IBM

Exploiting IBM PowerHA SystemMirror V6.1 for AIX Enterprise Edition

- Highlights the benefits of deploying PowerHA Enterprise Edition
- Includes multiple implementation scenarios
- Describes networking planning and design

Dino Quintero
Benny Abrar
Adriano Almeida
Michael Herrera
SoHee Kim
Bjorn Roden
Andrei Socoliuc
Tom Swart

Redbooks

ibm.com/redbooks

International Technical Support Organization

Exploiting IBM PowerHA SystemMirror V6.1 for AIX Enterprise Edition

May 2013

SG24-7841-01

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

Second Edition (May 2013)

This edition applies to IBM PowerHA SystemMirror for AIX Enterprise Edition Version 6.1 on IBM AIX Operating System Version 6.1 TL4.

Contents

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®
AIX 5L™
DB2®
DS4000®
DS6000™
DS8000®
Enterprise Storage Server®
ESCON®
FlashCopy®
Global Technology Services®
GPFS™
HACMP™
IBM®
Micro-Partitioning®
NetView®
Power Systems™
Power Systems Software™
POWER6®
POWER7®
PowerHA®
PowerVM®
Redbooks®
Redbooks (logo)  ®
System i®
System p®
System Storage®
SystemMirror®
Tivoli®

The following terms are trademarks of other companies:

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

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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

Preface

This IBM® Redbooks® publication positions the IBM PowerHA® SystemMirror® V6.1 for AIX® Enterprise Edition as the cluster management solution for high availability. This solution enables near-continuous application service and minimizes the impact of planned and unplanned outages.

The primary goal of this high-availability solution is to recover operations at a remote location after a system or data center failure, establish or strengthen a business recovery plan, and provide separate recovery location. The IBM PowerHA SystemMirror Enterprise Edition is targeted at multisite high-availability disaster recovery.

The objective of this book is to help new and existing PowerHA customers to understand how to plan to accomplish a successful installation and configuration of the PowerHA SystemMirror for AIX Enterprise Edition.

This book emphasizes the IBM Power Systems™ strategy to deliver more advanced functional capabilities for business resiliency and to enhance product usability and robustness through deep integration with AIX, affiliated software stack, and storage technologies. PowerHA SystemMirror is designed, developed, integrated, tested, and supported by IBM from top to bottom.

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 (ITSO), Poughkeepsie Center.

Dino Quintero is a Technical Project Leader and an IT Generalist with the ITSO in Poughkeepsie, NY. His areas of expertise include enterprise continuous availability planning and implementation, enterprise systems management, virtualization, and clustering solutions. He is an Open Group Master Certified IT Specialist - Server Systems. He holds a master degree in Computing Information Systems and a Bachelor of Science degree in Computer Science from Marist College.

Benny Abrar is a Senior Technical Sales Specialist - Power Systems in IBM Indonesia. He has almost 10 years of experience in the IT field. His areas of expertise include IBM AIX, IBM PowerVM®, IBM PowerHA, and AIX Performance Tuning. He is an Open Group Certified IT Specialist - Server Systems. He holds a degree in mechanical engineering from Bandung Institute of Technology.

Adriano Almeida is an IT Specialist with the Integrated Technology Delivery in Brazil. He has worked at IBM for 11 years. His areas of expertise include AIX, PowerVM Virtualization, and IBM PowerHA and HACMP™. He is a Certified Advanced Technical Expert IBM System p®. He has worked extensively on PowerHA, PowerVM, and AIX projects. Adriano holds a degree in Computing Technology from the Faculdade de Tecnologia em Processamento de Dados do Litoral (FTPDL).

Michael Herrera is a Certified IT Specialist in the Advanced Technical Skills team out of Dallas, TX. He has worked with IBM for 11 years and was formerly a support escalation contact in the IBM AIX Support Line. He specializes in AIX/PowerHA and SAN environments and is certified by IBM as an Advanced Technical Expert. He has coauthored two IBM

Redbooks publications. Michael holds a bachelor degree in Management Information Systems from the University of Connecticut and an MBA from the University of Dallas.

SoHee Kim is an IT Specialist currently working in IBM Maintenance and Technical Support in GTS, IBM Korea. She has four years of experience in Power Systems. She is a Certified Advanced Technical Expert IBM System p. She has expertise in storage systems, IBM Power Systems hardware, virtualization on Power Systems, and AIX and PowerHA solutions. SoHee holds a bachelor degree in electronic engineering from Dankook University in Korea.

Bjorn Roden of Sweden is the LBS MEA Enterprise Power Technical Lead in Dubai, United Arab Emirates. He has worked with AIX since Version 2.2.1 and 3.1. He works with AIX, PowerVM, PowerHA, IBM GPFS™, and IBM Tivoli® Storage Manager during technical deployment for business continuity. He focuses on information availability, preservation, and nondisclosure, but also with high-performance for commercial or scientific workloads. He holds more than 30 certificates, including IBM Certified Infrastructure Systems Architect, Certified PRINCE2 Project Manager and TOGAF Enterprise Architect, and IBM Certified Technical Leader. Bjorn has coauthored five other Redbooks publications and reviewed numerous other related publications. Bjorn has MSc, BSc, BCSc, DiplCSc, and DiplSSc degrees in Computer Science and Informatics.

Andrei Socoliuc is a Certified IT Specialist - Systems and Infrastructure, working in IBM Global Technologies Services Romania. He has more than 10 years of experience in IT infrastructure. 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 he is also a coauthor of various HACMP IBM Redbooks publications. Andrei holds a Master of Science in Computer Science from the University Politechnical of Bucharest.

Tom Swart is a Senior Software Engineer at IBM, currently with Global Technology Services® in the United States. He has 20 years of experience with IBM doing Level 2 customer support. His areas of expertise include installation and problem diagnosis of the HAGEO, GLVM, RSCT, and other HPC components. He holds a bachelor degree in Computer Science from SUNY Potsdam, NY.

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

David Bennin
Ella Buslovic
Richard Conway
Scott Vetter
ITSO, Poughkeepsie Center

Patrick Buah
Michael Coffey
Paul Herb
Robert McNamara
Paul Moyer
Skip Russell
Vee Savath
Steve Tovcimak
Mark Winrow
IBM Poughkeepsie

Octavian Lascu
IBM Romania

Guilherme Gallopini Felix
IBM Brazil

Shawn Bodily
Steven Finnes
Francisco Garcia
Nick Harris
Kevin Henson
Kam Lee
Glenn E. Miller
Rohit Krishna Prasad
Ravi Shankar
David Truong
Thomas Weaver
Joe Writz
IBM US

Claudio Marcantoni
IBM Italy

Tony Steel
IBM Australia

Michael Malicdem
IBM Philippines

Bernhard Buehler
IBM Germany

Joergen Berg, Christian Schmidt
IBM Denmark

Alex Abderrazag
Rosemary Killeen
IBM UK

Laszlo Niesz
IBM Hungary

Bernard Goelen
IBM Belgium

David Arrigo
Paul Bessa
James Murray
EMC Corporation

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/IBM-Redbooks
- 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

Summary of changes

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

Summary of Changes
for SG24-7841-01
for *Exploiting IBM PowerHA SystemMirror V6.1 for AIX Enterprise Edition*
as created or updated on February 18, 2014.

May 2013, Second Edition

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

New information

This publication contains the following additional chapters:

- Chapter 10, “Disaster recovery with DS8700 Global Mirror” on page 469
- Chapter 11, “Disaster recovery by using Hitachi TrueCopy and Universal Replicator” on page 515

Part 1

Introduction

This part provides an overview of the IBM PowerHA SystemMirror Enterprise Edition solution and details about the infrastructure considerations that clients must be aware of before implementing the solution.

This part includes the following chapters:

1

High-availability and disaster recovery overview

This book describes the disaster recovery offerings in the IBM PowerHA SystemMirror Enterprise Edition clustering software for AIX on Power Systems. It also demonstrates how the advanced capabilities can satisfy even the most challenging disaster recovery scenarios. Before you read this book, you must be familiar with the implementation and administration of the base PowerHA clustering software. This chapter describes the following topics:

- Progression of the PowerHA Enterprise Edition replication technologies
- Key differences between local high availability and multisite disaster recovery
- Options to consider when deciding whether to automate failovers in a multisite environment
- Architectures that are available when selecting the optimum cluster design for your highly available environment

This book educates IBM Power Systems customers who use the PowerHA SystemMirror solution and helps align their resources with the best disaster recovery model for their environment. With each technology refresh and new server consolidation efforts, it is not unreasonable for a client to consider using their existing servers for use in their recovery environment. Customers looking for an entry point into a high-availability solution that incorporates disaster recovery can use their older hardware and select the replication mechanism that best fits their needs.

This chapter provides an introduction to the topic. For additional information consult the product documentation and the references provided in “Related publications” on page 573. The chapter includes the following sections:

- Introduction
- Disaster recovery and PowerHA
- Application data integrity
- Selecting the correct solution
- PowerHA enterprise logistics
- What is new on PowerHA Enterprise Edition

1.1 Introduction

The IBM PowerHA clustering solution has undergone changes over the last couple of years. The most obvious change is the rebranding from its former HACMP name to PowerHA in 2008 with the 5.5 release. The 6.1 release in late 2009 updated the name to PowerHA SystemMirror for AIX. This change was a result of an IBM Power Systems Software™ initiative to provide an array of IBM solutions, such as the following examples, under the PowerHA name:

- IBM PowerHA SystemMirror for AIX
- IBM PowerHA SystemMirror for System i®
- IBM PowerHA DB2® PureScale

To simplify the offerings, the base clustering product is now sold as the *Standard Edition* and provides local clustering functions. The former HACMP/XD or Extended Distance that was intended for disaster recovery is now bundled into the *PowerHA SystemMirror Enterprise Edition*. Table 1-1 compares the features that are offered by these editions.

Table 1-1 Standard Edition versus Enterprise Edition features comparison

	Standard Edition	Enterprise Edition
Centralized management C-SPOC	✓	✓
Cluster resource management	✓	✓
Shared storage management	✓	✓
Cluster verification framework	✓	✓
Integrated heart beating	✓	✓
SMIT management interfaces	✓	✓
AIX event/error management	✓	✓
Disk heart beating	✓	✓
PowerHA Dynamic logical partition (LPAR) high-availability management	✓	✓
Smart assists	✓	✓
Multisite HA management	a	✓
PowerHA GLVM		✓
GLVM deployment wizard		✓
IBM Metro Mirror support		✓
IBM Global Mirror support		✓
EMC SRDF/S SRDF/A		✓

a. A cross-site LVM mirroring configuration is an exception to this line item.

Historically, the Enterprise Edition was mistaken for a single offering that addresses all areas as they pertain to data replication and site failover. This extension to the base product is in essence an umbrella of integrated solutions that can be fitted into various environments, depending on a multitude of factors.

In early versions of the High Availability Cluster Multi-Processing (HACMP) software solution, the mechanism that is used to replicate data between sites revolved around the Geographic Remote Mirroring (GeoRM) product. The GeoRM product used the concept of a Geographic Mirror Device (GMD). GMD was a logical device that the application wrote to. It pointed to a logical volume on either site and would commit the following types of writes to them based on the policy that is specified:

- Synchronous
- Synchronous with mirror write consistency
- Asynchronous

Used in a stand-alone fashion, you could replicate in one direction and reverse the mirroring flow manually if required. When integrated with the HACMP cluster software, GeoRM became known as HAGEO. The HAGEO construct allowed for the automated control of the GMDs and role reversal of the replication between the sites depending on where the resources were being hosted. Although the GeoRM product does a good job of moving the data in a consistent way between sites. Because it is more complex to implement, you considerations to make because it pertained to dynamic operations within the cluster.

To address certain HAGEO limitations and to simplify the logical device implementation when using IP replication, in the HACMP 5.3 release in 2005, the development team introduced Geographic Logical Volume Mirroring (GLVM). Similar to the original GeoRM product, GLVM can be installed from the base AIX 5.3 or AIX 6.1 media and used as a stand-alone product for IP replication between sites. However, its integration with the IBM PowerHA Enterprise Edition automates the management and role reversal of replicated resources if planned and unplanned movements occur. The initial release of GLVM provided a synchronous-only mode of IP replication as an alternative to HAGEO.

The GLVM product is much simpler to use because it is ultimately built on the concept of logical volume mirroring. After you install the GLVM file sets and establishing IP connectivity between the sites, you can make LUNs that are in a remote site appear as though they are local. By adding those LUNs into your volume groups and using them to create logical volume mirrors, you can have copies of the data at the local sites and remote sites. Because of its sync-only mode, at the time it was deemed necessary to continue to sustain the HAGEO offering. However, the current PowerHA 5.5 and 6.1 releases that are running on IBM AIX 6.1 can now replicate synchronously and asynchronously by using GLVM.

End-of-support dates: The older HAGEO functions in the HACMP/XD releases are no longer available as of the PowerHA 5.5 release. For information about the migration from HAGEO to GLVM, see 8.11, “Migration from HAGEO (AIX 5.3) to GLVM (AIX 6.1)” on page 423.

In its entirety, the IBM PowerHA SystemMirror Enterprise Edition provides IP and disk replication alternatives, including integration with IBM System Storage® SAN Volume Controller Metro and Global Mirror and Metro Mirror with the IBM TotalStorage Enterprise Storage Server® 800, IBM System Storage DS6000™, and IBM System Storage DS8000® storage subsystems. New in PowerHA 6.1 is the integration with EMC Symmetrix Remote Data Facility (SRDF) disk replication technology. In addition, you can replicate data by using the wide area network (WAN), which uses the IP replication functions of the GLVM product. The GLVM product as a standalone product is available in the AIX 5.3 and 6.1 releases.

Important: Asynchronous GLVM is available only with PowerHA 5.5 SP1 and later or with PowerHA 6.1 running on AIX 6.1. This availability is a result of the new mirror pool and asynchronous I/O (aio) cache logical volume constructs introduced in the AIX 6.1 release.

1.2 Disaster recovery and PowerHA

If you watch the news or experienced something as simple as a power outage at your work site, it is obvious why we need to consider backing up data to a remote location. Outages can result from natural catastrophic events, such as fire, tornado, or flood, or man-made incidents, such as a terrorist act. The concept of disaster recovery refers to the practice of providing an alternative, backup, or secondary site from which to resume the IT operations until the primary site is back in service. If a catastrophic business operation failure occurs, it might be required to run at the remote site indefinitely. 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. Today, little prevents you from sharing your disks between two servers and failing over manually, but something can be said for implementing a clustering solution that can detect failures and automate the acquisition and release of your resources.

Similarly, the use of one of the replication options within the PowerHA Enterprise Edition helps ensure that the data gets copied over. The clustering around it detects failures and triggers events based on need that orchestrate the reactivation of resources quickly if a site outage, hardware problem, or application failure (application event) occurs.

A common question is about the type of extra availability that can be attained by implementing a PowerHA solution, in particular one with the Enterprise Edition. The actual number varies from one environment to the next. High availability as a whole is intended offers the following advantages:

- Automate control over a set of critical resources.
- Harden the environment to avoid service interruptions if a component failure occurs. (PowerHA protects against a single hardware error.)
- Significantly shorten the amount of time that it takes to move those resources if an unexpected failure occurs.

The Enterprise Edition builds on the concept of a local cluster and incorporates the integration with either IP or disk replication technologies to copy data to a remote site.

The idea of a fast failover if a problem occurs, 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) are what make the design worth the investment. A customer should not enter a disaster recover planning session and expect to achieve the *five nines of availability* by solely implementing a clustering solution. Table 1-2 outlines the calculations for the uptime criteria.

Table 1-2 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

Planning a disaster recovery solution involves key considerations, such as accounting for the time that is for planned maintenance and whether that time deducted from the maximum

downtime figures for the year. The ability to nondisruptively update a cluster is available as of the PowerHA 5.4.0.1 release, although using this feature avoids the need for a restart when you update a cluster or upgrade to a new release. Therefore, you must consider the impact of other service interruptions in the environment, such as upgrades that involve the applications, the AIX operating system, and the system firmware, which often require the system to cycle.

The IBM PowerHA SystemMirror 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 failovers if an outage or natural disaster occurs. Figure 1-1 shows the various tiers of disaster recovery solutions and how the 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 Enterprise Edition provides. Looking over the recovery time axis, you can see how meeting an RTO of under four hours can be achieved with the implementation of an automated multisite clustering solution.

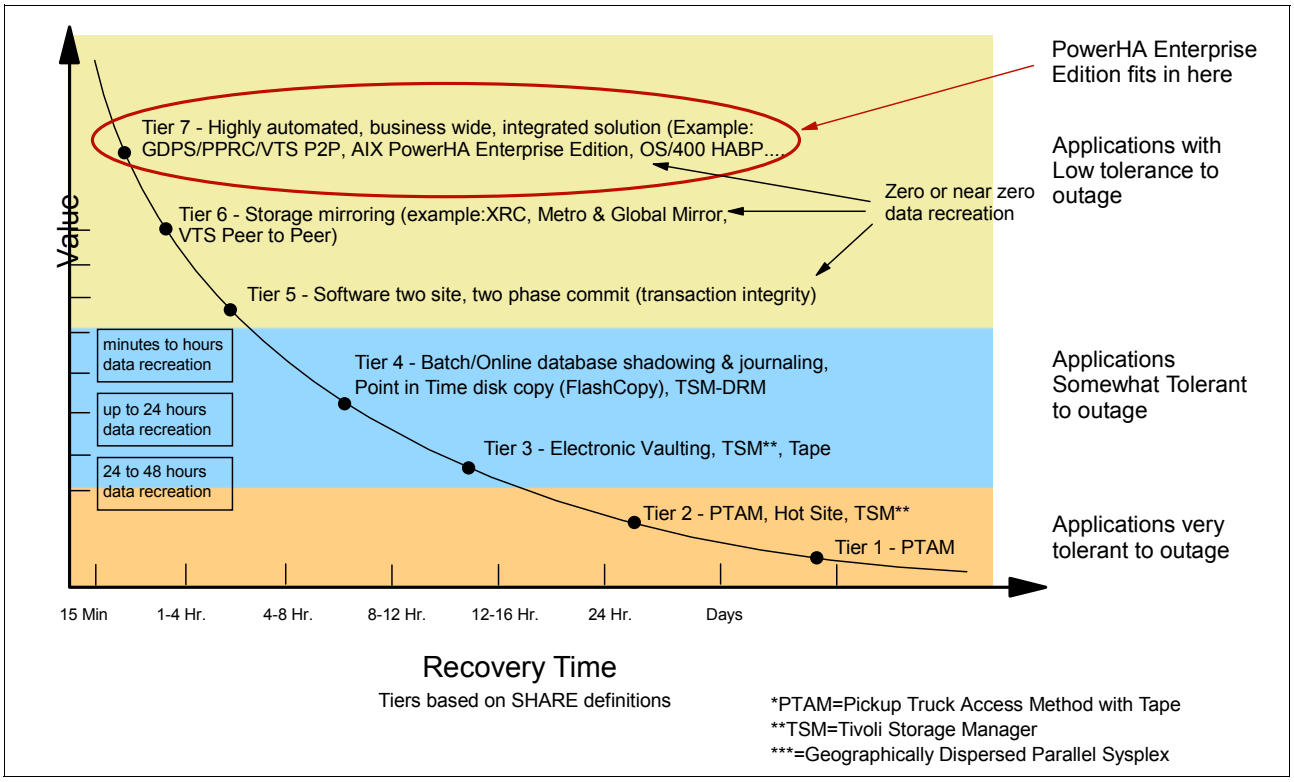


Figure 1-1 Tiers of disaster recovery solutions - IBM PowerHA SystemMirror Enterprise Edition

Table 1-3 describes the tiers of disaster recovery in more detail. Table 1-3 outlines different Power Systems solutions available in each Tier for disaster recovery solutions.

Table 1-3 Disaster recovery solution tiers - IBM disaster recovery planning model

Disaster recovery planning model reference	Power Systems solutions	100% recovery of application data possible?	Automatic detection of site failure?	Facility locations supported	Communication modes or protocols
Tier 7 Zero data loss. Recovery up to application restart in minutes.	PowerHA Enterprise Edition: ► Cross Site LVM ► GLVM ► Metro Mirror ► Global Mirror ► SRDF/S	Yes - minus data in transit at time of disaster	Yes - Failover and failback automated	All PowerHA Standard Edition suffices for a campus-style disaster recovery solution	Metropolitan area network (MAN) or WAN - IP SAN (dark fibre) - disk replication
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	All All All
Tier 5 Continuous electronic vaulting of backup data between active sites. Active data management at each site is provided.	IBM Tivoli Storage Manager with copy pool duplexing between sites, and active Tivoli Storage Manager servers at each site	No Recovery in days or weeks Must restore from backups	No	All	All
Tier 4 Electronic vaulting of critical backup data to a hot site. The hot site is not activated until a disaster occurs.	Tivoli Storage Manager with copy pool duplexing to the hot site, and Tivoli Storage Manager server at active site only	No Recovery in days or weeks	No		
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.	Tivoli Storage Manager with multiple local storage pools on disk and tape at active site	No Recovery in days or weeks	No		

Disaster recovery planning model reference	Power Systems solutions	100% recovery of application data possible?	Automatic detection of site failure?	Facility locations supported	Communication modes or protocols
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 hot site must be prepared. ► Backup data must be transported.	Tivoli Storage Manager with multiple local storage pools on disk and tape at active site	No Recovery in days or weeks	No		
Tier 1 Pickup Truck Access Method with tape.	Tape-backup-based solution	No	No		
Tier 0 No disaster recovery plan or protection.	Local backup solution may be in place, but no offsite data storage	No disaster recovery - site and data lost			

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

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

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

For various infrastructure considerations, see Chapter 2, “Infrastructure considerations” on page 39.

1.2.1 Local failover versus site failover

Customers who implement 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 that are at dispersed sites. Although from the cluster standpoint, the graceful release and reacquire functions are the same. In a scenario with remote sites, you always have a higher risk of data loss if a hard failure occurs. Therefore, it is often more appropriate to use a local HA cluster as the building block for a disaster recovery solution. Two local nodes are paired up within a site and at least one remote node is added to the cluster as part of a second site definition.

Restriction: Current PowerHA Enterprise Edition releases have a limitation of only two sites and a maximum eight nodes within a cluster.

We recognize that customers have service level agreements (SLAs) that require them to maintain a highly available environment after a site failure, and in turn we try to use 4-node cluster configurations for most of our test scenarios (Figure 1-2).

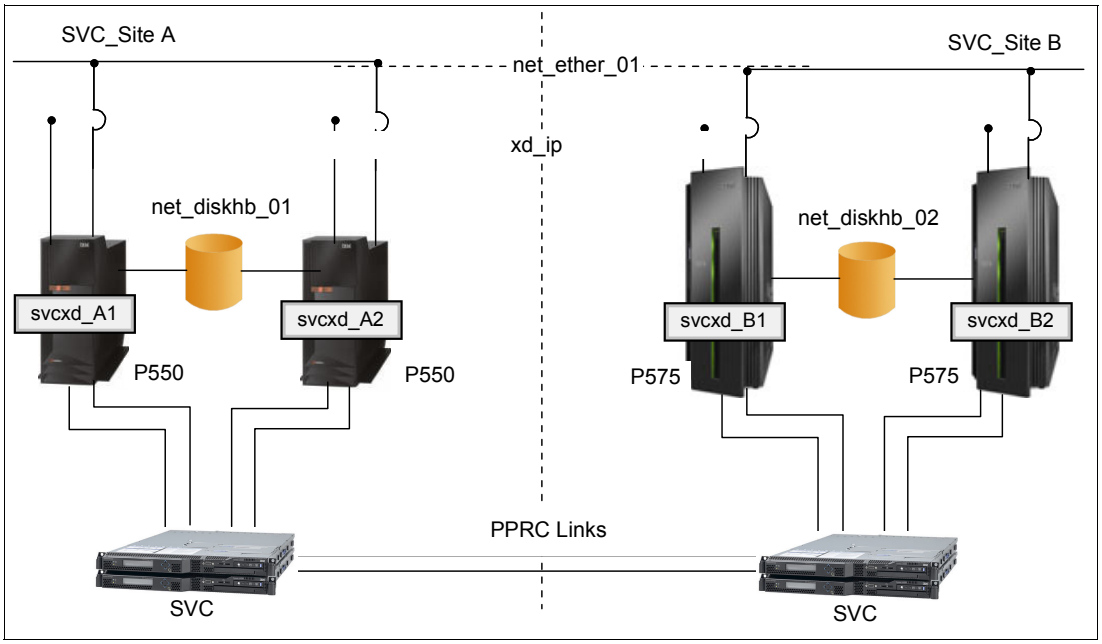


Figure 1-2 PowerHA Enterprise Edition with two sites that have a 4-node cluster

The line between local availability and disaster recovery is often denoted by the existence of multiple physical sites and distance between them, but with gray areas between the two. This setup is especially prevalent in environments that have multiple server rooms across different buildings within the same site. We typically consider these environments dispersed over a relatively close area to be campus-style hybrid disaster recovery environments. Banking institutions, universities, and hospitals are common examples, where an infrastructure is in place between separate server rooms, allowing IP and SAN connectivity between the different machines and independent storage subsystems (Figure 1-3).

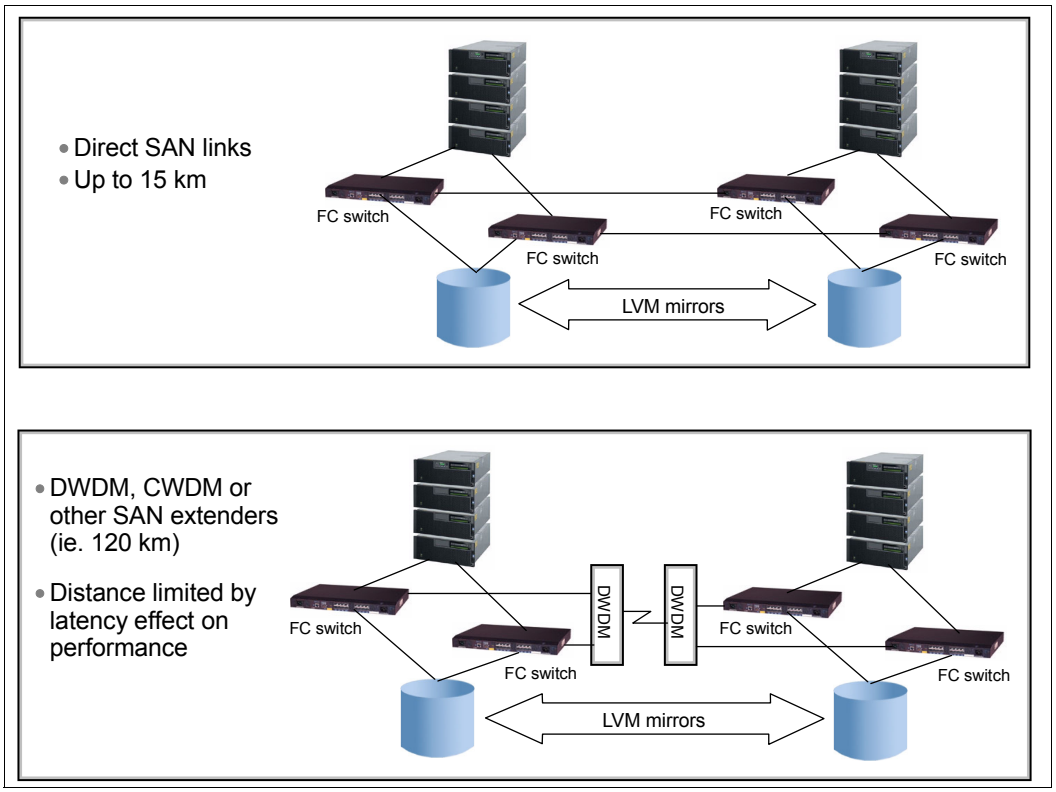


Figure 1-3 Cross Site LVM - direct SAN versus DWDM extended SAN

The PowerHA cluster nodes can be on different machine types and server classes if you maintain common AIX and cluster file set levels. This setup can be valuable in a scenario where a customer acquires newer machines and uses the older hardware to serve as the failover targets. In a campus environment, creating this type of a stretch-cluster can serve two roles:

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

These environments often present various options for how to configure the cluster. For example, they include a cross-site LVM mirrored configuration, one using disk-based metro-mirroring, or scenarios that use SAN Volume Controller VDisk mirroring with a split I/O group between two sites, new with the 5.1 SVC firmware release. Each option has its own merits and corresponding considerations that are explained more in Chapter 5, “Configuring PowerHA SystemMirror Enterprise Edition with Metro Mirror and Global Mirror” on page 153. Being inherently synchronous, all of these solutions experience minimal to zero data loss, similar to solutions in a local cluster that shares LUNs from the same storage subsystem.

Although every environment differs, more contention and disk latency are introduced the farther the sites are from each other. However, no hard set considerations dictate whether you

need to replicate synchronously or asynchronously. It can be difficult to provide an exact baseline for the distance that delineates synchronous versus asynchronous replication. We have customers that replicate synchronously between sites that are hundreds of miles apart and the configuration suits them well. The result is largely because their environments are mostly read intensive and writes occur only sporadically a few times a day. As a result, the impact of the application response time because of write activity is minimal. Therefore, consider such factors as the application read and write tendencies with current system usage.

Key differences between local high availability (HA) and disaster recovery revolve around the distance between the sites and ability, or inability, to extend the storage area network (SAN). Local failover provides a faster transition to another machine than a failover that goes to a geographically dispersed site. This fact is even clearer in a scenario where Logical Volume Manager (LVM) mirrored disks would mask the failure of one of the storage subsystems to the application and not require a failover. Environments that require the replication of data over greater distances, where asynchronous disk-based replication might be a better fit, might have a greater exposure for data loss. A larger delta might exist 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 were to go offline.

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

Physical resources: Only virtualized resources can be used in the VIO client (VIOC). However, if physical resources are necessary for any reason (for example, for performance considerations), you need to move it to virtualized adapters before migrating it by using Live Partition Mobility.

For more information about virtualization and how it pertains to PowerHA and disaster recovery, see 2.4, “PowerVM virtualization considerations” on page 76.

1.2.2 Independent versus integrated replication

The PowerHA software has been providing clustering and data replication functions for a several years. This software continues to strive to be the solution of choice for IBM customers who run on AIX. The software tightly integrates with various existing disk replication technologies under the IBM portfolio. It also recently extended integration of the EMC Symmetrix Remote Data Facility (SRDF) replication technology.

Historically, the main limiting factor for implementing disaster recovery solutions was cost. Today, as more clients reap the benefits of geographically dispersing their data, they face the following more apparent inhibitors:

- The risks that are associated with a fully automated clustered solution
- The associated added value of wrapping the cluster software around the replication

A multisite clustered environment is designed in such a way that, by default, if the primary site stops responding, the secondary site terminates the replication relationship and activates the

resources at the backup location. Plan for multiple redundant heart-beating links between the sites to avoid a separation. The PowerHA 5.5 release introduced the ability to disable the automated failover function for specific failures, for example, the loss of the connectivity between the storage units that manage the replication. With this option selected the cluster prevents a failover to retain data integrity and the cluster processing requires manual intervention. Figure 1-4 shows a sample of this option.

```
                        Change / Show SVC PPRC Resource

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

                                                      [Entry Fields]
  SVC PPRC Consistency Group Name                     svc_metro
  New SVC PPRC Consistency Group Name                 []
* Master SVC Cluster Name                             [B8_8G4]                +
* Auxiliary SVC Cluster Name                          [B12_4F2]               +
* List of Relationships                               [svc_disk2 svc_disk3 s> +
* Copy Type                                           METRO                   +
* HACMP Recovery Action                               MANUAL                  +
```

Figure 1-4 PowerHA manual recovery action option

The status of the replicated resources determines whether the MANUAL recovery action prevents a failover. The states vary between the replication types (Example 1-1).

Example 1-1 Replication types, depending on the status of the replicated resources

```
DS PPRC states that would not failover:
   Target-FullDuplex-Source-Unknown
   Source-Suspended-Source-Unknown
   Source-Unknown-Target-FullDuplex
   Source-Unknown-Source-Suspended

SVC PPRC states that would not failover:
   idling_disconnected
   consistent_disconnected
```

Often times after a disaster is declared in a non-automated environment, the technical staff take longer to decide whether to fail over. The time that is passed when making that decision can significantly extend your recovery time. For the recovery steps for a partitioned multisite cluster within each scenario that is reviewed, see in Part 2, “Campus style disaster recovery” on page 95, and Part 3, “Extended distance disaster recovery short overview” on page 151.

If the cluster design has already accommodated for redundant links between the sites and in turn minimized the risk of a *false down*, we can continue to discuss the added value that the Enterprise Edition brings.

Many clients use the various disk and IP replication technologies without using the integration within the PowerHA Enterprise Edition only to discover 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. Even after scripting the commands to activate the disks in the proper order, in various cluster scenarios, you need to append extra logic to achieve the results. This situation in part is the reason that the GLVM file set is included in AIX (the concept of try and buy). After you identify that the replication technology meets your

needs and you sized the data links appropriately you quickly realize that the management is simpler when you use the integration within the Enterprise Edition. Similarly, using the integrated logic to pass instructions to the disk storage subsystems automatically based on the various events detected by the cluster is also more efficient when left to the integrated code.

A major benefit of Power HA Enterprise Edition is that it has been comprehensively tested to ensure that it works with the basic failover and failback scenarios and with rg_move and selective failover inherent cluster mechanisms, for example. In addition, using the Enterprise Edition automatically reverses the flow and restarts the replication after the original site is restored. The integrated cluster **`clverify`** functions also help to 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 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 occurs.

Enabling full integration can also facilitate the ability to gracefully move resources and the testing of your scripts. By 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. Using some of the more granular options within the cluster, such as resource group location dependencies, can also facilitate the destaging of lower priority test or development resources at the failover location whenever a production site failure occurs. By using the site-specific dependencies, you can also specify a set of resource groups to always coexist within the same site.

Support considerations

Customers and lab service staff have been customizing replication solutions within the PowerHA software for a while now. If the integration for a replication technology is not available, it can be automated by using custom events or within the application scripts. One example of this customization is the custom integration of EMC SRDF replication before the availability of PowerHA 6.1.

Many clients already replicate their clustered volumes between sites without cluster automation. In these environments, the data exists at the remote site and its activation in a failure scenario requires a manual or semi-scripted procedure. These clients are good candidates for a customized environment if their storage replication devices are not already integrated with the PowerHA Enterprise Edition. The major caveat with a customized environment is that support for any of the custom scripts ultimately falls on the administrator.

Any of the custom logic that is embedded within custom events or application scripts is not reviewed by the IBM support staff. Technically, it is not considered an unsupported configuration because basic cluster functions are beyond those custom scripts.

Summary

As mentioned in the previous sections, using the integrated replication solutions with the Enterprise Edition can help simplify the management and shorten the amount of time to recovery. The key here is that you must be confident in the clustering technology to handle all aspects of the failover process. For tiered applications, it especially makes sense to use automatic recovery to ensure the order in which the resources are brought back online.

As new technologies emerge and the complexity of the environment continues to increase, the product simplifies the cluster interface and assists the user in hardening the environment to build a robust disaster recovery solution.

The goal is to provide the most resilient solution that incorporates as much of the existing components within the environment as possible. If the goal is to replicate the data in a consistent way, each of the mutually exclusive replication technologies does so in an effective manner. However, merging the clustering and replication technologies provides a much more elegant and effective solution.

1.3 Application data integrity

One of the major concerns when you are considering a replication solution is whether it will maintain the integrity of the data after a hard site failover. The truth is that all of the solutions do a good job, but they are not perfect. Remember the concept of garbage in-garbage out (GIGO). Database corruption at the source site is replicated to the target copy, which is the main reason that the disaster recovery design does not stop after you select a replication technology.

Replicating the data addresses only one problem. In a well-designed disaster recovery solution, a backup and recovery plan must also exist. Tape backups, snapshots, and flash copies are still a part of an effective backup and recovery solution. For a thorough design, consider the frequency of those backups at both the primary and remote locations.

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

As we delve into how the separate solutions manage to maintain consistency, we first need to review how applications with dependent writes behave within a disaster recovery environment.

1.3.1 Disaster recover for applications with dependent writes

Many applications can manage the consistency of their data by using dependent writes. A common example of application-dependent writes is databases and their associated log files. Database data sets are related with values and pointers from indexes to data. The databases have pointers in the data sets, in the database catalog and directory data sets, and in the logs. It is easy to see why it is imperative to maintain data integrity across the various components.

As shown in Figure 1-5, committing a database write to disk has the following flow:

1. An update request is written to the log.
2. After the log is successfully updated, the data is updated.
3. After the database object is successfully updated, the log is again updated to mark the transaction as completed.

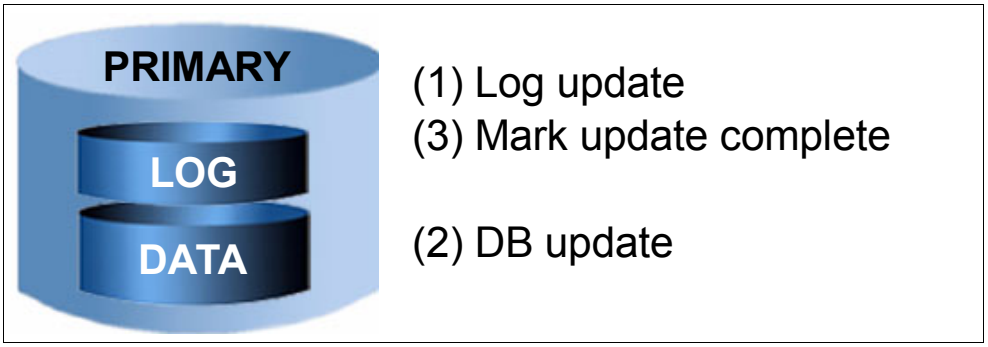


Figure 1-5 Database write flow

In an environment that replicates the volumes between storage subsystems, the flow has a few more updates. In a metro-mirrored environment, the database transaction sequence for updating DB objects normally has the following characteristics (Figure 1-6):

1a The update request is written to the database log on the primary volumes.

1b The DB log update is mirrored on the secondary volumes.

2a After the log is successfully updated, the database object is updated on the primary volumes.

2b The database object update is mirrored on the secondary volume.

3a After the database object is successfully updated, the database log is updated to mark that the transaction is completed.

3b The database log update is then mirrored on the secondary volume.

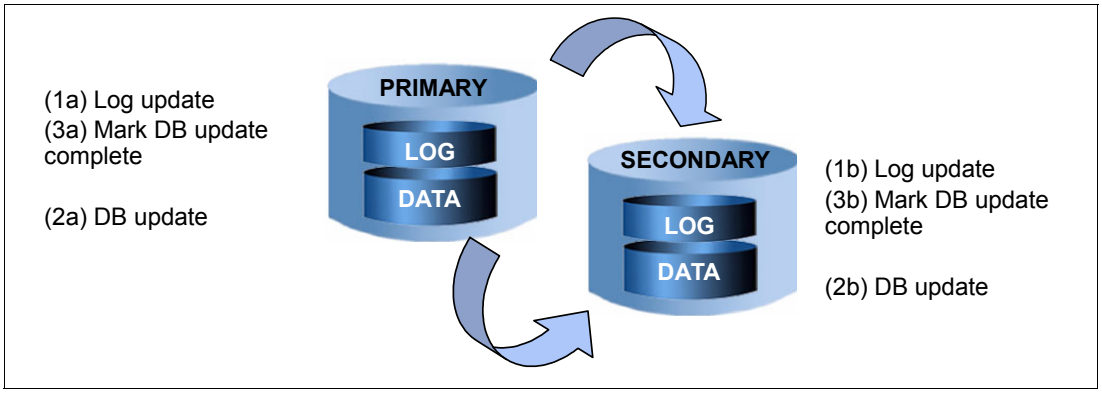


Figure 1-6 Database write flow in a replicated environment

One of the challenges in a replicated environment is that failures tend to be intermittent and gradual. It is unlikely that all of the components within a complex would fail at the same exact time. This type of event is called a rolling disaster. For example, consider a scenario where the storage links are suspended for a short period.

In the scenario shown in Figure 1-7, consider the order of events:

1. The link between the DB data volume and its mirrored pair at the recovery site is lost.
2. The write sequence of (1) - (2) - (3) completes on the primary devices.
3. Log writes (1) and (3) are mirrored on the log-sec device, however, the corresponding DB write (2) is not applied on db-sec.

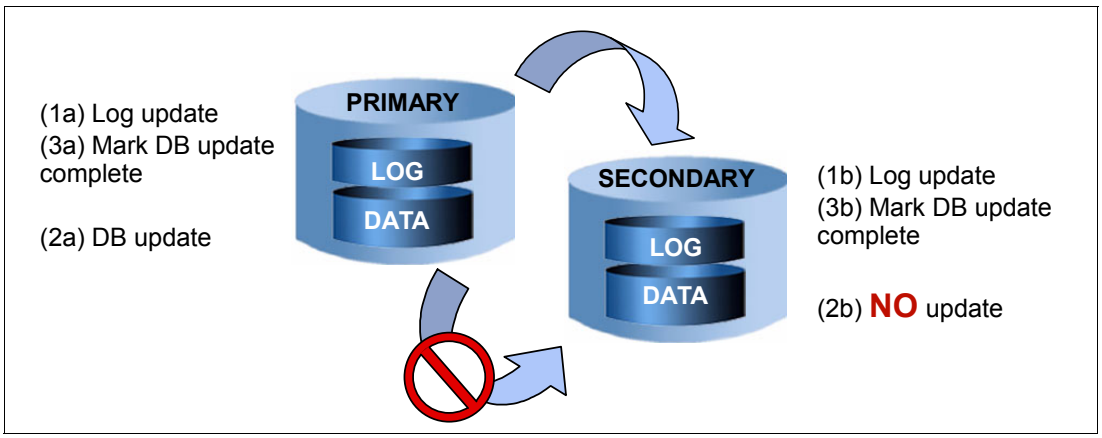


Figure 1-7 Database write failure scenario

With this type of failure, the database on the secondary site can end up having missing updates that might potentially go unnoticed for a time. This situation is considered unacceptable since there is no way to easily identify the problem. This area is one in which integration with the PowerHA Enterprise Edition can also provide added value. Depending on the storage type and replication technology that is implemented, separate methods are built in for mitigating these risks.

1.3.2 PowerHA Enterprise Edition SNMP trap support

Some environments might using the IBM TotalStorage Enterprise Storage Server 800, DS6000, or DS8000 subsystems with the PowerHA Enterprise Edition. If you enable the Simple Network Management Protocol (SNMP) trap function to share the traps with your cluster nodes, the cluster monitors any traps that are issued. Then it triggers an EVENT based on what the traps indicate.

The storage units must be enabled to send traps to the cluster nodes. The precise method is hardware-dependent. For more information, see the publications that are associated with the storage subsystem.

Before you use the following procedure to enable SNMP traps, ensure that any other subscriber of SNMP traps (such as IBM NetView®, Tivoli, or other network management software) is not already running. Otherwise, the PowerHA Cluster Information Daemon (clinfoES) is unable to receive traps. Conversely, if HACMP is configured to receive SNMP traps, no other management software is able to receive them.

PowerHA supports the receiving and handling of the following SNMP trap messages:

```
Generic Type = 6
```

The following SNMP trap types are possible for Metro Mirror consistency groups:

100 Link degraded
101 Link down
102 Link up
200 LSS pair consistency group error
201 Session consistency group error
202 LSS is suspended

PowerHA tests each SNMP trap that is received to ensure that it is from the following sources:

- A valid storage unit (checked by storage unit ID)
- A storage unit that is defined to PowerHA
- A logical subsystem (LSS) that was previously configured into a Peer-to-Peer Remote Copy (PPRC) resource group

If PowerHA receives an SNMP trap that does not meet this criteria, it is logged, but otherwise ignored. To enable SNMP traps in PowerHA, you must start the Cluster Information Daemon with SNMP traps enabled, which provides consistency group support.

Important: If clinfoES is already running, you must first stop and restart it with the corresponding flag for consistency group support as follows:

```
# stopsrc -s clinfoES
# startsrc -s clinfoES -a "-a"
```

When you restart clinfoES, you have the following options:

- False: Do not start cluster services.
- True: Start cluster services.
- True with consistency group support: Start cluster services with SNMP traps enabled.

Figure 1-8 shows the startup options for the clinfoES daemon when it is started from the SMIT menus with SNMP trap support.

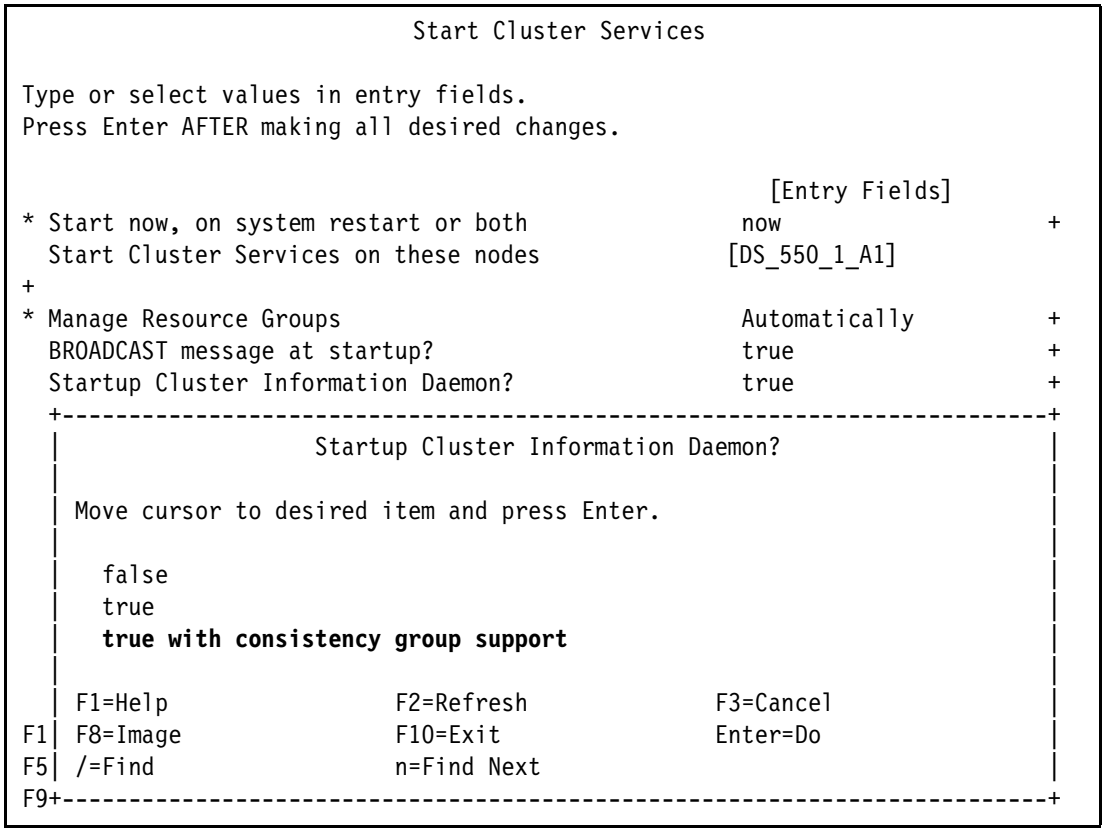


Figure 1-8 clinfoES start options for SNMP trap consistency group support

The clinfoES subsystem listens on port 162/tcp and uses the ERROR log to record the valid incoming trap. It then calls the resource_state_change script if the trap is for one of the online resources. It sets the global environment variable SNMP_DEVICE_INFO, which is used by the consistency group SNMP trap event script to process the trap.

There are different SNMP traps that trigger a cluster EVENT:

- PPRC link-degraded event
- PPRC link-down event
- PPRC link-up event
- LSS pair-consistency group PPRC-pair error event
- Session consistency group PPRC-pair error event
- Primary PPRC devices on an LSS that is suspended because of an error event

Figure 1-9 shows the following flow of a failure with the SNMP trap monitoring enabled:

1. An error is detected on storage subsystem and an SNMP trap notification is sent.
2. SNMP traps support (clinfoES) receives a trap and determines whether it is valid.
3. clinfoES writes the trap to the AIX errlog and sends the errlog ID and event type to the cluster manager.
4. The cluster manager parses the errlog for the ID, extracting the trap information, sends the event to the nodes in the cluster, targeting the node with the RG online, and sets the environment variable SNMP_DEVICE_INFO with the source storage ID and LSS pair information (Example 1-2).

Example 1-2 SNMP_DEVICE_INFO format

```
SNMP_DEVICE_INFO format is "storage_id:lss1:lss2"

storage_id is the full storage ID
LSS1 is the LSS ID on the storage unit where the trap originated from
LSS2 is the LSS ID paired with LSS1

Example: "IBM.2107-7516231:F7:15
```

5. The target node runs the resource_state_change script, which calls the **pprc_snmptrap_event** script (Figure 1-9).

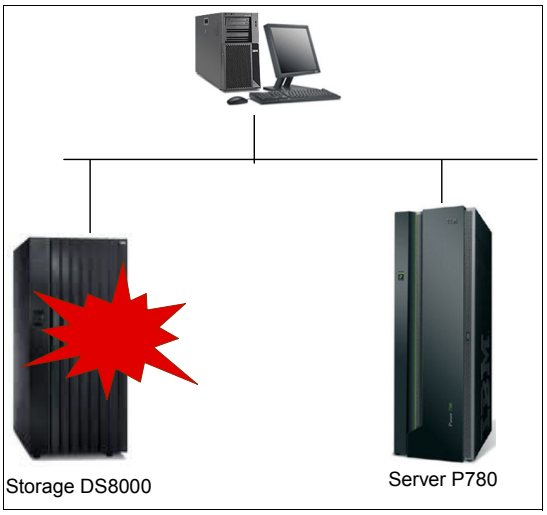


Figure 1-9 Flow of SNMP failure

resource_state_change event: All nodes run the resource_state_change event. On nodes that are not identified as the target node that is hosting the resource group, the script exits and does nothing.

In the scenario in Figure 1-10, a sent trap, which indicates that one of the links is down, is received and processed on the cluster node that manages the resource group. After the consistency group paths are evaluated and it is determined that for one of the paths the link is NotEstablished, the cluster freezes the replication and suspends all PPRC pairs in each consistency group.

