or mirror group and associated resource group.

7.7.1 Adding volumes to the XIV remote mirror replicated resource

We can add volumes to the XIV remote mirror replication resource or mirror group without affecting PowerHA SystemMirror cluster services using DARE.

First, we added a volume to a XIV remote mirror replicated resource or mirror group. In our example, we used a PowerHA_8 XIV volume on both master and slave storage systems. We then created a synchronous mirroring pair between the PowerHA_8 volume on the master and slave XIV storage system as shown in Example 7-39 on page 308. We then checked the mirror status of the XIV volume, as shown in Example 7-71 on page 327.

Example 7-71 Checking mirror status on a XIV volume

```
# xcli -u admin -p adminadmin -m 9.114.63.166 mirror_list vol=PowerHA_8
Name      Mirror Type      Mirror Object    Role   Remote System   Remote Peer   Active   Status
PowerHA_8  sync_best_effort  Volume          Master  XIV pPETXIV4  PowerHA_8     yes     Synchronized
Link Up
yes

(0) root @ c665f1sq09: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.163 mirror_list vol=PowerHA_8
Name      Mirror Type      Mirror Object    Role   Remote System   Remote Peer   Active   Status
PowerHA_8  sync_best_effort  Volume          Slave   XIV 1310088  PowerHA_8     yes     Consistent
Link Up
yes
```

Next add the synchronous volume mirroring pair to the synchronous consistency group redbook_sync_cg as shown in Example 7-72.

Example 7-72 Adding mirrored volume pair to a consistency group

```
(0) root @ c665f1sq09: /opt/xiv/XIVGUI
# xcli -u admin -p adminadmin -m 9.114.63.166 cg_add_vol cg=redbook_sync_cg
vol=PowerHA_8
Command executed successfully.
```

Check the hdisk names mapped to the PowerHA_8 XIV volume on all cluster nodes using the **clxd_list_xiv_luns** command as shown in Example 7-73. This command is located in the **/usr/es/sbin/cluster/xd_generic/xd_xiv_rm** directory. Execute this command on all cluster nodes.

Example 7-73 Physical disks mapped to corresponding XIV volumes

```
(0) root @ c665f1sq07: /usr/es/sbin/cluster/xd_generic/xd_xiv_rm
# clxd_list_xiv_luns admin adminadmin 9.114.63.166
HDISKS           S/N           XIV VOLUME ID       XIV VOLUME NAME VOLUME GROUP
=====           =====           ======           ======           ======           =====
hdisk23         276808F9      00173800276808F9  PowerHA_1 None
hdisk22         276808FA      00173800276808FA  PowerHA_2 None
hdisk21         276808FB      00173800276808FB  PowerHA_3 redbook_sync_vg
hdisk26         27680908      0017380027680908  PowerHA_4 None
hdisk25         27680909      0017380027680909  PowerHA_5 None
hdisk24         2768090A      001738002768090A  PowerHA_6 async_vg
hdisk13         27681F88      0017380027681F88  PowerHA_7 caavg_private
hdisk27         27681F89      0017380027681F89  PowerHA_8 None
hdisk14         27681F8A      0017380027681F8A  PowerHA_9 None
hdisk15         27681F8B      0017380027681F8B  PowerHA_10 None
```

Ideally, to simplify administration, the hdisk names should be the same on all cluster nodes. If not, then change the hdisk name using the **rendev** command on all cluster nodes as shown in Example 7-74.

Example 7-74 Changing a physical volume name

```
# rendev -l hdisk15 -n hdisk27
hdisk27
```

The **rendev** command syntax is as follows:

```
rendev -l Name [-n NewName] [-u]
```

After changing the name, remove the hdisk using the **rmdev** command and run **cfgmgr** on all nodes as shown in Example 7-75. Since the physical volumes are in a remote mirror relationship, they show up with the same PVID after running **cfgmgr**.

Example 7-75 Different PVIDs

```
(0) root @ c665f1sq08: /  
# rmdev -dl hdisk27  
hdisk27 deleted  
  
(0) root @ c665f1sq08: /  
# cfgmgr
```

Adding volume to volume group via SMIT/CSPOC

Now add the physical volume hdisk27 mapped to the PowerHA_8 XIV volume, which is associated with the redbook_sync_cg mirror relationship. Use the following SMIT path:

smit sysmirror → System Management (C-SPOC) → Storage → Volume Groups → Set Characteristics of a Volume Group → Add a Volume to a Group

Select the redbook_sync_vg volume group to extend it as shown in Figure 7-53.

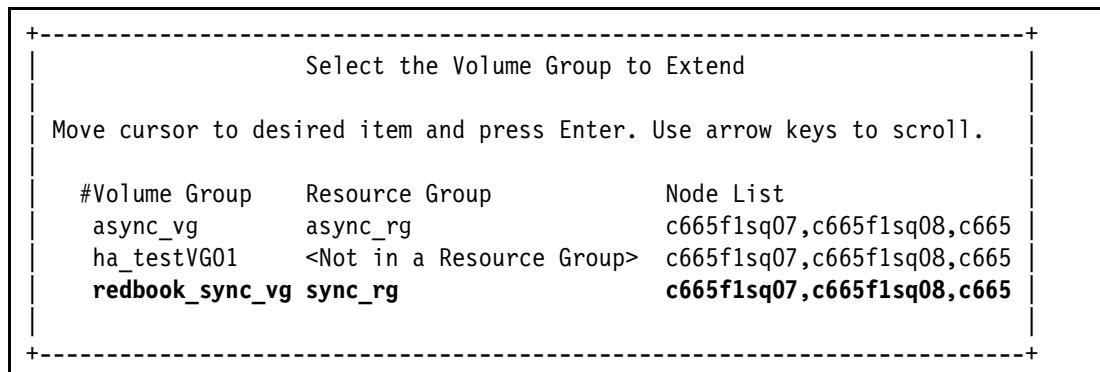


Figure 7-53 Select the volume group to extend it

Then choose the specified volume group and press Enter. This prompts you to select the physical volumes as shown in Figure 7-54 on page 329.

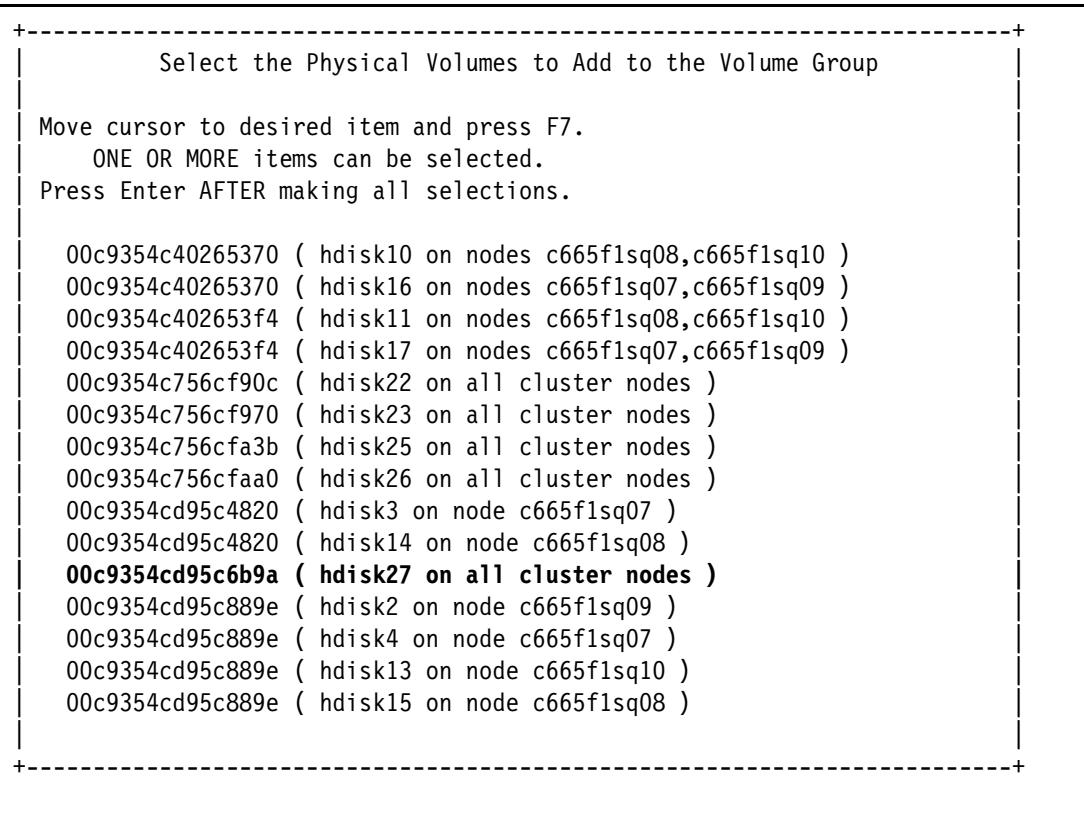


Figure 7-54 Selecting the physical volume to add

Choose the specified physical volume, hdisk27, on all cluster nodes and press Enter. This displays the SMIT panel shown in Figure 7-55.

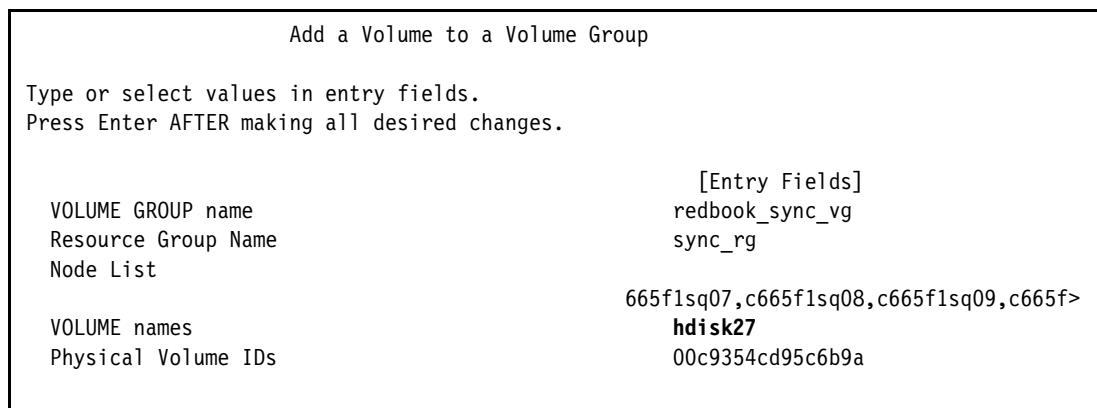


Figure 7-55 Adding the volume group SMIT panel

Verify that all the fields are correct before continuing. When you press Enter, this prompts for confirmation (Figure 7-56 on page 330). Press Enter again to continue.

COMMAND STATUS		
Command: OK	stdout: yes	stderr: no
Before command completion, additional instructions may appear below.		

Figure 7-56 Add a volume to the volume group status

Verify that the physical volume was added to the volume group successfully on the primary site nodes as shown in Example 7-76.

Example 7-76 Verifying a physical volume added to a volume group

```
# dsh lspv | grep hdisk27
c665f1sq07.ppd.pok.ibm.com: hdisk27 00c9354cd95c6b9a redbook_sync_vg concurrent
c665f1sq10.ppd.pok.ibm.com: hdisk27 00c9354cd95c6b9a redbook_sync_vg
c665f1sq08.ppd.pok.ibm.com: hdisk27 00c9354cd95c6b9a redbook_sync_vg concurrent
c665f1sq09.ppd.pok.ibm.com: hdisk27 00c9354cd95c6b9a redbook_sync_vg
```

We successfully added the physical volume to the XIV remote mirror replicated resource.

Attention: Notice that we did not change the resource group, nor synchronize the cluster. This is because technically the cluster resources have not changed.

Adding volume to volume group using clmgr

You can also add a volume to a volume group using the **clmgr** command shown in Example 7-77.

Example 7-77 Add volume to volume group using clmgr

```
(0) root @ c665f1sq07: /
# clmgr modify volume_group redbook_sync_vg ADD=hdisk27

ERROR: "00c9354cd95c6b9a" maps to different disks across c665f1sq07, , c665f1sq08, c665f1sq09,
c665f1sq10:

c665f1sq07: hdisk27, 00c9354cd95c6b9a, 49bc7338-1cb6-72a6-d454-45770fba006f
c665f1sq09: hdisk27, 00c9354cd95c6b9a, 01c1f63b-45d0-cfac-95fc-bce80bbd9f3a
c665f1sq10: hdisk27, 00c9354cd95c6b9a, 01c1f63b-45d0-cfac-95fc-bce80bbd9f3a
```

Either select a different disk, or try specifying the disk that you want by its PVID.

ERROR: "00c9354cd95c6b9a" maps to different disks across c665f1sq07, , c665f1sq08, c665f1sq09,
c665f1sq10:

```
c665f1sq07: hdisk27, 00c9354cd95c6b9a, 49bc7338-1cb6-72a6-d454-45770fba006f
c665f1sq09: hdisk27, 00c9354cd95c6b9a, 01c1f63b-45d0-cfac-95fc-bce80bbd9f3a
c665f1sq10: hdisk27, 00c9354cd95c6b9a, 01c1f63b-45d0-cfac-95fc-bce80bbd9f3a
```

Either select a different disk, or try specifying the disk that you want by its PVID.

Note: During our testing an error appeared and is shown in the previous example (Example 7-77 on page 330). However, our test completed successfully.

You can verify the hdisk addition to the volume group using the **clmgr** command or the **lsvg** command as shown in Example 7-78.

Example 7-78 Check that the volume hdisk27 was added to the volume group

```
(0) root @ c665f1sq07: /  
# clmgr modify volume_group redbook_sync_vg ADD=hdisk27  
  
(0) root @ c665f1sq07: /  
# lsvg -p redbook_sync_vg  
redbook_sync_vg:  
PV_NAME          PV STATE    TOTAL PPs   FREE PPs   FREE DISTRIBUTION  
hdisk21          active     4102        3838      821..556..820..820..821  
hdisk27          active     4102        4102      821..820..820..820..821  
  
(0) root @ c665f1sq07: /  
# clmgr query volume_group redbook_sync_vg  
NAME="redbook_sync_vg"  
TYPE="ORIGINAL"  
NODES="c665f1sq07,c665f1sq08,c665f1sq09,c665f1sq10"  
LOGICAL_VOLUMES="log1v04,fs1v04,lv02"  
PHYSICAL_VOLUMES="hdisk21@c665f1sq07@00c9354c47833c82,hdisk27@c665f1sq07@00c9354cd95c6b9a"  
MIRROR_POOLS=""  
STRICT_MIRROR_POOLS=""  
RESOURCE_GROUP="sync_rg"  
.....  
.....  
MAJOR_NUMBER="59"  
IDENTIFIER="00c9354c00004c000000013b47833cda"  
TIMESTAMP="50c9f39e319b9202"
```

7.7.2 Remove volumes from a XIV remote mirror replicated resource

We can remove volumes from the XIV remote mirror replication resource, or mirror group, without affecting the PowerHA SystemMirror cluster services using dynamic automatic reconfiguration (DARE). We performed this using both SMIT and the **clmgr** command.

Remove by using SMIT/CSPoC

In our scenario, we removed the physical volume, hdisk27, that we added in the previous section. First we removed the physical volume from the volume group redbook_sync_vg using the SMIT C-SPOC panel as follows:

smit sysmirror → System Management (C-SPOC) → Storage → Volume Groups → Set characteristics of a Volume Group → Remove a Volume from a Volume Group

A menu panel lists the volume group to reduce. Select the volume group you want to reduce. We selected the redbook_sync_vg volume group as shown in Figure 7-57 on page 332.

```

+-----+
          Select the Volume Group to Reduce
+-----+
Move cursor to desired item and press Enter. Use arrow keys to scroll.

#Volume Group      Resource Group          Node List
  async_vg         async_rg                c665f1sq07,c665f1sq08,c665
  ha_testVG01      <Not in a Resource Group> c665f1sq07,c665f1sq08,c665
  redbook_sync_vg  sync_rg                c665f1sq07,c665f1sq08,c665
+-----+

```

Figure 7-57 Volume groups to reduce

Choose the specified volume group and press Enter. This prompts for the physical volume selection as shown in Figure 7-58. We selected the physical volume hdisk27.

```

+-----+
          Select the Physical Volume to Remove    ||
+-----+
Move cursor to desired item and press Enter.

# Reference node      Physical Volume Name
  c665f1sq07           hdisk21
  c665f1sq07           hdisk27
+-----+

```

Figure 7-58 Selecting the physical volume to remove

Choose the desired physical volume and press Enter. This displays a removal panel as shown in Figure 7-59.

```

Remove a Volume from a Volume Group

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

[VOLUME GROUP name]          [Entry Fields]
VOLUME GROUP name             redbook_sync_vg
Resource Group Name           sync_rg
Node List                      c665f1sq07,c665f1sq08,c665f1sq09,c665f>
                               hdisk27
                               c665f1sq07
                               no +
VOLUME names
Reference node
FORCE deallocation of all partitions on
this physical volume?

```

Figure 7-59 Physical volume removal

Verify all the parameters before pressing Enter. This asks for confirmation. Press Enter to continue. Then remove the physical volume from the volume group. Verify the physical volumes as shown in Example 7-79.

Example 7-79 Physical volume verification

```
(0) root @ c665f1sq07: /  
# dsh lspv | grep hdisk27  
c665f1sq07.ppd.pok.ibm.com: hdisk27      00c9354cd95c6b9a None  
c665f1sq09.ppd.pok.ibm.com: hdisk27      00c9354cd95c6b9a None  
c665f1sq08.ppd.pok.ibm.com: hdisk27      00c9354cd95c6b9a None  
c665f1sq10.ppd.pok.ibm.com: hdisk27      00c9354cd95c6b9a None  
(0) root @ c665f1sq07: /
```

Remove by using clmgr

You can also remove physical volumes from the volume group using the **clmgr** command as shown in Example 7-80.

Example 7-80 Removing physical volumes from a volume group using clmgr

```
(0) root @ c665f1sq07: /  
# clmgr modify volume_group redbook_sync_vg REMOVE=hdisk27  
  
ERROR: "00c9354cd95c6b9a" maps to different disks across c665f1sq07, , c665f1sq08, c665f1sq09,  
c665f1sq10:  
  
c665f1sq07: hdisk27, 00c9354cd95c6b9a, 49bc7338-1cb6-72a6-d454-45770fba006f  
c665f1sq09: hdisk27, 00c9354cd95c6b9a, 01c1f63b-45d0-cfac-95fc-bce80bbd9f3a  
c665f1sq10: hdisk27, 00c9354cd95c6b9a, 01c1f63b-45d0-cfac-95fc-bce80bbd9f3a
```

Either select a different disk, or try specifying the disk that you want by its PVID.

Note: During our testing an error appeared and is shown in the previous example (Example 7-80). However, our test completed successfully.

Check the volume group details using the **clmgr** and **lsvg** commands as shown in Example 7-81.

Example 7-81 Checking that the volume was removed from the volume group

```
(0) root @ c665f1sq07: /  
# lsvg -p redbook_sync_vg  
redbook_sync_vg:  
PV_NAME          PV STATE        TOTAL PPs    FREE PPs    FREE DISTRIBUTION  
hdisk21         active          4102        3838        821..556..820..820..821  
  
(0) root @ c665f1sq07: /  
# clmgr query volume_group redbook_sync_vg  
NAME="redbook_sync_vg"  
TYPE="ORIGINAL"  
NODES="c665f1sq07,c665f1sq08,c665f1sq09,c665f1sq10"  
LOGICAL_VOLUMES="log1v04,fslv04,1v02"  
PHYSICAL_VOLUMES="hdisk21@c665f1sq07@00c9354c47833c82"  
MIRROR_POOLS=""  
STRICT_MIRROR_POOLS=""
```

```
RESOURCE_GROUP="sync_rg"
.....
.....
MAJOR_NUMBER="59"
IDENTIFIER="00c9354c00004c000000013b47833cda"
TIMESTAMP="50c9f6992a92a573"
```

Now remove and deactivate the mirroring pair PowerHA_8 XIV volume from the XIV remote mirror consistency group, as shown in Example 7-82.

Example 7-82 Removing volumes from a XIV consistency group

```
(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 cg_remove_vol
vol=PowerHA_8 -y
Command executed successfully.

(0) root @ c665f1sq07: /
# /opt/xiv/XIVGUI/xcli -u admin -p adminadmin -m 9.114.63.166 mirror_deactivate
vol=PowerHA_8 -y
Command executed successfully.
```

7.7.3 Adding a file system to a volume group

In this section, we add a new file system to an existing volume group. We show how to do this in two ways: First using C-SPOC and second via the `c1mgr` command line.

Adding a file system using SMIT/C-SPOC

You can add a file system to a volume group associated with the XIV remote mirror replicated resource using the SMIT C-SPOC panel. Change the resource group attributes to include the newly created file system. Use the following SMIT path:

smit sysmirror → System Management (C-SPOC) → Storage → File Systems → Add a File System

This adds a file system to a volume group associated with a XIV remote mirror replicated resource as shown in Figure 7-60.

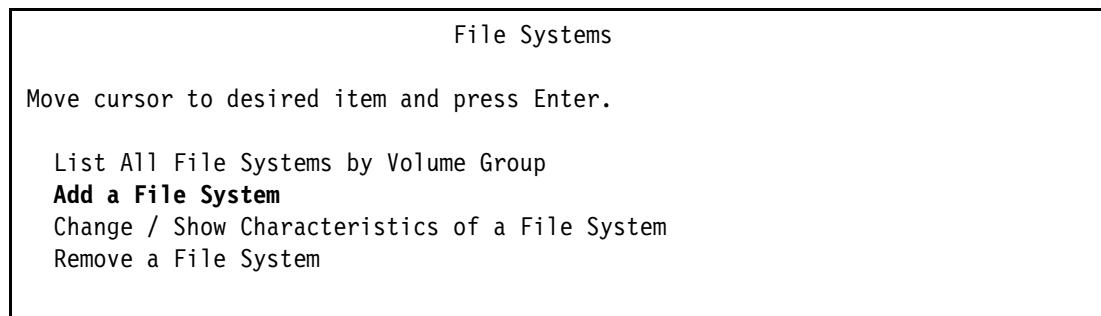


Figure 7-60 Adding a file system SMIT menu

The SMIT panel lists the volume groups on that node. Select the volume group associated with the XIV remote mirror replicated resource. In this example, we selected the redbook_sync_vg volume group as shown in Figure 7-61 on page 335.

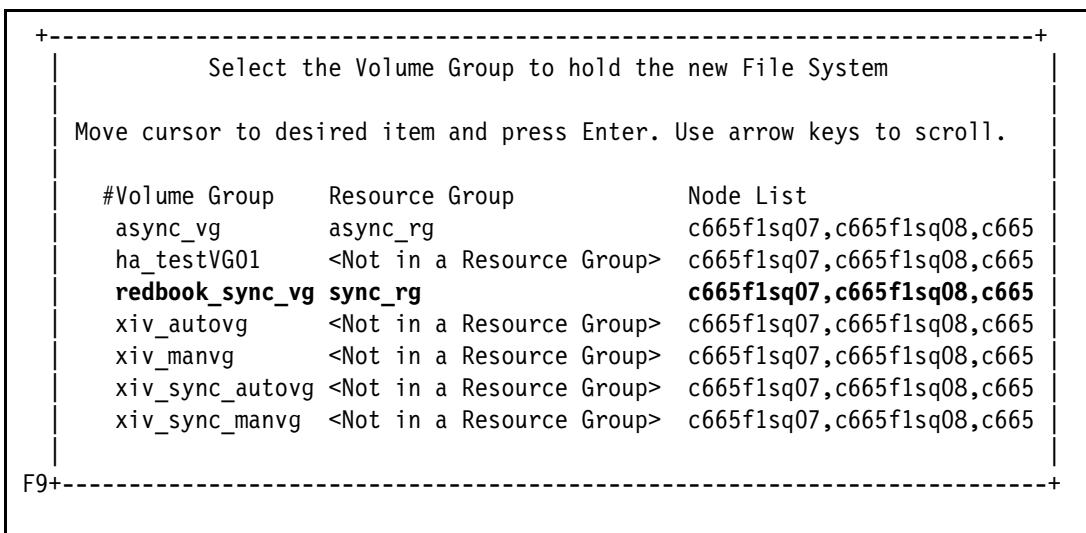


Figure 7-61 Select a volume group

After selecting the volume group, press Enter. This prompts for the type of file system to add. In our scenario, we selected Enhanced Journaled File System as shown in Figure 7-62.

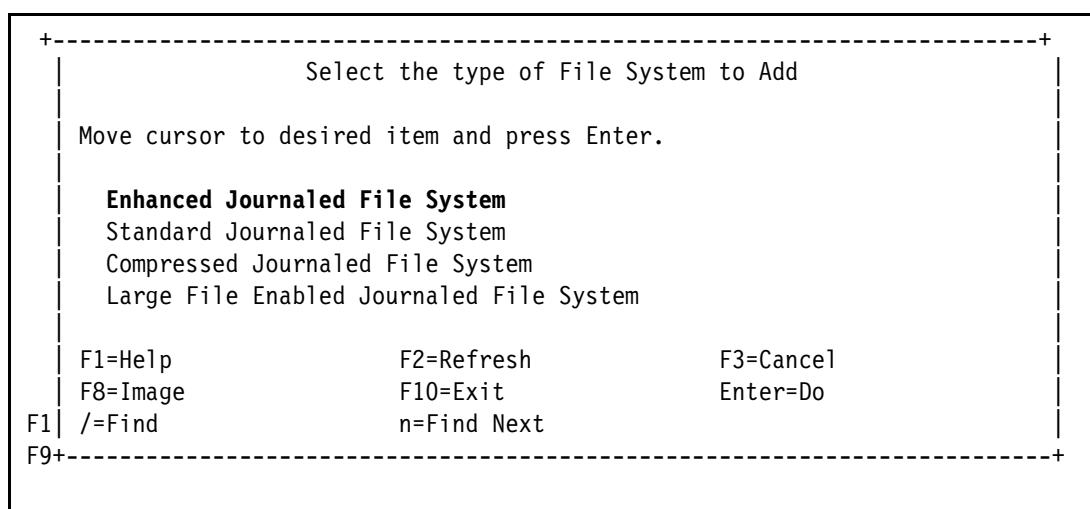


Figure 7-62 Select Enhanced Journaled File system

After selecting the file system type, press Enter. This displays the file system attributes as shown in Figure 7-63 on page 336. “Number of units” and “MOUNT POINT” are mandatory attributes. Specify the number of units or file size you want to create. Also specify the directory or location where you want to mount the file system.

Add an Enhanced Journaled File System

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

[Entry Fields]		
Resource Group	sync_rg	
* Node Names	c665f1sq07,c665f1sq08>	
Volume group name	redbook_sync_vg	
SIZE of file system	Megabytes	+
Unit Size	[50]	#
* Number of units		
* MOUNT POINT	[/testfs]	/
PERMISSIONS	read/write	+
Mount OPTIONS	[]	+
Block Size (bytes)	4096	+
Inline Log?	no	+
Inline Log size (MBytes)	[]	#
Logical Volume for Log		+
Extended Attribute Format	Version 1	+
ENABLE Quota Management?	no	+
Enable EFS?	no	+

Figure 7-63 File system attributes

After specifying all file system attributes, press Enter. This creates the file system shown in Figure 7-64.

COMMAND STATUS

Command: OK	stdout: yes	stderr: no
-------------	-------------	------------

Before command completion, additional instructions may appear below.

```

c665f1sq07: lv02
c665f1sq07: File system created successfully.
c665f1sq07: 53040 kilobytes total disk space.
c665f1sq07: New File System size is 106496
c665f1sq07: /testfs is now guarded against concurrent mounts.

```

Figure 7-64 File system created successfully

We added a file system called /testfs with 50 MB size on the redbook_sync_vg volume group. The volume group is associated with the XIV remote mirror replicated resource or mirror group, sync_mg.

Adding a file system using clmgr

You can also add file system using the **clmgr** command shown in Example 7-83 on page 337.

Example 7-83 Adding a file system using the clmgr command

```
(0) root @ c665f1sq07: /  
# clmgr add file_system /testfs VOLUME_GROUP=redbook_sync_vg TYPE=enforced  
UNITS=50 SIZE_PER_UNIT=megabytes PERMISSIONS=rw BLOCK_SIZE=4096  
c665f1sq07: 1v01  
c665f1sq07: File system created successfully.  
c665f1sq07: 53040 kilobytes total disk space.  
c665f1sq07: New File System size is 106496  
c665f1sq07: /testfs is now guarded against concurrent mounts.
```

```
(0) root @ c665f1sq07: /
```

Verify the newly created file system using the **lsvg** command shown in Example 7-84.

Example 7-84 Verifying newly created file system

```
(0) root @ c665f1sq07: /  
# lsvg -l redbook_sync_vg  
redbook_sync_vg:  
LV NAME          TYPE    LPs    PPs    PVs   LV STATE    MOUNT POINT  
log1v04         jfs2log 1       1       1   open/syncd  N/A  
fs1v04          jfs2     250    250    1   open/syncd  /redbook_sync_fs  
1v01            jfs2     13     13     1   open/syncd  /testfs
```

You can mount the newly created file system manually on the node where the resource group is online as follows:

```
# mount /testfs
```

This is needed only if you defined file systems explicitly in the resource group attributes as shown in Example 7-63 on page 320. You did not define file systems explicitly (or keep the file system field empty, which means ALL file systems in the volume group) during the initial configuration, thus the newly created file system will mount automatically.

If you defined file systems explicitly and you want to define a newly created file system in the resource group attributes, use the following SMIT path:

smit sysmirror → Custom Cluster Configuration → Resource Groups → Change/Show Resources and Attributes for a Resource Group

Then manually enter the file system name as shown in Figure 7-65 on page 338.

Change/Show All Resources and 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	sync_rg
Inter-site Management Policy	Prefer Primary Site
Participating Nodes from Primary Site	c665f1sq07 c665f1sq08
Participating Nodes from Secondary Site	c665f1sq09 c665f1sq10
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority
Fallback Policy	Node In The List
Fallback Timer Policy (empty is immediate)	Fallback To Higher
Service IP Labels/Addresses	Priority Node In The List
Volume Groups	[]
Filesystems (empty is ALL for VGs specified)	[f1sq08_ensvc] +
/testfs] +	[redbook_sync_vg] +
Filesystems Consistency Check	[/redbook_sync_fs
Filesystems Recovery Method	fsck +
XIV Replicated Resources	sequential +
DS8000-Metro Mirror (In-band) Resources +	sync_mg +
[BOTTOM]	

Figure 7-65 Change resource group attributes

After adding the file system, verify and synchronize the cluster using DARE. Use the following SMIT path:

smit sysmirror → Custom Cluster Configuration → Resource Groups → Verify and Synchronize Cluster Configuration.

Note: DARE operation is not required. You can mount the file system manually.

7.7.4 Change the recovery action

The recovery action is a mandatory field when creating a XIV remote mirror replicated resource. Refer to Example 7-61 on page 317 for creating XIV a remote mirror replicated resource or mirror group. The recovery action determines how the cluster reacts during site failure. There are two types of recovery actions:

- ▶ Automatic

The cluster automatically attempts to acquire the PowerHA SystemMirror resource group associated with the XIV remote mirror replicated resource or mirror group in case of a failure of the primary site for the resource group.

- ▶ Manual

In case of a site failover, the cluster provides an action plan in the log file (`/var/hacmp/log/hacmp.out`), and will not acquire the PowerHA SystemMirror resource

group on the secondary site. User intervention is required to manually failover the XIV remote mirroring relationships and activate the resource groups in the secondary site.

You can modify the recovery action for the mirror group using DARE. Follow the SMIT path:

smit sysmirror → Cluster Applications and Resources → Resources → Configure XIV Remote Mirror Resources → Configure Mirror Groups → Change/Show a Mirror Group

Then select the mirror group to modify the recovery action. Figure 7-66 shows the initial recovery action, set to manual. We changed it to automatic.

Change/Show a Mirror Group	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
* Mirror Group Name	[Entry Fields]
New Mirror Group Name	sync_mg
* Storage System Name	[]
* Mirror Mode	SS_site1 SS_site2 +
* Vendor Specific Identifier	Synchronous +
* Recovery Action	redbook_sync_cg +
	Manual +

Figure 7-66 Initial recovery action

For the recovery action press Tab to change the recovery action from manual to automatic as shown in Figure 7-67.

Change/Show a Mirror Group	
Type or select values in entry fields.	
Press Enter AFTER making all desired changes.	
* Mirror Group Name	[Entry Fields]
New Mirror Group Name	sync_mg
* Storage System Name	[]
* Mirror Mode	SS_site1 SS_site2 +
* Vendor Specific Identifier	Synchronous +
* Recovery Action	redbook_sync_cg +
	Automatic +

Figure 7-67 Changing the recovery action to automatic

After changing the recovery action to automatic, press Enter. The SMIT panel shows OK. Changes are not activated on a running cluster until you verify and synchronize the cluster configuration. Use the following SMIT path:

smit sysmirror → Cluster Applications and Resources → Resources → Verify and Synchronize Cluster Configuration

The SMIT panel asks for confirmation to continue with the verify and synchronize operation. Press Enter to continue the verification and synchronization.

7.8 IBM PowerHA SystemMirror demonstration with XIV replication

The following is a failover demonstration of PowerHA SystemMirror Enterprise Edition for AIX utilizing XIV async replication:

<http://youtu.be/RJ500030agM>

System administration, monitoring, maintenance, and management

This part contains information about how to manage, monitor, and maintain your PowerHA cluster.

The chapters are as follows:

- ▶ Chapter 8, “Migrating to PowerHA SystemMirror 7.1.2 Enterprise Edition” on page 343.
- ▶ Chapter 9, “PowerHA 7.1.2 for IBM Systems Director plug-in enhancements” on page 393.
- ▶ Chapter 10, “Cluster partition management” on page 449.



Migrating to PowerHA SystemMirror 7.1.2 Enterprise Edition

This chapter includes the following topics for migrating to PowerHA SystemMirror 7.1.2 Enterprise Edition:

- ▶ Migration planning
- ▶ Clmigcheck explained
- ▶ Migration scenarios
 - Rolling from PowerHA SystemMirror 6.1 Enterprise Edition
 - Snapshot from IBM PowerHA SystemMirror 6.1 Enterprise Edition
 - Offline migration from PowerHA 6.1 Enterprise Edition with IPv6 configuration

8.1 Migration planning

It is important to properly plan before migrating your existing cluster to the IBM PowerHA SystemMirror 7.1.2 Enterprise Edition (EE). Following are the basic requirements for a successful migration.

Before performing a migration, you should always have a backout plan in case any problems are encountered. Some general suggestions are:

- ▶ Create a backup of rootvg

In most cases of upgrading PowerHA, performing an update or upgrade of AIX is also required. So always save your existing rootvg. Our preferred method is to create a clone via `alt_disk_copy` to another free disk on the system. That way a simple change to the bootlist and a reboot returns the system back to the beginning state quite easily.

Other options are available, such as `mksysb`, `alt_disk_install` and `multibos`.

- ▶ Save the existing cluster configuration

Create a cluster snapshot before migrating. By default it is stored in `/usr/es/sbin/cluster/snapshots`. Make a copy of it and also save a copy off of the cluster nodes for additional insurance.

- ▶ Save any user-provided scripts

This most commonly refers to custom events, pre/post events, application controller, and application monitoring scripts.

Verify, via `1s1pp -h cluster.*`, that the current version of PowerHA is in the COMMIT state and not in the APPLY state. If not, run `smit install_commit` before installing the latest software version.

Note: The initial release of IBM PowerHA SystemMirror 7.1.2 Enterprise Edition had a support limitation for all third party storage replication previously supported with PowerHA 6.1 Enterprise Edition. However this limitation was removed and appropriate levels required can be found at:

<http://www-03.ibm.com/support/techdocs/atstr.nsf/WebIndex/FLASH10822>

8.1.1 Requirements

This section explains the software and hardware requirements.

Software

The minimum required software versions are:

- ▶ AIX 6.1 TL8 SP1 with RSCT 3.1.4.0
- ▶ AIX 7.1 TL2 SP1 with RSCT 3.1.4.0
- ▶ New additional requisite filesets of:
 - `bos.cluster`
 - `bos.ahafs`
 - `bos.clvm.enh`

This was optional in previous versions so you may already have this installed.

- `devices.common.IBM.storfwk` (for SAN heartbeat)
- `clic.rte` (for secured encryption communication options of clcomd)

- ▶ Other optional filesets suggested to install
 - cas.agent (for Systems Director plug-in)
- ▶ VIOS 2.2.0.1-FP24 SP01
- ▶ PowerHA SystemMirror 7.1.2 Enterprise Edition SP1

Important: We strongly recommend to always start with the latest service packs available for PowerHA, AIX, and VIOS.

Hardware

The hardware characteristics are as follows:

- ▶ Support is available only for POWER5 technologies and later.
- ▶ Shared disks for the cluster repository.

Choose an appropriate size. Usually 1 GB is sufficient for a two-node cluster. Ensure that the storage subsystem hosting the repository disk is supported. Also, make sure that the adapters and the multipath driver used for the connection to the repository disk are supported. Additionally, most PowerHA SystemMirror Enterprise Edition clusters will become *linked* clusters. In this case two repository disks are required, one for each site. The only requirement is for it to be accessible within each site and not across sites.

It is possible to repurpose an existing disk heartbeat device as the cluster repository disk. However, the disk must be clear of any contents other than a PVID.

For more information, see 3.1.4, “Cluster repository disk” on page 47.

Multicast must be enabled on your cluster network infrastructure. Ensure that the multicast traffic generated by any of the cluster nodes is properly forwarded by the network infrastructure between all cluster nodes. For more information, see “Cluster multicast IP address and PowerHA site configuration” on page 31.

HBA/SAN level heartbeating

Though it is a good practice to utilize as many different heartbeating lines of communication as possible, this is optional. It is only used within a site and not across sites. If desired, then you must have an adapter that has the *tme* attribute to enable. This usually applies to most 4 Gb and newer FC adapters. For more information about utilizing this feature, see “SAN-based heartbeat” on page 40.

Deprecated features

Starting with PowerHA SystemMirror 7.1, the following features are no longer available:

1. IP address takeover (IPAT) via IP replacement
2. Locally administered address (LAA) for hardware MAC address takeover (HWAT)
3. Heartbeat over IP aliases
4. The following IP network types:
 - ATM
 - FDDI
 - Token Ring
5. The following point-to-point (non-IP) network types:
 - RS232

- TMSCSI
 - TMSSA
 - Disk heartbeat (diskhb)
 - Multinode disk heartbeat (mndhb)
6. Two-node configuration assistant
 7. WebSMIT (replaced with the IBM Systems Director plug-in)

Though PowerHA Enterprise Edition was never supported with WebSMIT, PowerHA SystemMirror 7.1.2 Enterprise Edition is supported with the IBM Systems Director plug-in.

Important: If your cluster is configured with any of the features listed above in points 1 through 4, your environment cannot be migrated. You must either change or remove the features before migrating, or simply remove the cluster and configure a new one with the new version of PowerHA.

8.1.2 Migration options

In this section, we cover key terms and the options available for migrating.

Offline	A type of migration where PowerHA is brought offline on all nodes prior to performing the migration. During this time, resources are not available.
Rolling	A type of migration from one PowerHA version to another during which cluster services are stopped on one node at a time. That node is upgraded and reintegrated into the cluster before the next node is upgraded. It requires little down time, mostly by moving the resources between nodes while each node is being upgraded.
Snapshot	A type of migration from one PowerHA version to another during which you take a snapshot of the current cluster configuration, stop cluster services on all nodes, install the preferred version of PowerHA SystemMirror, and then convert the snapshot by running the c1convert_snapshot utility. Then restore the cluster configuration from the converted snapshot.
Non-Disruptive	A node can be “unmanaged” allowing all resources on that node to remain operational when cluster services are stopped. It generally can be used when applying service packs to the cluster. This option does <i>not</i> apply when migrating to version 7.1.x from a prior version.
Mixed cluster	Nodes in a cluster running two different versions of PowerHA. A cluster in this state may be operational for a long period, but the configuration cannot be changed until all nodes have been upgraded. This usually only applies while performing a rolling migration.

Tip: After upgrading, the following line is added to the /etc/syslog.conf file:

```
*.info /var/adm/ras/syslog.caa rotate size 1m files 10
```

We suggest enabling verbose logging by adding the following line:

```
*.debug /tmp/syslog.out rotate size 10m files 10
```

Then execute **refresh -s syslogd**. This provides valuable information if troubleshooting is required.

8.1.3 Offline method

The following is an overview of the steps required to perform an offline migration. These steps may often be performed in parallel since the entire cluster will be offline.

Important: We strongly recommend to always start with the latest service packs available for PowerHA, AIX, and VIOS.

1. Stop cluster services on all nodes.
Choose to bring resource groups offline.
2. Upgrade AIX (if needed).
3. Install additional requisite filesets as listed in “Software” on page 344.
Reboot.
4. Verify that **clcmd** is active:

```
lssrc -s clcmd
```
5. Update /etc/cluster/rhosts.
Enter either cluster node hostnames or IP addresses; only one per line.
6. **Refresh -s clcmd**.
7. Execute **clmigcheck** on one node.
 - Choose option 1 to verify that the cluster configuration is supported (assuming no errors).
 - Then choose option 3.
 - Choose the repository disk device to be used for each site.
 - Enter the multicast address for each site (or leave blank and one is assigned).
 - Exit the **clmigcheck** menu.
 - Review contents of /var/clmigcheck/clmigcheck.txt for accuracy.
8. Upgrade PowerHA on one node.
 - Install base level install images only (apply service packs later).
 - Review the /tmp/clconvert.log file.
9. Execute **clmigcheck** and upgrade PowerHA on the remaining node.
When executing **clmigcheck** on each additional node, the menu does not appear and no further actions are needed. On the last node it creates the CAA cluster.
10. Restart cluster services.

8.1.4 Rolling method

The following is an overview of the steps required to perform a rolling migration. These steps should be performed completely on one node at a time. An example of performing a rolling migration can also be found in 8.3.1, “Rolling from PowerHA SystemMirror 6.1 Enterprise Edition” on page 351.

Important: We strongly recommend to always start with the latest service packs available for PowerHA, AIX, and VIOS.

1. Stop cluster services on one node (move resource group as needed).

2. Upgrade AIX (if needed).
3. Install additional requisite filesets as listed in “Software” on page 344.
Reboot.
4. Verify that **clcmd** is active.
`lssrc -s clcmd`
5. Update /etc/cluster/rhosts.
Enter either cluster node hostnames or IP addresses; only one per line.
- 6. Refresh -s clcmd**
7. Execute **clmigcheck**.
 - First with option 1 (assuming no errors).
 - Then choose option 3.
 - Choose a repository disk device to be used for each site.
 - Enter the multicast address for each site (or leave blank and one is assigned).
 - Exit the **clmigcheck** menu.
 - Review the contents of /var/clmigcheck/clmigcheck.txt for accuracy.
8. Upgrade PowerHA.
 - Install base level install images only (apply Service Packs later.)
 - Review the /tmp/clconvert.log file.
9. Restart cluster services (move resource group back if needed).
10. Repeat the steps above for each node.

When executing **clmigcheck** on each additional node, a menu does not appear and no further actions are needed. On the last node it automatically creates the CAA cluster.

8.1.5 Snapshot method

The following is an overview of the steps required to perform a snapshot migration. Most of these steps may often be performed in parallel since the entire cluster will be offline.

Important: We strongly recommend to always start with the latest service packs available for PowerHA, AIX, and VIOS.

1. Stop cluster services on all nodes.
Choose to bring resource groups offline.
2. Create a cluster snapshot.
If you have not previously created one and saved copies of it.
3. Upgrade AIX (if needed).
4. Install additional requisite filesets as listed in “Software” on page 344.
Reboot.
5. Verify that **clcmd** is active.
`lssrc -s clcmd`
6. Update /etc/cluster/rhosts.
Enter either cluster node hostnames or IP addresses; only one per line.

7. Refresh -s **clcmd**
8. Execute **clmigcheck** on one node.
 - Choose option 1 to verify that the cluster configuration is supported (assuming no errors).
 - Then choose option 3.
 - Choose a repository disk device to be used for each site.
 - Enter the multicast address for each site (or leave blank and one is assigned).
 - Exit the **clmigcheck** menu.
 - Review contents of /var/clmigcheck/clmigcheck.txt for accuracy.
9. Uninstall the current version of PowerHA via **smitty remove** and specify **cluster.***
10. Install the new PowerHA version, including service packs on the same node **clmigcheck** was executed.
11. Execute **clmigcheck** on an additional node.
 - When executing **clmigcheck** on each additional node, the menu does not appear and no further actions are needed. On the last node it creates the CAA cluster.
 - After **clmigcheck** completes, you can install the new PowerHA version.
12. Execute **clconvert_snapshot**
 - This command is located in /usr/es/sbin/cluster/conversion.
 - Syntax example: **clconvert_snapshot -v <version migrating from> -s <snapshot>**
13. Restore the cluster configuration from the converted snapshot.
14. Restart cluster services on each node, one at a time.

8.2 Clmigcheck explained

Before migrating to PowerHA version 7, run the **clmigcheck** program to prepare the cluster for migration; see Figure 8-1 on page 350. The program has two functions:

- ▶ It validates the current cluster configuration (ODM via option 1 or snapshot via option 3) for migration. If the configuration is not valid, the program notifies you of any unsupported elements. If an error is encountered, it must be corrected or you cannot migrate. If a warning is displayed, such as for disk heartbeat, you may continue.
- ▶ It prepares for the new cluster by obtaining the disks to be used for the repository disks and multicast addresses.