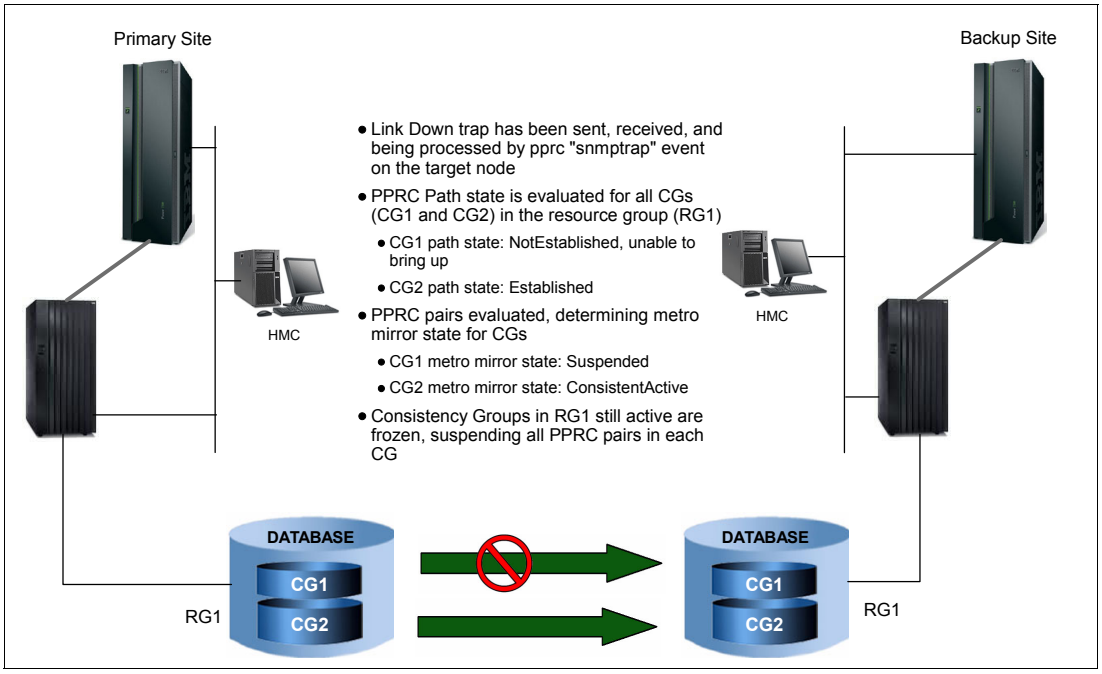


Figure 1-10 Diagram shows only one link down

Unavailable consistency group: When one consistency group becomes unavailable, mirrored writes to all other PPRC consistency group pairs within the resource group are frozen.

PPRC path states per consistency group (CG):

- Established: All paths in the CG are up (the case for CG1 in Figure 1-10)
- PartialEstablished: CG has more than one link (At least one link us up, which is acceptable.)
- NotEstablished: All paths in the CG are down (the case for CG2 in Figure 1-10)

Metro Mirror has the following states:

- ConsistentActive

  The source and target volumes are in a state of copy pending or full-duplex. This state indicates that the PPRC relationships are in sync (full duplex or copy pending). This state is the preferred state to be in for synchronous PPRC. Mirroring is already running primary to secondary (copy pending) and resync. Nothing is wrong with either the source volume or the target volume. They are consistent and active. No action is needed. (This state is the case for CG1 and CG2.)

- ConsistentSuspended

  All volumes in the LSS are in suspended state. The action that is needed is to remake and test links if they are not up. If the link is brought back up, the PPRC pair is resumed. If the PPRC pair cannot be resumed, all CGs in the RG are frozen. After pprc_snmptrap_event

freezes CGs, links are remade so we can receive future link-related traps (that is, LINK-UP).

- InconsistentSuspended

  This stat indicates that certain volumes are in the suspended state, while other volumes are in any other state (for example, full-duplex or copy pending). Certain volumes are mirroring while others are suspended. This case is handled the same as the ConsistentSuspended state.

- VolumesInactive

  This state indicates that the storage system/volume is unreachable. Either the storage system is down or the volume failed. This state triggers an rg_move to the remote site.

After a new trap is sent indicating that the link is reestablished, the nodes reevaluate the status by using the pprc_snmtrap_event and dynamically make the calls to the storage enclosure to unfreeze the consistency groups. The cluster then allows for the replication for all of the PPRC pairs to resume (Figure 1-11).

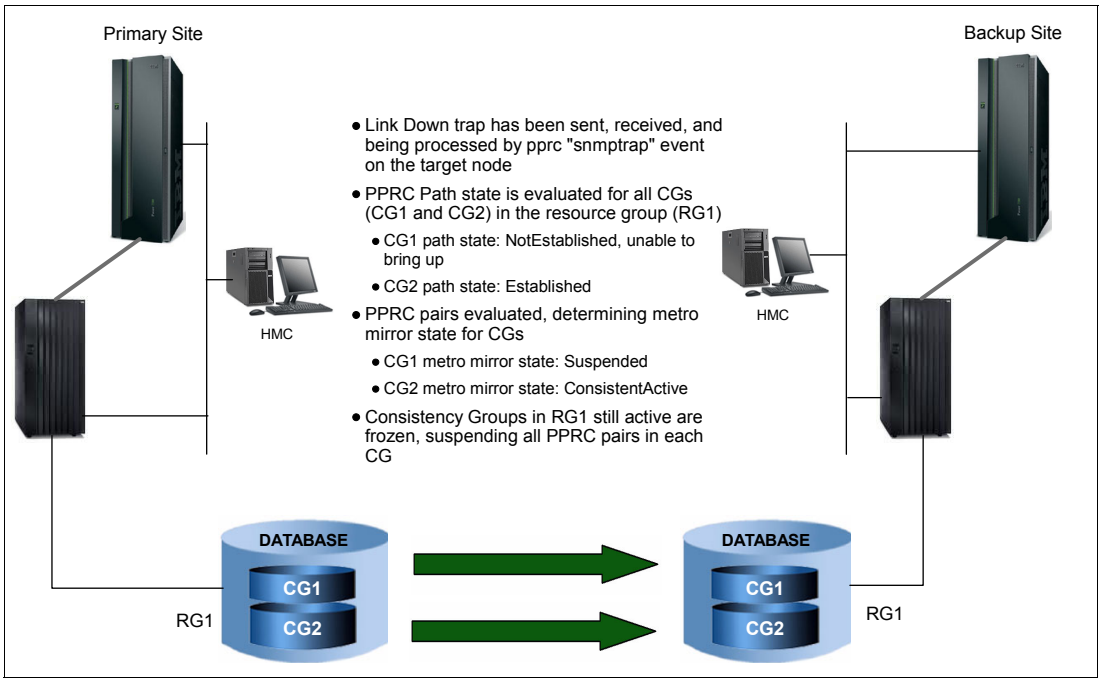


Figure 1-11 Replication links are restored

When using one of the alternative solutions, such as the sync/async replication with the SAN Volume Controller or the EMC subsystems, the built-in consistency group functions detect the loss of a link and freeze the relationships automatically. The SAN Volume Controller logs internal messages, indicating that the relationships are frozen. It is up to the client to configure the appropriate alerts for notification about the failure. The cluster software does not poll the paths to reactivate them. Clients can enable external monitoring to receive notification about the failure and must manually reestablish the links after the connectivity is resumed.

For more information about setting up external notification, see the *IBM System Storage SAN Volume Controller Troubleshooting Guide*, GC27-2227.

For more information about event monitoring for EMC Symmetrix, see the event daemon (storevntd) configuration that is described in the *EMC Solutions Enabler Version 7.0 - Installation Guide*, P/N 300-008-918.

1.4 Selecting the correct solution

Several facts can help you during the decision-making process to select the appropriate high-availability solution and in particular a solution that incorporates disaster recovery. First, you must identify certain key factors. Facility locations and telecommunication solutions are among the first factors or parameters to consider and finalize as part of building an effective disaster recovery plan. Communication path redundancy, bandwidth requirements, and application data volumes to be replicated are all critical factors that can have a large impact on the costs of the networking portion of the disaster recovery solution.

Synchronous transmission modes ensure data equivalence, but can impose a network performance penalty on the primary site application. This situation can affect local application response time and limit the distance between the sites.

1.4.1 Synchronous versus asynchronous replication

When you start to consider real distance for synchronous mirroring, realize that it is more a performance limit than an actual distance limitation.

Important: Synchronous mirroring is considered necessary if recovery of one 100% of intact data transactions is required.

Theoretically, if you assume an AIX I/O timeout baseline of 30 seconds (rw_timeout 30) for a disk, you can push a synchronous I/O over a 180,000-mile distance (Example 1-3). The response time would be incredibly slow for each I/O operation, but gets you thinking about the impact of introducing a synchronous mirror into the environment.

Example 1-3 Theoretical distance limit for synchronous mirroring

```
(M miles x .005 ms/km) / .6 mi/km = 30 seconds
M=360,000 miles round trip, or 180,000 miles
```

Along the same lines, if you have only a few transactions, a distance of 400 miles is not that significant. The latency of light through fiber at this distance, round trip, would translate to approximately 6.7 ms of added latency (Example 1-4).

Example 1-4 Sample overhead for 400-mile round-trip I/O

```
M x 2 (Round Trip) x .005ms/km / .6 mi/km
400 x 2 x.005 / .6 = 6.7 ms
```

Adding 6.7 ms to a transaction is insignificant regarding the actual time that it takes to complete the transaction. However, this setting affects the maximum system throughput in terms of transactions per second. If you then estimate the impact to the system throughput by relating it to the number of transactions per minute (tpm), you can speculate about it in more business-relevant terms.

Assume that each transaction occupies system memory and that you have an environment whose available memory supports 10,000 tpm by using locally attached drives. You introduce mirroring to a remote location, and it now takes those transactions 5 times longer to complete (2 ms for a local write and 10 ms for the corresponding remote write). In this case, you can estimate that the system now supports only 2000 tpm. If the application load is less than 2000 tpm, there is not a problem. However, if those counts were to rise, I/Os would queue up at the application layer as the application waited for memory to use for its next set of transactions.

Increasing performance: More processor and memory resources can help increase the performance but only if the application is designed to take advantage of it. An uncapped partition can help in this situation (with a specific weight factor). Also, to look at the license fee when using uncapped partitions (for example, processor license-based applications).

Therefore, when beginning to plan your environment, consider ratios such as current system utilization and latency impact (Example 1-5).

Example 1-5 Utilization and latency impact ratio formulas

```
Current throughput
--------------------- = Current System Utilization
Max system throughput

Time to local disk
--------------------- = Latency Impact
Time to remote disk
```

Most readily available reference material generalizes about synchronous distance caps for devices that extend the SAN at distances of 100–300 km. For more information about the different considerations, see Chapter 2, “Infrastructure considerations” on page 39.

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 might be out of sync with the primary site by a few minutes or more. This lag represents data that is unrecoverable if a disaster occurs 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.

Integrated asynchronous solutions: The following integrated asynchronous solutions are the only ones that are available in PowerHA Enterprise Edition 6.1:

- Asynchronous GLVM on AIX 6.1
- SAN Volume Controller Global Mirror
- EMC SRDF/A

Converting replication modes

If an environment begins replicating in synchronous mode and later wants to convert to asynchronous, you must consider the impact to the environment. Depending on the replication type that is being used, the conversion might or might not be performed dynamically.

The following three integrated replication mechanisms support asynchronous mode:

- SAN Volume Controller Global Mirror
- EMC SRDF/A
- Asynchronous GLVM

In an environment that uses SAN Volume Controller Global Mirror, the change can be performed from the cluster panels and the replication mode is immediately updated on the storage subsystem at the time of the synchronization. For more information, see Chapter 5,

"Configuring PowerHA SystemMirror Enterprise Edition with Metro Mirror and Global Mirror" on page 153.

By using the EMC SRDF replication, you can make the change from the cluster panels, and the change is reflected immediately on the storage side. After the change with a cluster, synchronization updates the changes in the cluster definitions across all of the nodes. For more information, see Chapter 7, "Configuring PowerHA SystemMirror Enterprise Edition with SRDF replication" on page 267.

Finally, the change in a GLVM environment is different because it depends on whether you were already running with AIX 6.1 for asynchronous support and whether the scalable volume groups and aio cache logical volumes existed. Therefore, going from synchronous to asynchronous requires manual steps and is disruptive. However, when you go from an asynchronous setup to a synchronous setup, a change to the mirror pool settings is dynamic. For more information, see Chapter 8, "Configuring PowerHA SystemMirror Enterprise Edition with Geographic Logical Volume Manager" on page 339.

1.4.2 Decision making

The tables that are outlined in this section highlight certain differences when you are trying to decipher which solution to implement. Certain key infrastructure design considerations revolve around the ability to either extend the fabric or to use dark fiber interconnects for the replication links. Also, the speed of the existing network infrastructure and whether the available bandwidth can accommodate one of the IP replication alternatives is another important factor to consider. If the immediate infrastructure cannot sustain the load, leasing lines from an external network provider is the next logical step. This setup can quickly become expensive depending on the amount of sustained bandwidth that is required to accommodate the environment.

One of the logical choices in campus environments is to configure mirroring between subsystems in different buildings within the complex (Figure 1-12).

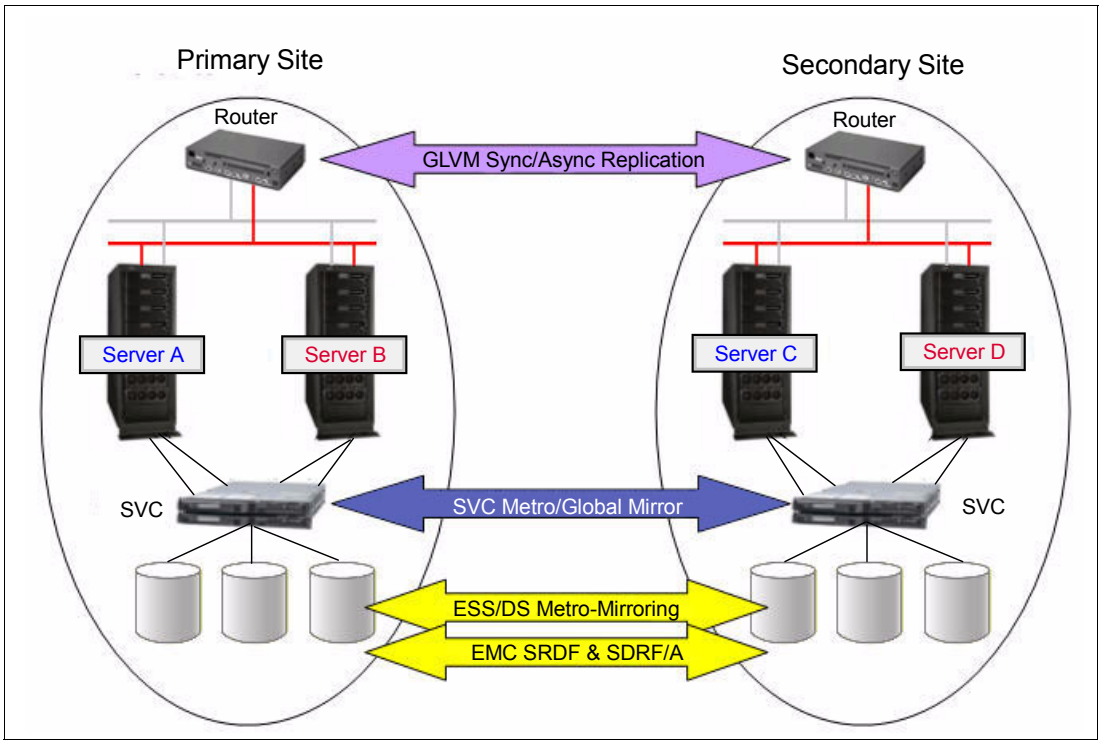


Figure 1-12 PowerHA Enterprise Edition replication options overview

Table 1-4 highlights considerations for choosing whether to use AIX LVM or the GLVM product.

Table 1-4 When to choose Cross-Site LVM versus GLVM

Logical Volume Mirroring (X-LVM)	Geographic Logical Volume Mirroring
The customer has good AIX LVM or PowerHA skills.	There is no inter-connected SAN infrastructure between the two sites.
SAN extends between the two sites.	There is IP WAN with sufficient bandwidth.
This is only available for synchronous distances.	Storage agnostic: There is no requirement for similar subsystems at each site.
	There is synchronous or asynchronous replication.

Using disk-level replication might make more sense in your environment if you are already using storage subsystems that have integrated support. For example, if you are already using the SAN Volume Controller virtualization across your enterprise, then use its Metro Mirror and Global Mirror functions.

Table 1-5 outlines considerations that are associated with each disk replication offering.

Table 1-5 When to choose SVC versus SRDF versus DS Metro/Global Mirroring

SAN Volume Controller Metro/Global Mirror	EMC SRDF/S SRDF/A	DS6000/8000 Metro Mirror
Extended SAN or dark fiber	Extended SAN or dark fiber	Extended SAN or dark fiber
Desire to use dissimilar disk subsystems at each site (fronted by SAN Volume Controller cluster)	Already using EMC DMX3, DMX4, or VMAX Symmetrix enclosures	Already using IBM TotalStorage Enterprise Storage Server 800, DS6K, or DS8K storage enclosures
Synchronous or asynchronous Integration	Synchronous or asynchronous integration	Synchronous-only integration
Provides one disk replication solution across most storage and server platforms	Already using SRDF replication between EMC enclosures	Already using Metro Mirroring Copy Services functions
Replication license for only amount of storage replicated	Details for SRDF licensing should be obtained from EMC	Replication license the full amount of storage on enclosure

Finally, when the environment calls for asynchronous replication, you have the following support choices:

- GLVM
- SAN Volume Controller Global Mirror
- SRDF/A

Table 1-6 shows the requirements for each asynchronous replication solution.

Table 1-6 When to choose async GLVM or async disk replication

Geographic Logical Volume Mirroring	SAN Volume Controller Global Mirror or EMC SRDF
IP-based-only connectivity between the sites	Extended SAN or dark fibre.
When distance makes synchronous replication unfeasible Running on AIX 6.1 and later for asynchronous replication	Distance makes synchronous replication unfeasible.

Geographic Logical Volume Mirroring	SAN Volume Controller Global Mirror or EMC SRDF
Desire to support dissimilar storage subsystems at each site	Already using SAN Volume Controller clusters at each site or EMC storage at each site.
Replication specific to AIX	Provides one solution across most server platforms.

Sample topologies and different cluster configurations are outlined in the test scenarios that are described in this book.

1.5 PowerHA enterprise logistics

This section provides information about the licensing considerations and options for the PowerHA solution.

1.5.1 IBM PowerHA SystemMirror Enterprise Edition licensing

The software licensing model in PowerHA Version 6.1 has been simplified and broken down into Standard and Enterprise Edition packages. Also, the pricing has also been changed to reflect the server class that the software might be running on.

Both the Standard Edition and the Enterprise Edition are licensed by processor and require a minimum of one processor license per server. No pricing model is available to accommodate micropartitioned LPARs in PowerHA. Therefore, an LPAR that uses only one-fourth of a processor still must license a minimum of one processor.

Table 1-7 lists the breakdown of the new licensing model.

Table 1-7 Licensing per core for PowerHA Editions

Standard Edition - PID 5765-H23 Enterprise Edition - PID 5765 - H24	Power Systems
Small tier	Blade and entry Power
Medium tier	Mid-range Power
Large tier	Enterprise servers

For more specific pricing details based on the server models that you deployed, contact your IBM sales representative.

Table 1-8 on page 27 outlines the orderable features codes for the PowerHA Standard and Enterprise Editions.

Table 1-8 PowerHA Standard and Enterprise Edition product identification numbers

Program PID number	Maintenance 1-year PID	Maintenance 3-year PID
IBM PowerHA SystemMirror Standard Edition - 5765-H23	SW Maintenance Regist/Renewal 1 Year - 5660-H23 SW Maintenance After License 1 Year - 5661-H23	SW Maintenance Registration 3 Year - 5662-H23 SW Maintenance Renewal 3 Year - 5663-H23 SW Maintenance After License 3 Year - 5664-H23
IBM PowerHA SystemMirror Enterprise Edition - 5765-H24	SW Maintenance Regist/Renewal 1 Year - 5660-H24 SW Maintenance After License 1 Year - 5661-H24	SW Maintenance Registration 3 Year - 5662-H24 SW Maintenance Renewal 3 Year - 5663-H24 SW Maintenance After License 3 Year - 5664-H24

The PowerHA SystemMirror for AIX Enterprise Edition includes all of the capabilities of the Standard Edition and more. The Enterprise Edition package enables you to extend your data center solution across multiple sites. It provides integrated support with the IBM TotalStorage Enterprise Storage Server (ESS), IBM System Storage (DS6000 and DS8000), and SAN Volume Controller replication types. In the 6.1 release the Enterprise Edition expands HA and disaster recovery support to integrate with the EMC SRDF replication. It also provides enhancements, including a wizard to configure and deploy host-based mirroring solutions (GLVM configuration wizard).

The base cluster software in the Standard Edition allows the use of sites within a local cluster. This is applicable to environments located within Metro distances that wanted to implement a campus-style disaster recovery solution. In an environment that uses AIX Logical Volume Mirroring, licensing only the PowerHA SystemMirror Standard Edition suffices. Even though the cluster would be configured between a two-site data center, it would behave more like a local stretch cluster. The use of site definitions within the cluster provides only extra cluster measures to check for the consistency of the configuration.

For an example of a four-node Cross Site LVM cluster that uses the Mirror Pool functions in AIX 6.1, see Chapter 4, “Configuring PowerHA Standard Edition with cross-site logical volume mirroring” on page 113.

1.5.2 Dynamic LPAR integration and licensing

The PowerHA 6.1 release introduced support for IBM Micro-Partitioning® within the integrated Dynamic LPAR functions. Before this release, you could use the integrated menus to only specify entire processor and memory values. In the current menus, you can specify fractions of a processor and alter the count of the virtual processors. The overall Dynamic LPAR functions in PowerHA were documented in prior IBM Redbooks publications. However, it is critical to understand the separate ways that an environment can be configured and its impact on licensing.

Figure 1-13 shows the changes to the Dynamic LPAR panel in the PowerHA 6.1 release.

```
              Add Dynamic LPAR and CUoD Resources for Applications

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

[TOP]                                                 [Entry Fields]
* Application Server Name                            oracle_db

* Minimum number of processing units                 [  0.00]
* Desired number of processing units                 [  0.00]

* Minimum number of CPUs                             [0]                      #
* Desired number of CPUs                             [0]                      #

* Minimum amount of memory (in megabytes)            [0]                      #
* Desired amount of memory (in megabytes)            [0]                      #

* Use CUoD if resources are insufficient?            [no]                    +
* I agree to use CUoD resources                      [no]                    +
  (Using CUoD may result in extra costs)

  You must ensure that
  * CoD enablement keys are activated
  * CoD resources are not used for any other purpose
```

Figure 1-13 DLPAR integration enhancements in PowerHA SystemMirror 6.1

To grasp the licensing considerations, one must first have a fundamental understanding of how Dynamic LPAR works in a PowerHA environment. The criteria for the minimum and desired values that you typically associate with the server that hosts your application are not bound to the individual machines. In PowerHA, these criteria are bound to the application server definition within the cluster. The cluster tries to meet these criteria based on where the application is being hosted. The Dynamic LPAR operations are invoked only during the acquisition or release of your application server. This means that an LPAR does not self-tune during normal operations. Its resource counts are only altered during the start and stop phases of the cluster EVENT processing. The cluster only deallocates any resources if it previously added them through the cluster functions. If you are trying to build a solution with minimal PowerHA licenses, you can license the standby partitions for only a single processor and lower your initial HA licensing costs.

To provide a better understanding of this concept, see the multisite scenario with two local nodes and one remote node that is shown in Figure 1-14.

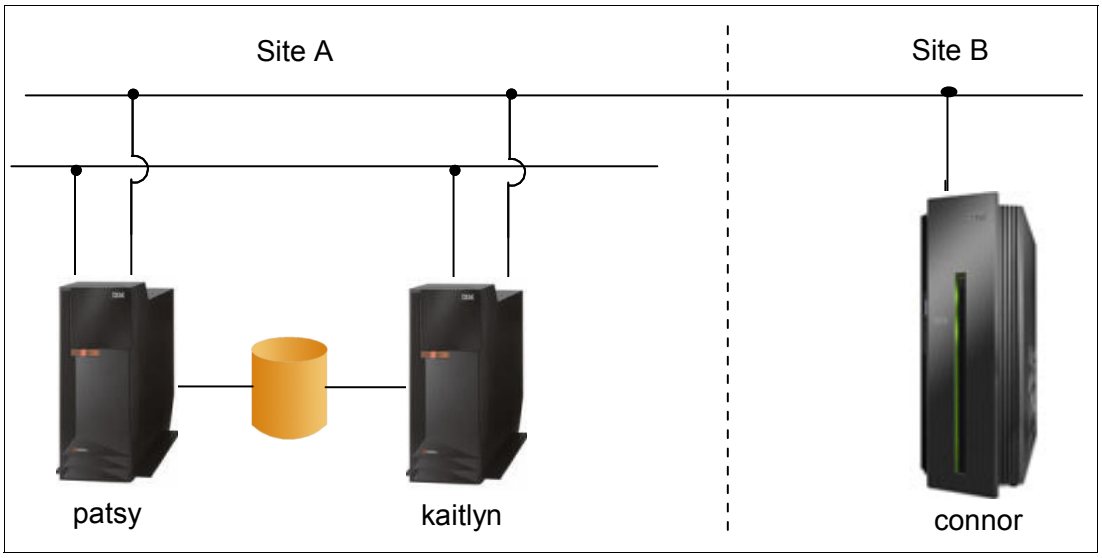


Figure 1-14 Dynamic LPAR example - multisite cluster diagram

These LPARs are configured with the partition profile settings shown in Table 1-9.

Table 1-9 Partition profiles for the dynamic LPAR example

LPAR name	Minimum	Desired	Maximum
Site 1 - patsy	1 processor	1 processor	4 processors
Site 1 - kaitlyn	1 processor	1 processor	4 processors
Site 2 - connor	1 processor	1 processor	4 processors

In Figure 1-14, all nodes are configured to come up with only one processor. During the start of the cluster, the Dynamic LPAR settings that are bound to the application server are evaluated and enforced. The desired resources for the application server are appended only if the shared pool has the available resources. Table 1-10 lists the application server dynamic LPAR settings.

Table 1-10 Dynamic LPAR example - application server settings

Application server	Minimum value	Desired value
app1	0	4

On cluster startup, the cluster evaluates the application server requirements and appends three more processors to meet the desired amount, or four processors. Therefore wherever the resource group and corresponding application server are being hosted, the cluster performs a *+3 processor* operation on acquisition or a *-3 processor* operation on the release. The key from a licensing standpoint is that you must license the number of active processors at any time. In this example, the number of processors that are licensed is six.

However, an exception to the rule is a scenario in which the production partition profile is set up with a minimum value of one and a desired value of four, as in Table 1-11 on page 30. This scenario can have different results.

Table 1-11 Dynamic LPAR example with modified partition profiles

LPAR name	Minimum value	Desired value	Maximum value
Site 1 - patsy	1	4	4
Site 1 - kaitlyn	1	1	4
Site 2 - connor	1	1	4

Assuming the same application server processor values as in Table 1-10, after a resource group move, you can end up with four active processors on the source and four on the target. This time is the only time in this type of configuration where you should run your system with a total of nine active processors when licensed for only six. For example, you are licensed for only four on production, one on local backup node, and one on the disaster recovery node.

A different scenario is an environment in which the LPARs are running uncapped. In such an environment, the partitions can exceed their desired partition processor counts and consume more of the available resources within their shared processor pool. To remain compliant in this situation, license the maximum number of processors that can ever be used by each cluster partition. Alternatively, create a shared processor pool that caps the processor resources that can be used by your cluster. Therefore, if a partition is set with a desired value of 4, but running uncapped enabled it to consume eight processors, the number of PowerHA licenses is eight. In this model, you pay more for cluster licensing, but the LPARs would be configured so that they can self-tune and truly share the processors with the other LPARs within the same CEC.

If insufficient resources are available in the free pool that can be allocated through the Dynamic LPAR, the Capacity on Demand (CoD) function can provide extra resources to the node. The proper CoD licenses must be in place for this function to work. Licensing revolves around the CoD terms and conditions. CoD is not available on all Power Systems.

Although the goal of the IBM clustering software team was never to tightly enforce processor license counts, you must follow these guidelines to remain fully compliant if an audit occurs.

More Dynamic LPAR considerations

The potential downside of Dynamic LPAR integration within the cluster is the introduction of a new point of failure. Although interruption to an HMC administrative console typically has no impact to an environment, it might now prevent an HA failover if it is not available. Also, if the environment is not carefully planned and not enough processors are in the free shared pool to meet the application server's minimum requirement, resources might be prevented from coming online.

Tip: The integrated PowerHA Dynamic LPAR panels allow multiple Hardware Management Console (HMC) definitions. Dual HMCs are favorable when you consider a fully redundant solution that uses dynamic LPAR.

For firewall configurations, the LPAR Resource Monitoring and Control (RMC) daemon listens on TCP port 657 for client/server communication and UDP port 657 for RMC daemon-to-RMC daemon (peer) communication.

RMC port registration: The RMC port is registered with the Internet Assigned Numbers Authority (IANA) and is listed at:

`http://www.iana.org/assignments/port-numbers`

For firewall configurations, the HMC-to-HMC discovery uses UDP port 9900, and the HMC-to-HMC commands use TCP port 9920. For firewall configuration, the HMC browser interface uses TCP port 443, TCP port 8443, and TCP port 9960 (browser applet communication).

For more information about HACMP configuration for dynamic LPAR and CoD, see the *HACMP for AIX: Administration Guide*, SC23-4862, and the *PowerHA for AIX Cookbook*, SG24-7739. For more information about HMC configurations, see the *Power Systems: Managing the Hardware Management Console* manual at:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/iphai_p5/iphaibook.pdf

1.6 What is new on PowerHA Enterprise Edition

The IBM PowerHA development team has expanded its scope beyond local availability and disaster recovery offerings with each new release. The following disaster-recovery-specific enhancements were introduced in each of the last five releases:

- PowerHA SystemMirror for AIX 6.1
  - GLVM Configuration Wizard
  - EMC SRDF/S and SRDF/A replication integration
  - DS8000 Storage (Metro Mirror) in the VIOS-based virtualized environments
- PowerHA Release 5.5
  - Asynchronous GLVM (with AIX 6.1)
  - Global Mirror Integration with SAN Volume Controller
  - Global Mirror for VSCSI volumes by using SAN Volume Controller
  - Manual and Auto Site failover Policy
  - SPPRC support for multiple storage units per site
  - SPPRC Password Encryption
- HACMP 5.4.1
  - Consistency groups
  - GLVM Monitoring enhancements
- HACMP 5.4.0
  - Metro Mirror support for intermixed environments (DS6000, DS8000, Enterprise Storage Server 800)
  - Multiple GLVM XD_Data networks
  - IPAT Across Sites

    Site-specific service labels solve the problem of having different subnets at different sites.
- HACMP 5.3
  - Location dependencies (for example, online on same site)
  - Synchronous GLVM

Notes about the new PowerHA features are available in the PowerHA Enterprise Edition release notes, which are installed in the `/usr/es/sbin/cluster/release_notes_xd` directory. For more information, see the PowerHA website at:

http://www.ibm.com/systems/power/software/availability/

1.6.1 PowerHA Enterprise Edition for Metro Mirror software

PowerHA Enterprise Edition for Metro Mirror increases data availability for IBM TotalStorage volumes that use synchronous PPRC to copy data to a remote site for disaster recovery purposes. PPRC allows mirroring to be suspended and restarted without affecting data integrity. PowerHA Enterprise Edition helps manage the PPRC instances.

This solution is available for the following IBM storage systems:

- IBM Enterprise Storage Server (ESS) Model 800
- IBM TotalStorage Server (DS) Models 6000 and 8000
- IBM TotalStorage SAN Volume Controller

PowerHA Enterprise Edition for Metro Mirror takes advantage of the following components to reduce downtime and recovery time during disaster recovery:

- PowerHA cluster management
- PPRC failover and failback functions
- Optional components:
  - ESSCLI PPRC support
  - DSCLI PPRC support
  - SAN Volume Controller PPRC support

1.6.2 ESSCLI-based PPRC support

PowerHA provides support for the task-based management of ESS PPRC by managing specified pair and path tasks as defined by the user on their ESS storage systems Copy Services Server. PowerHA provides monitoring, failover, and failback support for PPRC by issuing commands directly to the Copy Services Server by using the ESS command-line interface (CLI) interface.

1.6.3 DSCLI-based PPRC support

DSCLI management provides a simplified PPRC interface for both the Enterprise Storage Server 800 and DS storage hardware. The DSCLI interface provides more automated creation and management of PPRC paths and instances. It also has more flexibility by supporting both ESS and DS storage subsystems with the same interface.

1.6.4 DSCLI-based PPRC support for multiple storage units per site

DSCLI-based PowerHA Enterprise Edition PPRC supports more than one storage system per site. You can now configure and use more than one DSS on a single SPPRC site. Each PPRC replicated resource still has only one primary and one auxiliary storage per site, but you can use any configured DSS if there is a single one per site in a PPRC replicated resource group.

1.6.5 SVC-PPRC support

SVC-PPRC management provides both storage virtualization and an extra layer of disaster recovery by using the SAN Volume Controller cluster and hardware configuration. It enhances PPRC's ability to provide a fully automated, highly available disaster recovery management solution by taking advantage of the SAN Volume Controller's ability to provide virtual disks that are derived from varied disk subsystems. The PowerHA interface is designed so that, when the

basic SAN Volume Controller environment is configured, PPRC relationships are created automatically. No additional access to the SAN Volume Controller interface is needed.

1.6.6 PowerHA Enterprise Edition

PowerHA Enterprise Edition has several new enhancements.

Support for SVC Global Mirror copy type

Global Mirror support was added to PowerHA Enterprise Edition for SAN Volume Controller. A new field was introduced in the SVCPPRC consistency group SMIT interface that allows the configuration of Global Mirror.

If a SAN Volume Controller relationship exists but has a copy type of METRO MIRROR, you must remove the relationship and rerun sync. Then verify for the relationship to be re-created with the copy type of GLOBAL MIRROR. Figure 1-15 shows a panel with this update.

```
                        Change / Show SVC PPRC Resource

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

                                                      [Entry Fields]
  SVC PPRC Consistency Group Name                    svc_global
  New SVC PPRC Consistency Group Name                []
* Master SVC Cluster Name                            [B12_4F2]               +
* Auxiliary SVC Cluster Name                         [B8_8G4]                +
* List of Relationships                              [svc_disk6 svc_disk7 s> +
* Copy Type                                          GLOBAL                  +
* HACMP Recovery Action                              AUTO                    +
```

Figure 1-15 SAN Volume Controller Global Mirror copy type

MANUAL recovery option

A new SMIT interface is introduced to provide MANUAL and AUTO options for users to select the action that PowerHA Enterprise Edition should take in certain recovery situations. This option was introduced in the PowerHA 5.5 release for the SVC integration but is also available for the DS Metro Mirror integration and the EMC SRDF replication integration. This action is predetermined and can be specified to prevent the cluster from automatically failing over. The following two settings are available:

- If AUTO is selected and a total site failure occurs of the primary site, the secondary site automatically takes over the replicated resources.
- If MANUAL is selected, users are expected to perform certain actions manually before the failover can occur. Figure 1-15 on page 33 shows this option.

  Use of this option does not prevent a cluster failover in all situations. The MANUAL setting prevents a failover based only on the status of the replicated volumes at the time of the node failure. If all of the replication consistency groups display a consistent state, even with selecting the MANUAL" option, a failover still takes place.

The states vary based on the replication type that is being used between the sites. Example 1-6 lists the states that the MANUAL policy would be applicable.

Example 1-6 Applicable states for the MANUAL policy

```
DS PPRC states that would not failover:
   Target-FullDuplex-Source-Unknown
   Source-Suspended-Source-Unknown
   Source-Unknown-Target-FullDuplex
   Source-Unknown-Source-Suspended

SVC PPRC states that would not failover:
   idling_disconnected
   consistent_disconnected

EMC SRDF states that would not failover:
   SRDF/S - Partitioned
   SRDF/A - TransIdle
```

Most suitable setting for your environment: The default setting for the recovery action is AUTO and results in an automated failover if an outage occurs. Use careful consideration when you decide on the most suitable setting for your environment.

Support for IBM Virtual I/O clients

The use of a traditional VSCSI mapping from a VIO server to a client results in an abstracted view of the LUN on the host side. The current release introduces the ability to replicate VSCSI LUNs by using the SAN Volume Controller and the DS storage units.

Before this enhancement: Before this enhancement, the verification was unable to gather the information that it needed from the VSCSI view of the LUNs. No PowerHA disk replication support was available for VSCSI LUNs between sites.

An alternative approach to avoid this limitation is to use N-Port ID Virtualization (NPIV) capable fiber adapters on the VIO servers. NPIV requires the switches to be NPIV enabled. When you use virtual fiber adapters, the LUNs on the host are visible in the same fashion as when mapped to dedicated adapters.

Example 1-7 is sample output from an environment where the first two disks are VSCSI volumes and the next two come from an NPIV virtualized adapter.

Example 1-7 Output of an environment with VSCSI volumes

```
#lsdev -Cc disk
hdisk0 Available         Virtual SCSI Disk Drive
hdisk1 Available         Virtual SCSI Disk Drive
hdisk2 Available 33-T1-01 MPIO FC 2145
hdisk3 Available 33-T1-01 MPIO FC 2145


#lsdev -Cc adapter
ent0     Available       Virtual I/O Ethernet Adapter (l-lan)
ent1     Available       Virtual I/O Ethernet Adapter (l-lan)
fcs0   Available 23-T1 Virtual Fibre Channel Client Adapter
fcs1   Available 33-T1 Virtual Fibre Channel Client Adapter
vsa0    Available       LPAR Virtual Serial Adapter
```

```
vscsi0 Available       Virtual SCSI Client Adapter
vscsi1 Available       Virtual SCSI Client Adapter

lscfg -vl hdisk0
  hdisk0           U8204.E8A.10FE411-V5-C3-T1-L8100000000000000  Virtual SCSI Disk
Drive

#lscfg -vl hdisk2
  hdisk2           U8204.E8A.10FE411-V3-C33-T1-W5005076801304F45-L0 MPIO FC2145

        Manufacturer................IBM
        Machine Type and Model......2145
        ROS Level and ID............0000
        Device Specific.(Z0)........0000043268101002
        Device Specific.(Z1)........0200640
        Serial Number...............60050768019026C4000000000000000
```

If the base NPIV requirements are met in the environment, a configuration that uses disk replication works the same as in one that uses dedicated host bus adapters (HBAs). However, you must still confirm the NPIV qualification of the separate storage systems and their corresponding replication mechanisms.

Considerations

Consider the following items regarding the current PowerHA Enterprise Edition releases:

- A single PowerHA cluster supports only two sites. A single node can be part of only one PowerHA cluster site.
- A single PowerHA Enterprise Edition cluster supports up to eight nodes.
- Concurrent disk access within a cluster by using GLVM is supported only within sites, not between sites.
- PowerHA Version 6 base (cluster.es.server.rte 6.n.0.0) must be at the same release level (n) as PowerHA/XD.
- The GLVM two-site configuration wizard is not IPv6-enabled.
- The GLVM two-site configuration wizard does not support asynchronous replication.
- There is no support for enhanced concurrent mode volume groups when you use asynchronous GLVM on AIX 6.1.
- PPRC eRCMF support is no longer included with the Enterprise Edition.
- SRDF support is officially direct attach only.
- PowerHA for SRDF known considerations:
  - C-SPOC considerations

    C-SPOC cannot perform the following LVM operations on nodes at the remote site (that contain the target volumes):

    - Creating a volume group.
    - Operations that require nodes at the target site to read from the target volumes cause an error message in C-SPOC. This situation includes functions such as changing file system size, changing mount point, and adding LVM mirrors. However, nodes on the same site as the source volumes can successfully perform these tasks, and the changes then are propagated to the other site by using lazy update.

For C-SPOC operations to work on all other LVM operations, perform all C-SPOC operations with the SRDF pairs in synchronized or consistent states or the cluster ACTIVE on all nodes.

– Inter-site failover considerations

The following functions that are provided by PowerHA for local failover are not yet supported across sites:

- rg_move because of selective failover across sites
- User-directed rg_move across site

– SRDF functions considerations

- Multihop configurations are not supported.
- Mirroring to BCV devices is not supported.
- Concurrent RDF configurations are not supported.

For more information about these considerations, see the 6.1 Release Notes in the `/usr/es/sbin/cluster/release_notes_xd` file.

1.6.7 PowerHA/XD SPPRC DSCLI security enhancements

Security is enhanced by creating an encrypted password file for use with the **`dscli`** commands. This file is created based on the entries in the PowerHA configuration. The first time that sync and verify are run, the encrypted `passwd` file is created in the `/var/hacmp/logs/pprc/run/security` directory (or the directory that you specified for log file redirection). This directory has permissions of 600 and the files are added as a file collection that gets pushed to all the other nodes.

The file gets created only one time on the first node. Keep in mind the following maintenance considerations:

- If the HMC/SMC password is changed, you must change the password in the PowerHA configuration and remove the encrypted `passwd` file. If you do not change the password and more the `passwd` file, the next time that verify runs the wrong `passwd` file is used.
- If you have more than one `passwd` file and you remove one of them, you might end up removing that file from all the other nodes, causing the commands to fail. Do not remove any files from the `../pprc/run/security` directory unless you change your passwords, remove all files, and re-create them all to ensure that you have good `passwd` files.

1.6.8 PowerHA and Power Systems hardware and software support considerations

IBM POWER7® (750, 770, and 780 models) does not allow PowerHA to use the internal serial ports. Build-in Ethernet ports are supported.

- PowerHA support for POWER7 750:
  - V5.5 and AIX 6.1 TL4 SP2 RSCT 2.5.4.2
  - V6.1 and AIX 6.1 TL4 SP2 RSCT 2.5.4.2
- PowerHA support for POWER7 770 and 780:
  - V5.5 and AIX 5.3 TL11 SP1 RSCT 2.4.13.0
  - V5.5 and AIX 6.1 TL4 SP3 RSCT 2.5.5.0
  - V6.1 and AIX 5.3 TL11 SP1 RSCT 2.4.13.0
  - V6.1 and AIX 6.1 TL4 SP3 RSCT 2.5.5.0

1.6.9 Viewing and installing the documentation files

You can install the following documentation images and file sets in the System Management Interface Tool (SMIT):

- Image cluster.doc.en_US.es
  - cluster.doc.en_US.es.html HACMP web-based HTML Documentation - US English
  - cluster.doc.en_US.es.pdf HACMP PDF Documentation - US English
- Image cluster.doc.en_US.pprc
  - cluster.doc.en_US.pprc.html PPRC web-based HTML Documentation - US English
  - cluster.doc.en_US.pprc.pdf PPRC PDF Documentation - US English
- Image cluster.doc.en_US.glvm
  - cluster.doc.en_US.glvm.html PowerHA GLVM HTML Documentation - US English
  - cluster.doc.en_US.glvm.pdf PowerHA GLVM PDF Documentation - US English

If you install the documentation on a web server that is accessible through your network, you can view the documentation with your browser. To access the HTML file, in your browser, enter:

```
file:///usr/share/man/info/en_US/cluster/HAES/<directory name>/<filename>.htm
```

For example, enter the following file name and directory:

```
file:///usr/share/man/info/en_US/cluster/HAES/ha_concepts/ha_concepts.htm
```

2

Infrastructure considerations

The key to an effective high-availability solution goes well beyond the installation of the Systems software. Thorough planning and careful consideration of all potential single points of failure helps minimize the risk of an unforeseen outage of critical business applications. This chapter reviews many of the requirements and infrastructure considerations when you plan a disaster recovery solution with the PowerHA SystemMirror for AIX Enterprise Edition on Power Systems.

Under the infrastructure considerations for high availability, include the user network or the client network, the enterprise network, the local and global network (cloud technology), and the security needs for end-to-end operations of the business process.

This chapter includes the following sections:

- Network considerations
- Cluster topology considerations
- Storage considerations
- PowerVM virtualization considerations
- Server considerations

2.1 Network considerations

The network infrastructure between the servers and remote sites plays a critical role in how much data and how fast the cluster can replicate. It also dictates what communication paths are used for heartbeating.

2.1.1 Bandwidth

The performance of a geographically dispersed synchronized mirror storage system depends on the throughput data rate (bandwidth) and the latency of the communication links between the primary and the secondary sites. Use dedicated networks for synchronizing data.

For long-haul transmissions to extend storage networks, consider the following parameters regarding the required quality of service and performance levels. From a storage network extension perspective, consider the following quality of service parameters:

Bandwidth Can be viewed as the maximum transmission speed in one direction for the medium.

Latency Can be viewed as the delay from the source node to the access network before transmission can begin using the bandwidth.

End-to-End delay Can be viewed as the propagated latency from the source node to the target node with accumulated latency (end-to-end), which can vary from time to time and can include effects of congestion control.

Error rate Can be viewed as the frequency of information loss during transmission, possibly caused by noise from loss in signal power because of a photoelectric conversion process at the receiver.

Availability Can be viewed as the uptime of the local access network for storage network extension.

Reliability Can be viewed as the uptime of the entire storage network extension end-to-end.

The effective throughput depends on the bandwidth and all write I/O reductions. Examples include latency, end-to-end delay, error rates that cause local I/O queuing, and availability and reliability issues that can cause temporary synchronization suspension, depending on the deployed synchronization method and write verification policy.

2.1.2 Bandwidth sizing

To measure bandwidth requirements, start by identifying the logical volumes that contain the critical data to be synchronized, in which volume groups they are, and which disk they use. Measure the write I/O throughput and performance for these logical volumes. Use the data that is captured to help size the bandwidth requirements. Check also the write frequency and write sizes.

To identify the *initial data replication* from the primary to the secondary site, consider the amount of real data that is in the logical volumes and file systems. The actual initial synchronization can differ depending on whether it is host or storage-based synchronization, which storage platform, and the selected synchronization technology.

You can use several tools to gather the necessary disk I/O information to calculate site communication bandwidth requirements. For example, you can use the **gmdsizing** tool, the AIX **filemon** command, **iostat**, **sar**, or IBM Tivoli Monitoring.

The **gmdsizing** tool monitors disk utilization over a time. When the command completes or is interrupted, it produces a report on disk usage over the time the command was running.

The **gmdsizing** tool file set is available on the HACMP/XD for AIX installation media or can be downloaded from:

http://www.ibm.com/systems/resources/systems_p_advantages_ha_downloads_gmdsizing.tar

The **gmdsizing** command requires that the interval (-i) and time (-t) flags are supplied. The interval specifies how often disk activity is checked and the time specifies for how long the program runs.

The time flag defaults to seconds, but can be changed by appending letters after the number:

d Number of days
h Number of hours
m Number of minutes
s Number of seconds

For example, to check over five days, you can use 5d, 120h, or 7200m as an argument to the time flag.

In addition, at least one of either disk devices (-p) or volume groups (-v) must be specified. If the volume group flag is specified, the command converts the volume group argument into physical volume names. All data reported by **gmdsizing** is given in disk blocks.

If two-copy mirroring is enabled for the logical volumes, twice as many writes occur at the physical disk level because one logical write from the application generates two physical writes to the disk devices. Rather than selecting an entire volume group to be monitored, select just those disks that contain one copy of the mirrors. If the volume group is laid out such that selecting those disks is not feasible, potentially twice as much write activity is generated from the application perspective. Therefore, keep in mind the potential for twice the amount of write activity when you analyze the data.

The verbose (**-V**) flag results in a summary that is written at the end. The file flag (**-f**) makes the command write the output to the specified file. Always use both the **-V** and **-f** flags. The following flags are also available:

-D To change delimiter
-U To change units
-w For writes only
-A For aggregate
-T For time scale

Measure disk utilization at a representative time. If the system is running a nine-to-five operation, measuring disk utilization in the middle of the night does not give much information. Similarly, measuring over a short time is likely to give unrepresentative data. Consider any overnight batch processing activities because batch processing tends to be write intensive.

You must understand how the workload varies over time, such as during busy periods during a day, week, or month, or during year-end or quarter-end processing.

It is preferable to run the **gmdsizing** tool over a longer period than a shorter one because over a longer period, the tool is more likely to observe peaks and troughs, which monitoring for a shorter period might miss. However, when you specify the observation interval, keep the interval larger rather than smaller. One line of data is written per disk per interval. Therefore, a large amount of data is collected if you have a small interval, many disks, or both.

The following command monitors the disks in volume group datavg01 over 24 hours with a one-minute interval and saves the output to a file:

```
gmdsizing -I 1m -t 24h -v datavg01 -V -f /tmp/gmdout.48hx1m.$(date +%Y%m%d)
```

The following command monitors the disks in volume group datavg01 over 72 hours with a 10-minute interval and saves the output to a file:

```
gmdsizing -I 10m -t 72h -v datavg01 -V -f /tmp/gmdout.48hx1m.$(date +%Y%m%d)
```

The following command monitors the disks in volume group datavg01 over 60 days with a one-hour interval and saves the output to a file:

```
gmdsizing -I 60d -t 1h -v datavg01 -V -f /tmp/gmdout.60dx1h.$(date +%Y%m%d)
```

For sizing the networks, we are only interested in write traffic, but the read columns are useful to help determine the read:write ratio for the workload. Knowing how long the disk activity was measured for (the parameter that is passed with the -t flag) allows the determination of an average write rate. Because all data reported by **gmdsizing** is given in disk blocks, it is necessary to convert it to bytes, which can be done by multiplying the total of all the *total write* values by the block size of the device. This value can then be divided by the total time over which we were measuring. See the following example:

```
n-blocks-written x bytes-per-disk-block = total-volume-of-data-written
```

Then, divide the total volume of data that is written by the time for the measurement period:

```
total-volume-of-data-written / measurement-time = bytes/time-scale
```

See the following example:

```
130000 blocks x 512 bytes/block = 66560000 bytes
```

If the measurement period was 30 minutes, you make the following calculation:

```
66560000 / 1800 seconds = 36977 bytes/second
```

This value is an average, assuming that data is written at a constant rate, which is unlikely. To understand how accurate this average is, compare this value with the theoretical worst-case scenario. This value is the maximum volume of data that is written in one measurement interval (defined by the -i flag). Dividing this figure by the measurement interval gives us a worst-case rate. In this example:

```
4388 block x 512 bytes/block = 2246656 bytes / 60 seconds = 37444 bytes/second
```

Compare these two values (36977 and 37444). If they are relatively close (as in this example), the disk activity and the network bandwidth requirements are fairly uniform. These values probably can be used to get a good estimate for the network requirements. If these values differ widely, which is more likely (having a worst-case value 7 - 10 times that of the average), a more detailed analysis of the **gmdsizing** data is needed to determine the bandwidth requirements. More complex analysis of the data might be performed by using statistical techniques such as calculating the max, average, median, first, and third quartiles, and standard deviation based on a more detailed sample.

The network bandwidth determined by these methods does not take into account networking latency. Allowing a 20 - 30% overhead for the networking protocols can be sufficient for most networks. See the following example:

```
bytes/second / 0.75 = bandwidth requirement
```

In addition to the **gmdsizing** tool, the AIX **filemon** tool can be used to measure more granular metrics such as average write size and disk write times for each logical volume. Use

caution with the **filemon** tool because it can collect a huge amount of data during a relatively short period.

After you determine the bandwidth requirements for the current storage network utilization, consider the latency, end-to-end delay, error rate, and reliability (assuming that the local availability is near 100%). A mapping can be made to a network technology and to a network service provider.

Consider the protocol stack that is involved in long-haul transmissions to extend storage networks, and whether it is from host nodes or storage nodes. The following list is a simplified transmission stack from a source host node to a target host node:

1. SCSI layer
2. FCP/IBA
3. IP transport layer
4. IP network layer
5. Network interface
6. Local network
7. Wide area network interface
8. Wide area network
9. Wide area network interface
10. Local network
11. Network interface
12. IP network layer
13. IP transport layer
14. FCP/IBA
15. SCSI layer

2.1.3 Network technologies

The bandwidth for optical fiber is commonly defined as the range of frequencies within which a fiber optic waveguide or terminal device can transmit data, and is limited by the multimode dispersion phenomenon:

> Different reflection angles within the fiber core create different propagation paths for the light rays. Rays that travel nearest to the axis of the core propagate by what is called the *zeroth order mode*; other light rays propagate by higher-order modes. It is the simultaneous presence of many modes of propagation within a single fiber that creates multimode dispersion. Multimode dispersion causes a signal of uniform transmitted intensity to arrive at the far end of the fiber in a complicated spatial interference pattern. This pattern, in turn, can translate into pulse spreading or smearing and intersymbol interference at the optoelectronic receiver output. Pulse spreading worsens in longer fibers.[1]

Measure the actual link loss values once the link is established to identify any potential performance issues. The maximum distance of a particular fiber optic link depends on the following factors:

- Actual optical fiber attenuation per km
- Optical fiber design and age
- Quality of connectors and actual loss per pair
- Quality of splices and actual loss per splice
- Quality of splices and connectors in the link

[1] "Transmission media and the problem of signal degradation," Encyclopedia Britannica Online, 2010, http://www.britannica.com/EBchecked/topic/585825/telecommunications-media

- Transmitter type and variety
- Receiver sensitivity
- Selected margin to account for aging of fiber and link components, addition of devices along the link path, incidental twisting or bending of fiber, and extra splices to repair cable breaks

The actual long-haul communication implementation (Table 2-1) can be made with a service provider. Examples include using time-division multiplexing (TDM) over IP or similar, TDM over packet switched network (PSN), and time slot assignment (TSA). They might also include dense wavelength division multiplexing (DWDM), coarse wavelength division multiplexing (CWDM), or a private dark fiber network.

Table 2-1 Network technology bandwidth examples

Network technology	Protocol	Bandwidth Mbps	Bandwidth Gbps
10 Mbps Ethernet	TCP/IP	10	0.01
100 Mbps Ethernet	TCP/IP	100	0.1
1-Gbps Ethernet (GbE)	TCP/IP	1,000	1
10 Gbps Ethernet	TCP/IP	10,000	1
FC	FCP	1,000	1
FC	FCP	2,000	2
FC	FCP	4,000	4
FC	FCP	8,000	8
FC	FCP	10,000	10
FC	FCP	20,000	20
T1/DS1	TCP/IP	1,544	1.544
E1	TCP/IP	2,048	2.048
T2	TCP/IP	6,312	6.312
E2	TCP/IP	8,448	8.448
T3/DS3	TCP/IP	44,736	44.736
E3	TCP/IP	34,368	34.368
OC-1 (SONET)	ATM	51.84	0.05
OC-3 (SONET)/STM-1 (SDH)	ATM	155.52	0.16
OC-12 (SONET)/STM-4(SDH)	ATM	622	0.62
OC-24 (SONET)	ATM	1,244	1.24
OC-48 (SONET)/STM-16(SDH)	ATM	2,488	2.49
OC-192 (SONET)/STM-64(SDH)	ATM	10,000	10
OC-256 (SONET)	ATM	13,271	13.27
OC-768 (SONET)/STM-256(SDH)	ATM	40,000	40
OC-1536 (SONET)/STM-512(SDH)	ATM	80,000	80.00

Network technology	Protocol	Bandwidth Mbps	Bandwidth Gbps
OC-3072 (SONET)/STM-1024(SDH)	ATM	160,000	160.00
InfiniBand Single Data Rate (SDR)	SVP/IPoIB	2,000	2
InfiniBand Double Data Rate (DDR)	SVP/IPoIB	4,000	4
InfiniBand Quad Data Rate (QDR)	SVP/IPoIB	8,000	8
InfiniBand 4xSDR	SVP/IPoIB	8,000	8
InfiniBand 4xDDR	SVP/IPoIB	16,000	16
InfiniBand 4xQDR	SVP/IPoIB	32,000	32
InfiniBand 12xSDR	SVP/IPoIB	24,000	24
InfiniBand 12xDDR	SVP/IPoIB	48,000	48
InfiniBand 12xQDR	SVP/IPoIB	96,000	96

SONET and SDH: Synchronous Optical NEtwork (SONET) is the American National Standards Institute (ANSI) standard for synchronous data transmission on optical media.

Synchronous Digital Hierarchy (SDH) is the international standard for synchronous data transmission on optical media.

InfiniBand speed and data rate (bandwidth) differ because of the 8/10 bit ratio. Therefore, a 40 Gbps 4XQDR adapter has a data rate of 32 Gbps. Multiply the data rate by 1.25 for transmission speed and by 0.8 for the data rate from the transmission speed.

2.1.4 EtherChannel and IEEE 802.3ad link aggregation

EtherChannel and IEEE 802.3ad link aggregation are network port aggregation technologies that allow several Ethernet adapters to be aggregated together to form a single pseudo Ethernet device. For example, ent0 and ent1 can be aggregated into an EtherChannel adapter called ent3. Interface ent3 would then be configured with an IP address. The system considers these aggregated adapters as one adapter. Therefore, IP is configured over them as over any Ethernet adapter. In addition, all adapters in the EtherChannel or link aggregation are given the same hardware (MAC) address, so they are treated by remote systems as if they were one adapter. Both EtherChannel and IEEE 802.3ad link aggregation require support in the switch, so these two technologies are aware of which switch ports must be treated as one.

The main benefit of EtherChannel and IEEE 802.3ad link aggregation is that they have the network bandwidth of all of their adapters in a single network presence. If an adapter fails, network traffic is automatically sent on the next available adapter without disruption to existing user connections. The adapter is automatically returned to service on the EtherChannel or link aggregation when it recovers.

Like EtherChannel, IEEE 802.3ad requires support in the switch. Unlike EtherChannel, however, the switch does not need to be configured manually to know which ports belong to the same aggregation.

The advantages of using IEEE 802.3ad link aggregation instead of EtherChannel are that it creates the link aggregations in the switch automatically. Also, by using IEEE 802.3ad link

aggregation, you can use switches that support the IEEE 802.3ad standard but do not support EtherChannel.

In IEEE 802.3ad, the Link Aggregation Control Protocol (LACP) automatically tells the switch which ports should be aggregated. When an IEEE 802.3ad aggregation is configured, Link Aggregation Control Protocol Data Units (LACPDUs) are exchanged between the server machine and the switch. LACP assists the switch in detecting whether to consider the adapters that are configured in the aggregation as one on the switch without further user intervention.

Table 2-2 provides a brief comparison between the two modes of link aggregation.

Table 2-2 EtherChannel and IEEE 802.3ad differences

EtherChannel	IEEE 802.3ad
Requires switch configuration	Little, if any, configuration of the switch is required to form aggregation. Initial setup of the switch might be required.
Supports different packet distribution modes	Supports only standard distribution mode.

Details on the implementation and requirements for link aggregation mode can be found at the following website:

http://publib.boulder.ibm.com/infocenter/aix/v6r1/index.jsp?topic=/com.ibm.aix.commadmn/doc/commadmndita/etherchannel_intro.htm

Disaster recovery relevance

The benefit of link aggregation for interfaces in a multisite cluster is that you can minimize the number of routable IPs required for topology services to form the appropriate heartbeat rings in a network. Therefore, a lower number of logical adapters can help to simplify the topology while still achieving the same redundancy as a topology that has two or more adapters in the same network.

Take, for example, an XD_ip network that is used for heartbeating between the two sites. Rather than defining IPs for two physical adapters per node at each site, you can aggregate the adapters into a single logical one. The cluster thinks that it was working with a single adapter network, but the EtherChannel provides link redundancy and potentially load balance, depending on its implementation.

Standards: To load balance within an EtherChannel, all active links must be connected to the same switch. The backup adapter in the aggregate can be connected to a second switch to achieve switch redundancy. However, the backup adapter does not pass any traffic through it until all active links fail. The ability to load balance I/O on the network adapters might be most pertinent when you use Geographic Logical Volume Manager (GLVM) and XD_data networks because the data is replicated over the WAN. Although you can load balance between four separate XD_data networks by using GLVM, you might realize bandwidth benefits from having multiple active links within an EtherChannel in the network.

Another advantage of link aggregation is the transparency of an adapter failure. If a single physical adapter fails, it reports the loss of a link in the error report and automatically continues to pass traffic through the remaining links. Therefore, in a PowerHA XD_data network that is performing IP address takeover (IPAT), a failure bypasses a cluster SWAP_ADAPTER event and continues to operate normally without any disruption to the connections. This situation is similar to a virtualized environment. Here, a client's virtual

adapter does not experience any interruption if the physical adapters on one of the virtual I/O server (VIOS) is compromised or if only a single VIOS is taken offline.

Finally, since the introduction of the Dynamic Adapter Membership (DAM) functionality in AIX 5.2, if an adapter is created as an EtherChannel from the start, you can dynamically append extra links to it without disruption. Also, the replacement of a failed adapter is also non-disruptive, making the idea of using link aggregation, assuming that the environment is supported, even more appealing.

2.1.5 XD_rs232 networks and Serial over Ethernet

One approach to minimizing the risk of site isolation or cluster partitioning is to define as many networks as necessary to eliminate single points of failure. It is easier to implement serial networks in a local cluster by using the use of a null modem cable for an rs232 network or a disk heartbeating network over the shared SAN connection. It is more difficult to achieve this setup in a multisite environment that does not share any storage systems and presents a much larger distance between the nodes.

In these environments, the use of an XD_rs232 network helps establish an extra heartbeating link between the two sites for added redundancy. The problem with local rs232 networks in general is the distance considerations. For example, a maximum distance by specifications is 50 ft. (~15 m). A cable with a signal or enhancer converter can extend the length, possibly to 500 ft. (~15 2 m).

An option for shorter distances might be repeated rs232. This option might be useful for certain remote configurations, but most likely does not provide the distance that is required. rs232 without converters is up to 1.2 km (~.75 miles) at 9600 bps.

For longer distances, converted rs232 and Serial over Ethernet are the viable options. The choice is based on distance and comfort, while using Ethernet and TCP/IP for the serial extension.

Converted rs232 by using rs422/rs485

Using true serial requires rs422/rs485 converters. These converters take the distance to 1.2 km at 19,200 bps or 5 km (~3.1 miles) at 9600 bps (see “HACMP baud rate changes” on page 48 for changes that are used by the PowerHA Cluster Manager). In addition, the quality of the signal over the distance must be considered, making optically isolated repeaters an option. Opto-isolation uses optical isolators to isolate two serial devices. Optical isolator is a common element that is used for asynchronous data applications such as rs232, rs485, and rs422 systems. With an optical isolator, only light passes between two serial devices. This process is performed with an LED and a photosensitive transistor.

The usage of opto-isolated converters offers many benefits. Most importantly, it protects your rs232 devices from transient surges, ground loop, and remote lightning effectively. Optical isolation also eliminates ground loop and noise problems because the ground of two connected devices is separated.

Converted rs232 by using fiber optics

Using fiber optic modems/multiplexors for medium distances allows 20 - 100 km (~12 - 62 miles), but must conform to the vendor’s specifications to avoid signal loss. Such companies such as Black Box and TC Communications make devices that convert the electrical signals from an rs232 interface to light over a fiber connection, single mode, or multimode in some cases. These devices are referred to as fiber optic modems or multiplexors. The quality of the fiber between the sites determines the distances that can be achieved.

For the latest specifications, see the following vendor websites:

- Black Box

  http://www.blackbox.com/solutions/index.aspx

- TC Communications

  http://www.tccomm.com/

Devices are placed at each end of the fiber connection, and the AIX Power Systems is connected to the devices by using standard rs232 cables (and a null modem) to complete the rs232 network.

Serial over Ethernet

The Serial over Ethernet option provides the greatest distance by not defining any hard limitations, but is based on TCP/IP, which is one of the components that this type of network is designed to isolate. Planning the networks to avoid the use of the same networking infrastructure that is used for other networks, this options is viable. Also, it is key that the conversion to TCP/IP packets occurs in an external device to avoid the use of TCP/IP within the cluster node. An Internet search for Serial over Ethernet returns many vendors who provide these types of devices.

HACMP baud rate changes

By default, the tty ports that are defined as cluster communication devices and that are used as part of rs232 networks run at 38400 bps. To take advantage of the 1.2-km or 5-km distance that rs422 or rs485 extenders provide, change the default baud rate that is used:

1. Enter the **`smitty hacmp`** command.
2. Select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Network Modules** → **Change a Network Module using Custom Values** → **xd_rs232**.
3. Change the parameter settings to 9600. Setting **`Parameter=9600`** makes the cluster use a bps rate of 9600 instead of 38400.

Propagating changes: Changes that are made in this panel must be propagated to the other nodes by verifying and synchronizing the cluster.

2.1.6 Fibre Channel principles of distance

The Fibre Channel Industry Association (http://www.fibrechannel.org/) segments the Fibre Channel as follows:

Base2 Is used throughout all applications for the Fibre Channel (FC) infrastructure and devices

Base10 Is used for ISLs, core connections, and other high-speed applications that demand a maximum bandwidth

Fibre Channel over Ethernet (FCoE)
Is used to tunnel FC through Ethernet

Base2 and Base10 can use optical fiber for interconnection:

- Single-mode (SM) fiber allows only one pathway, or mode of light to travel within the fiber. The core size is typically referred to as 9 micron. Single-mode fibers are used in applications where low signal loss and high data rates are required such as on long

spans between two system or network devices where repeater/amplifier spacing must be maximized.

- Multi-mode (MM) fiber allows more than one mode of light and is suited for shorter-distance applications. Common core sizes are 50 micron and 62.5 micron.

The Fibre Channel architecture supports short-wave and long-wave optical transmitter technologies. The distance limitations for optical fiber depends on both, whether it is a multi-mode or single-mode cable and the light wave. The adapter, the equipment, and the service provider delivery specifications can override standardized distance limitations.

Specifications: International multimode and single-mode cables are specified by the following organizations:

- ISO/IEC IS11801 for multimode cables

  http://www.iso.org

- ITU-T for single-mode cables

  http://www.itu.int/ITU-T

For more information about FC and SAN, see the *Introduction to Storage Area Networks and System Networking*, SG24-5470.

2.1.7 DWDM

Nowadays, the expansion of IP-based technology is becoming faster and faster. Many devices and applications that did not use an IP-based layer for their data transmission before, such as TV, video, telephone (voice and data), and multiplayer gaming, have moved to an IP-based layer now.

Most of those devices and applications are bandwidth-intensive and real time, which results in the need for special methods and devices to support their requirements. Dense Wavelength Division Multiplexing (DWDM) is one of the solutions that can support that requirement, especially for data transmission. DWDM is the process of multiplexing signals of different wavelengths onto a single fiber. Through this operation, it creates many virtual fibers, each capable of carrying a different signal. At its simplest, a DWDM system can be viewed as a parallel set of optical channels, each using a slightly different light wavelength, but all sharing a single transmission medium. This new technical solution can increase the capacity of existing networks without the need for expensive recabling and can tremendously reduce the cost of network upgrades.

Internet Protocol (IP) over DWDM is the concept of sending data packets over an optical layer by using DWDM for its capacity and other operations. The optical network provides end-to-end services completely in the optical domain without having to convert the signal to the electrical domain during transit. Transmitting IP directly over DWDM supports bit rates of OC-192 (transmitting speed of up to 9953.28 Mbps). For more information about network technologies bandwidth examples, see Table 2-1 on page 44.

Over the past five years, many service providers have deployed DWDM networks as the underlying and enabling layer 0/1 (physical layer and data link layer) optical technology. This technology easily supports new protocols, such as GigE or 10 GigE and Fibre Channel, in their basic formats.

How does the DWDM work? Figure 2-1 illustrates a simple explanation.

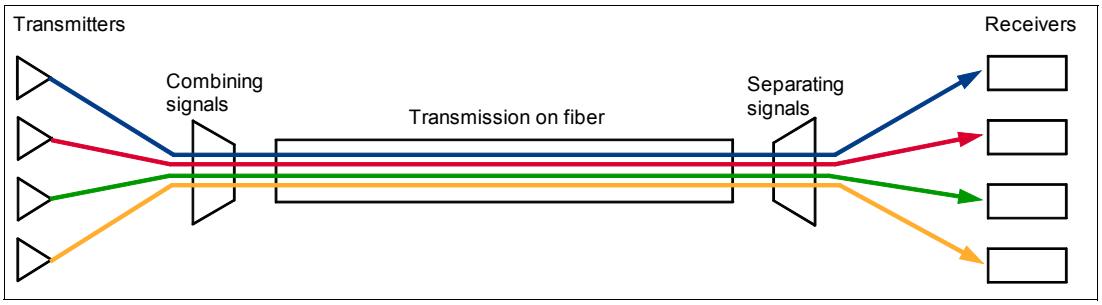


Figure 2-1 DWDM functional schematic

As illustrated in Figure 2-1, DWDM performs the following functions:

- Generates the signal. The source, a solid-state laser, must provide stable light within a specific narrow bandwidth that carries the digital data, which is modulated as an analog signal.
- Combines the signals. Modern DWDM systems employ multiplexers to combine the signals. A certain inherent loss is associated with multiplexing and demultiplexing. This loss depends on the number of channels, but can be mitigated with optical amplifiers, which boost all the wavelengths at once without electrical conversion.
- Transmits the signals. The effects of crosstalk and optical signal degradation or loss must be reckoned with in fiber optic transmission. These effects can be minimized by controlling variables such as channel spacings, wavelength tolerance, and laser power levels. Over a transmission link, the signal might need to be optically amplified.
- Separates the received signals. At the receiving end, the multiplexed signals must be separated out.
- Receives the signals. The demultiplexed signal is received by a photo detector.

In addition to these functions, a DWDM system must also be equipped with client-side interfaces to receive the input signal. This function is performed by transponders.

Figure 2-2 shows the complete operation of unidirectional DWDM with transponders.

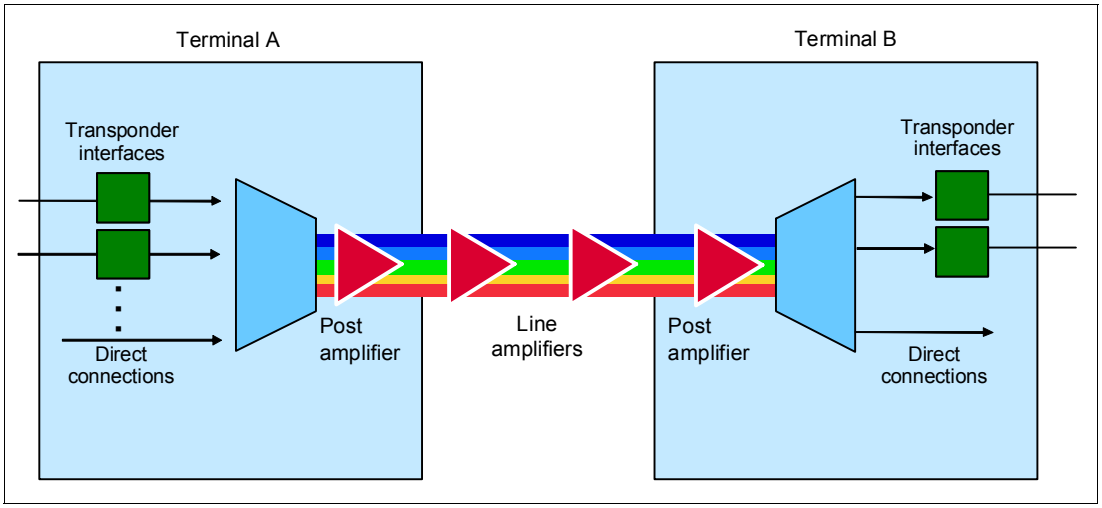


Figure 2-2 End-to-end process of a unidirectional DWDM system

Figure 2-2 illustrates the following process:

1. The transponder accepts input in the form of standard single-mode or multimode laser. The input can come from different physical media and different protocols and traffic types.
2. The wavelength of each input signal is mapped to a DWDM wavelength.
3. DWDM wavelengths from the transponder are multiplexed into a single optical signal and launched into the fiber. The system might also include the ability to accept direct optical signals to the multiplexer. For example, such signals can come from a satellite node.
4. A pre-amplifier boosts the strength of the optical signal as it leaves the system.
5. Optical amplifiers are used along the fiber span as needed.
6. A post-amplifier boosts the signal before it enters the end system.
7. The incoming signal is demultiplexed into individual DWDM lambdas (or wavelengths).
8. The individual DWDM lambdas are mapped to the required output type and sent out through the transponder.

Implementing DWDM technology, especially for IP over DWDM, offers the following advantages among others:

- Eliminate network layers and reduce complexities and equipment costs.

  A DWDM optical transport network can extend from the core to the edge and access networks. Reconfigurable Optical Add-Drop Multiplexers (Reconfigurable Optical Add-Drop Multiplexers (ROADMs)) can eliminate the complex optical-electrical-optical (O-E-O) layer, reducing the number of network elements. For example, traffic that needs switching or routing can be dropped to an appropriate device, whereas traffic that does not benefit from a sublambda packet processing can be optically switched. In effect, this collapsing of the optical network layers dramatically reduces network complexities and operating costs.

- Improve resource use to achieve optimal bandwidth efficiency.

  A fully integrated Layer 2 Ethernet (IP) over DWDM network can adapt to varying traffic demands. By combining the packet-processing intelligence and optical-wavelength assignment into a single system, service providers can avoid excessive inventory and the restrictions that are imposed by per-port protection. The result is a cost-effective system that makes better use of its resources.

- Simplify end-to-end provisioning to speed time to market.

  Most DWDM network devices are linked by using software interfaces to ensure that the link is installed exactly as planned. Work orders are issued automatically and sent from the network manager down to the network elements. Reconfiguration or installation of new channels is done continuously within the software suite, whereas the data can be monitored and logged automatically and verified by using the network.

- Automate network management for scalability and reduced operating expenses.

  By supporting open and standard interfaces, the DWDM network management system can be integrated into virtually any high-level or low-level management system, allowing for further reduction in operating expenses.

- Detect problems automatically and resolve them faster across the entire network.

With a Layer 2 Ethernet (IP) over DWDM network, you can achieve a much higher resiliency and simplified operations. For example, a dropped signal typically causes alarms to go off everywhere in the network across amplifiers, multiplexers, transponders, and other network elements. A DWDM network is able to correlate and isolate the problem to the direct source much faster, eliminating unnecessary alarms in the network and resolving problems more quickly.

2.1.8 Firewalls

A firewall is a device that protects private local area networks (LANs) from intrusion of computers that are accessing the Internet. It can be a hardware device or a software program that is running on a secure host that provides a gateway between two networks. The device has two network interfaces, one in the protected zone and one in the exposed one. The firewall examines all inbound and outbound traffic and filters it based on the specified criteria.

Firewalls can filter packets based on their source and destination addresses and different port numbers, which is known as *address filtering*. Firewalls can also filter specific types of network traffic, which is known as *protocol filtering*. The decision to forward or reject traffic depends on the protocol that is used, for example, HTTP, FTP, or telnet. Firewalls can also filter traffic by packet attribute or state.

Network architecture is designed around a seven-layer model. In a network, a single protocol can travel over more than one physical support (layer one) because the physical layer is dissociated from the protocol layers (layers three to seven). Similarly, a single physical cable can carry more than one protocol. The TCP/IP model is older than the OSI industry standard model, which is why it does not comply in every respect. The first four layers are so closely analogous to OSI layers, however, that interoperability is a day-to-day reality.

Figure 2-3 shows the contrast between the OSI and Internet Protocol network models.

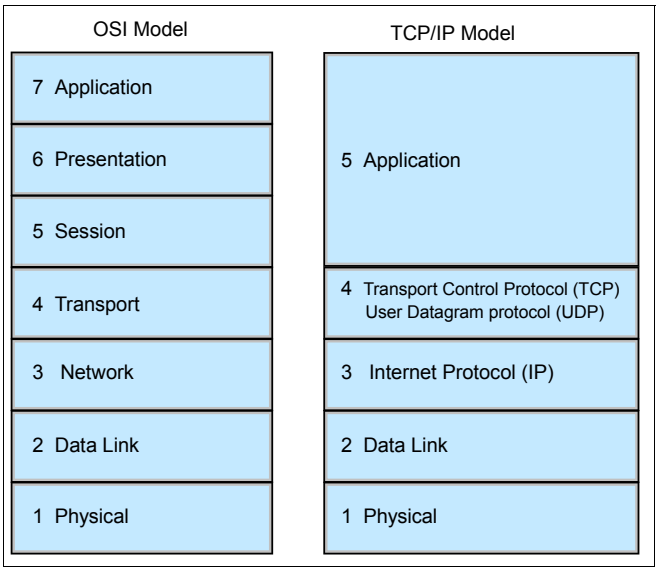


Figure 2-3 The OSI and Internet Protocol network models

Firewalls operate at different layers and can use different criteria to restrict traffic. The lowest layer at which a firewall can work is layer three. In the OSI model, this layer is the network layer. In TCP/IP, it is the Internet Protocol layer. This layer is concerned with routing packets to their destinations. At this layer, a firewall can determine whether a packet is from a trusted source, but it cannot be concerned with what it contains or what other packets it is associated with. Firewalls that operate at the transport layer know a little more about a packet and can grant or deny access, depending on more sophisticated criteria. At the application level, firewalls know a great deal about what is going on and can be selective in granting access.

Clustered environments might or might not have firewalls in place between their servers. Even when they do, they might not implement them in the same way. We provide details about what we used in our environment. Our test environments were configured to replicate

between buildings in our complex. A firewall was in place between each building and we opened up ports and enabled rules for our cluster communication to work.

The first step was to identify the cluster network topology. When we determined the routable IPs required at each site, we placed a request to the network team to allow communication between the corresponding pairs. Figure 2-4 shows a logical view of the firewalls in place between the sites that were used for our test scenarios.

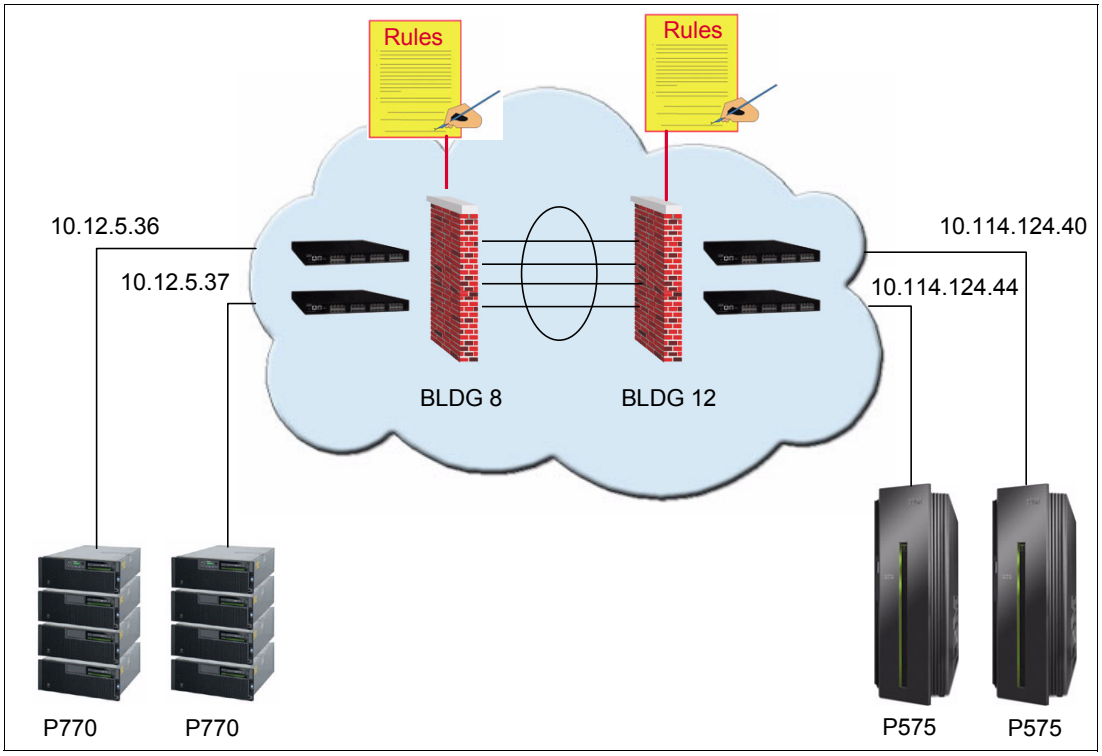


Figure 2-4 Firewall environment in our test scenarios

The following firewall rules were implemented at both sites:

- Inbound firewall rules for building 2:

```
10.12.5.36  to 10.114.124.40     -    All Ports Open
10.12.5.36  to 10.114.124.44     -    All Ports Open
10.12.5.37  to 10.114.124.40     -    All Ports Open
10.12.5.37  to 10.114.124.44     -    All Ports Open
A rule for no authentication for 10.12.5.36 to  10.114.124.40
A rule for no authentication for 10.12.5.36 to  10.114.124.44
A rule for no authentication for 10.12.5.37 to  10.114.124.40
A rule for no authentication for 10.12.5.37 to  10.114.124.44
```

- Outbound firewall rules for building 12:

```
10.114.124.40 to 10.12.5.36    - All Ports Open
10.114.124.44 to 10.12.5.36    - All Ports Open
10.114.124.40 to 10.12.5.37    - All Ports Open
10.114.124.44 to 10.12.5.37    - All Ports Open
A rule for no authentication for 10.114.124.40 to 10.12.5.36
A rule for no authentication for 10.114.124.44 to 10.12.5.36
A rule for no authentication for 10.114.124.40 to 10.12.5.37
A rule for no authentication for 10.114.124.44 to 10.12.5.37
```

We began our testing with all ports open to avoid any issues and later restricted it to only the required ports.

Required ports

For the ports that are used by Resource Monitoring and Control (RMC), Hardware Management Console (HMC), and Dynamic Logical Partition (DLPAR) operations, see “More Dynamic LPAR considerations” on page 30.

The `/etc/services` file defines the sockets and protocols that are used for network services on a system. The ports and protocols that are used by the PowerHA components are defined as follows:

```
clinfo_deadman 6176/tcp
clinfo_client 6174/tcp
clsmuxpd 6270/tcp
clm_lkm 6150/tcp
clm_smux 6175/tcp
godm 6177/tcp
topsvcs 6178/udp
grpsvcs 6179/udp
emsvcs 6180/udp
clcomd 6191/tcp
```

In addition, when the PowerHA Enterprise Edition is installed, the following entry for the port number and connection protocol is automatically added to the `/etc/services` file:

```
rpv 6192/tcp
```

The entry is added on each node on the local and remote sites on which you installed the software. This default value enables the RPV server and RPV client to start immediately after they are configured (that is, to be in the available state). For more information, see the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide,* SA23-1338.

2.2 Cluster topology considerations

Factors such as the number of nodes at each site and whether the network segments span between the sites affect the design of the cluster topology. This section highlights the factors, such as the failure detection rate settings for XD network types, and the concept of performing IP address takeover (IPAT) between two sites and how it can be accomplished.

2.2.1 Cluster topologies

Typical and supported topologies for disaster recovery for the PowerHA Enterprise Edition are with two sites. As with all types of high availability and continuity, the topology must reflect the business and continuity requirements, and strive to adhere to the principle of keeping it simple and straightforward.

Unforeseen circumstances can occur when a recovery solution is called for, and more complex designs carry even a greater degree of risk. Unforeseen circumstances can affect the reliability and capability to provide services when a disastrous situation arises. Therefore, deploying a more complex solution than necessary increases the risk of failure.

The following two-site topology designs are the most common:

- Two nodes at each site
- Two nodes at the primary site and one node at the recovery site
- One node at each site

The reasoning for the most common two-sites topology designs is simple. The risk that the primary site fails completely exists but does not occur frequently. The risk that part of the primary site fails is greater than the entire site and occurs with more frequency. Therefore, ensuring high availability at the primary site helps mitigate the higher probability of failure, and adding the disaster recovery capability at a secondary site mitigates the risk of a total site failure.

It is imperative to understand the communication paths for the selected topology and technology. Figure 2-5 illustrates this point.

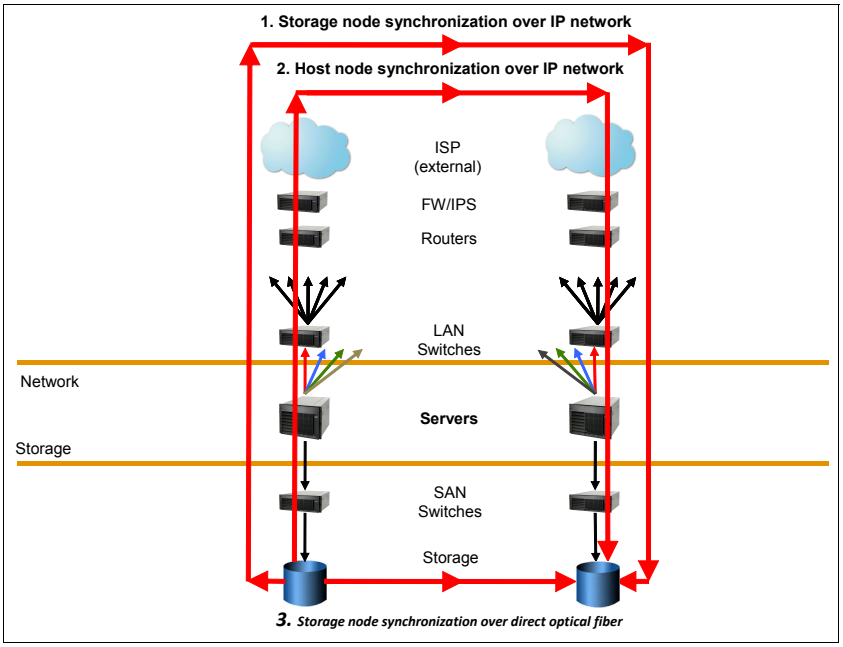


Figure 2-5 Three types of communication topology for synchronization

For more information, see the *HACMP for AIX Geographic LVM: Planning and Administration Guide*, SA23-1338. See also the selected and relevant storage solutions planning and administration guides.

As a base reference for topology, consider the basic dual-node PowerHA cluster that is illustrated in Figure 2-6.

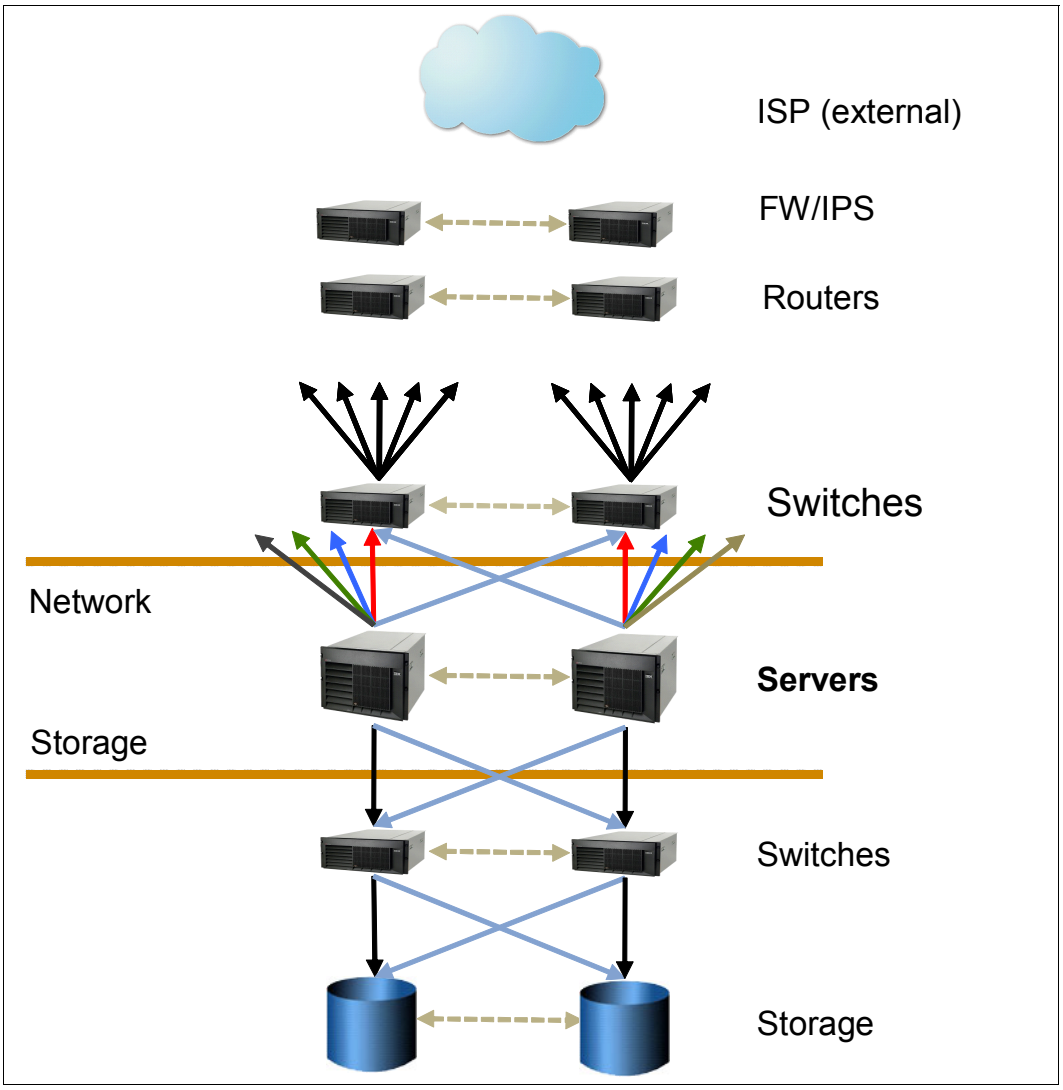


Figure 2-6 Basic PowerHA dual node cluster

Two nodes at each site

In the two nodes at each site in Figure 2-7, each site has a working pair of nodes that back up one another, and both sites can run services in one cluster across sites.

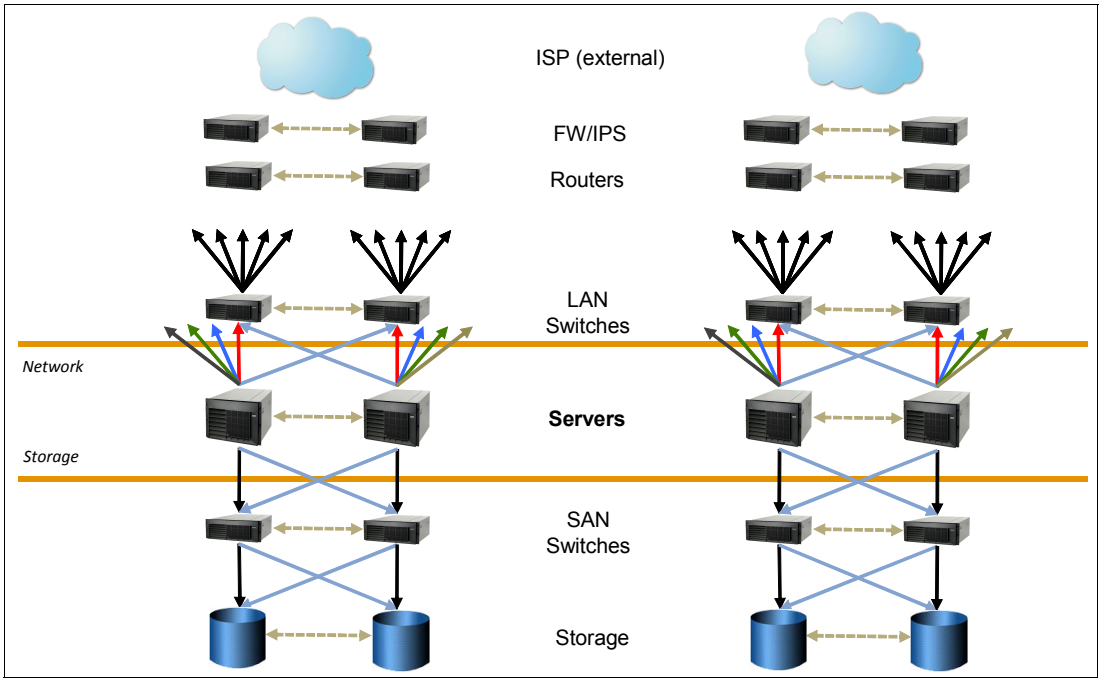


Figure 2-7 Two nodes at each site

Between the sites, the storage synchronization is performed from the active node at one site to the secondary site for that service. Therefore, if both sites have both active and passive services, the storage synchronization goes in both directions.

All component failures are handled locally at each site. Only site failures or planned site maintenance result in failover across sites. This configuration is preferred if you have applications active at both sites to maximize the utilization of the nodes. We used manual failback after the primary site failover to the remote or recovery site.

Two nodes at the primary site and one node at the recovery site

In the two nodes at the primary site shown in Figure 2-8, the site has a working PowerHA cluster. Also both nodes can run services with one node as the primary node and one node as the secondary node locally. The secondary-site node is commonly dedicated for recovery if the primary site fails. This node usually does not provide online services until such time.

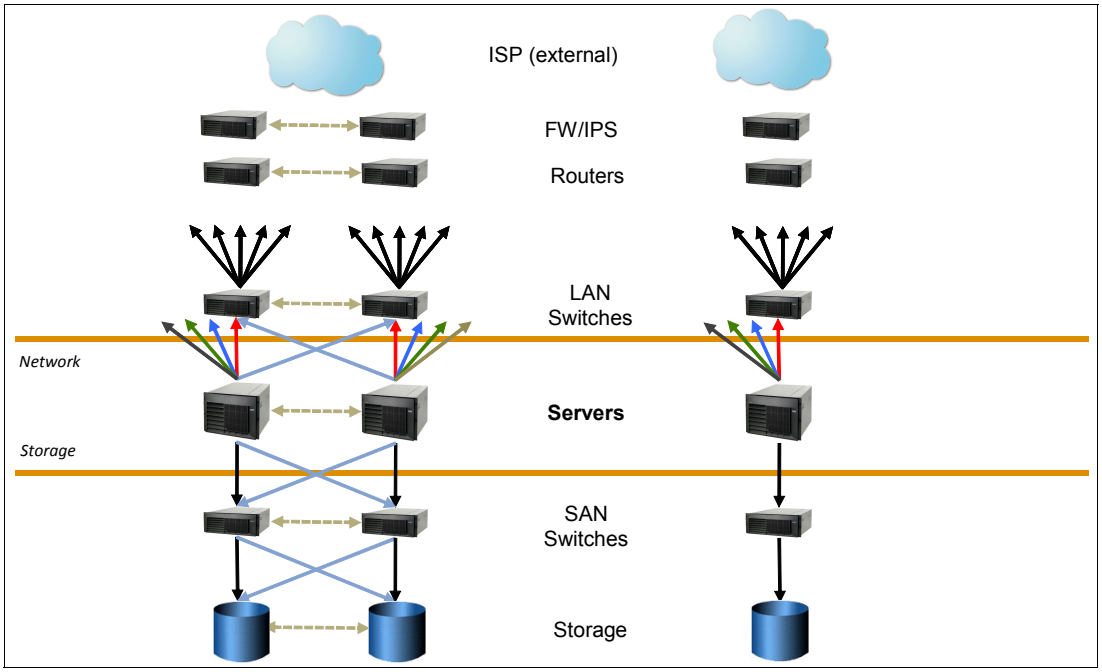


Figure 2-8 Two nodes at the primary site and one node at the recovery site

Between the sites, the storage synchronization is performed from the active node at the primary site to the secondary site. All component failures are handled locally only at the primary site. The secondary site elevates any node failure to a site failure.

This configuration is supported. However, if maintenance is being performed on the backup site, the nodes at the backup site might be down. Also the data that is updated at the primary site cannot be synchronized to the backup site if host synchronization is used. When the node at the backup site comes up, the primary site must synchronize all changes that occurred while the backup site was down.

One node at each site

In the one node at each site configuration, as shown in Figure 2-9, the nodes can be configured either as a PowerHA cluster with shared storage or as a PowerHA Enterprise Edition cluster with separate and synchronized storage. Both nodes can run services with one node as the primary node and one node as the secondary node. The secondary-site node is commonly dedicated for recovery if the primary site fails. This node usually does not provide online services until such time.

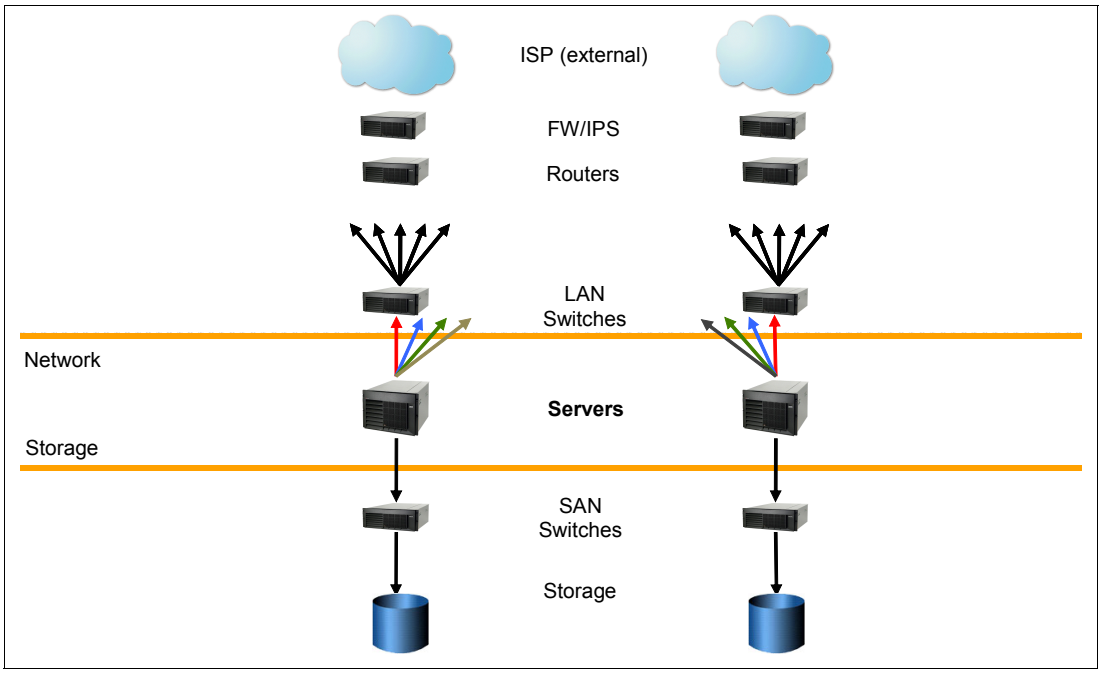


Figure 2-9 One node at each site

Between the sites, the storage synchronization is performed from the active node at the primary site to the secondary site. This design is more restrictive. It can handle disk or adapter failures locally, but node failures are propagated to site failures because no local peer nodes are available in this model.

2.2.2 Topology failure detection rates

The failure detection rate (FDR) is the number of consecutive seconds that it takes the cluster to classify an interruption as a problem and to trigger an automated action. The following parameters are involved in determining the failure detection rate:

- Heartbeat rate (in seconds): The frequency at which heartbeats (Keepalives) are sent between the nodes. You can check the status of the missed heartbeat by running the following command:

  ```
  # lssrc -ls topsvcs
  ```

- Failure detection cycle: The number of consecutive heartbeats that must be missed before failure is assumed.

The following formula is used to calculate the failure detection rate:

```
Heartbeat rate x Failure cycle x 2 = Failure Detection rate
2 x 10 x 2 = 20 seconds
```

The most commonly used PowerHA network types are *ether* for IP networks and *rs232* and *diskhb* for non-IP heartbeat traffic. Other network types are available for configuring clusters that span multiple sites. Other than when using an implementation with GLVM, the only difference between an ether type network and an XD_ip network is the amount of time that the cluster waits before declaring an interface failure and initiating an adapter swap.

Table 2-3 contrasts the failure detection rate between the IP network types. The FDR that is used by a cluster network interface module defaults to the predefined Normal value, as shown in the rows in Table 2-3.

Table 2-3 Failure detection rate - IP network types

Type ether network setting	**Seconds between heartbeats**	**Failure cycle**	**Failure detection rate**
Slow	2	12	48
Normal	1	10	20
Fast	1	5	10
XD_ip and XD_data Multisite networks	**Seconds between heartbeats**	**Failure cycle**	**Failure detection rate**
Slow	3	16	96
Normal	2.5	12	60
Fast	2	12	48

Table 2-4 contrasts the detection rate between serial non-IP network types commonly used in a local cluster and an XD_rs232 network that can be implemented between two sites.

Table 2-4 Failure detection rate for non-IP networks

rs232 and diskhb network types	**Seconds between heartbeats**	**Failure cycle**	**Failure detection rate**
Slow	3	8	48
Normal	2	5	20
Fast	1	5	10
XD_rs232 network type	**Seconds between heartbeats**	**Failure cycle**	**Failure detection rate**
Slow	3	10	60
Normal	2.5	6	30
Fast	2	6	24

The values that are documented in Table 2-3 and Table 2-4 correspond to the three predefined FDR values within the cluster software, where the default value is *Normal*. Although the default values are typically sufficient, if different values are desired, the setting for each network type can be customized to slower or faster settings.

2.2.3 IPAT across sites and DNS considerations

IP address takeover by IP aliases (Figure 2-10) is the default method for IP label and address recovery in PowerHA. For an IP to move between nodes, the cluster must have a service IP label that is defined within a resource group. The IP can then be moved by using aliasing between local nodes at the same site or across sites.