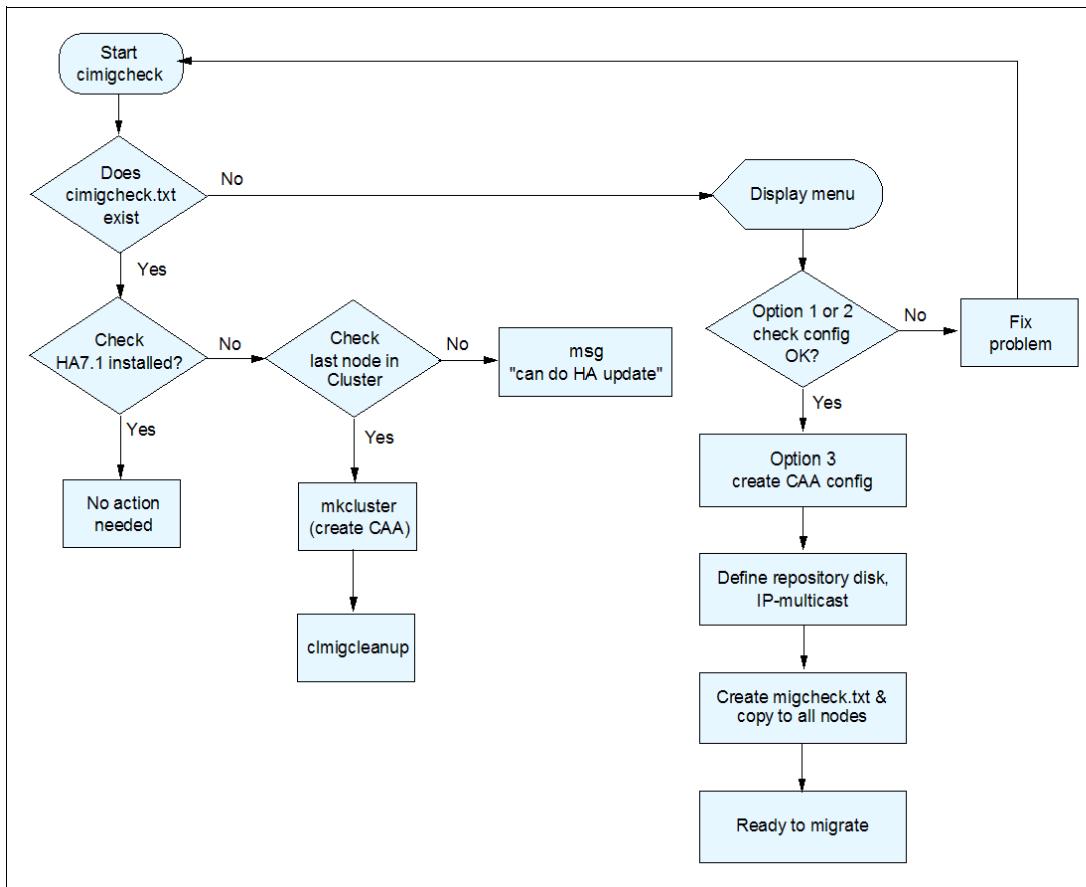


Figure 8-1 Clmigcheck command overview

The **clmigcheck** program (Figure 8-1) goes through the following stages:

1. Performing the first initial run

When the **clmigcheck** program runs, it checks whether it has been run before by looking for a /var/clmigcheck/clmigcheck.txt file. If this file does not exist, the program runs and opens the menu shown in Figure 8-10 on page 355.

2. Verifying that the cluster configuration is suitable for migration

From the **clmigcheck** menu, you can select options 1 or 2 to check your existing ODM or snapshot configuration to see whether the environment is valid for migration.

3. Creating the CAA required configuration

After performing option 1 or 2, choose option 3. Option 3 creates the /var/clmigcheck/clmigcheck.txt file with the information entered and is copied to all nodes in the cluster.

4. Performing the second run on the first node, or first run on any other node that is not the first or the last node in the cluster to be migrated.

If the program is run again and the clmigcheck.txt file already exists, a message is returned indicating that you can proceed with the upgrade of PowerHA.

5. Verifying whether the last node in the cluster is upgraded. When the **clmigcheck** program runs, apart from checking for the presence of the file, it checks whether it is the last node in the cluster to be upgraded. The **1s1pp** command runs against each node in the cluster to establish whether PowerHA has been upgraded. If all other nodes are upgraded, this

command confirms that this node is the last node of the cluster and can now create the CAA cluster.

The **c1migcheck** program uses the **mkcluster** command and passes the cluster parameters from the existing PowerHA cluster, along with the repository disk and multicast address.

8.3 Migration scenarios

In this section, we go through a complete scenario for each of the three options for migrating to IBM PowerHA SystemMirror 7.1.2 Enterprise Edition.

8.3.1 Rolling from PowerHA SystemMirror 6.1 Enterprise Edition

This section describes the rolling migration of PowerHA SystemMirror 6.1 Enterprise Edition.

Test environment overview

This scenario uses a three-node cluster, XIV_cluster, consisting of two nodes in the primary site, NewYork, and one in the secondary site, Texas. Nodes jay and maddi are in the NewYork site, and node peyton in the Texas site. We had two XIV storage subsystems, one for each site. Figure 8-2 provides an overview of the software and hardware tested for the rolling migration.

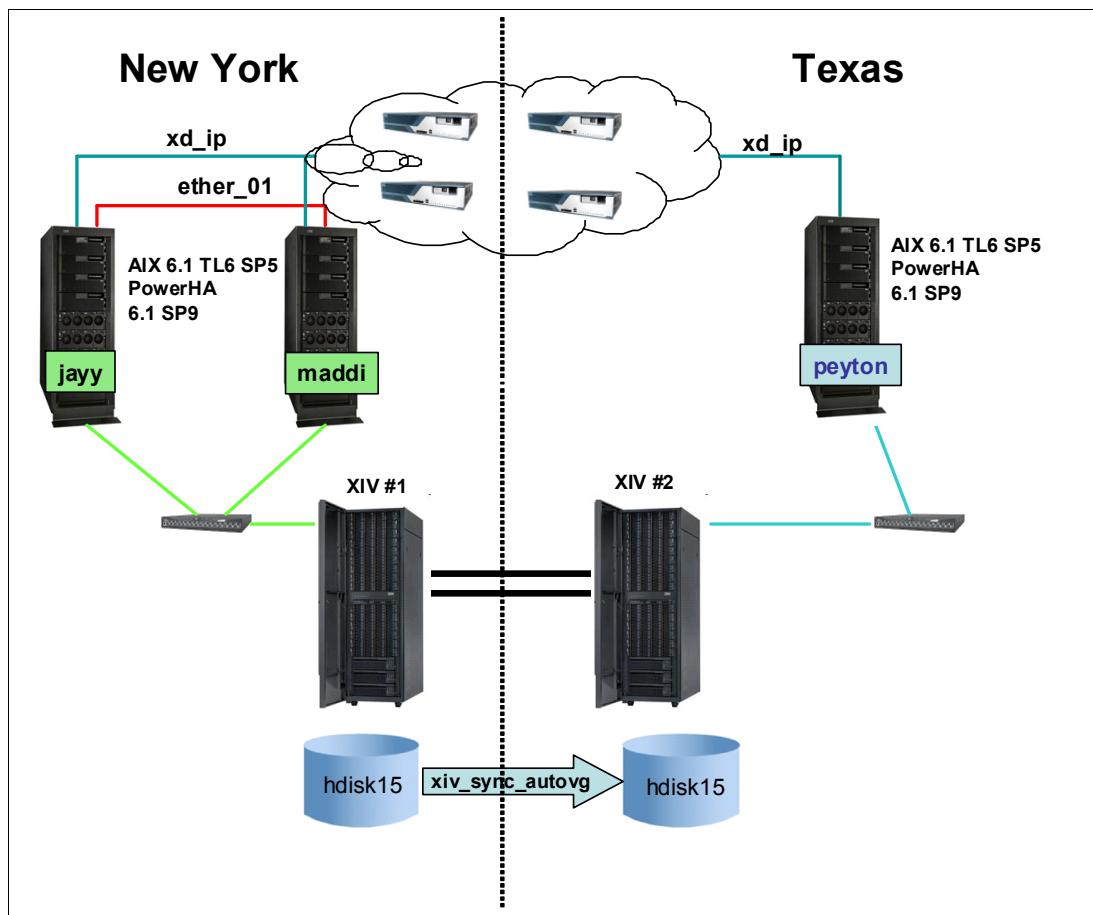


Figure 8-2 XIV replication test lab environment

We had two networks, an XD_ip and ether, named XD_net and ether_01, respectively. We also had a single resource group, RG01, with all three nodes participating as shown from the **cltopinfo** output in Figure 8-3.

```
#>cltopinfo
Cluster Name: XIV_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
Repository Disks: Site 1: None, Site 2:
Cluster IP Addresses: Cluster: , Site 1: , Site 2:
There are 3 node(s) and 2 network(s) defined

NODE jayy:
    Network XD_net
        jayy_xdip      24.48.88.5
    Network ether_01
        jayy_ensvc     30.30.30.65
        jayy_enboot    30.30.30.1
        jayy_enstby1   30.30.30.129

NODE maddi:
    Network XD_net
        maddi_xdip     24.48.88.55
    Network ether_01
        jayy_ensvc     30.30.30.65
        maddi_enstby1  30.30.30.130
        maddi_enboot   30.30.30.2

NODE peyton:
    Network XD_net
        peyton_xdip    24.48.88.25
    Network ether_01
        jayy_ensvc     30.30.30.65
        peyton_enboot  30.30.30.3
        peyton_enstby1 30.30.30.131

Resource Group RG01
    Startup Policy  Online On Home Node Only
    Fallback Policy Fallover To Next Priority Node In The List
    Fallback Policy Fallback To Higher Priority Node In The List
    Participating Nodes   jayy maddi peyton
    Service IP Label     jayy_ensvc
```

Figure 8-3 XIV cluster topology

The resource group contained a single service label, jayy_ensvc, a single volume group, xiv_sync_autovg, and a GENXD replicated resource called PowerHA_SYNC_AUTO(CG as shown from the parsed **c1showres** output in Figure 8-4 on page 353.

Resource Group Name	RG01
Participating Node Name(s)	jayy maddi peyton
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority Node
Fallback Policy	Fallback To Higher Priority Node
Site Relationship	Prefer Primary Site
Service IP Label	jayy_ensvc
Filesystems	ALL
Filesystems Consistency Check	fsck
Filesystems Recovery Method	sequential
Volume Groups	xiv_sync_autovg
Use forced varyon for volume groups, if necessary	false
GENXD Replicated Resources	PowerHA_SYNC_AUTO(CG)

Figure 8-4 XIV cluster resource group

Performing a rolling migration

We begin with all three nodes active in the cluster as seen from the **qha -nev** output in Figure 8-5.

Note: The **qha** command is not a PowerHA standard tool. It is a free utility available at:
<http://www.powerha.lpar.co.uk>

```
Cluster: XIV_cluster (6109)
15:46:44 13Nov12

jayy iState: ST_STABLE
RG01          ONLINE
    en0 jayy
    en1 jayy_enboot jayy_ensvc
    en2 jayy_enstby1
    en3 jayy_xdip
- xiv_sync_autovg(15) -

maddi iState: ST_STABLE
en0 maddi
en1 maddi_enboot
en2 maddi_enstby1
en3 maddi_xdip

peyton iState: ST_STABLE
RG01          ONLINE SECONDARY
    en0 peyton
    en1 peyton_enboot
    en2 peyton_enstby1
    en3 peyton_xdip
```

Figure 8-5 XIV stable cluster status

We used the steps shown in 8.1.4, “Rolling method” on page 347.

Stop cluster services on one node

In our case, we stopped cluster services on the remote node, peyton, first. This was accomplished by executing **smitty clstop** and choosing the options shown in Figure 8-6.

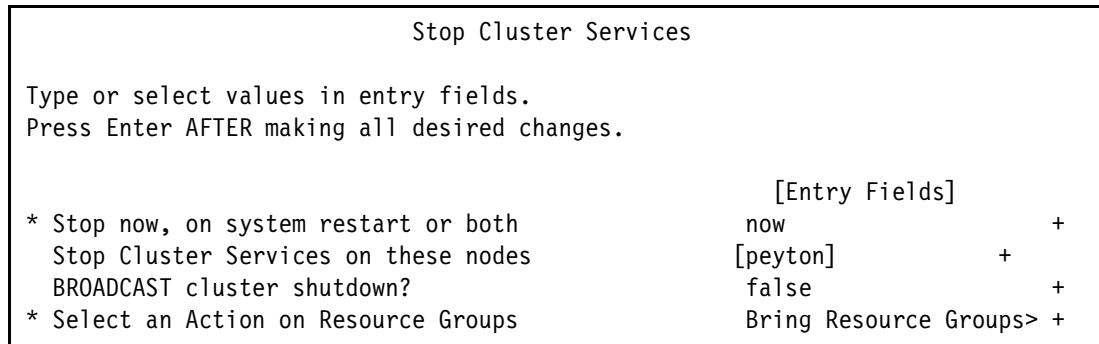


Figure 8-6 Stopping cluster services

After executing, the OK response appears quickly. Make sure the cluster node is in the ST_INIT state. While we show this in Figure 8-7 from **qha -nev**, this can also be found from the **lssrc -ls clstrmgrES** output.

```
Cluster: XIV_cluster (6109)
15:48:24 13Nov12

jayy iState: ST_STABLE
RG01      ONLINE
    en0 jayy
    en1 jayy_enboot jayy_ensvc
    en2 jayy_enstby1
    en3 jayy_xdip
- xiv_sync_autovg(15) -

maddi   iState: ST_STABLE
    en0 maddi
    en1 maddi_enboot
    en2 maddi_enstby1
    en3 maddi_xdip

peyton  iState: ST_INIT
    en0 peyton
    en1 peyton_enboot
    en2 peyton_enstby1
    en3 peyton_xdip
```

Figure 8-7 Verifying that the first node is down

Upgrade AIX

Now you need to update AIX to at least 6.1 TL8 SP1. In our case, since we started with 6.1 TL6 SP5, we simply performed a **smitty update_all** to update to the appropriate level. After updating, we installed the following filesets:

- ▶ bos.cluster
- ▶ bos.ahafs
- ▶ devices.common.IBM.storfwk (for SAN HB)

- ▶ cas.agent
- ▶ clic.rte

Attention: Though it is expected that most existing clusters already have it, make sure that *bos.clvm.enh* is also installed because it is no longer optional.

Upon completing the update and installs, reboot the system.

Clmigcheck

After rebooting, the CAA-specific communications daemon, *clcomd*, should now be running. Verify this by executing the **lssrc -s clcomd** command as shown in Figure 8-8.

```
#peyton>lssrc -s clcomd
Subsystem      Group          PID      Status
clcomd        caa            4980924    active
```

Figure 8-8 CAA *clcomd* service active

Next, edit the CAA-specific communication file, */etc/cluster/rhosts*. You can enter either the hostname for each node, or the IP address that resolves to the hostname. But there must only be one entry per line. We entered the IP addresses as shown in Figure 8-9.

```
#>peyton vi /etc/cluster/rhosts
9.114.135.71
9.114.135.72
9.114.135.73
```

Figure 8-9 */etc/cluster/rhosts* contents

After completing this edit, it is necessary to refresh *clcomd* by executing **refresh -s clcomd**. Then execute **clmigcheck** and the menu shown in Figure 8-10 is displayed.

```
-----[ PowerHA System Mirror Migration Check ]-----
Please select one of the following options:
1      = Check ODM configuration.
2      = Check snapshot configuration.
3      = Enter repository disk and multicast IP addresses.

Select one of the above,"x"to exit or "h" for help:1
```

Figure 8-10 *Clmigcheck* menu

Because this was a rolling migration, we chose *option 1* and press Enter. In most environments, where there is more than one node within a site, it is common to have a disk heartbeat network configured. If that is the case, a warning appears as shown in Figure 8-11 on page 356. This is normal because it is removed during the last phase of the migration. In our scenario, we did not have one configured.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

CONFIG-WARNING: The configuration contains unsupported hardware: Disk Heartbeat network. The PowerHA network name is net_diskhb_01. This will be removed from the configuration during the migration to PowerHA System Mirror 7.1.

Hit <Enter> to continue

Figure 8-11 Clmigcheck disk heartbeat warning

In our case, we also had no errors with unsupported elements, as shown in Figure 8-12.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

The ODM has no unsupported elements.

Hit <Enter> to continue

Figure 8-12 Clmigcheck ODM is error free

After pressing Enter to continue, the panel returns to the main **clmigcheck** menu shown in Figure 8-10 on page 355. This time we chose *option 3* and pressed Enter. We were presented with the site-specific options of choosing either a *stretched* or *linked* cluster as shown in Figure 8-13. For more information about the difference between the two, refer to “Linked cluster and multiple site support” on page 19.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

Your multi-site cluster can be based on a single AIX CAA cluster (a Stretched cluster) or a linked AIX CAA cluster (a Linked cluster).

Select the type of multi-site cluster you want:

- | | |
|---|-------------|
| 1 | = STRETCHED |
| 2 | = LINKED |

Select one of the above or "h" for help or "x" to exit:2

Figure 8-13 Clmigcheck stretched or linked cluster menu option

In our case, we chose *option 2* and pressed Enter. It is expected that almost all existing Enterprise Edition clusters will use this option. Once chosen, a discovery is executed to create a list of shared disks that are currently not members of an exiting volume group. Once the discovery is completed, you get a list of candidate disks to choose from for the repository disk. In our case, it showed the choice for the second site, Texas, first. We chose hdisk22 as shown in Figure 8-14 on page 357.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

Select the disk to use for the repository on site Texas

- 1 = 00c9354cd95d61e2(hdisk19)
- 2 = 00c9354cd95d7b44(hdisk20)
- 3 = 00c9354cd95d9e23(hdisk21)
- 4 = 00c9354cd95dbb26(hdisk22)

Select one of the above or "h" for help or "x" to exit:4

Figure 8-14 Choosing the repository disk for the second site

Another similar menu appeared for choosing a repository disk for the primary site, NewYork. In this case, we chose hdisk10 as shown in Figure 8-15.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

Select the disk to use for the repository on site NewYork

- 1 = 00c9354c16c2488a(hdisk6)
- 2 = 00c9354c3fef6268(hdisk7)
- 3 = 00c9354c3fefec22(hdisk8)
- 4 = 00c9354c402652a8(hdisk9)
- 5 = 00c9354c40265370(hdisk10)

Select one of the above or "h" for help or "x" to exit: 5

Figure 8-15 Choosing the repository disk for the first site

Once choosing the repository disks for each site is complete, the option of entering a multicast address for a site is displayed, as shown in Figure 8-16. In our case, we simply pressed Enter twice, once for each site because the menu refreshes with the next site to let it default and creates one of its own at the time of cluster creation. This now returns to the main **c1migcheck** menu, and you can enter *x* and press Enter to exit out of the menu, as shown in Figure 8-17 on page 358.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

PowerHA System Mirror uses multicast address for internal cluster communication and monitoring. These must be in the multicast range, 224.0.0.0 - 239.255.255.255.

If you make a NULL entry, AIX will generate an appropriate address for you. You should only specify an address if you have an explicit reason to do so, but are cautioned that this address cannot be changed once the configuration is activated (i.e. migration is complete).

h = help

Enter the multicast IP address to use for site Texas:

Figure 8-16 Choosing a multicast address for each site

```
-----[ PowerHA System Mirror Migration Check ]-----  
You have requested to exit clmigcheck.  
Do you really want to exit? (y) y  
Note - If you have not completed the input of repository disks and  
multicast IP addresses, you will not be able to install  
PowerHA System Mirror  
Additional details for this session may be found in  
/tmp/clmigcheck/clmigcheck.log.
```

Figure 8-17 Clmigcheck menu exit

You can now verify the information entered in the **clmigcheck** menu by viewing the contents of the `/var/clmigcheck/clmigcheck.txt` file shown in Figure 8-18.

```
CLUSTER_TYPE:LINKED  
SITE2_REPOSITORY_DISK:Texas:00c9354cd95dbb26  
SITE1_REPOSITORY_DISK:NewYork:00c9354cd95c889e  
SITE1_MULTICAST:site_1:NULL  
SITE2_MULTICAST:site_2:NULL
```

Figure 8-18 Clmigcheck.txt file contents

Upgrade PowerHA

To upgrade PowerHA, simply execute **smitty update_all**, choosing the v7.1.2.0 install images, and always set *ACCEPT new license agreements?* to *yes* as shown in Figure 8-19 on page 359.

Important: Always complete a migration using base levels. Then come back and install any service packs later.

Once the upgrade has completed, check `/tmp/c1convert.log` for any errors. Shown in Example 8-2 on page 359 are the first few lines and last few lines of our log file.

Restart cluster services on peyton by executing **smitty c1start**, leaving the defaults, and press Enter. Normal verification is skipped because this results in a mixed cluster, as shown in Example 8-1.

Example 8-1 Mixed cluster ignores verification

```
---- start ----
```

```
Cluster services are running at different levels across  
the cluster. Verification will not be invoked in this environment.
```

Update Installed Software to Latest Level (Update All)		
Type or select values in entry fields. Press Enter AFTER making all desired changes.		
[TOP]	[Entry Fields]	
* INPUT device / directory for software	.	
* SOFTWARE to update	_update_all	+
PREVIEW only? (update 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	+
VERIFY install and check file sizes?	no	+
DETAILED output?	no	+
Process multiple volumes?	yes	+
ACCEPT new license agreements?	yes	+
Preview new LICENSE agreements?	no	+

Figure 8-19 Update_all of PowerHA

Attention: During our upgrade testing we encountered the following error:

0513-075 - The new subsystem name is already on file. Failed to register the clxd subsystem with AIX SRC.install: Failed while executing the /usr/lpp/cluster.es.genxd/inst_root/cluster.es.genxd.rte.config script.

This is a known issue specific to replicated resources using genxd, like our XIV configuration. The following is documented in release_notes_xd.

It is possible to encounter an error on the install of the cluster.genxd.rte file set. If a previous version of the product was installed or the uninstall of the product was unsuccessful or incomplete, the clxd SRC subsystem could be left defined. In this case, the install of filessetcluster.es.genxd.rte will fail. If this happens, run the following command:

```
rmmssys -s clxd
```

and then re-install the cluster.es.genxd.rte fileset.

Example 8-2 Clconvert.log

```
----- log file for cl_convert: Tue Nov 13 19:06:48 CST 2012
Command line is:
/usr/es/sbin/cluster/conversion/cl_convert -F -v 6.1
No source product specified.
Assume source and target are same product.
Parameters read in from command line are:
    Source Product is HAES.
    Source Version is 6.1.0.
    Target Product is HAES.
    Target Version is 7.1.2.
    Force Flag is set.
```

```
exiting cl_convert.  
Exiting with error code 0. Completed successfully.
```

Once node peyton rejoins the cluster and stabilizes, we could see via the **lssrc -ls clstrmgrES** command that the cluster version had not changed. The reason for this is that all nodes in the cluster had not been upgraded yet. When the last upgraded node rejoins the cluster, the version number is updated. We also saw that the new CAA services were also not active, via Example 8-3.

Example 8-3 Cluster version not updated

```
lssrc -ls clstrmgrES  
  
Current state: ST_STABLE  
sccsid = "@(#)36 1.135.1.112  
src/43haes/usr/sbin/cluster/hacmpd/main.C,hacmp.pe,61haes_r712,1221A_hacmp712  
5/22/1"  
build = "Nov 5 2012 16:31:36 1242F_hacmp712"  
i_local_nodeid 2, i_local_siteid 2, my_handle 3  
m1_idx[1]=0 m1_idx[2]=1 m1_idx[3]=2  
There are 0 events on the Ibcast queue  
There are 0 events on the RM Ibcast queue  
CLversion: 11 <-----  
local node vrmf is 7121 <-----  
cluster fix level is "1" <-----  
  
CAA Cluster Capabilities <-----  
Cluster services are not active. <-----
```

Upgrading the remaining nodes

We repeated the steps previously performed on node peyton, and also covered in 8.1.4, “Rolling method” on page 347 on each node. The only difference was that when executing **clmigcheck** on the remaining nodes, the menu no longer appeared. When executed on the second node, *maddi* in our scenario, the output shown in Figure 8-20 is displayed. Afterwards, another message is displayed as shown in Figure 8-22 on page 363.

```
-----[ PowerHA System Mirror Migration Check ]-----  
  
clmigcheck: This is not the first node or last node clmigcheck was run on.  
No further checking is required on this node. You can install the new  
version of PowerHA System Mirror.  
  
Hit <Enter> to continue
```

Figure 8-20 clmigcheck on the second node

When beginning the upgrade on the last node, stop cluster services moving the resource group. When **clmigcheck** is executed on the last node, that is when the CAA cluster is created, as shown in Figure 8-21 on page 361.

```

Saving existing /tmp/clmigcheck/clmigcheck.log to
/tmp/clmigcheck/clmigcheck.log.bak
Verifying clcomd communication, please be patient.

Verifying multicast IP communication, please be patient.

Verifying IPV4 multicast communication with mping.

/usr/sbin/mping version 1.1
Connecting using IPv4.
Listening on 228.168.101.43/4098:

Replies to mping from 9.114.135.71 bytes=32 seqno=1 ttl=1
Replies to mping from 9.114.135.71 bytes=32 seqno=2 ttl=1
Replies to mping from 9.114.135.71 bytes=32 seqno=3 ttl=1
Replies to mping from 9.114.135.71 bytes=32 seqno=4 ttl=1
Replies to mping from 9.114.135.71 bytes=32 seqno=0 ttl=1
clmigcheck: Running
/usr/sbin/rsct/install/bin/ct_caa_set_disabled_for_migration on each node in
the cluster

Creating CAA cluster, please be patient.

```

Figure 8-21 CAA cluster creation

You can also view the newly created CAA cluster via the **lscuster -m** command shown in Example 8-4.

Example 8-4 lscuster output

```

lscuster -m
Calling node query for all nodes...
Node query number of nodes examined: 3

        Node name: jayy.ppd.pok.ibm.com
        Cluster shorthand id for node: 1
        UUID for node: 34f51478-3008-11e2-ad80-daf770007002
        State of node: UP NODE_LOCAL
        Smoothed rtt to node: 0
        Mean Deviation in network rtt to node: 0
        Number of clusters node is a member in: 1
        CLUSTER NAME      SHID      UUID
        XIV_cluster       0         34f14ece-3008-11e2-ad80-daf770007002
        SITE NAME         SHID      UUID
        site_1            1         34e11a90-3008-11e2-ad80-daf770007002

        Points of contact for node: 0

```

```

        Node name: maddi.ppd.pok.ibm.com
        Cluster shorthand id for node: 2
        UUID for node: 34fe718a-3008-11e2-ad80-daf770007002
        State of node: UP
        Smoothed rtt to node: 7

```

```

Mean Deviation in network rtt to node: 3
Number of clusters node is a member in: 1
CLUSTER NAME      SHID      UUID
XIV_cluster       0         34f14ece-3008-11e2-ad80-daf770007002
SITE NAME        SHID      UUID
site_1            1         34e11a90-3008-11e2-ad80-daf770007002

```

Points of contact for node: 5

Interface	State	Protocol	Status
dpcm	DOWN	none	RESTRICTED
en0	UP	IPv4	none
en1	UP	IPv4	none
en2	UP	IPv4	none
en3	UP	IPv4	none

```

Node name: peyton.ppd.pok.ibm.com
Cluster shorthand id for node: 3
UUID for node: 491ba944-3008-11e2-a883-daf770007002
State of node: UP
Smoothed rtt to node: 7
Mean Deviation in network rtt to node: 3
Number of clusters node is a member in: 1
CLUSTER NAME      SHID      UUID
XIV_cluster       0         34f14ece-3008-11e2-ad80-daf770007002
SITE NAME        SHID      UUID
Texas            2         48729ae8-3008-11e2-a883-daf770007002

```

Points of contact for node: 1

Interface	State	Protocol	Status
tcpsock->03	UP	none	none

You can also see that the disks chosen at each site are now part of a caavg_private volume group, as shown in Example 8-5.

Example 8-5 Repository disk on each node

```

jayy#>c1cmd lspv |grep caa
hdisk22      00c9354cd95dbb26          caavg_private  active
hdisk10      00c9354c40265370          caavg_private  active
hdisk10      00c9354c40265370          caavg_private  active

```

After the successful completion of creating the CAA cluster, the menu displays a message to proceed with upgrading PowerHA as shown in Figure 8-22 on page 363.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

```
You can install the new version of PowerHA System Mirror.
```

```
Hit <Enter> to continue
```

Figure 8-22 Clmigcheck message to upgrade PowerHA

When restarting cluster services on the last node, *jayy* in our scenario, the migration completes. Verify that the cluster version number has been updated and that CAA services are running, as shown in Example 8-6.

Example 8-6 Cluster version updated

```
Current state: ST_STABLE
sccsid = "@(#)36 1.135.1.112
src/43haes/usr/sbin/cluster/hacmpd/main.C,hacmp.pe,61haes_r712,1221A_hacmp712
5/22/1"
build = "Nov 5 2012 16:31:36 1242F_hacmp712"
i_local_nodeid 2, i_local_siteid 2, my_handle 3
m1_idx[2]=1 m1_idx[3]=2
There are 0 events on the Ibcast queue
There are 0 events on the RM Ibcast queue
CLversion: 14 <-----
local node vrmf is 7121
cluster fix level is "1"

CAA Cluster Capabilities
CAA Cluster services are active <-----
```

Now that the migration is complete, the cluster should be tested. An overview of our migrated cluster is shown in Figure 8-23 on page 364.

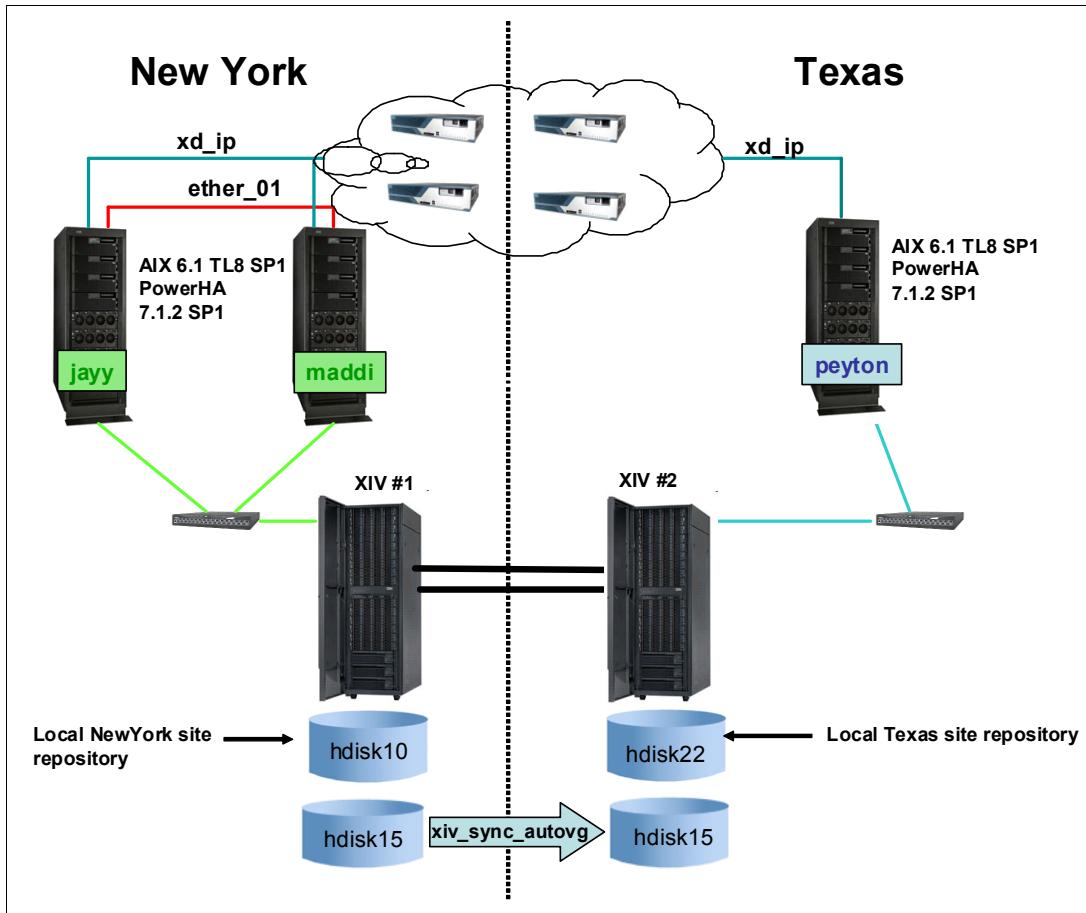


Figure 8-23 Migrated XIV cluster

8.3.2 Snapshot from IBM PowerHA SystemMirror 6.1 Enterprise Edition

This section describes the snapshot migration from PowerHA 6.1 Enterprise Edition.

Test environment overview

This scenario uses a two-node GLVM cluster consisting of one node per site. The primary site has Node 1 and the secondary site Node 2. The cluster is using vSCSI storage for each site. Figure 8-24 on page 365 provides an overview of the software and hardware tested for this snapshot migration.

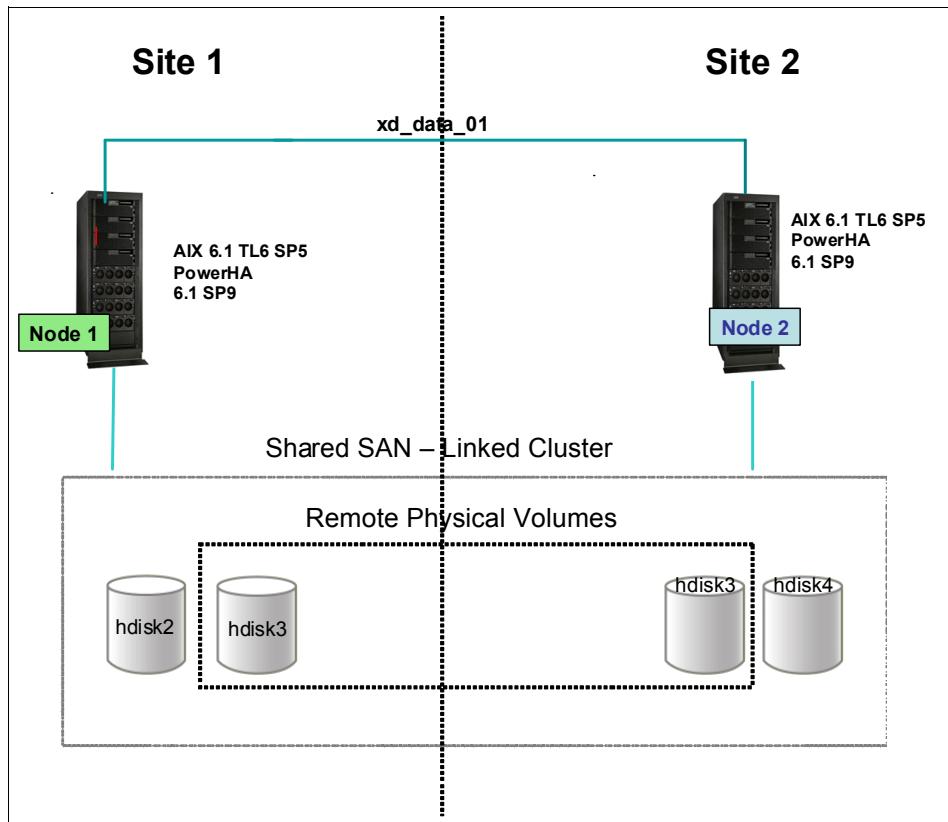


Figure 8-24 Snapshot migration test environment

The topology information for this cluster is shown in Example 8-7.

Example 8-7 cltopinfo on the first node

```

hacmp41:/home/root> #cltopinfo
Cluster Name: xdtest
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
There are 2 node(s) and 1 network(s) defined
NODE node1:
    Network net_XD_data_01
        RG1svc 10.2.10.21
        hacmp41bt 10.1.10.21
NODE node2:
    Network net_XD_data_01
        RG1svc 10.2.10.21
        hacmp42bt 10.1.10.22

Resource Group xdrg1
    Startup Policy  Online On Home Node Only
    Failover Policy  Failover To Next Priority Node In The List
    Fallback Policy  Fallback To Higher Priority Node In The List
    Participating Nodes      node1 node2
    Service IP Label1      RG1svc

```

Total Heartbeats Missed: 0
Cluster Topology Start Time: 11/19/2012 19:42:08

Performing a snapshot migration

Before migration, the cluster status is as shown in Example 8-8.

Example 8-8 clRGinfo output on the first node

Group Name	Group State	Node
xdrg1	ONLINE SECONDARY	node1@site1
	ONLINE	node2@site2

Snapshot creation

Before migrating the cluster, it is important to first take a snapshot of the cluster configuration. It is advisable, once the snapshot is created, to also back it up to a different location in case of problems.

Steps to create a snapshot:

1. **smitty hacmp → Extended Configuration → Snapshot Configuration**

You should then see the Snapshot menu shown in Figure 8-25.

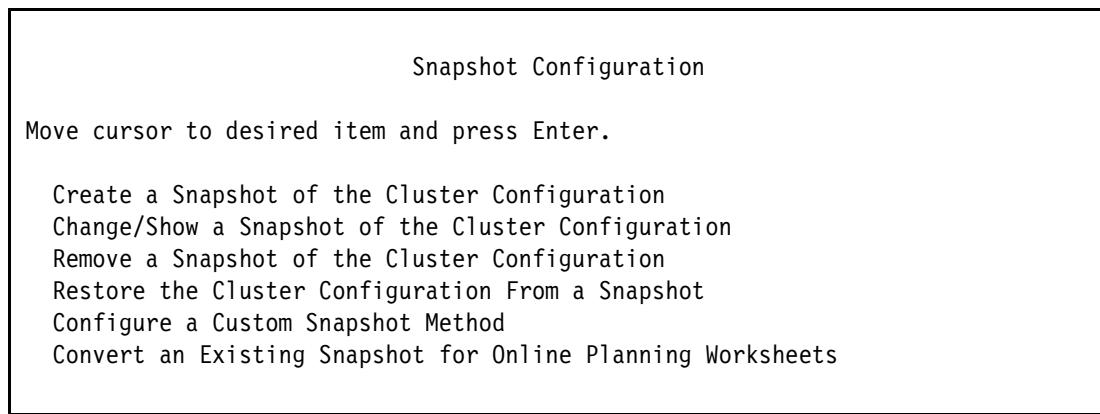


Figure 8-25 SMIT snapshot menu

To proceed, select the first option, “Create a Snapshot of the Cluster Configuration”. The next panel appears, as shown in Figure 8-26 on page 367.

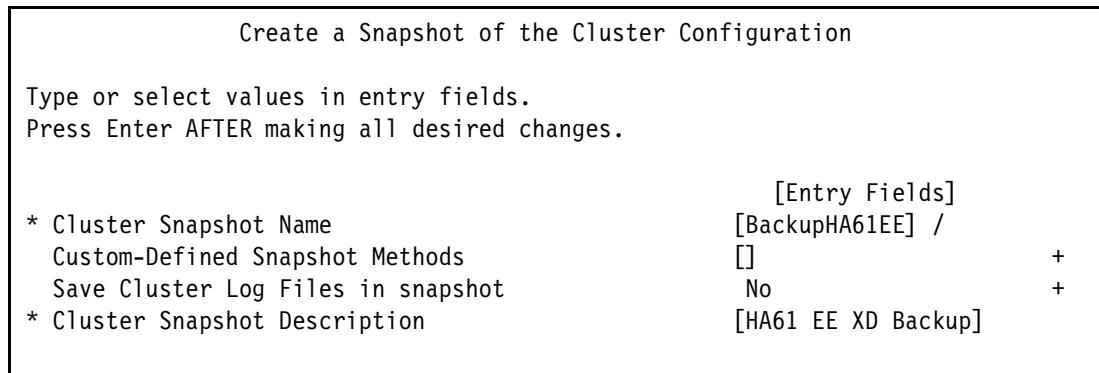


Figure 8-26 Creating a cluster snapshot in SMIT

When you press *Enter*, it should come back OK with:

```
clsnapshot: Succeeded creating Cluster Snapshot: BackupHA61EE
```

2. Back up snapshot files to another location.

The snapshot file is stored in /usr/es/sbin/cluster/snapshots. We advise that you copy your snapshot to another location as backup.

Upgrade AIX

IBM PowerHA SystemMirror 7.1.2 Enterprise Edition requires AIX to be at least 6.1 TL8 SP1. This must be done before you can proceed with PowerHA migration to 7.1.2.

The method of performing the AIX upgrade depends on whether you are performing both AIX and PowerHA in the same maintenance window. If so, then you can perform AIX upgrade with the cluster offline.

If you are performing this upgrade at an earlier maintenance slot in preparation for the PowerHA migration, then you can choose to perform a rolling AIX upgrade by moving resources from the active site to the standby site in order to perform the AIX upgrade. See 8.1.4, "Rolling method" on page 347 for additional details.

Stop the PowerHA cluster

Before you can proceed with the AIX upgrade, make sure cluster services are stopped on both nodes. Execute **smitty clstop** and select both nodes in the cluster, as shown in Figure 8-27.

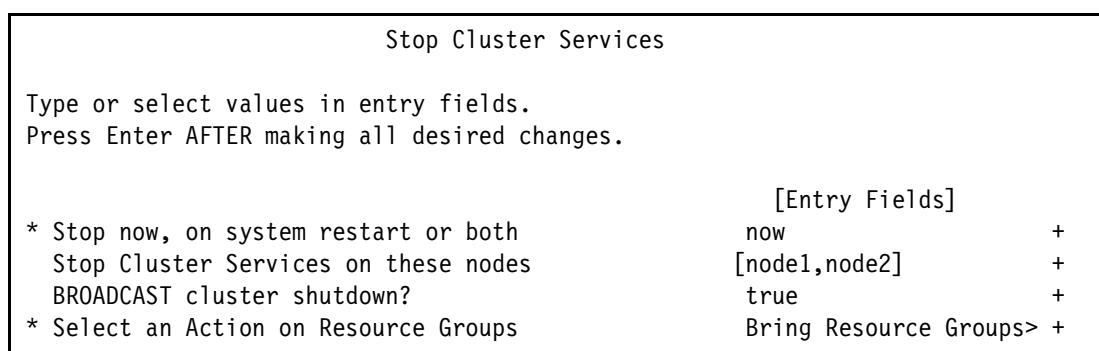


Figure 8-27 Smitty clstop menu

Next verify that cluster services have stopped on each node. As shown in Example 8-9, you need to make sure that the cluster state has gone to ST_INIT.

Attention: Though the OK appears quickly from SMIT, it may take a while for the cluster services to completely stop. You may need to wait and rerun this a few times until it shows ST_INIT.

Example 8-9 Cluster status

```
hacmp41:/home/root> #lssrc -ls clstrmgrES
Current state: ST_INIT
sccsid = "@(#)" 1.135.6.1 src/43haes/usr/sbin/cluster/hacmprd/main.C,
hacmp.pe, 53haes_r610, 1135G_hacmp610 11/30/11 08:50:54"
```

Upgrade AIX, as needed. Once AIX has been updated, a reboot of the node is required.

Clmigcheck

Before the migration to PowerHA 7.12 Enterprise Edition can be performed, a program called **clmigcheck** needs to be executed to ensure that the cluster configuration is supported for migration. For more information, see 8.2, “Clmigcheck explained” on page 349.

Check rhosts

When you have completed the AIX upgrade on both nodes, make sure that they are both rebooted. Verify that **clcmd** is running on both nodes after rebooting with the **lssrc -s clcmd** command shown in Figure 8-8 on page 355.

Ensure that /etc/cluster/rhosts is configured correctly with the IP address or hostname of the nodes on each site, as shown in Example 8-10.

Example 8-10 rhosts configuration

```
hacmp41:/home/root> #cat /etc/cluster/rhosts
9.175.210.77
9.175.210.78
```

After completing the editing of the rhosts configuration file, it is necessary to refresh **clcmd** by executing the command **refresh -s clcmd**.

Running clmigcheck

When you first run **clmigcheck**, it checks for the presence of **clmigcheck.txt**. This file is generated after **clmigcheck** has been run and contains the information necessary for migration. On the first run the file should not exist. If for some reason you have run this before or if this was previously a PowerHA 7.1 cluster, then you may get the message shown in Example 8-11.

Example 8-11 Clmigcheck error if previously run

```
-----[ PowerHA System Mirror Migration Check ]-----
```

clmigcheck: This is not the first node or last node clmigcheck was run on.
No further checking is required on this node. You can install the new
version of PowerHA System Mirror.

Hit <Enter> to continue

To resolve this, check in /var/clmigcheck/ for clmigcheck.txt and delete the file. Then rerun **clmigcheck**.

When you run **clmigcheck**, the menu shown in Example 8-12 is displayed.

Example 8-12 The clmigcheck menu

-----[PowerHA System Mirror Migration Check]-----

Please select one of the following options:

- 1 = Check ODM configuration.
- 2 = Check snapshot configuration.
- 3 = Enter repository disk and multicast IP addresses.

Select one of the above, "x" to exit or "h" for help:

Since this is a snapshot migration, you need to select *option 2*. This checks the previously created snapshot file to ensure that it is compliant for migration, as shown in Example 8-13.

Example 8-13 Option 2 checking the snapshot configuration

-----[PowerHA System Mirror Migration Check]-----

h = help

Enter snapshot name (in /usr/es/sbin/cluster/snapshots): BackupHA61EE

Example 8-14 shows that no errors were encountered verifying the snapshot.