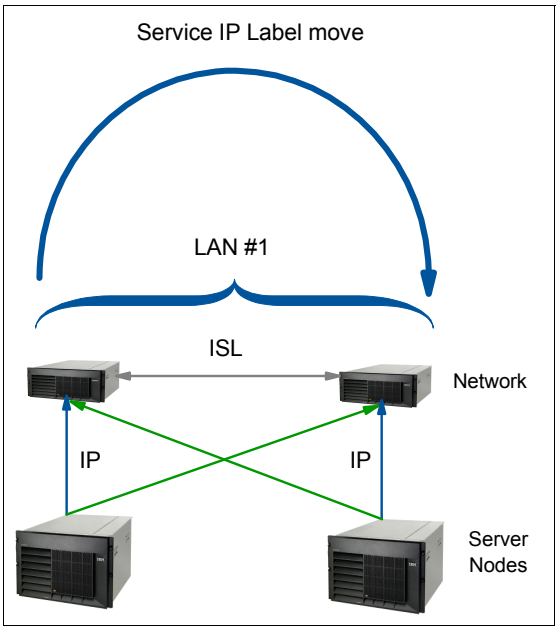


Figure 2-10 Default IP address takeover by using IP aliases

This method is the most common deployment in single-site PowerHA clusters. It can sometimes be extended between sites by using virtual LAN (VLAN) technology (VLANs are often associated with IP subnetworks), such as VLAN trunking protocol. This capability usually is made up of one or more network devices that are interconnected with trunks such as with Inter-Switch Link (ISL). When an IP service label move is requested or required, the server nodes are all in the same subnet (VLAN). Also the client nodes outside of the cluster can still use this service IP label to access the resource group service.

An IP address at one site might not be valid at the other site because of subnet issues. Therefore, it might not be possible or desirable to interconnect the network devices, although it is possible to configure site-specific service IP labels to handle this situation.

Site-specific service IP labels (Figure 2-11) were introduced to activate only on a corresponding site. The service IP label can fail over to other nodes within the same site. In a cluster configuration where each site is on a separate network segment, you can define a service IP address for each one and append both service labels into the same resource group. Based on where the resource group is being hosted, only the appropriate service IP address is brought online.

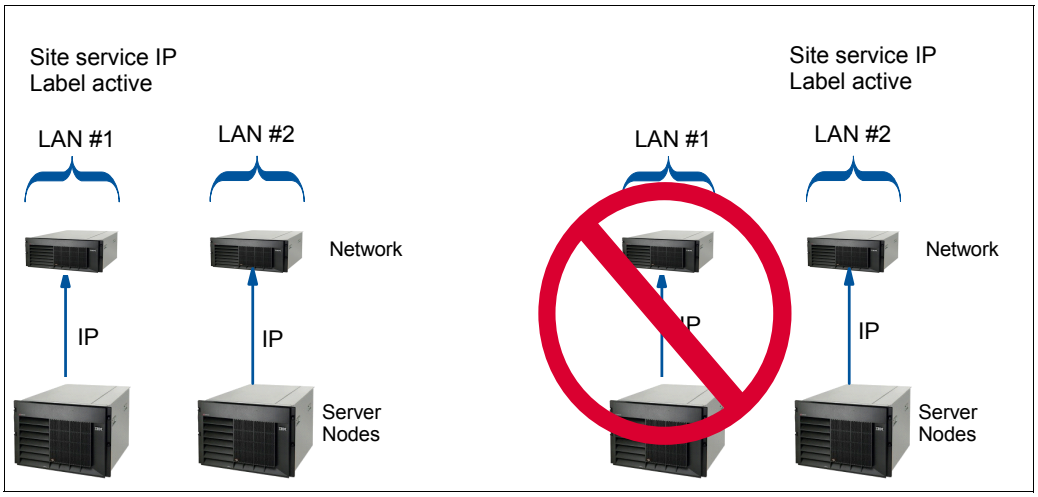


Figure 2-11 Site-specific service IP labels

Except for the site-specific limitation, these service IP labels have all the same functions as regular service IP labels, except the ability to work with NFS cross-mounts between the sites. The panel in Figure 2-12 shows the additional attribute that is available when you define service IP labels in PowerHA.

```
                Change/Show a Service IP Label/Address (extended)

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

                                                      [Entry Fields]
* IP Label/Address                                    svcxd_a1_sv
  New IP Label/Address                               []                      +
  Netmask(IPv4)/Prefix Length(IPv6)                  [24]
* Network Name                                       [net_ether_01]          +
  Alternate HW Address to accompany IP Label/Address []
  Associated Site                                    svc_sitea               +
```

Figure 2-12 PowerHA cluster panel showing a site-specific service IP

On the same cluster, Example 2-1 shows how the resource group service IP definition might look.

Example 2-1 Resource group service IP definition

```
# clshowres
Resource Group Name                                 RG_sitea
Participating Node Name(s)                          svcxd_a1 svcxd_a2 svcxd_b2
svcxd_b1
Startup Policy                                      Online On First Available Node
Failover Policy                                     Failover To Next Priority Node
                                                       In The List
Failback Policy                                     Never Failback
Site Relationship                                   Prefer Primary Site
                                                       Node Priority
Service IP Label                                    svcxd_a1_sv svcxd_b1_sv
.............
```

For more information about setting up site-specific service IP labels, see the *HACMP for AIX 6.1 Administration Guide*, SC23-4862

http://publib.boulder.ibm.com/infocenter/aix/v6r1/topic/com.ibm.aix.hacmp.admngd/hacmpadmngd_pdf.pdf

With site-specific service IP labels, consider how the client nodes outside of the cluster become aware that the service is provided over another IP address. The most efficient and reliable solutions are when the client node application that uses the cluster node service is cluster aware and can detect and reroute communication from the failing node or site to the active node or site as required.

You can use several ways to inform client nodes of IP service label changes. However, examine how and when the client application handles IP address resolution, whether it is on startup or when needed. Upon startup, it usually requires a restart of the application. When needed, it might introduce name resolution delays, which in turn might be handled outside of the application by using either static host tables (/etc/hosts) or DNS caching only on the local client node. Dynamic alteration of DNS records is another possibility, but avoid manipulating DNS structures if possible.

Time settings: DNS changes are not propagated in real time throughout a segmented DNS structure. Individual records either have the Start of Authority (SOA) time settings, or might have individual record-specific time settings. In either case, they are not real time.

Example 2-2 shows a simple nslookup of the ibm.com domain. The expire attribute in this case is seven days. It indicates how long the information can be held before it is no longer considered authoritative, such as with a secondary or cache-only domain name server. In this case however, the refresh is once every hour, which is the time interval between polling checks from the secondary to the primary for record changes (indicated by a higher serial number).

Example 2-2 Using nslookup to view domain name SOA record

```
bjro@paco24:/Users/bjro: nslookup
> set querytype=soa
> ibm.com
Server:192.168.0.254
Address:192.168.0.254#53
```

```
Non-authoritative answer:
ibm.com
   origin = ns.watson.ibm.com
   mail addr = dnstech.us.ibm.com
   serial = 2010040100
   refresh = 3600
   retry = 1800
   expire = 604800
   minimum = 10800

Authoritative answers can be found from:
ibm.comnameserver = internet-server.zurich.ibm.com.
ibm.comnameserver = ns.watson.ibm.com.
ibm.comnameserver = ns.austin.ibm.com.
ibm.comnameserver = ns.almaden.ibm.com.
ns.almaden.ibm.cominternet address = 198.4.83.35
internet-server.zurich.ibm.comhas AAAA address 2001:620:20:fe01:100::2000
ns.austin.ibm.cominternet address = 192.35.232.34
ns.watson.ibm.cominternet address = 129.34.20.80
```

If the client nodes can use a domain name server with support for the RFC 2782, you can use the DNS Resource Record (RR) SRV (SeRVice) to move a service from host to host and to “designate some hosts as primary servers for a service and others as backups.” For more information, go to RFC 2782 at:

http://tools.ietf.org/html/rfc2782

The SRV RR has the following format:

```
_Service._Proto.Name TTL Class SRV Priority Weight Port Target
```

As stated in RFC 2782, “A client *must* attempt to contact the target host with the lowest-numbered priority it can reach”.

2.3 Storage considerations

The storage configuration is one of the most important tasks that you have to consider when planning for the PowerHA Enterprise Edition cluster configuration. In a local cluster configuration, you must consider a few elements, such as the following examples, when you are planning for the storage configuration:

- The shared and non-shared data for your application environment
- Concurrent or non-concurrent access requirements to the shared volumes
- The volume groups
- Logical volumes
- File system configuration

In most cluster configurations, application data is stored on disk volumes. It is shared between the cluster nodes to enable access to the data from the backup node in case the primary node running an application and performing the I/O operations of the data fails.

When you are planning for a PowerHA Enterprise Edition solution, the data availability is the key factor for a successful recovery in a secondary location if a disaster occurs at the primary site. Because data in separate locations is in separate storage, data availability at the

secondary location can be provided by various data replication technologies. PowerHA Enterprise Edition offers support for two data replication options across extended distance limits:

- Using IBM GLVM (synchronous and asynchronous), based on underlaying TCP/IP replication to send the data updates from the primary nodes to nodes in a remote location
- Integrating storage replication technologies in a flexible and reliable cluster framework to provide the highest tier disaster recovery solution

The storage subsystems that are supported by PowerHA cover all IBM disk subsystem classes, starting with the entry-level modes to the enterprise storage. The particular models that are supported are documented in the sales manual for PowerHA on AIX. PowerHA supports these IBM storage devices that have passed IBM qualification efforts, and for which IBM development and service are prepared to provide support. You can also use third-party (original equipment manufacturer (OEM)) storage devices and subsystems, although most of them are not directly certified by IBM for PowerHA usage. For these devices, check the manufacturer's respective websites for information that supports the IBM PowerHA SystemMirror solution.

With the PowerHA Enterprise Edition, clients who use EMC storage with Symmetrix Remote Data Facility (SRDF) can take advantage of the PowerHA SystemMirror capabilities for HA/DR operations. EMC clients might also deploy PowerHA SystemMirror Standard Edition for data center operations.

The following sections highlight considerations for applying each type of environment that involves site configuration:

- "PowerHA using cross-site LVM mirroring" on page 66
- "PowerHA Enterprise Edition with storage replication" on page 67
- "PowerHA Enterprise Edition with GLVM" on page 75

2.3.1 PowerHA using cross-site LVM mirroring

PowerHA using cross-site LVM mirroring is a campus-style solution. It is supported by the base PowerHA product. The configuration is similar to a local PowerHA cluster, except that nodes and mirrored physical volumes are associated with sites. The data on the storage devices in each site is shared transparently between the cluster nodes by using the SAN devices that are interconnected between sites. Figure 2-13 illustrates a cross-site LVM cluster.

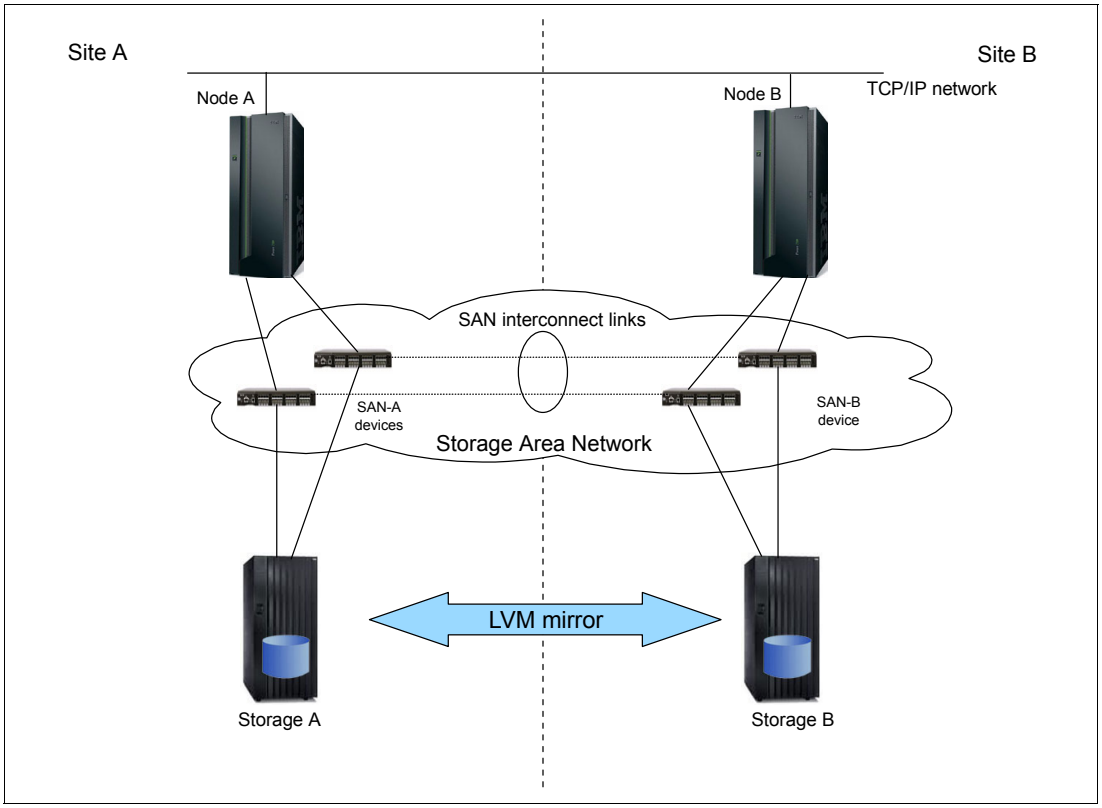


Figure 2-13 Storage sharing for cross-site LVM configurations

Independent storage subsystems are on each site. Storage subsystems might be of different models and providers. Although it is not a requirement, for performance and functionality considerations, we implemented storage environments with similar characteristics at both sites.

SAN zoning and volume mapping on each storage environment must be configured to enable each node in the cluster to access both local and remote storage volumes.

The SAN can be expanded beyond one site by using communication-extending technologies that also determine the maximum distance of the FC connections between the sites. The following list provides examples of the types of technologies that might be used:

- Direct FC links between FC switches that use long-wave gigabit interface converter (GBIC) modules in switches at both sides and single-mode interconnect optical links
- Fibre Channel over IP (FCIP) connections that use independent modules, or specific modules on SAN switches, to transport the FC frames over TCP/IP communication links
- Wave division multiplexing (WDM) devices, which includes coarse wavelength division multiplexing (CWDM) and dense wave-length division multiplexing (DWDM)

  For more information about WDM devices, see 2.1.7, “DWDM” on page 49.

When you use cross-site LVM, consider changing the default values of the following attributes for the FC SCSI protocol devices (fscsi) on all host bus adapter (HBA) ports:

- **fc_err_recov = fast_fail** to enable fast I/O failure recovery
- **dyntrk = yes** to enable the FC device drivers to reroute the traffic to the target device in case the SCSI ID has changed

To check the current attributes of the HBA ports, see the **lsattr** command that is shown in Example 2-3.

Example 2-3 State of the FC SCSI protocol devices

```
# lsattr -El fscsi0
attach       switch    How this adapter is CONNECTED         False
dyntrk       yes       Dynamic Tracking of FC Devices        True
fc_err_recov fast_fail FC Fabric Event Error RECOVERY Policy True
scsi_id      0x70400   Adapter SCSI ID                       False
sw_fc_class  3         FC Class for Fabric                   True
```

For more information about the two attributes and their interaction, see *Fast I/O Failure and Dynamic Tracking interaction* in the System p and AIX Information Center at:

http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/fast_fail_dynamic_interaction.htm

The shared data in the cluster is mirrored between the storage volumes that are placed in different sites by the AIX LVM. LVM data mirroring keeps both copies of the data in sync. PowerHA drives automatic LVM mirroring synchronization, and after the failed site joins the cluster, it automatically fixes removed and missing volumes (physical volumes, or PV, states removed and missing) and synchronizes the data. Automatic synchronization is not possible for all cases, but you can use C-SPOC to synchronize the data from the surviving mirrors to stale mirrors after a disk or site failure.

2.3.2 PowerHA Enterprise Edition with storage replication

In an environment that uses storage replication, the data replication services are provided by the storage subsystem that is driven by the PowerHA Enterprise Edition software. The cluster software that runs on the nodes communicates with the storage subsystem to perform specific operations on the replicated volume pairs during the resource group processing.

Figure 2-14 illustrates a typical environment that uses storage replication.

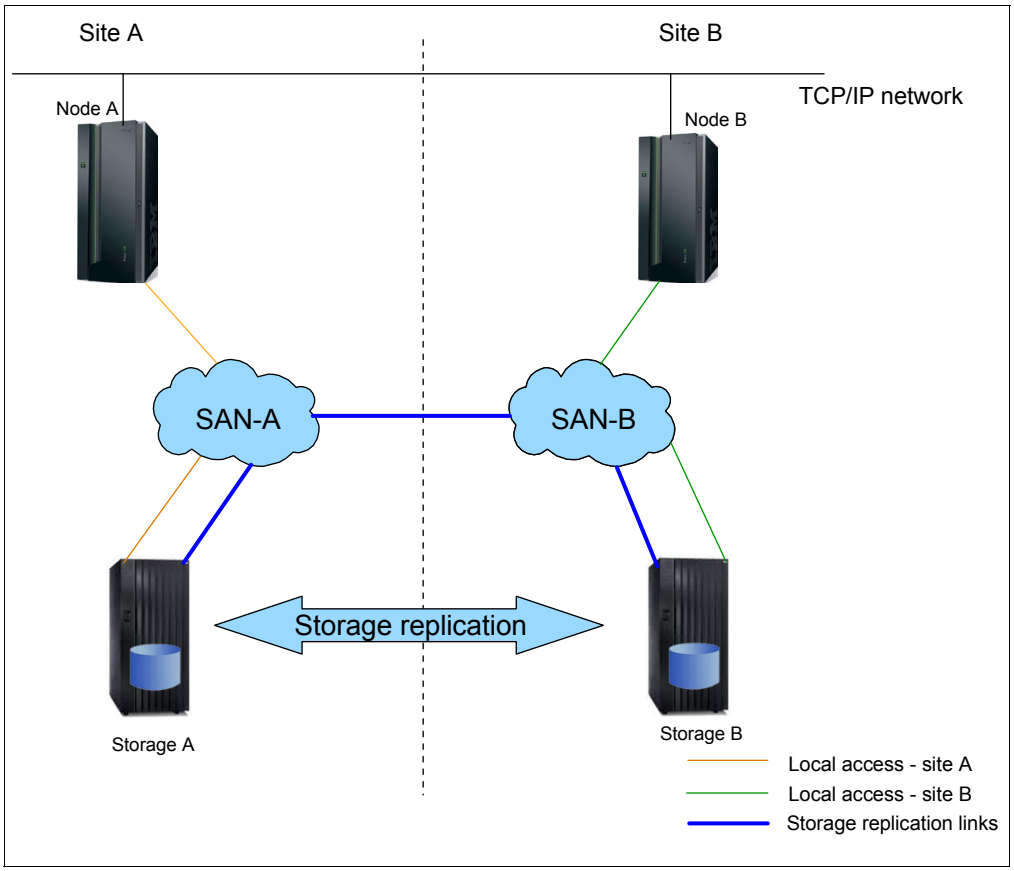


Figure 2-14 Typical PowerHA Enterprise Edition environment using storage replication

Nodes in each site have access to the local storage by using either SCSI or FC links. No path is defined between a node in a site, and the storage from the remote site to access the data volumes on the remote storage during a failover or failback of the replicated volume pairs.

In general, storage subsystems from both sites are connected for data replication by using dedicated communication links. Depending on the storage type, there are a couple of technologies available for the replication links (see 2.1.6, “Fibre Channel principles of distance” on page 48, and 2.1.7, “DWDM” on page 49):

- Fibre Channel (FC) and FCIP (transports the FC frames over an Internet Protocol network)
- IBM ESCON®

Consult the storage product and the PowerHA Enterprise Edition documentation for the supported types of technologies available for a specific product and PowerHA Enterprise Edition version.

The cluster nodes need to access the storage subsystems at both sites to manage the replicated pairs. Depending on the storage subsystem and the management software that is used with PowerHA Enterprise Edition, this type of connection can be supported in the following ways:

- An Internet Protocol network that uses one or two IP segments that are routed between sites with PowerHA Enterprise Edition that uses IBM storage subsystems
- FC/SCSI links between the nodes and the storage in each site, and the storage replication links for EMC Symmetrix replication (SRDF)

PowerHA Enterprise Edition 6.1 supports the PowerHA Enterprise Edition for Metro Mirror Software environment for storage-based replication (as per the 6.1 Release Notes, only ESS Model 800 is supported). This solution is available for the following IBM storage systems:

- IBM Enterprise Storage Server (ESS) Model 800
- IBM System Storage Server (DS) Models 6000 and 8000
- IBM System Storage SAN Volume Controller
- EMC Symmetrix with Synchronous and Asynchronous SRDF

Overview of the management type for PPRC replication

Peer-to-Peer Remote Copy (PPRC) is a hardware mirroring technique that is used by IBM TotalStorage ESS for data replication. It mirrors the data at the disk subsystem level, making it transparent to the hosts. Throughout this book, this term is used to refer the data replication technologies on the IBM Enterprise Storage Subsystems (DS8000, DS6000, ESS) and SAN Volume Controller.

PowerHA Enterprise Edition supports both synchronous and asynchronous PPRC replicated volumes, depending on the hardware configuration. The method that is used to manage the PPRC pairs also depends on the available hardware configuration. Table 2-5 shows the type of PPRC configuration that you can choose to manage your cluster.

Table 2-5 Options for PPRC management in a PowerHA Enterprise Edition cluster

Mirror type	Hardware type	PPRC is managed by	How HACMP manages PPRC pairs
Synchronous	ESS 800	Copy Services Server[a] (CSS, on storage controller)	Directly manages the failover and resynchronization of the PPRC pairs by issuing commands directly to the ESS systems. It is referred as *direct management*.
Synchronous	ESS (800) or DS (8000, 6000) or intermix of any of these	DSCLI management, via ESSNI[b] Server on either storage controller or HMC	Relies on the ESSNI server to manage PPRC pairs. Directly manages the failover and resynchronization of the PPRC pairs by issuing commands directly to the storage systems. It is referred as *DSCLI*[c] *management*.
Synchronous/ asynchronous	SAN Volume Controller (hardware as supported by SAN Volume Controller services)	SVC management of Copy Services on SVC-specific hardware	Relies on the Copy Services Server to manage the replication function. Directly manages the failover and resynchronization of the PPRC pairs by issuing commands directly to the Copy Services Server. It is referred as *SAN Volume Controller management*.

a. Copy Services Server (CSS) represents the service running on the ESS controller or SVC node, managing the copy services on the storage.
b. ESSNI stands for Enterprise Storage Server Network Interface. The ESSNI server runs on HMC (DS6000/8000) or storage controller (ESS).
c. DSCLI represents the command-line interface used to manage the storage configuration and copy services operations.

PowerHA Enterprise Edition with different PPRC management types can coexist on the same PowerHA cluster only if the PPRC pairs are managed by one of the PPRC solutions at a time. See the latest support information for which PPRC solutions can successfully coexist on a single PowerHA cluster.

ESS/DS Metro Mirror environment

Metro Mirror (also known as synchronous PPRC) is the synchronous data replication technology that is used on ESS and DS6000/DS8000 to maintain consistent copies of data across two storage subsystems.

The data replication between storage subsystems can be performed on FC or ESCON links. PowerHA Enterprise Edition requires that inter-storage links be available to carry data in both directions at the same time. If ESCON links are used, a minimum of two ESCON links is required because each link can carry data in only one direction at a time. Have at least four links to improve throughput and to provide redundancy of the ESCON cables and adapters.

PowerHA Enterprise Edition supports ESS/DS Metro Mirror replication in the following configurations:

- Direct management PPRC (only for ESS systems)
- DSCLI PPRC management (for a cluster that uses DS6000, DS8000, or ESS Model 800)

Direct management PPRC

Direct management PPRC is the longest-running type of support through the XD PowerHA solutions and it is designed to provide basic PPRC management for ESS systems. For PowerHA Enterprise Edition only the ESS800 model is supported.

In this type of configuration, the cluster nodes directly communicate with the ESS disk controllers in both sites for PPRC operations, by using the ESS administrative network. The ESS CLI program is used by PowerHA Enterprise Edition software to manage the PPRC relationships during cluster event processing, and it must be installed on all cluster nodes. PowerHA Enterprise Edition expects the ESS CLI file sets to be installed in the `/usr/opt/ibm2105cli` directory.

You must install the version of the ESS CLI shipped with the microcode version that is applied on your storage subsystem. We used the latest version of the microcode and the ESS CLI available for the ESS800 storage model. You can download the ESS CLI from the following support website:

`ftp://ftp.software.ibm.com/storage/storwatch/esscli`

Dual active copy service feature: The microcode level vrmf 2.4.x.x supports the dual active copy services server feature. If this feature is not enabled, the copy services server that is running on the primary controller in the ESS must be manually started on the secondary controller, if the primary copy services server is unavailable.

DSCLI PPRC management

DSCLI management for HACMP/XD Metro Mirror allows more flexibility by supporting both ESS and DS storage subsystems with the same interface. It provides a simplified PPRC interface for both the ESS and DS storage hardware in the following ways:

- Using a simplified interface to the IBM TotalStorage PPRC services on ESS or DS storage systems to allow management and reporting on PPRC instances and paths
- Monitoring the status of PPRC relationships and consistency groups on the volumes that are being mirrored

  It reports any change in status, such as a volume that is moving to an offline state.

The DSCLI client software interfaces with the ESSNI server that is running on HMC (DS8000), Storage Management Console (DS6000), or ESS controller (ESS) to issue the PPRC commands. The ESSNI server, in turn, communicates the CLI commands to the ESS disk controllers and retrieves the status information. Use the version of the DSCLI shipped with the microcode version applied on your storage subsystem. For a matrix of DSCLI versions that are used with a particular version of microcode and to download the DSCLI file sets, see the DSCLI - DS8000 Command Line Interface program Download page at:

`http://www.ibm.com/support/docview.wss?&uid=ssg1S4000641`

ESS storage can also be used in DSCLI management configurations. In this case, you also must install the ESS CLI software on the cluster nodes. Cluster software for DSCLI configurations expects the ESS CLI file sets to be installed in the `/opt/ibm/ibm2105cli` non-default directory.

DSCLI PPRC resources: ESS storage resources (LSS and LUNs) are considered DSCLI PPRC resources in this type of configuration because they are managed by using the DSCLI interface and not the ESS CLI.

The XD Release Notes always have the most up-to-date information about the required software levels. They are available in the `/usr/es/sbin/cluster/release_notes_xd` file.

Multiple storage subsystems in a site

Starting with PowerHA/XD Version 5.5, you can have multiple storage units in a site that is used in a DSCLI PPRC management configuration. Keep in mind the following considerations for environments that use multiple storage subsystems in a site:

- Multiple storage subsystems can be configured in a site, but each PPRC replicated resource has only one primary and one auxiliary storage environment per site. PPRC pairs that are spread over multiple storage subsystems in a site cannot be part of the same replicated resource.
- A PPRC path should not be shared between multiple resource groups.
- If a resource group contains more than one PPRC replicated resource, each replicated resource can use a different combination of disk storage subsystems.
- When you use a resource group that contains multiple replicated resources on different disk storage pairs between sites, if any storage fails on a site, the entire resource group does not come online. Have all replicated resources in a resource group use the same disk storage.

The cluster verification scripts are no longer limited to one disk storage per site. The **`spprc_verify_config`** script checks all configured disk storage subsystems.

SAN Volume Controller environment

The IBM TotalStorage SAN Volume Controller is a virtualization appliance solution that maps virtualized volumes visible to hosts and applications to physical volumes on storage devices. SAN Volume Controller can use varied disk subsystems as backed storage devices. Volumes on the backend storage devices are mapped to the SAN Volume Controller and known in the SAN Volume Controller configuration as *managed disks* (MDisks). MDisks are used to provide logical disks, which are known as *virtual disks* (VDisks), which are presented to the final host by SAN Volume Controller.

For more information about the SAN Volume Controller architecture and the current features provided, see *Implementing the IBM System Storage SAN Volume Controller V5.1*, SG24-6423.

PowerHA Enterprise Edition with SAN Volume Controller PPRC management provides a fully automated, highly available disaster recovery management solution. It takes advantage of the ability of SAN Volume Controller to provide virtual disks that are derived from varied disk subsystems.

For PowerHA Enterprise Edition integration, the SAN Volume Controller file set version must be 4.2 or later. The SAN Volume Controller PPRC relations require an inter-cluster relationship that is defined between a SAN Volume Controller in the primary site and a second SAN Volume Controller in a remote site.

SAN Volume Controller 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 FC data frames over an Internet Protocol network between the sites.

Management of the SAN Volume Controller PPRC pairs is performed by using SSH over an Internet Protocol network. Each cluster node must have the **openssh** package installed and configured to access the SAN Volume Controllers in both sites.

PowerHA Enterprise Edition with SAN Volume Controller replication supports the following options for data replication between the SAN Volume Controller 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 starts a point-in-time copy at the primary site without impacting the I/O to the source volumes. This feature was introduced in SAN Volume Controller Version 4.1.
- Global Mirror is used for greater distances. Such factors as distance, bandwidth, and latency of the communication links between the sites can affect application performance when using synchronous mirroring. For these considerations, Global Mirror can be an alternative replication solution to SAN Volume Controller Metro Mirror. According to the latest flash, SAN Volume Controller 5.1 support is now available.

EMC SRDF environment

PowerHA Enterprise Edition 6.1 introduces support for EMC Symmetrix storage-based replication provided by Symmetrix Remote Data Facility (SRDF). The following SRDF facilities are supported for the PowerHA Enterprise Edition integration:

- Synchronous replication (SRDF/S): The primary array responds to host writes only after the target array acknowledges that it received and checked the data. Source and target devices contain identical copies.
- Asynchronous replication (SRDF/A): Point-in-time images are copied from source to target device at predefined timed cycles. Delta sets ensure that data at the remote site is dependent write consistent. In asynchronous SRDF mode, the Symmetrix array acknowledges the write operations to the source volumes immediately before it sends the update to the remote array.
- SRDF consistency groups: Application data can be spread over multiple volume pairs. Consistency groups define a relationship between volumes such that write ordering is preserved at the target (secondary) site. Consistency can be maintained in both SRDF operating modes:
  - In asynchronous mode by using multi-session consistency (MSC)
  - In synchronous mode by using ingenuity consistency assist (RDF/ECA)

Consistency: You must enable consistency when you use PowerHA Enterprise Edition with SRDF asynchronous replicated volumes. We turned on or used consistency on SRDF synchronous replicated volumes.

For more information about SRDF features, see the EMC website at:

http://www.emc.com/products/detail/software/srdf.htm

In the current integration release, the following Symmetrix models are supported:

- DMX-3
- DMX-4
- V-MAX

For more information about the Symmetrix models and firmware versions that are supported for the PowerHA Enterprise Edition integration, contact the EMC support representative. Also consult the available bulletins for RFA. For more documentation resources, see the EMC Powerlink website at:

`https://powerlink.emc.com`

You can achieve EMC SRDF storage management with the EMC-supplied Symmetrix command-line interface (SYMCLI) from the PowerHA environment. Using SYMCLI allows PowerHA software to automatically manage SRDF links and manage switching the direction of the SRDF relationships when a site failure occurs. If a primary site failure does occur, the target (secondary) site can take control of the managed resource groups that contain SRDF replicated resources from the primary site. For more information about the prerequisites and configuration of PowerHA Enterprise Edition cluster that uses EMC SRDF, see Chapter 7, “Configuring PowerHA SystemMirror Enterprise Edition with SRDF replication” on page 267.

The XD Release Notes have the most up-to-date information about the required software levels. They are available in the `/usr/es/sbin/cluster/release_notes_xd` file.

2.3.3 SAN considerations in storage replication environments

In a PowerHA Enterprise Edition that uses storage-based replication direct or SAN-switched connections can be used between the storage subsystems in the sites. Refer to the specific storage replication requirements that apply to your environment. Consult the documentation of the storage subsystem for the replication type that you use in your environment.

In most cases, switched environments are used, providing better flexibility for the configuration. When you are planning for PowerHA Enterprise Edition with storage replication, consider the zoning configuration on your SAN. For non-SVC environments consider the following types of zones:

- Host-to-storage. The HBA ports on the cluster nodes need to access the local storage ports to enable host I/O operations to the storage volumes in the same site. Configure zoning to allow multiple paths between the host and the storage subsystem. Consult the multipath software documentation for specific requirements.

  Using N-PIV: When you use N-PIV on a logical partition that is configured for Live Partition Mobility, you must include the primary and the secondary WWPNs that are associated with the N-PIV adapter in the same zoning configuration for accessing the storage devices. For more information, see “Support for IBM Virtual I/O clients” on page 34.

- Storage-to-storage. Storage ports at the primary site need to access ports on the storage at the secondary site for the data replication services. This zoning configuration does not apply for point-to-point connections such as ESCON links or FC direct connections. Use dedicated ports on the storage for the hardware replication services. In certain environments, such as EMC Symmetrix arrays, hardware-specific ports might be used for the remote copy services. For the storage model that is used in your environment, see your hardware documentation.

For a SAN Volume Controller environment, specific zoning requirements apply because of the storage virtualization role of the SAN Volume Controller. Figure 2-15 shows a typical configuration that uses SAN Volume Controller inter-cluster communication between two sites for remote copy services.

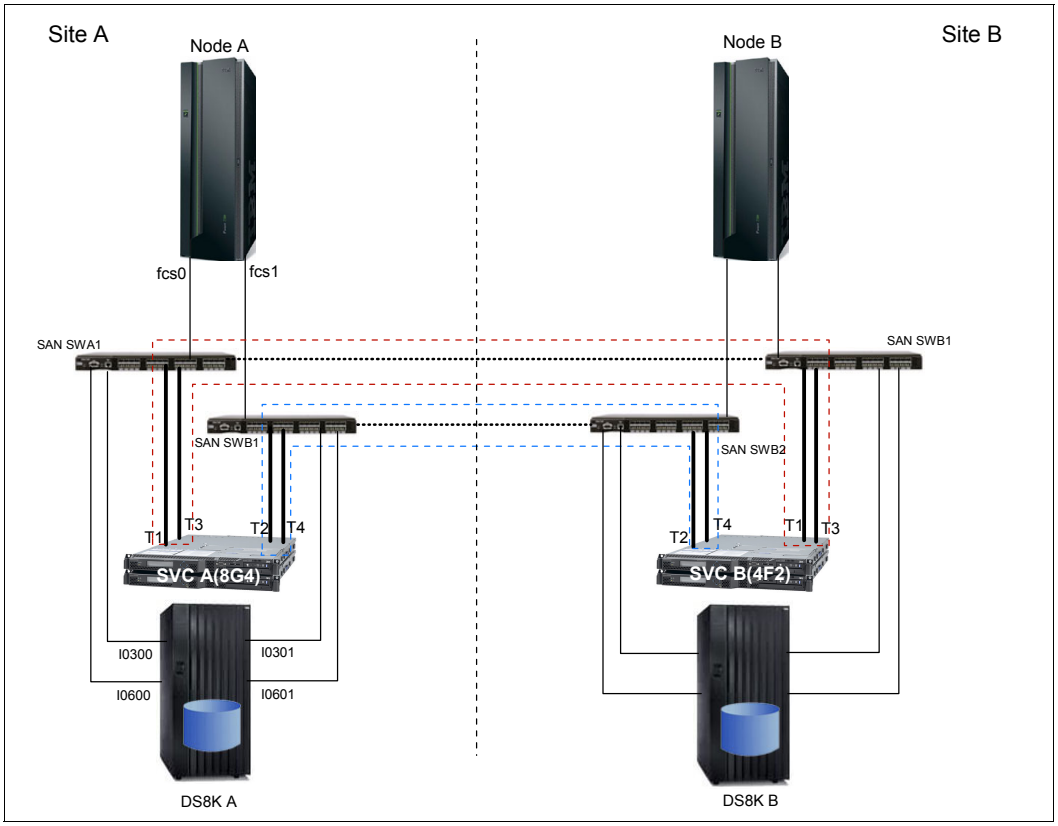


Figure 2-15 Configuration example with SAN Volume Controller inter-cluster communication

Our example has two SAN Volume Controller clusters, each with two internal nodes (one I/O group). Four FC switches are grouped in two pairs. Each pair contains a switch from site A connected to a switch at site B in the same fabric. Each SAN Volume Controller internal node uses four ports to connect to the SAN switches:

- T1 and T3 from each node are connected to the fabric 1 (SWA1-SWB1).
- T2 and T4 ports are connected to the fabric 2 (SWA2-SWB2).

The following examples show the relevant part of the **`cfgshow`** command output that is used in a Brocade SAN environment. The zone definitions use aliases that are associated with the WWPN of the HBAs on the nodes, DS8000 storage I/O ports, and SAN Volume Controller ports.

The following types of zones are defined:

- Host-to-SVC. Each host HBA is zoned with a port on each internal node of the SAN Volume Controller. Example 2-4 shows the zone configuration that is defined on each fabric for node A and SVC A (8G4). A similar configuration applies for node B accessing SVC B.

Example 2-4 Zoning configurations for node access to SAN Volume Controller (site A)

```
Fabric 1:

zone: nodeA_fcs0__svc_8g4_T1
                nodeA_fcs0; svc_8g4_n1_T1;
```

```
                svc_8g4_n2_T1

Fabric 2:

zone: nodeA_fcs1__svc_8g4_T2
                nodeA_fcs1; svc_8g4_n1_T2;
                svc_8g4_n2_T2
```

- SVC-to-storage. All ports of the SAN Volume Controller cluster nodes and storage within the same fabric are part of the same zone. Example 2-5 shows the zoning configuration between the SAN Volume Controller ports and DS8K storage at site A. A similar configuration applies to site B.

Example 2-5 Zoning configuration for SAN Volume Controller to storage access (site A)

```
Fabric 1:

zone:  svc_8g4__ds8ka_fabric1
                svc_8g4_n1_T1; svc_8g4_n1_T3;
                svc_8g4_n2_T1; svc_8g4_n2_T3;
                ds8ka_IO300; ds8ka_IO600

Fabric 2:

zone:  svc_8g4__ds8ka_fabric2
                svc_8g4_n1_T2; svc_8g4_n1_T4;
                svc_8g4_n2_T2; svc_8g4_n2_T4;
                ds8ka_IO301; ds8ka_IO601
```

- SVC-to-SVC for the inter-cluster communication relationship and remote copy services: We define all SAN Volume Controller ports from both sites, part of the same fabric in the same zone on each fabric (Example 2-6).

Example 2-6 Zoning configuration for SAN Volume Controller inter-cluster relationship

```
Fabric 1:

zone:  svc_8g4__svc_4f2_fabric1
                svc_8g4_n1_T1; svc_8g4_n1_T3;
                svc_8g4_n2_T1; svc_8g4_n2_T3;
                svc_4f2_n1_T1; svc_4f2_n1_T3;
                svc_4f2_n2_T1; svc_4f2_n2_T3

Fabric 2:

zone:  svc_8g4__svc_4f2_fabric2
                svc_8g4_n1_T2; svc_8g4_n1_T4;
                svc_8g4_n2_T2; svc_8g4_n2_T4;
                svc_4f2_n1_T2; svc_4f2_n1_T4;
                svc_4f2_n2_T4; svc_4f2_n2_T4
```

2.3.4 PowerHA Enterprise Edition with GLVM

Geographical Logical Volume Mirroring is another feature of the PowerHA Enterprise Edition that uses host-based data replication for disaster recovery over TCP/IP communication

networks. GLVM can use synchronous or asynchronous modes to replicate the data between the local and the remote copy of the data.

Storage volumes in a GLVM environment might be allocated to a single host or shared between multiple cluster nodes within the same site. The data replication layer is integrated with AIX LVM mirroring functions to provide flexible and simplified management. For more information about the GLVM configuration, see Chapter 8, “Configuring PowerHA SystemMirror Enterprise Edition with Geographic Logical Volume Manager” on page 339.

2.4 PowerVM virtualization considerations

The PowerVM virtualization features on IBM Power servers are proven to be effective at maximizing the usage of server resources. The implementation of such features as Micro-Partitioning, dynamic LPAR, virtual I/O servers, N-Port ID virtualization (NPIV), and even more advanced functions with Live Partition Mobility are becoming common within PowerHA clustered environments.

Figure 2-16 shows a virtualized Live Partition Mobility capable environment for the primary site in a cluster. The vfc adapter definitions represent virtual fiber adapters that come from an NPIV-capable HBA that is assigned to each VIOS.

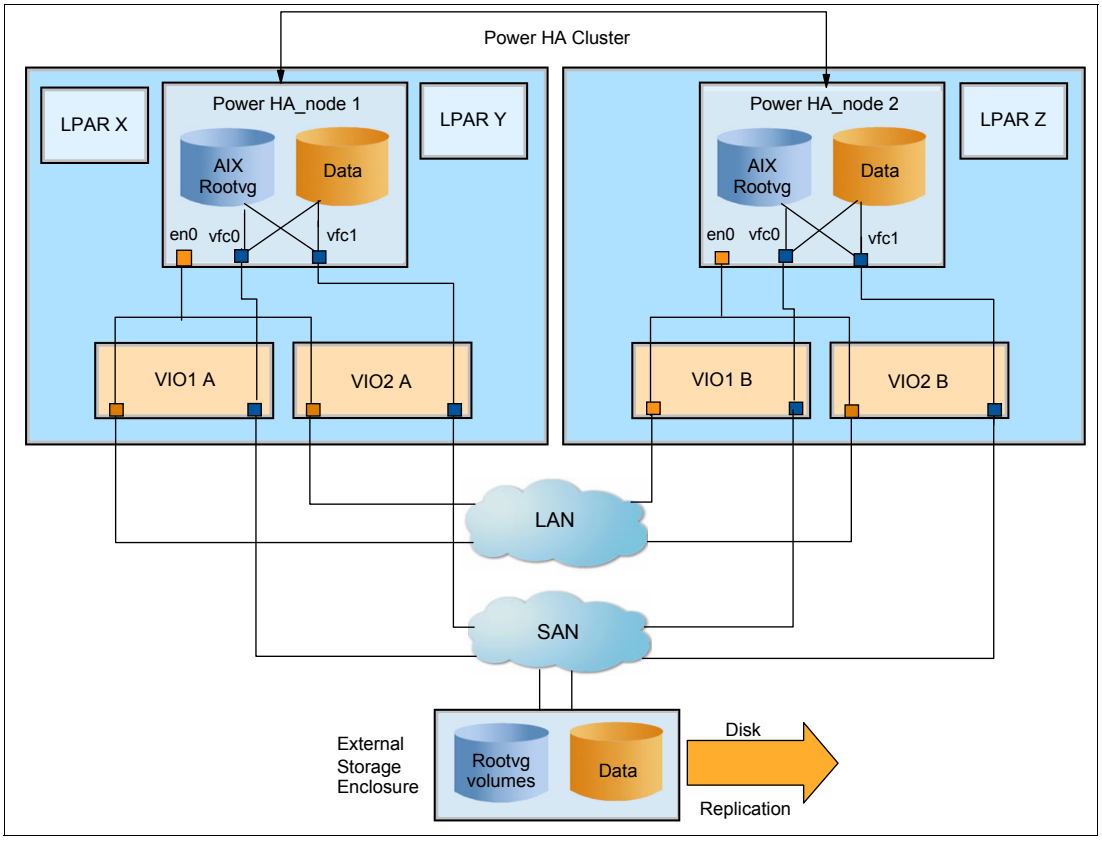


Figure 2-16 Live Partition Mobility capable virtualized environment diagram

Designing virtualized cluster solutions with best practices in mind ensures that the environments have the same or better availability as ones that use dedicated adapters. This approach includes the use of dual VIO servers, Shared Ethernet Adapter (SEA) failover, and

link aggregation for IP connectivity if applicable. Redundant SAN connections and multipathing also ensure maximum availability for I/O traffic.

Environments that use virtualization functions can create test environments much faster. If the bandwidth on the existing physical adapters is not already maxed out, clients can easily provide new virtual resources for new test LPARs. The test LPARs can be defined with minimal processor resources and a low priority to share processor time within the shared processor pool. When you implement such an environment, no new cable drops are required. Only the LUNs for the new LPAR boot image and the virtual resources are required. Sample data can be brought into the test environment from a backup or flash copy and be used in parallel to the production environment. The other advantage is that if your production environment is already virtualizing you now have a matching environment to run tests against.

The following sections review the following areas:

- “Network virtualization considerations” on page 77
- “Storage considerations in a virtualized environment” on page 80
- “Virtualization performance considerations” on page 84
- “Live Partition Mobility and PowerHA” on page 84
- “Virtualization and migrating to new hardware” on page 87

For more information and considerations about the IBM PowerVM Virtual I/O Server, see *Power Systems: Virtual I/O Server* at:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/iphb1/iphb1.pdf

2.4.1 Network virtualization considerations

A clustered environment can benefit from virtualization in various ways. The first benefit is a simpler topology. If the client LPARs are backed by redundant VIOS, which are configured with SEA failover, there is no need to mimic traditional topology configurations with multiple Ethernet adapters. In a virtualized environment, a single virtual Ethernet adapter would suffice on each client partition. Similar to when you use an EtherChannel, a physical NIC failure results in the loss of a link and causes the traffic to automatically reroute through the backup adapter. The failure is not apparent to the cluster and keeps it from having to run a local adapter swap.

Figure 2-17 shows an example of how the environment backing the PowerHA cluster nodes can look in a virtualized environment.

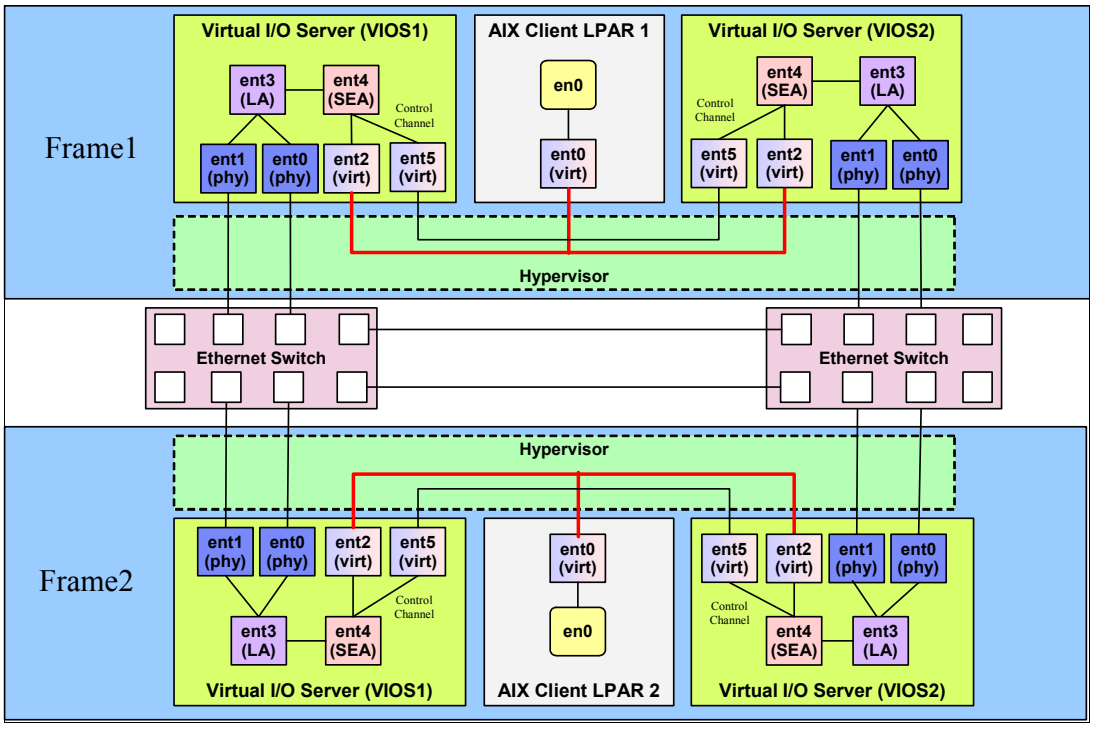


Figure 2-17 PowerHA virtual Ethernet configuration - SEA failover

The following rules apply for Ethernet configurations that use the Standard and Enterprise Editions of the IBM PowerHA solution:

- IPAT by way of aliasing must be used.
- All virtual Ethernet interfaces that are defined to HACMP should be treated as single-adapter networks and use the ping_client_list attribute to monitor and detect failure of the network interfaces (`/usr/es/sbin/cluster/netmon.cf` file).
- The Shared Ethernet Adapter bridge is active on one VIOS until failover. The VIOS with the lowest priority value is used as the primary link bridge.

Link aggregation of EtherChannel configurations on VIOS: Link aggregation or EtherChannel configurations on the VIOS have been implemented for extra bandwidth.

In a PowerHA environment that use Live Partition Mobility between two frames, there are times when the primary and failover LPARs might end up on the same side. In this scenario, if IP connectivity to the outside world is lost, heartbeats between the nodes can continue to pass through the hypervisor unless entries in the `/usr/es/sbin/cluster/netmon.cf` file with the appropriate formatting were implemented.

Figure 2-18 shows how this situation can occur if maintenance was being performed on the primary machine. In this example, both the primary and secondary cluster nodes are temporarily hosted on the same physical machine. While in this state, the physical target server introduces a temporary single point of failure.

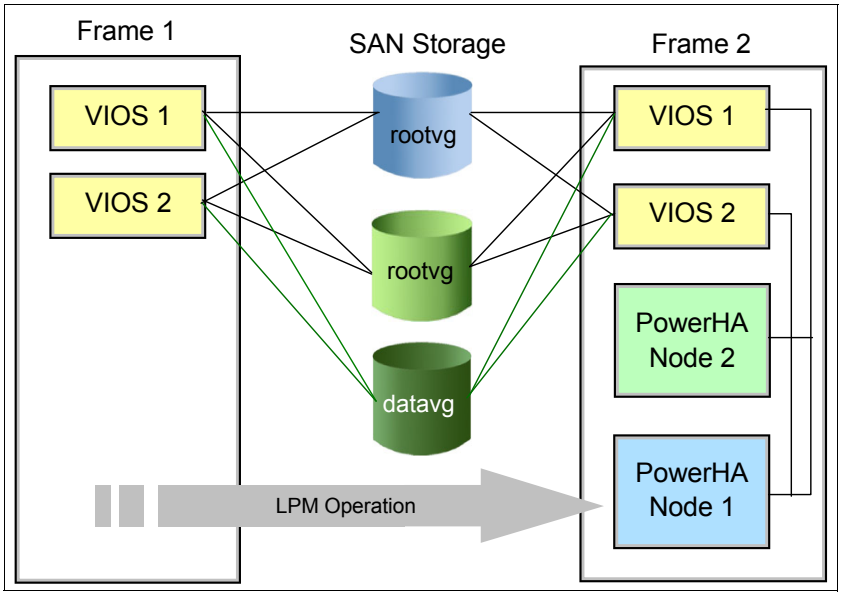


Figure 2-18 Live Partition Mobility operation - both cluster nodes on same frame

The new format for entries within the `netmon.cf` file assists the netmon logic probe IPs and interfaces outside of the machine to establish an accurate status.