Example 8-14 Clmigcheck correct check snapshot output

-----[PowerHA System Mirror Migration Check]-----

The ODM has no unsupported elements.

Hit <Enter> to continue

Selecting *option 1* for a *stretched* cluster is shown in Example 8-15. You then get the panel to choose the physical disk that may be used for the CAA repository disk, as shown in Example 8-16 on page 370. In our case, this disk is a single disk that is shared between the sites.

Example 8-15 Stretched or linked cluster

-----[PowerHA System Mirror Migration Check]-----

Your multi-site cluster can be based on a single AIX CAA cluster (a Stretched cluster) or a linked AIX CAA cluster (a Linked cluster).

Select the type of multi-site cluster you want:

```
1      = STRETCHED  
2      = LINKED
```

Select one of the above or "h" for help or "x" to exit:

Important: For information on stretched and linked clusters refer to Part 2, “Campus style disaster recovery (stretched clusters)” on page 57, and Part 3, “Extended disaster recovery (linked clusters)” on page 193. Most PowerHA Enterprise Edition clusters would be linked as they are on remote sites. However, in our example we chose a stretched cluster in order to test this type of migration and also because our PowerHA Enterprise Edition test cluster was a local cluster with shared storage, so this was the obvious choice to make in this situation.

Example 8-16 Choosing the repository disk

-----[PowerHA System Mirror Migration Check]-----

Select the disk to use for the repository

```
1      = 00ccfe74ec0d0199(hdisk4)
```

Select one of the above or "h" for help or "x" to exit:

Next choose a multicast address as shown in Example 8-17. You can press Enter and one is assigned for you based on your IP address. For a stretched cluster only choose one. If using a linked cluster, you are prompted to do this for both sites.

Example 8-17 Choosing a multicast address

-----[PowerHA System Mirror Migration Check]-----

PowerHA System Mirror uses multicast address for internal cluster communication and monitoring. These must be in the multicast range, 224.0.0.0 - 239.255.255.255.

If you make a NULL entry, AIX will generate an appropriate address for you. You should only specify an address if you have an explicit reason to do so, but are cautioned that this address cannot be changed once the configuration is activated (i.e. migration is complete).

h = help

Enter the multicast IP address to use for site site1:

Next exit the **c1migcheck** menu. You are now ready to migrate to PowerHA SystemMirror 7.1.2 Enterprise Edition.

Note: **C1migcheck** only needs to be run on the first site.

Clmigcheck creates a special file called `clmigcheck.txt`, which is located in `/var/clmigcheck`. When you exit **clmigcheck**, this file is created on the local node and copied to remote site nodes. The file contains information necessary to perform the migration. Example 8-18 shows the contents of `clmigcheck.txt`.

Example 8-18 clmigcheck.txt contents

```
CLUSTER_TYPE:STRETCHED
CLUSTER_REPOSITORY_DISK:00ccfe74ec0d0199
CLUSTER_MULTICAST:NULL
```

Installing PowerHA SystemMirror 7.1.2 Enterprise Edition

You are now ready to perform the upgrade to PowerHA 7.1.2 Enterprise Edition. The steps are as follows:

1. Stop cluster services.

If not already stopped, stop cluster services on both sites using **smitty clstop**.

2. Uninstall PowerHA 6.1 Enterprise Edition.

Since this is a snapshot migration, now uninstall PowerHA 6.1 Enterprise Edition from both sites. Since AIX has already been upgraded, a reboot is not required.

Tip: Only choose to uninstall the `cluster.x` filesets. Do not select the `bos.cluster` fileset. This fileset is part of CAA and not of PowerHA 6.1 Enterprise Edition. Ensure that all cluster services are stopped before attempting to uninstall.

3. Install PowerHA 7.1.2 Enterprise Edition.

Next install PowerHA 7.1.2 Enterprise Edition on both nodes. Ensure that you choose the relevant filesets for your XD cluster. In our case, because we were using GLVM, we chose `cluster.xd.glvm`, and also `glvm.rpv` to ensure that we had the right support installed in addition to the base filesets.

4. Convert the snapshot.

Once the install has been performed on both nodes and updated to PowerHA 7.1.2 Enterprise Edition SP1, you next need to convert your snapshot. To do this, run:

```
/usr/es/sbin/cluster/conversion/clconvert_snapshot
```

Example 8-19 shows the syntax of the previous command.

Example 8-19 Syntax for clconvert_snapshot

```
hacmp41:/usr/es/sbin/cluster/conversion> #./clconvert_snapshot
```

```
Usage: ./clconvert_snapshot -v [release] [-s [snap_file]]
```

```
Options:
```

```
    -v [version]    version being migrated from
    -s [snap_file]  snapshot file
```

```
-----Warning-----
```

```
If you do not know your previous
version DO NOT run this command.
```

```
ERROR: For details please look in /tmp/clconvert.log
```

There is also a man page for this command that explains it in more detail. The tool can convert from the following releases:

- PowerHA Extended Distance version 5.5
- PowerHA SystemMirror 6.1 Enterprise Edition

Attention: Though the man page shows PowerHA version 7, there is no equivalent Enterprise Edition levels available prior to version 7.1.2.

Example 8-20 shows the conversion utility running.

Example 8-20 Running clconvert_snapshot

```
#./clconvert_snapshot -v 6.1.0 -s BackupHA61EE
```

```
Extracting ODMs from snapshot file... done.  
Converting extracted ODMs... done.  
Rebuilding snapshot file... done.  
hacmp21:/usr/es/sbin/cluster/conversion> #
```

5. Restore the snapshot configuration

Next we need to restore our converted snapshot. The clconvert_snapshot overwrites the original snapshot file with the converted one.

smitty sysmirror → Cluster Nodes and Networks → Manage the Cluster → Snapshot Configuration

This takes you to the snapshot configuration SMIT menu shown in Example 8-21.

Example 8-21 PowerHA 7.1.2 Enterprise Edition snapshot configuration menu

Snapshot Configuration

Move cursor to desired item and press Enter.

```
Create a Snapshot of the Cluster Configuration  
Change/Show a Snapshot of the Cluster Configuration  
Remove a Snapshot of the Cluster Configuration  
Restore the Cluster Configuration from a Snapshot  
Configure Custom Snapshot Method
```

Select Restore the Cluster Configuration from a Snapshot as shown in Example 8-22.

Example 8-22 Restoring the cluster snapshot

Restore the Cluster Configuration from a Snapshot

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

Cluster Snapshot Name	[Entry Fields]
Cluster Snapshot Description	BackupHA61EE
Un/Configure Cluster Resources?	HA61 XD snapshot
Force apply if verify fails?	[Yes] + [No] +

While running the restore, we encountered a problem. The error encountered is shown in Example 8-23 on page 373.

Example 8-23 Snapshot restore error

ERROR: unable to verify outbound clcomd communication from the local node, "hacmp21", to node "hacmp21".

ERROR: Internode Communication check failed, check clcomd.log file for more information.

Tip: PowerHA 7.12 Enterprise Edition uses **clcomd** instead of **clcomdES**. To resolve this problem, you need to ensure that both nodes of the cluster sites are rebooted before you attempt to restore your snapshot configuration. The restarting of the nodes also starts **clcomd** up correctly.

You can see a successful snapshot restored in Example 8-24.

Example 8-24 Successful snapshot restore

COMMAND STATUS

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

[MORE...81]

Adding any necessary PowerHA SystemMirror entries to /etc/inittab and /etc/rc.net for IPAT on node hacmp22.

Verification has completed normally.

== INFO >>

Invoked in the context of SNAPSHOT, GENXD Configuration will not perform any operation. User need to run verify and sync after snapshot restoration to make genxd configuration in sync.

clsnapshot: Succeeded applying Cluster Snapshot: HA61XD1.

[BOTTOM]

6. Testing the migrated cluster

Once migrated, start the cluster to check for any errors. The cluster should come up online, and you should also see a shared repository disk online on both nodes. This is because you are using a *stretched* cluster configuration.

8.3.3 Offline migration from PowerHA 6.1 Enterprise Edition with IPv6 configuration

This section describes the offline migration from PowerHA 6.1 Enterprise Edition with the IPv6 configuration.

Test environment overview

This scenario covers migrating a four-node cluster that has been configured with both IPv6 and IPv4. This is also known as a dual stack configuration. The cluster name is

ipv6mig_cluster. It consists of two nodes, glvma1ip6 and glvma2ip6 in siteA, and two nodes, glvmb1ip6 and glvmb2ip6, in siteB. The disks are replicated through synchronous GLVM. The AIX level is AIX 7.1 TL2 SP1, which does not require an update during the PowerHA migration. Figure 8-28 shows an overview of the cluster configuration.

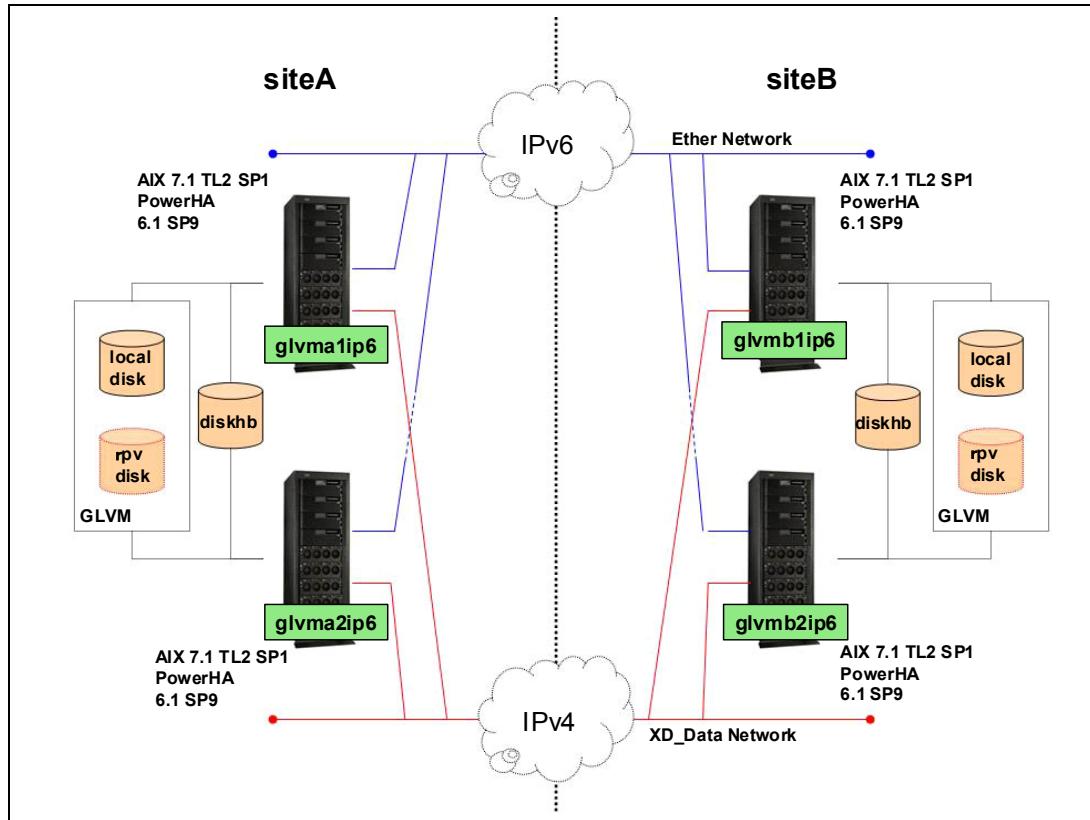


Figure 8-28 IPv6 IPv4 dual stack configuration test environment

We configured two networks, an XD_data and an ether network. The XD_data network is connected with IPv4 addresses, and the ether network is connected with IPv6 addresses. Between sites we configured an IPv4 and IPv6 router so that different IP subnet addresses can be configured. Within each site a dataless disk, hdisk2, was configured for a diskhb network. We planned to convert this disk into a CAA repository disk later in “Execute clmigcheck on one node” on page 377. We also had a single resource group, glvm_ip6. Example 8-25 shows the cluster topology with the **cltopinfo** command.

Example 8-25 IPv6 configuration topology

```
# /usr/es/sbin/cluster/utilities/cltopinfo
Cluster Name: ipv6mig_cluster
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
There are 4 node(s) and 4 network(s) defined
NODE glvma1ip6:
    Network net_XD_data_ip4
        glvma1 192.168.100.55
    Network net_diskhb_siteA
        glvma1_hdsk2 /dev/hdisk2
    Network net_diskhb_siteB
```

```

Network net_ether_01
    glvmbsrv      2000:cccc::b0a:643d
    glvmasrv      2000:bbbb::d0a8:643c
    glvma1ip6     2000::c0a8:6437

NODE glvma2ip6:
    Network net_XD_data_ip4
        glvma2  192.168.100.56
    Network net_diskhb_siteA
        glvma2_hdisk2 /dev/hdisk2
    Network net_diskhb_siteB
    Network net_ether_01
        glvmbsrv      2000:cccc::b0a:643d
        glvmasrv      2000:bbbb::d0a8:643c
        glvma2ip6     2000::c0a8:6438

NODE glvmb1ip6:
    Network net_XD_data_ip4
        glvmb1  10.10.100.57
    Network net_diskhb_siteA
    Network net_diskhb_siteB
        glvmb1_hdisk2 /dev/hdisk2
    Network net_ether_01
        glvmbsrv      2000:cccc::b0a:643d
        glvmasrv      2000:bbbb::d0a8:643c
        glvmb1ip6     2000:aaaa::a0a:6439

NODE glvmb2ip6:
    Network net_XD_data_ip4
        glvmb2  10.10.100.58
    Network net_diskhb_siteA
    Network net_diskhb_siteB
        glvmb2_hdisk2 /dev/hdisk2
    Network net_ether_01
        glvmbsrv      2000:cccc::b0a:643d
        glvmasrv      2000:bbbb::d0a8:643c
        glvmb2ip6     2000:aaaa::a0a:643a

Resource Group glvm_ip6
    Startup Policy          Online On Home Node Only
    Fallover Policy         Fallover To Next Priority Node In The List
    Fallback Policy         Never Fallback
    Participating Nodes     glvma1ip6 glvma2ip6 glvmb1ip6 glvmb2ip6
    Service IP Label        glvmasrv
    Service IP Label        glvmbsrv

```

Note: Ideally, in a GLVM environment, additional heartbeating through the XD_ip network XD_rs232 should be configured. In our test environment, an ether network was configured instead. This was due to a limitation such that we could not configure an IPv6 service address on a XD_ip network. The cluster verification shows the following warning, but in this environment, the message was ignored.

WARNING: An XD_data network has been defined, but no additional XD heartbeat network is defined. It is strongly recommended that an XD_ip or an XD_rs232 network be configured in order to help prevent cluster partitioning if the XD_data network fails. Cluster partitioning may lead to data corruption for your replicated resources.

Be also aware that XD_rs232 no longer exists in PowerHA SystemMirror 7.1.2 Enterprise Edition.

The resource group contains two site-specific single service labels, glvmasrv for siteA, and glvmbsrv for siteB. For disks, a single volume group, glvmvg, and a GMVG replicated resource called glvmvg were configured. Figure 8-29 shows the `clshowres` command output for an outline of the resource group configuration.

Resource Group Name	glvm_ip6
Participating Node Name(s)	glvma1ip6 glvma2ip6 glvmb1ip6 glvmb2ip6
Startup Policy	Online On Home Node Only
Fallover Policy	Fallover To Next Priority Node In The List
Fallback Policy	Never Fallback
Site Relationship	Prefer Primary Site
Service IP Label	glvmasrv glvmbsrv
Filesystems	ALL
Filesystems Consistency Check	fsck
Filesystems Recovery Method	sequential
Volume Groups	glvmvg
Use forced varyon for volume groups, if necessary	true
GMVG Replicated Resources	glvmvg

Figure 8-29 IPv6 cluster resource group

Note: Tips for configuring a PowerHA with IPv6 environment are discussed further in Appendix A, “Configuring IBM PowerHA SystemMirror with IPv6” on page 479.

Performing an offline migration

We performed the steps shown in 8.1.3, “Offline method” on page 347.

Stop cluster services on all nodes

In this case, we stopped all of the cluster services. This is accomplished by executing `smitty clstop` and entering every node in the cluster in the “Stop Cluster Services on these nodes” field shown in Figure 8-30.

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?	[glvma1ip6,glvma2ip6,glvmb1ip6,glvmb2ip6]	+
* Select an Action on Resource Groups	false	+
	Bring Resource Groups>	+

Figure 8-30 Stopping cluster services on every node

After executing, make sure every node in the cluster is in the ST_INIT state, from the `lssrc -ls clstrmgrES` command output shown in Figure 8-31 on page 377.

```
# lssrc -ls clstrmgrES
Current state: ST_INIT
sccsid = "@(#)36 1.135.6.1 src/43haes/usr/sbin/cluster/hacmpd/main.C,
hacmp.pe, 53haes_r610, 1135G_hacmp610 11/30/11 08:50:54"
```

Figure 8-31 Checking the ST_INIT state

Upgrading AIX

In our scenario, the AIX level was already at the minimum level of PowerHA 7.1.2 Enterprise Edition, so we skipped this section.

Execute `clmigcheck` on one node

Now verify that `clcomd` is running by executing `lssrc -s clcomd` as shown in Figure 8-32.

```
# lssrc -s clcomd
Subsystem      Group          PID      Status
clcomd        caa           6357202    active
```

Figure 8-32 CAA `clcomd` service active

Next edit the CAA-specific communication file, `/etc/cluster/rhosts`. In our environment, the hostname was configured with the IPv6 boot label. We entered the IP addresses as shown in Figure 8-33.

```
# cat /etc/cluster/rhosts
glvma1ip6
glvma2ip6
glvmb1ip6
glvmb2ip6
```

Figure 8-33 `/etc/cluster/rhosts` file contents

After completing the edit, it is necessary to refresh `clcomd` by executing `refresh -s clcomd`. Then execute `clmigcheck` and the menu is displayed as in Figure 8-34.

```
-----[ PowerHA System Mirror Migration Check ]-----
Please select one of the following options:
1      = Check ODM configuration.
2      = Check snapshot configuration.
3      = Enter repository disk and multicast IP addresses.

Select one of the above,"x"to exit or "h" for help:1
```

Figure 8-34 `Clmigcheck` menu

For offline migration choose *option 1* and press Enter. Since this environment has a disk heartbeat network configured, a warning is displayed as shown in Figure 8-35 on page 378.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

CONFIG-WARNING: The configuration contains unsupported hardware: Disk Heartbeat network. The PowerHA network name is net_diskhb_siteA. This will be removed from the configuration during the migration to PowerHA System Mirror 7.1.

Hit <Enter> to continue

Figure 8-35 Clmigcheck disk heartbeat warning

As previously mentioned, communication devices for this disk heartbeat network are dataless. In our scenario, we reused this disk for the CAA repository disk. The enhanced concurrent volume group used for disk heartbeating must be removed prior to **clmigcheck**. This can be achieved with the **exportvg** command. Confirm with **lspv** that the disk is not associated to a volume group, as shown in Example 8-26. This needs to be performed on every node in the cluster.

Example 8-26 Removing the enhanced concurrent volume group used for disk heartbeat

```
# exportvg diskhbA_vg  
# lspv  
hdisk2 00f70c992405114b None
```

Note: Using unclean disks for the CAA repository disk may lead to configuration errors. These configuration errors mainly occur when the disk has CAA contents left on its physical disk. The disk /dev/hdisk2 was never used as a CAA disk, so only executing the **exportvg** command is sufficient for this scenario. For safety, you can additionally clean the disk with the following procedure:

1. Delete the VGDA content with the **dd** command. For example, **dd if=/dev/zero of=/dev/rhdisk2 bs=1m count=100**
2. Re-initialize the PVID by executing **chdev -l hdisk2 -a pv=yes**

Once you have confirmed that the disk heartbeat /dev/hdisk2 has not been associated to any disk, press Enter (Figure 8-35). Since the diskhb network will be removed, there are no unsupported elements, as shown in Figure 8-36.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

The ODM has no unsupported elements.

Hit <Enter> to continue

Figure 8-36 Clmigcheck indicating the ODM has no unsupported elements

After pressing Enter, the panel returns to the **clmigcheck** menu as shown in Figure 8-10 on page 355. Now select 3 and press Enter. A panel similar to Figure 8-37 on page 379 appears.

```
-----[ PowerHA System Mirror Migration Check ]-----  
  
Your multi-site cluster can be based on a single AIX CAA cluster  
(a Stretched cluster) or a linked AIX CAA cluster (a Linked cluster).  
  
Select the type of multi-site cluster you want:  
  
1 = STRETCHED  
2 = LINKED  
  
Select one of the above or "h" for help or "x" to exit:2
```

Figure 8-37 Clmigcheck stretched or linked cluster menu option

Because you are configuring this cluster as a linked cluster, choose *option 2* and press Enter. Once chosen, a panel to choose the repository disk for siteA appears. In our case hdisk2 was previously used for disk heartbeating and now was used for the repository disk. We chose *option 1* and pressed Enter. Refer to Figure 8-38.

```
-----[ PowerHA System Mirror Migration Check ]-----  
  
Select the disk to use for the repository on site siteA  
  
1 = 00f70c992405114b(hdisk2)  
  
Select one of the above or "h" for help or "x" to exit:
```

Figure 8-38 Choosing the repository disk for siteA

Another similar menu appears to choose a repository disk for siteB. We also wanted hdisk2 for the repository disk so we chose *option 1* and pressed Enter. Refer to Figure 8-39.

```
-----[ PowerHA System Mirror Migration Check ]-----  
  
Select the disk to use for the repository on site siteB  
  
1 = 00f6f5d023ec219d(hdisk2)  
  
Select one of the above or "h" for help or "x" to exit:
```

Figure 8-39 Choosing the repository disk for siteB

Once choosing the repository disks for each site is complete, the option for entering a multicast address for a site is displayed in Figure 8-40 on page 380. In our case we entered 224.168.100.55 for siteA.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

PowerHA System Mirror uses multicast address for internal cluster communication and monitoring. These must be in the multicast range, 224.0.0.0 - 239.255.255.255.

If you make a NULL entry, AIX will generate an appropriate address for you. You should only specify an address if you have an explicit reason to do so, but are cautioned that this address cannot be changed once the configuration is activated (i.e. migration is complete).

h = help

Enter the multicast IP address to use for site siteA: **224.168.100.55**

Figure 8-40 Choosing the multicast addresses for siteA

After entering the multicast addresses for siteA, another prompt appears for multicast addresses for siteB. Enter 224.10.100.57 for siteB as shown in Figure 8-41.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

PowerHA System Mirror uses multicast address for internal cluster communication and monitoring. These must be in the multicast range, 224.0.0.0 - 239.255.255.255.

If you make a NULL entry, AIX will generate an appropriate address for you. You should only specify an address if you have an explicit reason to do so, but are cautioned that this address cannot be changed once the configuration is activated (i.e. migration is complete).

h = help

Enter the multicast IP address to use for site siteB: **224.10.100.57**

Figure 8-41 Choosing multicast addresses for siteB

Note: You are not prompted for an IPv6 multicast address. Instead, CAA converts the IPv4 multicast address to a hexadecimal format adding 0xFF05 to the first octet. In our case, the rules were applied as follows:

- ▶ 224.168.100.55 or 0xE0A86437, the IPv6 multicast address will be ff05::e0a8:6437.
- ▶ 224.10.100.57 or 0xE00A6439, the IPv6 multicast address will be ff05::e00a:6439.

After entering the multicast addresses **clmigcheck** returns to the main menu, as shown in Figure 8-42 on page 381. Since we had completed the steps, we selected *x* and pressed Enter to exit the menu. Another panel appears, as shown in Figure 8-17 on page 358. We selected *y* and pressed Enter.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

```
You have requested to exit clmigcheck.
```

```
Do you really want to exit? (y) y
```

```
Note - If you have not completed the input of repository disks and  
multicast IP addresses, you will not be able to install  
PowerHA System Mirror
```

```
Additional details for this session may be found in  
/tmp/clmigcheck/clmigcheck.log.
```

Figure 8-42 Clmigcheck menu exit

We now verified the information entered in the **clmigcheck** menu by viewing the contents of the /var/clmigcheck/clmigcheck.txt file shown in Figure 8-43.

```
# cat /var/clmigcheck/clmigcheck.txt
CLUSTER_TYPE:LINKED
SITE1_REPOSITORY_DISK:siteA:00f70c992405114b
SITE2_REPOSITORY_DISK:siteB:00f6f5d023ec219d
SITE1_MULTICAST:siteA:224.168.100.55
SITE2_MULTICAST:siteB:224.10.100.57
```

Figure 8-43 Clmigcheck.txt file contents

Upgrading PowerHA

To upgrade PowerHA we simply executed **smitty update_all**, choosing the v7.1.2.0 install images and setting *ACCEPT new license agreements?* to yes as shown in Figure 8-44 on page 382.

Important: Always complete a migration using base levels. Then come back and install any service packs later.

Once the upgrade has completed, check the /tmp/c1convert.log file for any errors. The first few lines and the last few lines of our log file are shown in Example 8-27 on page 382.

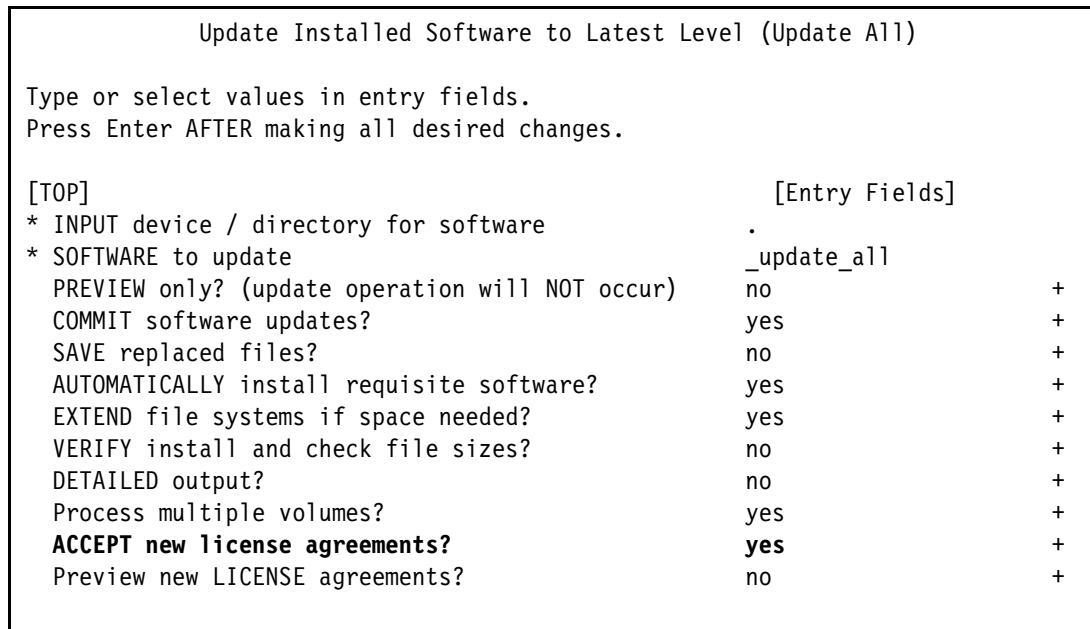


Figure 8-44 *Update_all* of PowerHA

Example 8-27 Clconvert.log

```
----- log file for cl_convert: Wed Nov 21 12:32:40 EST 2012

Command line is:
/usr/es/sbin/cluster/conversion/cl_convert -F -v 6.1

No source product specified.
Assume source and target are same product.
Parameters read in from command line are:
    Source Product is HAES.
    Source Version is 6.1.0.
    Target Product is HAES.
    Target Version is 7.1.2.
    Force Flag is set.

.....
Exiting cl_convert.

Exiting with error code 0. Completed successfully.

----- end of log file for cl_convert: Wed Nov 21 12:33:03 EST 2012
```

Execute clmigcheck and upgrade PowerHA on the remaining nodes

Repeat the steps previously performed on the first node on each remaining node as covered in 8.1.3, “Offline method” on page 347. The only difference is that when executing **clmigcheck** on the remaining nodes the menu no longer appears. When executed on the second node, the output shown in Figure 8-45 on page 383 is displayed.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

```
clmigcheck: This is not the first node or last node clmigcheck was run on.  
No further checking is required on this node. You can install the new  
version of PowerHA System Mirror.
```

```
Hit <Enter> to continue
```

Figure 8-45 Clmigcheck on the second node

When **clmigcheck** is executed on the last node, a prompt appears asking to create the CAA cluster as shown in Figure 8-46. Press Enter.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

```
About to configure a 4 node CAA cluster, this can take up to 3 minutes.
```

```
Hit <Enter> to continue
```

Figure 8-46 CAA cluster creation

After the successful creation of the CAA cluster, you are prompted to update PowerHA for this node as shown in Figure 8-47. Press Enter; it exits out of the menu and returns to the command line prompt.

```
-----[ PowerHA System Mirror Migration Check ]-----
```

```
You can install the new version of PowerHA System Mirror.
```

```
Hit <Enter> to continue
```

Figure 8-47 Prompt to update PowerHA

Complete the update of the last node by executing the steps in “Upgrading PowerHA” on page 381.

Finally confirm that the CAA cluster has been created by executing the **lscuster -i** command. Example 8-28 shows parts of the command output where you can see the device states and multicast addresses configured through the CAA.

Example 8-28 lscuster output

```
# lscuster -i  
Network/Storage Interface Query  
  
Cluster Name: ipv6mig_cluster  
  
Node glvmlip6  
Number of interfaces discovered = 5  
    Interface number 1, en0  
        Interface state = UP  
        Number of regular addresses configured on interface = 1
```

```

IPv4 ADDRESS: 192.168.100.55 broadcast 192.168.100.255 netmask
255.255.255.0
    Number of cluster multicast addresses configured on interface = 1
    IPv4 MULTICAST ADDRESS: 224.168.100.55
Interface number 2, en1
    Interface state = UP
    Number of regular addresses configured on interface = 3
    IPv6 ADDRESS: fe80::7840:c3ff:fe0b:1f03/64
    IPv6 ADDRESS: 2000::c0a8:6437/64
    IPv6 ADDRESS: 2000:bbbb::d0a8:643c/64
    Number of cluster multicast addresses configured on interface = 1
    IPv6 MULTICAST ADDRESS: ff05::e0a8:6437
Interface number 3, dpcom
    Interface state = UP RESTRICTED AIX_CONTROLLED
Interface number 4, tcpsock->08
    Interface state = UP
Interface number 5, tcpsock->07
    Interface state = UP

Node glvmb2ip6
Node UUID = 4523eb52-358b-11e2-a31b-eeaf03910902
Number of interfaces discovered = 5
    Interface number 1, en0
        IPv4 ADDRESS: 10.10.100.58 broadcast 10.10.100.255 netmask 255.255.255.0
        Number of cluster multicast addresses configured on interface = 1
        IPv4 MULTICAST ADDRESS: 224.10.100.57
    Interface number 2, en1
        Interface state = UP
        Number of regular addresses configured on interface = 2
        IPv6 ADDRESS: fe80::ecaf:3ff:fe91:903/64
        IPv6 ADDRESS: 2000:aaaa::a0a:643a/64
        Number of cluster multicast addresses configured on interface = 1
        IPv6 MULTICAST ADDRESS: ff05::e00a:6439
    Interface number 3, dpcom
        Interface state = UP RESTRICTED AIX_CONTROLLED
    Interface number 4, tcpsock->06
        Interface state = UP
    Interface number 5, tcpsock->05
        Interface state = UP

```

Restarting cluster services

Now start the cluster on all nodes. This is accomplished by executing **smitty clstart** and entering every node in the “Start Cluster Services on these nodes” field shown in Figure 8-48 on page 385.

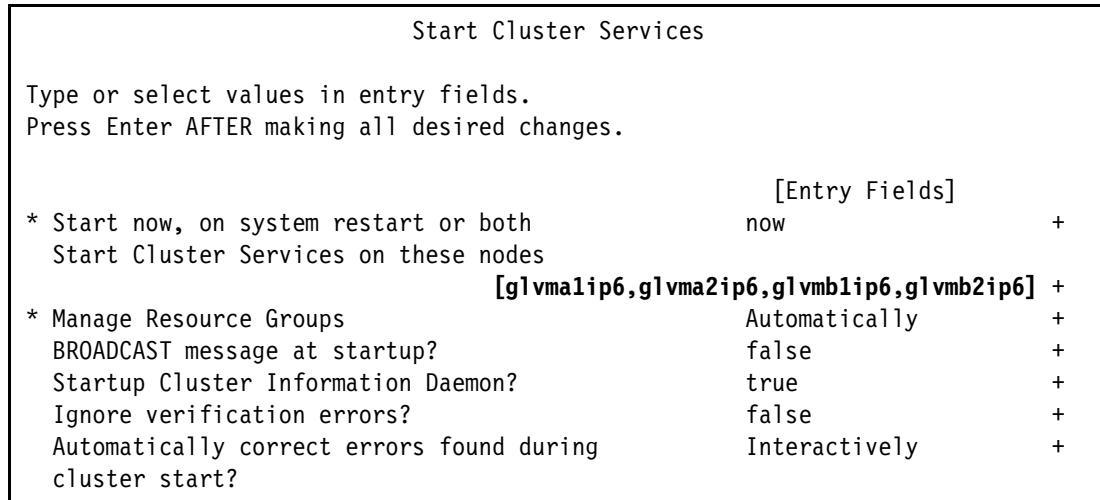


Figure 8-48 Starting the cluster

On the initial startup, it shows that cluster services are not on the same level, as shown in Example 8-29. When all the clusters are in the ST_STABLE state from the **lssrc -ls clstrmgrES** command, the cluster versions are synchronized, as shown in Example 8-30.

Example 8-29 Clistart initial message

---- start ----

Cluster services are running at different levels across the cluster. Verification will not be invoked in this environment.

Example 8-30 Confirming the upgrade

```
# lssrc -ls clstrmgrES
Current state: ST_STABLE
sccsid = "@(#)36 1.135.1.112
src/43haes/usr/sbin/cluster/hacmprd/main.C,hacmp.pe,61haes_r712,1221A_hacmp712
5/22/1"
build = "Nov  5 2012 16:31:36 1242F_hacmp712"
i_local_nodeid 3, i_local_siteid 2, my_handle 4
m1_idx[1]=0    m1_idx[2]=1    m1_idx[3]=2    m1_idx[4]=3
There are 0 events on the Ibcast queue
There are 0 events on the RM Ibcast queue
CLversion: 14 <-----
local node vrmf is 7121
cluster fix level is "1" <-----
```

Now that the migration is complete, the cluster should be tested. Figure 8-49 on page 386 shows the migrated IPv6/IPv4 dual stack cluster ready for testing.

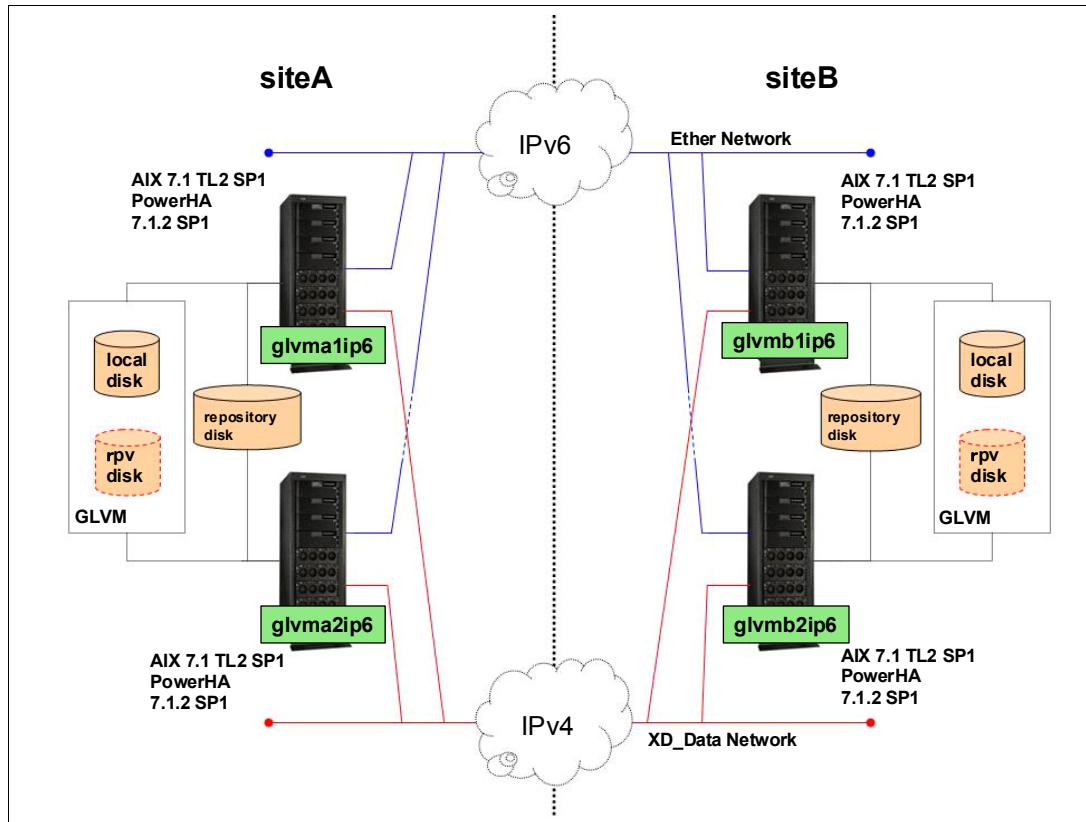


Figure 8-49 Migrated IPv6 IPv4 dual stack cluster

8.4 Solving migration errors

In this section, we cover how to fix common migration problems that may be encountered.

8.4.1 Node name not set to hostname

Starting with PowerHA 7.1.0, The PowerHA node name defaults to the hostname. Though it is not a requirement. The key piece is that the defined *Communication Path* for the node must be the hostname IP address.

During our migration testing we actually ran into this error (Example 8-31) on two separate clusters. However, the cause in each case was not exactly the same.

Example 8-31 Error when checking snapshot

-----[PowerHA System Mirror Migration Check]-----

ERROR: Communications Path for node node1 must be set to hostname

The first time we encountered it on our XIV cluster, we discovered that there were additional boot IP addresses in /etc/hosts that also had the hostname listed as an alias. Once we deleted those aliases on each node, the error was no longer encountered.

The second time we encountered the error on our GLVM cluster, we discovered that neither the node name, nor the communication path resolved to the hostname IP address, as shown in Example 8-32.

Example 8-32 Check node communication path

```
hacmp41:/> #odmget HACMPnode |grep -p COMMUNICATION
HACMPnode:
    name = "node1"
    object = "COMMUNICATION_PATH"
    value = "10.1.10.21"
    node_id = 1
    node_handle = 1
    version = 11

HACMPnode:
    name = "node2"
    object = "COMMUNICATION_PATH"
    value = "10.1.11.21"
    node_id = 2
    node_handle = 2
    version = 11
```

Ultimately, to resolve this challenge, the only thing that needs to be accomplished is to have the node communication path point to the hostname IP address. However, in our scenario we:

1. Changed the node name to match the hostname.
2. Changed the boot IP to be the hostname IP.

Step 1 - Change node name to hostname

Our cluster becomes nodes hacmp41 and hacmp42 instead of node1 and node2.

Go to **smitty hacmp** → **Extended Topology Configuration** → **Configure HACMP Nodes** → **Change/Show a Node in the HACMP Cluster**.

You are presented with the SMIT panel shown in Example 8-33.

Example 8-33 Change the node name in SMIT

Change/Show a Node in the HACMP Cluster

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

* Node Name New Node Name Communication Path to Node	[Entry Fields] node1 [hacmp41] [10.1.10.21] +
--	--

Step 2 - Change the boot address to match the hostname and update topology

Edit your /etc/hosts file and change the boot address hostname to match the nodename. Before you can do this step, you need to remove the resource group and topology from the cluster configuration.

- ▶ Make a note of your resource group configuration, then remove it:
smitty hacmp → Extended Configuration → Extended Resource Configuration → HACMP Extended Resource Group Configuration → Remove a Resource Group
- ▶ Next remove the network. This also removes your network interface configuration. In our case it was XD_DATA:
smitty hacmp → Extended Configuration → Extended Topology Configuration → Configure HACMP Networks → Remove a Network from the HACMP Cluster

Important: You must remove your resource group configuration before you can remove the network.

Edit the /etc/hosts file on both nodes. Then change the boot interface to match the node name. Example 8-34 shows our before and after host table entries

Example 8-34 /etc/hosts before and after changes needed

BEFORE:>

```

127.0.0.1           loopback localhost      # loopback (lo0) name/address
9.175.210.77        hacmp41
9.175.210.78        hacmp42
9.175.211.187       aix2.usceth.farn.uk.ibm.com

#HACMP config
#boot addresses
10.1.10.21          hacmp41bt
10.1.11.21          hacmp41bt2

10.1.10.22          hacmp42bt
10.1.11.22          hacmp42bt2

#Service address
10.2.10.21          RG1svc
10.2.10.22          RG2svc

```

AFTER:>

```

127.0.0.1           loopback localhost      # loopback (lo0) name/address
9.175.210.77        hacmp41pers
9.175.210.78        hacmp42pers
9.175.211.187       aix2.usceth.farn.uk.ibm.com

#HACMP config
#boot addresses
10.1.10.21          hacmp41
10.1.11.21          hacmp41bt2

10.1.10.22          hacmp42
10.1.11.22          hacmp42bt2

#Service address
10.2.10.21          RG1svc
10.2.10.22          RG2svc

```

Once this has been changed on both nodes, you can then go back into SMIT and recreate the XD_DATA network. When re-adding the en1 interfaces if it does not discover the new name,

you need to use the predefined option to manually enter it. Once this has been done, the service address and resource group need to be recreated as per the previous configuration. Rerun **c1migcheck** and verify that the error no longer exists.

8.4.2 Stuck in migration

When migration is completed, you might not progress to the update of the Object Data Manager (ODM) entries until the `node_up` event is run on the last node of the cluster. If you have this problem, start the node to see whether this action completes the migration protocol and updates the version numbers correctly. For PowerHA 7.1.2, the version number must be 14 in the HACMPcluster class. You can verify this number with the **odmget** command as shown in Example 8-35. If the version number is less than 14, you are still stuck in migration. Usually a complete stop, sync and verify, and restart of the cluster completes the migration. If not, contact IBM support.

Example 8-35 odmget to check the version

```
# odmget HACMPcluster|grep version
    cluster_version = 14
```

8.4.3 Non-IP network not deleted after migration completed

Here we provide details about problems with existing non-IP networks that are not removed. This section describes a possible workaround to remove disk heartbeat networks if they were not deleted as part of the migration process.

After the migration, the output of the **cltopinfo** command might still show the disk heartbeat network shown in Example 8-36.

Example 8-36 Cltopinfo showing that diskhb still exists

```
Cluster IP Address: 228.19.51.194
There are 2 node(s) and 2 network(s) defined

NODE jessica:
    Network shawn_dhb_01
        jess_hdisk1      /dev/hdisk1
    Network net_ether_01
        ha_svc 192.168.100.100
        jess_boot1     192.168.100.1
    Network net_ether_02
        jessica 9.19.51.193

NODE jordan:
    Network shawn_dhb_01
        jordan_hdisk1    /dev/hdisk1
    Network net_ether_01
        ha_svc 192.168.100.100
        jordan_boot1   192.168.100.2
    Network net_ether_02
        jordan 9.19.51.194

Resource Group testsvcip
    Startup Policy  Online On Home Node Only
    Fallover Policy Fallover To Next Priority Node In The List
```

Fallback Policy	Never Fallback
Participating Nodes	jessica jordan
Service IP Label	ha_svc

To remove the disk heartbeat network, follow these steps:

1. Stop PowerHA on all cluster nodes.

In most circumstances, you can add or remove networks dynamically. However, this does not work in this case because the migration has technically completed. Any attempt to remove it with services active results in the error shown in Figure 8-50.

COMMAND STATUS		
Command: failed	stdout: yes	stderr: no
Before command completion, additional instructions may appear below.		
cldare: Migration from PowerHA SystemMirror to PowerHA SystemMirror/ES detected.		
A DARE event cannot be run until the migration has completed.		

Figure 8-50 Removing the network failed

2. Remove the network.
 - a. Using SMIT execute **smitty sysmirror** → **Cluster Nodes and Networks** → **Manage Networks and Network Interfaces** → **Networks** → **Remove a Network**.
 - b. On the SMIT panel, similar to the one shown in Figure 8-51 on page 391, select the disk heartbeat network you want to remove.

You might have to repeat these steps if you have more than one disk heartbeat network.

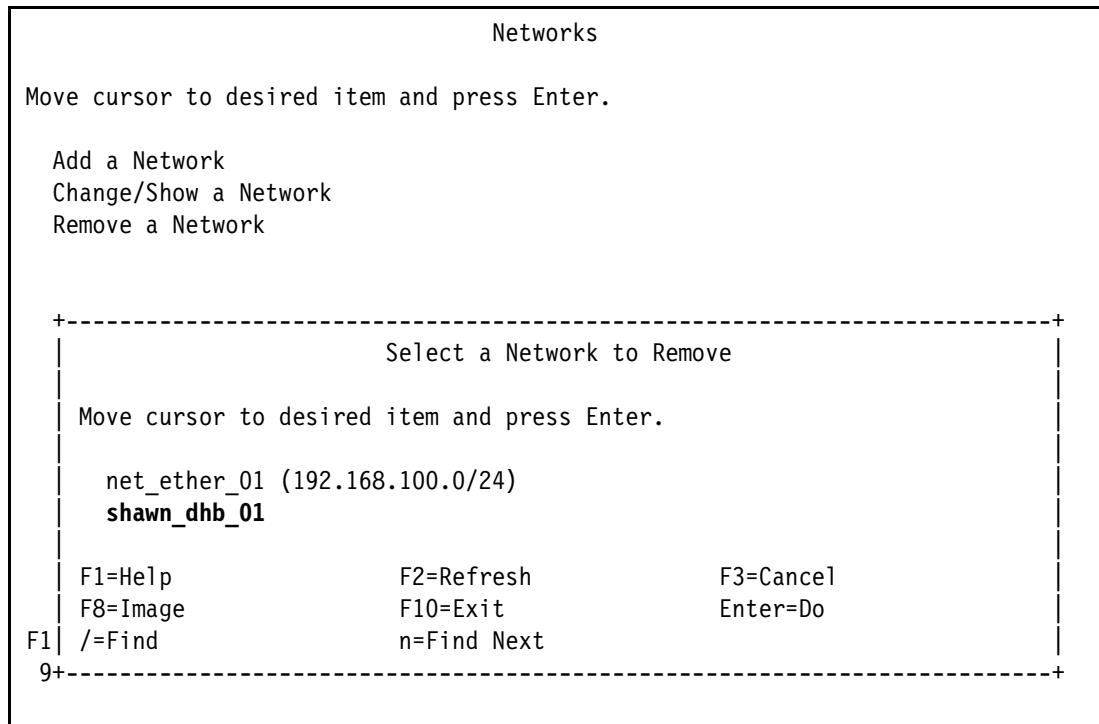


Figure 8-51 Choose diskhb to delete

3. Synchronize your cluster by executing **smitty sysmirror** → **Custom Cluster Configuration** → **Verify and Synchronize Cluster Configuration (Advanced)**.
4. Verify that the disk heartbeat has been removed from **c1topinfo**, as shown in Example 8-37.

Example 8-37 C1topinfo after removing the disk heartbeat manually

Cluster IP Address: 228.19.51.194
There are 2 node(s) and 2 network(s) defined

```

NODE jessica:
    Network net_ether_01
        ha_svc 192.168.100.100
        jess_boot1      192.168.100.1
    Network net_ether_02
        jessica 9.19.51.193

NODE jordan:
    Network net_ether_01
        ha_svc 192.168.100.100
        jordan_boot1      192.168.100.2
    Network net_ether_02
        jordan 9.19.51.194

Resource Group testsvcip
    Startup Policy  Online On Home Node Only
    Fallover Policy Fallover To Next Priority Node In The List
    Fallback Policy Never Fallback
    Participating Nodes      jessica jordan

```

8.4.4 Clodmget not found

When migrating from a pre-6.1 version and executing **clmigcheck** on the last node, you may encounter a clodmget error, as shown in Example 8-38.

Example 8-38 Clodmget error

```
/usr/sbin/clmigcheck[3743]: mk_cluster: line 55:  
/usr/es/sbin/cluster/utilities/clodmget: not found  
/usr/sbin/clmigcheck[3743]: mk_cluster: line 62:  
/usr/es/sbin/cluster/utilities/clodmget: not found  
/usr/sbin/clmigcheck[3743]: mk_cluster: line 88: =: not found  
  
ERROR: Missing cluster name or node name
```

If this error is encountered, simply copy the information from /usr/es/sbin/cluster/utilities from an upgraded PowerHA 7.1.2 Enterprise Edition node and execute **clmigcheck** again.



PowerHA 7.1.2 for IBM Systems Director plug-in enhancements

This chapter covers the following topics:

- ▶ Installing the IBM Systems Director environment
- ▶ Environment configuration
- ▶ Configuring IBM PowerHA SystemMirror 7.1.2 Enterprise Edition using IBM Systems Director
- ▶ Administering a cluster

9.1 Installing the IBM Systems Director environment

This section describes how to deploy IBM Systems Director.

9.1.1 Installing IBM Systems Director

In this section, we explain the IBM Systems Director setup and the IBM PowerHA SystemMirror plug-in installation. We also describe how to manage PowerHA SystemMirror capabilities by using a web browser. IBM Systems Director 6.3 is required to install PowerHA SystemMirror 7.1.2 for IBM Systems Director plug-in.

Download information: You can download IBM Systems Director 6.3 from:

<http://www-03.ibm.com/systems/software/director/downloads/mgmtservers.html>

Extract the file SysDir6_3_Server_AIX_a.tar.gz, and use the checkds.sh script to find the file system and page size requirement as shown in Example 9-1. Configure the server according to the requirement shown by the script before installation.

Example 9-1 checkds.sh - report example

```
# ./checkds.sh
Java: /dirinst/Director/checkds/jvm/aix/bin/java
Starting IBM Systems Director Pre-Installation Utility...
Finished analysing system
Creating reports...
Install Readiness Text report being written to
/tmp/checkds/reports/checkDS_Text_11232012_010452.txt
Install Readiness Error Text report being written to
/tmp/checkds/reports/checkDS_Error.txt
Install Readiness Detailed HTML report being written to
/tmp/checkds/reports/checkDS_Detailed_11232012_010453.html
Install Readiness Summary HTML report being written to
/tmp/checkds/reports/checkDS_Summary_11232012_010454.html
Unable to launch the default browser, please view the text or summary HTML report
manually.
```

Overall Report Return Code: 0

Example 9-2 shows the installation procedure.

Example 9-2 dirinstall.server - IBM Systems Director installation

```
# ./dirinstall.server
=====
Start of product installation on newdir
=====
Variables will be used during the installation:
  PRE_INSTALL_CHECKS : 1
  PRE_INSTALL_WARN_ABORT : 0
  PortNumber : 8421
  SecurePortNumber : 8422
  AGENT_MANAGER_PORT : 20000
```

```

MIGRATE_DATA : 1
UPDATES_PATH : /dirinst/Director/packages/updates
-Managed DB2 is supported and its prerequisites are met.
DB_INST_TYPE : 1
DB_DATAPATH : /home/dirinst1
DB_PWD : default.
DB_SERVER : localhost
DB_PORT : default
=====
Warning:
-There is no valid definition of SNMPv1_COMMUNITY in /etc/snmpdv3.conf.
-dirsnmpd can not be enabled.
=====
Attempting to install Managed DB2...done
Attempting to install sysmgt.cimserver.pegasus Director.install.msg ....done
Attempting to install sysmgt.cim.providers...done
Attempting to install sysmgt.cim.smisproviders.hba_hdr...done
Attempting to install sysmgt.cim.smisproviders.vblkssrv...done
Attempting to install sysmgt.cim.smisproviders.hhr...done
Attempting to install DirectorPlatformAgent...done
Attempting to install tivoli.tivguid cas.agent ....done
Attempting to install DirectorCommonAgent...done
=====
Attempting to install DirectorServer
=====
          Pre-installation Verification...
=====
Verifying selections...done
Verifying requisites...done
Results...

SUCCESSES
-----
Filesets listed in this section passed pre-installation verification
and will be installed.

Selected Filesets
-----
DirectorServer 6.3.0.0                      # All required files of Direct...
<< End of Success Section >>

=====
          BUILDDATE Verification ...
=====
Verifying build dates...done
FILESET STATISTICS
-----
1 Selected to be installed, of which:
  1 Passed pre-installation verification
-----
1 Total to be installed
=====
          Installing Software...

```

```

+-----+
installp: APPLYING software for:
    DirectorServer 6.3.0.0

. . . . . << Copyright notice for DirectorServer >> . . . . .
Licensed Materials - Property of IBM

5765-DRP
    Copyright International Business Machines Corp. 2010, 2011.

All rights reserved.
US Government Users Restricted Rights - Use, duplication or disclosure
restricted by GSA ADP Schedule Contract with IBM Corp.
. . . . << End of copyright notice for DirectorServer >>. . .

Restoring files, please wait.
1030 files restored.
Starting server runtime....done
Attempting to install features.....done
Stopping the server runtime...done
Configuring database.....done

Openssh is not installed on the system. It is required if you plan to deploy
agents with the

"Agent Installation Wizard".

Finished processing all filesets. (Total time: 23 mins 34 secs).

+-----+
Summaries:
+-----+

Installation Summary
-----
Name          Level      Part   Event     Result
-----
DirectorServer 6.3.0.0  USR    APPLY    SUCCESS
DirectorServer 6.3.0.0  ROOT   APPLY    SUCCESS
Installation of IBM Systems Director Server completed successfully.
This installation log file can be found in /var/log/dirinst.log.
You must configure the agent manager prior to starting the server.
To configure the agent manager, run
    /opt/ibm/director/bin/configAgtMgr.sh
To start the server manually, run
    /opt/ibm/director/bin/smstart
-----
```

After installing the IBM Systems Director server, you need to configure the agent manager. Example 9-3 on page 397 shows the agent manager configuration, and illustrates how to start IBM Systems Director server and status verification.

Example 9-3 Agent manager configuration

```
# /opt/ibm/director/bin/configAgtMgr.sh
Enter 1 to use the Agent Manager installed with this server (recommended)
Enter 0 to use an existing Agent Manager (advanced) : 1
Enter Resource Manager username : usmi
Enter Resource Manager password :
Re-Enter Resource Manager password :
Enter Agent Registration password :
Re-Enter Agent Registration password :
[Add] [Element]: AgentManagerUserID [Value]: usmi
[Add] [Element]: AgentManagerPassword [Value]:

{aes:3C5SnKQL63SjkEy44Gs+vHF6nQzC+Dl1NzNvSiAzzk=}It1CiZmQR0bdj9vkt02BqA==
[Add] [Element]: ManagerRegistrationPassword [Value]:


{aes:3C5SnKQL63SjkEy44Gs+vHF6nQzC+Dl1NzNvSiAzzk=}It1CiZmQR0bdj9vkt02BqA==

DataSourceConfig.sh=0
DataStoreInstall.sh=0
GenerateCertificates.sh=0
EncryptAMProps.sh=0
WebConfig.sh=0
usmi-cas-setup.sh=0

# /opt/ibm/director/bin/smstart
Starting IBM Director...
The starting process may take a while. Please use smstatus to check if the server
is active.

# smstatus -r
Starting
Active
```

Configuration consideration: While configuring the agent manager, the resource manager username, password and agent registration password were given *usmi*.

9.1.2 Installing IBM Systems Director plug-in for PowerHA SystemMirror

Now you can login into the IBM Systems Director. You need to install the IBM PowerHA SystemMirror plug-in to get on the IBM Systems Director web access. Example 9-4 shows the PowerHA SystemMirror plug-in installation details.

Downloading the plug-in: You can download the PowerHA SystemMirror 7.1.2 IBM Systems Director plug-in from:

<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>

Example 9-4 PowerHA SystemMirror plug-in installation

```
# ./IBMSystemsDirector_PowerHA_sysmirror_Setup.bin
Preparing to install...
Extracting the installation resources from the installer archive...
Configuring the installer for this system's environment...

Launching installer...
```

Graphical installers are not supported by the VM. The console mode will be used instead...

=====

Choose Locale...

- 1- Deutsch
- >2- English
- 3- Español
- 4- Français
- 5- Italiano
- 6- Português (Brasil)

CHOOSE LOCALE BY NUMBER: 2

=====

IBM PowerHA SystemMirror (created with InstallAnywhere)

Preparing CONSOLE Mode Installation...

=====

Introduction

InstallAnywhere will guide you through the installation of IBM PowerHA SystemMirror.

It is strongly recommended that you quit all programs before continuing with this installation.

Respond to each prompt to proceed to the next step in the installation. If you want to change something on a previous step, type 'back'.

You may cancel this installation at any time by typing 'quit'.

PRESS <ENTER> TO CONTINUE:

=====

International Program License Agreement

Part 1 - General Terms

BY DOWNLOADING, INSTALLING, COPYING, ACCESSING, CLICKING ON AN

"ACCEPT" BUTTON, OR OTHERWISE USING THE PROGRAM, LICENSEE AGREES TO THE TERMS OF THIS AGREEMENT. IF YOU ARE ACCEPTING THESE TERMS ON BEHALF OF LICENSEE, YOU REPRESENT AND WARRANT THAT YOU HAVE FULL AUTHORITY TO BIND LICENSEE TO THESE TERMS. IF YOU DO NOT AGREE TO THESE TERMS,

- DO NOT DOWNLOAD, INSTALL, COPY, ACCESS, CLICK ON AN "ACCEPT" BUTTON, OR USE THE PROGRAM; AND
- PROMPTLY RETURN THE UNUSED MEDIA, DOCUMENTATION, AND PROOF OF ENTITLEMENT TO THE PARTY FROM WHOM IT WAS OBTAINED FOR A REFUND OF THE AMOUNT PAID. IF THE PROGRAM WAS DOWNLOADED, DESTROY ALL COPIES OF THE PROGRAM.

Press Enter to continue viewing the license agreement, or enter "1" to accept the agreement, "2" to decline it, "3" to print it, "4" to read non-IBM terms, or "99" to go back to the previous screen.: 1

=====

IBM Director Start

IBM Systems Director is currently running. Do you want IBM Systems Director to be restarted automatically after
IBM PowerHA SystemMirror is installed? Although it does not need to be stopped in order to install IBM PowerHA SystemMirror, it will need to be restarted before IBM PowerHA SystemMirror functions are available.

- 1- Yes
- >2- No

ENTER THE NUMBER FOR YOUR CHOICE, OR PRESS <ENTER> TO ACCEPT THE DEFAULT:: 1

=====

Installing...

[=====|=====|=====|=====]
[-----|-----|-----]

smstatus -r
Starting
Active

In our case, we have now have successfully installed both IBM Systems Director server and IBM PowerHA SystemMirror plug-in and restarted IBM Systems Director. Figure 9-1 on

page 400 displays the web access of the IBM Systems Director and PowerHA SystemMirror plug-in.

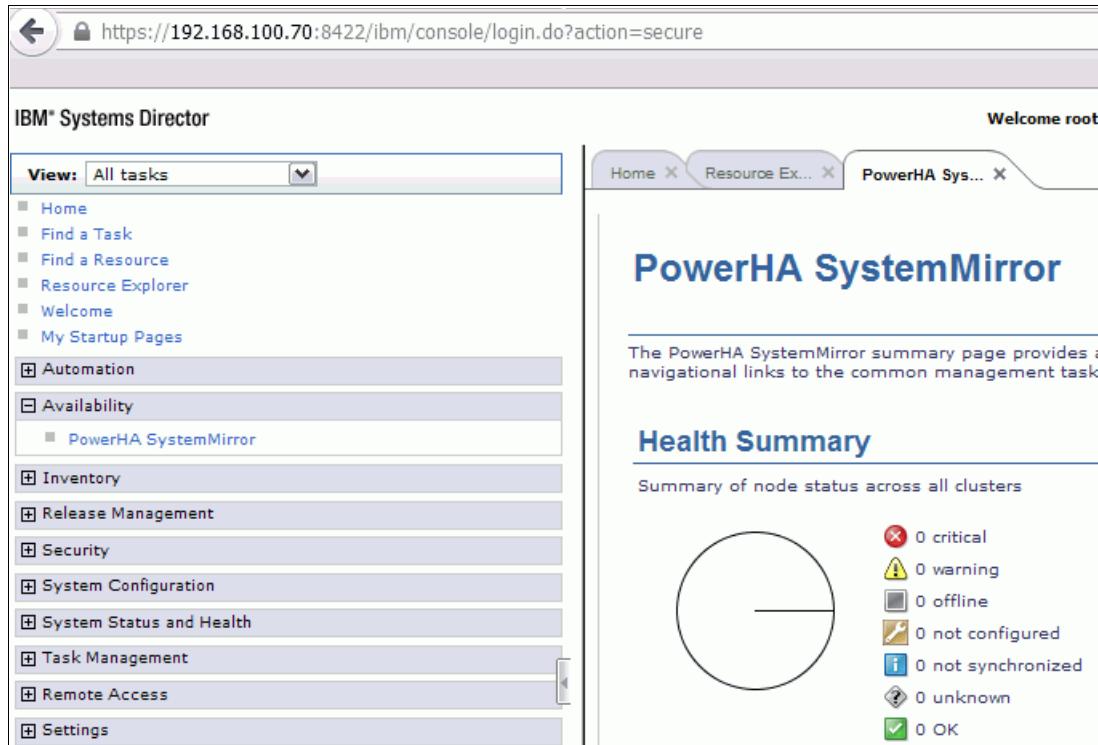


Figure 9-1 PowerHA SystemMirror plug-in on IBM Systems Director

Version 7.1.2.1 (or higher): To obtain the features and latest corrections of the plug-in and agent, download and install version 7.1.2.1 (or higher).

9.1.3 Installing PowerHA SystemMirror plug-in fixes

Navigate to Fix Central at the following address and select to download the fix for IBM Systems Director:

<http://www-933.ibm.com/support/fixcentral/>

Then choose the correct IBM Systems Director server for your environment. In our case, we chose AIX, as shown in Figure 9-2 on page 401. After choosing the options, click **Continue**.

Fix Central

Fix Central provides fixes and updates for your system's software, hardware, and operating system.

For additional information, click on the following link.

[Getting started with Fix Central](#)

[Select product](#) [Find product](#)

Select the product below.

When using the keyboard to navigate the page, use the **Alt** and **down arrow** keys to navigate the selection lists.

Product Group

IBM Systems Director ▾

Product

IBM Systems Director ▾

Installed Version

6.3 ▾

Platform

AIX ▾

Continue

Figure 9-2 Choose IBM Systems Director updates

On the next window, choose the fix for the PowerHA SystemMirror plug-in. Select the fix as shown in Figure 9-3 on page 402 and then click **Continue**.

Select fixes

The following results match your request. Select the fixes you want to download.

To try a different query, go to the [Identify fixes](#) page.

[Share this download list](#)

Continue [Clear selections](#) [Show fix details](#) | [Hide fix details](#)

1-10 of 28 results [Next →](#) Results per page: [10](#) | [20](#) | [All](#)

<input checked="" type="checkbox"/> 1. fix pack: com.ibm.director.power.ha.systemmirror.server.feature_7.1.2.1 → com.ibm.director.power.ha.systemmirror.server.feature	Nov 9, 2012
<input checked="" type="checkbox"/> 2. fix pack: com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1 → com.ibm.director.power.ha.systemmirror.agent.feature	Nov 9, 2012
<input checked="" type="checkbox"/> 3. fix pack: com.ibm.director.power.ha.systemmirror.server.feature_7.1.2.0 → com.ibm.director.power.ha.systemmirror.server.feature	Nov 9, 2012

Figure 9-3 Selecting the fixes for the PowerHA SystemMirror plug-in

A page with the download method options appears (Figure 9-4). Choose your preferred method and click **Continue**.

Download options
IBM Systems Director, IBM Systems Director (6.3, AIX)

Select download options

Select the download method to be used to download fixes.

Download using Download Director (requires Java enabled browser) [What is this?](#)

Download using bulk FTP [What is this?](#)

Download using your browser (HTTP)

CAUTION: Do not assume that Fix Central will show you all the prerequisites you need.
Be sure to always click the **More information** link for additional prerequisite and other important fix information. Click [here](#) for an explanation of what prerequisites you can expect Fix Central to provide.

Include prerequisites and co-requisite fixes (you can select the ones you need later)

Continue [Back](#)

Figure 9-4 Download options

A confirmation page appears. Verify that all fixes are selected and click **Download Now** as shown in Figure 9-5.

The screenshot shows a confirmation page for downloading files using Download Director. At the top, it says "Download files using Download Director" and "IBM Systems Director, IBM Systems Director (6.3, AIX)". Below that, a section titled "Select files to download using Download Director" instructs the user to "Select the fixes you want to download and click the **Download now** button." It lists three fix packs:

Fix Pack Details	Date
1. fix pack: com.ibm.director.power.ha.systemmirror.server.feature_7.1.2.1 (8.13 MB) com.ibm.director.power.ha.systemmirror.server.feature	Nov 9, 2012
2. fix pack: com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1 (570.62 KB) com.ibm.director.power.ha.systemmirror.agent.feature	Nov 9, 2012
3. fix pack: com.ibm.director.power.ha.systemmirror.server.feature_7.1.2.0 (8.11 MB) com.ibm.director.power.ha.systemmirror.server.feature	Nov 9, 2012

At the bottom, there are two buttons: "Download now" and "Back".

Figure 9-5 Confirmation page

Upload these fixes to a location on your IBM Systems Director server and open the Console. On the Home page select the **Plug-in** tab and click **Update IBM Systems Director** (Figure 9-6 on page 404).

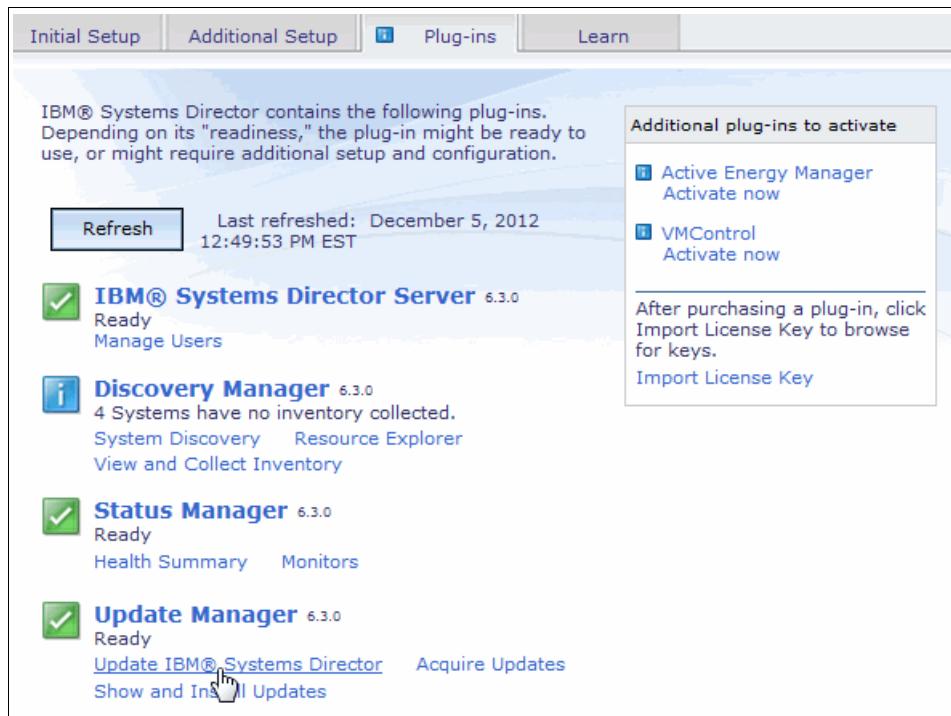


Figure 9-6 Updating the IBM Systems Director

Then it tries to connect to download fixes from the Internet. Click **Stop** as shown in Figure 9-7.

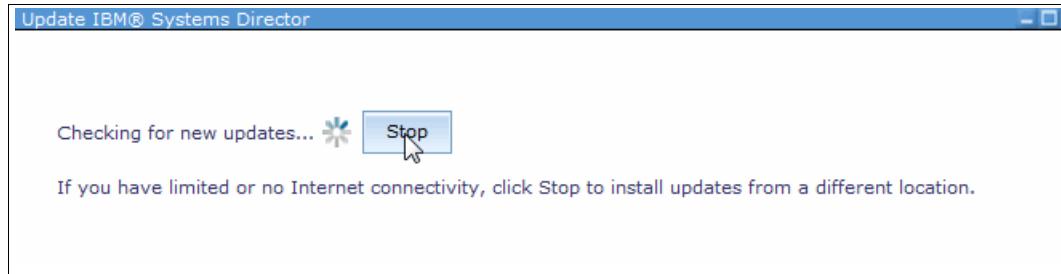


Figure 9-7 Stop Internet attempt

Fill the Path input with the location where you uploaded the fixes (Figure 9-8 on page 405). Then click **Import and Install**.

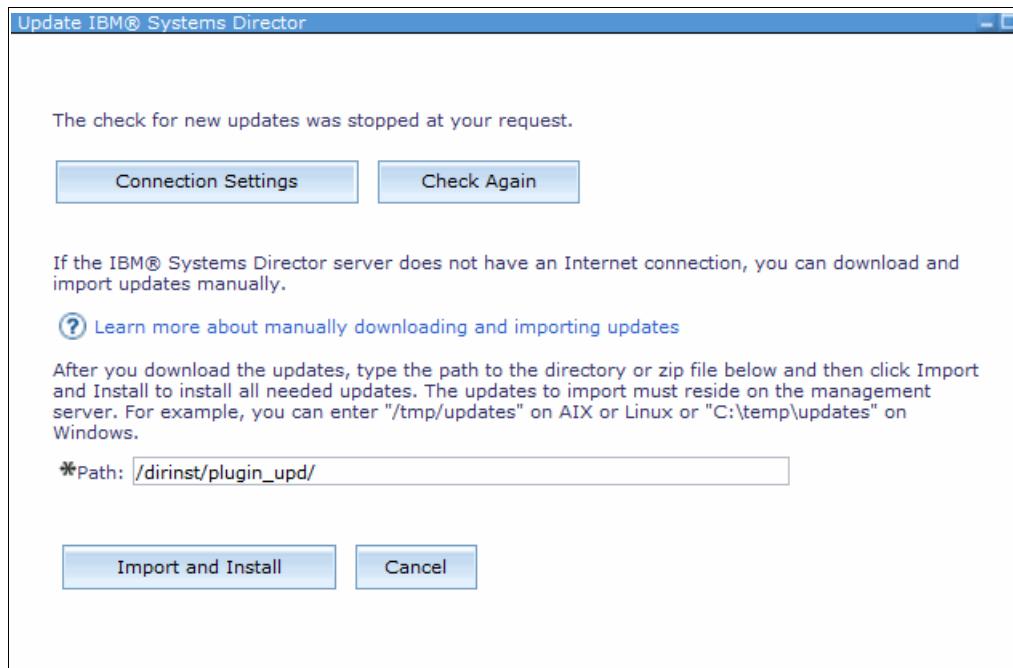


Figure 9-8 Path to the updates

Then an Update IBM System Director panel appears. Ensure that **Run Now** is selected (Figure 9-9) and click **OK**.

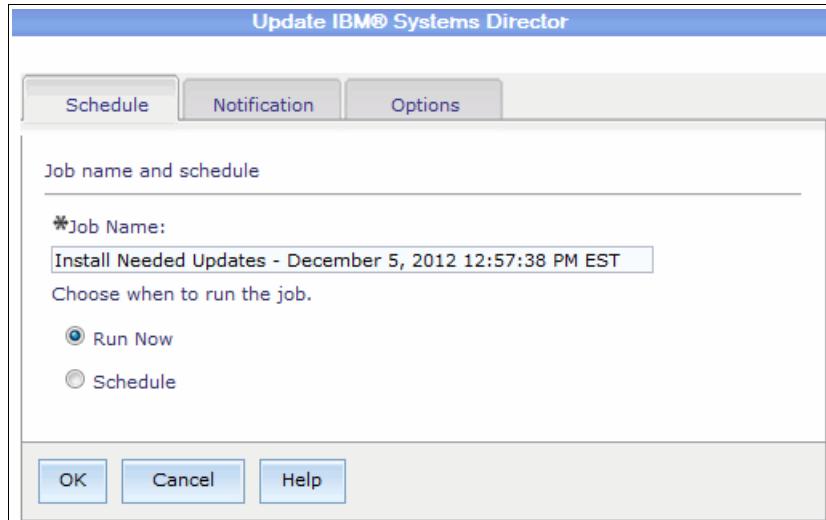


Figure 9-9 Run now panel

The logs of your task are displayed. Search for and verify the correct installation of the packages (Example 9-5).

Example 9-5 Verify package installation

```
December 5, 2012 1:08:21 PM EST-Level:1-MEID:0--MSG: Job "Install Needed Updates - December 5, 2012 12:57:38 PM EST" activated.  
December 5, 2012 1:08:21 PM EST-Level:200-MEID:0--MSG: Subtask "Install Needed Updates" activated.
```

December 5, 2012 1:08:21 PM EST-Level:200-MEID:0--MSG: Starting clients
December 5, 2012 1:08:21 PM EST-Level:100-MEID:0--MSG: Clients started for task "Install Needed Updates"
December 5, 2012 1:08:21 PM EST-Level:200-MEID:0--MSG: Subtask activation status changed to "Active".
December 5, 2012 1:08:21 PM EST-Level:200-MEID:0--MSG: Subtask activation status changed to "Starting".
December 5, 2012 1:08:21 PM EST-Level:1-MEID:0--MSG: Job activation status changed to "Active".
December 5, 2012 1:08:22 PM EST-Level:200-MEID:0--MSG: Subtask activation status changed to "Active".
December 5, 2012 1:08:22 PM EST-Level:150-MEID:0--MSG: ATKUPD489I Collecting inventory for one or more systems.
December 5, 2012 1:09:40 PM EST-Level:100-MEID:2903--MSG: ibmdirector client job status changed to "Active".
December 5, 2012 1:09:46 PM EST-Level:1-MEID:2987--MSG: Job "Install Updates" activated.
December 5, 2012 1:09:46 PM EST-Level:200-MEID:2987--MSG: Subtask "Install Updates" activated.
December 5, 2012 1:09:46 PM EST-Level:200-MEID:2987--MSG: Starting clients
December 5, 2012 1:09:46 PM EST-Level:100-MEID:2987--MSG: Clients started for task "Install Updates"
December 5, 2012 1:09:46 PM EST-Level:200-MEID:2987--MSG: Subtask activation status changed to "Active".
December 5, 2012 1:09:46 PM EST-Level:200-MEID:2987--MSG: Subtask activation status changed to "Starting".
December 5, 2012 1:09:46 PM EST-Level:1-MEID:2987--MSG: Job activation status changed to "Active".
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD725I The update install task has started.
December 5, 2012 1:09:46 PM EST-Level:200-MEID:2987--MSG: Subtask activation status changed to "Active".
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD487I The download task has finished successfully.
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD629I Installation staging will be performed to 1 systems.
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD632I The Installation Staging task is starting to process system "IBM 8233E8B 100C99R 3".
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD633I The Installation Staging task has finished processing system "IBM 8233E8B 100C99R 3".
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD630I The update installation staging has completed.
December 5, 2012 1:09:46 PM EST-Level:150-MEID:2987--MSG: ATKUPD760I Start processing update "com.ibm.director.power.ha.systemmirror.server.feature_7.1.2.1" and system "IBM 8233E8B 100C99R 3".
December 5, 2012 1:09:58 PM EST-Level:150-MEID:2987--MSG: ATKUPD764I Update "com.ibm.director.power.ha.systemmirror.server.feature_7.1.2.1" was installed on system "IBM 8233E8B 100C99R 3" successfully.
December 5, 2012 1:09:58 PM EST-Level:150-MEID:2987--MSG: ATKUPD795I You must manually restart the IBM Systems Director management server after this install completes for the updates to take effect.
December 5, 2012 1:09:58 PM EST-Level:100-MEID:2987--MSG: IBM 8233E8B 100C99R 3 client job status changed to "Complete".
December 5, 2012 1:09:58 PM EST-Level:150-MEID:2987--MSG: ATKUPD739I Collecting inventory on system "IBM 8233E8B 100C99R 3".

December 5, 2012 1:10:53 PM EST-Level:150-MEID:2987--MSG: ATKUPD572I Running compliance on system "IBM 8233E8B 100C99R 3".
December 5, 2012 1:10:53 PM EST-Level:200-MEID:2987--MSG: Subtask activation status changed to "Complete".
December 5, 2012 1:10:53 PM EST-Level:1-MEID:2987--MSG: Job activation status changed to "Complete".
December 5, 2012 1:10:53 PM EST-Level:150-MEID:2987--MSG: ATKUPD727I The update install task has finished successfully.
December 5, 2012 1:10:59 PM EST-Level:100-MEID:2903--MSG: ibmdirector client job status changed to "Complete".
December 5, 2012 1:10:59 PM EST-Level:150-MEID:2987--MSG: ATKUPD288I The install needed updates task has completed successfully.
December 5, 2012 1:10:59 PM EST-Level:200-MEID:0--MSG: Subtask activation status changed to "Complete".
December 5, 2012 1:10:59 PM EST-Level:1-MEID:0--MSG: Job activation status changed to "Complete".

Next, a warning displays showing that a restart of Systems Director is required (Figure 9-10).

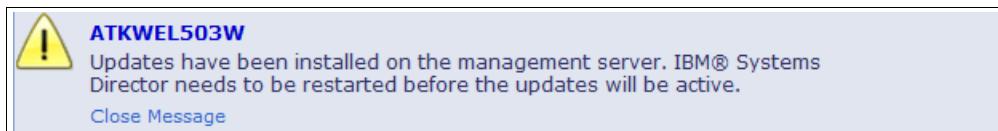


Figure 9-10 Update warning

To restart Systems Director, connect to the server and follow the instructions as shown in Example 9-6.

Example 9-6 Restarting IBM Systems Director

```
root@ibmdir / # smstop
Shutting down IBM Director...
root@ibmdir / # smstart
Starting IBM Director...
The starting process may take a while. Please use smstatus to check if the server
is active.

root@ibmdir / # smstatus -r
Starting
Active
```

After the upgrade, the plug-in tab displays the updated version (Figure 9-11).



Figure 9-11 Plug-in updated

9.1.4 Installing PowerHA SystemMirror agent fixes

After this procedure, you are ready to install the agent update. At the Home window, under the plug-in tab, click **Show and Install Updates** under Update Manager (Figure 9-12).

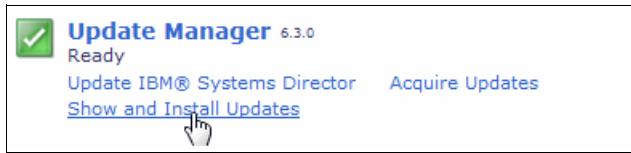


Figure 9-12 Show and install updates

Click **Browse** to select the systems where you want to install the agent update (Figure 9-13).

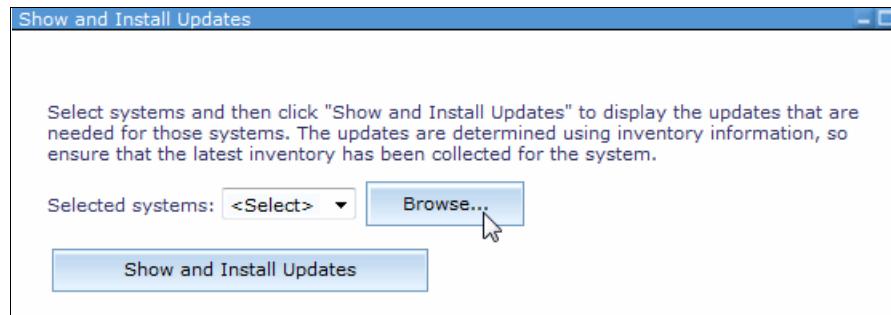


Figure 9-13 Browsing the systems to install the Systems Director update

Choose all systems to be updated by selecting them on the Systems table and click **Add**. They now appear as Selected (Figure 9-14). Click **OK**.

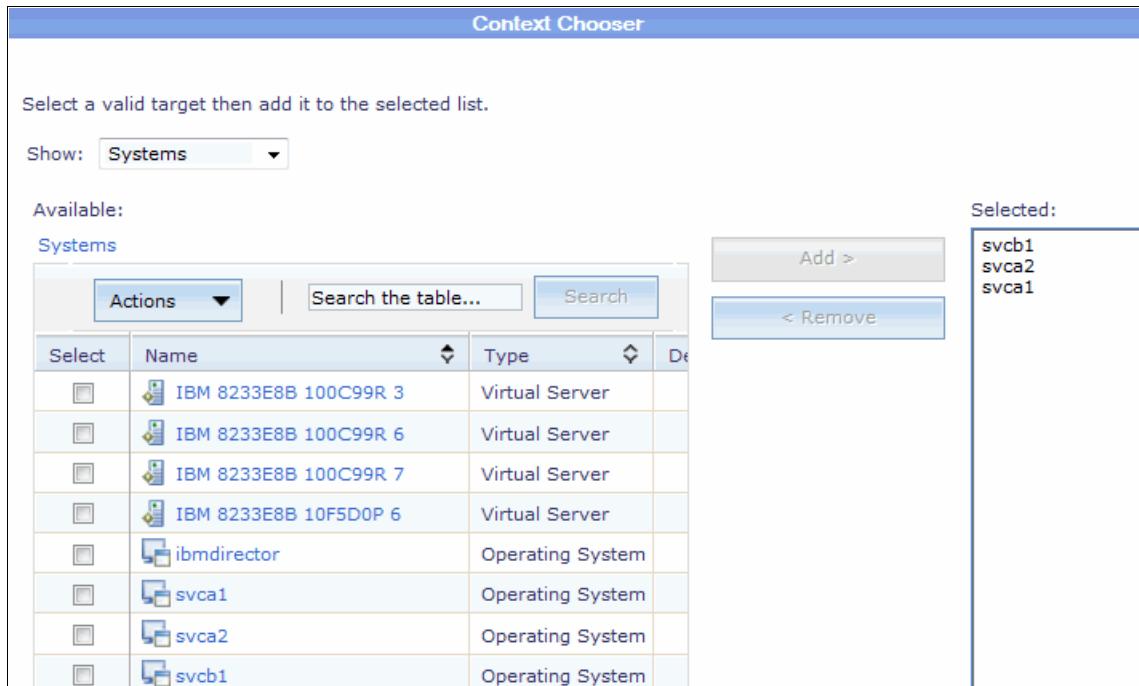


Figure 9-14 Selecting systems to install update

Click **Show and Install Updates** (Figure 9-15 on page 409).

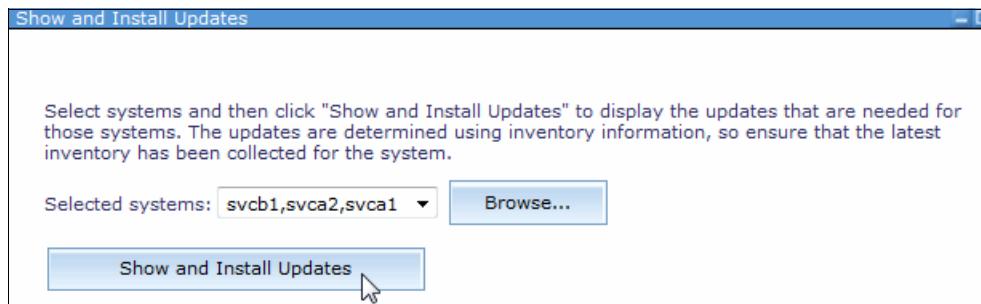


Figure 9-15 Show and install updates

If a warning appears stating that inventory has not been collected for some systems (Figure 9-16), click **Collect Inventory** to collect inventory for all the systems.



Figure 9-16 Collect inventory

Now select the fixes that apply to the systems and click **Install** (Figure 9-17).

Updates needed for "svcb1,svca2,svca1":								
		Actions		Search the table...		Search		
Select	Name	Installs updates on one or more systems		System	Version	Severity	Product	
<input checked="" type="checkbox"/>	 com.ibm.director.power.ha.systemmirror.agent.feature			3 systems	7.1.2.1	Medium	IBM Sys	
<input type="button" value="Install..."/> <input type="button" value="Actions"/> <input type="button" value="Search the table..."/> <input type="button" value="Search"/>								
<input type="button" value="Page 1 of 1"/> <input type="button" value="Page 2 of 1"/> <input type="button" value="Page 3 of 1"/> <input type="button" value="Page 4 of 1"/> <input type="button" value="Page 5 of 1"/> <input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> Selected: 1 Total: 1 Filtered: 1								

Figure 9-17 Selecting fixes for the PowerHA SystemMirror agent

Follow the installation wizard to complete the installation of the selected fixes. Click **Next** on the welcome page (Figure 9-18 on page 410).

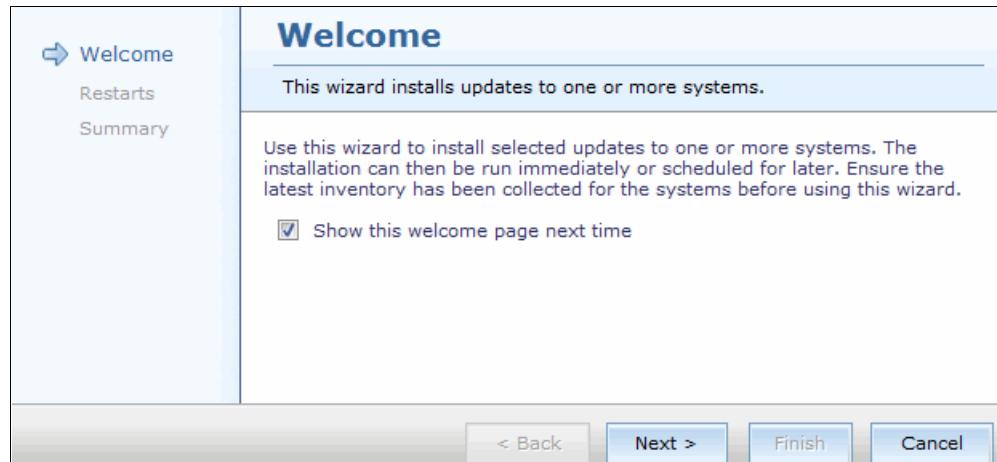


Figure 9-18 Welcome page

Restarting the Common Agents is required, so select **Automatically restart as needed during installation** and click **Next** (Figure 9-19).

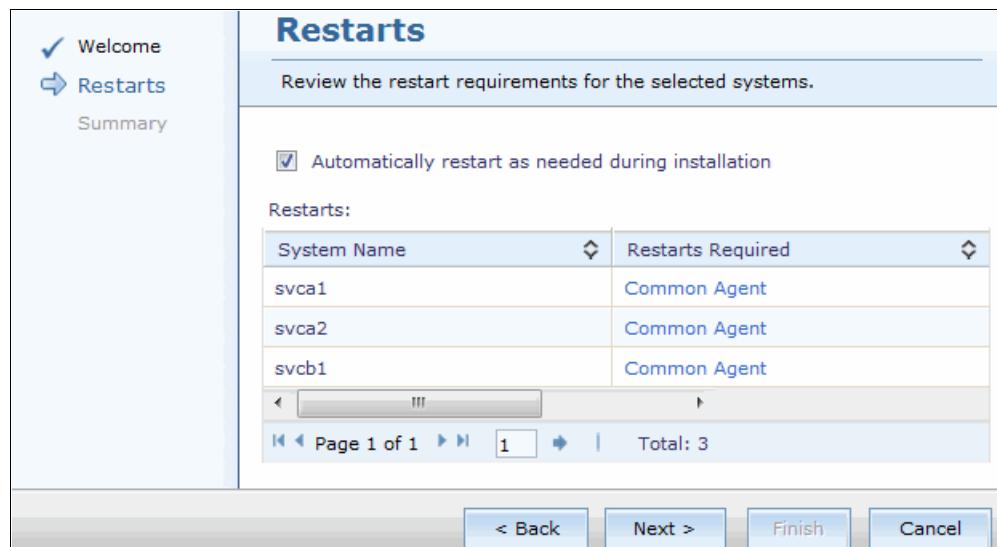


Figure 9-19 Re-starting requirements for the selected systems

Verify that all the information is correct and then click **Finish** on the Summary page (Figure 9-20 on page 411).

Summary

The updates will now be installed on the selected systems. Verify the installation settings below.

Selected updates:

Name	Version	Severity
com.ibm.director.power.ha.system	7.1.2.1	Medium

Page 1 of 1 | Total: 1

Selected systems:

Name	Type	Description
svca1	Operating System	
svca2	Operating System	
svcb1	Operating System	

Page 1 of 1 | Total: 3

< Back | Next > | **Finish** | Cancel

Figure 9-20 Summary

Verify that **Run Now** is selected then click **OK** on the launch job panel (Figure 9-21).

Launch Job

Schedule Notification Options

Job name and schedule

*Job Name:
Install Updates - December 6, 2012 6:41:41 PM EST

Choose when to run the job.

Run Now
 Schedule

OK | Cancel | Help

Figure 9-21 Launch job panel

The Job Properties window is displayed. Wait for the completion of the job (Figure 9-22 on page 412).

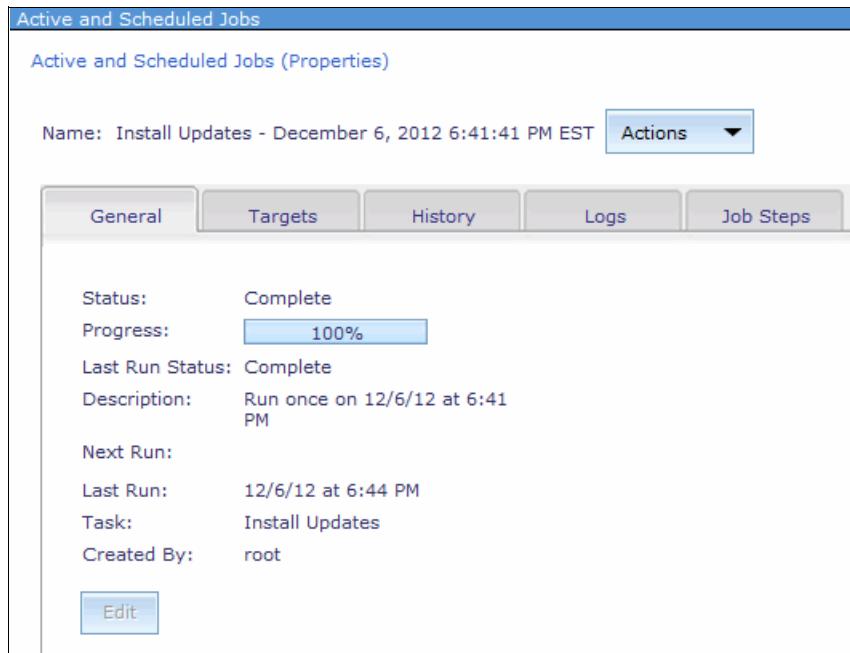


Figure 9-22 Job properties

The logs tab shows the status of the update (Example 9-7).