Format: For single-adapter virtualized networks in a PowerHA cluster, verify that you are using the new `/usr/es/sbin/cluster/netmon.cf` format, as explained in APAR IZ01332: NEW NETMON FUNCTIONALITY TO SUPPORT HACMP ON VIO at:

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

See the following sample format:

```
10.12.4.11
10.12.4.13
!REQD en2 100.12.7.9
!REQD en2 100.12.7.10
!REQD host1.ibm 100.12.7.9
!REQD host1.ibm host4.ibm
!REQD 100.12.7.20 100.12.7.10
```

In a multisite cluster, the use of virtual interfaces can be advantageous. Because the two sites cannot be on the same network segment, the use of a single interface on the clients can help reduce the number of IPs required to fulfill the PowerHA subnetting requirements.

IP requirement: Another requirement for IPs is that they must be pingable from the Boot IP. For more information, see *Choosing IP addresses for the netmon.cf file* in the AIX 6/1 Information Center at:

`http://publib.boulder.ibm.com/infocenter/aix/v6r1/index.jsp?topic=/com.ibm.aix.hacmp.plangd/ha_plan_netmon_cf_file.htm`

With the use of a single static address at each site, the cluster can establish a network heartbeat ring while still providing adapter redundancy and load balancing.

If only a single virtual interface was defined on the client, load balancing can be achieved by aggregating multiple interfaces for the SEA defined on the VIOS. The simple choice is to configure the link aggregate in a network interface backup (NIB) fashion and connect each adapter to a different switch. This approach provides switch redundancy, but no load balancing. To set the round-robin policy and load balance within that link aggregate, you must connect the active links to the same switch. Therefore, when you evaluate this method, consider a minimum of three adapters for the link aggregate. The first two active links connect to the first switch, and the third backup adapter connects to a second switch. Load balancing occurs between the active links, and the backup adapter is not used until either the first switch fails or all of the active adapters in the link aggregate go offline.

Tip: There are no limitations in PowerHA System Mirror regarding whether the network interfaces are virtual or dedicated at each site. Therefore, one site can be use dedicated adapters, and the other site can use virtual ones.

2.4.2 Storage considerations in a virtualized environment

Virtualizing the network adapters is only one of the ways that virtualization introduces new complexities. Storage virtualization introduces its own set of considerations.

Initially, the VIOS only provided the ability to virtualize the storage through a VSCSI mechanism, which required the LUNs to be mapped to the VIO server first and then remapped to the clients. To avoid locking on the LUNs mapped to the redundant VIO servers, it was required to set the no_reserve policy for each of the shared disks. These VSCSI mappings provided an abstracted view of the LUNs at the client level. Therefore, a LUN that might show up as a DS8000 disk in an **lsdev** output might now show up as a VSCSI disk. On the client, only the MPIO driver can be used for the management of the different paths, and it introduced the requirement for the use of enhanced concurrent mode volume groups.

The new NPIV capable Fibre Channel Adapters (feature code 5735) on IBM POWER6® systems can define virtual fiber adapters to each of the VIO clients. These 2-port, 8-GB HBAs can create 64 new WWPNs per port.

Reference: For more information about NPIV configurations, see 2.9, “N_Port ID virtualization,” in *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

Figure 2-19 illustrates a logical view of the NPIV function

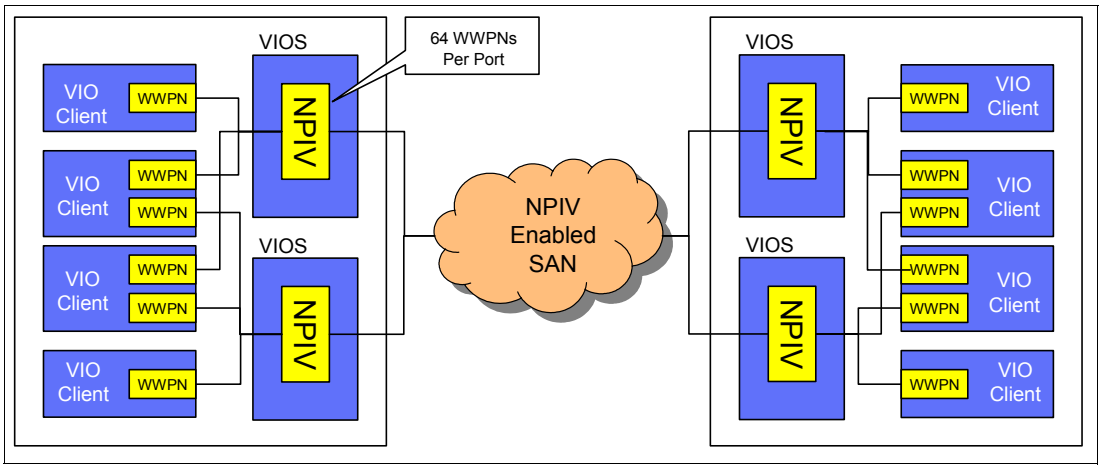


Figure 2-19 N-Port ID virtualization logical diagram

The physical adapter is allocated to the VIO server, but in contrast to the VSCSI model. After the virtual fiber adapters are defined to the clients and the new WWPNs are noted in the zoning and mapping definitions, the LUNs are assigned to the clients. They are assigned the same way that they use the dedicated adapters. Therefore, the LUN mappings bypass the VIOS and allow the physical volumes to be detected on the client because they are in a dedicated environment.

Figure 2-20 shows a sample virtualization configuration that contrasts VSCSI volumes with virtual fiber volumes. In this example, the first three volumes, which include the rootvg, are attached by using VSCSI server (vhost) and client definitions (vscsi). The next two volumes (hdisk3 and hdisk4) in the npiv_vg volume group are presented through virtual fiber adapters. The disk characteristics for these two LUNs show up on the host in the same fashion as when using dedicated HBAs.

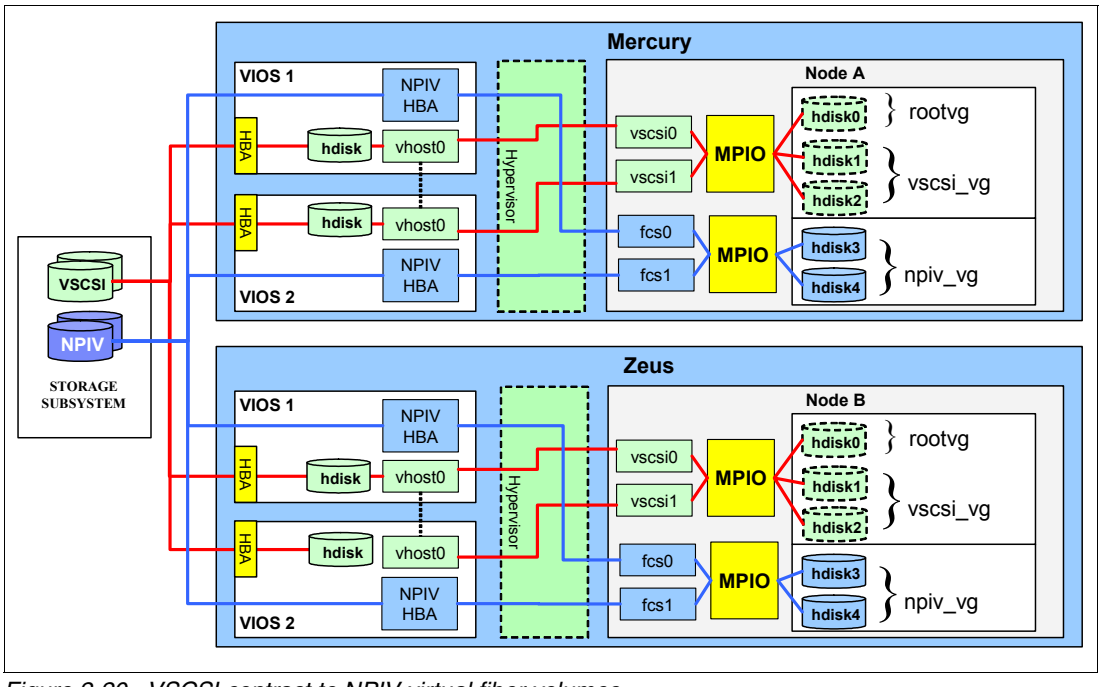


Figure 2-20 VSCSI contrast to NPIV virtual fiber volumes

PowerHA Enterprise Edition is supported on IBM Virtual I/O clients for SAN Volume Controller, DS/ESS PPRC, and GLVM environments. The VIOS allows a machine to be divided into LPARs, with each LPAR running a separate OS image, allowing the sharing of physical resources between the LPARs including virtual SCSI and virtual Ethernet. The VIOS can also support the NPIV feature, which provides a virtual Fiber Channel adapter to the client partition, allowing transparent access to external storage subsystems. PowerHA Enterprise Edition 6.1supports both VSCSI and NPIV storage configurations with VIOS 2.1.

Figure 2-21 shows an example of using Virtual SCSI for a SAN Volume Controller replication environment.

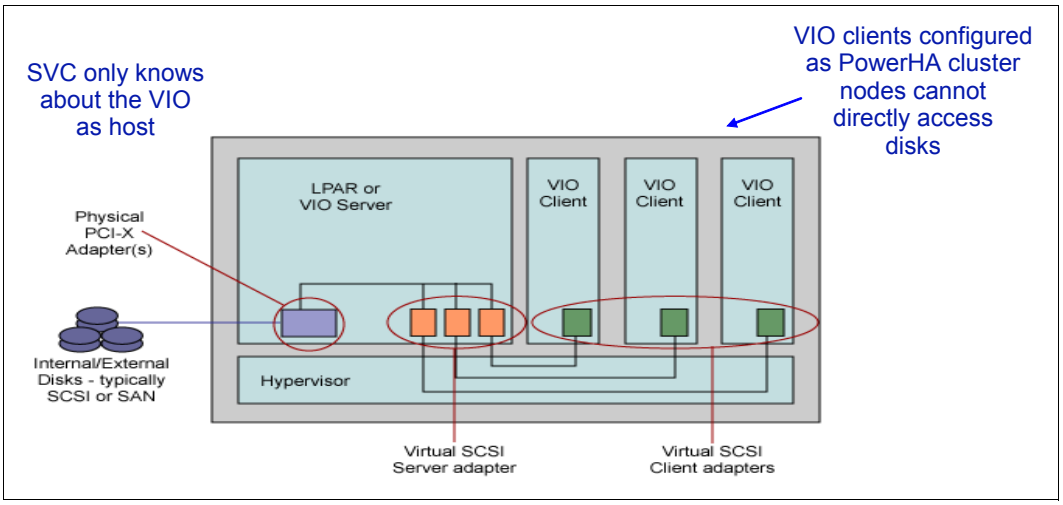


Figure 2-21 Using virtual SCSI in a SAN Volume Controller PPRC environment

The VIOS has a few disks that can be SCSI or Fibre Channel SAN disks. The VIO clients use the VIO client device driver just as they would with a regular local device disk to communicate with the matching server VIO device driver. Then, the VIOS does the disk transfers on behalf of the VIO client. Because the SAN Volume Controller devices are not directly attached to the VIO clients, normal query commands, such as **`lscfg`**, **`lsvpcfg`**, and **`datapath query device`**, cannot be used to extract the necessary SAN Volume Controller vdisk information.

The PowerHA Enterprise Edition solution has the support for using the disk resources that are defined on the VIO clients in the cluster environment, for example:

- It allows disks that are defined as HACMP resources to be traced back to the source physical disk on the SAN Volume Controller.
- When you define the SAN Volume Controller PPRC resources, you use VDisks from the SAN Volume Controller configuration.
- Cluster events run SAN Volume Controller commands against VDisks. It maps hdisk on VIOS client to the VDisk on SAN Volume Controller.

  The disk definition in VIOS to the client partition is one-to-one between the VDisk that is assigned to the VIOS and the hdisk that is defined in the VIO client. Configurations that use logical volumes or file-backed devices that are mapped to the VIO clients are not allowed.

No special configuration steps are required for defining the disks to the client partitions. Use the following steps to configure the disks on the client partition. We assume that the SAN Volume Controller configuration is prepared and that the mdisks are already defined:

1. On the SAN Volume Controller clusters:
   a. Identify the managed disk's MDisks:

      ```
      svcinfo lsmdisk
      ```

   b. Identify or create the managed disk group's MDisk group by using one of the following commands:

      ```
      svcinfo lsmdiskgrp
      svctask mkmdiskgrp
      ```

   c. Identify or create the virtual disks by using one of the following commands:

      ```
      svcinfo lsvdisk
      svctask mkvdisk
      ```

   d. Map the VDisks to the VIO servers as hosts:

      ```
      svctask mkvdiskhostmap
      ```

2. On the VIOS:
   a. Access to the regular AIX command-line interface on the VIO server:

      ```
      oem_setup_env
      ```

   b. Run **cfgmgr** or, to speed up the process, run with the flag on the vio#:

      ```
      # cfgmgr -vl vio0
      ```

   c. Identify the hdisks/vpaths mapped to the SAN Volume Controller VDisks on the servers:

      ```
      odmget -q "id=unique_id" CuAt
      ```

   d. Select the disk to export by running **lsdev** to show the virtual SCSI server adapters that can be used for mapping with a physical disk.
   e. Run the **mkvdev** command by using the appropriate hdisk numbers to create the virtual target device. (This command maps the LUNs to the virtual I/O clients.)

      ```
      $ mkvdev -vdev hdisk# -vadapter vhost# -dev vhdisk#
      ```

   f. To bring the vtd and vhosts from the available state to the defined state, run the following commands:

      ```
      # rmdev -l  <vtd devices>
      # rmdev -l  vhost#
      ```

3. On the VIO clients

   After the **mkvdev** command runs successfully on the VIO server, the LUNs are exported to the VIO clients.

   a. Identify the LUN information on the client:

      ```
      lsdev -Cc disk
      ```

   b. Identify the SAN Volume Controller VDisk - LUN mapping on the VIO clients:

      ```
      cl_vpath_to_vdisk
      ```

2.4.3 Virtualization performance considerations

Apart from limits in distance and bandwidth, for virtualization, it is also important to know the maximum transfer size for the target storage device (LUN). It can be viewed as a route-specific MTU size between the host and target device if compared to TCP/IP, keeping in mind that the maximum SCSI payload in an FC frame is 2112 bytes.

With host-based synchronization or mirroring, all cluster nodes are LPARs and have access to shared storage. In this case, the I/O and memory load on the managed system is limited because the VIOS design uses DMA and RDMA from the LPAR to access virtual disks that are mapped over a virtual SCSI adapter over the VIOS to the LPAR.

Theoretical values for the number of virtual SCSI adapters per managed system are 64 K, and the number of mapped virtual disks per adapter is approximately 500. With a physical SCSI adapter, there is a limit of 16 devices per adapter.

One view about the number of mapped virtual disks per adapter is that a similar reasoning can be applied for virtual SCSI adapters and mapped target devices as for physical SCSI. By looking at the storage connections from a storage system perspective, the intention is to limit the risk of one malfunctioning storage system or link to cause interference with other devices and connections.

With this consideration in mind, it can be proposed that disks from different storage subsystems should not be mixed over the same virtual SCSI adapter, especially if the maximum transfer size differs between the separate storage devices. In a SAN Volume Controller environment, two virtual SCSI adapters per LPAR result, one to the primary VIOS and one to the secondary VIOS. In a scenario where the LPAR has a virtual target disk from VIOS DASD, and from DS4700 and DS8300, the LPAR can have six virtual SCSI adapters, three to each VIOS in the VIOS cluster.

Another consideration for VIOS with considerable disk I/O and mapped target devices is collocation on the same VIOS that is used for bridging the hypervisor virtual Ethernet to the physical network. Consider whether it is more reliable if the secondary VIOSs are used for disk I/O while the primary VIOSs are used as a network bridge.

If a primary VIOS failure occurs because of resource starvation that is caused by excessive network load, the disk I/O is not affected immediately. Also, consider asymmetric adapter allocation between the primary and secondary VIOS for increased throughput through the primary or secondary VIOS during normal operation, allowing for lower throughput during a service window for servicing either VIOS.

2.4.4 Live Partition Mobility and PowerHA

With Power Systems environments, the natural progression is to use the advanced PowerVM functions features such as Live Partition Mobility. Live Partition Mobility allows for the movement of a running LPAR from one system to another without disruption to the active applications.

For list of the PowerHA requirements for Live Partition Mobility, see the support flash at:

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10640

On AIX, the combination of PowerHA clustering and Live Partition Mobility is powerful because it can dynamically migrate a running LPAR between physical systems and provide protection against an unforeseen outage. In a virtualized implementation that is designed with disaster recovery, configure the primary site with Live Partition Mobility before replicating to a remote site.

NPIV/Live Partition Mobility: The SAN Volume Controller Global Mirror and DS Metro Mirror scenarios in this book were built by using NPIV/Live Partition Mobility capable configurations at the primary site that replicated to a non-virtualized remote site.

From a positioning standpoint, Live Partition Mobility is designed for planned maintenance. For the movement to work, all the resources must be virtualized. The minimum HMC requirement for Live Partition Mobility was V7 R3.2, which required all of the systems to be visible by the same HMC. Inter-HMC Live Partition Mobility operations became available in the V7 R3.4 HMC update where that requirement was lifted and each system can now be managed by its own HMC. If ssh authentication is enabled between the HMCs, then during the Live Partition Mobility selections, there is an option to specify the target HMC and server to which to migrate. When the move is initialized, a validation process must complete and then the server's memory is cached and moved over when fully synchronized. The cutover results in approximately a 2-second pause in the processing.

In contrast, PowerHA is designed for both planned and unplanned outages. The cluster software can move the application by initiating a graceful stop of the resources and by invoking the start scripts at the failover target node. If an unexpected outage occurs, the cluster software detects the failure, initiates a takeover, and manages the activation of resources automatically.

One of the assumptions for Live Partition Mobility to function is that all nodes start from the SAN. You can begin to see where NPIV function is really beneficial in DR implementations. Such an environment has the same look and feel as an environment that uses dedicated adapters, and provides a view of the cluster LUNs that is not affected by the VSCSI limitations.

Virtualization and SAN booting

In the rapid movement to take advantage of the many virtualization features, clients often overlook the risks that are involved with making the move. Before the virtualization movement, it was common to start the cluster nodes by using internal drives and sharing data volumes that come from SAN-attached storage. It was also common for a copy of rootvg to come from VSCSI internal LUNs on one VIOS and a second mirrored copy from a second LUN from a backup VIOS.

However, for Live Partition Mobility to migrate an LPAR to another frame, the requirement is that it must start from the SAN. Clients who used to perform failure testing by pulling out cables found that the cluster software does not behave as expected whenever complete access to the rootvg is lost. Topology services heartbeating is brought into cache during the cluster startup. Whenever access to the rootvg is lost, the ability for the host to react is neutralized. Because no runtime files are available, there is no way to trigger a failover. A failover still occurs, but the times might vary based on the activity that occurs on the node and how long before the cache allows the heartbeating to continue. The use of non-IP disk heartbeat networks does not provide any benefits of identifying the loss of the rootvg volumes.

Result: What happens is that AIX is running in memory and the kernel and applications work as long as the root file system is not accessed.

This feature is an AIX feature that, after the loss of the root volume, does not cause immediate kernel panic. The loss of the root volume is difficult to encounter when you use internal drives or direct-attached storage.

To better understand this concept, we must first identify the various components that back the environment. Figure 2-22 shows a common Live Partition Mobility capable clustered environment.

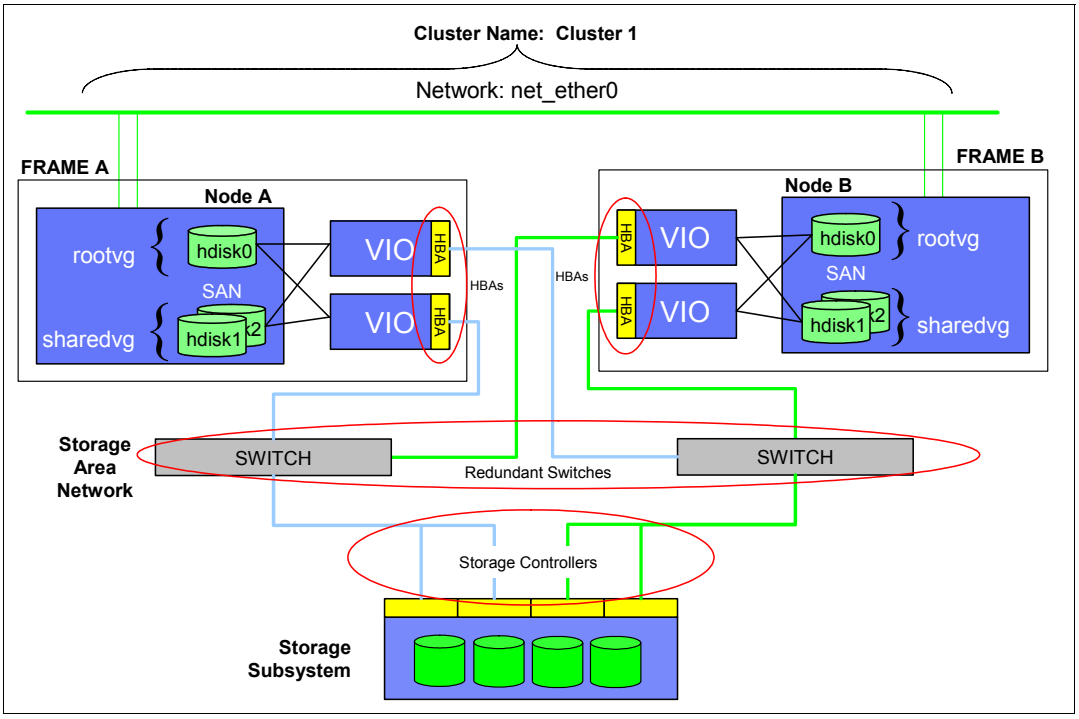


Figure 2-22 Live Partition Mobility with SAN booting points of failure

Multitude of clients: Even though Figure 2-22 only shows one client LPAR behind the VIO servers, typical environments back a multitude of clients.

Figure 2-22 highlights how many layers of redundancy are overlooked by simply pulling all of the fiber interconnects from one side. Working our way from the storage subsystem up, we can make the following assumptions:

- LUNs presented from SAN-attached enclosures are typically raided on the backend and protected from physical disk drive failures.
- The use of multiple storage controller connections also provides protection against internal component failures.
- The use of a redundant fabric provides protection against the loss of a SAN switch.
- The use of dual VIOS provides multipathing and protects against a VIO or physical HBA failure.

Virtualization and Live Partition Mobility can provide significant benefits in most environments. This configuration helps you understand the significance of implementing a redundant infrastructure and certain new complexities about component failure testing in a virtualized environment.

2.4.5 Virtualization and migrating to new hardware

Virtualization can also have a major impact on the amount of time that is required to migrate to new machines. Consider a traditional non-virtualized environment where one of the existing servers is scheduled to be replaced by a newer machine. In this scenario, you have a couple of choices for performing the system migration:

- Back up and restore the cluster node.
- Start the new machine and introduce a new node into the cluster, fail over the resources to it, and then, remove the original node from the cluster.

There is a common misunderstanding of the statement that LPAR cloning is not a supported backup and restore procedure for a PowerHA cluster node. In general terms, this means that the cluster definitions in the ODM are duplicated onto another server during the cloning. Therefore, we can still clone the LPAR, but to avoid duplication we need to ensure that the cluster definitions are deleted. You invoke the cluster menus and remove the cluster definition, which removes all other associated definitions with it. For example, enter the **`smitty hacmp`** command, select **Extended Configuration** → **Extended Topology Configuration** → **Configure an HACMP Cluster** → **Remove an HACMP Cluster**, and press Enter.

The remove operation removes only the ODM stanzas from the node on which the operation ran. Although redefining a cluster and synchronizing it overwrites the ODMs on all other cluster nodes, you might consider deleting the cluster definitions on all members individually. This can also be done with a NIM post-install script and the PowerHA **`/usr/es/sbin/cluster/utilities/clrmclstr`** command.

Either of the two methods is disruptive and requires careful planning to ensure that all of the cluster components are properly shared and accounted for. In a virtualized environment, the movement is simpler because a Live Partition Mobility operation can non-disruptively migrate the LPAR onto the new machine. This method provides a significant advantage over a dedicated environment because it does not require the installation of AIX, PowerHA, and the base application images on the new LPAR. It also avoids the effort that is required to properly share the volume group and file system information and the resetting of the various tunables and configuration files that are associated with the cluster. In addition, in a multisite environment, this approach is appealing because zero disruption happened on the primary site during a Live Partition Mobility move. The replication can continue running between the sites.

2.5 Server considerations

Many hardware and software combinations can be used in the implementation of a PowerHA cluster. This section reviews the fileset level combinations and provides a point of reference when trying to decipher the requirements for your environment.

2.5.1 Software considerations

AIX clustering with the PowerHA software has a multitude of variables that affect its interoperability between all of the solution components. The current PowerHA Enterprise Edition 6.1 release has the following minimum AIX software requirements:

- AIX 5.3 TL 9, with RSCT 2.4.12.0 or later
- AIX 6.1 TL 2, SP1 with APAR IZ31208 with RSCT 2.5.4.0 or later

Test scenarios: All of the test scenarios in this book were performed using PowerHA Enterprise Edition 6.1 SP1 on AIX 6.1 TL2 - SP3, with the exception of the GLVM migration section.

The Enterprise Edition requires the installation and acceptance of license agreements for both the Standard Edition `cluster.license` file set and the Enterprise Edition `cluster.xd.license` file set, as shown in Table 2-6, for the remainder of the file sets to install.

Table 2-6 PowerHA Enterprise Edition - required file set

Required package	File sets to install
Enterprise Edition License	`cluster.xd.license`

The base file sets in the Standard Edition are required to install the Enterprise Edition file sets. The Enterprise package levels must match those of the base runtime level (cluster.es.server.rte). Table 2-7 displays the itemized list of file sets for each of the integrated offerings.

Table 2-7 PowerHA Enterprise Edition - integrated offering solution file sets

Replication management type	File sets to install

Direct management	`cluster.es.pprc.rte`
	`cluster.es.pprc.cmds`
	`cluster.msg.en_US.pprc`

DSCLI	`cluster.es.spprc.cmds`
	`cluster.es.spprc.rte`
	`cluster.es.pprc.rte`
	`cluster.es.pprc.cmds`
	`cluster.msg.en_US.pprc`

SAN Volume Controller	`cluster.es.svcpprc.cmds`
	`cluster.es.svcpprc.rte`
	`cluster.msg.en_US.svcpprc`

Geographic	`glvm.rpv.client`
	`glvm.rpv.msg.en_US`
	`glvm.rpv.server`
	`glvm.rpv.util`
	`glvm.rpv.man.en_US`
	`glvm.rpv.msg.en_US`
	`cluster.msg.en_US.glvm`

EMC SRDF	`cluster.es.sr.cmds`
	`cluster.es.sr.rte`
	`cluster.msg.en_US.sr`

Installation notes

If you are using IPv6 networking:

- Contact IBM support to have the latest version of the HMC microcode (bundle 64.30.78.0 or later) installed.
- For customers who are using SNMP traps notification (for asynchronous alerts of changes in the PPRC and consistency group status), a known problem exists with the handling of traps in a network that use only IPv6. IBM intends to remove this restriction in a future service pack. See APAR IZ60486.

2.5.2 AIX level requirements

The integrated replication solutions can be implemented by using both AIX 5.3 and 6.1 levels. Table 2-8 shows the minimum OS level requirements for each integrated offering and the current virtualization support statement. For the most up-to-date information, see the following website:

http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105440

Table 2-8 PowerHA Enterprise Edition support cross-reference

Replication options	PowerHA min. version	AIX 5.3	AIX 6.1	VIOS/NPIV
GLVM Synchronous	5.2 + IY66555	Yes - ML2	Yes	Yes
GLVM Asynchronous	5.5 SP1	No	Yes - TL2 - SP3	Yes
SVC Metro Mirroring	5.2	Yes	Yes	Yes - HA 5.4.1 + IZ13774 VIOS 1.5 HA 5.5/6.1 - VIOS 2.1
SVC Global Mirror	5.5	Yes - TL9 RSCT 2.4.10.0	Yes - TL2 RSCT 2.5.2.0	Yes - VIOS 2.1 No HA 5.4.1 support
ESS Metro Mirroring	5.1	Yes	Yes	Yes with HA v6.1
DS6K/DS8K Metro Mirroring	5.2 + IY73937 5.3 + IY74112	Yes - ML1	Yes	Yes with HA V6.1
EMC SRDF ► DMX - 3 ► DMX - 4 ► V-MAX	6.1	Yes - TL9 RSCT 2.4.12.0	Yes - TL2 - SP1 RSCT 2.5.4.0	No

2.5.3 Multipath driver requirements

SAN attachment to IBM storage units by using redundant fiber connections requires the use of a multi-pathing driver on the host. Use careful consideration when you select the appropriate driver for your environment.

Tip: Most current implementations use MPIO with the corresponding path control module (PCM) to simplify the management of the available paths.

On AIX, clients can choose one of the following options:

- Multipath I/O (MPIO) with the corresponding PCM
- Subsystem Device Driver (SDD)
- Redundant Disk Array Controller (RDAC) for the IBM System Storage DS4000®

Important: The SDD driver and SDDPCM module cannot coexist on same server.

Attachment to EMC Symmetrix enclosures is the exception, requiring the use of the Powerpath multipath driver for the integrated replication support.

Multipath I/O requirements

The IBM Multipath Subsystem Device Driver Path Control Module (SDDPCM) requires **devices.sdd.*XX*.rte** Version 2.2.0.0 or later, where *XX* corresponds to the associated level of AIX. For AIX 5.3, the version is **devices.sddpcm.53.rte**.

The host attachment for SDDPCM adds 2105/2145/1750/2107 device information to allow AIX to properly configure 2105/2145/1750/2107 hdisks. The 2105/2145/1750/2107 device information allows AIX to perform the following tasks:

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

AIX has the following supported AIX levels:

- AIX 610
- AIX 530 TL2 and later

Table 2-9 outlines the AIX SDDPCM levels for supported device types and operating system levels.

Table 2-9 SDDPCM supported device types and operating system (OS) levels

AIX Operating System	ESS	DS8000	DS6000	SVC
6.1	2.2.0.0 (2, 3)	2.2.0.4 (2)	2.2.0.0 (2)	2.2.0.3 (2, 3)
5.3	2.2.0.0 (1, 2)	2.2.0.4 (1, 2)	2.2.0.0 (1, 2)	2.2.0.3 (2, 3)

1. Requires AIX 5.3 TL04 (with all the latest PTFs) or later or AIX 5.3 TL6 SP1 (5300-06-01) or with VIOS V1.4 or later.
2. Persistent Reservation with PowerHA is not supported. Shared volume groups that are managed by HACMP and accessed through SDDPCM must be in enhanced concurrent mode.
3. If you are running the SAN Volume Controller release that includes APAR IC55826 (SAN Volume Controller V4.2.1.6 and later) or the SAN Volume Controller release before V4.2.1.6 with an interim fix on LONG BUSY, the following AIX APARs and AIX interim fix are required for your AIX TL:
   AIX53 TL06: APAR IZ06622 IZ20198 IZ28285 (IZ28285.080728.epkg.Z)
   AIX53 TL07: APAR IZ06490 IZ19199 IZ26561 (IZ26561.080728.epkg.Z)
   AIX53 TL08: APAR IZ07063 IZ20199 IZ26655 (IZ26655.080728.epkg.Z)
   AIX61 TL00: APAR IZ09534 IZ20201 IZ26657 (IZ26657.080728.epkg.Z)
   AIX61 TL01: APAR IZ06905 IZ20202 IZ26658 (IZ26658.080728.epkg.Z)
   The third column of APARs might not be published yet. However, they are available in the interim fix format. You can download these interim fixes from the AIX interim fix website, which is an anonymous FTP site.

Before you install SDDPCM, you must uninstall the **ibm2105.rte** (Version 32.6.100.*XX*) and **devices.fcp.disk.ibm.rte** (Version 1.0.0.XX) SDD host attachment packages along with the SDD package. The SDD driver and SDDPCM module cannot coexist on one server.

You must install the ESS, DS8000, DS6000, or SAN Volume Controller host attachment for SDDPCM:

```
devices.fcp.disk.ibm.mpio.rte (version 1.0.0.12)
```

```
This version of SDDPCM package requires the host to run with least AIX53 TL04 with
the latest PTFs. SVC attachment can be achieved with AIX 5.3 TL03 with the
following APARs installed:

AIX53 TL03 APAR: IY79165, IY81545
```

Important: ESS SCSI devices are not supported by SDDPCM, nor is an ESS host attachment for SCSI MPIO.

You might have installed an earlier version (earlier than v33.6.100.9) of SDDPCM host attachment and are going to update to a new SDDPCM host attachment Version 1.0.0.0 or later. In this case, you must remove all SDDPCM mpio devices before you start to update this package. The reason is that the ESS device type changed, starting from version 33.6.100.9.

Subsystem Device Driver requirements

The following SDD software and microcode levels are required:

- IBM 2105 SDD **`ibmSdd_510nchacmp.rte`** 1.3.3.6 or later
- APAR IY83601 for AIX V5.3 if you are upgrading from AIX V5.3
- IBM SDD FCP host attachment script, **`devices.fcp.disk.ibm.rte`**, version 1.0.0.12 or later

  File set: During installation, it might not be clear whether this file set is necessary. However, this file set is critical to ensure that failover works correctly. Without this file set installed, issues might occur with disk reserves not being managed correctly.

- IBM SDD: **`devices.sdd.`*`XX`*`.rte`** Version 1.6.3.0 or later, where *XX* corresponds to the associated level of AIX, for example **`devices.sdd.53.rte`** for AIX 5.3.

Table 2-10 lists the latest AIX SDD levels for supported device types and supported operating system levels.

Table 2-10 SDD supported device types and OS levels

AIX operating system	ESS	DS8000	DS6000	SVC
6.1	1.7.1.0	1.7.2.0	1.7.0.0	1.7.1.0
5.3	1.7.1.0 *	1.7.2.0 *	1.7.0.0 *	1.7.1.0 [a]

a. Requires AIX 5.3 TL04 or later

For the most current information about the latest release of SDD, see the Troublshoot page at:

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

2.5.4 DSCLI requirements

The following IBM TotalStorage microcode and ESSCLI/DSCLI levels are required for the specified storage subsystems:

- General software:
  - IBM ESS microcode level 2.4.4.120 or later
  - IBM 2105 command-line interface (**`ibm2105cli.rte 32.6.100.13`**) or later
  - IBM 2105 command-line interface (**`ibm2105esscli.rte 2.1.0.15`**) or later

CLI location: We assume that the command-line interface is installed in its old default location, `/opt/ibm/ibm2105cli`. You might be required to create a link from the current default installation directory to this location.

- IBM ESS host attachment script disk device driver: **`ibm2105.rte`** version 32.6.100.25 or later. This driver supports all DS and ESS storage units.

For DS8000 when used with PowerHA/XD for Metro Mirror with DSCLI Management:

- DSCLI Version 5.3.1.236 or later on AIX 6.1
- DSCLI Version 5.1.730.216 or later on AIX 5.3

For DS6000 when used with PowerHA/XD for Metro Mirror with DSCLI Management:

- DSCLI Version 5.3.1.236 or later on AIX 6.1
- DSCLI Version 5.1.730.216 or later on AIX 5.3

For ESS 800 when used with PowerHA/XD for Metro Mirror with DSCLI Management:

- ESSCLI Version 2.4.4.129 or later when running on AIX 6.1
- ESSCLI Version 2.4.4.129 or later when running on AIX 5.3

Before you install or migrate to PowerHA/XD for Metro Mirror with DSCLI, ensure that you installed the following prerequisite software:

- PowerHA/XD for Metro Mirror with DSCLI requires AIX 5.3 or later.
- PowerHA/XD for Metro Mirror with DSCLI requires HACMP Version 5.4 or later.

When you combine heterogeneous storage units (ESS800, DS6000, and DS8000):

- The ESS CLI version as stated for ESS800 support must be installed on all nodes in the cluster.
- DSCLI versions must be installed as stated for each storage type.
- SDD or SDDPCM drivers must be installed on the nodes local to each storage type.

2.5.5 Requirements for PowerHA Enterprise Edition for Metro Mirror with SAN Volume Controller

The openssh Version 3.6.1 or later (for access to SAN Volume Controller interfaces) software and microcode levels are required.

- When using SDDPCM:
  - Subsystem Device Driver Path Control Module SDDPCM): v 2.2.0.0 or later
  - IBM Host attachment scripts:
    - **`devices.fcp.disk.ibm.mpio.rte`** 1.0.0.10 or later
    - **`ibm2105.rte`** 32.6.100.25 or later (as specified by SAN Volume Controller support)
- When running SAN Volume Controller Version 4.x
  - Storage microcode/LIC versions as per SAN Volume Controller support requirements

    Subsystem Device Driver (SDD) v 1.6.3.0 or later
  - 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 SAN Volume Controller support)

- When using Virtual I/O Server 1.5.1.x:
  - SDDPCM v 2.2.0.0 or later
  - IBM Host attachment scripts: **devices.fcp.disk.ibm.mpio.rte** 1.0.0.10 or later

2.5.6 Hardware considerations

The combinations for a cluster configuration vary based on the existing infrastructure and environment requirements. Device support is also a common concern that clients have when you design a solution.

When using PowerHA, the nodes within the same cluster do not need to be on the same machine type. For example, you can have a pair of POWER7 machines at your primary site and replicate to a remote environment by using POWER5 systems. Likewise, the local nodes can be a mix of POWER6 and POWER7 machines. The only potentially limiting factor when intermixing old and new nodes is the inability to take advantage of several of the newer features. An example is the inability to perform Live Partition Mobility unless you have a pair of POWER6 servers or later model servers, at a minimum.

Part 2

Campus style disaster recovery

This part describes campus style disaster recovery solutions with the IBM PowerHA SystemMirror Enterprise Edition and specific disaster recovery implementation details with the IBM PowerHA SystemMirror Enterprise Edition and cross-site Logical Volume Mirroring.

This part includes the following chapters:

3

Campus-style disaster recovery solutions

This chapter highlights the options that are available for campus-style disaster recovery solutions. In general, campus-style disaster recovery uses the synchronous mirroring method. Synchronous mirroring works in such a way that Logical Volume Manager (LVM) returns control to the application after writing both the local and the remote copies of the data, keeping our disaster recovery site up- to-date. Where having both sites up-to-date is advantageous, the time that it takes to write to the remote physical volumes can have a large impact on the application response time. For this reason, campus-style disaster recovery requires a standard environment that has low-latency access, high bandwidth, or gigabit Ethernet connections between the facilities. This approach is applicable for a short distance, up to 100 km, such as for towns and universities.

The main advantage of synchronous mirroring is that when a disaster occurs, no data loss occurs at the recovery site because the data at both sites is always kept the same.

This chapter includes the following sections:

- IBM cross-site LVM mirroring
- IBM SAN Volume Controller VDisk split I/O group
- IBM Metro Mirror
- IBM GLVM
- Performance implications

3.1 IBM cross-site LVM mirroring

Cross-site LVM mirroring, generally, is built from the same components that are used for local cluster solutions with storage area network (SAN)-attached storage. Local clusters and cross-site LVM mirroring have the following main differences:

- In local clusters, all nodes and storage subsystems are in the same location. Usually, the location consists of one external storage and is shared between two nodes.
- In cross-site LVM mirroring, cluster nodes and storage subsystems are on separate locations (sites). Each site has at least one cluster node and one storage subsystem with all necessary IP network and SAN infrastructure.

Figure 3-1 illustrates a general logical diagram of a cross-site LVM mirroring implementation.

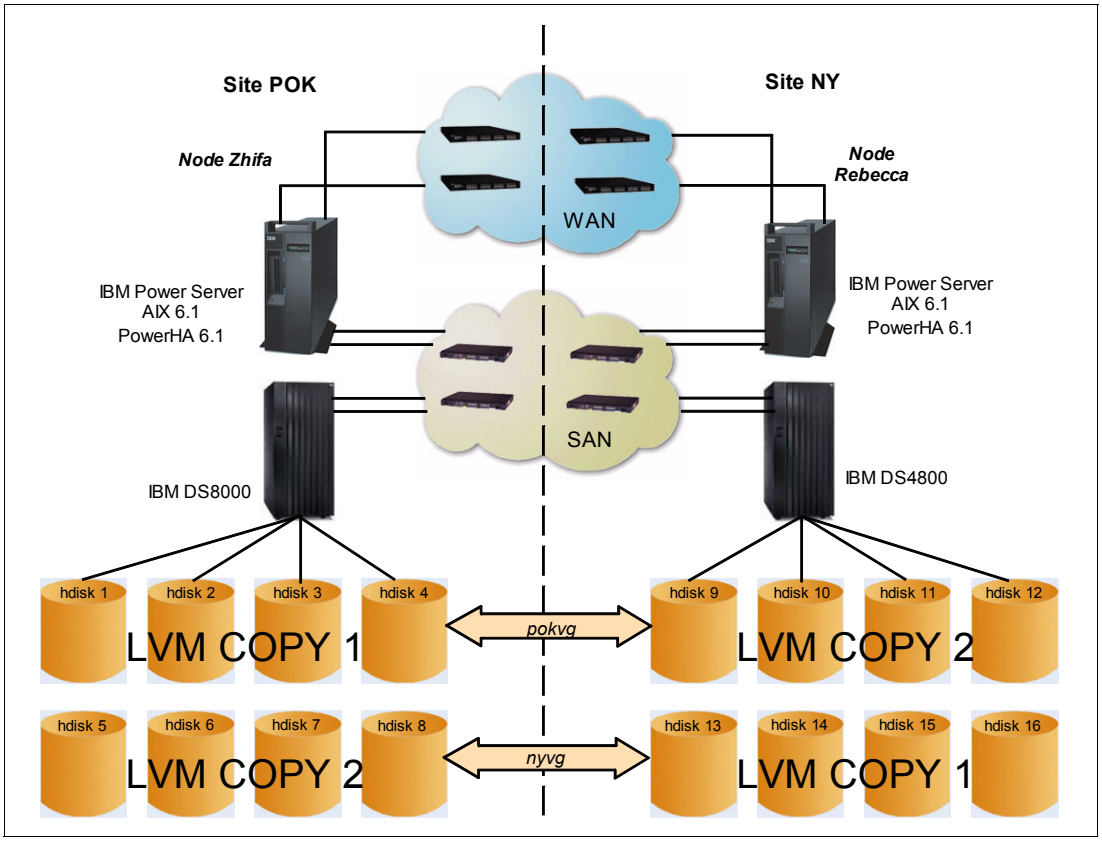


Figure 3-1 Cross-site LVM mirroring diagram

Figure 3-1 shows that cross-site LVM mirroring consists of at least two sites with one node server that is connected to an external storage at each site. These sites are connected by using an IP network and a SAN network. The data availability is ensured through the LVM mirroring between the volumes that are on separate storage subsystems on separate sites. These remote disks can be combined into a volume group by using the AIX Logical Volume Manager. This volume group can be imported to the nodes at separate sites. You can create logical volumes and set up an LVM mirror with a copy at each site. The number of active sites in a cross-site LVM mirroring that is supported in PowerHA environment is limited to two.

If a site failure occurs, PowerHA SystemMirror 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. If one storage subsystem fails, data access is not

interrupted and applications can access data from the active mirroring copy on the surviving disk subsystem.

PowerHA SystemMirror also drives automatic LVM mirroring synchronization after site or storage failure. When the failed site joins the cluster, it automatically fixes removed and missing volumes (PV states *removed* and *missing*) and synchronizes the data. Automatic synchronization is not possible for all cases, but we can use the C-SPOC menu to synchronize the data from the surviving mirrors to stale mirrors after a disk or site failure.

The limitation of this solution is the distance between the primary site and the secondary site. The distance is closely related to the SAN technology because it uses synchronous mirroring that requires high bandwidth. The following protocol technologies are usually used for connection between sites in a cross-site LVM mirroring:

- Fibre Channel over IP (FCIP)

  FCIP is a protocol specification that was developed by the Internet Engineering Task Force (IETF). It allows a device to transparently tunnel Fibre Channel frames over an IP network. An FCIP gateway or edge device attaches to a Fibre Channel switch and provides an interface to the IP network. At the remote SAN, another FCIP device receives incoming FCIP traffic and places Fibre Channel frames back onto the SAN. FCIP devices provide Fibre Channel expansion port connectivity, creating a single Fibre Channel fabric.

- Wave division multiplexing (WDM) devices

  This technology includes the following multiplexing types:

  - Coarse wavelength division multiplexing (CWDM), which is the less expensive component among the WDM technology
  - Dense wave length division multiplexing (DWDM), which is explained in 2.1.7, “DWDM” on page 49.

Cross-site LVM mirroring has the following advantages:

- It reduces license costs because it requires only the IBM PowerHA SystemMirror standard license.
- Easy setup and maintenance, especially for a customer who has good AIX LVM or PowerHA skills.

Cross-site LVM mirroring has the following considerations:

- It supports only synchronous replication.
- You must provide SAN or Fibre Connection between sites.

3.2 IBM SAN Volume Controller VDisk split I/O group

You have a few choices when you consider cluster configurations for campus-style disaster recovery implementations. Beyond cross-site LVM, Metro Mirroring, and synchronous Geographic Logical Volume Mirroring (GLVM), a SAN Volume Controller offering in the Version 5.1uses VDisk Mirroring (VDM). By using the VDM functionality, SAN Volume Controller can present a single VDisk to the host, which points to two copies of MDisk extents on the back-end storage.

The new SAN Volume Controller offering provides support for a split I/O group (Figure 3-2) configuration. It allows the SAN Volume Controller cluster nodes in the same I/O group to be split across buildings within a complex. This offering was formerly available through RPQ with older SAN Volume Controller versions. It is now considered a supported configuration in the SAN Volume Controller 5.1 code update.

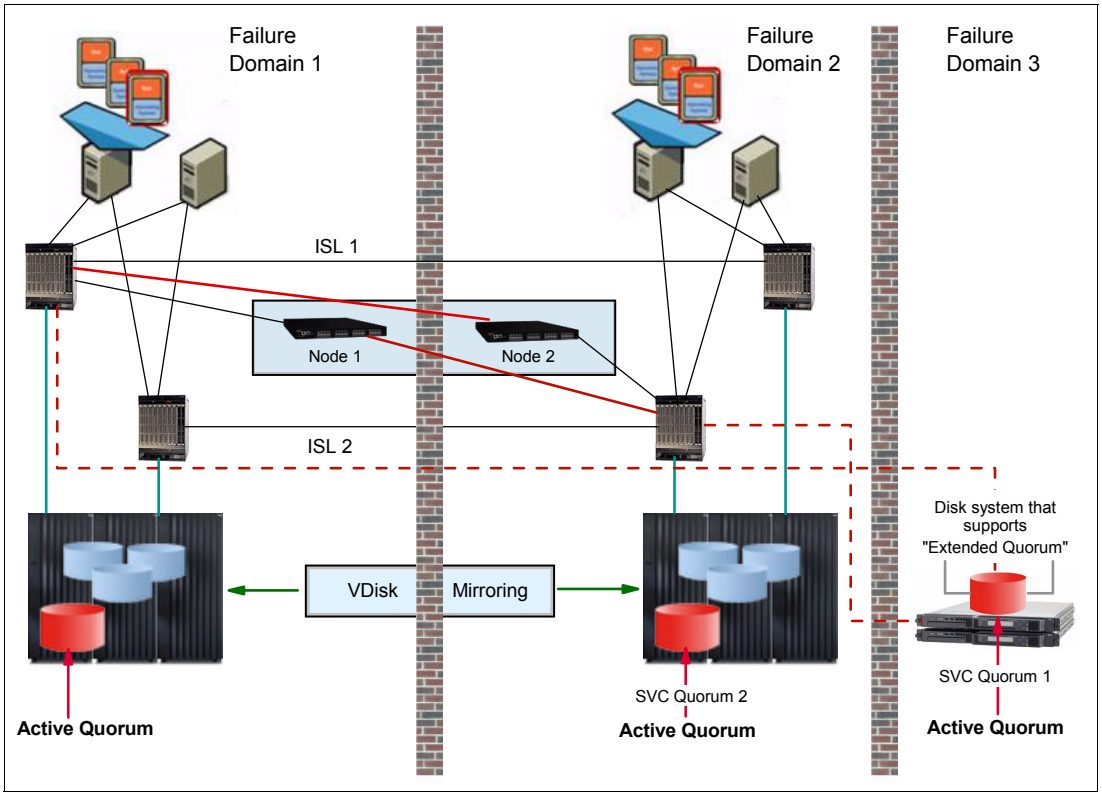


Figure 3-2 SVC 5.1 - split I/O group configuration

This solution assumes the use of the SAN Volume Controller as the storage front end within the environment and is comparable to a configuration that uses AIX logical volume mirroring. The difference is that, instead of detecting two physical volumes, only a single VDisk is visible at the host level. The mirroring is not visible. To the cluster, it looks like a traditional shared-disk environment. However, SAN Volume Controller provides protection against the loss of a storage enclosure, and, if a failure occurs, any stale partitions are not an issue for the host.

In implementations that use AIX 6.1, use mirror pools to harden logical volume mirroring and prevent the accidental spanning of logical partitions onto the wrong mirror copy whenever extending logical volumes through the command line. A solution that uses VDisk mirroring avoids this extra step and can simplify the management of the environment at the host level. It also avoids the extra processor utilization at the host level to maintain the mirrors and manage tasks such as mirror write consistency (MWC).

This approach has the following advantages:

- VDisks mirrored to both locations on separate disk storage subsystems
- VDisk view unchanged to host regardless of SAN Volume Controller node or disk failure
- No server shutdown required
- No maintenance of Metro Mirror or Global Mirror automation scripts for high availability (HA)
- No manual intervention required
- Automatic incremental resynchronization of VDisk copies

- Good fit in virtualized server environment:
  - PowerHA
  - AIX Live Partition Mobility
  - MSCS
  - VMware VMotion

This approach has the following disadvantages:

- Mix between HA and disaster recovery solution, but not a true disaster recovery solution
- No redundant SAN Volume Controller cluster involved

Keep in mind the following considerations:

- VDisk is owned by an I/O group, so if the I/O group goes offline the VDisk access is lost.
- If all quorum disks are inaccessible and VDisk Mirroring is unable to update its state information, a mirrored VDisk is taken offline to maintain data integrity.
- It is not a true disaster recovery solution, as there is only one SAN Volume Controller cluster involved.