Example 9-7 Log output

```

December 6, 2012 6:44:56 PM EST-Level:1-MEID:0--MSG: Job "Install Updates - 
December 6, 2012 6:41:41 PM EST" activated.
December 6, 2012 6:44:56 PM EST-Level:200-MEID:0--MSG: Subtask "Install Updates" 
activated.
December 6, 2012 6:44:56 PM EST-Level:200-MEID:0--MSG: Starting clients
December 6, 2012 6:44:56 PM EST-Level:100-MEID:0--MSG: Clients started for task 
"Install Updates"
December 6, 2012 6:44:56 PM EST-Level:200-MEID:0--MSG: Subtask activation status 
changed to "Active".
December 6, 2012 6:44:56 PM EST-Level:200-MEID:0--MSG: Subtask activation status 
changed to "Starting".
December 6, 2012 6:44:56 PM EST-Level:1-MEID:0--MSG: Job activation status changed 
to "Active".
December 6, 2012 6:44:56 PM EST-Level:150-MEID:0--MSG: ATKUPD725I The update 
install task has started.
December 6, 2012 6:44:56 PM EST-Level:200-MEID:0--MSG: Subtask activation status 
changed to "Active".
December 6, 2012 6:44:56 PM EST-Level:150-MEID:0--MSG: ATKUPD487I The download 
task has finished successfully.
December 6, 2012 6:44:56 PM EST-Level:150-MEID:0--MSG: ATKUPD629I Installation 
staging will be performed to 1 systems.
December 6, 2012 6:44:56 PM EST-Level:150-MEID:0--MSG: ATKUPD629I Installation 
staging will be performed to 1 systems.
December 6, 2012 6:44:56 PM EST-Level:150-MEID:0--MSG: ATKUPD629I Installation 
staging will be performed to 1 systems.
December 6, 2012 6:44:56 PM EST-Level:150-MEID:5513--MSG: ATKUPD632I The 
Installation Staging task is starting to process system "svca2".
December 6, 2012 6:44:56 PM EST-Level:150-MEID:5610--MSG: ATKUPD632I The 
Installation Staging task is starting to process system "svcb1".

```

December 6, 2012 6:44:56 PM EST-Level:150-MEID:5414--MSG: ATKUPD632I The Installation Staging task is starting to process system "svca1".

December 6, 2012 6:45:04 PM EST-Level:150-MEID:5414--MSG: ATKUPD686I The update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" has been staged for installation to
"/tmp/updatemanager/staging/Director/com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1/" on the managed resource "svca1" successfully.

December 6, 2012 6:45:04 PM EST-Level:150-MEID:5414--MSG: ATKUPD633I The Installation Staging task has finished processing system "svca1".

December 6, 2012 6:45:04 PM EST-Level:150-MEID:0--MSG: ATKUPD630I The update installation staging has completed.

December 6, 2012 6:45:04 PM EST-Level:150-MEID:5513--MSG: ATKUPD686I The update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" has been staged for installation to
"/tmp/updatemanager/staging/Director/com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1/" on the managed resource "svca2" successfully.

December 6, 2012 6:45:04 PM EST-Level:150-MEID:5513--MSG: ATKUPD633I The Installation Staging task has finished processing system "svca2".

December 6, 2012 6:45:04 PM EST-Level:150-MEID:0--MSG: ATKUPD630I The update installation staging has completed.

December 6, 2012 6:45:05 PM EST-Level:150-MEID:5610--MSG: ATKUPD686I The update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" has been staged for installation to
"/tmp/updatemanager/staging/Director/com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1/" on the managed resource "svcb1" successfully.

December 6, 2012 6:45:05 PM EST-Level:150-MEID:5610--MSG: ATKUPD633I The Installation Staging task has finished processing system "svcb1".

December 6, 2012 6:45:05 PM EST-Level:150-MEID:0--MSG: ATKUPD630I The update installation staging has completed.

December 6, 2012 6:45:06 PM EST-Level:150-MEID:5414--MSG: ATKUPD760I Start processing update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" and system "svca1".

December 6, 2012 6:45:06 PM EST-Level:150-MEID:5610--MSG: ATKUPD760I Start processing update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" and system "svcb1".

December 6, 2012 6:45:06 PM EST-Level:150-MEID:5513--MSG: ATKUPD760I Start processing update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" and system "svca2".

December 6, 2012 6:45:07 PM EST-Level:150-MEID:5414--MSG: ATKUPD764I Update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" was installed on system "svca1" successfully.

December 6, 2012 6:45:07 PM EST-Level:150-MEID:5610--MSG: ATKUPD764I Update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" was installed on system "svcb1" successfully.

December 6, 2012 6:45:07 PM EST-Level:150-MEID:5513--MSG: ATKUPD764I Update "com.ibm.director.power.ha.systemmirror.agent.feature_7.1.2.1" was installed on system "svca2" successfully.

December 6, 2012 6:45:08 PM EST-Level:150-MEID:5414--MSG: ATKUPD793I Restarting the Common Agent on system "svca1".

December 6, 2012 6:45:08 PM EST-Level:150-MEID:5610--MSG: ATKUPD793I Restarting the Common Agent on system "svcb1".

December 6, 2012 6:45:08 PM EST-Level:150-MEID:5513--MSG: ATKUPD793I Restarting the Common Agent on system "svca2".

December 6, 2012 6:48:54 PM EST-Level:150-MEID:5610--MSG: ATKUPD798I Update Manager successfully restarted the Common Agent on system "svcb1".
December 6, 2012 6:48:54 PM EST-Level:100-MEID:5363--MSG: svcb1 client job status changed to "Complete".
December 6, 2012 6:48:54 PM EST-Level:150-MEID:5363--MSG: ATKUPD739I Collecting inventory on system "svcb1".
December 6, 2012 6:48:56 PM EST-Level:150-MEID:5513--MSG: ATKUPD798I Update Manager successfully restarted the Common Agent on system "svca2".
December 6, 2012 6:48:56 PM EST-Level:100-MEID:5349--MSG: svca2 client job status changed to "Complete".
December 6, 2012 6:48:56 PM EST-Level:150-MEID:5349--MSG: ATKUPD739I Collecting inventory on system "svca2".
December 6, 2012 6:49:16 PM EST-Level:150-MEID:5414--MSG: ATKUPD798I Update Manager successfully restarted the Common Agent on system "svca1".
December 6, 2012 6:49:16 PM EST-Level:100-MEID:5335--MSG: svca1 client job status changed to "Complete".
December 6, 2012 6:49:16 PM EST-Level:150-MEID:5335--MSG: ATKUPD739I Collecting inventory on system "svca1".
December 6, 2012 6:50:14 PM EST-Level:150-MEID:5363--MSG: ATKUPD572I Running compliance on system "svcb1".
December 6, 2012 6:50:16 PM EST-Level:150-MEID:5349--MSG: ATKUPD572I Running compliance on system "svca2".
December 6, 2012 6:50:36 PM EST-Level:150-MEID:5335--MSG: ATKUPD572I Running compliance on system "svca1".
December 6, 2012 6:50:36 PM EST-Level:200-MEID:0--MSG: Subtask activation status changed to "Complete".
December 6, 2012 6:50:36 PM EST-Level:1-MEID:0--MSG: Job activation status changed to "Complete".
December 6, 2012 6:50:36 PM EST-Level:150-MEID:0--MSG: ATKUPD727I The update install task has finished successfully.

Version display: Even though the 7.1.2.1 installation completed successfully, the version shown through `1s1pp` is 7.1.2.0.

9.2 Environment configuration

The cluster design for this chapter is the same one used in 6.6.1, “SVC-V7000 mixed environment” on page 230. One DNS server to provide all name resolution and one IBM Systems Director server were installed in the environment to support our cluster configuration.

Name resolution: Name resolution is important for IBM Systems Director. It uses this feature to translate the hostnames of the director’s endpoints being connected to it to the IP address used for managing the connections.

After the IBM V7000 is configured as a replicated resource with the IBM SVC, follow the steps described in 6.3.2, “Environment configuration for PowerHA” on page 205, and you will be ready to begin configuring the cluster using Systems Director.

9.3 Configuring IBM PowerHA SystemMirror 7.1.2 Enterprise Edition using IBM Systems Director

This section describes how to configure IBM PowerHA SystemMirror 7.1.2 Enterprise Edition using IBM Systems Director.

9.3.1 Discovering servers

The PowerHA director CAS agent software is needed to configure the PowerHA cluster using the Director graphical interface. This means the fileset `cluster.es.director.agent.rte` must be installed as shown in Example 9-8.

Example 9-8 Software required to configure PowerHA with IBM Systems Director

root@svca1:/>lslpp -l grep cluster.es sort -u			
cluster.es.client.clcomd	7.1.2.0	COMMITTED	Cluster Communication
cluster.es.client.lib	7.1.2.1	COMMITTED	PowerHA SystemMirror Client
cluster.es.client.rte	7.1.2.1	COMMITTED	PowerHA SystemMirror Client
cluster.es.client.utils	7.1.2.1	COMMITTED	PowerHA SystemMirror Client
cluster.es.client.wsm	7.1.2.0	COMMITTED	Web based Smit
cluster.es.cspoc.cmds	7.1.2.1	COMMITTED	CSPOC Commands
cluster.es.cspoc.dsh	7.1.2.0	COMMITTED	CSPOC dsh
cluster.es.cspoc.rte	7.1.2.0	COMMITTED	CSPOC Runtime Commands
cluster.es.cspoc.rte	7.1.2.1	COMMITTED	CSPOC Runtime Commands
cluster.es.director.agent.rte	7.1.2.0 COMMITTED	PowerHA SystemMirror Director CAS agent	
cluster.es.genxd.cmds	7.1.2.0	COMMITTED	PowerHA SystemMirror
cluster.es.genxd.cmds	7.1.2.1	COMMITTED	PowerHA SystemMirror
cluster.es.genxd.rte	7.1.2.0	COMMITTED	PowerHA SystemMirror
cluster.es.genxd.rte	7.1.2.1	COMMITTED	PowerHA SystemMirror
cluster.es.migcheck	7.1.2.0	COMMITTED	PowerHA SystemMirror Migration
cluster.es.server.cfgast	7.1.2.0	COMMITTED	Two-Node Configuration
cluster.es.server.diag	7.1.2.0	COMMITTED	Server Diags
cluster.es.server.diag	7.1.2.1	COMMITTED	Server Diags
cluster.es.server.events	7.1.2.0	COMMITTED	Server Events
cluster.es.server.events	7.1.2.1	COMMITTED	Server Events
cluster.es.server.rte	7.1.2.1	COMMITTED	Base Server Runtime
cluster.es.server.testtool	7.1.2.0	COMMITTED	Cluster Test Tool
cluster.es.server.utils	7.1.2.1	COMMITTED	Server Utilities
cluster.es.svcpprc.cmds	7.1.2.0	COMMITTED	PowerHA SystemMirror
cluster.es.svcpprc.rte	7.1.2.0	COMMITTED	PowerHA SystemMirror
cluster.es.worksheets	7.1.2.0	COMMITTED	Online Planning Worksheets

Repeat the required software check procedure on all nodes and verify the software is correctly installed.

Now you must discover, give access, and collect inventory for the first time on Systems Director as shown in Example 9-9. Repeat the procedure from the Systems Director server until the inventory is collected from all nodes.

Example 9-9 Collecting inventory information for the first time

```
root@ibmdirior /dirinst # smcli discover -H svca1
```

```
Total discovery modules:
```

```
22
```

```

Discovery completion percentage 82%
Discovery completion percentage 96%
Discovery completion percentage 100%
Discovery completed:
100%
root@ibmdirector /dirinst # smcli accesssys -i svcal
Type the user ID:root
Type the password:
DNZCLI0727I : Waiting for request access to complete on... svcal
Result Value: DNZCLI0734I : Request access was successful on : svcal
root@ibmdirector / # smcli collectinv -p "All Inventory" -n svcal
Inventory collection percentage 0%
Inventory collection percentage 84%
Inventory collection percentage 95%
Inventory collection completed:
100%

```

9.3.2 Creating the cluster

On the IBM Systems Director left navigation panel, select **Availability** → **PowerHA SystemMirror**. Then, on the main panel within the Cluster Management Section, find and click the link to the **Create Cluster** utility (Figure 9-23).



Figure 9-23 Creating the cluster wizard

Click **Next** on the Create Cluster Wizard welcome screen and fill in the required information. In our case, we have a two-site stretched cluster named `ihs_cluster` (Figure 9-24 on page 417).

✓ Welcome

➡ Configure the cluster

Choose nodes

Configure nodes

Choose repository

Summary

Configure the cluster

Provide a name, the level of security and, if necessary, provide a multicast address for the new cluster.

*Cluster name:

Configure security:

Choose the sites: No sites 2 sites

Choose cluster type: Linked Stretched

Cluster multicast address (IPv4):

*Site 1 name:

*Site 2 name:

Inter-site recovery:

Notify script or executable:

[? What is the multicast address used for?](#)

[? How do I choose cluster type ?](#)

< Back Finish Cancel

The screenshot shows the 'Configure the cluster' step of a PowerHA setup wizard. On the left, a sidebar lists steps: 'Welcome' (marked with a checkmark), 'Configure the cluster' (marked with a right arrow), 'Choose nodes', 'Configure nodes', 'Choose repository', and 'Summary'. The main area has a title 'Configure the cluster' and a sub-instruction 'Provide a name, the level of security and, if necessary, provide a multicast address for the new cluster.' It contains fields for 'Cluster name' (set to 'ihs_cluster'), 'Configure security' (set to 'No'), 'Choose the sites' (set to '2 sites'), 'Choose cluster type' (set to 'Stretched'), 'Cluster multicast address (IPv4)' (empty), 'Site 1 name' (set to 'SiteA'), 'Site 2 name' (set to 'SiteB'), 'Inter-site recovery' (set to 'Automatically failover'), and 'Notify script or executable' (empty). Below these fields are two help links: '? What is the multicast address used for?' and '? How do I choose cluster type ?'. At the bottom are buttons for '< Back', 'Next >', 'Finish', and 'Cancel'.

Figure 9-24 Initial cluster and site configuration

Click **Next** and then choose the nodes from each site (Figure 9-25 on page 418).

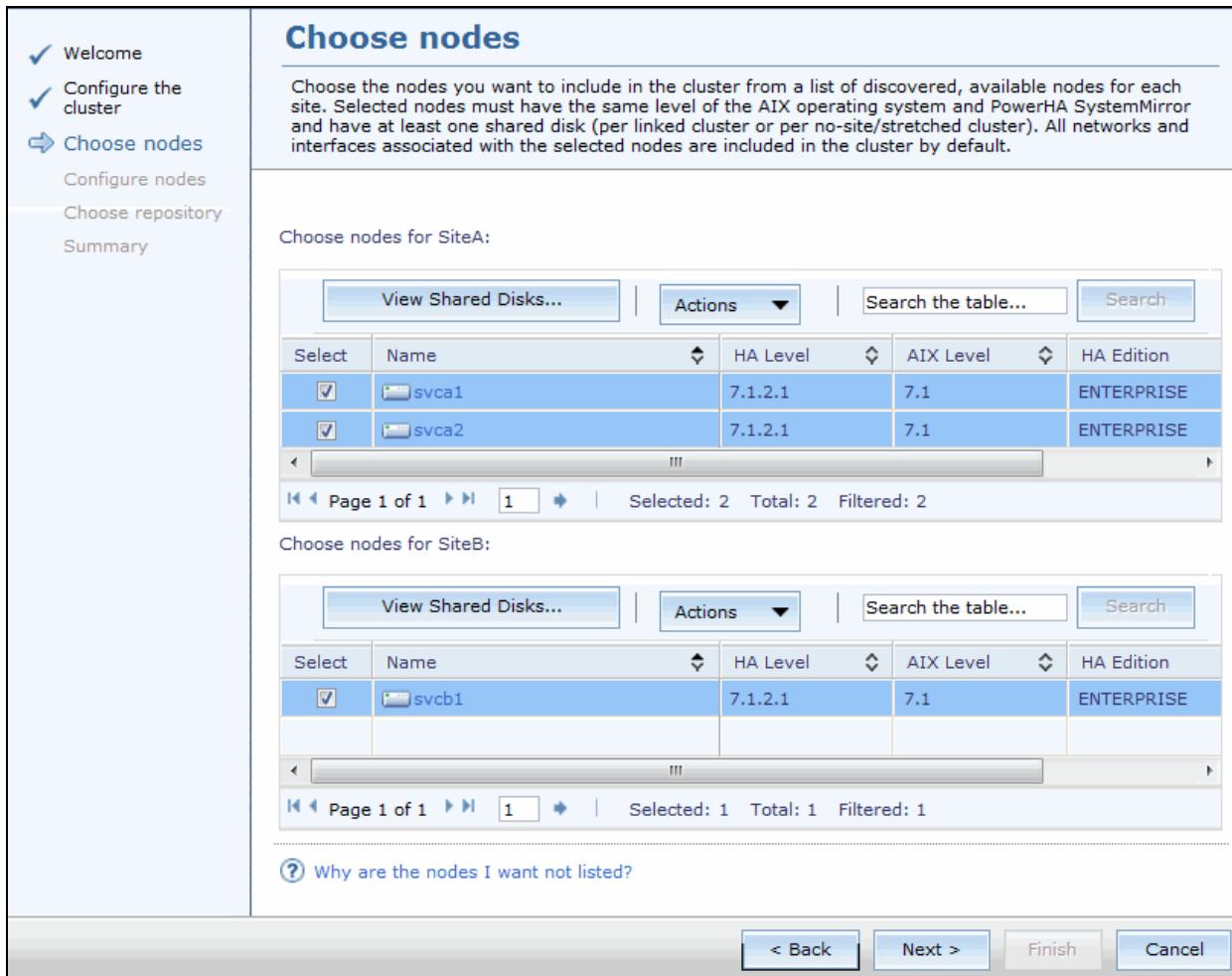


Figure 9-25 Choose nodes from each site

Select the IP address for each node and controlling node (Figure 9-26 on page 419) and click **Next**.

The controlling node is the primary contact point between Systems Director and PowerHA on which all commands are executed. If the specified controlling node is unreachable, Systems Director automatically changes to another node.

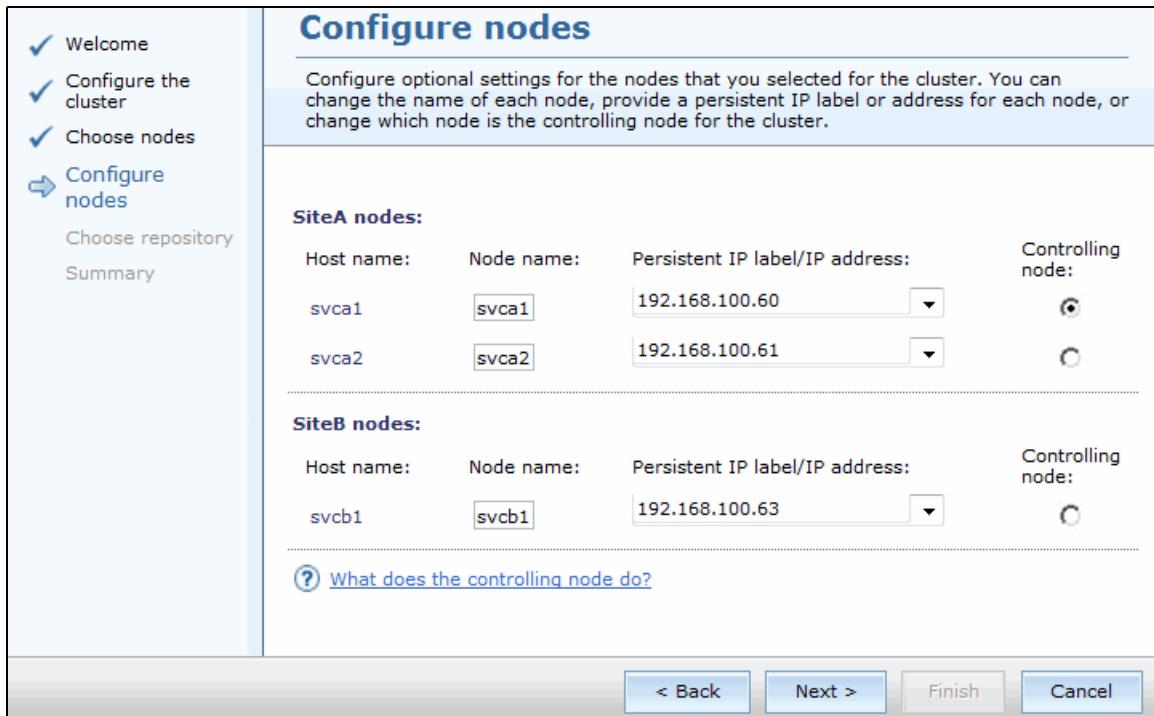


Figure 9-26 Configuring nodes for the cluster

Select the repository disk that will be used for the CAA data (Figure 9-27) and click **Next**.

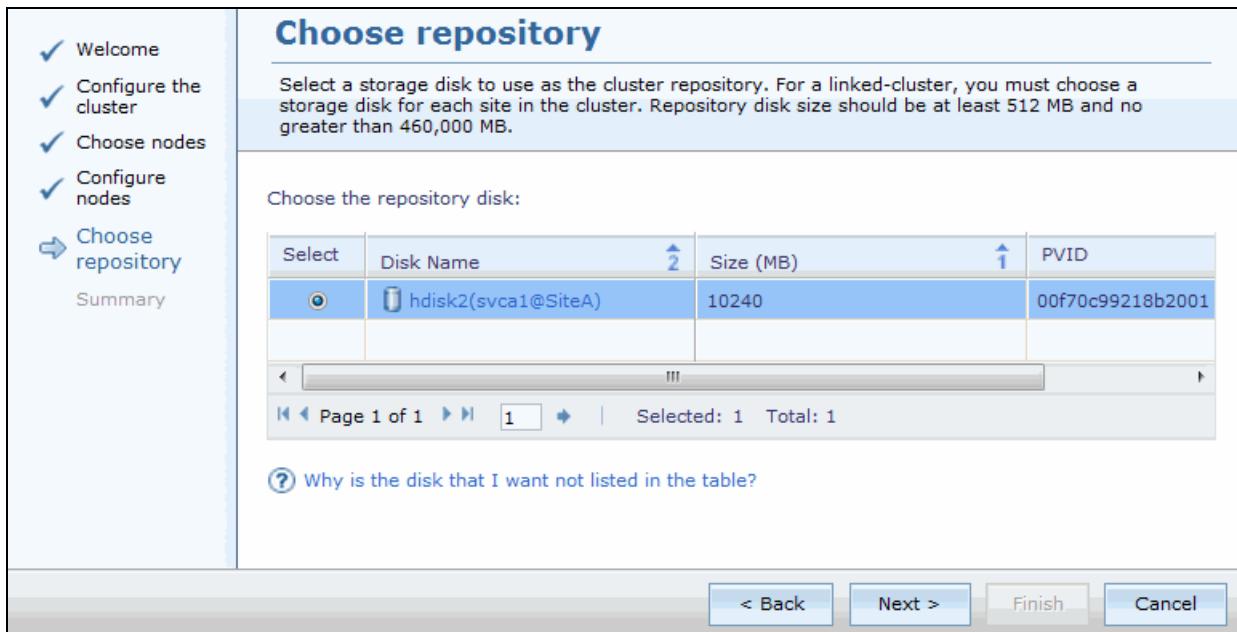


Figure 9-27 Choosing the CAA repository disk

Systems Director shows the current cluster configuration (Figure 9-28 on page 420). Verify all the information is correct and then click **Finish**.

Welcome

Configure the cluster

Choose nodes

Configure nodes

Choose repository

Summary

Summary

Review your cluster configuration settings and specify whether to start cluster services when you finish creating the new cluster.

Start cluster services

Note: Configuring clusters can take a long time. You can specify whether to start the configuration now or schedule as a job later after you click the "Finish" button.

ihs_cluster details:

Cluster type:	Stretched
Cluster multicast address (IPv4):	Take system default
HA level:	7.1.2.1
AIX level:	7.1
Controlling node:	svca1
Repository:	hdisk2(svca1@SiteA) 00f70c99218b2001
Inter-site recovery:	Automatically failover

Site 1:

Name:	SiteA
Node 1:	
Host name:	svca1
Node name:	svca1
Persistent IP label/IP address:	192.168.100.60
Node 2:	
Host name:	svca2
Node name:	svca2
Persistent IP label/IP address:	192.168.100.61

Site 2:

Name:	SiteB
Node 3:	
Host name:	svcb1
Node name:	svcb1
Persistent IP label/IP address:	192.168.100.63

[< Back](#) [Next >](#) [Finish](#) [Cancel](#)

Figure 9-28 Cluster configuration summary

The Launch Job pop-up appears (Figure 9-29 on page 421). Select **Run Now** and click **OK**.

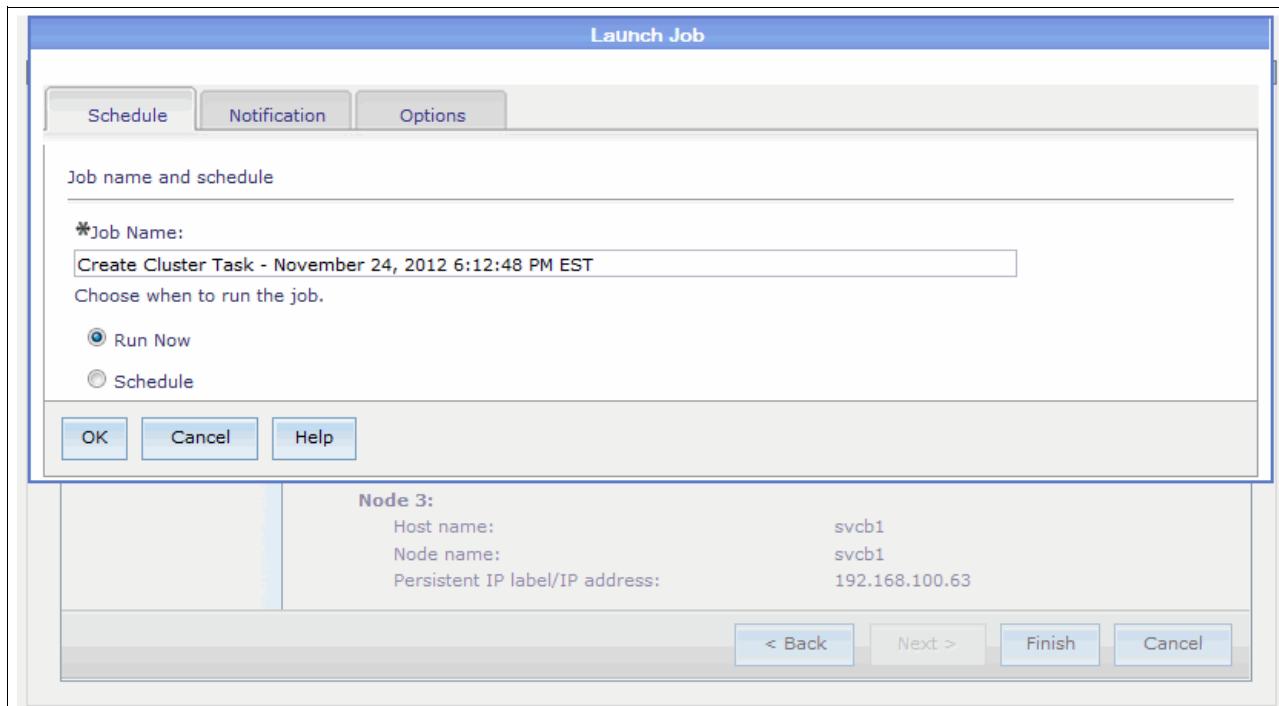


Figure 9-29 Launching the job

After the confirmation appears that the scheduled job is running, click **Display Properties**. This displays the job status on the Systems Director HTTP (Figure 9-30).

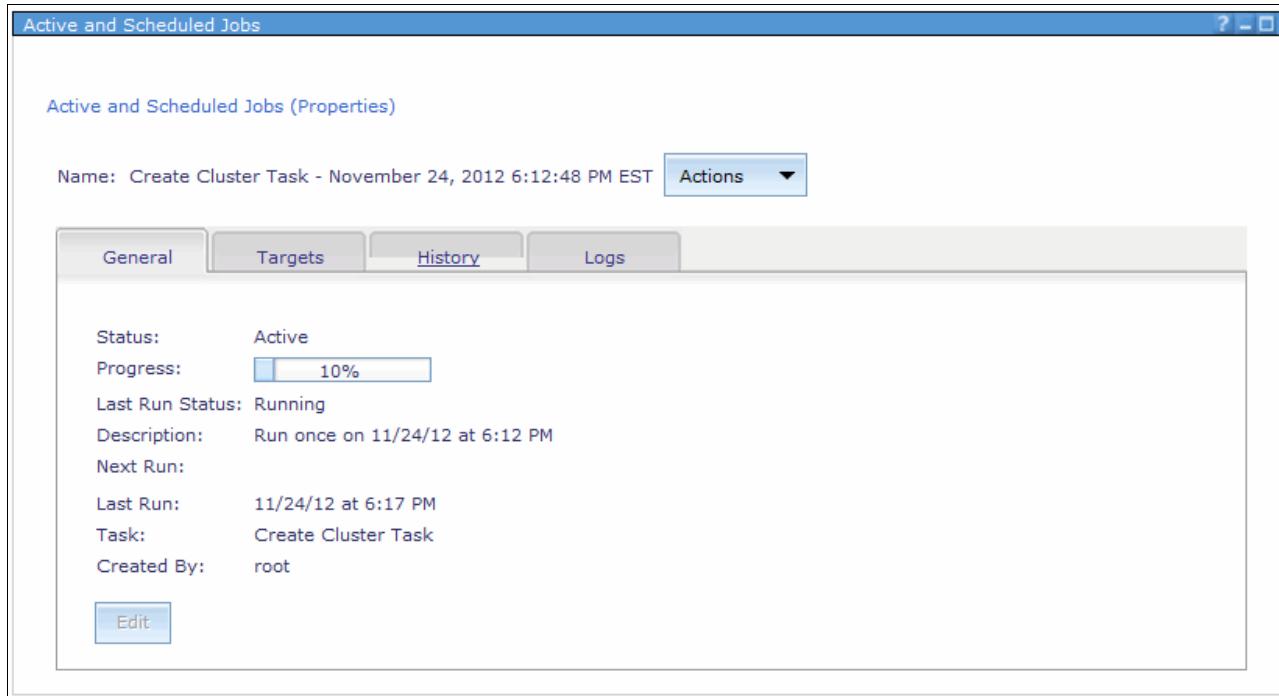


Figure 9-30 Job running

This procedure can take some time to complete. When finished, the cluster configuration job status displays complete (Figure 9-31 on page 422).

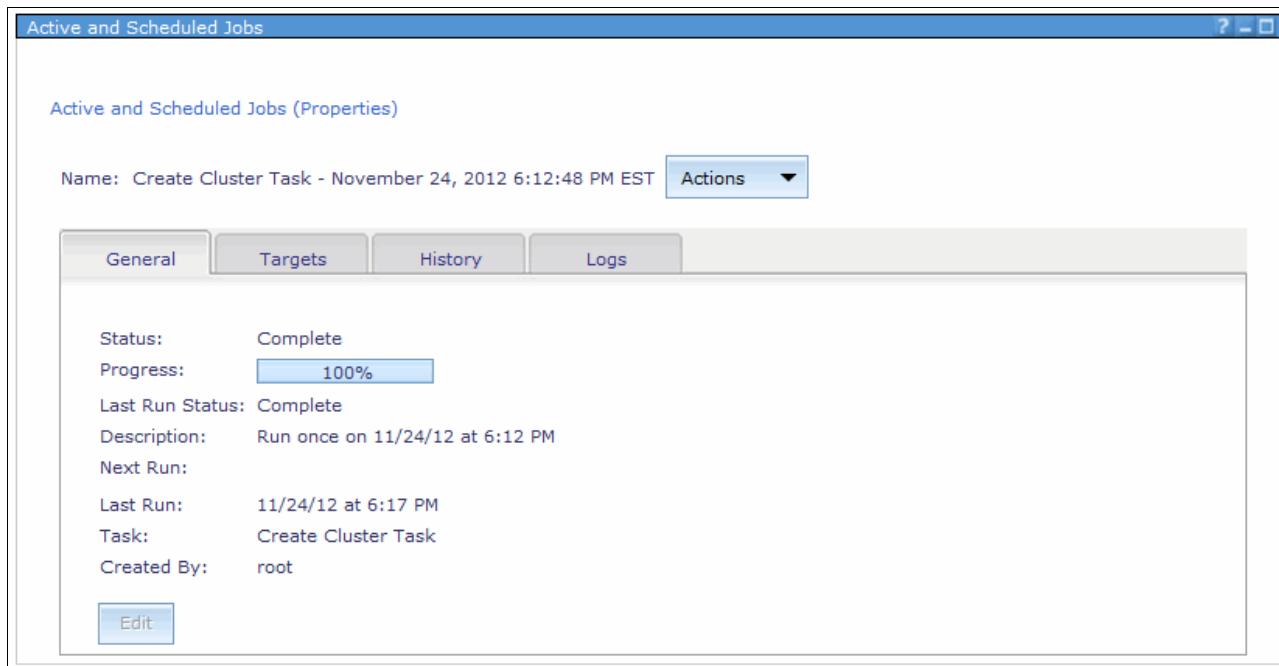


Figure 9-31 Job completed

Now if you navigate to **Availability** → **PowerHA SystemMirror** → **Cluster Management** → **Manage Cluster**, notice the cluster with the not configured status (Figure 9-32). This status will only be changed to Offline after you synchronize the cluster on the next step.

The screenshot shows the 'Clusters' tab in the PowerHA SystemMirror management interface. A table lists clusters: 'ihs_cluster' (selected), 'SiteA', 'svca1', 'svca2', 'SiteB', and 'svcb1'. The 'ihs_cluster' row shows 'Not configured' under HA Status. A detailed view panel on the right shows:

- General:** Name: ihs_cluster, Status: Not configured, Type: Stretched cluster.
- Software:** PowerHA SystemMirror version: 7.1.2.1, edition: ENTERPRISE, AIX version: 7.1.
- Resources:** Active repository disk: hdisk2@svca1, Controlling node: svca1, Cluster multicast address: svca1.
- Security:** Security Level: Node Security Configuration: Tuning.
- Tuning:** Heartbeat frequency: 20 seconds, Grace period: 10 seconds.
- Other:** Synchronize file collections every: 10 minutes, Inter-Site recovery: Automatically failover, Automatically verify cluster configuration: Yes.

Figure 9-32 Cluster not configured

9.3.3 Verifying and synchronizing the cluster

To verify and synchronize the cluster, select **Availability** → **PowerHA SystemMirror** → **Cluster Management** → **Manage Cluster** → **Actions** → **Verify and Synchronize**.

Change **Correct errors found during verification** to **Yes** and click **OK** (Figure 9-33).

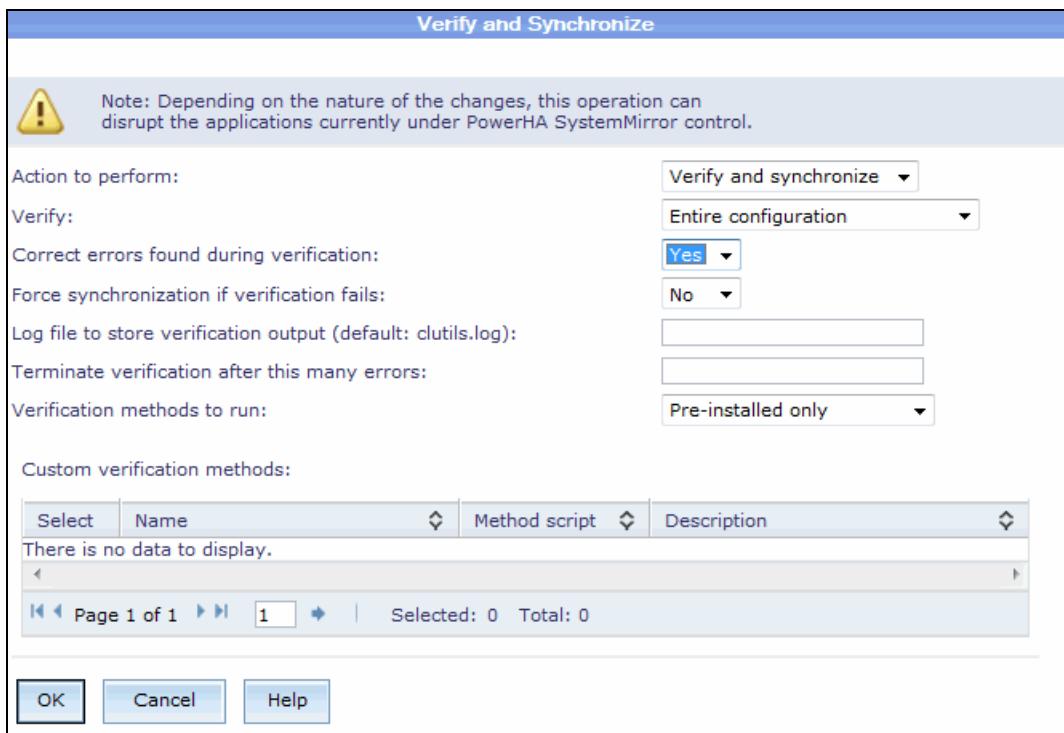


Figure 9-33 Verifying and synchronizing the cluster

The successful completion results are shown in Figure 9-34.

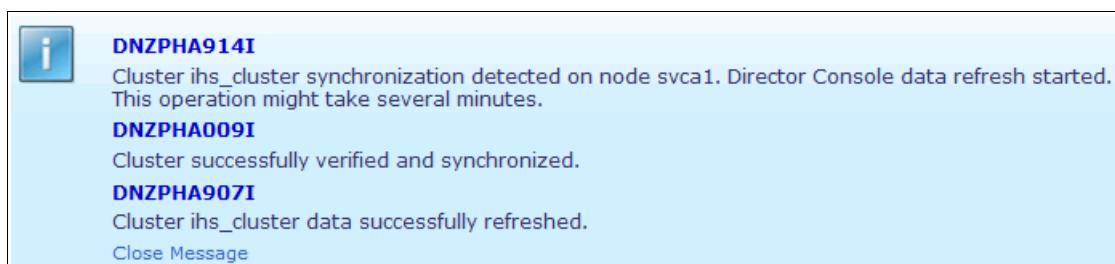


Figure 9-34 Verifying and synchronizing the cluster results

Now the cluster status appears as offline (Figure 9-35 on page 424).

Figure 9-35 Cluster configured and offline

9.3.4 Configuring resources

In the same view shown in Figure 9-35, choose the **Networks** Tab. The network configuration is automatically created by the discovery process of the cluster configuration (Figure 9-36).

Figure 9-36 Network configuration

After obtaining the information about the network, choose **Resources Groups Tab** → **Resources Tab** → **Actions** → **Create a Resource** → **Service IP Label** (Figure 9-37 on page 425) to configure the service IP to the cluster.

Figure 9-37 Adding the service IP to the cluster configuration

Choose the IP address and the correct network in the Service IP Label box (Figure 9-38). Because there are different services on each site, configure site-specific service IPs. Repeat this procedure for each service IP as needed.

Name:	svca_srv (10.10.10.10)
Network name:	net_ether_01
Netmask (IPv4) / Prefix length (IPv6):	
Associated site:	SiteA

Figure 9-38 Adding a service IP label

After all the service IPs are added, they appear as shown in Figure 9-39.

Select	Name	Type	Resource Groups
○	svca_srv	Service IP label	
○	svcb_srv	Service IP label	

Figure 9-39 Services IP configured

To configure an application controller, choose **Actions** → **Create a Resource** → **Application controller** and configure it with a start and stop script (Figure 9-40).

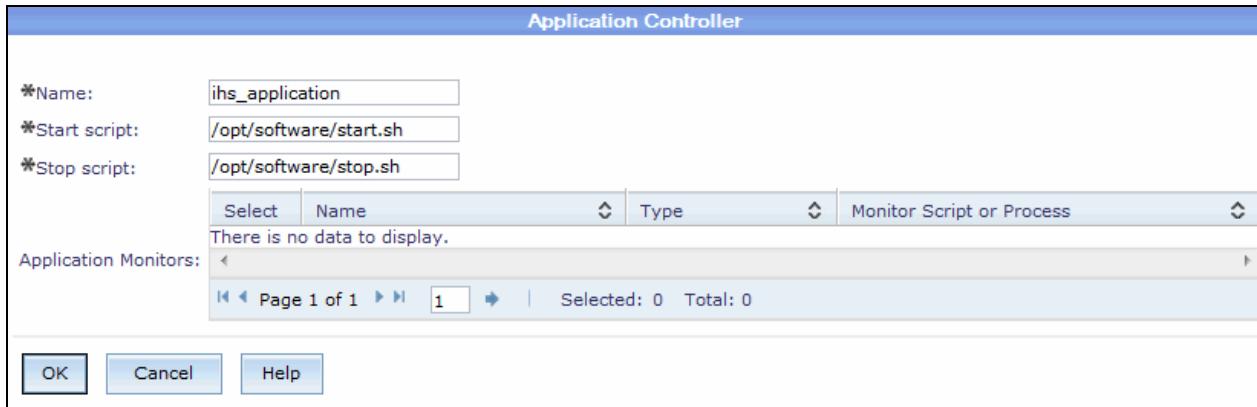


Figure 9-40 Application controller

9.3.5 Configuring replicated mirror groups

In this section we demonstrate how to configure a mirror group to control an SVC replication. Select **HyperSwap and replicated Storage** → **Create a replicated mirror group** and click **Next** on the Welcome to the Create a Replicated Mirror Group wizard screen. Then select your storage type on the Choose Storage Type Panel (Figure 9-41) and click **Next**.

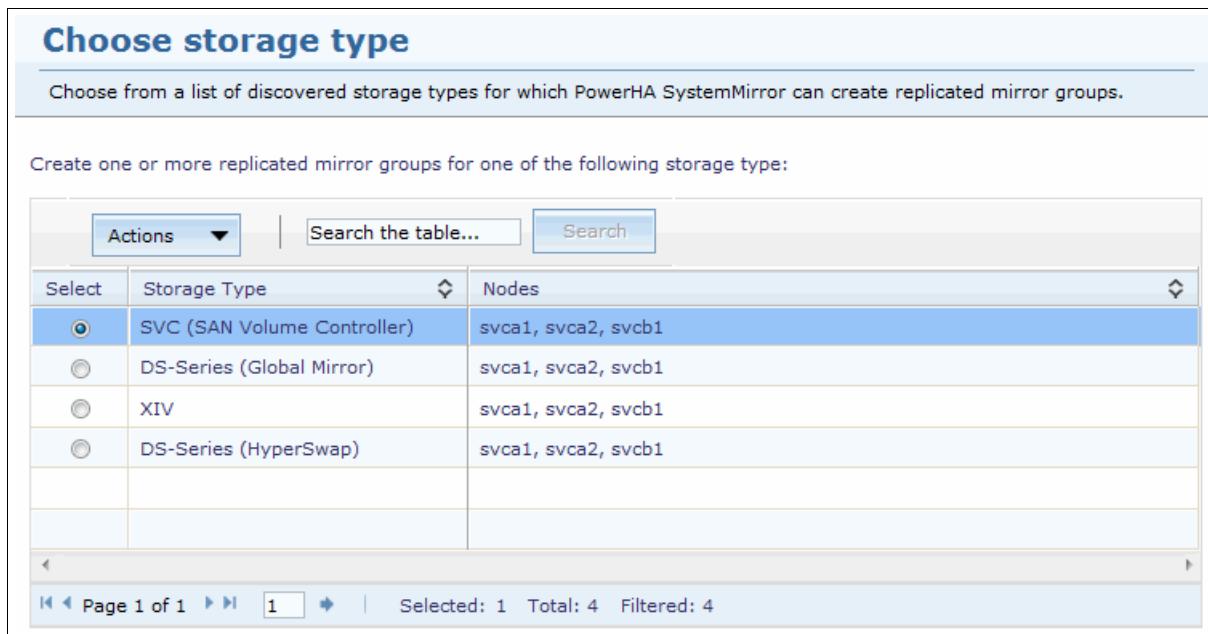


Figure 9-41 Select SVC

Fill in the required information for the first storage. Then click **Add Another**, fill in the required information for the other storage, and click **Next** (Figure 9-42 on page 427).

Configure storage systems

Provide properties for one or more new storage systems, or choose from a list of existing storage systems if storage agents were defined in the previous page of this wizard. Click the "Add another" button to add another resource of that type. Remove a storage agent and its parameters by clicking the "Remove" button.

Storage Systems represent the storage unit/controller that hosts/controls one or more disks/volumes that are being replicated.