This approach has the following restrictions:

- A node and its uninterruptible power supply (UPS) must be in the same rack.
- Nodes in an I/O group greater than 100 m apart must use long-wave connections directly to the switch or director.
- No ISLs can be used for node-to-node communication within an I/O group in any SAN Volume Controller cluster configuration.
- A third failure domain is required for the active quorum disk, and it must be on an IBM disk system that is listed as supporting *extended quorum*.
- Have a design reviewed by the IBM SAN Volume Controller development by using SCORE or by ATS if you are considering SAN Volume Controller split I/O groups with VDM.

For more information about splitting an I/O group between multiple sites, and the considerations that are involved with the implementation of quorum disks, write cache mirroring, and failure domains see the current SAN Volume Controller documentation.

3.3 IBM Metro Mirror

Synchronous disk-level replication is a disaster recovery solution that uses storage area network or Fibre Channel technology to replicate the data synchronously. IBM provides the solution for disaster recovery with synchronous disk-level replication by using IBM PowerHA SystemMirror with Metro Mirror.

Metro Mirror (previously known as synchronous Peer-to-Peer Remote Copy (PPRC)) provides real-time mirroring of logical volumes between two IBM ESS/DS6000/DS8000 family storage that can be located up to 300 km from each other. It uses a synchronous copy solution where write operations are completed on both copies (local and remote site) before they are considered to be complete. It is typically used for applications that cannot suffer any data loss if a failure occurs.

As data is synchronously transferred, the distance between the local and the remote disk subsystems determines the effect on application response time. Figure 3-3 illustrates the sequence of a write update with Metro Mirror.

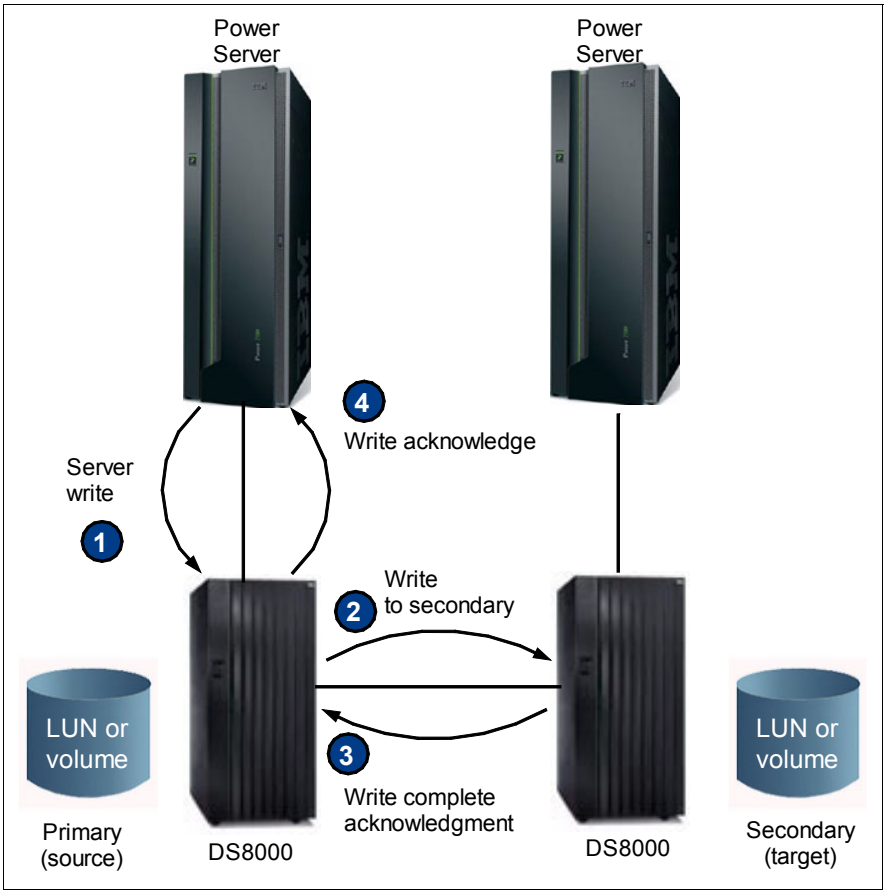


Figure 3-3 Writing process in synchronous Metro Mirror

When the application performs a write update operation to a source volume, the process has the following flow:

1. Write to the source volume.
2. Write to the target volume.
3. Signal write complete from the remote target.
4. Post I/O complete to the host server.

The Fibre Channel connection between the local and the remote disk subsystems can be direct, through a switch, or through other supported distance solutions such as DWDM.

The PowerHA SystemMirror with Metro Mirror provides automated copy split in case of primary site failure and automated reintegration when the primary site becomes available. Figure 3-4 shows a typical configuration for PowerHA SystemMirror with Metro Mirror.

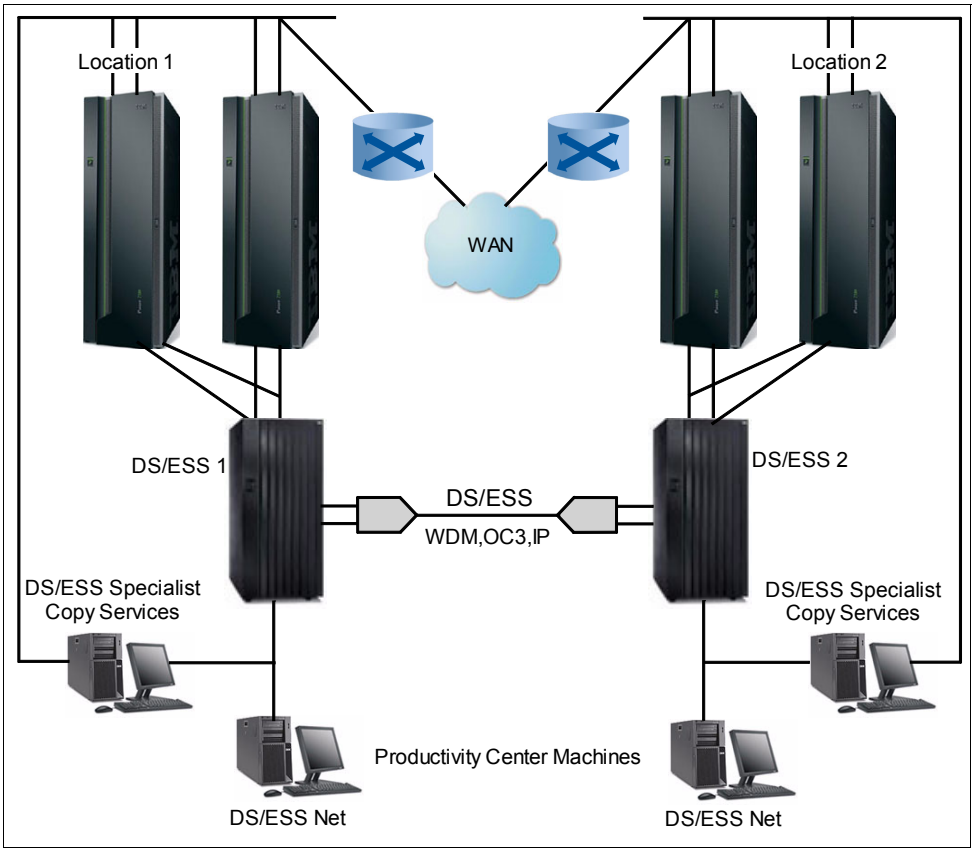


Figure 3-4 Example of PowerHA SystemMirror with Metro Mirror configuration

Advantages of Metro Mirror are that it:

- Is suitable for a customer who already has the IBM Storage DS8000/DS6000/ESS series.
- Can be converted to asynchronous replication mode for unlimited distance replication.
- Can be implemented in different models of IBM storage DS/ESS series.

Keep in mind the following considerations about Metro Mirror:

- You must provide a SAN or Fibre Channel connection between sites.
- You need an IBM PowerHA SystemMirror Enterprise license.

3.4 IBM GLVM

The IBM solution for synchronous IP-level replication is the IBM PowerHA SystemMirror with GLVM. This solution can mirror data across a standard IP network and provide automated failover and failback support for the applications that is using the geographically mirrored data.

Figure 3-5 illustrates the logical diagram of a PowerHA SystemMirror solution with GLVM.

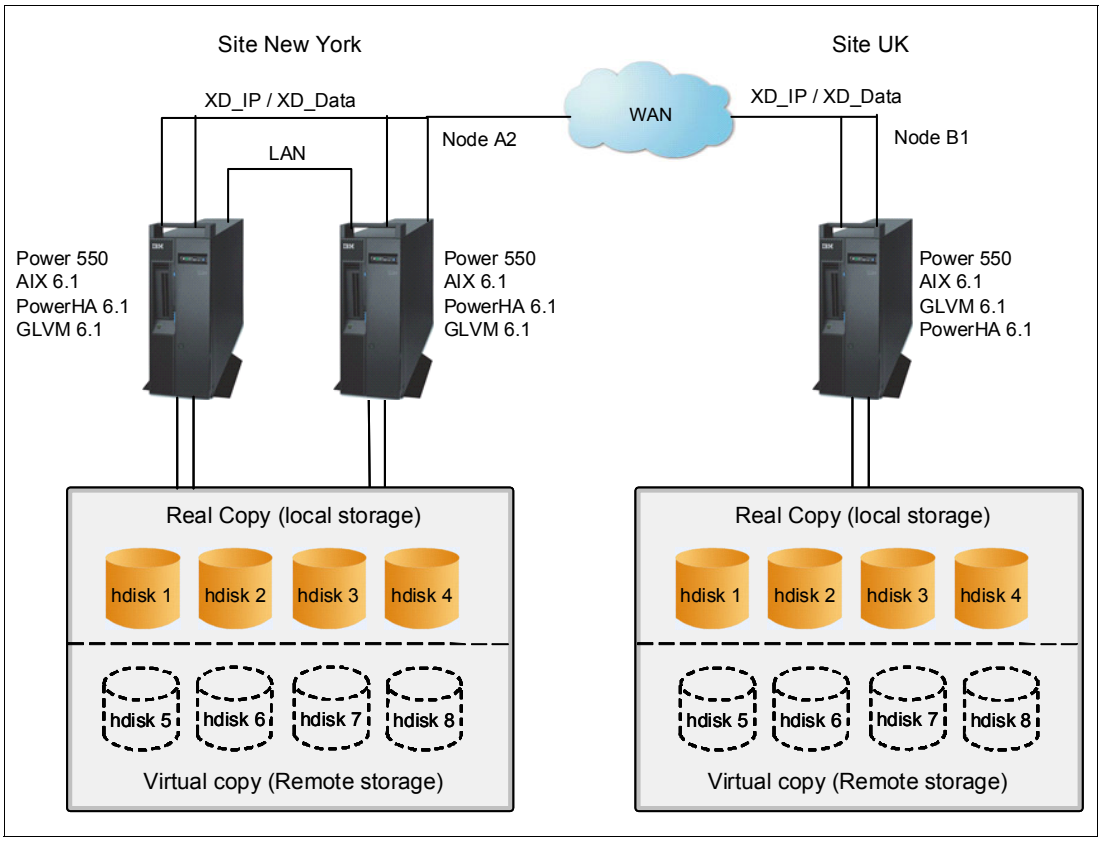


Figure 3-5 PowerHA SystemMirror with GLVM technology

GLVM performs the remote mirroring of AIX logical volumes by using the basic AIX LVM functions for optimal performance and ease of configuration and maintenance.

PowerHA/XD GLVM provides two essential functions:

- Remote data mirroring
- Automated failover and failback

Together these functions provide high-availability support for applications and data across a standard Internet Protocol network to a remote site.

PowerHA SystemMirror with GLVM provides the following capabilities:

- It allows automatic detection of and response to site and network failures in the geographic cluster without user intervention.
- It performs automatic site takeover and recovery and keeps mission-critical applications highly available through application failover and monitoring.
- It allows for simplified configuration of volume groups, logical volumes, and resource groups.
- It uses the Internet Protocol network for remote mirroring over an unlimited distance.
- It supports maximum sized logical volumes.

Figure 3-6 shows a diagram of how PowerHA SystemMirror with GLVM works.

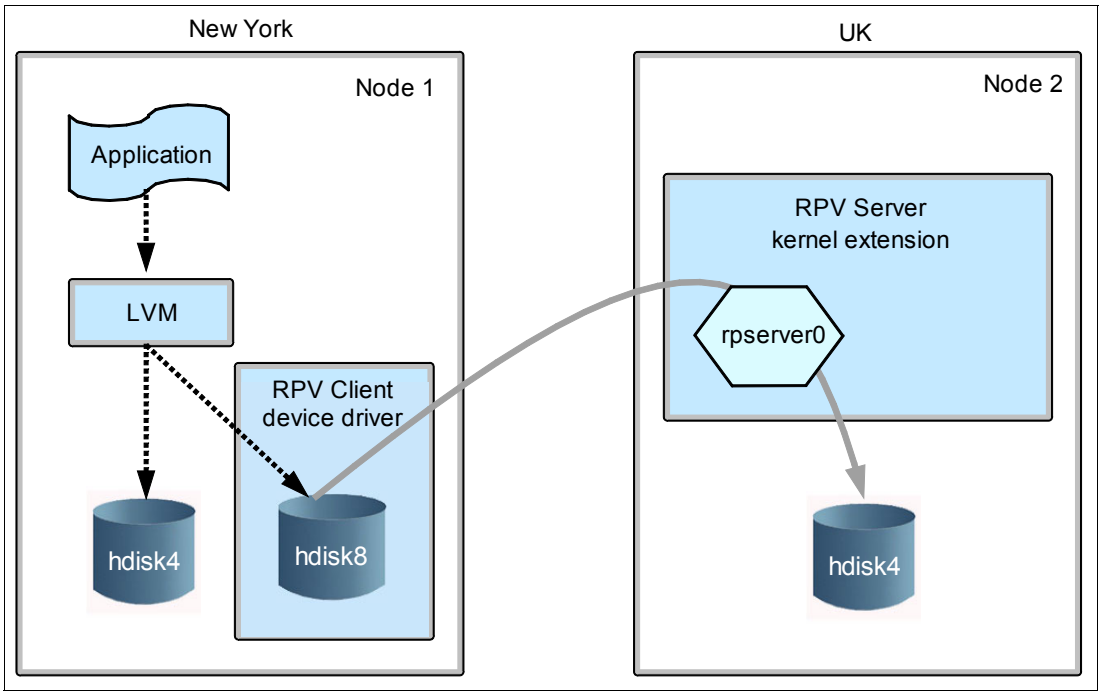


Figure 3-6 PowerHA SystemMirror with GLVM work process when node 1 is active

Figure 3-6 shows an example of two sites, with one node at each site. Node 1 has one physical volume, hdisk4, and the remote node has one physical volume also, hdisk4. These disks are configured to be one volume group and mirror each other's. Viewing the replication from node 1, we see that the destination physical volume hdisk4 on node 2 is presented on Node1 as hdisk8. Currently, node 2 functions as an RPV server, and node 1 functions as an RPV client.

The reverse condition happens when node 1 is down or the application moves to node 2 (Figure 3-7). The replication occurs from node 2 to node 1 and the physical volume hdisk4 from node 1 is presented in node 2 as hdisk8.

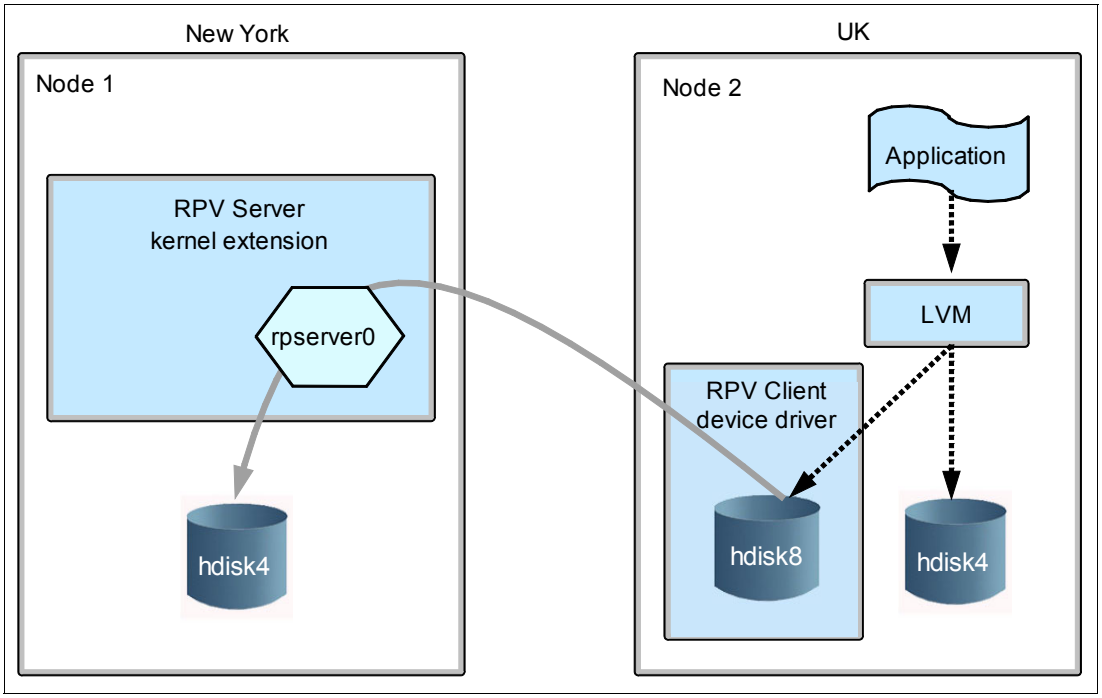


Figure 3-7 PowerHA SystemMirror with GLVM work process when application moves to node 2

PowerHA SystemMirror Enterprise Edition with GLVM has the following key components, among others:

- Remote physical volume (RPV)

  RPV is a pseudo device driver that provides access to the remote disks as though they are locally attached. The remote system must be connected by using a Internet Protocol network. The distance between the sites is limited by the latency and bandwidth of the connecting networks.

  The RPV consists of two parts:

  – RPV client

    This pseudo device driver runs on the local machine and allows the AIX LVM to access remote physical volumes as though they are local. The RPV clients are seen as hdisk devices, which are logical representations of the remote physical volume.

    The RPV client device driver appears as an ordinary disk device (for example, RPV client device hdisk8) and has all its I/O directed to the remote RPV server. It is unaware of the nodes and networks.

    In PowerHA/XD, concurrent access is not supported for GLVM. Therefore, when you are accessing the RPV clients, the local equivalent RPV servers and remote RPV clients must be in a defined state.

  – RPV server

    The RPV server runs on the remote machine, one for each physical volume that is being replicated. The RPV server can listen to a number of remote RPV clients on separate hosts to handle failover. The RPV server is an instance of the kernel extension of the RPV device driver with names such as rpvserver0 and is not an actual physical device.

– Geographically mirrored volume group (GMVG)

A GMVG is a volume group that consists of local and remote physical volumes. If you use IBM AIX 6, you can implement a mirror pool configuration to make sure that you have a complete copy of the mirror at each site.

PowerHA GLVM also expects each logical volume in a GMVG to be mirrored. GMVGs are managed by PowerHA/XD and are recognized as a separate class of replicated resources, so they have their own events. PowerHA SystemMirror verification issues a warning if there are resource groups that contain GMVG resources that do not have the forced **varyon** flag set and if quorum is not disabled.

PowerHA with GLVM solution has the following advantages:

- The infrastructure is simple. You need only IP connectivity between the sites.
- It supports up to four IP-based networks that are configured to provide the data stream between mirrored volume group copies at the two sites.
- The local and remote storage systems do not have to be the same equipment.
- It can be converted to asynchronous replication and supports unlimited distance replication.

PowerHA with GLVM requires the following considerations:

- It supports up two sites, a local site and a remote site.
- It needs an IBM PowerHA SystemMirror Enterprise Edition license.
- The rootvg volume group cannot be geographically mirrored.

3.5 Performance implications

Because of the nature of separate application behaviors and the myriad of devices that are used in client environments, a discussion about performance implications can be subjective and open to interpretation. Therefore, the goal in this section is not to indicate whether one solution is better than the next, but to document findings based on our testing and reference the results that were found in another study for a customer proof of concept.

Looking past the availability merits of each solution, the key metrics that clients focus on are the new I/O characteristics that are introduced when one of the mirroring solutions is implemented. Having an idea about the read and write behaviors in your own environment can help steer you toward the most appropriate solution.

In our test environment, we configured various independent clusters for testing the replication technologies. For the purposes of this section, we elected to focus on only three of these clusters. The three scenarios that we compared were AIX Logical Volume Mirroring, SAN Volume Controller Metro and Global Mirroring, and DS Metro Mirroring. To show the impact of the mirroring, we compared an unmirrored SAN-attached volume to a second mirrored LUN in each cluster configuration.

3.5.1 Test environment details

All three cluster configurations in our two sites were separated between DWDMs and had a one-way SAN distance of approximately 45 km. Our testing was performed with a simulated write load by using the **ndisk64** utility, which is part of the **nstress** package.

You can download the package from:

https://www.ibm.com/developerworks/mydeveloperworks/wikis/home?lang=en#/wiki/Power%20Systems/page/nstress

Our storage subsystems included a mix of DS4800, DS8000, and SAN Volume Controller clusters at each remote site. The storage subsystems were not dedicated to our clusters and had other varying loads running on them. For this reason, our results should not be interpreted as an actual benchmark or an accurate indication of what is expected in a real-life scenario.

The fabric used Brocade switches at each site, running at a speed of 2 GB. The SAN attachment from each host was virtualized by using 8-GB NPIV capable adapters. No additional tuning was performed.

Example 3-1 shows one of the scripts that is used to generate the 100% random write load to a raw logical volume named `mikeylv`. The output in Example 3-1 shows the script setup to run with 30 threads.

Example 3-1 The ndisk64 script that was used to generate disk I/O load

```
#vi rand.ksh

#!/bin/ksh
export DIR=.
export TIME=$1
if [[ $TIME < 1 ]]
 then
   export TIME=60
fi
DATE=`date +"%m.%d.%Y"`
#$DIR/ndisk64 -R -r 0 -M 1 -s 12G -b 4k -t $TIME -f /dev/rmikeylv -o "All Writes" >rsync1.$DATE &
#$DIR/ndisk64 -R -r 0 -M 2 -s 12G -b 4k -t $TIME -f /dev/rmikeylv -o "All Writes" >rsync2.$DATE &
#$DIR/ndisk64 -R -r 0 -M 5 -s 12G -b 4k -t $TIME -f /dev/rmikeylv -o "All Writes" >rsync3.$DATE &
#$DIR/ndisk64 -R -r 0 -M 10 -s 12G -b 4k -t $TIME -f /dev/rmikeylv -o "All Writes" >rsync4.$DATE &
$DIR/ndisk64 -R -r 0 -M 30 -s 12G -b 4k -t $TIME -f /dev/rmikeylv -o "All Writes" >rsync5.$DATE &
#$DIR/ndisk64 -R -r 0 -M 60 -s 12G -b 4k -t $TIME -f /dev/rmikeylv -o "All Writes" >rsync6.$DATE &
#$DIR/ndisk64 -S -r 0 -M 1 -s 12G -b 512k -t $TIME -f /dev/rmikeylv -o "All
Writes">rsync7.$DATE&
```

Example 3-2 shows the results from a 5-minute run of the same **rand.ksh** script. The output shows the total IOPS and the overall throughput that is driven through the LUN that is hosting the unmirrored logical volume.

Example 3-2 Output from ndisk64 test that is run against unmirrored volume

```
# ./rand.ksh 300

# cat rsync5.03.20.2010
Command: All Writes -R -r 0 -M 30 -s 12G -b 4k -t 300 -f /dev/rmikeylv
        Synchronous Disk test (regular read/write)
        No. of processes = 30
        I/O type         = Random
        Block size       = 4096
        Read-Write       = Write Only
        Sync type: none  = just close the file
        Number of files  = 1
        File size        = 12884901888 bytes = 12582912 KB = 12288 MB
```

```
        Run time         = 300 seconds
        Snooze %         = 0 percent
----> Running test with block Size=4096 (4KB) .............................
Proc - <-----Disk IO----> | <-----Throughput------> RunTime
 Num -      TOTAL   IO/sec |    MB/sec       KB/sec  Seconds
   1 -     330990   1103.3 |      4.31      4413.22   300.00
   2 -     334098   1113.7 |      4.35      4454.62   300.00
.............
.............
  29 -     334247   1114.2 |      4.35      4456.62   300.00
  30 -     333953   1113.2 |      4.35      4452.66   300.00
TOTALS    9982634  33275.5 |    129.98 Rand procs= 30 read=  0% bs=  4KB
```

The total IOPS show a value of 33275.5, which resulted from a test that was run by using 30 threads against an unmirrored SAN-attached SAN Volume Controller LUN. In our scenarios, more tests were performed by using different thread counts and different local logical volumes, which had mirroring in place. The summary of the results is described in the next section.

3.5.2 Mirroring performance penalty test results

In the first cluster we tested, we used a cross-site LVM configuration, which used the AIX Logical Volume Mirroring between a DS4800 storage enclosure and a DS8000 subsystem.

The load was initiated by using the same **ndisk64** utility that was started over various 5-minute intervals runs. We used thread counts ranging from 1, 2, 5, 10, to 30. In this scenario, we compared the impact of a stand-alone DS4800 LUN, a second mirrored locally, and a third mirrored to a LUN presented from the remote site 45 km apart. Figure 3-8 shows how the IOPS rises as the number of threads increases. The additional lines show the penalty that is imposed by the local and remote mirrors.

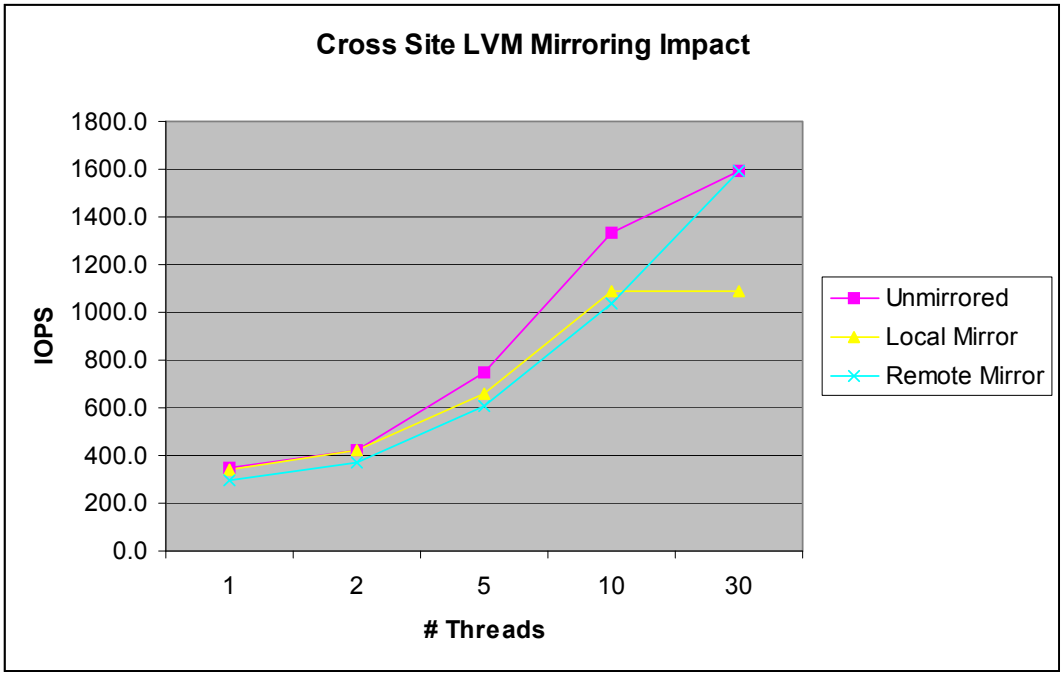


Figure 3-8 Cross-site LVM mirroring impact

In the second cluster, we tested SAN-attached volumes from a SAN Volume Controller backed by DS8000 storage. The mirrored LUNs used for our testing were being replicated by using Metro Mirror and Global Mirror Copy Service functions. The distance between our two sites was the same as in our other test scenarios, but we had no control over the additional load that was being imposed on the SAN Volume Controller by other servers that were attached.

Figure 3-9 shows the mirroring penalty and how the Global Mirror replication provides better throughput than the Metro Mirrored volumes.

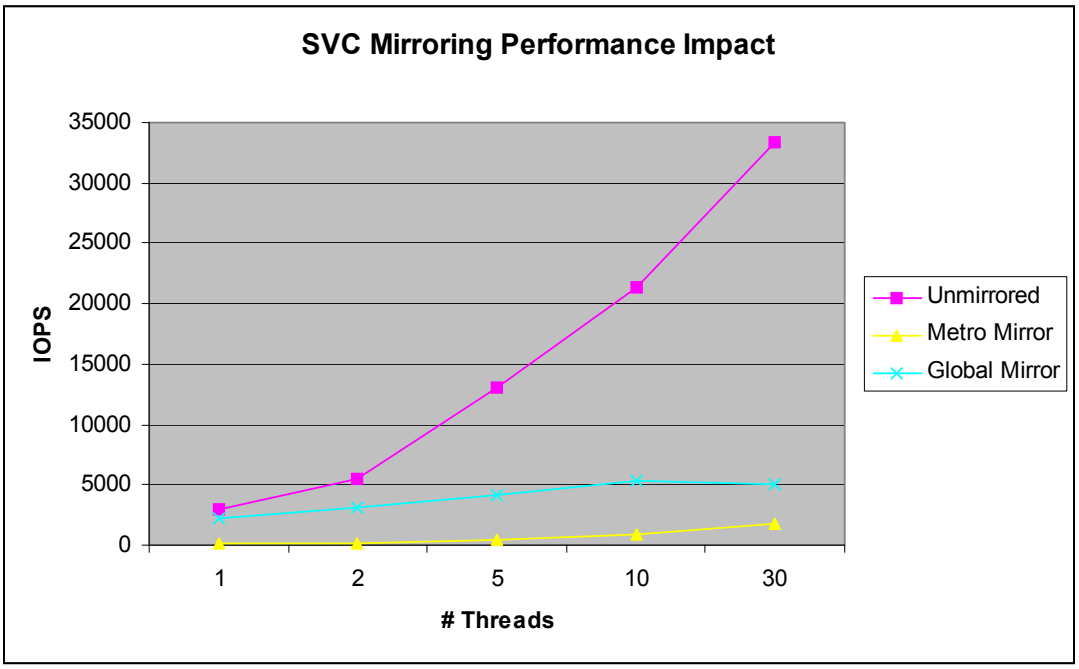


Figure 3-9 SAN Volume Controller mirroring impact

In our last clustered environment, we tested DS8000 mirrored and unmirrored LUNs that replicated between separate storage units, which were at each site. The distance and SAN infrastructure were also the same as in the previous tests.

Figure 3-10 shows the affect of the Metro Mirror replication as it pertains to the number of IOPS. We elected not to test DS Global Mirror because it currently has no integration with the PowerHA Enterprise Edition.

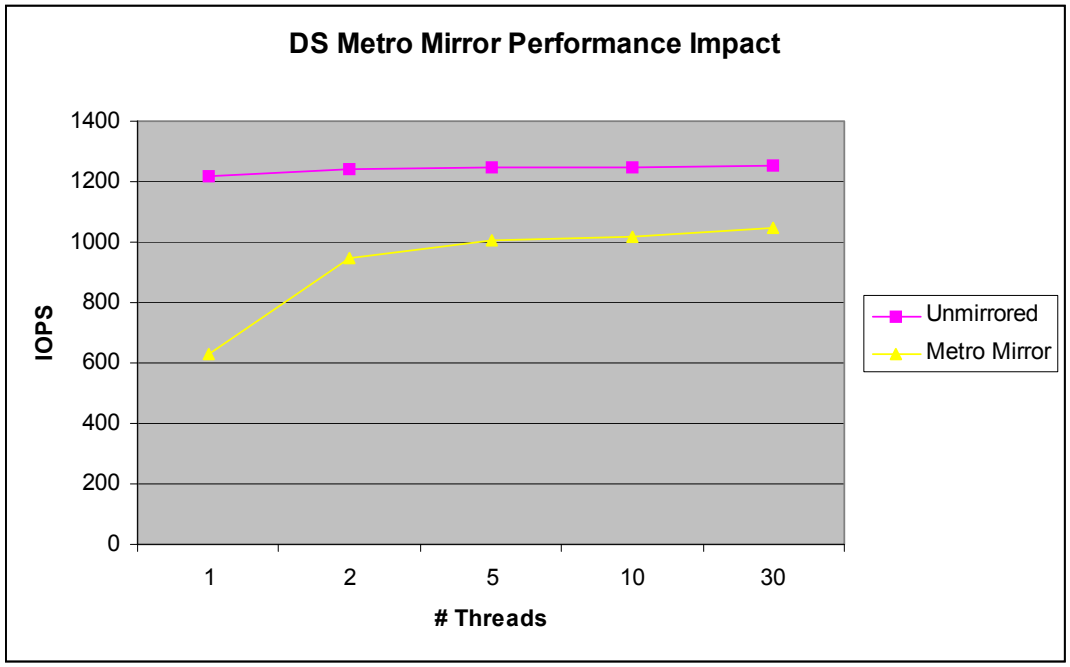


Figure 3-10 DS Metro Mirroring with the mirroring impact

3.5.3 Test results

The goal of these tests was not to compare the replication between the solutions but to highlight the impact that is imposed whenever replication is introduced into an environment. All three scenarios had an obvious mirroring penalty. As expected, the use of asynchronous mirroring showed an improvement over the synchronous replication. In a real-life scenario, the results might vary based on several characteristics.

Our testing was performed with a simulated random load that did all writes. However, real application loads might have different random and sequential tendencies and might be more write or read intensive. More factors, such as the fabric speed, the hardware components that were used, and the extra load from other servers in the environment can also impact the findings. The key is to evaluate the environments that are being considered and to perform a proof of concept to ensure that the proposed solution yields the desired results.

3.5.4 Customer case study

A separate study was documented in 2008 that compared cross-site LVM mirroring with a Metro Mirrored environment approximately 10 kilometers (6.1 miles) apart. The detailed study was documented within the *Understanding the Performance Implications of Cross-Site Mirroring with AIX's Logical Volume Manager* white paper and is available for download at:

`http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101269`

In this study, the response time of disk writes and transaction throughput was measured under different load conditions. In contrast to our testing, the study used a simulated load by using the **`Swingbench`** utility on an Oracle database. The study draws a comparison with the

servers running at 60% utilization and 100% utilization, highlighting the difference between the two mirroring solutions.

Ultimately, the study determined that writes ended up taking less time in the cross-site LVM configuration than in the Metro-Mirrored environment, and slightly longer for reads (Figure 3-11).

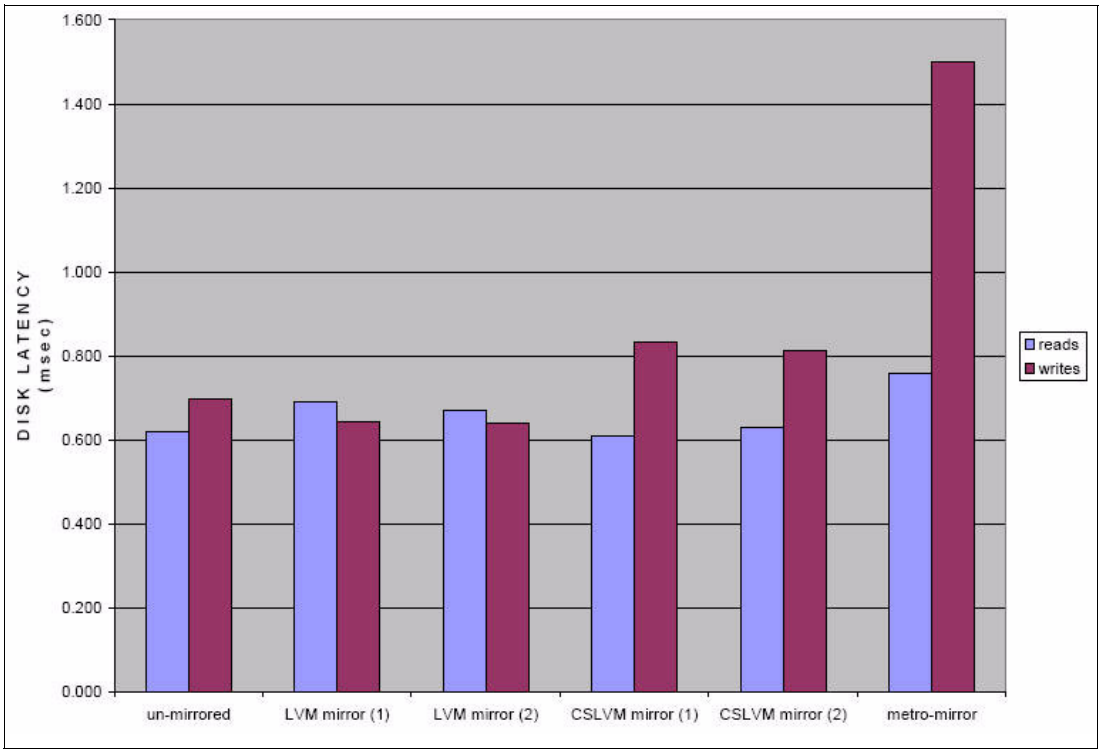


Figure 3-11 Remote mirroring performance results from customer proof of concept

The study explains that this is a result of AIX LVM Mirroring submitting both write requests at nearly the same time. The write acknowledgement is sent back from both disk subsystems as soon as they both have the data in the cache. When both acknowledgements are received by the host, the application is signaled to go ahead to the next transaction. In Metro Mirroring, the host system sends a write request to local storage and the enclosure then sends a second write request to the remote subsystem. Acknowledgement must be received back at the host before the application can proceed.

3.5.5 Summary

There are different solutions available to replicate the data in a campus-style disaster recovery environment. Evaluating the I/O characteristics in the environment and considering the performance impact that is imposed by the replication should all be part of the design planning. Our test results seem to fall in line with the results documented in the case study.

Although we did not focus on the read aspect, the write characteristics showed AIX Logical Volume Mirroring providing more consistent results. Since logical volume mirroring is a function specific to AIX and the individual host, make further considerations when you select a technology for data replication. The disk mirroring functions are not platform-specific and can be used to provide a common replication method between servers in the environment. The use of IBM resources to assist with the design planning and performance evaluation are encouraged.

4

Configuring PowerHA Standard Edition with cross-site logical volume mirroring

This chapter explains how to set up cross-site Logical Volume Manager (LVM) mirroring cluster using IBM AIX 6.1 and PowerHA SystemMirror 6.1. This is one of the scenarios that can be implemented for campus-style disaster recovery solutions.

In this chapter, we use mirror pool configuration for cross-site LVM mirroring, which is one of the new features in AIX6L. We highlight mirror pool definition and function and explain the steps to configure a mirror pool.

The chapter includes the following sections:

- Configuring the cross-site LVM mirroring cluster
- Testing cross-site LVM mirroring cluster
- Maintaining cross-site LVM mirroring cluster

4.1 Configuring the cross-site LVM mirroring cluster

To help you understand how to set up and configure a cross-site LVM mirroring cluster by using IBM PowerHA System Mirror for AIX 6.1, follow the example provided in this section. We set up the cross-site LVM cluster as a new cluster implementation. You can also implement it with an existing local cluster by adding a site with at least one node and one external storage and integrate it with the cross-site LVM cluster.

Figure 4-1 shows our testing environment.

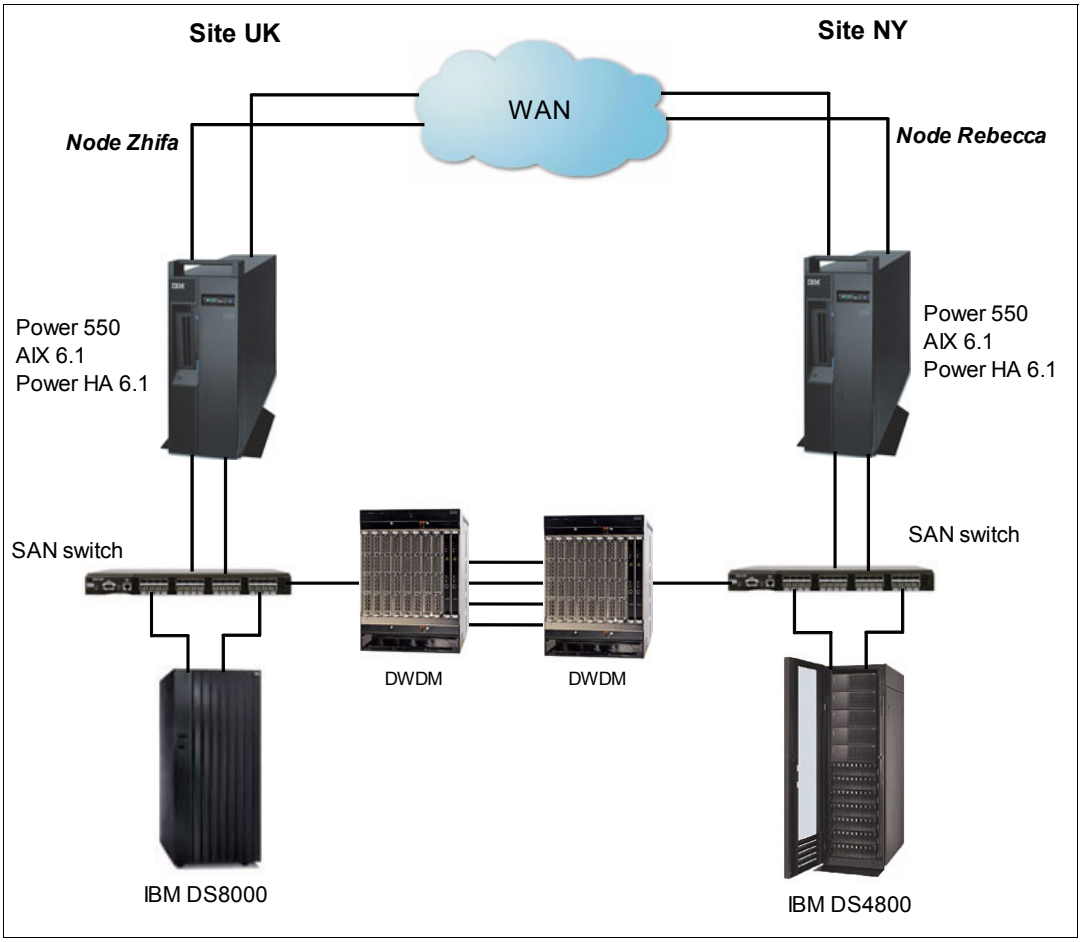


Figure 4-1 Testing environment

4.1.1 Configuring the cluster topology

In our environment, we have two sites (Figure 4-1), and each site has one server/node and one external storage. We use an IBM Power 550 server that is connected to an IBM DS8000 storage at site 1, and an IBM Power 550 connected to an IBM DS4800 storage at site 2. Both servers are configured with a Virtual Input/Output Server (VIOS) by using N-Port ID Virtualization (NPIV) (feature code 5735) Fibre Channel (FC) adapters.

In an ideal environment, such as a production environment, each component must be redundant to avoid single point of failures. The VIOS, Ethernet adapter, Fibre Channel adapter, and power supply, for example, must be redundant. In our test environment, because of hardware limitations, we implement only one VIOS, while the others components are redundant.

Prerequisites for installation of the environment

Before you configure the cluster topology, check that the necessary software is installed on all cluster nodes. The software includes the following components:

- IBM PowerHA SystemMirror Standard Edition 6.1.
- IBM AIX6L TL2.
- The latest version of IBM RSCT.
- Storage device driver for MPIO.
- 1 MB of disk space, which is used by the PowerHA software (*only* if you install just the XD file sets)

  However, ensure that the `/usr` file system has 82 MB of free disk space for the full PowerHA base product installation.

Creating a cluster

Create a cluster and name it `XLVM_cluster`. You create it by using SMIT menus (Figure 4-2). We enter the **`smitty hacmp`** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure an HACMP Cluster** → **Add/Change/Show an HACMP Cluster**.

```
                    Add/Change/Show an HACMP Cluster

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

                                                    [Entry Fields]
* Cluster Name                                     [XLVM_Cluster]

  NOTE: HACMP must be RESTARTED
  on all nodes in order for change to take effect
```

Figure 4-2 SMIT menu for creating the cluster

Creating the cluster node and the cluster site

Create a two-site cluster with one node at each site. Site 1 is POK and site 2 is NY. The node at site POK is Zhifa, while the node at site NY is Rebecca.

Use the SMIT menu to create the site and the node (Figure 4-3). Enter the **`smitty hacmp`** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Nodes** → **Add a node to the HACMP Cluster**.

```
                    Add a Node to the HACMP Cluster

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

                                                    [Entry Fields]
* Node Name                                        [Zhifa]
  Communication Path to Node                       [192.168.8.101]        +
```

Figure 4-3 SMIT menu for adding a cluster node

Run the **smitty hacmp** command once again to add the second node. Then, add the cluster site by running the following SMIT command. Enter the **smitty hacmp** command, and then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Sites** → **Add a Site**. Figure 4-4 shows the Add Site panel.

```
                                Add Site

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                    [Entry Fields]
* Site Name                                         [POK]
* Site Nodes                                         Zhifa                +
```

Figure 4-4 SMIT menu for adding a cluster site

When adding a cluster site, complete the following required parameters:

Site Name Define a name of a site by using no more than 64 alphanumeric and underscore characters.

Site Nodes For each site, define at least one node that is in the site.

You can add multiple nodes by leaving a blank space between the names. Each node can only belong to one site.

Run the **smitty hacmp** command again to create site NY.

Creating the cluster network

Two kinds of networks are need to be configured in the cluster.

IP network

The IP network uses TCP/IP communication and has the following functions:

IP services An IP label that associates with a service address and that is used by users for accessing an application.

IP base An IP label that associates with a non-service address and that is configured by AIX at boot time and is not used for accessing the application.

IP persistent An IP label that associates with a persistent address and that is used for cluster administration and maintenance.

In this example, we use the configuration that is shown in Table 4-1 for the IP network.

Table 4-1 List of IP addresses in our environment

	Node Zhifa	**Node Rebecca**
IP base	192.168.8.101	192.168.8.102
IP services	192.168.100.53	192.168.100.92
IP persistent	192.168.100.171	192.168.100.171

Register the IP addresses from Table 4-1 in the /etc/hosts file (Example 4-1).

Example 4-1 List of IP addresses and host names in the /etc/hosts file

```
#  service addresses
192.168.100.53 zhifa_svc    XLVM_550_1_A_SVC
192.168.100.92 rebecca_svc  XLVM_550_2_B_SVC

# boot addresses
192.168.8.101 zhifa_boot    XLVM_550_1_A_boot
192.168.8.102 rebecca_boot  XLVM_550_2_B_boot

# persistent addresses
192.168.100.171 rebecca_per XLVM_550_2_B
192.168.100.171 zhifa_per XLVM_550_1_A  Zhifa
```

To configure the IP network, run the SMIT command (Figure 4-5). Enter the **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Networks** → **Add a Network to the HACMP Cluster**.

```
                Add an IP-Based Network to the HACMP Cluster

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

                                                  [Entry Fields]
* Network Name                                   [net_ether_01]
* Network Type                                    ether
* Netmask(IPv4)/Prefix Length(IPv6)              [255.255.255.0]
* Enable IP Address Takeover via IP Aliases      [Yes]                +
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 4-5 SMIT menu for adding a network

Then, configure the communication network for node Zhifa and node Rebecca by running the **smitty hacmp** SMIT command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Communication Interfaces/Devices** → **Add Communication Interfaces/Devices** (Figure 4-6). As shown in Figure 4-6, we registered the IP base of the communication interfaces.

```
                        Add a Communication Interface

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

                                                  [Entry Fields]
* IP Label/Address                               [zhifa_boot]         +
* Network Type                                    ether
* Network Name                                    net_ether_01
* Node Name                                      [Zhifa]              +
  Network Interface                              []
```

Figure 4-6 SMIT menu for creating the network interfaces

Add the IP persistent address during the network configuration by running the **smitty hacmp** SMIT command. Then, select **Extended Configuration → Extended Topology Configuration → Configure HACMP Persistent Node IP Label/Addresses → Add a Persistent Node IP Label/Addresses**. Figure 4-7 shows the SMIT menu for adding the IP persistent address.

```
                    Add a Persistent Node IP Label/Address

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

                                                      [Entry Fields]
* Node Name                                          Zhifa
* Network Name                                      [net_ether_01]          +
* Node IP Label/Address                             [zhifa_per]             +
  Netmask(IPv4)/Prefix Length(IPv6)                 [255.255.255.0]
```

Figure 4-7 SMIT menu for creating the IP persistent address

Run the **smitty hacmp** command to create the IP persistent address for node Rebecca. After you create the IP base and IP persistent addresses, create the non-IP network.

Non-IP network

A non-IP network is used for cluster heartbeat to prevent the split-brain condition in a cluster environment. Certain devices can be used as a non-IP network. In this case, we use a hard disk as a cluster non-IP network.

Two kinds of disk heartbeats are possible:

- Traditional disk heartbeat

  Disk heartbeat is a form of non-IP heartbeat that uses the existing shared disks of any disk type. This feature, which was introduced in HACMP 5.1, is the most common and preferred method of non-IP heartbeat. It eliminates the need for serial cables or 8-port asynchronous adapters. Also, it can easily accommodate greater distances between nodes when using a SAN environment.

  This feature requires usage of enhanced concurrent volume groups to allow access to the disk by each node. It uses a special reserved area on the disks to read and write the heartbeat data. Because it uses a reserved area, it allows the use of existing data volume groups without losing any additional storage space.

  It is possible to use a dedicated disk or LUN for disk heartbeat. However, because disk heartbeat uses the reserved space, the remaining data storage is not used. The bigger the disk/LUN you use solely for this purpose, the more space is not used. However, you can use it later for extra storage space if needed.

  A traditional disk heartbeat network is a point-to-point network. If more than two nodes exist in your cluster, you need a minimum of *N* number of non-IP heartbeat networks, where *N* represents the number of nodes in the cluster. For example, a three-node cluster requires at least three non-IP heartbeat networks.

- Multinode disk heartbeat

  HACMP 5.4.1 introduced another form of disk heartbeating called *multi-node disk heartbeat (mndhb)*. Unlike traditional disk heartbeat network, it is not a single point-to-point network. Instead, as its name implies, it allows multiple nodes to use the same disk. However, it requires configuring a logical volume on an enhanced concurrent volume group.

Where this heartbeat can reduce the total number of disks that are required for non-IP heartbeating, we configured it with multiple disks to eliminate a single point of failure.

Multinode disk heartbeating also offers the ability to start one of the following actions:

Halt	Halts the node (default).
Fence	Fences this node from the disks**.**
Shutdown	Stops PowerHA on this node (gracefully).
Takeover	Moves the resource groups to a backup node.

In our environment, we create a traditional disk heartbeat by using two disks, one in each node, for redundancy purposes. This redundancy is set to keep alive the cluster in case of one disk subsystem fails. The first network uses disks from the IBM DS8000 storage, and the second network uses disks from the IBM DS4800 storage. The first network uses physical volume *hdisk1*, and the second network uses physical volume *hdisk5*. These disks are assigned into a volume group, *pokvg*, that is created in 4.1.2, “Configuring the cross-site LVM disk mirroring dependency” on page 121.

To create a disk heartbeat network:

1. Add a disk heartbeat network. Run **`smitty hacmp`**.
2. Select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Networks** → **Add a Network to the HACMP Cluster**.
3. Enter the Network Name `net_diskhb_01`, select the Network Type **diskhb**, and press Enter (Figure 4-8).