Add another

* Name:	ITSO_SVC_CF8	Remove
* Site:	SiteA	
* IP Addresses:	9.12.5.67	
* Role:	Master	
* Remote Storage System:	POWT3V7000	

* Name:	POWT3V7000	Remove
* Site:	SiteB	
* IP Addresses:	9.12.5.6	
* Role:	Master	
* Remote Storage System:	ITSO_SVC_CF8	

Figure 9-42 Configuring storage systems

Next, configure the remote copy relationship (Figure 9-43 on page 428) by giving a name to the relationship and complete the field mirror disk pair name. Then select the source storage on the first field and the destination storage on the second field, and complete the disk ID fields with the LUN name given at the storage. Click **add another** to repeat the procedure to configure other relationships. Complete as many as needed. After all mirror relationships are added, click **Next**.

Configure mirror disk pairs

Provide properties for one or more mirror disk pairs. Click the "Add another" button to add an additional set of parameters for another storage agent. Remove a storage agent and its parameters by clicking the "Remove" button.

Mirror Disk Pairs represent the replication relationship (already established at the storage level) between two disks/volumes. One of the disks resides in the primary site and the other resides on the secondary site.

Add another

* Mirror disk pair name:	itso_siteab_rel	Remove
Mirror disks for SiteA Mirror disks for SiteB		
* Storage system name:	ITSO_SVC_CF8	POWT3V7000
* Disk ID	itso_sitea	itso_siteb

Figure 9-43 Configuring mirror pairs

Configure the Mirror Group Name (Figure 9-44) and click **Next**.

Configure replicated mirror groups

Provide properties for one or more replicated mirror groups. Click the "Add another" button to add an additional set of parameters for another replicated mirror groups. Remove a replicated mirror groups and its parameters by clicking the "Remove" button.

Add another

* Name:	itso_siteab_grp	Remove
Recovery action:	Automatic	
Mirror type:	Synchronous	
* Mirror disk pairs:	itso_siteab_rel	

Figure 9-44 Configuring replicated mirror groups

Review the configuration as shown in Figure 9-45 on page 429 and click **Finish**. Verify that **Run now** is selected in the Launch Job Panel and click **OK**.

Summary		
Review the replicated mirror group summary, and click Finish to create the replicated mirror group(s).		
itso_siteab_grp details:		
Cluster: ihs_cluster		
Storage type:		
Name: itso_siteab_grp		
Recovery action: Automatic		
Mirror type: Synchronous		
Mirror disk pairs: itso_siteab_rel		
Mirror disk pairs details:		
Mirror disk pair name: itso_siteab_rel		
Mirror disks for SiteA		
Storage system name: ITSO_SVC_CF8		
Disk ID: itso_sitea		
Mirror disks for SiteB		
Storage system name: POWT3V7000		
Disk ID: itso_siteb		
Storage systems details:		
Name: ITSO_SVC_CF8		
Site: SiteA		
IP Addresses: 9.12.5.67		
Role: Master		
Remote Storage System: POWT3V7000		
Name: POWT3V7000		
Site: SiteB		
IP Addresses: 9.12.5.6		
Role: Master		
Remote Storage System: ITSO_SVC_CF8		

Figure 9-45 Replicated mirror group summary

The previous steps create all the needed resources for PowerHA. The next tasks are to verify and synchronize the cluster again using the procedure described in 9.3.3, “Verifying and synchronizing the cluster” on page 423.

9.3.6 Create a resource group

To create the resource group, navigate to the tab **Resource Group** → **Select the cluster** and click **Actions** → **Create Resource Group** (Figure 9-46 on page 430). Then click **Next** on the welcome screen.

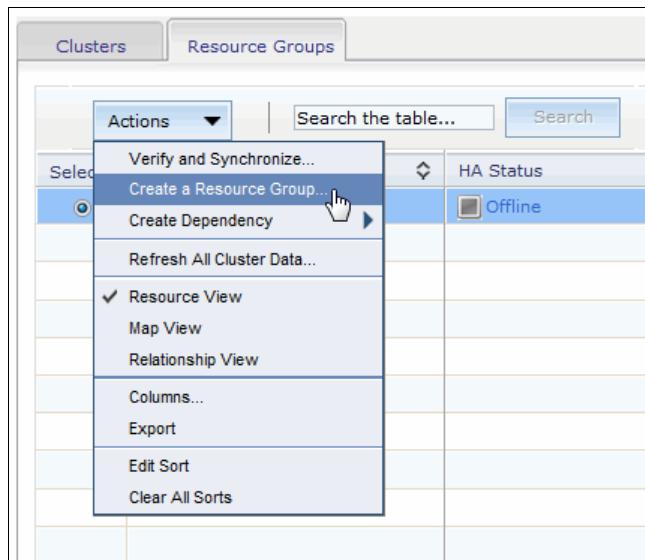


Figure 9-46 Creating a resource group

Fill in the resource group name field (Figure 9-47) and click **Next**.

Configure a resource group

Configure a resource group by selectively choosing policies and resources using a guided process, or choose to create one or more resource groups using resource group application assistants that IBM Systems Director discovers on your cluster.

Create a resource group using a guided process
 Create one or more resource groups using discovered application assistants (SmartAssist)
[Which process should I choose?](#)

*Resource group name:

Figure 9-47 Resource group name

Select nodes for both sites (Figure 9-48 on page 431) and click **Next**.

Choose nodes

Choose the nodes for which you want to configure this custom resource group.

Available site SiteA nodes:	<input type="button" value="Add >"/>	Participating site SiteA nodes (in order of priority):
<input type="button" value="< Remove"/>	svca1 svca2	<input type="button" value="Up"/> <input type="button" value="Down"/>
Available site SiteB nodes:	<input type="button" value="Add >"/>	Participating site SiteB nodes (in order of priority):
<input type="button" value="< Remove"/>	svcb1	<input type="button" value="Up"/> <input type="button" value="Down"/>

Figure 9-48 Choose nodes

Choose the policies and attributes for the resource group (Figure 9-49). These are the same options as the ones found in SMIT.

Choose policies and attributes

Choose the policies that you want to add to this custom application resource group.

Startup policy:	Online on home node only
Fallover policy:	Fallover to next priority node in the list
Dynamic node priority policy:	Next node in the list
Run file or script:	<input type="text"/>
Timeout (seconds):	<input type="text"/>
Fallback policy:	Never fallback
Fallback timer:	Immediately
Time (0:00 - 23:59):	<input type="text"/>
Day of the week:	Monday
Day of the month (1 - 31):	<input type="text"/>
Month:	January
Year:	<input type="text"/>
Inter-site management policy:	Ignore
Enable WPAR:	No
WPAR name:	<input type="text"/>

Figure 9-49 Resource group policies and attributes

Select the resources that will be part of the resource group (RG). The first tab is the service IP, so select all that apply as shown in Figure 9-50 on page 432. If you have two service IPs, as we have in our environment, click **Add Another**.

The screenshot shows the 'Service IP Label' tab selected in a software interface. There are two entries listed:

- Entry 1:** Service IP label/IP address: svca_srv, Network name: net_ether_01, Netmask (IPv4) / Prefix length (IPv6): 255.255.255.0, Associated site: SiteA.
- Entry 2:** Service IP label/IP address: svcb_srv, Network name: net_ether_01, Netmask (IPv4) / Prefix length (IPv6): 255.255.255.0, Associated site: SiteB.

Each entry has a 'Remove' button to its right.

Figure 9-50 Choosing the service IP label

If you want application monitors, use the Application Monitor tab. In our case, we skip this tab and go to the Application Controller tab.

In the Application Controller tab, we select the application controller we previously created in 9.3.4, “Configuring resources” on page 424 and we maintain the check box Keep scripts synchronized on all nodes (Figure 9-51).

The screenshot shows the 'Application Controller' tab selected. A single entry is listed:

- Application controller name: ihs_application, Start script: /opt/software/start.sh, Stop script: /opt/software/stop.sh.

A checkbox labeled 'Keep scripts synchronized on all nodes' is checked.

Below the entry is a table titled 'Application monitors' with the following columns: Actions, Name, Type, and Monitor Script or Process. The table displays the message: 'There is no data to display.'

Figure 9-51 Choose application controller

We also skip the Tape tab because we do not have tapes in our environment. Now we configure the volume group (VG) on the resource group (RG); see Figure 9-52 on page 433.

Select	Name	Type	File Systems	NFS Exports	Edit File Systems and NFS
<input checked="" type="checkbox"/>	httpvg	ORIGINAL	1 (of 1)	0 (of 1)	Edit File Systems and NFS

Version of network file system exports: NFS version 2/3 NFS version 4

Stable storage path (NFS version 4 only):

NFS mounts:

Preferred network for NFS mounts: net_ether_01

Use forced varyon: No

Figure 9-52 Choosing the volume group (VG)

NFS consideration: At the time of writing, for clusters with more than two nodes, it is not possible to add a volume group and decline to export an NFS4 file system. Also, you must configure a stable storage path. If you do not need an NFS, you always receive error messages like the one shown in Figure 9-53. This is a known problem and there is an intention from the IBM PowerHA SystemMirror development team to fix this in future software releases.

However, in our configuration we do not have an NFS to export, so we avoid this error message by simply adding the volume group to the resource group later, after the Create Resource Group Wizard completes.

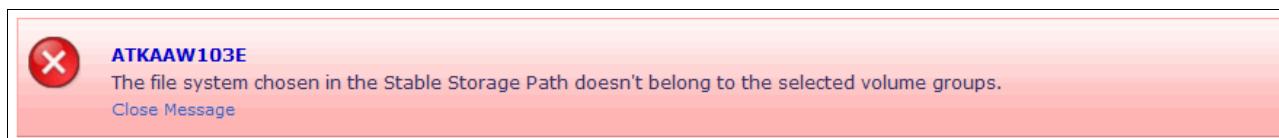


Figure 9-53 Stable storage path message

Because the Mirror Group depends on the volume group, we add both resources later using Systems Director.

We create the resource group with the service IPs and application controller as seen in Figure 9-54 on page 434. Do *not* select “Bring resource group online now.” Simply verify the configuration and click **Finish**. Verify that **Run Now** is selected on the Launch Job panel and click **OK**.

Summary

Review the resource group summary and click Finish to create the resource group.

Bring the resource groups online now

Note: Configuring resource groups can take a long time. You can specify whether to start the configuration now or schedule as a job later after you click the "Finish" button.

ihs_app_rg details:

Cluster name:	ihs_cluster
Participating site SiteA nodes (in order of priority):	svca1,svca2
Participating site SiteB nodes (in order of priority):	svcb1
Startup policy:	Online on home node only
Failover policy:	Failover to next priority node in the list
Fallback policy:	Never fallback
Inter-site management policy:	Ignore
Enable WPAR:	No
Service IP label/IP address:	svca_srv
Service IP label/IP address:	svcb_srv
Application controller name:	ihs_application

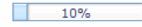
Figure 9-54 Resource group (RG) summary

Now the job status is displayed (Figure 9-55).

Active and Scheduled Jobs (Properties)

Name: Create Resource Group Task - November 25, 2012 3:14:01 PM EST Actions ▾

General Targets History Logs

Status: Active
 Progress:  10%
 Last Run Status: Running
 Description: Run once on 11/25/12 at 3:14 PM
 Next Run:
 Last Run: 11/25/12 at 3:15 PM
 Task: Create Resource Group Task
 Created By: root

[Edit](#)

Figure 9-55 Job running

When the resource group is configured on the controller node, the status is changed to complete (Figure 9-56 on page 435).

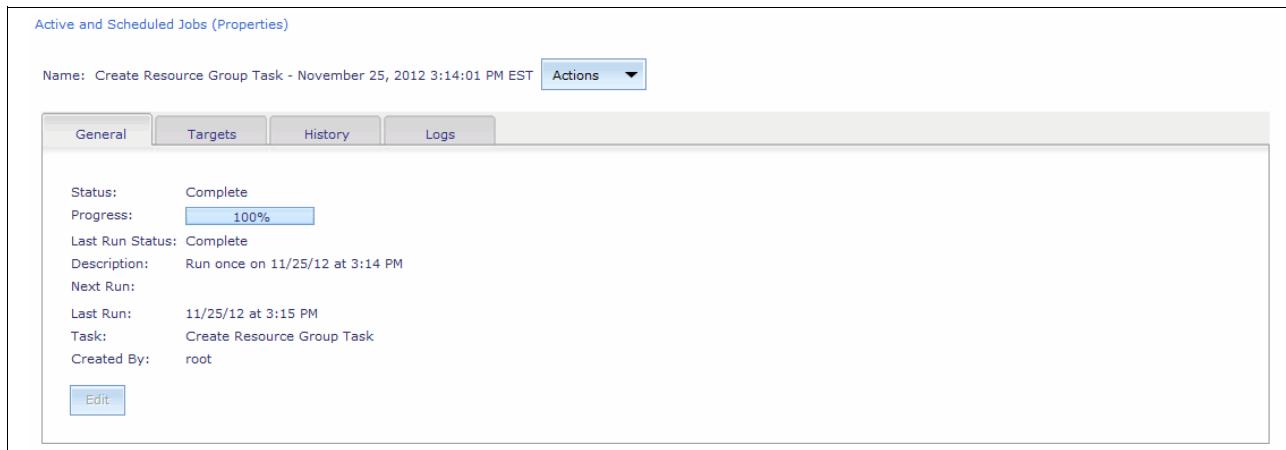


Figure 9-56 Job completed

Now verify and synchronize the cluster again, as demonstrated in 9.3.3, “Verifying and synchronizing the cluster” on page 423.

The next task is to add the volume group to the resource group. First, select the RG as shown in Figure 9-57.

Cluster and Resource Group Management			
		Clusters	Resource Groups
		Actions	Search the table... Search
Select	Name	HA Status	
<input type="radio"/>	ihs_cluster	<input type="checkbox"/> Offline	
<input checked="" type="radio"/>	ihs_app_rg	<input type="checkbox"/> Unknown	

Figure 9-57 Selecting the resource group (RG)

Then click **Actions** → **Add Storage** → **Add a volume group** (Figure 9-58 on page 436).

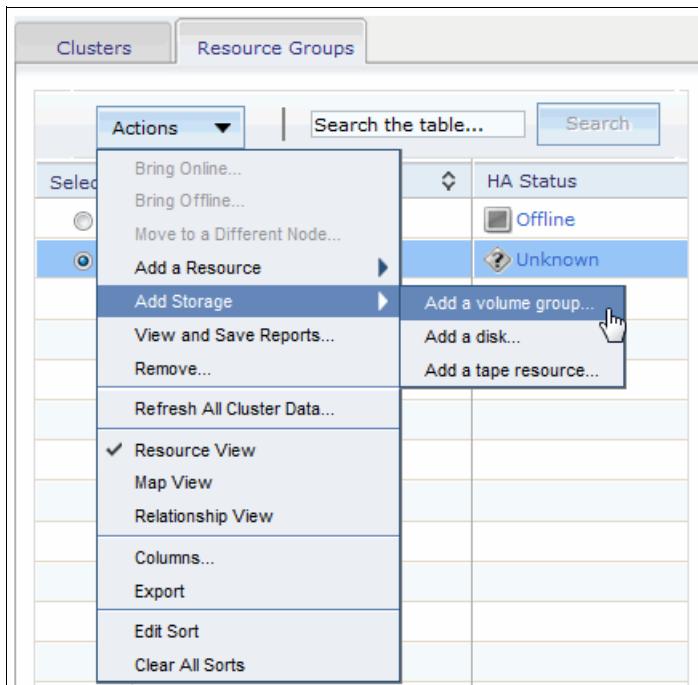


Figure 9-58 Add a volume group menu

Next, select the volume group to add on the resource group tab (Figure 9-59) and click **OK**.



Figure 9-59 Selecting the volume group

The completion message appears in Systems Director (Figure 9-60).

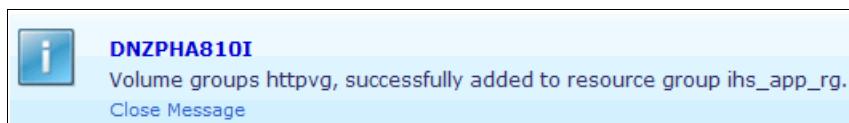


Figure 9-60 Completion message

To add the replicated mirror group to the resource group, select the Replicated Storage tab and click **Add a replicated mirror group** (Figure 9-61 on page 437).

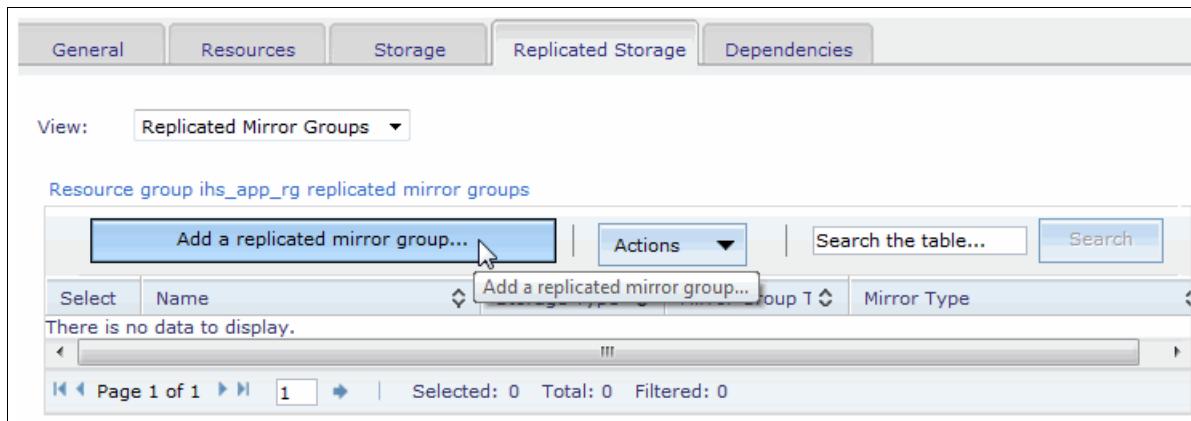


Figure 9-61 Add a replicated mirror group

Select the mirror group to add to the resource group and click **OK** (Figure 9-62).



Figure 9-62 Select mirror group

Upon synchronizing the cluster again, the resource group is ready to use.

9.4 Administering a cluster

IBM Systems Director helps manage clusters using a graphical user interface. In this section, we demonstrate how to:

- ▶ Start and stop cluster services
- ▶ Create a cluster snapshot
- ▶ Move resource groups between nodes or sites

9.4.1 Bringing cluster services online and offline

To bring the cluster services online, click **Action** under the Clusters tab and select the option **Bring Cluster Services Online** (Figure 9-63 on page 438).

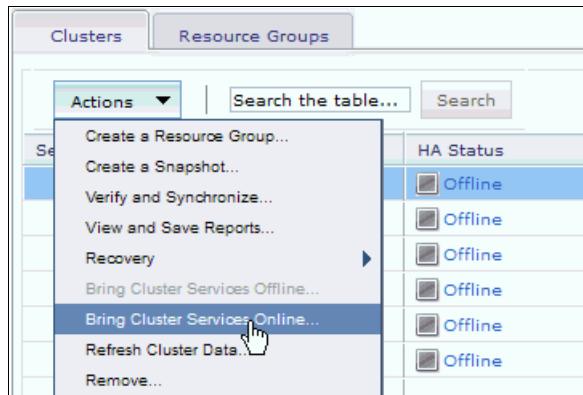


Figure 9-63 Bringing the cluster services online

This action opens another window to select the nodes on which to start the cluster services and options for manage resource groups (Figure 9-64).

Select all nodes and click **OK**.

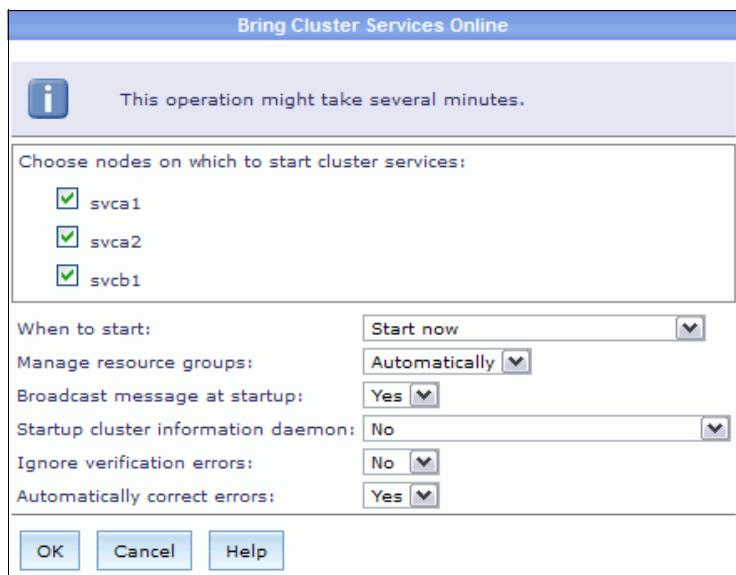


Figure 9-64 Starting cluster services on the nodes

The cluster, sites and nodes will be online after the cluster services on all nodes starts successfully (Figure 9-65 on page 439).

Select	Name	HA Status
<input checked="" type="radio"/>	ihs_cluster	OK
<input type="radio"/>	SiteA	OK
<input type="radio"/>	svca1	OK
<input type="radio"/>	svca2	OK
<input type="radio"/>	SiteB	OK
<input type="radio"/>	svcb1	OK

Figure 9-65 Cluster online

To bring the cluster services offline, click **Action** under the Clusters tab and select the option **Bringing Cluster Services Offline** (Figure 9-66).

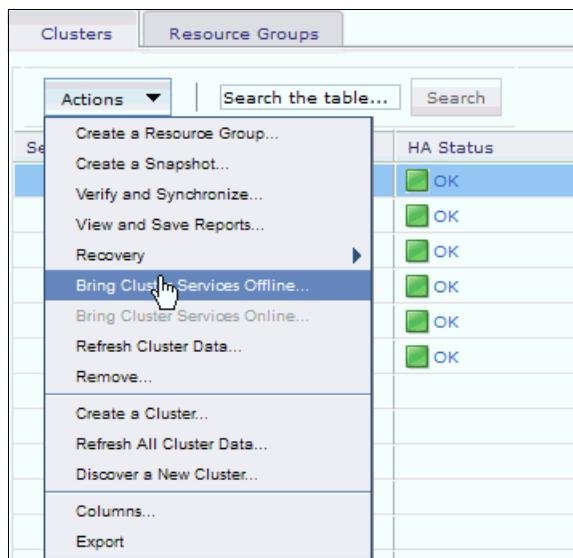


Figure 9-66 Bringing the cluster services offline

This action opens another window allowing you to choose to select either all nodes or the desired node (Figure 9-67 on page 440).

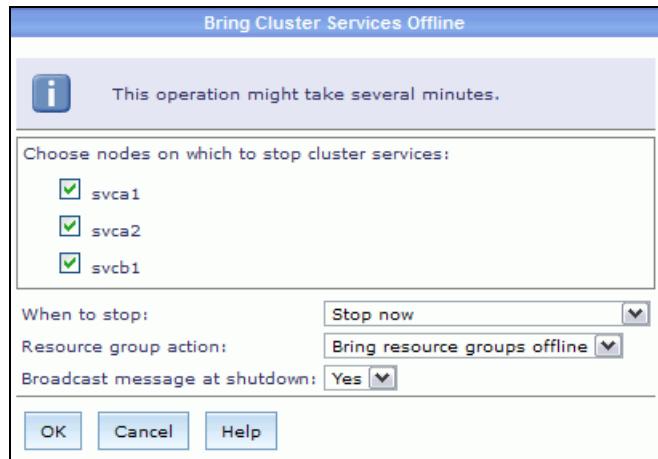


Figure 9-67 Stopping cluster services on nodes

Figure 9-68 displays the cluster, sites, and nodes in the offline state.

Clusters		
	Name	HA Status
<input checked="" type="radio"/>	ihs_cluster	<input type="checkbox"/> Offline
<input type="radio"/>	SiteA	<input type="checkbox"/> Offline
<input type="radio"/>	svca1	<input type="checkbox"/> Offline
<input type="radio"/>	svca2	<input type="checkbox"/> Offline
<input type="radio"/>	SiteB	<input type="checkbox"/> Offline
<input type="radio"/>	svcb1	<input type="checkbox"/> Offline

Figure 9-68 Cluster offline

9.4.2 Cluster snapshot

Login to IBM Systems Director and click **PowerHA SystemMirror** under the Availability tab. This provides the PowerHA SystemMirror health summary, cluster management, and resource group management options.

Under cluster management, click **Manage cluster** to view the cluster and resource groups tabs. Click **Actions** → **Create a snapshot** (Figure 9-69 on page 441).

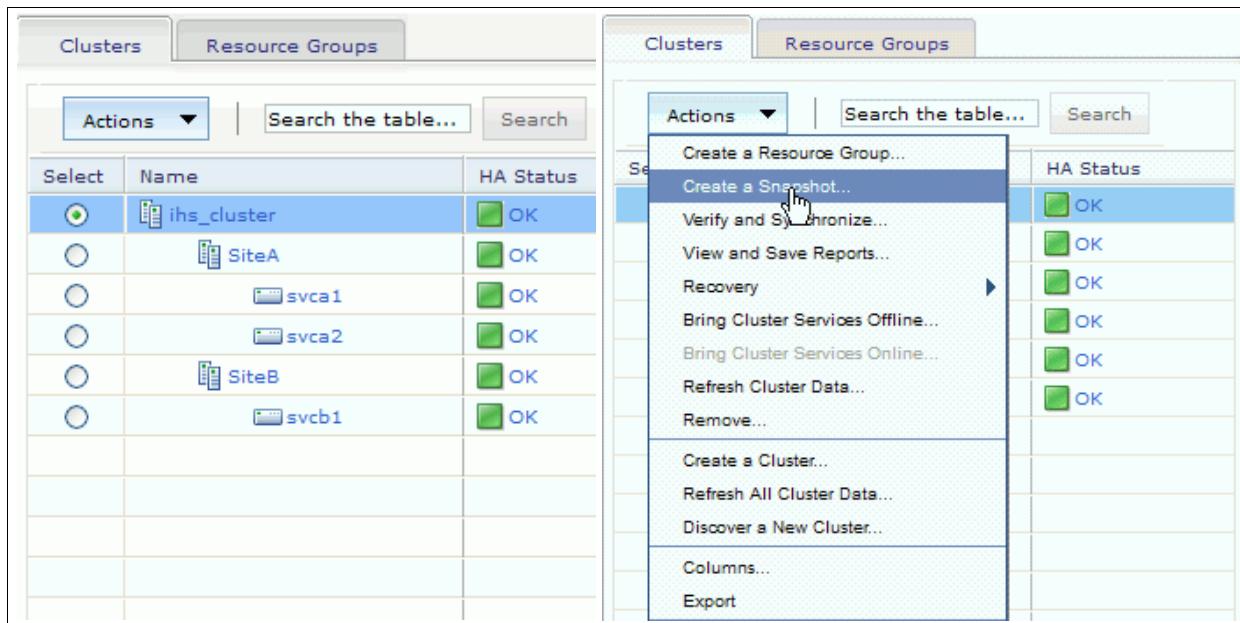


Figure 9-69 Managing the cluster - snapshot creation

The action opens a window where you provide the Name and the Description to create a snapshot (Figure 9-70). After providing this information, click **OK**.



Figure 9-70 Name and description for snapshot

After successful creation of the snapshot, a message is displayed(Figure 9-71).



Figure 9-71 Successful creation of snapshot

The snapshot is stored in the controller node in the path /usr/es/sbin/cluster/snapshots. If a controller node is down, the IBM Systems Director reassigns the role to another node.

Figure 9-72 on page 442 displays the snapshot backup list.

General	Topology	Networks	Snapshots	Federated Security	Reports and Logs	Events and Alerts	Other
View:	Snapshots						
Cluster ihs_cluster snapshots							
Create a Snapshot... Actions <input type="text" value="Search the table..."/> <input type="button" value="Search"/>							
Select	Name	Date captured	Hosting node	Description			
<input checked="" type="radio"/>	 active.0	Dec 05 22:47	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 active.1	Dec 05 20:45	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 active.2	Nov 30 20:02	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 active.3	Nov 23 19:27	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 active.4	Nov 11 20:55	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 active.5	Nov 11 20:51	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 active.6	Nov 11 20:44	svca1	Cluster Snapshot of active cluster configuration			
<input checked="" type="radio"/>	 ihs_cluster_23_Nov_2012_14_44	Dec 06 22:37	svca1	snapshot backup			
◀ ◀ Page 1 of 1 ▶ ▶ <input type="button" value="1"/> Selected: 0 Total: 8 Filtered: 8							

Figure 9-72 Snapshot backup

Restoring snapshots: Restoring snapshots with Systems Director is not possible with the cluster definition removed. It is also not possible to create a cluster from a snapshot.

Use SMIT to restore snapshots.

When creating a snapshot, both the snapshot itself and a report containing the cluster configuration is created in /usr/es/sbin/cluster/snapshots. The commands issued during the snapshot creation process are shown in Example 9-10.

Example 9-10 Commands on the report file

```
root@svca1:/>grep COMMAND:
/usr/es/sbin/cluster/snapshots/ihc_cluster_23_Nov_2012_14_44.info
COMMAND: cl1scf
COMMAND: cl1snw
COMMAND: cl1sif
COMMAND: clshowres
COMMAND: /usr/bin/hostname
COMMAND: /usr/bin/host svca1
COMMAND: netstat -i
COMMAND: netstat -in
COMMAND: no -a
COMMAND: clchsynccd
COMMAND: lsdev -Cc if
COMMAND: lsdev -Cc disk
COMMAND: lsvg
COMMAND: lspv
COMMAND: lsrvserver -H
COMMAND: lsrvclient -H
COMMAND: df
COMMAND: mount
COMMAND: exportfs
```

```
COMMAND: lsfs
COMMAND: lslpp -h "cluster.*"
COMMAND: ifconfig en0
COMMAND: ifconfig en1
COMMAND: ifconfig en2
COMMAND: ifconfig lo0
COMMAND: lsvg -l httpvg
COMMAND: lsvg -l caavg_private
COMMAND: lsvg -l rootvg
COMMAND: lslv -l apploglv
COMMAND: lslv -l applv
COMMAND: lslv -l caalv_private1
COMMAND: lslv -l caalv_private2
COMMAND: lslv -l caalv_private3
COMMAND: lslv -l powerha_crlv
COMMAND: lslv -l hd5
COMMAND: lslv -l hd6
COMMAND: lslv -l hd8
COMMAND: lslv -l hd4
COMMAND: lslv -l hd2
COMMAND: lslv -l hd9var
COMMAND: lslv -l hd3
COMMAND: lslv -l hd1
COMMAND: lslv -l hd10opt
COMMAND: lslv -l hd11admin
COMMAND: lslv -l lg_dumplv
COMMAND: lslv -l livedump
COMMAND: odmget HACMPdaemons
COMMAND: odmget HACMPsp2
COMMAND: /usr/bin/hostname
COMMAND: /usr/bin/host svca2
COMMAND: netstat -i
COMMAND: netstat -in
COMMAND: no -a
COMMAND: clchsynccd
COMMAND: lsdev -Cc if
COMMAND: lsdev -Cc disk
COMMAND: lsvg
COMMAND: lspv
COMMAND: lsrvpserver -H
COMMAND: lsrvpclient -H
COMMAND: df
COMMAND: mount
COMMAND: exportfs
COMMAND: lsfs
COMMAND: lslpp -h "cluster.*"
COMMAND: ifconfig en0
COMMAND: ifconfig en1
COMMAND: ifconfig en2
COMMAND: ifconfig lo0
COMMAND: lsvg -l caavg_private
COMMAND: lsvg -l rootvg
COMMAND: lslv -l caalv_private1
COMMAND: lslv -l caalv_private2
COMMAND: lslv -l caalv_private3
```

```
COMMAND: lslv -l powerha_crlv
COMMAND: lslv -l hd5
COMMAND: lslv -l hd6
COMMAND: lslv -l hd8
COMMAND: lslv -l hd4
COMMAND: lslv -l hd2
COMMAND: lslv -l hd9var
COMMAND: lslv -l hd3
COMMAND: lslv -l hd1
COMMAND: lslv -l hd10opt
COMMAND: lslv -l hd11admin
COMMAND: lslv -l lg_dumplv
COMMAND: lslv -l livedump
COMMAND: odmget HACMPdaemons
COMMAND: odmget HACMPsp2
COMMAND: /usr/bin/hostname
COMMAND: /usr/bin/host svcb1
COMMAND: netstat -i
COMMAND: netstat -in
COMMAND: no -a
COMMAND: clchsynccd
COMMAND: lsdev -Cc if
COMMAND: lsdev -Cc disk
COMMAND: lsvg
COMMAND: lspv
COMMAND: lsrvserver -H
COMMAND: lsrvclient -H
COMMAND: df
COMMAND: mount
COMMAND: exportfs
COMMAND: lsfs
COMMAND: lspp -h "cluster.*"
COMMAND: ifconfig en0
COMMAND: ifconfig en1
COMMAND: ifconfig en2
COMMAND: ifconfig lo0
COMMAND: lsvg -l caavg_private
COMMAND: lsvg -l rootvg
COMMAND: lslv -l caalv_private1
COMMAND: lslv -l caalv_private2
COMMAND: lslv -l caalv_private3
COMMAND: lslv -l powerha_crlv
COMMAND: lslv -l hd5
COMMAND: lslv -l hd6
COMMAND: lslv -l hd8
COMMAND: lslv -l hd4
COMMAND: lslv -l hd2
COMMAND: lslv -l hd9var
COMMAND: lslv -l hd3
COMMAND: lslv -l hd1
COMMAND: lslv -l hd10opt
COMMAND: lslv -l hd11admin
COMMAND: lslv -l lg_dumplv
COMMAND: lslv -l livedump
COMMAND: odmget HACMPdaemons
```

COMMAND: odmget HACMPsp2

9.4.3 Moving cluster resource groups

This section contains information that explains how you can move a resource group while bringing it online or offline on certain cluster nodes.

In our example, we have two nodes (svca1 and svca2) on SiteA. We have one node (svcb1) on SiteB.

Node svca1 has the IBM HTTP Server application running with a service IP and volume group (Example 9-11).

Example 9-11 Service IP and volume group on svca1

```
root@svca1:/>ifconfig -a
en0: flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GRUO
T,64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 192.168.100.60 netmask 0xfffffc00 broadcast 192.168.103.255
        tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en1: flags=1e080863,10480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROU
PRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 10.10.10.10 netmask 0xffffffff00 broadcast 10.10.10.255
    inet 10.10.10.11 netmask 0xffffffff00 broadcast 10.10.10.255
        tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en2: flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GRUO
T,64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 9.12.5.7 netmask 0xfffffff000 broadcast 9.12.15.255
        tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0: flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GRUOPRT,64
BIT,LARGESEND,CHAIN>
    inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
    inet6 ::1%1/0
        tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1

root@svca1:/>lspv
hdisk0      00f70c99ec834c85          rootvg      active
hdisk1      00f70c99ed029db8          rootvg      active
hdisk2      00f70c99218b2001         caavg_private active
hdisk3      00f70c992185bde3         httpvg      concurrent

root@svca1:/>lsvg -l httpvg
httpvg:
LV NAME      TYPE     LPs     PPs     PVs   LV STATE      MOUNT POINT
applog1v     jfs2log    20      20      1    open/syncd    N/A
applv       jfs2       40      40      1    open/syncd    /httpapp
```

To move the resource groups from node svca1 to node svca2, click the resource groups tab, then click **Action** and select the option **move to a different node**. This action opens another window and where you are asked for confirmation to move the resource group to svca2 (Figure 9-73 on page 446).

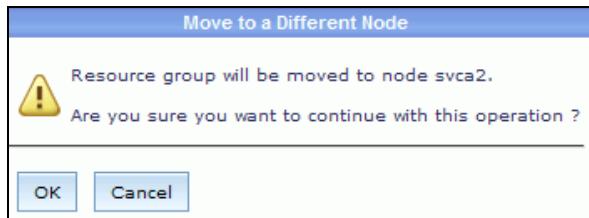


Figure 9-73 Resource group to svca2

Example 9-12 shows the resource group has moved from svca1 to svca2.

Example 9-12 Resource group from svca1 to svca2

```
root@svca2:/>ifconfig -a
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 192.168.100.61 netmask 0xfffffc00 broadcast 192.168.103.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en1:
flags=1e080863,10480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64B
IT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 10.10.10.10 netmask 0xffffffff00 broadcast 10.10.10.255
        inet 10.10.10.12 netmask 0xffffffff00 broadcast 10.10.10.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en2:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 9.12.5.8 netmask 0xfffffff000 broadcast 9.12.15.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0:
flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,LAR
GESEND,CHAIN>
        inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
        inet6 ::1%1/0
                tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
root@svca2:/>lspv
hdisk0          00f70c99ec83d574                  rootvg      active
hdisk1          00f70c99ed030959                  rootvg      active
hdisk2          00f70c99218b2001                caavg_private active
hdisk3          00f70c992185bde3                httpvg      concurrent
root@svca2:/>lsvg -l httpvg
httpvg:
LV NAME        TYPE     LPs    PPs    PVs   LV STATE      MOUNT POINT
applog1v       jfs2log   20     20     1    open/syncd    N/A
applv          jfs2      40     40     1    open/syncd    /httpapp
root@svca2:/>
```

Resource group move consideration: After a successful resource group (`ihs_app_rg`) move from svca1 to svca2, the RG appears to be offline because it turns gray on the Systems Director GUI, but in fact it is working.

To bring it online, thus turning it green, navigate to the Clusters tab and click **Action**. Then select **refresh clusters data**.

Using the option Move Resource Groups to Another Node: From the IBM Systems Director, there is no option to move the resource group to svcb1 or another site. It works only from smitty when you use the Move Resource Groups to Another Node option and select **svcb1**.

Using the Move Resource Groups to Another Site option did not work.

Example 9-13 shows the movement of the resource group from svca2 to svcb1.

Example 9-13 Resource group from svca2 to svcb1

```
root@svcb1:/>ifconfig -a
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 192.168.100.62 netmask 0xfffffc00 broadcast 192.168.103.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en1:
flags=1e080863,10480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64B
IT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 10.10.20.10 netmask 0xffffffff00 broadcast 10.10.20.255
        inet 10.10.20.21 netmask 0xffffffff00 broadcast 10.10.20.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en2:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 9.12.5.9 netmask 0xfffffff000 broadcast 9.12.15.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0:
flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,LAR
GESEND,CHAIN>
        inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
        inet6 ::1%1/0
                tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
root@svcb1:/>lspv
hdisk0          00f6f5d0ec83cf8e                  rootvg      active
hdisk1          00f6f5d0ed030e9b                  rootvg      active
hdisk2          00f70c99218b2001                 caavg_private active
hdisk3          00f70c992185bde3                 httpvg      concurrent

root@svcb1:/>lsvg -l httpvg
httpvg:
LV NAME          TYPE     LPs    PPs    PVs   LV STATE      MOUNT POINT
applog1v         jfs2log   20     20     1    open/syncd    N/A
applv            jfs2      40     40     1    open/syncd    /httpapp
root@svcb1:/>
```

Service IP consideration: The service IP on SiteA was **10.10.10.10**. When the resource group moved to SiteB, the service IP changed to **10.10.20.10**, due to the cluster configuration for SiteB and the use of Site Specific Service IP labels.

To make this change dynamically available to the IBM HTTP Server application, the DNS record must be updated with the change of IP. This action is explained in Appendix B, “DNS change for the IBM Systems Director environment with PoweHA” on page 495.



Cluster partition management

This chapter contains information about the following topics for cluster partition management on PowerHA SystemMirror 7.1.2 Enterprise Edition (EE):

- ▶ Cluster partitioning
- ▶ Methods to avoid cluster partitioning
- ▶ Planning to avoid cluster partitioning
- ▶ Detailed behavior of cluster partitioning

10.1 Cluster partitioning

This section describes cluster partitioning considerations.

The terms *cluster partitioning*, *split-brain*, and *node isolation* all refer to a situation where more than one cluster node activates resources as though it were the primary node. Such a situation can occur in the following scenarios:

- ▶ Failure of all links between sites
- ▶ Multiple failures within a site
 - Requires failures of Ethernet, SAN, and repository access

Although cluster partitioning can occur within a site, cluster partitioning between sites is considered to be of a higher probability because heartbeating relies on IP network connectivity.

Figure 10-1 illustrates an overview of cluster partitioning between sites.

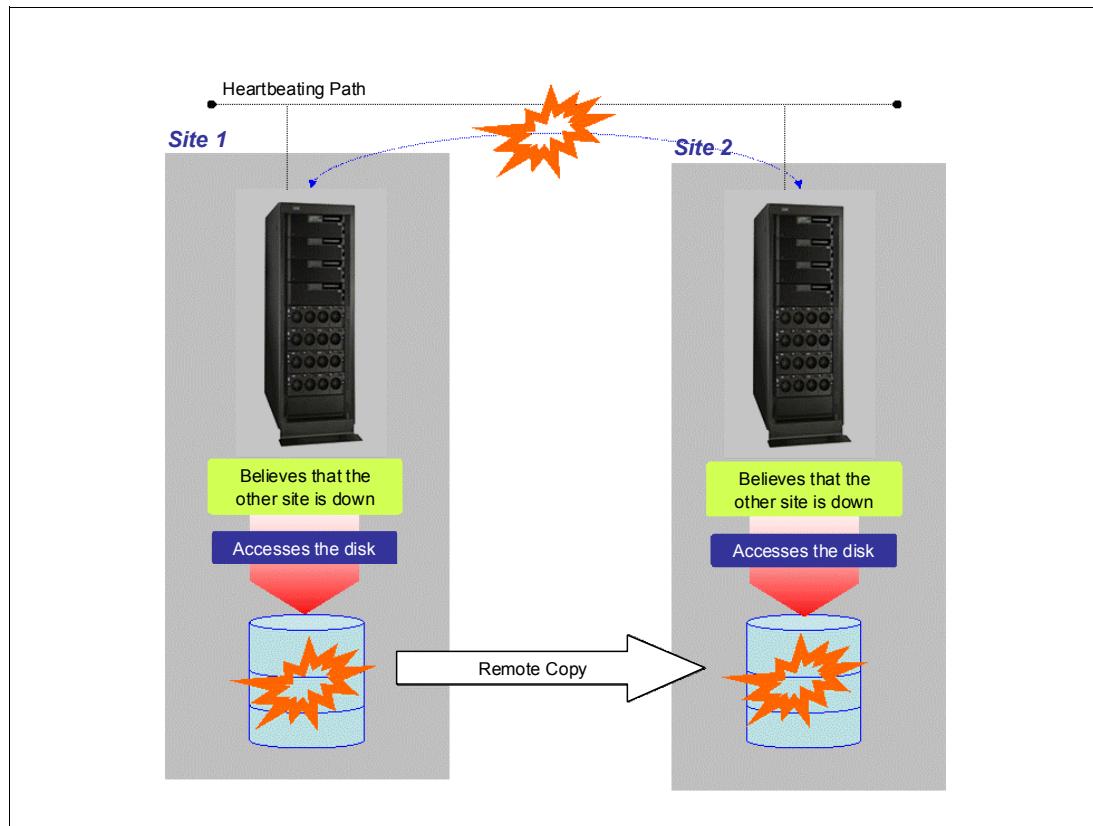


Figure 10-1 Cluster partitioning

IBM PowerHA defines the following terminologies for these specific situations:

Split	When communication is lost between the cluster nodes, and each site believes that it is the only one still online.
Merge	When the split partitions in the cluster attempt to reform as a single cluster because the links between them have been restored.

When utilizing sites with a PowerHA cluster, avoid unplanned split events. Not only can such events lead to undesirable results, but data divergence is also a risk. Both sites might bring

the resources online, leading to each site accessing its own local disks and thus creating inconsistent copies. Manual intervention is required to recover from this situation, but it might be difficult to perform.

In general, cluster partitions are not detected until a merge event occurs. Although previous versions of PowerHA implemented safeguard functions for merge events, during the split there is no function to prevent data divergence.

10.2 Methods to avoid cluster partitioning

A partitioned cluster is one of the worst scenarios that can occur in a clustered environment. This condition can be dangerous because each node can independently run the applications, and can acquire the data from their own storage copies. If this occurs you risk losing access to the disks, and potentially experiencing data divergence.

The basic way to prevent partitioning between sites is to define multiple IP networks between the sites. Having multiple network communication paths between the sites minimizes the risk of a false failover. To achieve true redundancy, the networks must be backed by a separate network infrastructure.

If that is not possible, there might be no benefit to having a separate network defined. Instead, the multiple interfaces can be aggregated to form a single redundant logical interface, which adds redundancy to the communication interface.

However, even with the method previously mentioned, we cannot prevent cluster partitions from total loss of IP connectivity between sites.

IBM PowerHA SystemMirror 7.1.2 Enterprise Edition now provides a new function: the tie breaker disk. This new function allows PowerHA to decide which partition of the cluster should survive in case of split events.

Tie breaker disk overview

You can use the tie breaker option to specify a SCSI-3 Persistent Reserve (PR)-capable disk that is used by the split and merge policies.

A tie breaker disk is used when a group of nodes in a cluster cannot communicate with each other. This communication failure results in the cluster splitting the nodes into two or more partitions. If failure occurs because the cluster communication links are not responding, both partitions attempt to lock the tie breaker disk. The partition that acquires the tie breaker disk continues to function, while the other partition reboots.

Tie breaker accessibility considerations: The disk that is identified as the tie breaker must be accessible to all nodes in the cluster.

When partitions that were part of the cluster are brought back online after the communication failure, they must be able to communicate with the partition that owns the tie breaker disk. If a partition that is brought back online cannot communicate with the tie breaker disk, it does not join the cluster. The tie breaker disk is released when all nodes in the configuration rejoin the cluster.

10.3 Planning to avoid cluster partitioning

This section discusses how to plan to avoid cluster partitioning.