```
                Add a Serial Network to the HACMP Cluster

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

                                                      [Entry Fields]
* Network Name                                        [net_diskhb_01]
* Network Type                                         diskhb
```

Figure 4-8 Creating the network for disk heartbeat

We create net_diskhb_01 for the heartbeat network by using the DS8000, and then create net_diskhb_02 for the heartbeat network by using the DS4800.

4. To add the heartbeat devices, run the **`smitty hacmp`** SMIT command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Communication Interfaces/Devices** → **Add Communication Interfaces/Devices**.

   Figure 4-9 on page 120 shows the creation of a communication device in the DS8000 for node Zhifa. Continue this process to create a communication device in the DS8000 for node Rebecca. Then, create a communication device in the DS4800 for nodes Zhifa and Rebecca.

```
                        Add a Communication Device

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

                                                [Entry Fields]
* Device Name                                   [ds8k_hb]
* Network Type                                   diskhb
* Network Name                                   net_diskhb_01
* Device Path                                   [/dev/hdisk1]
* Node Name                                     [Zhifa]            +
```

Figure 4-9 Menu for adding heartbeat communication devices

5. Test the disk heartbeat connectivity.

   After the network and device definitions are created, we test it to make sure that the communications work properly. We use the RSCT **/usr/sbin/rsct/bin/dhb_read** command to test the validity of the diskhb connection. Table 4-2 shows the functions of the **dhb_read** command.

Table 4-2 Commands for checking heartbeat functions

Command	Action
dhb_read -p devicename	Dumps diskhb sector contents.
dhb_read -p devicename -r	Receives data over diskhb network.
dhb_read -p devicename -t	Transmits data over diskhb network.

To test the diskhb network connectivity, we set node Zhifa as the receiver, and set node Rebecca as the transmitter:

1. On node Zhifa, enter:

   ```
   dhb_read -p hdisk1 -r
   ```

2. On node Rebecca, enter:

   ```
   dhb_read -p hdisk1 -t
   ```

If the link between the nodes is operational, both nodes show `Link operating normally` (Example 4-2).

Example 4-2 Disk heartbeat link testing

```
Node Zhifa:
root@Zhifa / > dhb_read -p hdisk1 -r
DHB CLASSIC MODE
 First node byte offset: 61440
Second node byte offset: 62976
Handshaking byte offset: 65024
       Test byte offset: 64512

Receive Mode:
Waiting for response . . .
Magic number = 0x87654321
Magic number = 0x87654321
```

```
Magic number = 0x87654321
Magic number = 0x87654321
Link operating normally

Node Rebecca
root@Rebecca / > dhb_read -p hdisk1 -t
DHB CLASSIC MODE
 First node byte offset: 61440
Second node byte offset: 62976
Handshaking byte offset: 65024
       Test byte offset: 64512

Transmit Mode:
Magic number = 0x87654321
Detected remote utility in receive mode.  Waiting for response . . .
Magic number = 0x87654321
Magic number = 0x87654321
Link operating normally
```

The volume groups that are associated with the disks used for the disk heartbeating network do not have to be defined as resources within a resource group. However, if it uses the shared volume group for disk heartbeating, as in our case, then it can be defined as a resource in the resource group.

Disk heartbeat testing: Disk heartbeat testing can be done only when PowerHA is not running on the nodes.

4.1.2 Configuring the cross-site LVM disk mirroring dependency

In our environment, each node is connected to one external storage. Node Zhifa connects to an IBM DS8000 external storage, and node Rebecca connects to an IBM DS4800 external storage.

The storage is accessed through one VIOS, which has a redundant single-port HBA. Each HBA is configured with an NPIV configuration and is connected to four LPARs (another three LPARs are used for other testing). In a production implementation, it is helpful to have redundant Virtual I/O servers.

SAN zoning is configured to allow both HBAs from the VIOS in node Zhifa and node Rebecca to access the IBM DS8000 and also IBM DS4800 storage. The storage LUN from the DS8000 and the DS4000 are mapped to host group Zhifa and Rebecca. Therefore, both servers have shared disks from both storage.

In addition, we provide redundant HBAs in each server for redundancy and increase performance through dual-path access. To configure the Multipath Input/Output (MPIO) access in AIX 6.1, install the MPIO driver in each node.

Example 4-3 shows the driver.

Example 4-3 MPIO device driver

```
devices.common.IBM.mpio.rte
devices.fcp.disk.ibm.mpio.rte
devices.common.IBM.mpio.rte
```

Example 4-4 shows the disk configuration for each node after we configure the storage and installing the MPIO device driver. hdisk1 - hdisk4 are on IBM DS8000 storage, and hdisk5 - hdisk8 are on IBM DS4800 storage.

Example 4-4 Disk configuration on node Zhifa and node Rebecca

```
Node Zhifa:
root@Zhifa / > lsdev -Cc disk
hdisk0 Available          Virtual SCSI Disk Drive
hdisk1 Available 24-T1-01 IBM MPIO FC 2107
hdisk2 Available 24-T1-01 IBM MPIO FC 2107
hdisk3 Available 24-T1-01 IBM MPIO FC 2107
hdisk4 Available 24-T1-01 IBM MPIO FC 2107
hdisk5 Available 24-T1-01 IBM MPIO DS4800 Array Disk
hdisk6 Available 24-T1-01 IBM MPIO DS4800 Array Disk
hdisk7 Available 24-T1-01 IBM MPIO DS4800 Array Disk
hdisk8 Available 24-T1-01 IBM MPIO DS4800 Array Disk
hdisk9 Available 24-T1-01 IBM MPIO FC 2107

Node Rebecca:
root@Rebecca / > lsdev -Cc disk
hdisk0 Available          Virtual SCSI Disk Drive
hdisk1 Available 13-T1-01 IBM MPIO FC 2107
hdisk2 Available 03-T1-01 IBM MPIO FC 2107
hdisk3 Available 03-T1-01 IBM MPIO FC 2107
hdisk4 Available 03-T1-01 IBM MPIO FC 2107
hdisk5 Available 03-T1-01 IBM MPIO DS4800 Array Disk
hdisk6 Available 03-T1-01 IBM MPIO DS4800 Array Disk
hdisk7 Available 03-T1-01 IBM MPIO DS4800 Array Disk
hdisk8 Available 03-T1-01 IBM MPIO DS4800 Array Disk
```

We also must check the PVID of each disk to see the corresponding disk in node Zhifa with disk in node Rebecca. We need this information later when we configure disk and site dependency. Example 4-5 shows our disk PVID configuration.

hdisk1 - hdisk4 in node Zhifa and Rebecca are the same disks that come from IBM DS8000 storage, and hdisk5 - hdisk8 are the same disks in node Zhifa and Rebecca that come from IBM DS48000 storage. In the disk/site definition for cross-site LVM mirroring, we select hdisk1 - hdisk4 owned by site POK and hdisk5 - hdisk8 owned by site NY.

Example 4-5 Disk PVID in node Zhifa and node Rebecca

```
Node Zhifa:
root@Zhifa / > lspv
hdisk0          000fe411492d4136
hdisk1          000fe4110742ad37
hdisk2          000fe4110742ad9f
hdisk3          000fe4110742ae0f
hdisk4          000fe4110742ae6f
hdisk5          000fe41107345f77
hdisk6          000fe41107346f8a
hdisk7          000fe41107347b1e
hdisk8          000fe4110734851b
hdisk9          000fe41164116629

Node Rebecca:
```

```
root@Rebecca / > lspv
hdisk0          000fe401a77b45d4
hdisk1          000fe4110742ad37
hdisk2          000fe4110742ad9f
hdisk3          000fe4110742ae0f
hdisk4          000fe4110742ae6f
hdisk5          000fe41107345f77
hdisk6          000fe41107346f8a
hdisk7          000fe41107347b1e
hdisk8          000fe4110734851b
```

Before you configure the cross-site disk LVM mirroring dependencies, run the cluster discovery:

1. Enter the **smitty hacmp** command.
2. Select **Extended Configuration → Discover HACMP-related Information from Configured Nodes**.

   PVIDs: All disks must have PVIDs assigned before you run the Discover HACMP-related Information from Configured Nodes to have the complete disk information that is stored into the disk discovery file. To do this task, run the following command on each node:

   ```
   chdev -l hdiskX -a pv=yes
   ```

3. Continue to define the disk/site definition by entering the **smitty hacmp** command on the SMIT menu.
4. Select **System Management (C-SPOC) → Storage → Physical Volume → Configure Disk/Site Locations for Cross-Site LVM Mirroring → Add Disk/Site Definition for Cross-Site LVM Mirroring**.
5. Choose site **POK**, and press Enter.

As shown in Figure 4-10, we select hdisk1 - hdisk4 in node Zhifa and node Rebecca owned by site POK. Then, continue configuring the cross-site LVM mirroring disk definition for node Rebecca.

```
              Add Disk/Site Definition for Cross-Site LVM Mirroring

Ty+--------------------------------------------------------------------------+
Pr|                               Disks PVID                                 |
  |                                                                          |
  | Move cursor to desired item and press F7.                                |
* |     ONE OR MORE items can be selected.                                   |
* | Press Enter AFTER making all selections.                                 |
  |                                                                          |
  | > 000fe4110742ad37        hdisk1  Rebecca                                |
  | > 000fe4110742ad37        hdisk1  Zhifa                                  |
  | > 000fe4110742ad9f        hdisk2  Rebecca                                |
  | > 000fe4110742ad9f        hdisk2  Zhifa                                  |
  | > 000fe4110742ae0f        hdisk3  Rebecca                                |
  | > 000fe4110742ae0f        hdisk3  Zhifa                                  |
  | > 000fe4110742ae6f        hdisk4  Zhifa                                  |
  | > 000fe4110742ae6f        hdisk4  Rebecca                                |
  |   000fe41107345f77        hdisk5  Zhifa                                  |
  |   000fe41107346f8a        hdisk6  Zhifa                                  |
  |   000fe41107345f77        hdisk5  Rebecca                                |
  |   000fe41107347b1e        hdisk7  Zhifa                                  |
  |   000fe4110734851b        hdisk8  Zhifa                                  |
  |   000fe41107346f8a        hdisk6  Rebecca                                |
  |   000fe41107347b1e        hdisk7  Rebecca                                |
  |   000fe4110734851b        hdisk8  Rebecca                                |
  |                                                                          |
F9+--------------------------------------------------------------------------+
```

Figure 4-10 Disk selection for cross-site LVM disk/site definition

Current disk configuration: The Configure Disk/Site Locations for Cross-Site LVM Mirroring menu selection functions correctly only if the disk discovery file reflects the current disk configuration.

We can change the disk/site dependency later by entering the **smitty cl_xslvmm** command and selecting **Change/Show Disk/Site Definition for Cross-Site LVM Mirroring**. We can also remove the site and disk dependencies later by using the **smitty cl_xslvmm** command and selecting **Remove Disk/Site Definition for Cross-Site LVM Mirroring**.

4.1.3 Mirror pool disk

Mirror pools provide a way to group two or more disks together within a volume group. Figure 4-11 shows the diagram description of a mirror pool.

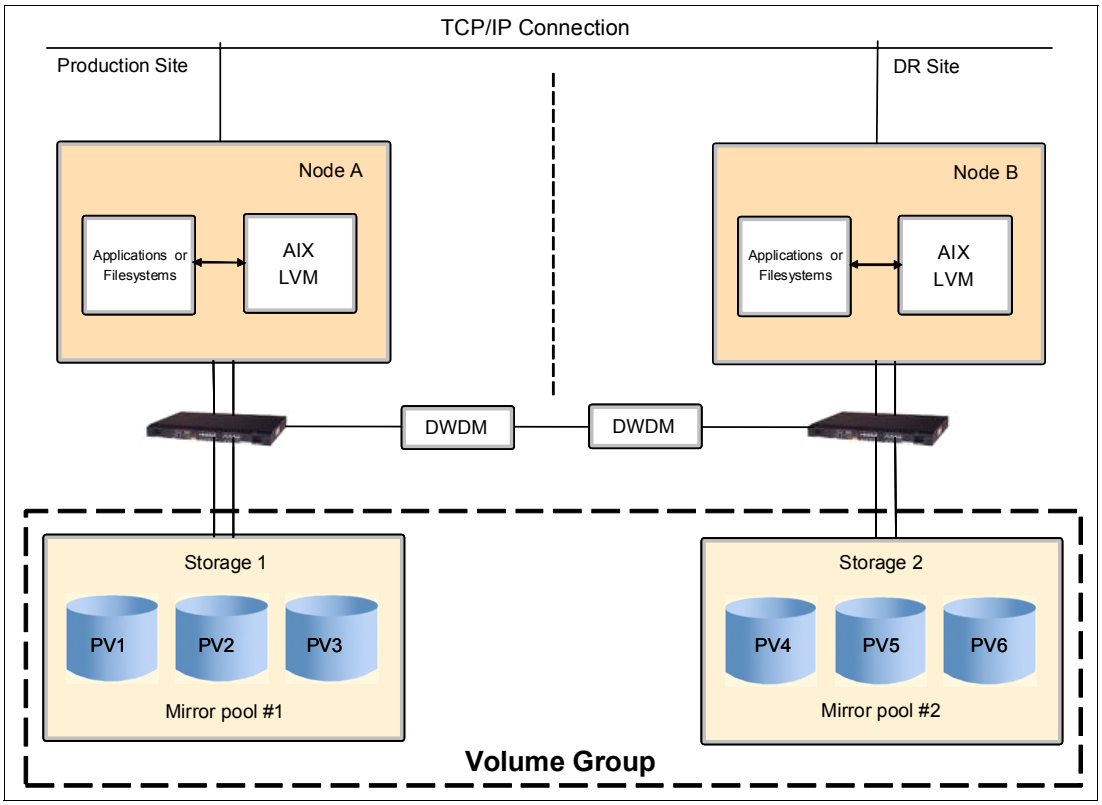


Figure 4-11 Mirror pool diagram

Figure 4-11 shows a geographically mirrored volume group where the disks at the production site are placed into mirror pool #1, and the disks at the disaster recovery site are placed into mirror pool #2. Both mirror pools are configured to become one volume group that is used for synchronous cross-site LVM mirroring.

To support a mirror pool configuration, certain parameters must be followed when you create a volume group. The mirror pool requires that volume groups enable the *super strict* feature.

The super strict function performs the following tasks:

- Checks that the local and remote physical volumes cannot belong to the same mirror pool.
- Checks for no more than three mirror pools per volume group.
- Checks that each mirror pool contains at least one copy of each logical volume:
  - When we create a logical volume, we must configure it so that each mirror pool gets a copy. However, if we create a mirror pool in a volume group where logical volumes exist, logical volume copies are not automatically created in the new mirror pool. We must create them by running the **`mirrorvg`** or **`mklvcopy`** commands.
  - Asynchronous GLVM mirroring requires a new type of logical volume for caching of asynchronous write requests. This logical volume should not be mirrored across sites. Super strict mirror pools handle this new aio_cache logical volume type as a special case.

In addition this function has the following requirements:

- We must disable the auto-on and bad-block relocation options of the volume group.
- The volume group cannot be a snapshot volume group. The volume group cannot contain active paging-space logical volumes.
- The volume group must be varied off to make mirror pool changes.
- We cannot remove or reduce an aio_cache type logical volume after it is part of the asynchronous mirroring setup.
- The rootvg volume group cannot be configured for asynchronous mirroring.

Mirror pools provide extra benefits to the mirroring function:

- Rather than having to configure individual RPV devices, mirror pools provide a convenient way for users to manage mirroring at a higher level.
- The decision of whether to mirror synchronously or asynchronously is made at the mirror pool level. Therefore, we can decide to mirror from the production site to the disaster recovery site asynchronously, and then mirror from the disaster recovery site back to the production site synchronously. This task can be accomplished by configuring the mirror pool that contains the disaster recovery site disks as asynchronous when you configure the mirror pool that contains the production site disks as synchronous.

Figure 4-12 shows a geographically mirrored volume group that uses a mirror pool configuration.

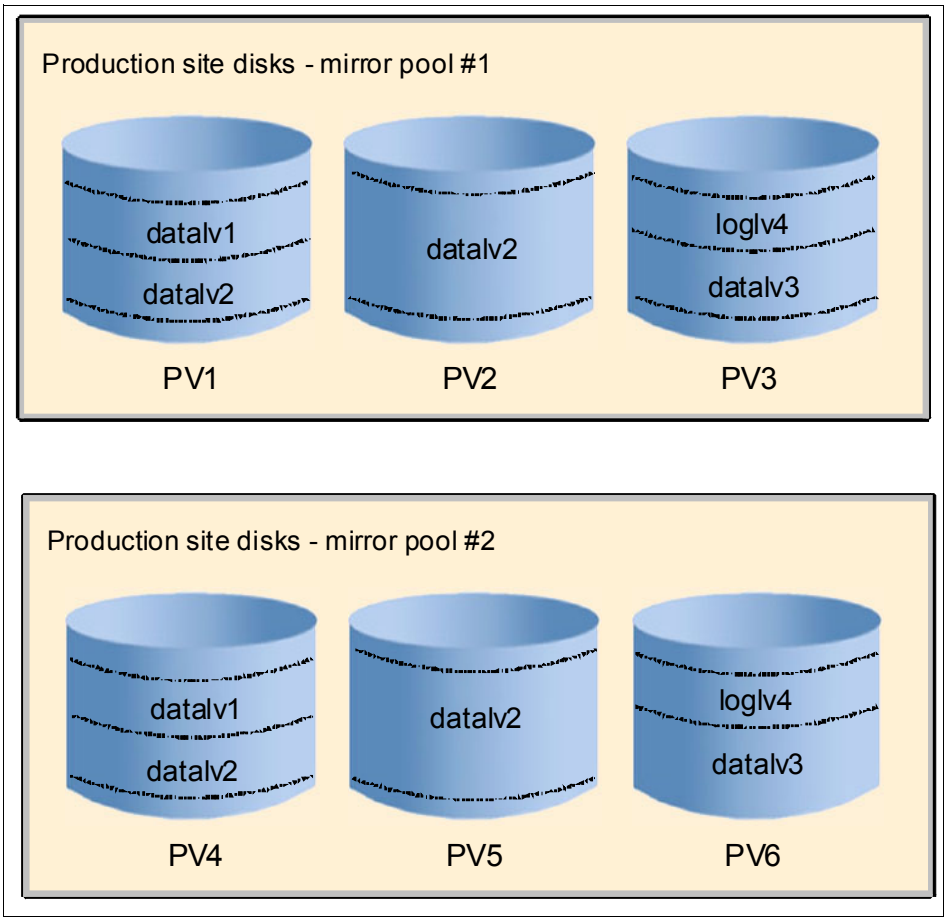


Figure 4-12 Geographically mirrored volume group

The volume group has a total of five logical volumes. The user data is stored in three logical volumes. Logical volumes datalv1, datalv2, and datalv3 contain file systems, and logical volume loglv4 contains the file system log. These four logical volumes are mirrored across both sites because they have copies in both mirror pools.

In another case, such as in GLVM asynchronous mirroring, we need to create a cache logical volume, aiocache. We can create logical volume aiocachelv1 in a local site and logical volume aiocachelv2 in a remote site. Both are used to cache asynchronous write requests. They are not mirrored across both sites.

In Figure 4-12, the volume group is varied online at the production site. Writes to the local disks in mirror pool #1 and mirror pool #2 occur synchronously.

Creating a volume group

We create two volume groups, pokvg and nyvg, which consisting of four disks for pokvg and two disks for nyvg. To create the volume groups, on the SMIT C-SPOC menus, enter the **`smitty hacmp`** command. Select **System Management (C-SPOC)** → **Storage** → **Volume Group** → **Create a Volume Group**.

Then, on the SMIT menu (Figure 4-13), select the node as the owner for the volume group.

```
                                Volume Groups

Move cursor to desired item and press Enter.

  List All Volume Groups
  Create a Volume Group
  Create a Volume Group with Data Path Devices

  Set Characteristics of a Volume Group
  Enable a Volume Group for Fast Disk Takeover or Concurrent Access

  Import a Volume Group
  Mirror a Volume Group
  Unmirror a Volume Group

  Manage Concurrent Access Volume Groups for Multi-Node Disk Heartbeat
  +--------------------------------------------------------------------------+
  |                               Node Names                                 |
  |                                                                          |
  | Move cursor to desired item and press F7.                                |
  |     ONE OR MORE items can be selected.                                   |
  | Press Enter AFTER making all selections.                                 |
  |                                                                          |
  | > Rebecca                                                                |
  | > Zhifa                                                                  |
F9+--------------------------------------------------------------------------+
```

Figure 4-13 Node selection for creating a volume group

Select both nodes for creating volume group pokvg. Then, select the physical volume for this volume group (Figure 4-14).

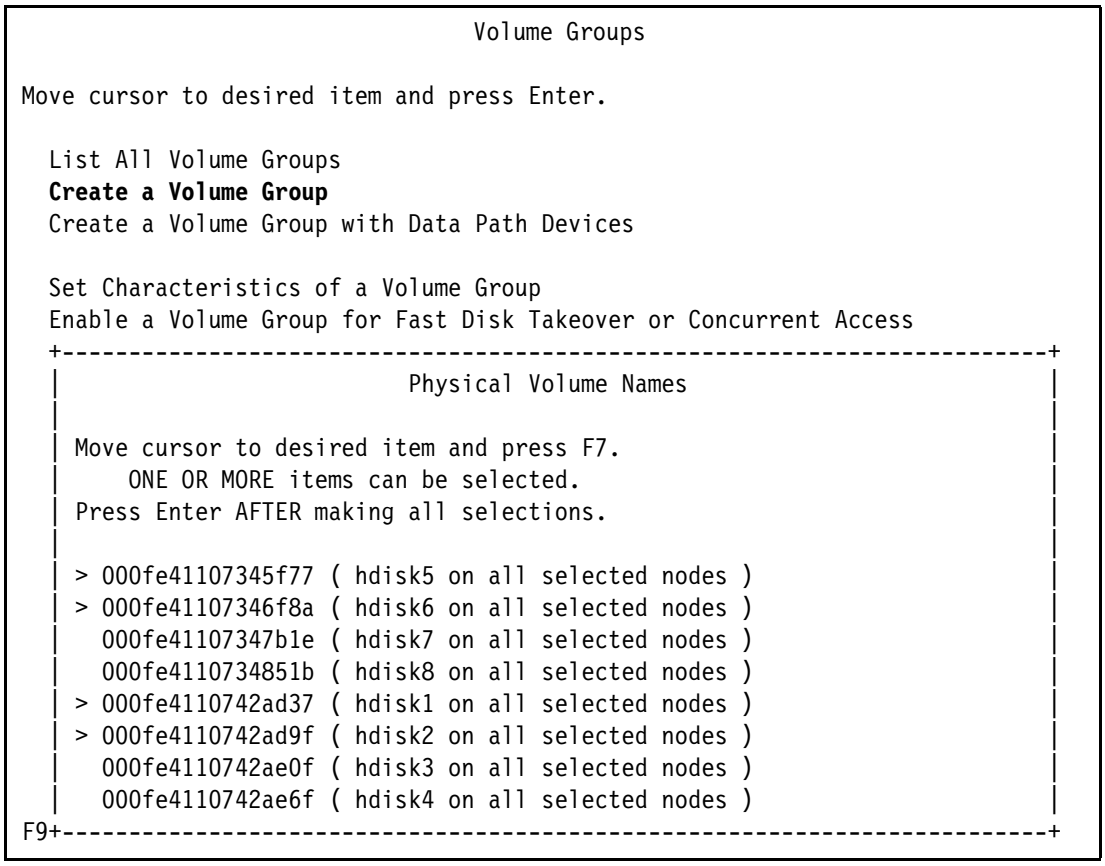


Figure 4-14 Disk selection for creating a volume group

Press F7 to select the appropriate disks. The volume group pokvg is in hdisk1 and hdisk2 in node Zhifa and is mirrored to hdisk5 and hdisk6 in node Rebecca.

After you select the physical volume, select the volume group type. Four options are available, and, in our case, we choose a scalable volume group. The scalable volume group is a prerequisite for creating a mirror pool, which we use in our scenario (Figure 4-15).

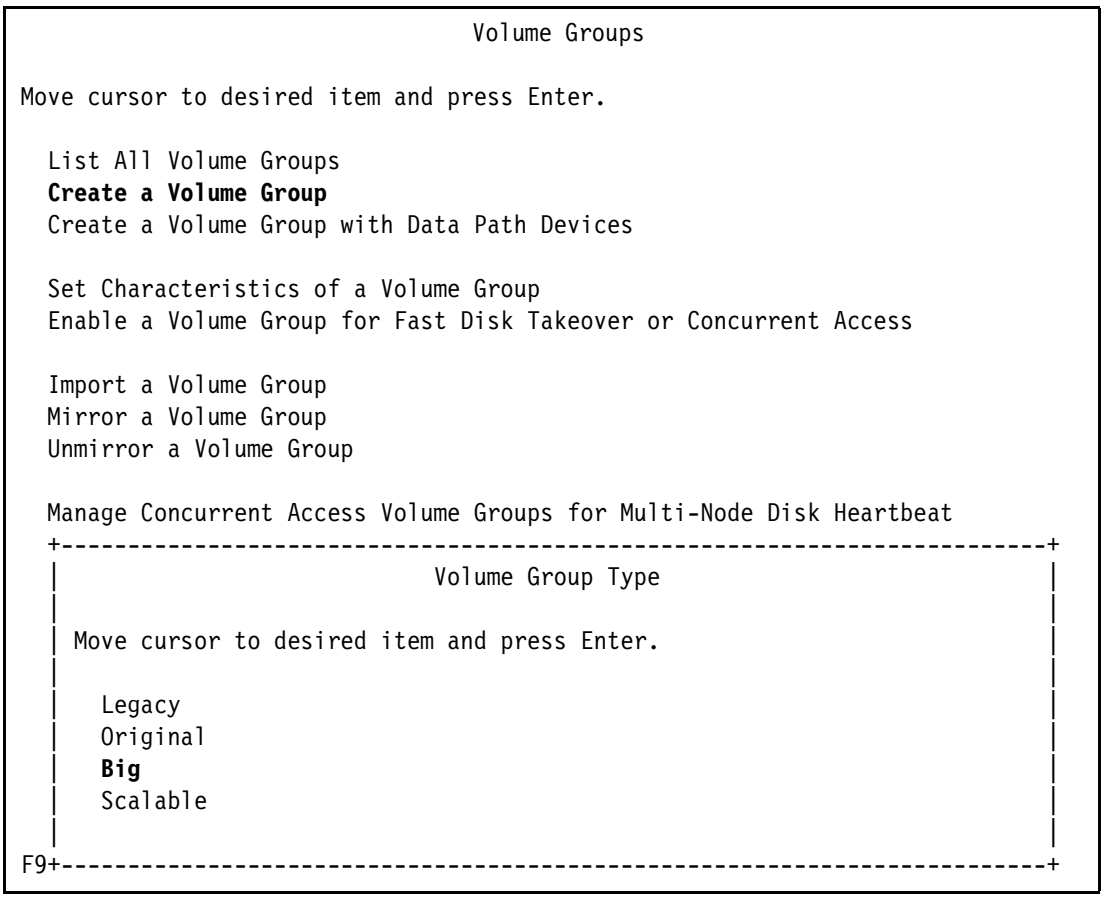


Figure 4-15 Selecting volume group type

We must choose **Scalable**, but if we chose the big volume group, as shown in Figure 4-15, we can change it to a scalable volume group by using the following AIX command:

```
# chvg -G pokvg
```

You can also use the SMIT menu to change the volume group to the scalable volume group. Run **smitty chvg**, choose the volume group name, and change the Change to scalable VG format? parameter to yes.

chvg man page: The volume group must be varied offline before you run the **chvg** command. Other considerations are described in the **chvg** man page. For more information about converting your existing volume groups to scalable VG format, see the AIX documentation.

After you select the scalable volume group, complete the volume group parameters (Figure 4-16).

```
                          Create a Big Volume Group

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

                                                      [Entry Fields]
  Node Names                                          Rebecca,Zhifa
  Resource Group Name                                 []                          +
  PVID                                                000fe41107345f77 000f>
  VOLUME GROUP name                                   [pokvg]
  Physical partition SIZE in megabytes                64                          +
  Volume group MAJOR NUMBER                           [34]                         #
  Enable Cross-Site LVM Mirroring Verification        false                       +
  Enable Fast Disk Takeover or Concurrent Access      Fast Disk Takeover or>      +
  Volume Group Type                                   Big
```

Figure 4-16 Creating a volume group

Set the mirror pool strictness volume group parameter to superstrict. If have not set it yet, you can use the SMIT menu to enable it. Run **smitty chvg**, and then choose the name of the volume groups that must be superstrict enabled (Figure 4-17).

```
                          Change a Volume Group

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

                                                      [Entry Fields]
* VOLUME GROUP name                                  pokvg
* Activate volume group AUTOMATICALLY                no                          +
    at system restart?
* A QUORUM of disks required to keep the volume      no                          +
   group on-line ?
  Convert this VG to Concurrent Capable?             no                          +
  Change to big VG format?                           no                          +
  Change to scalable VG format?                      no                          +
  LTG Size in kbytes                                 128                         +
  Set hotspare characteristics                       n                           +
  Set synchronization characteristics of stale       n                           +
   partitions
  Max PPs per VG in units of 1024                    32                          +
  Max Logical Volumes                                256                         +
  Mirror Pool Strictness                             Superstrict                 +
```

Figure 4-17 Changing a volume group to superstrict mirror pool

When you are finished creating the volume group pokvg, continue by creating the second volume group nyvg.

Creating a mirror pool disk

In our environment, we create two mirror pools in volume group pokvg. Two physical volumes at site POK are grouped in one mirror pool with the name mp_pok and another two physical volumes are grouped in a mirror pool with the name mp_ny. Run the AIX command line to create this mirror pool (Example 4-6).

Example 4-6 Creating mirror pool

```
root@Zhifa / > chpv -p mp_pok hdisk1 hdisk2
root@Zhifa / > chpv -p mp_ny hdisk5 hdisk6
```

Check the mirror pool configuration by using the **lsvg** and **lsmp** commands (Example 4-7).

Example 4-7 Checking the mirror pool configuration

```
root@Zhifa / > lsvg -P pokvg
Physical Volume   Mirror Pool
hdisk5            mp_ny
hdisk6            mp_ny
hdisk1            mp_pok
hdisk2            mp_pok

root@Zhifa / > lsmp -A pokvg
VOLUME GROUP:       pokvg               Mirror Pool Super Strict: yes
MIRROR POOL:        mp_pok              Mirroring Mode:           SYNC
MIRROR POOL:        mp_ny               Mirroring Mode:           SYNC
root@Zhifa / >
```

Example 4-7 shows that hdisk1 and hdisk2 are configured as mirror pool mp_pok, hdisk5 and hdisk6 are configured as mirror pool mp_ny, and the replication method that is used is synchronous replication.

Creating a logical volume

As described in 4.1.3, “Mirror pool disk” on page 125, we do not need to create the aio_cache logical volume for synchronous replication. We create only a jfs2 logical volume and check that the mirror pool configuration works well.

To create a jfs2 logical volume with mirror pool configuration, run the **smitty mklv** SMIT command (Figure 4-18).

```
                                Add a Logical Volume

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

[TOP]                                                   [Entry Fields]
  Logical volume NAME                                  [lv1_mp]
* VOLUME GROUP name                                     pokvg
* Number of LOGICAL PARTITIONS                        [10]                    #
  PHYSICAL VOLUME names                                []                     +
  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                      2                     +
    partition
  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?                          no                    +
  SCHEDULING POLICY for writing/reading                 parallel              +
     logical partition copies
  Enable WRITE VERIFY?                                  no                    +
  File containing ALLOCATION MAP                       []
  Stripe Size?                                         [Not Striped]          +
  Serialize IO?                                         no                    +
  Mirror Pool for First Copy                            mp_pok                +
  Mirror Pool for Second Copy                           mp_ny                 +
  Mirror Pool for Third Copy                                                  +
```

Figure 4-18 Creating a logical volume with mirror pool configuration

Consider the following parameters (Figure 4-18) when you create a logical volume with the mirror pool configuration:

- Logical volume type

  This parameter relates with the type of logical volume. We can choose jfs, jfs2, sysdump, paging, jfslog, jfs2log, boot, and aio_cache by pressing F4.

- Number of COPIES of each logical partition

  This parameter must be completed with the number of mirror pools on that volume group because each mirror pool has a copy of every logical volume on that volume group, except for the aio_cache logical volume.

- Enable BAD BLOCK relocation.

  This parameter must be set to *no*. AIX is intelligent enough to detect bad blocks and then relocate the data on the disk drives. Whenever a problem occurs on the disks, data relocation takes place automatically for data protection. In a mirror pool environment, this

feature must be set to no. The reason is that we have a group of disks, and each group of disks has every copy of the logical volume and cannot be mixed with each other.

- Mirror Pool for First Copy

  This parameter indicates the mirror pool name where the first copy of this logical volume is.

- Mirror Pool for Second Copy

  This parameter indicates the mirror pool name where the second copy of this logical volume is.

- Mirror Pool for Third Copy

  This parameter indicates the mirror pool name where the third copy of this logical volume is. You can leave it blank if you do not have the third mirror pool.

However, if you create a mirror pool in a volume group where logical volumes exist, the logical volume copies are not automatically created in the new mirror pool. You must copy them by running the **mirrorvg** or **mklvcopy** commands.

Logical volume with mirror pool configuration: Creating a logical volume with a mirror pool configuration in PowerHA 6.1 can be performed by using the AIX command line or **smitty lv**. The C-SPOC menu does not support this feature yet.

Creating file systems

We create jfs2 file systems with a mirror pool configuration from the available logical volumes as shown in Figure 4-19. By using the SMIT menu, enter the **smitty hacmp** command. Then, select **System Management (C-SPOC)** → **Storage** → **File Systems** → **Add a File Systems**.

```
                                File Systems

Move cursor to desired item and press Enter.

  List All File Systems by Volume Group
  Add a File System
  Change / Show Characteristics of a File System
  Remove a File System
  +--------------------------------------------------------------------------+
  |          Select the Logical Volume to hold the new File System           |
  |                                                                          |
  | Move cursor to desired item and press Enter. Use arrow keys to scroll.   |
  |                                                                          |
  |   # To Add a new File System to volume group pokvg,                      |
  |   # you must either chose to                                             |
  |                                                                          |
  |     Create a new Logical Volume for this File System                     |
  |                                                                          |
  |   # Or select an existing logical volume from the list below             |
  |   # LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MO |
  |     lv1_mp              jfs2       10      20      2    closed/syncd  N/ |
F9+--------------------------------------------------------------------------+
```

Figure 4-19 Selecting the previous logical volume for creating a file system

Choose the volume group where the logical volume resides, and then choose the file system type. In this environment, we choose Enhanced Journal File Systems (Figure 4-20).

```
  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.

                                                      [Entry Fields]
  Resource Group                                     pokrg
* Node Names                                         Rebecca,Zhifa
  Logical Volume name                                lv1_mp
  Volume Group                                       pokvg

* MOUNT POINT                                       [/data_zhifa]            /
  PERMISSIONS                                        read/write              +
  Mount OPTIONS                                      []                      +
  Block Size (bytes)                                 4096                    +
  Inline Log?                                        yes                     +
  Inline Log size (MBytes)                           []                       #
  Logical Volume for Log                                                     +
  Extended Attribute Format                          Version 1               +
  Enable Quota Management?                           no                      +
```

Figure 4-20 Creating file systems with mirror pool configuration

Set the inline log parameter to `yes`. The inline log parameter distributes the log over all disks that are involved in the logical volume. You need to also apply this parameter in a mirror pool configuration.

After you create the file system, check whether the file system is already distributed within the mirror pools with the **lsvg** command. Figure 4-21 shows that the file system *data_zhifa* resides in separate mirror pools and in separate physical volumes.

```
root@Zhifa / > lsvg -M pokvg
pokvg
hdisk5:1-637
hdisk6:1-134
hdisk6:135      lv1_mp:1:2
hdisk6:136      lv1_mp:2:2
hdisk6:137      lv1_mp:3:2
hdisk6:138      lv1_mp:4:2
hdisk6:139      lv1_mp:5:2
hdisk6:140      lv1_mp:6:2
hdisk6:141      lv1_mp:7:2
hdisk6:142      lv1_mp:8:2
hdisk6:143      lv1_mp:9:2
hdisk6:144      lv1_mp:10:2
hdisk6:145-637
hdisk1:1-637
hdisk2:1-134
hdisk2:135      lv1_mp:1:1
hdisk2:136      lv1_mp:2:1
hdisk2:137      lv1_mp:3:1
hdisk2:138      lv1_mp:4:1
hdisk2:139      lv1_mp:5:1
hdisk2:140      lv1_mp:6:1
hdisk2:141      lv1_mp:7:1
hdisk2:142      lv1_mp:8:1
hdisk2:143      lv1_mp:9:1
hdisk2:144      lv1_mp:10:1
root@Zhifa / > lsvg -l pokvg
pokvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lv1_mp              jfs2       10      20      2    open/syncd    /data_zhifa
```

Figure 4-21 Checking file system distribution

4.1.4 Configuring a resource group

After you create the cluster topology and the LVM with mirror pool configuration, create a cluster resource group. We create two resource groups, pokrg and nyrg. Each resource group owns one application server and one volume group.

Creating a cluster resource group

Run the SMIT **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resource Group Configuration** → **Add a Resource Group**.

Completet the parameters that are shown in Figure 4-22. Certain parameters are the same parameters that are used in the local cluster configuration. The additional parameter for the cross-site LVM mirroring cluster is the *Inter-Site Management Policy*. This parameter relates with the resource group recovery policy to allow or disallow the cluster manager to move a resource group to another site in case the resource group goes into an error state.

```
                        Add a Resource Group (extended)

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

                                                      [Entry Fields]
* Resource Group Name                                 [pokrg]

  Inter-Site Management Policy                        [Prefer Primary Site]   +
* Participating Nodes from Primary Site               [Zhifa]                 +
  Participating Nodes from Secondary Site             [Rebecca]               +

  Startup Policy                                       Online On Home Node O> +
  Failover Policy                                      Failover To Next Prio> +
  Failback Policy                                      Failback To Higher Pr> +
```

Figure 4-22 Creating a resource group

This option has the following features:

Ignore This is default selection and ignores the site dependency settings for the resource group.

Prefer Primary Site The resource group may be assigned to be taken over by multiple sites in a prioritized manner. When a site fails, the active site with the highest priority acquires the resource. When the failed site rejoins, the site with the highest priority acquires the resource.

Online On Either Site The resources group may be acquired by any site in its resource chain. When a site failure occurs, the resource group is acquired by the highest priority standby site. When the failed site rejoins, the resource group remains with its new owner.

Online On Both Sites The resource group is acquired by both sites. This selection defines the concurrent capable resource group.

After you create the resource group pokrg, create the second resource group nyrg.

Creating a cluster application server

In HACMP terms, an *application server* is a cluster resource that is made highly available by the HACMP software. An application server has a start script and a stop script. The start script starts the application server. The **stop** script stops the application server so that the application resource can be released, allowing the second node to take it over and restart the application.

Create the cluster application server by using the SMIT menu (Figure 4-23). Enter the **smitty hacmp** command. Then, select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resource Configuration → Configure HACMP Applications Servers → Configure HACMP Application Servers → Add an Application Server** (Figure 4-23).

Enter the application server name and the link to the start script file. Then, continue with the creation of the second application server, ny_app.

```
                          Add Application Server

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

                                                      [Entry Fields]
* Server Name                                        [pok_app]
* Start Script                                       [/IBM/start_pok.sc]
* Stop Script                                        [/IBM/stop_pok.sc]
  Application Monitor Name(s)
```

Figure 4-23 Creating cluster application server

Creating a cluster service IP label or address

Creating a cluster service IP label or address is an optional step. It makes the resource group's IP-service highly available to give the clients the possibility of always connecting to the same IP address.

We use a SMIT command to create a cluster service IP label. Enter the **smitty hacmp** command. Then, select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resource Configuration → Configure HACMP Service IP Labels / Addresses → Add a Service IP Label / Address**.

The following two options for creating the service IP label/address are displayed:

- Configurable on multiple nodes
- Bound to a single node

  A node-bound service IP label is a specific type of service IP label that is configured on a non-aliased network. Therefore, a network must first be configured to use IP address takeover (IP replacement).

  These IP labels do not float with a resource group, but they are kept highly available on the node to which they are assigned.

  Node-bound service IP labels can be useful for the following purposes:

  – A concurrent resource group might have node-bound service IP addresses configured for it.

  – A node-bound service IP label can also be used for administrative purposes. Several of its functions can be achieved by using other capabilities in PowerHA (such as using persistent service IP labels).

In this scenario, we choose to configure on multiple nodes.

Figure 4-24 shows the SMIT menu after we choose the service IP label/address mode.

```
    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.

                                                      [Entry Fields]
* IP Label/Address                                    zhifa_svc             +
  Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.255.0]
* Network Name                                        net_ether_01
  Alternate Hardware Address to accompany IP Label/A []
  ddress
  Associated Site                                     ignore                +
```

Figure 4-24 Creating a cluster service IP label/address

Then, continue with creating the second service IP label/address for nyrg.

Changing resource group properties

Changing the resource group properties integrates the application server and the service IP label that was created into the resource group. On the SMIT menu, enter the **smitty hacmp** command. Select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resource Group Configuration** → **Change/Show Resources and Attributes for a Resource Group**. Choose the resource group.

Figure 4-25 shows the parameters to enter for the integration, such as service IP labels/addresses, application servers, and volume groups.

```
          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                                pokrg
  Inter-site Management Policy                       Prefer Primary Site
  Participating Nodes from Primary Site              Zhifa
  Participating Nodes from Secondary Site            Rebecca

  Startup Policy                                     Online On Home Node O>
  Failover Policy                                    Failover To Next Prio>
  Failback Policy                                    Failback To Higher Pr>
  Failback Timer Policy (empty is immediate)         []                     +

  Service IP Labels/Addresses                        [zhifa_svc]            +
  Application Servers                                [pok_app]              +

  Volume Groups                                      [pokvg ]               +
  Use forced varyon of volume groups, if necessary   true                   +
  Automatically Import Volume Groups                 false                  +
  Default choice for data divergence recovery        ignore                 +
      (Asynchronous GLVM Mirroring Only)
  Allow varyon with missing data updates?            true                   +
```

Figure 4-25 Resource group parameter

After you change the property of the first resource group, continue with the second resource group, nyrg.

Cluster verification and synchronization

Check that the cluster configuration is correct and that it works. The cluster verification and synchronization process checks the entire cluster configuration, such as topology, logical volumes, and resource groups.

Run SMIT to do this task by entering the **`smitty hacmp`** command. Then, select **Extended Configuration** → **Extended Verification and Synchronization**.

If the output of this cluster verification and synchronization is successful, start the cluster.

4.2 Testing cross-site LVM mirroring cluster

We created and configured a cluster topology with its resources. Now test the cluster to verify that it works properly.

4.2.1 Adding file systems

Check whether the mirror pool configuration is working properly. We create a file system, called testfs, by using the C-SPOC menu (Figure 4-26). Check on which physical volume the new file system is located. Enter the **smitty hacmp** command. Select **System Management (C-SPOC) → Storage → File Systems → Add a File System**. Choose the corresponding volume group and file system type.

```
                    Add an Enhanced Journaled File System

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

                                                      [Entry Fields]
  Resource Group                                     pokrg
* Node Names                                         Rebecca,Zhifa
  Volume group name                                  pokvg

  SIZE of file system
         Unit Size                                   M                       +
*        Number of units                            [10]                     #
* MOUNT POINT                                       [/test]                  /
  PERMISSIONS                                        read/write              +
  Mount OPTIONS                                      []                      +
  Block Size (bytes)                                 4096                    +
  Inline Log?                                        yes                     +
  Inline Log size (MBytes)                           []                       #
  Logical Volume for Log                                                     +
  Extended Attribute Format                          Version 1               +
  Enable Quota Management?                           no                      +
```

Figure 4-26 Adding a file system by using C-SPOC

As shown in Figure 4-26, enter the necessary parameters, and then press Enter. In this case, the command fails (Figure 4-27).

```
                                COMMAND STATUS

Command: failed        stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Zhifa: 0516-1829 mklv: Every mirror pool must contain a copy of
Zhifa:  the logical volume.
Zhifa: 0516-822 mklv: Unable to create logical volume.
Zhifa: cl_rsh had exit code = 1, see cspoc.log and/or clcomd.log for more
inform
ation
Error creating logical volume
Exiting due to errors.
User action required to correct or complete changes.
```

Figure 4-27 Error message when you create file systems by using the C-SPOC menu

Also when trying to create a file system by using the **smitty jfs2** command with *Add an Enhanced Journaled File System*, the result fails with the message shown in Figure 4-28.

```
                                COMMAND STATUS

Command: failed        stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

0516-1829 mklv: Every mirror pool must contain a copy of
        the logical volume.
0516-822 mklv: Unable to create logical volume.
crfs: Cannot create logical volume for log device.
```

Figure 4-28 Error message when you create a file system by using the smitty jfs2 menu

Finally, we found that we must create the *logical volume* first by using AIX commands or the **smitty mklv** menu. Then, we can continue with the task of creating file systems by using the cluster C-SPOC menu.

Mirror pool configuration: In a mirror pool configuration, create the logical volume by using AIX command or **smitty mklv** first. Then, continue by creating file systems by using the cluster C-SPOC menu. We cannot create a file system directly by using the C-SPOC menu.

Check the distribution of new file systems by using the **lsvg -m pokvg** AIX command. The result shows that the new file system is created in hdisk1 and hdisk5, where these two disks are part of a separate mirror pool, which means that the mirror pool configuration works well.

4.2.2 Changing a file system size

The next test is to change the file system size of *testfs*. We tried to enlarge the file system by using the C-SPOC menu, but the result fails (Figure 4-29). We must enter the resource group name. However, when in the C-SPOC menu, that parameter cannot be changed because it is automatically selected by the system.

```
 COMMAND STATUS

Command: failed        stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

cl_chfs: _get_rgnodes: A resource group must be specified.
```

Figure 4-29 Error message when you change a file system size by using the C-SPOC menu

PowerHA 6.1 SP1 APAR: This APAR is a known PowerHA 6.1 SP1 APAR:

IZ70894 C-SPOC CHANGE/SHOW JFS2 FAILS WITH "CL_CHFS: _GET_RGNODES"

Then, change the file system size by using the **smitty chjfs2** command, and add the capacity. Run this command on the node where the corresponding file systems is mounted at

the time. The result is that the file system size increased and it also expanded to a different mirror pool. By using this method, the other node in the cluster recognizes that the file system has changed, and we do not need to resynchronize the cluster to update the other node.

4.2.3 Moving cluster resource group

In this test, we move the cluster resource group, pokrg, from node Zhifa to node Rebecca without stopping the cluster. Run the SMIT **smitty hacmp** command. Then, select **System Management (C-SPOC) → Resource Groups and Applications → Move Resource Group to Another Site / Node → Move Resource Groups to Another Node**. Choose the resource group name and the destination node.

The result, the resource group pokrg, is activated on site NY and node Rebecca. We run the same command to move the resource group back from node Rebecca to node Zhifa. The result, resource group pokrg, is acquired by node Zhifa, and the applications become active.

Run the **/usr/es/sbin/cluster/utilities/clRGinfo** command to check the status of the resource group before and after you move the resource group (Figure 4-30).