Stretched cluster versus linked cluster

Although split and merge can happen in either stretched clusters or linked clusters, a linked cluster is considered crucial because it usually involves data replication.

Adding a tie breaker disk to a stretched cluster is, in general, expected to be not that beneficial. This is based on the assessment that partitions are more likely in a configuration with additional heartbeat paths through SAN or the repository disk.

In comparison, the risk of total loss of IP connectivity between sites is higher in a linked cluster. In such configurations, using a tie breaker disk configuration is suggested to prevent cluster partitions when a split event occurs.

Split handling policy

The split handling policy attribute describes the type of handling performed by PowerHA when a split occurs within a cluster. The possible choices for this attribute are described here:

None: (Default) With this option, each partition that is created by the cluster split event becomes an independent cluster. Each partition can start a workload independent of the other partition. This option is the default setting.

Note that, for linked clusters, do *not* use this option if your environment is configured to use HyperSwap for PowerHA SystemMirror.

TieBreaker With this option, each partition attempts to acquire the tie breaker by placing a lock on the tie breaker disk. The tie breaker is a SCSI disk that is accessible to all nodes in the cluster. The partition that cannot lock the disk is rebooted.

Note that if you use this option, the merge policy configuration must also use the tie breaker option.

Merge handling policy

The merge handling policy attribute controls the behavior of the PowerHA cluster when the cluster partitions or splits, and later, when the cluster partitions are attempting to merge. The possible choices for this attribute are described here:

Majority (Default) With this option, the partition with the highest number of nodes remains online. If each partition has the same number of nodes, then the partition that has the shortest node ID is chosen. The partition that does not remain online is rebooted.

This is the default policy, and it is identical to that provided by PowerHA 6.1 and prior releases.

TieBreaker With this option, each partition attempts to acquire the tie breaker by placing a lock on the tie breaker disk. The tie breaker is a SCSI disk that is accessible to all nodes in the cluster. The partition that cannot lock the disk is rebooted.

Note that if you use this option, your split policy configuration must also use the tie breaker option.

Important: The current release only supports the following combinations:

- ▶ Split Policy: None, Merge Policy: Majority
- ▶ Split Policy: TieBreaker, Merge Policy: TieBreaker

Split and merge action plan

This policy setting describes what action is taken on the nodes in the losing partitions when the split or merge happens. The only configurable method in the current release is listed here:

Reboot (Default) In this case, the AIX operating system is rebooted.

Tie breaker disk requirements

Note the following requirements for tie breaker disks:

- ▶ A disk supporting SCSI-3 persistent reserve that is accessible by all nodes.
- ▶ The repository cannot be used as a tie breaker.

You can verify whether the disks are SCSI-3 persistent reserve-capable with the `lsattr -Rl hdiskX -a reserve_policy` command. The *PR_exclusive* value should appear as shown in Example 10-1.

Example 10-1 Checking reserve capability of the disks

```
# lsattr -Rl hdisk9 -a reserve_policy
no_reserve
single_path
PR_exclusive
PR_shared
```

10.4 Detailed behavior of cluster partitioning

In this section, we discuss the detailed behavior of cluster partitioning through actual testing.

10.4.1 Test environment overview

This scenario used a three-node cluster that was configured with IPv6 and IPv4 (dual stack configuration). The cluster name is *glvma1_cluster*. The cluster consisted of one node on siteA, and two nodes on siteB.

Two networks were configured: an XD_data network and an Ether network. The XD_data network was connected with IPv4 addresses. The Ether network was connected with IPv6 addresses.

The disks were replicated through a synchronous GLVM function. The tie breaker disk was configured as a Fibre Channel-attached DS3400 device that was accessible from both sites. Figure 10-2 shows the overview of the cluster configuration.

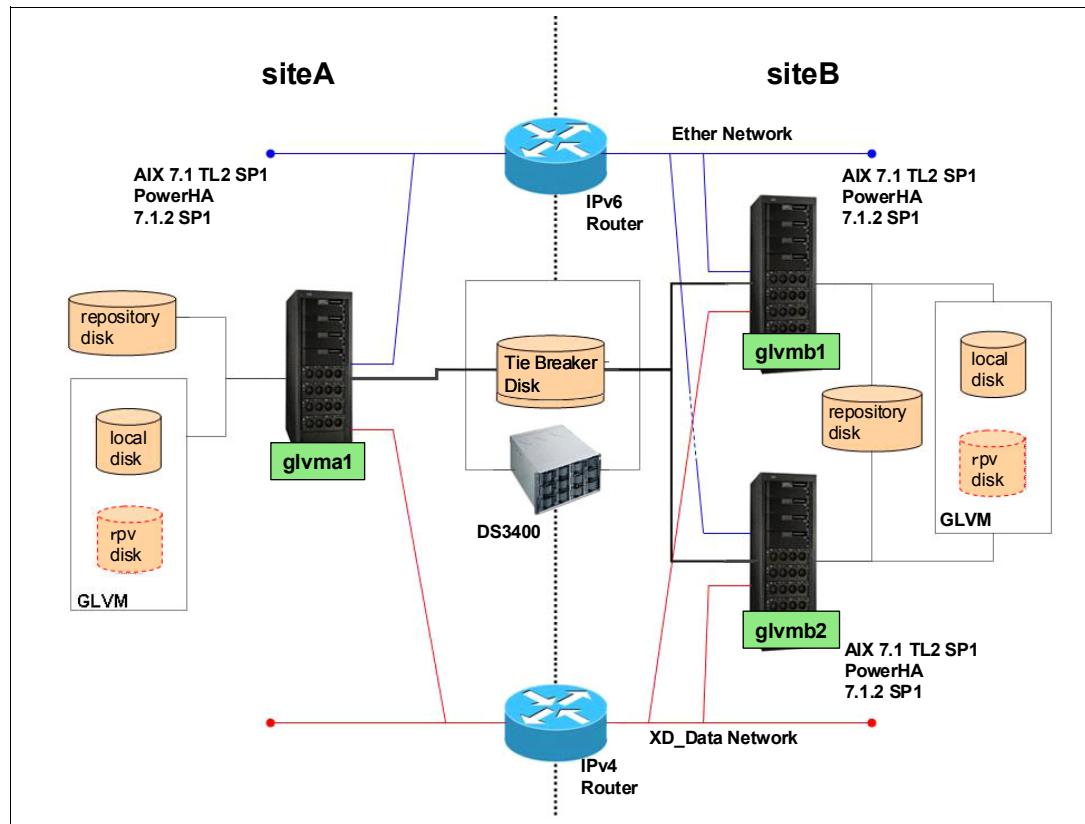


Figure 10-2 Test environment overview

A routed network was configured between the two sites. To test the cluster partitioning, we shut down the routers (Figure 10-3). This isolated the two sites, initiating the split event.

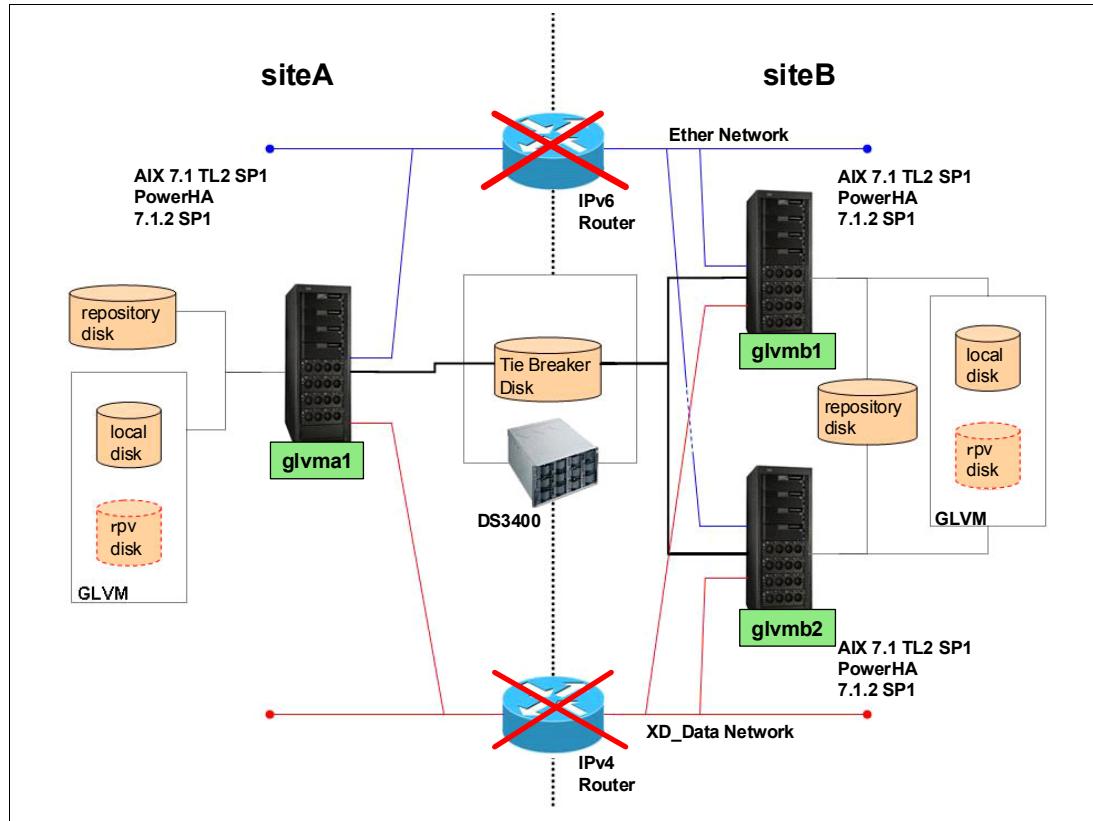


Figure 10-3 Disabling the routers

Next, we reactivated the routers for the two sites to re-establish the IP link. This initiated the merge event.

10.4.2 Configuring the split and merge policy

To configure the split and merge policy, we used `smitty sysmirror` → **Custom Cluster Configuration** → **Cluster Nodes and Networks** → **Initial Cluster Setup (Custom)** → **Configure Cluster Split and Merge Policy**.

Figure 10-4 shows the corresponding SMIT panel of the window.

Configure Cluster Split and Merge Policy for a Linked Cluster		
Type or select values in entry fields. Press Enter AFTER making all desired changes.		
[Entry Fields]		
Split Handling Policy	None	+
Merge Handling Policy	Majority	+
Split and Merge Action Plan	Reboot	+
Select Tie Breaker		+

Figure 10-4 SMIT panel for split and merge policy

To configure the tie breaker disks, we pressed **F4** while on the Select Tie Breaker menu. A list of available disks that can be configured as the tie breaker disk was displayed (Figure 10-5).

For information about how to get your disk to show up in this list, see “Tie breaker disk requirements” on page 453.

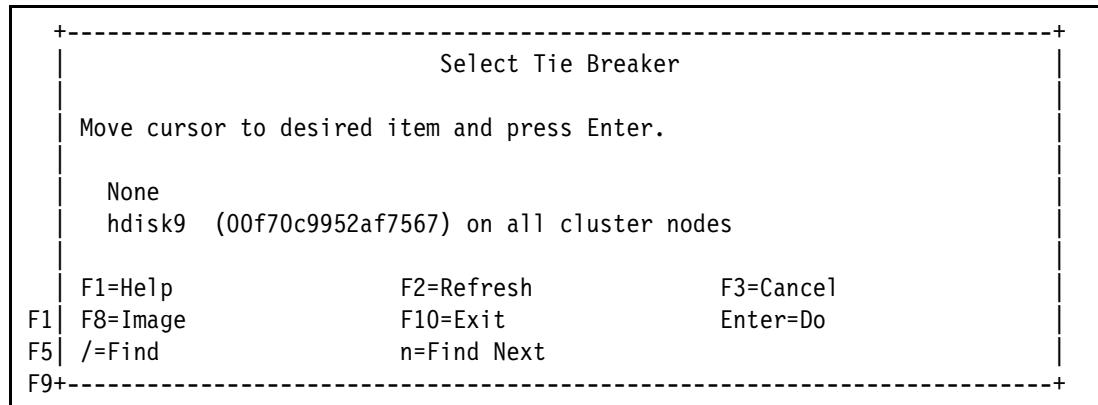


Figure 10-5 Tie breaker disk selection

To confirm this setting, we checked the HACMPsplitmerge ODM database (Example 10-2).

Example 10-2 HACMPsplitmerge ODM database

```
# odmget HACMPsplitmerge

HACMPsplitmerge:
    id = 0
    policy = "action"
    value = "Reboot"

HACMPsplitmerge:
    id = 0
    policy = "split"
    value = "TieBreaker"

HACMPsplitmerge:
    id = 0
    policy = "merge"
    value = "TieBreaker"

HACMPsplitmerge:
    id = 0
    policy = "tiebreaker"
    value = "00f70c9952af7567"
```

10.4.3 Split policy: None, merge policy: Majority

Our first test was performed using the configuration shown in Figure 10-6.

Configure Cluster Split and Merge Policy for a Linked Cluster

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

[Entry Fields]		
Split Handling Policy	None	+
Merge Handling Policy	Majority	+
Split and Merge Action Plan	Reboot	+
Select Tie Breaker		+

Figure 10-6 Configuring split policy: None, merge policy: Majority

When the cluster was first activated, glvma1 acquired the resource group rg1 (Example 10-3).

Example 10-3 Resource group acquisition

```
# /usr/es/sbin/cluster/utilities/clRGinfo -p
```

Cluster Name: glvma1_cluster

Resource Group Name: rg1

Node	Primary State	Secondary State
glvma1@siteA	ONLINE	OFFLINE
glvmb1@siteB	OFFLINE	ONLINE SECONDARY
glvmb2@siteB	OFFLINE	OFFLINE

Testing the split

We created a split situation by disabling the routers. As soon as the router was disabled, an event `split_merge_prompt_split` occurred on every node. Table 10-1 on page 458 lists the events that occurred on node glvma1 and glvmb1.

Notice that false and mismatching events such as `site_down` and `node_down` events occurred on each node. Because glvmb1 falsely indicated that siteA was down, it tried to acquire the resource without disabling it on glvma1.

Table 10-1 Events during the split

Cluster events during the split on glvma1	Cluster events during the split on glvmb1
EVENT START: split_merge_prompt split EVENT START: node_down glvmb1 EVENT COMPLETED: node_down glvmb1 0 EVENT START: site_down siteB EVENT COMPLETED: split_merge_prompt split 0 EVENT START: site_down_remote siteB EVENT COMPLETED: site_down_remote siteB 0 EVENT COMPLETED: site_down siteB 0 EVENT START: node_down glvmb2 EVENT COMPLETED: node_down glvmb2 0 EVENT START: rg_move_release glvma1 1 EVENT START: rg_move glvma1 1 RELEASE EVENT COMPLETED: rg_move glvma1 1 RELEASE 0 EVENT COMPLETED: rg_move_release glvma1 1 0 EVENT START: rg_move_fence glvma1 1 EVENT COMPLETED: rg_move_fence glvma1 1 0 EVENT START: node_down_complete glvmb1 EVENT COMPLETED: node_down_complete glvmb1 0 EVENT START: node_down_complete glvmb2 EVENT COMPLETED: node_down_complete glvmb2 0	EVENT START: split_merge_prompt split EVENT COMPLETED: split_merge_prompt split 0 EVENT START: site_down siteA EVENT START: site_down_remote siteA EVENT COMPLETED: site_down_remote siteA 0 EVENT COMPLETED: site_down siteA 0 EVENT START: node_down glvma1 EVENT COMPLETED: node_down glvma1 0 EVENT START: rg_move_release glvmb1 1 EVENT START: rg_move glvmb1 1 RELEASE EVENT COMPLETED: rg_move glvmb1 1 RELEASE 0 EVENT COMPLETED: rg_move_release glvmb1 1 0 EVENT START: rg_move_fence glvmb1 1 EVENT COMPLETED: rg_move_fence glvmb1 1 0 EVENT START: rg_move_release glvmb1 1 EVENT START: rg_move glvmb1 1 RELEASE EVENT COMPLETED: rg_move glvmb1 1 RELEASE 0 EVENT COMPLETED: rg_move_release glvmb1 1 0 EVENT START: rg_move_fence glvmb1 1 EVENT COMPLETED: rg_move_fence glvmb1 1 0 EVENT START: rg_move glvmb1 1 ACQUIRE EVENT START: acquire_takeover_addr EVENT COMPLETED: acquire_takeover_addr 0 EVENT COMPLETED: rg_move glvmb1 1 ACQUIRE 0 EVENT COMPLETED: rg_move_acquire glvmb1 1 0 EVENT START: rg_move_complete glvmb1 1 EVENT COMPLETED: rg_move_complete glvmb1 1 0 EVENT START: node_down_complete glvma1 EVENT COMPLETED: node_down_complete glvma1 0

The error log on each node showed a split occurred (Example 10-4).

Example 10-4 Split errlog

```
# errpt
# IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
4BDDFBCC 1207041112 I S ConfigRM The operational quorum state of the acti
A098BF90 1207041112 P S ConfigRM The operational quorum state of the acti
77A1A9A4 1207041112 I O ConfigRM ConfigRM received Site Split event notif
```

```
# errpt -a
```

```
LABEL: CONFIGRM_HASQUORUM_
# IDENTIFIER: 4BDDFBCC
```

```
Date/Time: Fri Dec 7 04:11:13 2012
Sequence Number: 883
Machine Id: 00F70C994C00
Node Id: glvma1
Class: S
Type: INFO
WPAR: Global
Resource Name: ConfigRM
```

Description

The operational quorum state of the active peer domain has changed to HAS_QUORUM. In this state, cluster resources may be recovered and controlled as needed by management applications.

Probable Causes

One or more nodes have come online in the peer domain.

User Causes

One or more nodes have come online in the peer domain.

Recommended Actions

None

Detail Data

DETECTING MODULE

RSCT,PeerDomain.C,1.99.22.110,18993

ERROR ID

REFERENCE CODE

LABEL: CONFIGRM_PENDINGQUO
IDENTIFIER: A098BF90

Date/Time: Fri Dec 7 04:11:13 2012
Sequence Number: 882
Machine Id: 00F70C994C00
Node Id: glvma1
Class: S
Type: PERM
WPAR: Global
Resource Name: ConfigRM

Description

The operational quorum state of the active peer domain has changed to PENDING_QUORUM.

This state usually indicates that exactly half of the nodes that are defined in the peer domain are online. In this state cluster resources cannot be recovered although none will be stopped explicitly.

Probable Causes

One or more nodes in the active peer domain have failed.

One or more nodes in the active peer domain have been taken offline by the user.
A network failure is disrupted communication between the cluster nodes.

Failure Causes

One or more nodes in the active peer domain have failed.

One or more nodes in the active peer domain have been taken offline by the user.
A network failure is disrupted communication between the cluster nodes.

Recommended Actions

Ensure that more than half of the nodes of the domain are online.

Ensure that the network that is used for communication between the nodes is functioning correctly.

Ensure that the active tie breaker device is operational and if it set to 'Operator' then resolve the tie situation by granting ownership to one of the active sub-domains.

Detail Data
DETECTING MODULE
RSCT,PeerDomain.C,1.99.22.110,18997
ERROR ID

REFERENCE CODE

LABEL: CONFIGRM_SITE_SPLIT
IDENTIFIER: 77A1A9A4

Date/Time: Fri Dec 7 04:11:12 2012
Sequence Number: 880
Machine Id: 00F70C994C00
Node Id: glvma1
Class: 0
Type: INFO
WPAR: Global
Resource Name: ConfigRM

Description
ConfigRM received Site Split event notification

Probable Causes
Networks between sites may have been disconnected

Failure Causes
Networks between sites may have been disconnected

Recommended Actions
Check the network connectivity between sites

Detail Data
DETECTING MODULE
RSCT,ConfigRMGROUP.C,1.331,1398
ERROR ID

REFERENCE CODE

DIAGNOSTIC EXPLANATION

At this point, resource groups were now acquired on both siteA and siteB. The volume group was activated on both nodes (Example 10-5 on page 461). The cluster state was *not* synced between sites, and each node was able to access write-enabled its own copies.

Table 10-2 Resource group state after the split

Resource group state in glvma1			Resource group state in glvmb1																				
<pre># /usr/es/sbin/cluster/utilities/c1RGinfo -p Cluster Name: glvma1_cluster Resource Group Name: rg1 Node State Primary State Secondary State -----</pre> <table> <tbody> <tr> <td>glvma1@siteA</td> <td>ONLINE</td> <td>OFFLINE</td> </tr> <tr> <td>glvmb1@siteB</td> <td>OFFLINE</td> <td>OFFLINE</td> </tr> <tr> <td>glvmb2@siteB</td> <td>OFFLINE</td> <td>OFFLINE</td> </tr> </tbody> </table> <pre># lspv hdisk0 00f70c99e24ff9ff altinst_rootvg hdisk1 00f70c9901259917 rootvg active hdisk2 00f70c992405114b caavg_private active hdisk3 00f70c990580a411 glvmvg active hdisk4 00f70c990580a44c glvmvg active hdisk5 00f70c990580a486 glvmvg active hdisk6 00f6f5d005808d31 glvmvg active hdisk7 00f6f5d005808d6b glvmvg active hdisk8 00f6f5d005808da5 glvmvg active hdisk9 00f70c9952af7567 None</pre>			glvma1@siteA	ONLINE	OFFLINE	glvmb1@siteB	OFFLINE	OFFLINE	glvmb2@siteB	OFFLINE	OFFLINE	<pre># /usr/es/sbin/cluster/utilities/c1RGinfo -p Cluster Name: glvma1_cluster Resource Group Name: rg1 Node State Primary State Secondary State -----</pre> <table> <tbody> <tr> <td>glvma1@siteA</td> <td>OFFLINE</td> <td>OFFLINE</td> </tr> <tr> <td>glvmb1@siteB</td> <td>ONLINE</td> <td>OFFLINE</td> </tr> <tr> <td>glvmb2@siteB</td> <td>OFFLINE</td> <td>OFFLINE</td> </tr> </tbody> </table> <pre># lspv hdisk0 00f6f5d0e24f6303 altinst_rootvg hdisk1 00f6f5d0012596cc rootvg active hdisk2 00f6f5d023ec219d caavg_private active hdisk3 00f6f5d005808d31 glvmvg active hdisk4 00f6f5d005808d6b glvmvg active hdisk5 00f6f5d005808da5 glvmvg active hdisk6 00f70c990580a411 glvmvg active hdisk7 00f70c990580a44c glvmvg active hdisk8 00f70c990580a486 glvmvg active hdisk9 00f70c9952af7567 None</pre>			glvma1@siteA	OFFLINE	OFFLINE	glvmb1@siteB	ONLINE	OFFLINE	glvmb2@siteB	OFFLINE	OFFLINE
glvma1@siteA	ONLINE	OFFLINE																					
glvmb1@siteB	OFFLINE	OFFLINE																					
glvmb2@siteB	OFFLINE	OFFLINE																					
glvma1@siteA	OFFLINE	OFFLINE																					
glvmb1@siteB	ONLINE	OFFLINE																					
glvmb2@siteB	OFFLINE	OFFLINE																					

In summary, the cluster partition was not prevented during total IP connection lost. This is the expected behavior for setting the split handling policy to None.

Testing the merge

Next, we created a merge situation by reactivating the routers. Because siteB was configured with two nodes, the expected behavior was the resource group moved to siteB.

Shortly after the router had been activated, an event split_merge_prompt merge occurred on every node. Table 10-3 shows the events that occurred on node glvma1 and glvmb1.

Table 10-3 Events during the merge

Cluster events during the merge on glvma1	Cluster events during the merge on glvmb1
EVENT START: split_merge_prompt merge EVENT COMPLETED: split_merge_prompt merge 0	EVENT START: split_merge_prompt merge EVENT COMPLETED: split_merge_prompt merge 0

The error log on each node indicated that a merge occurred(Example 10-5).

Example 10-5 Merge errlog

```
# errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
1BD32427 1207042512 I O ConfigRM ConfigRM received Site Merge event notif

# errpt -a
-----
LABEL: CONFIGRM_SITE_MERGE
IDENTIFIER: 1BD32427

Date/Time: Fri Dec 7 04:25:41 2012
Sequence Number: 764
Machine Id: 00F6F5D04C00
Node Id: glvmb2
Class: 0
Type: INFO
```

WPAR: Global
Resource Name: ConfigRM

Description
ConfigRM received Site Merge event notification

Probable Causes
Networks between sites may have been reconnected

Failure Causes
Networks between sites may have been reconnected

Recommended Actions
Verify the network connection between sites

Detail Data
DETECTING MODULE
RSCT,ConfigRMGroup.C,1.331,1436
ERROR ID

REFERENCE CODE

DIAGNOSTIC EXPLANATION

Shortly after these events occurred, a reboot occurred on glvma1. The errlog indicated the cause of the reboot (Example 10-6).

Example 10-6 Merge errlog

```
# errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
AFA89905 1207042712 I 0 cthags Group Services daemon started
DE84C4DB 1207042612 I 0 ConfigRM IBM.ConfigRM daemon has started.
A6DF45AA 1207042612 I 0 RMCdaemon The daemon is started.
2BFA76F6 1207042612 T S SYSPROC SYSTEM SHUTDOWN BY USER
9DBCFDEE 1207042612 T 0 errdemon ERROR LOGGING TURNED ON
24126A2B 1207042512 P 0 cthags Group Services daemon exit to merge/split
F0851662 1207042512 I S ConfigRM The sub-domain containing the local node
1BD32427 1207042512 I 0 ConfigRM ConfigRM received Site Merge event notif
```

```
# errpt -a
```

```
LABEL: GS_SITE_DISSOLVE_ER
IDENTIFIER: 24126A2B
```

```
Date/Time: Fri Dec 7 04:25:44 2012
Sequence Number: 886
Machine Id: 00F70C994C00
Node Id: glvma1
Class: 0
Type: PERM
WPAR: Global
Resource Name: cthags
```

Description
Group Services daemon exit to merge/split sites

Probable Causes

Network between two sites has repaired

Failure Causes

CAA Services has been partitioned and/or merged.

Recommended Actions

Check the CAA policies.

Verify that CAA has been restarted

Call IBM Service if problem persists

Detail Data

DETECTING MODULE

RSCT,NS.C,1.107.1.68,4894

ERROR ID

6fca2Y.MMPkE/kn8.rE4e.1.....

REFERENCE CODE

DIAGNOSTIC EXPLANATION

NS::Ack(): The master requests to dissolve my domain because of the merge with other domain 65535.Ni

LABEL: CONFIGRM_MERGE_ST
IDENTIFIER: F0851662

Date/Time: Fri Dec 7 04:25:44 2012

Sequence Number: 885

Machine Id: 00F70C994C00

Node Id: glvma1

Class: S

Type: INFO

WPAR: Global

Resource Name: ConfigRM

Description

The sub-domain containing the local node is being dissolved because another sub-domain has been detected that takes precedence over it. Group services will be ended on each node of the local sub-domain which will cause the configuration manager daemon (IBM.ConfigRMd) to force the node offline and then bring it back online in the surviving domain.

Probable Causes

A merge of two sub-domain is probably caused by a network outage being repaired so that the nodes of the two sub-domains can now communicate.

User Causes

A merge of two sub-domain is probably caused by a network outage being repaired so that the nodes of the two sub-domains can now communicate.

Recommended Actions

No action is necessary since the nodes will be automatically synchronized and brought online in the surviving domain.

Detail Data

DETECTING MODULE
RSCT,ConfigRMGroup.C,1.331,919
ERROR ID

REFERENCE CODE

In summary, the resource group stayed online on siteB, and siteA was brought offline on a merge event. This is the expected behavior for setting the merge policy to Majority because siteB had the majority of the nodes in the cluster.

10.4.4 Split policy: TieBreaker, merge policy: TieBreaker

Then we tested with the configuration shown in Figure 10-7.

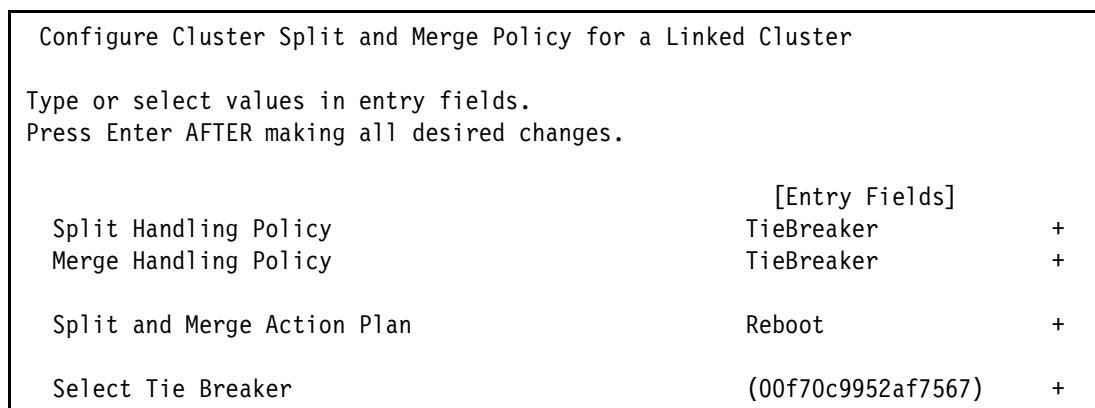


Figure 10-7 Configuring split policy: TieBreaker, merge policy: TieBreaker

The disk with PVID 00f70c9952af7567 was a DS3400 disk configured to all nodes in the cluster. From the DS3400, the LUN is named *glvm_tiebreaker*. Example 10-7 shows some of the tie breaker disks properties.

Example 10-7 Tie breaker disk names

```
# lspv | grep 00f70c9952af7567
hdisk9          00f70c9952af7567           None

# lsdev -Cc disk -l hdisk9
hdisk9 Available 80-T1-01 MPIO Other DS3K Array Disk

# mpio_get_config -l hdisk9
Storage Subsystem Name = 'DS3400POK-1'
      hdisk#        LUN #  Ownership          User Label
      hdisk9          0     A (preferred)    glvm_tiebreaker
```

When the cluster was first activated, *glvma1* acquires the resource group *rg1* (Example 10-8).

Example 10-8 Resource group acquisition

```
# /usr/es/sbin/cluster/utilities/clRGinfo -p

Cluster Name: glvma1_cluster

Resource Group Name: rg1
```

Node	Primary State	Secondary State
glvma1@siteA	ONLINE	OFFLINE
glvmb1@siteB	OFFLINE	ONLINE SECONDARY
glvmb2@siteB	OFFLINE	OFFLINE

To check whether there was a persistent reserve on the tie breaker disk, we used the **devrsrv** command. The output after the cluster activation is shown in Example 10-9.

Notice that the ODM Reservation Policy displayed PR EXCLUSIVE, and the Device Reservation State displayed NO RESERVE.

Example 10-9 The devrsrv command output after cluster activation

```
# devrsrv -c query -l hdisk9
Device Reservation State Information
=====
Device Name          : hdisk9
Device Open On Current Host?   : NO
ODM Reservation Policy      : PR EXCLUSIVE
ODM PR Key Value    : 5071576869747560161
Device Reservation State   : NO RESERVE
Registered PR Keys  : No Keys Registered
PR Capabilities Byte[2] : 0xd SIP_C ATP_C PTPL_C
PR Capabilities Byte[3] : 0x0
PR Types Supported  : NOT VALID
```

We also checked whether there was a persistent reserve on the DS3400 after the cluster was activated. To check the persistent reserve, we right-clicked the storage subsystem on the storage manager GUI and pressed **Execute Script** (Figure 10-8).

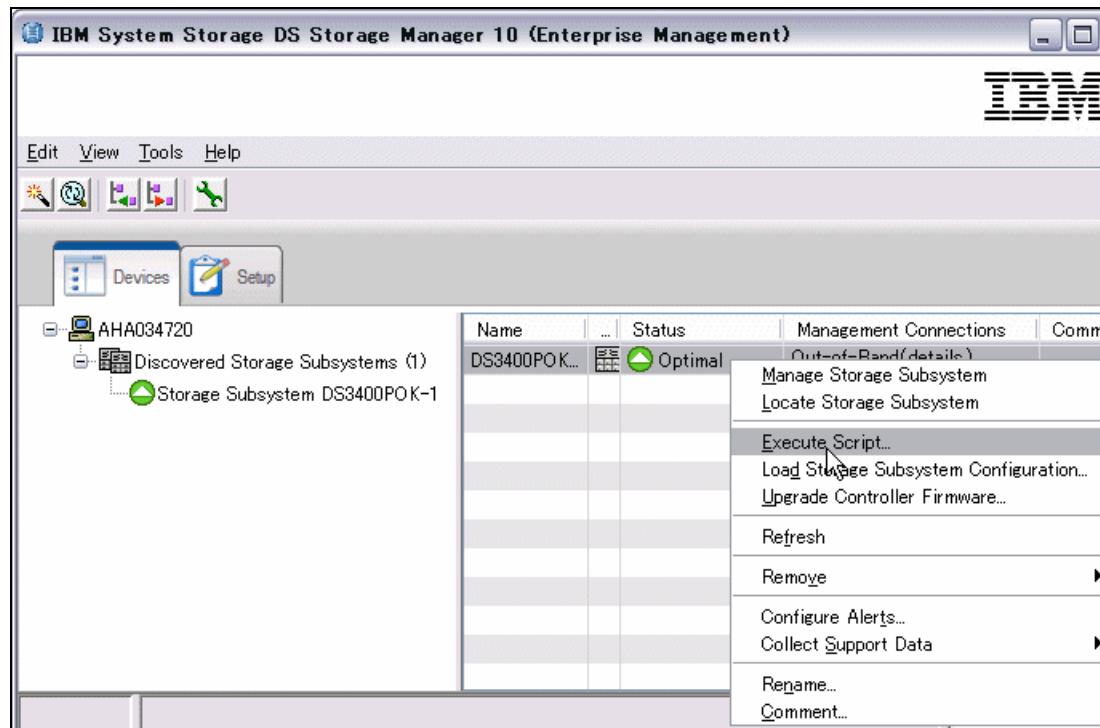
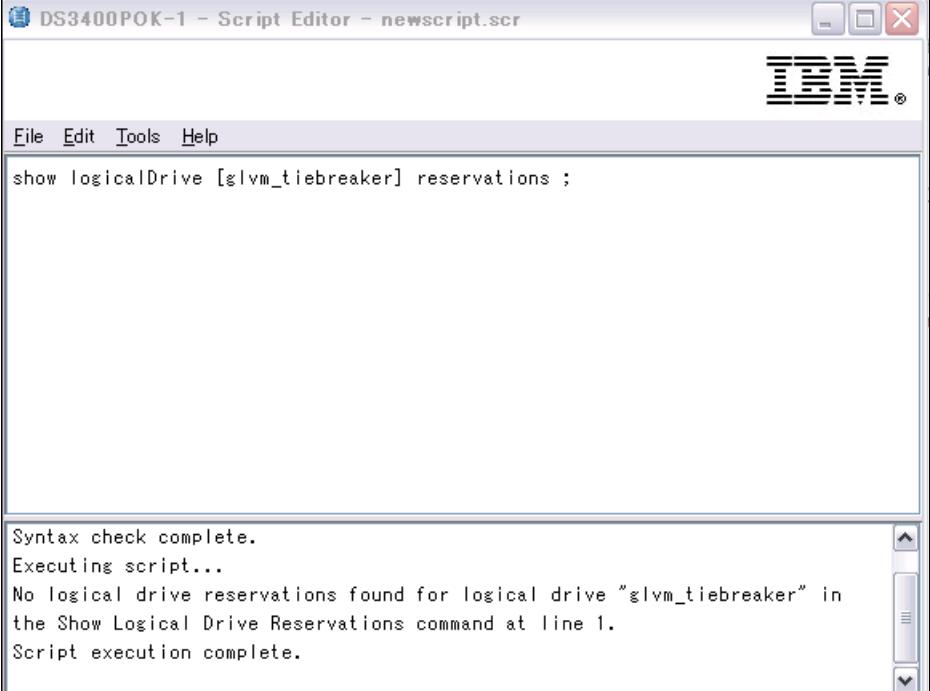


Figure 10-8 Execute Script from the Storage Manager GUI

Then we executed shows logicalDrives [logicalDrivelabel] reservations (Figure 10-9).



The screenshot shows a Windows-style application window titled "DS3400POK-1 - Script Editor - newscript.scr". The window has an "IBM" logo in the top right corner. A menu bar at the top includes "File", "Edit", "Tools", and "Help". Below the menu is a text input area containing the command: "show logicalDrive [glvmb_tiebreaker] reservations ;". In the bottom pane, the output of the command is displayed: "Syntax check complete.", "Executing script...", "No logical drive reservations found for logical drive \"glvmb_tiebreaker\" in the Show Logical Drive Reservations command at line 1.", and "Script execution complete.".

Figure 10-9 Checking the persistent reserve

From the output display we confirmed that no persistent reserve existed on this LUN.

Checking persistent reservation: Methods of checking the persistent reservation differ on the storage subsystem and device driver you are using.

Testing the split

Next we created a split situation by disabling the routers. Shortly after the routers were disabled, an event `split_merge_prompt split` occurred on every node. Table 10-4 on page 467 shows the events that occurred on node `glvma1` and `glvmb1` after the split.

Notice that the `split_merge_prompt` event is the only event on `glvma1`. This is because a reboot occurred shortly after the event happened. In comparison, from the “Split policy: None, merge policy: Majority” on page 456, there is no mismatch or false events, because all nodes on `siteA` have been rebooted.

Table 10-4 Events during the split

Cluster events during the split on glvma1	Cluster events during the split on glvmb1
EVENT START: split_merge_prompt split EVENT COMPLETED: split_merge_prompt split 0	EVENT START: split_merge_prompt split EVENT COMPLETED: split merge_prompt split 0 EVENT START: site_down_remote siteA EVENT COMPLETED: site_down_remote siteA 0 EVENT COMPLETED: site_down siteA 0 EVENT START: node_down glvma1 EVENT COMPLETED: node_down glvma1 0 EVENT START: rg_move_release glvmb1 1 EVENT START: rg_move glvmb1 1 RELEASE EVENT COMPLETED: rg_move glvmb1 1 RELEASE 0 EVENT COMPLETED: rg_move_release glvmb1 1 0 EVENT START: rg_move_fence glvmb1 1 EVENT COMPLETED: rg_move_fence glvmb1 1 0 EVENT START: rg_move_release glvmb1 1 EVENT START: rg_move glvmb1 1 RELEASE EVENT COMPLETED: rg_move glvmb1 1 RELEASE 0 EVENT COMPLETED: rg_move_release glvmb1 1 0 EVENT START: rg_move_fence glvmb1 1 EVENT COMPLETED: rg_move_fence glvmb1 1 0 EVENT START: rg_move_acquire glvmb1 1 EVENT START: rg_move glvmb1 1 ACQUIRE EVENT START: acquire_takeover_addr EVENT COMPLETED: acquire_takeover_addr 0 EVENT COMPLETED: rg_move glvmb1 1 ACQUIRE 0 EVENT COMPLETED: rg_move_acquire glvmb1 1 0 EVENT START: rg_move_complete glvmb1 1 EVENT COMPLETED: rg_move_complete glvmb1 1 0 EVENT START: node_down_complete glvma1 EVENT COMPLETED: node_down_complete glvma1 0

Additionally, other than the errlog in Example 10-4 on page 458 and before rebooting, failing attempts to access the tie breaker disk (hdisk9) can be observed on glvma1 (Example 10-10).

Example 10-10 Split errlog

```
# errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
AFA89905 1207053112 I 0 cthags Group Services daemon started
DE84C4DB 1207052812 I 0 ConfigRM IBM.ConfigRM daemon has started.
A6DF45AA 1207052812 I 0 RMCdaemon The daemon is started.
2BFA76F6 1207052812 T S SYSPROC SYSTEM SHUTDOWN BY USER
9DBCFDEE 1207052812 T 0 errdemon ERROR LOGGING TURNED ON
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207052712 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
A098BF90 1207052712 P S ConfigRM The operational quorum state of the acti
77A1A9A4 1207052712 I 0 ConfigRM ConfigRM received Site Split event notif

# errpt -a
-----
LABEL: SC_DISK_ERR9
IDENTIFIER: B0EE9AF5
```

Date/Time: Fri Dec 7 05:27:49 2012

Sequence Number: 925
Machine Id: 00F70C994C00
Node Id: glvma1
Class: S
Type: TEMP
WPAR: Global
Resource Name: hdisk9

Description
REQUESTED OPERATION CANNOT BE PERFORMED

Probable Causes
MEDIA

User Causes
MEDIA DEFECTIVE
RESOURCE NOT AVAILABLE

Recommended Actions
FOR REMOVABLE MEDIA, CHANGE MEDIA AND RETRY
PERFORM PROBLEM DETERMINATION PROCEDURES

Failure Causes
MEDIA
DISK DRIVE

Recommended Actions
FOR REMOVABLE MEDIA, CHANGE MEDIA AND RETRY
PERFORM PROBLEM DETERMINATION PROCEDURES

Detail Data

PATH ID

1

SENSE DATA

0600 1A00 7F00 FF00 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0118 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 009C 009C

Tips: We can determine that there is reserve conflict from the sense data of the SC_DISK_ERR errlog. Refer to the following layout:

SENSE DATA LAYOUT

LL00 CCCC CCCC CCCC CCCC CCCC CCCC RRRR RRRR RRRR VVSS AARR DDDD KKDD

In Example 10-10 on page 467, notice VV=01, SS=18. This indicates the following:

- ▶ VV 01 indicates that the SCSI status field (SS) is valid.
- ▶ SS 18 indicates that the SCSI device is reserved by another host.

For details about SCSI3 protocol errlog layout, refer to PCI Fibre Channel Adapter, SCSI-3 Protocol (Disk, CD-ROM, Read/Write Optical Device) Error Log Sense Information in *Fibre Channel Planning and Integration: User's Guide and Service Information* at:

<http://publibfp.boulder.ibm.com/epubs/pdf/c2343293.pdf>

The **devrsrv** command shows a persistent key reservation on the disk (Example 10-11). The Device Reservation State now shows PR EXCLUSIVE. The node that acquired the tie breaker can be determined by comparing the ODM PR Key Value (which is unique on each cluster node) and the PR Holder Key Value.

Example 10-11 The devrsrv command after the split event

```
# devrsrv -c query -l hdisk9
Device Reservation State Information
=====
Device Name          : hdisk9
Device Open On Current Host?   : NO
ODM Reservation Policy      : PR EXCLUSIVE
ODM PR Key Value        : 5071576869747560161
Device Reservation State    : PR EXCLUSIVE
PR Generation Value       : 1274
PR Type                 : PR_WE_RO (WRITE EXCLUSIVE, REGISTRANTS ONLY)
PR Holder Key Value      : 5071576869747560161
Registered PR Keys        : 5071576869747560161 5071576869747560161
PR Capabilities Byte[2]    : 0xd SIP_C ATP_C PTPL_C
PR Capabilities Byte[3]    : 0x1 PTPL_A
PR Types Supported        : NOT VALID
```

Also, using the Storage Manager GUI, we now observed the persistent reserve on hdisk9 (Figure 10-10). As discussed in “Tie breaker disk overview” on page 451, this reservation is designed to last until every node rejoins the cluster.

The screenshot shows a window titled "DS3400POK-1 - Script Editor - newscript.scr". The window has an "IBM." logo in the top right corner. The menu bar includes "File", "Edit", "Tools", and "Help". The main text area contains the command:

```
show logicalDrive [glvm_tiebreaker] reservations ;
```

The output pane below shows the results of the command execution:

```
Executing script...
Logical Drive Name: glvm_tiebreaker
 Accessible by: glvm_tiebreaker 1
 Registrations: 2
 Reservation type: WE - RO
 Persistent through power loss?: Yes
Script execution complete.
```

Figure 10-10 Persistent reserve on the tie breaker disk

Because glvma1 rebooted, glvmb1 is now the only node that has access to the data (Example 10-12).

Example 10-12 Resource group state on glvmb1

```
# /usr/es/sbin/cluster/utilities/clRGinfo -p

Cluster Name: glvma1_cluster

Resource Group Name: rg1
Node Primary State Secondary State
-----
glvma1@siteA OFFLINE OFFLINE
glvmb1@siteB ONLINE OFFLINE
glvmb2@siteB OFFLINE OFFLINE

# lspv
hdisk0 00f6f5d0e24f6303 altinst_rootvg
hdisk1 00f6f5d0012596cc rootvg active
hdisk2 00f6f5d023ec219d caavg_private active
hdisk3 00f6f5d005808d31 glvmvg active
hdisk4 00f6f5d005808d6b glvmvg active
hdisk5 00f6f5d005808da5 glvmvg active
hdisk6 00f70c990580a411 glvmvg active
```

hdisk7	00f70c990580a44c	glvmsg	active
hdisk8	00f70c990580a486	glvmsg	active
hdisk9	00f70c9952af7567	None	

In summary, this scenario shows that the cluster partitioning was prevented. However, although the resource group was initially started at siteA, siteB acquired the resource group making an unplanned resource group movement. Because having a split handling policy to TieBreaker is a “winner takes all” policy, this outcome is working as designed.

Testing the merge

After the reboot completed on glvma1, we reactivated the cluster services with **smitty clstart**. Because the IP-connectivity had not been restored between sites, the two nodes acquired the resource group at the same time (Table 10-5).

Table 10-5 Resource group state before the merge

Resource group state in glvma1			Resource group state in glvmb1		
# /usr/es/sbin/cluster/utilities/clRGinfo -p Cluster Name: glvma1_cluster Resource Group Name: rg1 Node Primary State Secondary State			# /usr/es/sbin/cluster/utilities/clRGinfo -p Cluster Name: glvma1_cluster Resource Group Name: rg1 Node Primary State Secondary State		
glvma1@siteA ONLINE OFFLINE glvmb1@siteB OFFLINE OFFLINE glvmb2@siteB OFFLINE OFFLINE			glvma1@siteA OFFLINE OFFLINE glvmb1@siteB ONLINE OFFLINE glvmb2@siteB OFFLINE OFFLINE		
# lsvp hdisk0 00f70c99e24ff9ff altinst_rootvg hdisk1 00f70c9901259917 rootvg active hdisk2 00f70c992405114b caavg_private active hdisk3 00f70c990580a411 glvmsg active hdisk4 00f70c990580a44c glvmsg active hdisk5 00f70c990580a486 glvmsg active hdisk6 00f6f5d005808d31 glvmsg active hdisk7 00f6f5d005808d6b glvmsg active hdisk8 00f6f5d005808da5 glvmsg active hdisk9 00f70c9952af7567 None			# lsvp hdisk0 00f6f5d0e24f6303 altinst_rootvg hdisk1 00f6f5d0012596cc rootvg active hdisk2 00f6f5d023ec219d caavg_private active hdisk3 00f6f5d005808d31 glvmsg active hdisk4 00f6f5d005808d6b glvmsg active hdisk5 00f6f5d005808da5 glvmsg active hdisk6 00f70c990580a411 glvmsg active hdisk7 00f70c990580a44c glvmsg active hdisk8 00f70c990580a486 glvmsg active hdisk9 00f70c9952af7567 None		

Important: From the previous test result, the split handling policy TieBreaker cannot prevent the cluster partitioning before regaining the IP connectivity between sites. In real-world scenarios, confirm that the IP connectivity is restored for the two sites before you perform additional operations.

Upon reactivating the routers, an event `split_merge_prompt merge` occurred on every node. Table 10-6 shows the events that occurred on node glvma1 and glvmb1 after the merge.

Table 10-6 Events during the merge

Cluster events during the merge on glvma1	Cluster events during the merge on glvmb1
EVENT START: <code>split_merge_prompt merge</code> EVENT COMPLETED: <code>split_merge_prompt merge 0</code>	EVENT START: <code>split_merge_prompt merge</code> EVENT COMPLETED: <code>split_merge_prompt merge 0</code>

Shortly after these events, node glvma1 rebooted. The errlog is shown in Example 10-13 on page 472. Observe that after the merge events, there are several failing attempts to access the tie breaker disk, hdisk9, and initiate a reboot.

Example 10-13 Merge errlog

```
# errpt
IDENTIFIER TIMESTAMP T C RESOURCE_NAME DESCRIPTION
AFA89905 1207054812 I 0 cthags Group Services daemon started
DE84C4DB 1207054712 I 0 ConfigRM IBM.ConfigRM daemon has started.
A6DF45AA 1207054712 I 0 RMCdaemon The daemon is started.
2BFA76F6 1207054712 T S SYSPROC SYSTEM SHUTDOWN BY USER
9DBCFDEE 1207054712 T 0 errdemon ERROR LOGGING TURNED ON
24126A2B 1207054612 P 0 cthags Group Services daemon exit to merge/spli
F0851662 1207054612 I S ConfigRM The sub-domain containing the local node
65DE6DE3 1207054612 P S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
B0EE9AF5 1207054612 T S hdisk9 REQUESTED OPERATION CANNOT BE PERFORMED
1BD32427 1207054512 I 0 ConfigRM ConfigRM received Site Merge event notif

# errpt -a
-----
LABEL: GS_SITE_DISSOLVE_ER
IDENTIFIER: 24126A2B

Date/Time: Fri Dec 7 05:46:29 2012
Sequence Number: 1652
Machine Id: 00F70C994C00
Node Id: glvma1
Class: 0
Type: PERM
WPAR: Global
Resource Name: cthags

Description
Group Services daemon exit to merge/split sites

Probable Causes
Network between two sites has repaired

Failure Causes
CAA Services has been partitioned and/or merged.

Recommended Actions
Check the CAA policies.
Verify that CAA has been restarted
Call IBM Service if problem persists

Detail Data
DETECTING MODULE
RSCT,NS.C,1.107.1.68,4894
ERROR ID
6fca2Y.3YQkE/f1o/rE4e.1.....
REFERENCE CODE
```

DIAGNOSTIC EXPLANATION

NS::Ack(): The master requests to dissolve my domain because of the merge with other dom
ain 65535.Ni

LABEL: CONFIGRM_MERGE_ST
IDENTIFIER: F0851662

Date/Time: Fri Dec 7 05:46:29 2012
Sequence Number: 1651
Machine Id: 00F70C994C00
Node Id: glvma1
Class: S
Type: INFO
WPAR: Global
Resource Name: ConfigRM

Description

The sub-domain containing the local node is being dissolved because another sub-domain has been detected that takes precedence over it. Group services will be ended on each node of the local sub-domain which will cause the configuration manager daemon (IBM.ConfigRMd) to force the node offline and then bring it back online in the surviving domain.

Probable Causes

A merge of two sub-domain is probably caused by a network outage being repaired so that the nodes of the two sub-domains can now communicate.

User Causes

A merge of two sub-domain is probably caused by a network outage being repaired so that the nodes of the two sub-domains can now communicate.

Recommended Actions

No action is necessary since the nodes will be automatically synchronized and brought online in the surviving domain.

Detail Data

DETECTING MODULE
RSCT,ConfigRMDGroup.C,1.331,919
ERROR ID

REFERENCE CODE

LABEL: SC_DISK_ERR10
IDENTIFIER: 65DE6DE3

Date/Time: Fri Dec 7 05:46:25 2012
Sequence Number: 1650
Machine Id: 00F70C994C00
Node Id: glvma1
Class: S
Type: PERM
WPAR: Global
Resource Name: hdisk9

Description
REQUESTED OPERATION CANNOT BE PERFORMED

Probable Causes
DASD DEVICE

User Causes
RESOURCE NOT AVAILABLE
UNAUTHORIZED ACCESS ATTEMPTED

Recommended Actions
FOR REMOVABLE MEDIA, CHANGE MEDIA AND RETRY
PERFORM PROBLEM DETERMINATION PROCEDURES

Failure Causes
MEDIA
DISK DRIVE

Recommended Actions
FOR REMOVABLE MEDIA, CHANGE MEDIA AND RETRY
PERFORM PROBLEM DETERMINATION PROCEDURES

Detail Data
PATH ID

1

SENSE DATA
0A00 2A00 0000 0000 0000 0804 0000 0000 0000 0000 0000 0000 0000 0000 0118 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0004 0200 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0093
0000 003D 0017

LABEL: CONFIGRM_SITE_MERGE
IDENTIFIER: 1BD32427

Date/Time: Fri Dec 7 05:45:26 2012
Sequence Number: 1595
Machine Id: 00F70C994C00
Node Id: glvma1
Class: 0
Type: INFO
WPAR: Global
Resource Name: ConfigRM

Description
ConfigRM received Site Merge event notification

Probable Causes
Networks between sites may have been reconnected

Failure Causes
Networks between sites may have been reconnected

Recommended Actions
Verify the network connection between sites

Detail Data
DETECTING MODULE
RSCT,ConfigRMGroup.C,1.331,1436
ERROR ID

REFERENCE CODE

DIAGNOSTIC EXPLANATION

After the merge event, observe that the persistent reservation has been released from the **devrsrv** command (Example 10-14), and from the Storage Manager GUI, as shown Figure 10-11.

Example 10-14 Devrsrv command output after merge event

```
# devrsrv -c query -l hdisk9
Device Reservation State Information
=====
Device Name          : hdisk9
Device Open On Current Host?   : NO
ODM Reservation Policy      : PR EXCLUSIVE
ODM PR Key Value    : 5071576869747560161
Device Reservation State    : NO RESERVE
Registered PR Keys   : No Keys Registered
PR Capabilities Byte[2]  : 0xd SIP_C ATP_C PTPL_C
PR Capabilities Byte[3]  : 0x0
PR Types Supported   : NOT VALID
```

The screenshot shows a window titled "DS3400POK-1 - Script Editor - newscript.scr". The window has a menu bar with "File", "Edit", "Tools", and "Help". The main area contains the command:

```
show logicalDrive [glvm_tiebreaker] reservations ;
```

Below the command, a message box displays the execution results:

```
Executing script...
No logical drive reservations found for logical drive "glvm_tiebreaker" in the Show Logical Drive Reservations command at line 1.
Script execution complete.
```

Figure 10-11 Persistent reserve release

Important: Although there are several methods to manually unlock the persistent reserve, do *not* perform a reserve operation outside of PowerHA. If, in any way, the persistent reserve leads to a problem, contact IBM support first.

In summary from this test scenario, we can observe that the tie breaker disk prevents other nodes from acquiring the resources after the split events occurs.



Part 5

Appendices

This part provides appendices with additional information discovered during the residency which we believe has informational value for our readers. The following appendices are offered:

- ▶ Appendix A, “Configuring IBM PowerHA SystemMirror with IPv6” on page 479.
- ▶ Appendix B, “DNS change for the IBM Systems Director environment with PoweHA” on page 495.



A

Configuring IBM PowerHA SystemMirror with IPv6

This appendix describes how to configure IBM PowerHA SystemMirror with IPv6.

The following topics are presented:

- ▶ Configuring IBM PowerHA SystemMirror with IPv6
- ▶ Enabling IPv6 on AIX
- ▶ Configuration tips for PowerHA with IPv6
- ▶ CAA command enhancements
- ▶ Migrating steps from IPv4 to IPv6

Configuring IBM PowerHA SystemMirror with IPv6

This section includes tips and considerations discovered while configuring IBM PowerHA SystemMirror with IPv6.

Enabling IPv6 on AIX

This section covers the basic steps to enable IPv6 on AIX. This must be done prior to performing any PowerHA configuration steps.

Autoconf6

On AIX, IPv6 can be enabled by the **autoconf6** command. If no flags are specified, this command enables IPv6 on every Ethernet adapter that is configured. The **-i** flag is used to specify that certain adapters will be configured as IPv6. After executing the command, the IPv6 link local address is configured on the adapter that has been specified. Also SIT interface sit0 (used for IPv4-compatible IPv6 addresses) is configured as shown in Example A-1.

Example A-1 Executing autoconf6

```
# autoconf6 -i en1

# ifconfig -a
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 192.168.100.55 netmask 0xffffffff00 broadcast 192.168.100.255
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en1:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet6 fe80::7840:c3ff:fe0b:1f03/64
                tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
sit0: flags=8100041<UP,RUNNING,LINKO>
        inet6 ::192.168.100.55/96
lo0:
flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,LARGESEND,CHAIN>
        inet 127.0.0.1 netmask 0xffffffff broadcast 127.255.255.255
        inet6 ::1%1/128
                tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
```

For these configuration changes to persist after a reboot, execute **smitty chauto6**. Choose the adapter to enable IPv6. Confirm your setting in the SMIT panel as shown in Figure A-1 on page 481, and press Enter.

Change / Show Restart Characteristics of Autoconf6 Process

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

[Entry Fields]		
* Start the Autoconf6 Process with VERBOSE on?	no	+
* Start the Autoconf6 Process with SIT on?	yes	+
Network INTERFACE	en1	

Figure A-1 Changing the autoconf6 process

After executing, the /etc/rc.tcpip file is modified as shown in Example A-2.

Example A-2 Activating the autoconf6 through /etc/rc.tcpip

```
# cat /etc/rc.tcpip | grep autoconf6
# Start up autoconf6 process
start /usr/sbin/autoconf6 "" " -i en1"
```

Configuring a static IPv6 address

To configure a static IPv6 address, the **ifconfig** command can be used with the **inet6** flag as shown in Example A-3.

Example A-3 Configuring a static IP with the ifconfig command

```
# ifconfig en1 inet6 2000::c0a8:6437 prefixlen 64
```

For this change to persist after a reboot, **smitty chinetc6** can be used as shown in Figure A-2.

Change / Show an IPV6 Standard Ethernet Interface

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

[Entry Fields]		
Network Interface Name	en1	
IPV6 ADDRESS (colon separated)	[2000::c0a8:6437]	
Prefixlength	[64]	
Current STATE	up	+

Figure A-2 SMIT panel for configuring the IPv6 address

This modifies the attributes of the specified device. The **lsattr** command is used to display the configuration, as shown in Example A-4.

Example A-4 Modified IP attribute

```
# lsattr -El en1 -a netaddr6 -a prefixlen -a state
netaddr6 2000::c0a8:6437      IPv6 Internet Address      True
prefixlen 64                  prefix Length for IPv6 Internet Address True
state     up                  Current Interface Status  True
```

Configuring local name resolutions

To configure local name resolutions, modify the /etc/hosts file as shown in Example A-5.

Example A-5 Adding IPv6 entries in /etc/hosts

```
# cat /etc/hosts
# IPv6 boot
2000::c0a8:6437 glvma1ip6
2000::c0a8:6438 glvma2ip6
2000:aaaa::a0a:6439 glvmb1ip6
2000:aaaa::a0a:643a glvmb2ip6
```

If the DNS is configured, and /etc/resolv.conf exists, remember to edit the /etc/netsvc.conf file to prefer local name resolutions by adding **local** or **local6** to the hosts variable.

Configuring IPv6 static routes

To configure IPv6 static routes, the **route** command can be used with an **inet6** flag as shown in Example A-6.

Example A-6 Adding routes through the route command

```
# route add -inet6 default 2000::c0a8:643b
```

For this to persist across a reboot, **smitty mkroute6** can be used as shown in Figure A-3.

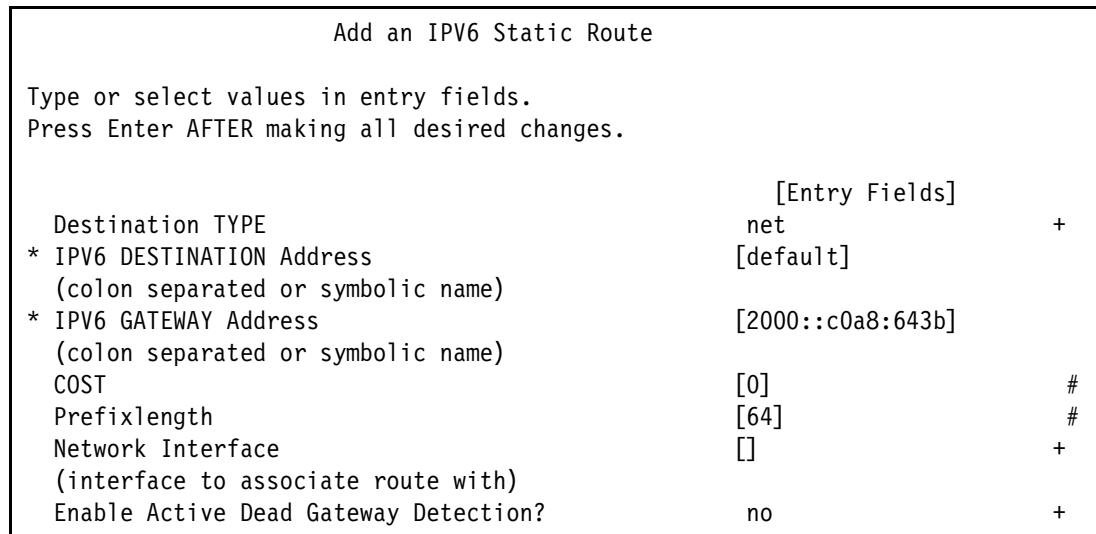


Figure A-3 SMIT panel for adding routes

This modifies the attributes of the specified device. The **lsattr** command is used to display the configuration, as shown in Example A-7.

Example A-7 Route labels that have been modified

```
# lsattr -El inet0 -a rout6
rout6 net,-hopcount,0,,,,-static,:,:,:2000::c0a8:643b IPv6 Route True
```

Configuration tips for PowerHA with IPv6

This section covers tips and limitations for configuring PowerHA with IPv6.

Persisting a static IPv6 address

In our test environment, we were required to perform the following for the static IPv6 address to persist during system reboots and cluster activation:

- ▶ Add the **-g** flag to the **ndpd-host** daemon.
- ▶ Disable the Router Advertisement (RA) functions on the network routers.

The reason these were required is that IPv6 introduces IPv6 stateless address auto configuration. This is a function that the client communicates with the network routers and automatically configures a global IPv6 address, which is based on the following:

- ▶ The network prefix provided from the network routers
- ▶ The client network card MAC address

Note: Refer to *RFC 4862 - IPv6 Stateless Address Auto configuration* for more details at:

<http://tools.ietf.org/html/rfc4862>

In AIX the **ndpd-host** daemon is responsible for the stateless IPv6 addresses.

When PowerHA is installed and the IPv6 addresses are configured, the **ndpd-host** daemon starts up on system startup. The *hacmp* entry in the *inittab* file is responsible for this as shown in Example A-8.

Example A-8 ndp-host startup from the hacmp entry in the inittab file

```
# lsitab hacmp
hacmp:2:once:/usr/es/sbin/cluster/etc/rc.init >/dev/console 2>&1

# cat /usr/es/sbin/cluster/etc/rc.init
    IPV6_BOOT_INTERFACES=$(cllsif -S -J "$OP_SEP" | grep
"${OP_SEP}${NODENAME}${OP_SEP}" | \
                grep "${OP_SEP}boot${OP_SEP}" | grep AF_INET6 | cut
-d"$OP_SEP" -f9)
    [[ -n "$IPV6_BOOT_INTERFACES" ]] && {
        for INTERFACE in $IPV6_BOOT_INTERFACES
        do
            /usr/sbin/autoconf6 -i $INTERFACE
        done

        /usr/bin/startsrc -s ndpd-host
        sleep 60
    }
...

```

Upon activating **ndpd-host**, we discovered that **ndpd-host** deletes the static IPv6 address, and instead the stateless IPv6 address becomes the global IPv6 address as shown in Example A-9.

Example A-9 Static IP address getting deleted

```
# lssrc -s ndpd-host
```

Subsystem	Group	PID	Status
ndpd-host	tcpip		inoperative

```
# ifconfig en1
en1:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet6 2000::c0a8:6437/64
        inet6 fe80::7840:c3ff:fe0b:1f03/64
            tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1

# startsrc -s ndpd-host
0513-059 The ndpd-host Subsystem has been started. Subsystem PID is 8782050.

# ifconfig en1
en1:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet6 2000::7840:c3ff:fe0b:1f03/64 <---- static IP address gets deleted
and stateless IP address is being used
        inet6 fe80::7840:c3ff:fe0b:1f03/64
            tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
```

This is a known issue; APAR IV16141 has introduced the **-g** flag to avoid deletion of the static IPv6 address. Since the *hacmp* entry in the *inittab* starts the **ndpd-host** daemon with no flags specified, we needed to add this flag with the **chssys** command as shown in Example A-10. This must be executed on all nodes in the cluster.

Example A-10 Changing the attributes in ndpd-host

```
# chssys -s ndpd-host -a "-g"

# odmget -q "subsysname = ndpd-host" SRCsubsys

SRCsubsys:
    subsysname = "ndpd-host"
    synonym = ""
    cmdargs = "-g" <---- confirm the setting
    path = "/usr/sbin/ndpd-host"
    uid = 0
    auditid = 0
    standin = "/dev/console"
    standout = "/dev/console"
    standerr = "/dev/console"
    action = 2
    multi = 0
    contact = 3
    svrkey = 0
    svrmtype = 0
    priority = 20
    signorm = 0
    sigforce = 0
    display = 1
    waittime = 20
    grpname = "tcpip"
```

The execution of the command shown in Example A-10 on page 484 prevents the static IPv6 address from being deleted. However, now the stateless IPv6 address and the static IPv6 will coexist on the same adapter, as shown in Example A-11.

Example A-11 Coexistence of stateless and static IP addresses

```
# ifconfig en1
en1:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet6 fe80::7840:c3ff:fe0b:1f03/64
    inet6 2000::c0a8:6437/64
        tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1

# startsrc -s ndpd-host
0513-059 The ndpd-host Subsystem has been started. Subsystem PID is 8782052.

# ifconfig en1
en1:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet6 2000::7840:c3ff:fe0b:1f03/64 <---- stateless IP address
    inet6 fe80::7840:c3ff:fe0b:1f03/64
    inet6 2000::c0a8:6437/64 <---- static IP address
        tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
```

This also led to another issue: it changes the source IP addresses of packets. In the event of performing a migration, this made **clmigcheck** fail to create the CAA cluster due to incoming IPs not listed in /etc/cluster/rhosts. Example A-12 shows the error observed during the **clmigcheck** command.

Example A-12 clmigcheck error

```
# cat /etc/hosts
# IPv6 boot
2000::c0a8:6437 glvma1ip6
2000::c0a8:6438 glvma2ip6
2000:aaaa::a0a:6439 glvmb1ip6
2000:aaaa::a0a:643a glvmb2ip6

# cat /etc/cluster/rhosts
glvma1ip6
glvma2ip6
glvmb1ip6
glvmb2ip6

# clmigcheck
Saving existing /tmp/clmigcheck/clmigcheck.log to
/tmp/clmigcheck/clmigcheck.log.bak
Verifying clcomd communication, please be patient.

ERROR: COMM-ERROR: Unable to verify inbound clcomd communication to node:
glvma1ip6
```

Example A-13 on page 486 shows the **tcpdump** command on glvma1ip6 during the **clmigcheck** error. You can see, in bold, that the communication is done using the stateless IP addresses.

Example A-13 *tcpdump from glvma1ip6 during the clmigcheck error*

```
# ifconfig -a
en1:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,CHECKSUM_OFFL
OAD(ACTIVE),CHAIN>
    inet6 2000::7840:c3ff:fe0b:1f03/64 <---- stateless IP address.
    inet6 fe80::7840:c3ff:fe0b:1f03/64
    inet6 2000::c0a8:6437/64 <---- static IP address.
        tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1

# tcpdump -i en1 ip6
11:26:27.894223 IP6 2000::7840:c3ff:fe0b:1f03.1023 > glvmb2ip6.clcomd: . ack 147 win 32844
<nop,nop,timestamp 1353688754 1353688744>
11:26:27.894258 IP6 2000::7840:c3ff:fe0b:1f03.1023 > glvmb2ip6.clcomd: F 72:72(0) ack 147 win
32844 <nop,nop,timestamp 1353688754 1353688744>
11:26:27.894399 IP6 glvmb2ip6.clcomd > 2000::7840:c3ff:fe0b:1f03.1023: . ack 73 win 32844
<nop,nop,timestamp 1353688744 1353688754>
11:26:27.894897 IP6 glvmb2ip6.clcomd > 2000::7840:c3ff:fe0b:1f03.1023: S 814143724:814143724(0)
ack 814079652 win 65535 <mss 1440,nop,wscale 3,nop,nop,timestamp 1353688744 1353688754>
```

In our test scenarios, to avoid further issues, we decided to disable the network routers' Router Advertisement (RA) function, which is responsible for creating the stateless IPv6 addresses. Consult with your network administrator to inquire whether the same is possible in your environment.

Tips: Disabling the router advertisement differs depending on which network router you use. For CISCO IOS version 15.1 or later, issue the following command:

```
> ipv6 nd ra suppress all
```

The IBM PowerHA SystemMirror development team is currently addressing these issues, and intends to solve them in future software releases.

You may use the stateless IPv6 address instead of a static IPv6 address. However, we do *not* suggest doing so. Since a stateless IPv6 address is based on network router settings and the nodes MAC address, there is no guarantee that this IP address will persist.

Manual configuration of IPv6 labels

In our test environment we encountered issues where IPv6 interfaces and networks were not configured during cluster configuration through **smitty sysmirror** → **Cluster Nodes and Networks** → **Multi Site Cluster Deployment** → **Setup a Cluster, Sites, Nodes and Networks**. If these situations occur, configure IPv6 labels manually through **smitty sysmirror** → **Cluster Nodes and Networks** → **Manage Networks and Network Interfaces**. When configuring network interfaces, although optional, we advise to specify the interfaces in the SMIT panel as shown in Figure A-4 on page 487.

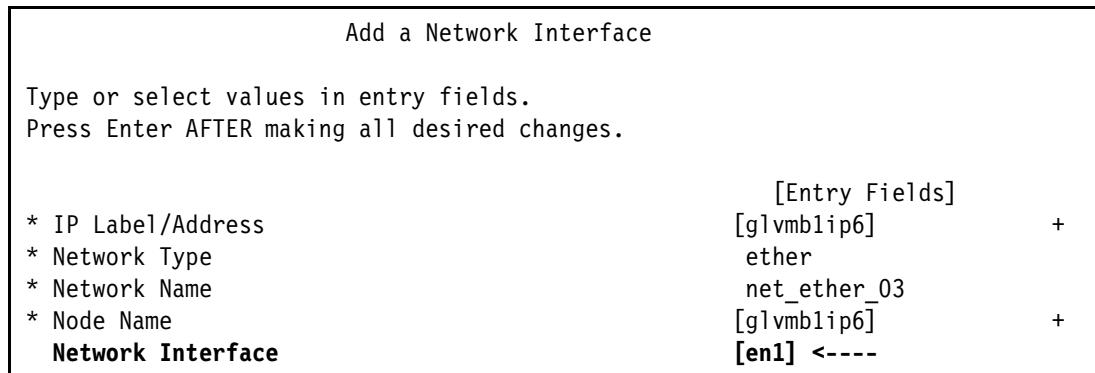


Figure A-4 Specifying network interfaces

When these interfaces are not specified, we see verification errors stating that a link local address has not been detected, as shown in Figure A-5. This still occurs even if we follow the steps in “Autoconf6” on page 480.

WARNING: No Link local addresses detected on interface of node glvmb1ip6.
Do you want to auto-configure link-local address on interface of glvmb1ip6?
[Yes / No]:

Figure A-5 Verification error for a null interface device

The warning disappeared after specifying the network interface, as shown in Figure A-4.

IPv6 service IP label limitations

IPv6 service labels can only be configured on XD_data and ether networks. Figure A-6 shows the verification errors if configured for different types of networks. Currently XD_ip networks cannot be configured with IPv6 service IPs.

ERROR: Resource group "glvm_ip6" has IPv6 service label "glvmasrv" configured on unsupported network type. IPv6 service IP can only be configured on network of type "ether and XD_data".

Figure A-6 Verification error when configuring IPv6 service labels on an XD_ip network

The IBM PowerHA SystemMirror development team is currently addressing this limitation, and intends to solve it in future software releases.

IPv6 service IP label takeover issues

We encountered an issue when configuring the boot and service IPv6 label to the same prefix. When a service IP takeover occurs, the routing table gets deleted making the node inaccessible. Example A-14 shows the routing table before the service IP takeover.

Example A-14 IPv6 route before IP takeover

```
Route tree for Protocol Family 24 (Internet v6):
::/96          0.0.0.0      UC      0      0 sit0      -      -      =>
default        2000:::c0a8:643b  UGS     4      4055 en1      -      - 
::1%1          ::1%1        UH      1      56 loo      -      - 
::192.168.100.55 192.168.100.55 UHLW    0      8 loo      -      - 
2000:::c0a8:6400/120 link#3      UC      1      0 en1      -      - 
2000:::c0a8:6437      UHLW1    1      1750 loo      -      - 
2000:::c0a8:6438      UHL      1      1798 en1      -      -
```

2000::c0a8:6439		UHLW1	0	12	lo0	-	-
2000::c0a8:643a	7a:40:cf:a:ea:3	UHLW	0	3	en1	-	-
2000::c0a8:643b	ee:af:bb:b8:2	UHLW	1	0	en1	-	-
fe80::/64	link#3	UCX	2	0	en1	-	-
fe80::7840:cfffe0a:ea03	7a:40:cf:a:ea:3	UHL	0	7	en1	-	-
fe80::ecaf:bf:feb:b802	ee:af:bb:b8:2	UHL	0	17	en1	-	-
ff01::%2/16	::1	U	0	3	lo0	-	-
ff02::/16	fe80::7840:c3ff:fe0b:1f03	U	0	36	en1	-	-
ff11::%2/16	::1	U	0	0	lo0	-	-
ff12::/16	fe80::7840:c3ff:fe0b:1f03	U	0	0	en1	-	-

Example A-15 shows the routing table after the service IP takeover, which shows the highlighted routes getting deleted.

Example A-15 IPv6 routes after IP takeover

Route tree for Protocol Family 24 (Internet v6):							
::/96	0.0.0.0	UC	0	0	sit0	-	- =>
default	2000::c0a8:643b	UGS	10	40752	en1	-	-
::1%	::1%	UH	1	136	lo0	-	-
::192.168.100.55	192.168.100.55	UHLW	0	8	lo0	-	-
fe80::/64	link#3	UCX	6	0	en1	-	-
fe80::7840:cfffe0a:ea03	7a:40:cf:a:ea:3	UHL	0	5	en1	-	-
fe80::ecaf:bf:feb:b802	ee:af:bb:b8:2	UHL	0	5	en1	-	-
ff01::%2/16	::1	U	0	3	lo0	-	-
ff02::/16	fe80::7840:c3ff:fe0b:1f03	U	0	41	en1	-	-
ff11::%2/16	::1	U	0	0	lo0	-	-
ff12::/16	fe80::7840:c3ff:fe0b:1f03	U	0	0	en1	-	-

Currently this can be avoided in one of two ways:

- ▶ Configure the IPv6 service labels other than the prefix of the IPv6 boot label.
- ▶ Use pre-post scripts to manually rebuild the routing table.

Important: There will be no IBM support for using this method. Testing will be your responsibility.

Development is currently aware of this issue. Contact your IBM support for any available fixes.

Multicast packets in an IPv6/IPv4 dual stack environment

When CAA is configured in a dual stack environment with isolated IPv4 and IPv6 networks, you observe the following behavior:

- ▶ IPv4 multicast is only sent through IPv4 adapters.
- ▶ IPv6 multicast is only sent through IPv6 adapters.

You can use the `tcpdump` command to confirm this behavior. Example A-16 shows the output in our environment.

Example A-16 Tcpdump in IPv6/IPv4 dual stack configuration

Multicast addresses:

IPv4 224.10.100.57

IPv6 ff05::e00a:6439

en0 is IPv4 isolated network

en1 is IPv6 isolated network

```

# tcpdump -i en0 ip multicast
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on en0, link-type 1, capture size 96 bytes
08:14:25.626648 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:25.626678 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:25.926655 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:25.926685 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:26.226669 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:26.226698 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:26.466735 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 208
08:14:26.526679 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
08:14:26.526705 IP glvmb1.drmsfsd > 224.10.100.57.drmsfsd: UDP, length 384
^C
165 packets received by filter
0 packets dropped by kernel
-----
# tcpdump -i en1 ip multicast <----- no IPv4 multicast through en1
tcpdump: WARNING: en1: no IPv4 address assigned
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on en1, link-type 1, capture size 96 bytes
^C
150 packets received by filter
0 packets dropped by kernel
-----
# tcpdump -i en0 ip6 multicast <----- no IPv6 multicast through en0
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on en0, link-type 1, capture size 96 bytes
^C
377 packets received by filter
0 packets dropped by kernel
-----
# tcpdump -i en1 ip6 multicast
tcpdump: WARNING: en1: no IPv4 address assigned
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on en1, link-type 1, capture size 96 bytes
08:15:20.828613 IP6 glvmb1ip6.drmsfsd > ff05::e00a:6439.drmsfsd: UDP, length 384
08:15:20.828636 IP6 glvmb1ip6.drmsfsd > ff05::e00a:6439.drmsfsd: UDP, length 384
08:15:21.128621 IP6 glvmb1ip6.drmsfsd > ff05::e00a:6439.drmsfsd: UDP, length 384
08:15:21.128645 IP6 glvmb1ip6.drmsfsd > ff05::e00a:6439.drmsfsd: UDP, length 384
^C
20 packets received by filter
0 packets dropped by kernel

```

The **mping** command can be used to verify that your network is correctly configured for multicast packets. For an overview of how to use the **mping** command, refer to [multicast in a network verification at](#):

http://pic.dhe.ibm.com/infocenter/aix/v6r1/topic/com.ibm.aix.powerha.trgd/ha_trgd_test_multicast.htm

Example A-17 on page 490 shows the results of the IPv4 multicast address verification in our test environment.

Important: When the **hostname** is configured to an IPv6 label, we observed that the **mping** command fails with the following error:

```
-----  
# host `hostname`  
glvmaip6 is 2000::c0a8:6437  
  
# mping -v -s -c 5 -a 224.10.100.57  
mping version 1.1  
gethostbyname() failed in mping_init_socket(): Error 0  
-----
```

APAR IV34031 is opened for this issue.

The output shown in Example A-17 was taken when the **hostname** command was configured with an IPv4 label.

Example A-17 IPv4 multicast verification with the mping command

```
glvmb1ip6:  
# mping -v -r -c 5 -a 224.10.100.57  
mping version 1.1  
Connecting using IPv4.  
Listening on 224.10.100.57/4098:  
  
Replies to mping from 10.10.100.58 bytes=32 seqno=0 ttl=1  
Replies to mping from 10.10.100.58 bytes=32 seqno=1 ttl=1  
Replies to mping from 10.10.100.58 bytes=32 seqno=2 ttl=1  
Replies to mping from 10.10.100.58 bytes=32 seqno=3 ttl=1  
Replies to mping from 10.10.100.58 bytes=32 seqno=4 ttl=1  
#  
-----  
glvmb2ip6:  
# mping -v -s -c 5 -a 224.10.100.57  
mping version 1.1  
Connecting using IPv4.  
mpinging 224.10.100.57/4098 with ttl=1:  
  
32 bytes from 10.10.100.57 seqno=0 ttl=1 time=0.196 ms  
32 bytes from 10.10.100.57 seqno=1 ttl=1 time=0.160 ms  
32 bytes from 10.10.100.57 seqno=2 ttl=1 time=0.161 ms  
32 bytes from 10.10.100.57 seqno=3 ttl=1 time=0.153 ms  
32 bytes from 10.10.100.57 seqno=4 ttl=1 time=0.156 ms  
Sleeping for 1 second to wait for any additional packets to arrive.  
  
--- 224.10.100.57 mping statistics ---  
5 packets transmitted, 5 packets received, 0% packet loss  
round-trip min/avg/max = 0.153/0.165/0.196 ms
```

If the **mping** command is issued with the **-a** flag, the IPv6 multicast addresses can also be specified. Example A-18 shows the results of an IPv6 multicast address in our test environment.

Example A-18 IPv6 multicast verification with the mping command

```
glvmb1ip6:  
# mping -v -r -c 5 -a ff05::e00a:6439  
  
mping version 1.1  
Connecting using IPv6.  
Listening on ff05::e00a:6439/4098:  
  
Replies to mping from 2000:aaaa::a0a:643a bytes=48 seqno=0 ttl=1  
Replies to mping from 2000:aaaa::a0a:643a bytes=48 seqno=1 ttl=1  
Replies to mping from 2000:aaaa::a0a:643a bytes=48 seqno=2 ttl=1  
Replies to mping from 2000:aaaa::a0a:643a bytes=48 seqno=3 ttl=1  
Replies to mping from 2000:aaaa::a0a:643a bytes=48 seqno=4 ttl=1  
-----  
glvmb2ip6:  
# mping -v -s -c 5 -a ff05::e00a:6439  
mping version 1.1  
Connecting using IPv6.  
mpinging ff05::e00a:6439/4098 with ttl=1:  
  
48 bytes from 2000:aaaa::a0a:6439 seqno=0 ttl=1 time=0.247 ms  
48 bytes from 2000:aaaa::a0a:6439 seqno=1 ttl=1 time=0.210 ms  
48 bytes from 2000:aaaa::a0a:6439 seqno=2 ttl=1 time=0.199 ms  
48 bytes from 2000:aaaa::a0a:6439 seqno=3 ttl=1 time=0.239 ms  
48 bytes from 2000:aaaa::a0a:6439 seqno=4 ttl=1 time=0.202 ms  
Sleeping for 1 second to wait for any additional packets to arrive.  
--- ff05::e00a:6439 mping statistics ---  
5 packets transmitted, 5 packets received, 0% packet loss  
round-trip min/avg/max = 0.199/0.219/0.247 ms
```

CAA command enhancements

The **lscluster** command has been enhanced to provide the following new support for IPv6:

```
lscluster { -i | -d | -c [ -n clustername ] } | { -m [ nodename ] | -s | -i  
interfacename | -d diskname }  
  
-m option will display PROTOCOL as ipv4 or ipv6.  
-i option has no change.  
-s option will display network statistics with ipv6 additions.
```

Example A-19 shows **lscluster -m** output displaying the PROTOCOL information.

Example A-19 lscluster -m example showing new PROTOCOL information

```
Node name: glvma2ip6  
Cluster shorthand id for node: 3  
UUID for node: 4109e892-42be-11e2-a5a7-7a40c30b1f03  
State of node: UP  
Smoothed rtt to node: 7  
Mean Deviation in network rtt to node: 3
```

```

Number of clusters node is a member in: 1
CLUSTER NAME      SHID          UUID
glvma1_cluster    0             40f4497e-42be-11e2-a5a7-7a40c30b1f03
SITE NAME         SHID          UUID
siteA              1             40ec25f0-42be-11e2-a5a7-7a40c30b1f03

```

Points of contact for node: 3

Interface	State	Protocol	Status
dpcom	DOWN	none	RESTRICTED
en0	UP	IPv4	none
en1	UP	IPv6	none

The **lscuster -s** command now shows IPv6 statistics in Example A-20.

Example A-20 Extract from lscuster -s showing IPv6 statistic information

IPv6 pkts sent: 1040346	IPv6 pkts recv: 4857164
IPv6 frags sent: 63	IPv6 frags recv: 0

Migrating steps from IPv4 to IPv6

If you are planning to migrate your cluster network configuration from IPv4 to IPv6, then the following steps should be taken:

1. Add your new IPv6 addresses to your configuration.

Without removing any IPv4 configuration, use the **mktcpip** command to add an IPv6 address to each of the nodes in your cluster.

Tip: IPv4 addresses at this point should still exist. Your node hostnames should still resolve to IPv4 addresses.

2. Stop cluster services on each node and make the changes.

On one node at a time, stop cluster services and make the changes:

```
clstartstop -stop -n <clustername> -m <node hostname>
```

Ensure that the changes are made in:

- Node network configuration
- NDPD router
- DNS Server

Note: Enable IPv6 address resolution of the node to the same hostname which earlier resolved to a IPv4 address. Ensure that forward and reverse lookup are configured the same.

Restart cluster services after the changes using:

```
clstartstop -start -n <clustername> -m <node-hostname>
```

3. Removing an IPv4 configuration.

Once you have completed the changes on all nodes, use the **rmtcpip** command to remove the existing IPv4 addresses from each node that previously resolved to the hostname. Migration is now complete.

DNS change for the IBM Systems Director environment with PowerHA

This appendix complements the implementation of a web server in the environment described in 6.3, “Configuring PowerHA SystemMirror 7.1.2 Enterprise Edition with SVC remote copy” on page 204, and configured in 9.3, “Configuring IBM PowerHA SystemMirror 7.1.2 Enterprise Edition using IBM Systems Director” on page 415. We installed a web browser on the IBM Systems Director server so we could connect to the web server using the DNS to the correct “webserver” IP address.

This appendix contains the following:

- ▶ Configuring DNS on an IBM Systems Director server
- ▶ Configuring the application server scripts
- ▶ Testing the application

Configuring DNS on an IBM Systems Director server

The IBM Systems Director server uses name resolution to show names on the GUI and connect to the clients. The same DNS used for Director is also used in this appendix to change the IP address for the “webserver” name when the application is running on one site or the other. To do that we configured /etc/resolv.conf as shown in Example B-1.

Example B-1 IBM Systems Director /etc/resolv.conf

```
root@ibmdirector / # cat /etc/resolv.conf
nameserver 192.168.100.64
domain itsolab.redbook.com
```

Our DNS server was configured on 192.168.100.64 on a Linux server. However, the Linux server configuration is not included in this appendix.

Configuring the application server scripts

We installed IBM IHS to be the application in our environment. This provided an easy tool to demonstrate the IP change concept. The application start script not only contains the command to start our application, but before it does that the script issues a command to the DNS server that changes the name table to the new IP address, as shown in Example B-2.

Example B-2 Start script

```
root@svca1:/opt/software>cat /opt/software/start.sh
if [[ `hostname` == svca* ]]
then
    ssh 192.168.100.64 "/root/chdns.sh sitea"
else
    ssh 192.168.100.64 "/root/chdns.sh siteb"
fi
/opt/IBM/HTTPServer/bin/apachectl start
```

The script on the DNS server depends on the DNS server you are using.

Testing the application

We started with cluster services up and the Resource Group online on SiteA as shown in Figure B-1.

root@svca1:/>/usr/es/sbin/cluster/utilities/cLRGinfo		
Group Name	Group State	Node
ihc_app_rg	ONLINE	svca1
	OFFLINE	svca2
	OFFLINE	svcb1

Figure B-1 Services online on SiteA

Now we opened the Web browser installed on the IBM Systems Director server and pointed it to the web server. This connected to the IP 10.10.10.10 as shown in Figure B-2.

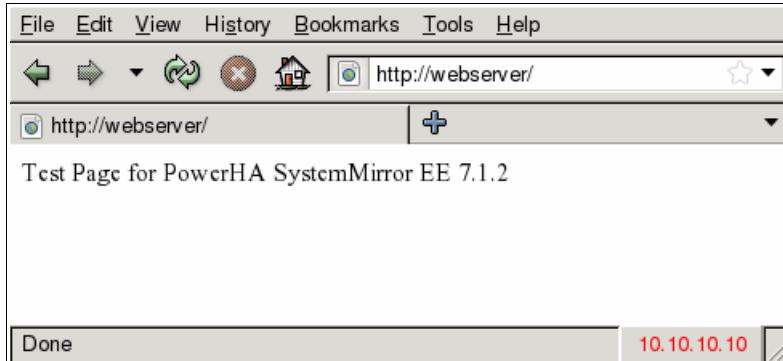


Figure B-2 Web server on SiteA

Now we moved the resource Group to the svcb1 node. To do this, we executed **smitty sysmirror** → **System Management (C-SPOC)** → **Resource Group and Applications** → **Move Resource Groups to Another Node**. We selected *ihs_app_rg* that was online on node svca1, and chose the *svcb1* node as the destination node, as shown in Figure B-3.

Move Resource Group(s) to Another Node	
Type or select values in entry fields. Press Enter AFTER making all desired changes.	
Resource Group(s) to be Moved	[Entry Fields]
Destination Node	ihs_app_rg svcb1

Figure B-3 Moving the resource group

Figure B-4 shows that the resource group successfully moved from svca1 to svcb1.

root@svca1:/>/usr/es/sbin/cluster/utilities/cTRGinfo		
Group Name	Group State	Node
ihs_app_rg	OFFLINE	svca1
	OFFLINE	svca2
	ONLINE	svcb1

Figure B-4 Resource group on SiteB

We next opened another browser window. Since the connection was broken, the old connection had the IP cached and failed to load the application, and pointed it again to the web server. Now it connected to the 10.10.20.10 IP address as shown in Figure B-5 on page 498.



Figure B-5 Web server on SiteB

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- ▶ *Implementing the IBM System Storage SAN Volume Controller*, SG24-6423
- ▶ *IBM XIV Storage System: Copy Services and Migration*, SG24-7759
- ▶ *IBM PowerHA SystemMirror Standard Edition 7.1.1 for AIX Update*, SG24-8030
- ▶ *PowerHA Enterprise Edition in the Exploiting IBM PowerHA SystemMirror Enterprise Edition*, SG24-7841
- ▶ *IBM System Storage DS8000: Copy Services in Open Environments*, SG24-6788
- ▶ *IBM System Storage DS8000: Architecture and Implementation*, SG24-8886

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- ▶ *IBM PowerHA SystemMirror* manuals
<http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.navigation/powerha.htm>
- ▶ *Fibre Channel Planning and Integration: User's Guide and Service Information*
<http://publibfp.boulder.ibm.com/epubs/pdf/c2343293.pdf>

Online resources

These websites are also relevant as further information sources:

- ▶ Storage-based high availability and disaster recovery for PowerHA SystemMirror Enterprise Edition:
http://pic.dhe.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.powerha.pprc/ha_pprc_main.htm

- ▶ PowerHA SystemMirror planning publication:
http://pic.dhe.ibm.com/infocenter/aix/v6r1/topic/com.ibm.aix.powerha.plangd/hacmpplangd_pdf.pdf
- ▶ IBM FixCentral pageDescription3:
<http://www-933.ibm.com/support/fixcentral/>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

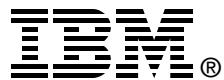
IBM



Redbooks

IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX

(1.0" spine)
0.875" <-> 1.498"
460 <-> 788 pages



IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX



**Unleashes IBM
PowerHA
SystemMirror**

**Describes new
features and
functionalities**

**Includes
implementation
scenarios**

This IBM Redbooks publication helps strengthen high availability solutions for IBM Power Systems with IBM PowerHA SystemMirror Enterprise Edition (hardware, software, and tools) with a well-defined and documented deployment model within an IBM Power Systems environment, offering clients a planned foundation for a dynamic, highly available infrastructure for their enterprise applications.

This book addresses topics to leverage the strengths of IBM PowerHA SystemMirror 7.1.2 Enterprise Edition for AIX on IBM Power Systems to solve client application high availability challenges, and maximize system availability, and management. The book examines the tools, utilities, documentation, and other resources available to help the IBM technical teams provide solutions and support for IBM high availability solutions with IBM PowerHA in an IBM Power Systems environment.

This book is targeted toward technical professionals (consultants, technical support staff, IT Architects, and IT Specialists) responsible for providing high availability solutions and support with IBM Power Systems and IBM PowerHA.

**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**