```
Before moving the resource group:
root@Zhifa /IBM > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
pokrg          ONLINE                       Zhifa@POK
               ONLINE Rebecca@NY

nyrg           ONLINE                       Rebecca@NY
               ONLINE Zhifa@POK

After moving the resource group:
root@Rebecca / > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
pokrg          OFFLINE                      Zhifa@POK
               ONLINE                       Rebecca@NY

nyrg           ONLINE                       Rebecca@NY
               OFFLINE                      Zhifa@POK
```

Figure 4-30 Checking the resource group status

4.2.4 Node failure

This test simulates one node that is suddenly down. We performed this test by using the **halt -q** AIX command in node Zhifa. This command stops the processor and powers off the server, similar to pulling out the power cable from the server.

It takes around two minutes for the cluster to fail over from node Zhifa to node Rebecca. The resource group pokrg, including its volume group, IP services, and application, is ready in node Rebecca within two minutes. The time for takeover varies because it depends on the cluster configuration (for example, number of file systems, file systems size, and length of time for application to be started).

You can check the status of the resource group before and after node failure by entering the **/usr/es/sbin/cluster/utilities/clRGinfo** command.

4.2.5 Storage connection failure

The storage connection failure test is performed to check the cluster behavior when the storage connection from the node to the storage is lost. Loss of connection to the storage can happen because of an adapter failure, link failure, or SAN switch failure.

In our testing, we simulate the case in which the connection from node Zhifa to the DS8000 storage and the DS4800 storage fail. We remove the Fibre Channel ports of node Zhifa to the DS8000 and the DS4800 from the zoning configuration in the SAN switch, which causes the connection from the node to the storage to be broken suddenly.

You can also do this test by removing the Fibre Channel adapter from the server or pulling out the Fibre Channel cable from the SAN switch. However, this test cannot be applied in our environment because we use NPIV configuration, and each adapter and Fibre Channel cable is also shared by other partitions.

The result of our test showed that, when the Fibre Channel ports of node Zhifa are removed from zoning configuration in the SAN switch, node Rebecca detects the heartbeat failure in a few seconds, but does not take it over automatically. The status of the resource group *pokrg* is still available in node *Zhifa* for around 5 minutes, and during that time the application appears to be hung and the users cannot write or read to the disks.

When recovering from a disk storage subsystem failure, I/O to each LUN might stall for more than a minute. Please note that AIX can process failures of multiple LUNs simultaneously, but in the test 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 failures:

- 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 (as was the workload used in the testing described above), then to help AIX process failures of multiple LUNs simultaneously, stripe logical volumes and file systems across LUNs using the **-S** flag on the **mklv** command. Please note that it will not help to configure logical volumes with PP striping by specifying **-e x** on the **mklv** command. If PP striping is used with such a workload, AIX is still likely to process a disk storage subsystem failure one LUN at a time.

It takes a few minutes for node Rebecca to ensure that there is a failure with the storage in node Zhifa and then perform the takeover. It takes around five minutes to complete the takeover process and make the resource group pokrg, including the application, IP, and volume group, available in node Rebecca.

You can check the status of the resource group before and after the storage connection failure by entering the **/usr/es/sbin/cluster/utilities/clRGinfo** command.

4.2.6 Storage failure

The storage failure test is performed to check the cluster behavior when there is a failure of storage at one site. We simulate this test by removing all connections from node *Zhifa* and node *Rebecca* to the DS8000 storage.

To complete this test, remove the SAN switch zoning configuration, which causes the sudden total loss of connection from all nodes to the DS8000 storage. The result is that the

applications continue to work without interruption and the volume groups and file systems remain available. After the failure, check the availability of the disks and the status of the logical volume copy synchronization. The hdisk1 and hdisk2 from the DS8000 storage are marked as *missing*. The status of the logical volume that uses these disks is *stale* (Example 4-8).

Example 4-8 Status of hdisk after one storage failure

```
root@Zhifa / > lsvg -o
pokvg
mikevg
rootvg

root@Zhifa / > lsvg -p pokvg
pokvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk5            active            637         226         128..00..00..00..98
hdisk6            active            637         631
128..121..127..127..128
hdisk1            missing           637         226         128..00..00..00..98
hdisk2            missing           637         631
128..121..127..127..128

root@Zhifa / > lsvg -l pokvg
pokvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lv_1                jfs2       6       12      2    open/stale    /pok_data1
mplv2               jfs2       6       12      2    open/stale    /data2
mikeylvm            jfs        400     800     2    closed/syncd  N/A
test_lv             jfs2       5       10      2    open/stale    /testfs
```

After this test, rezone the SAN switch to make the DS8000 storage available again. Then, use the C-SPOC to synchronize and make all hdisk devices available. Enter the **smitty hacmp** command. Select **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Synchronize LVM Mirrors** → **Synchronize by Volume Group**. Then, select the appropriate VG.

When you check the status of the logical volume copy in all volume groups after synchronization, the result is in the `syncd` status (Example 4-9).

Example 4-9 Synchronization and disk status after the storage synchronization

```
Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

Zhifa: Mar 16 2010 18:21:39 Starting execution of /usr/es/sbin/cluster/events/ut
ils/cl_disk_available with parameters: -s hdisk1 hdisk2
Zhifa: cl_fscsilunreset[969]: ioctl SCIOLSTART id=0X660200 lun=0X40A040000000000
0: Invalid argument
Zhifa: cl_fscsilunreset[969]: ioctl SCIOLSTART id=0X660200 lun=0X40A040010000000
0: Invalid argument
cl_syncvg: Volume group pokvg has the disks in missing state. Trying to activate
 them

root@Zhifa / > lsvg -p pokvg
```

```
pokvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk5            active            637         226         128..00..00..00..98
hdisk6            active            637         631
128..121..127..127..128
hdisk1            active            637         226         128..00..00..00..98
hdisk2            active            637         631
128..121..127..127..128
root@Zhifa / > lsvg -l pokvg
pokvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lv_1                jfs2       6       12      2    open/syncd    /pok_data1
mplv2               jfs2       6       12      2    open/syncd    /data2
mikeylvm            jfs        400     800     2    closed/syncd  N/A
test_lv             jfs2       5       10      2    open/syncd    /testfs
root@Zhifa / >
```

4.2.7 Site failure

In the test, we simulate the POK site failure by simultaneously crashing the node Zhifa and disconnecting all the connections to the DS8000 storage. We run the AIX **`halt -q`** command in node Zhifa. At the same time, we remove the connection to the DS8000 external storage in the SAN switch zoning configuration. This action causes a server and storage failure at the same time at site POK.

The cluster at site NY detects a site failure and performs a takeover to site NY. Then, it activates the resources on node Rebecca. In this test, because the DS8000 is unreachable, takeover forces a varyon to the volume group. After 3 minutes, all resources are available on node Rebecca, including the volume group, IP services, and application in the surviving disk copy (DS4800 storage).

After this test, we reconnected the DS8000 disk subsystem so that the disk resources are available again and power on the node Zhifa. We used the C-SPOC Synchronize Shared LVM Mirrors option, which automatically makes all hdisk devices available and synchronizes all logical volumes. Run the SMIT **`smitty hacmp`** command. Select **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Synchronize LVM Mirrors** → **Synchronize by Volume Group**. Then, select the appropriate VG.

After we verify that the data is back in synchronization, we move a resource group back to node Zhifa at the POK site. The resources become available on node Zhifa, and the application is started.

4.3 Maintaining cross-site LVM mirroring cluster

You must properly maintain the configuration of your cluster. In general, use C-SPOC to maintain the storage management in our environment. However, you must know how to do this task correctly because parts of C-SPOC are not cross-site aware. This procedure entails the following most common tasks:

- Creating a volume group
- Adding volumes into an existing volume group
- Adding new logical volumes
- Adding more space to an existing logical volume

- Adding a file system
- Increasing the size of a file system

4.3.1 Creating a volume group

To create a volume group in a cross-site LVM mirroring, see “Creating a volume group” on page 127.

4.3.2 Adding and removing volumes into an existing volume group

When you add disks into an existing volume group, add them in pairs, one disk for each site. Ensure that the PVID of the disks are known to each system, run the discovery process, and define the disks to their appropriate sites, as described in 4.1.2, “Configuring the cross-site LVM disk mirroring dependency” on page 121.

In our example, we want to add hdisk4 from the DS8000 storage to volume group pokvg. Run the SMIT **smitty hacmp** command. Select **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Set Characteristic of a Volume Group** → **Add a Volume to a Volume Group**. Then, select the appropriate volume group, and press Enter.

You then see a list of physical volumes. We add hdisk4 to volume group nyvg (Figure 4-31).

```
                    Set Characteristics of a Volume Group

Move cursor to desired item and press Enter.

  Add a Volume to a Volume Group
  Change/Show characteristics of a Volume Group
  Remove a Volume from a Volume Group
  Enable/Disable a Volume Group for Cross-Site LVM Mirroring Verification
+--------------------------------------------------------------------------+
  |                          Physical Volume Names                         |
  |                                                                        |
  | Move cursor to desired item and press Enter.                           |
  |                                                                        |
  |   000fe4110734851b ( hdisk8 on all selected nodes )                    |
  |   000fe4110742ae6f ( hdisk4 on all selected nodes )                    |
F9+--------------------------------------------------------------------------+
```

Figure 4-31 Selecting a disk to add to a shared volume group

After you choose the appropriate disks, verify the fields in the final menu, and press Enter to add the additional disk to the volume group process. You must have a pair of disks at another site to enable the mirroring policy. You also must add one more disk from the storage at the second site. In this case, we add hdisk8 from the DS4800 storage at site NY as a pair of hdisk4.

To remove a volume from a volume group, enter **smitty hacmp**. Select **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Set Characteristic of a Volume Group** → **Remove a Volume from a Volume Group**. Then, select the appropriate volume group.

Then, choose the appropriate disk to remove (Figure 4-32). Verify the fields in the final menu, and press Enter to remove the disks from a volume group.

```
                    Set Characteristics of a Volume Group

Move cursor to desired item and press Enter.

  Add a Volume to a Volume Group
  Change/Show characteristics of a Volume Group
  Remove a Volume from a Volume Group
  Enable/Disable a Volume Group for Cross-Site LVM Mirroring Verification
+--------------------------------------------------------------------------+
  |                          Physical Volume Names                         |
  |                                                                        |
  | Move cursor to desired item and press Enter.                           |
  |                                                                        |
  |   Zhifa   hdisk3                                                       |
  |   Zhifa   hdisk4                                                       |
  |   Zhifa   hdisk7                                                       |
  |   Zhifa   hdisk8                                                       |
F9+--------------------------------------------------------------------------+
```

Figure 4-32 Removing a disk from a shared volume group

Repeat this process to remove the disk pair that is in another site.

4.3.3 Adding new logical volumes

To add a new logical volume in a cross-site LVM mirroring with mirror pool configuration, see “Creating a logical volume” on page 131.

4.3.4 Adding space to an existing logical volume

Allocate more space properly to maintain the mirrored copies at each site when you add extra space to an existing logical volume. To add more space, enter the **smitty hacmp** command. Then, select **System Management (C-SPOC)** → **Storage** → **Logical Volumes** → **Set Characteristics of a Logical Volume** → **Increase the Size of a Logical Volume**

Choose the appropriate volume group and logical volume in the list that is displayed. Finally, increase the size of a shared logical volume (Figure 4-33).

```
                    Increase the Size of a Logical Volume

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

                                                      [Entry Fields]
  Volume Group Name                                  pokvg
  Resource Group Name                                pokrg
  LOGICAL VOLUME name                                test_lv

  Reference node
* Number of ADDITIONAL logical partitions            [3]                    #
  PHYSICAL VOLUME names
  POSITION on physical volume                        outer_middle           +
  RANGE of physical volumes                          minimum                +
  MAXIMUM NUMBER of PHYSICAL VOLUMES                 [1024]                 #
    to use for allocation
  Allocate each logical partition copy               yes                    +
    on a SEPARATE physical volume?

  File containing ALLOCATION MAP                     []                     /
```

Figure 4-33 Increasing the size of a shared logical volume

After you add the space, verify that the partition mapping is correct by running the **lslv -m lvname** AIX command. Example 4-10 shows the output of this command.

Example 4-10 Checking the logical volume disk mapping in a shared volume group

```
root@Zhifa /IBM > lslv -m test_lv
test_lv:/testfs
LP    PP1  PV1               PP2  PV2               PP3  PV3
0001  0129 hdisk1            0129 hdisk5
0002  0130 hdisk1            0130 hdisk5
0003  0131 hdisk1            0131 hdisk5
0004  0132 hdisk1            0132 hdisk5
root@Zhifa /IBM >
```

4.3.5 Adding a file system

To add a file system in a cross-site LVM mirroring with mirror pool configuration, see 4.2.1, “Adding file systems” on page 140.

4.3.6 Increasing the size of a file system

To increase the size of a file system in cross-site LVM mirroring, see 4.2.2, “Changing a file system size” on page 141.

Part 3

Extended distance disaster recovery short overview

This part describes storage integration features and implementation options available with the IBM PowerHA SystemMirror Enterprise Edition for disaster recovery across sites.

This part includes the following chapters:

5

Configuring PowerHA SystemMirror Enterprise Edition with Metro Mirror and Global Mirror

The PowerHA Enterprise Edition-based SAN Volume Controller has several copy services. This chapter explains the steps to plan, configure, and test the disaster recovery solution.

This chapter includes the following sections:

- Scenario description
- Planning and prerequisites overview
- Installing and configuring PowerHA Enterprise Edition for SAN Volume Controller
- Adding and removing disks to PowerHA Enterprise Edition for SAN Volume Controller
- Testing PowerHA Enterprise Edition with SAN Volume Controller
- Troubleshooting PowerHA Enterprise Edition for SAN Volume Controller

5.1 Scenario description

To configure and test the PowerHA Enterprise Edition with SVC, we use a scenario in which we implemented a four-node cluster in two sites (Figure 5-1).

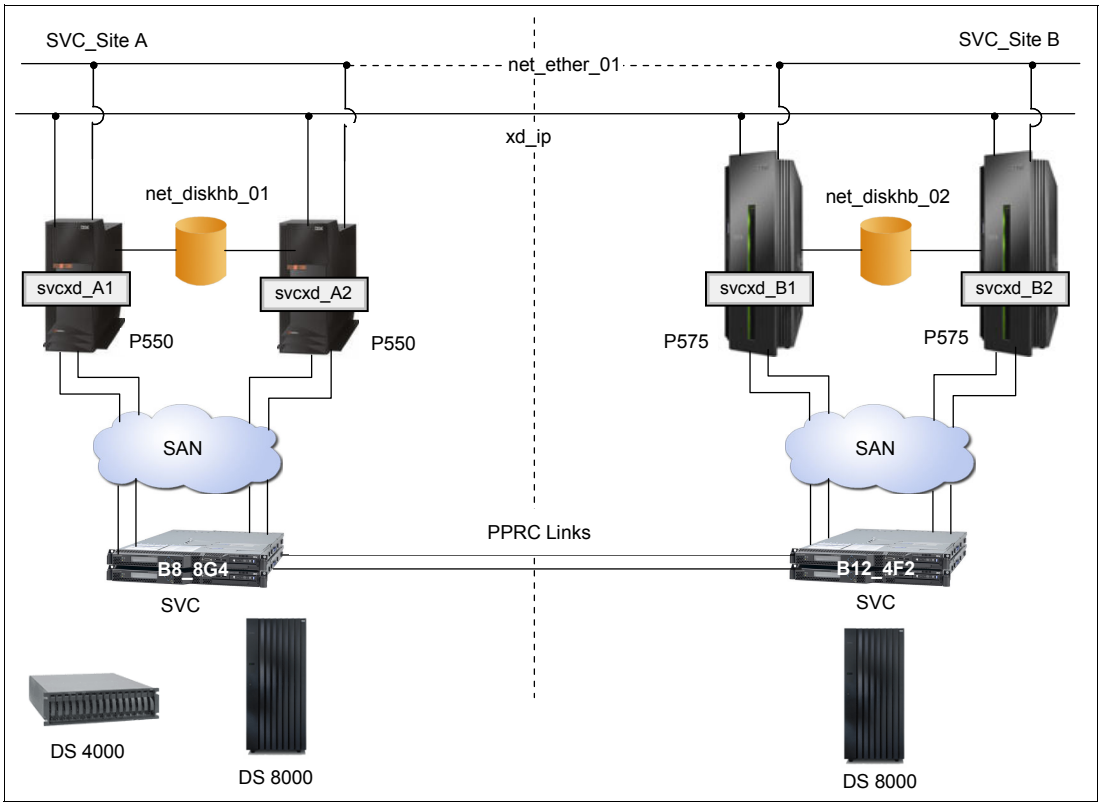


Figure 5-1 PowerHA Enterprise Edition for the SAN Volume Controller scenario

The svc_sitea includes the following components:

- svcxd_a1 and svcxd_a2 nodes
- A SAN Volume Controller B8_8G4
- A DS4000 for boot and DS8000 to use in PPRC scenarios

The svc_siteb includes the following components:

- svcxd_b1 and svcxd_b2 nodes
- A SAN Volume Controller B12_4F2
- A DS8000 to use in PPRC scenarios

Each site contains three networks definitions:

- net_XD_ip_01 is used with RSCT protocols, heartbeating, and client communication.
- net_ether_01 is used for boot, persistent, and service IP address.
- net_diskhb_01 is used between nodes in svc_sitea, and net_diskhb_02 is used between nodes in svc_siteb for disk heartbeating.

Multiple networks: To avoid site isolation and split-brain situations, have multiple networks between the sites.

In this scenario, each site has seven disks that are defined through each SAN Volume Controller cluster. The svc_sitea site uses four disks for a Metro Mirror PPRC configuration, and the svc_siteb site uses three disks for a Global Mirror PPRC configuration (Figure 5-2).

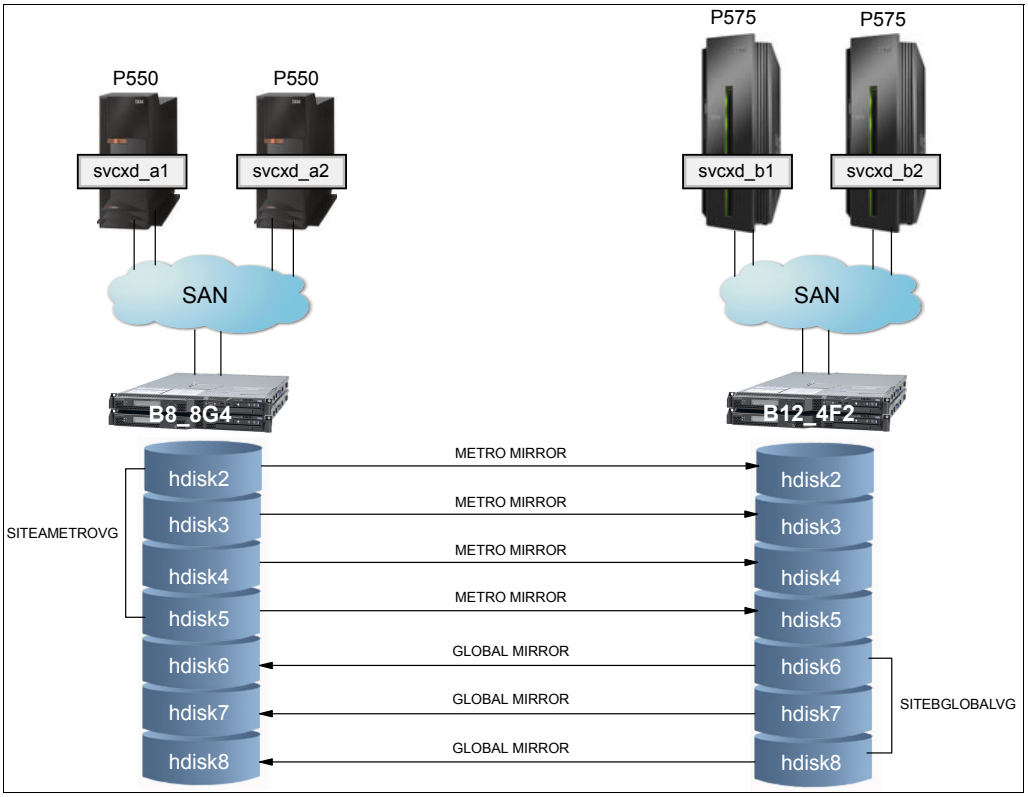


Figure 5-2 SVC PPRC configuration scenario

We configure a mutual takeover cluster between svc_sitea and svc_siteb. In svc_sitea, we have a Metro Mirror resource group, and in svc_siteb, we have Global Mirror resource group configuration.

Considerations for Live Partition Mobility capability

By using the IBM PowerVM Live Partition Mobility feature, you can migrate running AIX, applications, and data from one physical server to another server without disrupting infrastructure services. A prerequisite to use a Live Partition Mobility environment is that both servers must be managed by the same Hardware Management Console (HMC). HMC Version 7 Release 3.4 supports a source server that is managed by one HMC and a destination server that is managed by a different HMC.

The operating system, application, and data from mobile partitions must be on virtual storage on an external storage subsystem. No physical adapters can be used, and all resources must be virtualized by using one or more Virtual I/O Servers (VIOS).

Reference: For more information about how to implement Live Partition Mobility, see *IBM PowerVM Live Partition Mobility*, SG24-7460.

In our scenario, the nodes from site svc_sitea (svcxd_a1 and svcxd_a2) are implemented on POWER6 with the idea to use the Logical Partition Mobility capability. Next, we describe both nodes (svcxd_a1 and svcxd_a2).

We configure two VIOS. In each VIOS, we configure two Shared Ethernet Adapters failover (SEA-failover), one for net_XD_ip_01 and another for net_ether_01 (Figure 5-3). For virtual disk storage, we use NPIV in each VIO, providing Virtual Fibre Channel adapters for the partitions.

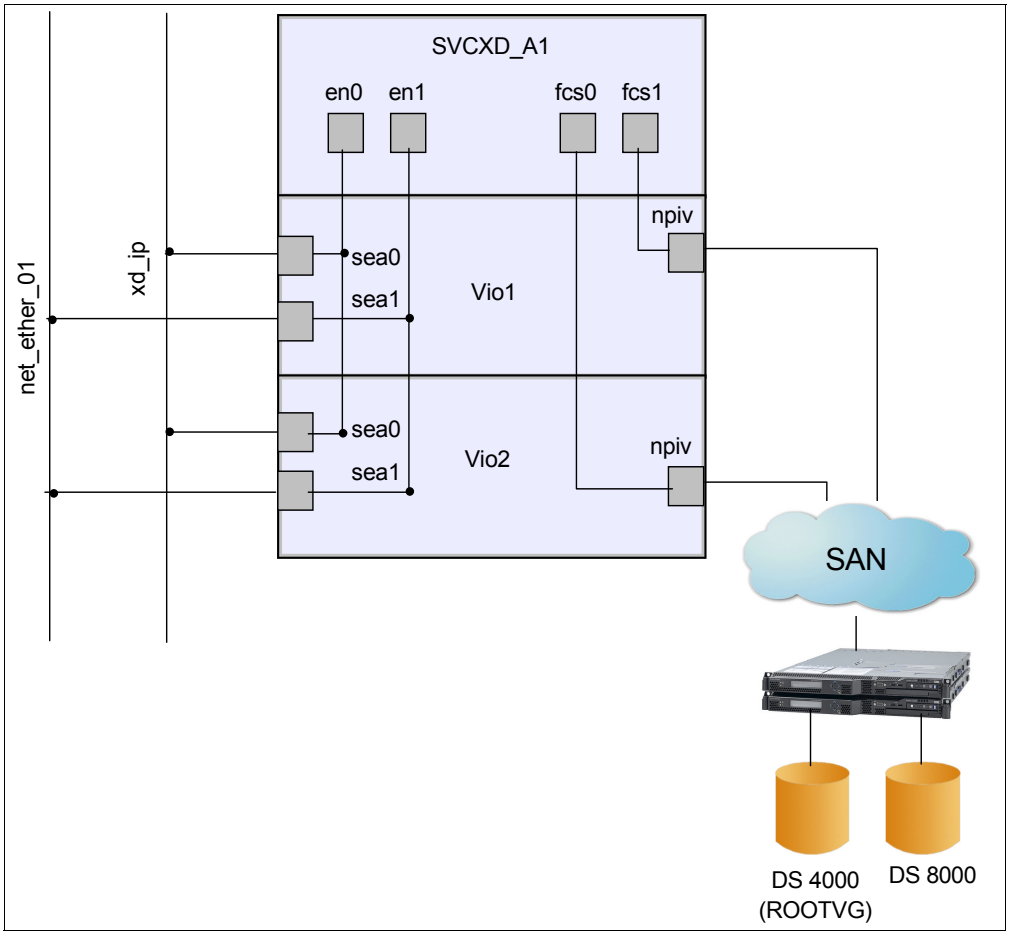


Figure 5-3 Live Partition Mobility implementation scenario

References: For more information about Power Systems virtualization features, see the *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

For information about NPIV configuration, see also “N_Port ID virtualization,” in the *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

5.2 Planning and prerequisites overview

To install and configure PowerHA Enterprise Edition for SAN Volume Controller, you must plan for the configuration and ensure that you have the prerequisites.

5.2.1 Planning

Before you configure PowerHA Enterprise Edition for SAN Volume Controller, check the following areas:

- PowerHA Enterprise Edition sites and nodes are identified.
- SAN Volume Controller cluster and PPRC licenses are configured.
- SAN Volume Controller virtual disks (VDisk), relationships, and consistency groups are identified.
- The resource groups that will contain the SAN Volume Controller-managed PPRC resources are planned.

Considerations for PowerHA Enterprise Edition for SAN Volume Controller

Keep in mind the following considerations for the current release of PowerHA Enterprise Edition for SAN Volume Controller:

- Volumes groups

  Resource groups to be managed by HACMP cannot contain volume groups with both SVC PPRC-protected and non-SVC-PPRC-protected disks.

- 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. The operations sinclude such functions 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 then propagated to the other site by using a lazy update.

- Replicated resources

  You cannot mix both SAN Volume Controller Global Mirror and Metro Mirror in the same resource group.

Considerations on resource groups with PPRC replicated resources

In general, the following restrictions apply to PowerHA Enterprise Edition resource groups that will manage PPRC replicated resources because of the way that PPRC instances are managed. For example, source site nodes have I/O access and target site nodes do not. The outcome of this is that the PowerHA policy for a resource group to come online on more than one site at a time is not supported:

- Inter-site management policy of online both sides is not supported.
- Startup policies of online using distribution policy and online on all available Nodes are not supported.
- Failover policy failover using dynamic node priority is not supported.

Reference: For more information and corresponding planning worksheets, see the *HACMP Enterprise Metro Mirror: Planning and Administration Guide*, SC23-4863.

5.2.2 Prerequisites overview

Installing and configuring PowerHA Enterprise Edition for SAN Volume Controller has several prerequisites.

Software requirement

To implement PowerHA Enterprise Edition with SAN Volume Controller, you must ensure that you have the required software.

You must install the following file sets:

- `cluster.es.svcpprc.cmds`
- `cluster.es.svcpprc.rte`
- Optional: `cluster.msg.en_US.svcpprc`

The required software and microcode levels are openssh Version 3.6.1 or later for access to SAN Volume Controller interfaces.

When you run SAN Volume Controller Version 4.x, you must install the following components:

- Storage microcode/LIC versions as per SAN Volume Controller support requirements
- Subsystem Device Driver (SDD) V1.6.3.0 or later
- 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 SAN Volume Controller support)

When you use SDDPCM, you must install the following components:

- Subsystem Device Driver Path Control Module SDDPCM): V2.2.0.0 or later
- IBM Host attachment scripts:
  - **`devices.fcp.disk.ibm.mpio.rte`** 1.0.0.10 or later
  - **`ibm2105.rte`** 32.6.100.25 or later (as specified by SAN Volume Controller support)

When you use Virtual I/O Server 1.5.1.x, you must install the following components:

- Subsystem Device Driver Path Control Module SDDPCM): v 2.2.0.0 or later
- IBM Host attachment scripts: **`devices.fcp.disk.ibm.mpio.rte`** 1.0.0.10 or later

SAN Volume Controller configuration

To eliminate a single point of failure, in each site to have two SAN Volume Controllers (clusters), and have each cluster comprise at least two SAN Volume Controller nodes. Each cluster can be both master and auxiliary copy services.

SSH connection configuration

PowerHA runs remote commands to communicate with SVC PPRC that require network connectivity and an ssh client that is configured to each SAN Volume Controller cluster.

Important: All PowerHA nodes must have an ssh client that is configured to access each SAN Volume Controller cluster from both sites.

Tip: For information about how to configure ssh client access to the SAN Volume Controller, see *Implementing the IBM System Storage SAN Volume Controller V5.1*, SG24-6423.

Identifying VDisk SAN Volume Controller to hdisk AIX client

Before you configure SVC PPRC replicated resources, identify the corresponding VDisk on the SAN Volume Controller and each hdisk device on the AIX client node.

If you are using vpath devices instead of hdisks, see *HACMP/XD Metro Mirror:Planning and Administration Guide*, SC23-4863-06.

To determine which devices correspond to each other, a SAN Volume Controller VDisk has a unique_id (UID) on the SAN Volume Controller and is part of the disk device definition in AIX. To find this information from the SAN Volume Controller master console under Work with Virtual Disks → Virtual Disks (Figure 5-4).

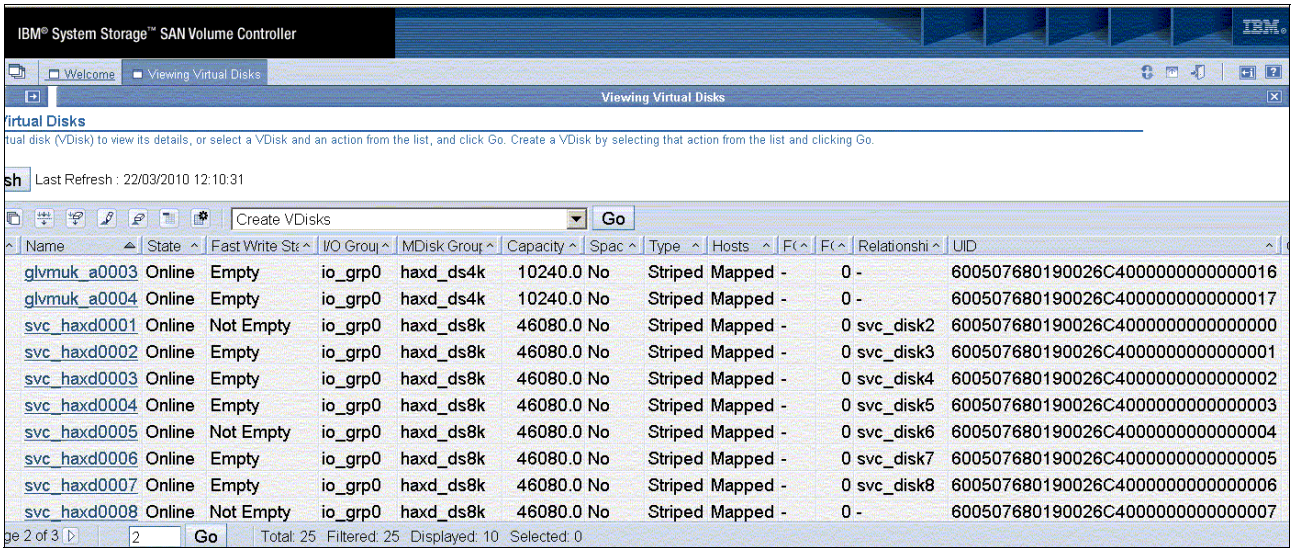


Figure 5-4 VDisks list

You can also check each UID by using the command line. Assuming that ssh access from the client to the SAN Volume Controller is configured, run:

```
ssh admin@svc_cluster_ip svcinfo lshostvdiskmap |more
```

You can also **grep** on the host alias name to narrow the list (Example 5-1).

Example 5-1 Using grep to narrow the VDisk list

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lshostvdiskmap | grep svc_haxd0001
0                SVC_550_1_A1      0              0              svc_haxd0001      C05076004FAA0027
600507680190026C4000000000000000
```

Collect this information from each VDisk to be used for the PPRC relationship from the SAN Volume Controller cluster in both sites.

On the AIX clients, the UID is in the ODM. To get it running, enter the following command:

```
odmget -q "attribute=unique_id" CuAt
```

The VDisk UID is contained in this attribute at the fifth numeric position. Example 5-2 shows the VDisk UID in bold. In this example, hdisk2 matches VDisk svc_haxd0001. Repeat the command to match and record to create proper replicated relationships.

Example 5-2 VDisk UID in ODM attribute

```
CuAt:
        name = "hdisk2"
        attribute = "unique_id"
        value = "33213600507680190026C4000000000000000004214503IBMfcp"
        type = "R"
        generic = "D"
        rep = "nl"
        nls_index = 42
```

Because the nodes from site svc_sitea share disks, you need only to get the VDisk UID from one of these nodes. You can do the same on a node from site svc_siteb.

5.3 Installing and configuring PowerHA Enterprise Edition for SAN Volume Controller

The following file sets to be installed for PowerHA Enterprise Edition for SAN Volume Controller configuration are required:

- cluster.es.svcpprc.cmds
- cluster.es.svcpprc.rte
- cluster.msg.en_US.svcpprc
- cluster.xd.license

After you install the required file sets, you see the cluster file sets that are installed (Example 5-3).

Example 5-3 Cluster file sets listing

```
[svcxd_a1][/etc]> lslpp -l | grep -i cluster
  cluster.adt.es.client.include
  cluster.adt.es.client.samples.clinfo
  cluster.adt.es.client.samples.clstat
  cluster.adt.es.client.samples.libcl
  cluster.adt.es.java.demo.monitor
  cluster.es.cfs.rte         6.1.0.0  COMMITTED  ES Cluster File System Support
  cluster.es.client.clcomd   6.1.0.1  COMMITTED  ES Cluster Communication
  cluster.es.client.lib      6.1.0.1  COMMITTED  ES Client Libraries
  cluster.es.client.rte      6.1.0.1  COMMITTED  ES Client Runtime
  cluster.es.client.utils    6.1.0.0  COMMITTED  ES Client Utilities
  cluster.es.client.wsm      6.1.0.0  COMMITTED  Web based Smit
  cluster.es.cspoc.cmds      6.1.0.1  COMMITTED  ES CSPOC Commands
  cluster.es.cspoc.dsh       6.1.0.0  COMMITTED  ES CSPOC dsh
  cluster.es.cspoc.rte       6.1.0.1  COMMITTED  ES CSPOC Runtime Commands
  cluster.es.server.cfgast   6.1.0.0  COMMITTED  ES Two-Node Configuration
  cluster.es.server.diag     6.1.0.1  COMMITTED  ES Server Diags
  cluster.es.server.events   6.1.0.1  COMMITTED  ES Server Events
  cluster.es.server.rte      6.1.0.1  COMMITTED  ES Base Server Runtime
  cluster.es.server.testtool 6.1.0.0  COMMITTED  ES Cluster Test Tool
  cluster.es.server.utils    6.1.0.1  COMMITTED  ES Server Utilities
  cluster.es.svcpprc.cmds    6.1.0.1  COMMITTED  ES HACMP - SVC PPRC Commands
```

```
  cluster.es.svcpprc.rte     6.1.0.1  COMMITTED  ES HACMP - SVC PPRC Runtime
  cluster.license            6.1.0.0  COMMITTED  HACMP Electronic License
  cluster.xd.license         6.1.0.0  COMMITTED  HACMP XD Feature License
  cluster.es.client.clcomd   6.1.0.1  COMMITTED  ES Cluster Communication
  cluster.es.client.lib      6.1.0.1  COMMITTED  ES Client Libraries
  cluster.es.client.rte      6.1.0.1  COMMITTED  ES Client Runtime
  cluster.es.client.wsm      6.1.0.0  COMMITTED  Web based Smit
  cluster.es.cspoc.rte       6.1.0.0  COMMITTED  ES CSPOC Runtime Commands
  cluster.es.server.diag     6.1.0.0  COMMITTED  ES Server Diags
  cluster.es.server.events   6.1.0.0  COMMITTED  ES Server Events
  cluster.es.server.rte      6.1.0.1  COMMITTED  ES Base Server Runtime
  cluster.es.server.utils    6.1.0.1  COMMITTED  ES Server Utilities
  cluster.es.svcpprc.rte     6.1.0.0  COMMITTED  ES HACMP - SVC PPRC Runtime
```

This remainder of this section explains how to configure PowerHA Enterprise Edition for SAN Volume Controller.

5.3.1 Topology

In this scenario, configure a four-node cluster (two at each site) by using two networks (net_ether_01 and net_XD_ip_01). Because svc_sitea and svc_siteb are in different segments of the network, we use site-specific service IP label configuration for the service IP address. Table 5-1 lists the IP address topology.

Table 5-1 IP address topology

Site	LPAR	Boot	Persistent	Service	XD_IP
svc_sitea	svcxd_a1	192.168.8.103	192.168.100.173	192.168.100.54	10.12.5.36
svc_sitea	svcxd_a2	192.168.8.104	192.168.100.174	192.168.100.94	10.12.5.37
svc_siteb	svcxd_b1	192.168.12.103	10.10.12.103	10.10.12.113	10.114.124.40
svc_siteb	svcxd_b2	192.168.12.104	10.10.12.104	10.10.12.114	10.114.124.44
		255.255.255.0	255.255.255.0	255.255.255.0	255.255.252.0

Example 5-4 shows the `/etc/hosts` configured in all nodes.

Example 5-4 The /etc/hosts that are configured on all nodes

```
# persistent addresses
192.168.100.173 svcxd_a1
192.168.100.174 svcxd_a2
10.10.12.103    svcxd_b1
10.10.12.104    svcxd_b2
#  service addresses
192.168.100.54 svcxd_a1_sv
192.168.100.94 svcxd_a2_sv
10.10.12.113   svcxd_b1_sv
10.10.12.114   svcxd_b2_sv
# boot addresses
192.168.8.103  svcxd_a1_boot
192.168.8.104  svcxd_a2_boot
192.168.12.103 svcxd_b1_boot
192.168.12.104 svcxd_b2_boot
```

```
#XD_IP network
10.12.5.36     svcxd_a1_xdip
10.12.5.37     svcxd_a2_xdip
10.114.124.40  svcxd_b1_xdip
10.114.124.44  svcxd_b2_xdip
#SVC_CLUSTER
10.12.5.55 B8_8G4
10.114.63.250 B12_4F2
```

We use the disk with the PVID “000fe4112579ef78” on the nodes svcxd_a1 and svcxd_a2 to configure net_diskhb_01. We use the disk with the PVID “00ca02ef25b24924” on the nodes svcxd_a1svcxd_b1 and svcxd_b2 to configure net_diskhb_02 (Example 5-5).

Example 5-5 Disks that are used for the heartbeat configuration

```
[svcxd_a1][/]> lspv | grep vg_hb
hdisk9          000fe4112579ef78                    vg_hba
[svcxd_a2][/]> lspv | grep vg_hb
hdisk9          000fe4112579ef78                    vg_hba
[svcxd_b1][/]> lspv | grep vg_hb
hdisk9          00ca02ef25b24924                    vg_hbb
[svcxd_b2][/]> lspv | grep vg_hb
hdisk9          00ca02ef25b24924                    vg_hbb
```

Important: These disks are used for heartbeating only between nodes that belong to the same site.

5.3.2 Configuring PowerHA Enterprise Edition for SAN Volume Controller

To configure this scenario by using SVC PPRC copy services between sites:

1. Add four nodes.
2. Add two sites.
3. Add the net_ether_01 network.
4. Add the net_XD_ip_01 network.
5. Add the net_diskhb_01 network.
6. Add the net_diskhb_02 network.
7. Add service IP labels.
8. Add SAN Volume Controller cluster definitions.
9. Add SVC PPRC relationships.
10. Add the SVC PPRC replicated Resource Configuration.
11. Create volume groups in the svc_sitea:
    a. Create logical volumes.
    b. Create file systems.
12. Create a temporary SVC PPRC relationship. Import volume groups to the remote site in the svc_siteb.
13. Create resource groups.
14. Add volumes groups, replicated resources, and services IP labels into resources groups.

5.3.3 Adding the cluster

To add the cluster, run the **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure an HACMP Cluster** → **Add/Change/Show an HACMP Cluster** (Figure 5-5).

```
                Add/Change/Show an HACMP Cluster

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

                                                      [Entry Fields]
* Cluster Name                                       [svcxd]

  NOTE: HACMP must be RESTARTED
  on all nodes in order for change to take effect
```

Figure 5-5 Adding the cluster to the topology configuration

5.3.4 Adding four nodes

To add the nodes, run the **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Nodes** → **Add a Node to the HACMP Cluster** (Figure 5-6).

```
                    Add a Node to the HACMP Cluster

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

                                                      [Entry Fields]
* Node Name                                          [svcxd_a1]
  Communication Path to Node                         [svcxd_a1_xdip]          +
```

Figure 5-6 Adding nodes to the topology configuration

Because svc_sitea and svc_siteb are in separate network segments, we use the IP address that is used in XD_IP network as the communication path. Repeat the procedure that is shown in Figure 5-6 to add each node to the cluster. After you complete the task, you see the nodes as shown in Example 5-6.

Example 5-6 Nodes names

```
[svcxd_a1][/]> clnodename
svcxd_a1
svcxd_a2
svcxd_b1
svcxd_b2
```

5.3.5 Adding two sites

To add the sites, enter the **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Sites** → **Add a Site** (Figure 5-7).

```
                                Add Site

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

                                                      [Entry Fields]
* Site Name                                          [svc_sitea]             +
* Site Nodes                                          svcxd_a1 svcxd_a2      +
+
```

Figure 5-7 Adding sites to the topology configuration

In this scenario, two sites, svc_sitea and svc_siteb, are added. Nodes svcxd_a1 and svcxd_a2 are part of svc_sitea, and svcxd_b1 and svcxd_b2 are part of svc_sitea. Repeat the procedure that is shown in Figure 5-7 for each site in the cluster. After you complete the definition of the sites, they are listed as shown in Example 5-7.

Example 5-7 Sites listing

```
[svcxd_a1][/]> cllssite
---------------------------------------------------
Sitename            Site Nodes          Dominance           Protection Type
---------------------------------------------------
svc_sitea           svcxd_a1 svcxd_a2   yes                 NONE
svc_siteb           svcxd_b1 svcxd_b2   no                  NONE
```

5.3.6 Adding the net_ether_01 network

To add the network:

1. Run the **smitty hacmp** command.
2. Select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Networks** → **Add a Network to the HACMP Cluster** → **Pre-defined IP-based Network Types** → **ether**.

3. Change the Enable IP Address Takeover via Alias to `YES` (Figure 5-8).

```
 Add an IP-Based Network to the HACMP Cluster

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

                                                      [Entry Fields]
* Network Name                                       [net_ether_01]
* Network Type                                        ether
* Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.255.0]
* Enable IP Address Takeover via IP Aliases          [Yes]
+
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 5-8 Adding net_ether_01 in the topology configuration

4. After you add the network net_ether_01, add the network interfaces by entering the SMIT **`smitty hacmp`** command. Then, select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Add Pre-defined Communication Interfaces and Devices → Communication Interfaces → net_ether_01**.
5. In the Add a Communication Interface panel (Figure 5-9), for IP Label/Address, select each correspondent boot address for each node.
6. Repeat the procedure that is shown in Figure 5-9 for each node in the cluster.

```
                        Add a Communication Interface

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

                                       [Entry Fields]
* IP Label/Address                     [svcxd_a1_boot]               +
* Network Type                          ether
* Network Name                          net_ether_01
* Node Name                             [svcxd_a1]             +
  Network Interface                     []
```

Figure 5-9 Adding communication Interfaces for net_ether_01

Example 5-8 shows the cluster topology now.

Example 5-8 Cluster topology after configuration

```
[svcxd_a1][/]> cltopinfo
Cluster Name: svcxd
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 1 network(s) defined
NODE svcxd_a1:
               Network net_ether_01
                       svcxd_a1_boot   192.168.8.103
```

```
NODE svcxd_a2:
                Network net_ether_01
                        svcxd_a2_boot   192.168.8.104
NODE svcxd_b1:
                Network net_ether_01
                        svcxd_b1_boot   192.168.12.103
NODE svcxd_b2:
               Network net_ether_01
                       svcxd_b2_boot   192.168.12.104
```

7. Add the persistent IP address by entering the SMIT **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Persistent Node IP Label/Addresses** → **Add a Persistent Node IP Label/Address** → **Select a Node** (Figure 5-10).

```
                  Add a Persistent Node IP Label/Address

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

                                                      [Entry Fields]
* Node Name                                           svcxd_a1
* Network Name                                       [net_ether_01] +
* Node IP Label/Address                              [svcxd_a1] +
  Netmask(IPv4)/Prefix Length(IPv6)                  []
```

Figure 5-10 Adding a persistent node IP label

8. Repeat the procedure that is shown in Figure 5-10 for each node in the cluster.

5.3.7 Adding the net_XD_ip_01 network

To add the network:

1. Run the **smitty hacmp** command.
2. Select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Networks** → **Add a Network to the HACMP Cluster** → **Pre-defined IP-based Network Types** → **XD-IP**.
3. Change the Enable IP Address Takeover via Alias field to NO (Figure 5-11).

```
              Add an IP-Based Network to the HACMP Cluster

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

                                                      [Entry Fields]
* Network Name                                       [net_XD_ip_01]
* Network Type                                        XD_ip
* Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.252.0]
* Enable IP Address Takeover via IP Aliases     [NO]
+
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 5-11 Adding network net_XD_ip_01

4. After you create the net_XD_ip_01 network, add the network interfaces. Enter the **smitty hacmp** command. Then, select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Add Pre-defined Communication Interfaces and Devices → Communication Interfaces → net_XD_ip_01** (Figure 5-12).
5. Repeat the procedure that is shown in Figure 5-12 for each node in the cluster.

```
                        Add a Communication Interface

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

                                                        [Entry Fields]
* IP Label/Address                                     [svcxd_a1_xdip] +
* Network Type                                          XD_ip
* Network Name                                          net_XD_ip_01
* Node Name                                            [svcxd_a1] +
  Network Interface                                    []
```

Figure 5-12 Adding communication Interfaces for net_XD_ip_01

6. After you complete this procedure, synchronize the cluster. Example 5-9 shows the configuration.

Example 5-9 Cluster topology

```
[svcxd_a1][/]> cltopinfo
Cluster Name: svcxd
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 2clnetwork(s) defined
NODE svcxd_a1:
        Network net_XD_ip_01
                svcxd_a1_xdip   10.12.5.36
        Network net_ether_01
                svcxd_a1_boot   192.168.8.103
NODE svcxd_a2:
        Network net_XD_ip_01
                svcxd_a2_xdip   10.12.5.37
        Network net_ether_01
                svcxd_a2_boot   192.168.8.104
NODE svcxd_b1:
        Network net_XD_ip_01
                svcxd_b1_xdip   10.114.124.40
        Network net_ether_01
                svcxd_b1_boot   192.168.12.103
NODE svcxd_b2:
        Network net_XD_ip_01
                svcxd_b2_xdip   10.114.124.44
        Network net_ether_01
                svcxd_b2_boot   192.168.12.104
```

5.3.8 Adding the net_diskhb_01 network

To add the network:

1. Enter the **smitty hacmp** command.
2. Select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Communication Devices**.
3. Select both disks that are related to PVID 000fe4112579ef78 (Figure 5-13).

```
>   svcxd_a1                          hdisk9   000fe4112579ef78
>   svcxd_a2                          hdisk9   000fe4112579ef78
```

Figure 5-13 Selecting disks to net_diskhb_01

4. Select the disks to be shared between the svc_sitea.

5.3.9 Adding the net_diskhb_02 network

To add the network:

1. Enter the **smitty hacmp** command.
2. Select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Communication Devices**.
3. Select both disks that are related to PVID 00ca02ef25b24924 (Figure 5-14).

```
>   svcxd_b1                          hdisk9   00ca02ef25b24924
>   svcxd_b2                          hdisk9   00ca02ef25b24924
```

Figure 5-14 Selecting disks to net_diskhb_02

4. Select the disks to be shared between the svc_siteb.

After this procedure is completed, synchronize the cluster. Example 5-10 shows the configuration.

Example 5-10 Custer topology

```
[svcxd_a1][/]> cltopinfo
Cluster Name: svcxd
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 svcxd_a1:
        Network net_XD_ip_01
                svcxd_a1_xdip   10.12.5.36
        Network net_diskhb_01
                svcxd_a1_hdisk9_01      /dev/hdisk9
        Network net_diskhb_02
        Network net_ether_01
                svcxd_a1_boot   192.168.8.103
```

```
NODE svcxd_a2:
        Network net_XD_ip_01
                svcxd_a2_xdip   10.12.5.37
        Network net_diskhb_01
                svcxd_a2_hdisk9_01      /dev/hdisk9
        Network net_diskhb_02
        Network net_ether_01
                svcxd_a2_boot   192.168.8.104
NODE svcxd_b1:
        Network net_XD_ip_01
                svcxd_b1_xdip   10.114.124.40
        Network net_diskhb_01
        Network net_diskhb_02
                svcxd_b1_hdisk9_01      /dev/hdisk9
        Network net_ether_01
                svcxd_b1_boot   192.168.12.103
NODE svcxd_b2:
        Network net_XD_ip_01
                svcxd_b2_xdip   10.114.124.44
        Network net_diskhb_01
        Network net_diskhb_02
                svcxd_b2_hdisk9_01      /dev/hdisk9
        Network net_ether_01
                svcxd_b2_boot   192.168.12.104
```

5.3.10 Adding the service IP label

To add the service IP label:

1. Enter the SMIT **`smitty hacmp`** command.
2. Select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resources Configuration** → **Configure HACMP Service IP Labels/Addresses** → **Add a Service IP Label/Address** → **Configurable on Multiple Nodes** → **net_ether_01** (Figure 5-15).
3. Repeat the procedure that is shown in Figure 5-15 for each service IP label. We repeat it, and add the service label `svcxd_a2_sv` at site svc_sitea, add the service label `svcxd_b1_sv` at site svc_siteb, and add the service label `svcxd_b2_sv` at the svc_siteb site.

```
 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.

                                                        [Entry Fields]
* IP Label/Address                                     svcxd_a1_sv+
  Netmask(IPv4)/Prefix Length(IPv6)                    []
* Network Name                                         net_ether_01
  Alternate Hardware Address to accompany IP Label/A   []
  ddress
  Associated Site                                      svc_sitea +
```

Figure 5-15 Adding a site-specific service IP label

Example 5-11 shows the current topology configuration.

Example 5-11 Cluster topology

```
[svcxd_a1][/]> cltopinfo
Cluster Name: svcxd
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 svcxd_a1:
        Network net_XD_ip_01
                svcxd_a1_xdip   10.12.5.36
        Network net_diskhb_01
                svcxd_a1_hdisk9_01      /dev/hdisk9
        Network net_diskhb_02
        Network net_ether_01
                svcxd_b1_sv     10.10.12.113
                svcxd_a2_sv     192.168.100.94
                svcxd_a1_sv     192.168.100.54
                svcxd_b2_sv     10.10.12.114
                svcxd_a1_boot   192.168.8.103
NODE svcxd_a2:
        Network net_XD_ip_01
                svcxd_a2_xdip   10.12.5.37
        Network net_diskhb_01
                svcxd_a2_hdisk9_01      /dev/hdisk9
        Network net_diskhb_02
        Network net_ether_01
                svcxd_b1_sv     10.10.12.113
                svcxd_a2_sv     192.168.100.94
                svcxd_a1_sv     192.168.100.54
                svcxd_b2_sv     10.10.12.114
                svcxd_a2_boot   192.168.8.104
NODE svcxd_b1:
        Network net_XD_ip_01
                svcxd_b1_xdip   10.114.124.40
        Network net_diskhb_01
        Network net_diskhb_02
                svcxd_b1_hdisk9_01      /dev/hdisk9
        Network net_ether_01
                svcxd_b1_sv     10.10.12.113
                svcxd_a2_sv     192.168.100.94
                svcxd_a1_sv     192.168.100.54
                svcxd_b2_sv     10.10.12.114
                svcxd_b1_boot   192.168.12.103
NODE svcxd_b2:
        Network net_XD_ip_01
                svcxd_b2_xdip   10.114.124.44
        Network net_diskhb_01
        Network net_diskhb_02
                svcxd_b2_hdisk9_01      /dev/hdisk9
        Network net_ether_01
                svcxd_b1_sv     10.10.12.113
                svcxd_a2_sv     192.168.100.94
```

```
svcxd_a1_sv     192.168.100.54
svcxd_b2_sv     10.10.12.114
svcxd_b2_boot   192.168.12.104
```

5.3.11 Adding SAN Volume Controller cluster definitions

To add the SAN Volume Controller cluster, run the SMIT **`smitty svcpprc_def`** command, and then select **SVC Clusters Definition to HACMP → Add an SVC Cluster**.

In the panel that opens (see Figure 5-16), complete the following parameters:

1. In the SVC Cluster Name parameter, enter the same name for the SAN Volume Controller cluster. The name cannot be more than 20 alphanumeric characters.
2. In the SVC Cluster Role parameter, select Master or Auxiliary. The Master SVC Cluster is defined at the primary PowerHA site, the auxiliary SVC Cluster, and the backup PowerHA site.
3. In the SVC Cluster IP Address parameter, enter the IP address of the cluster that is used by PowerHA to submit PPRC management commands.
4. In the Remote SVC Partner parameter, enter the name of the SAN Volume Controller cluster that will host VDisks from the other site of the SVC PPRC link.

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

                                                  [Entry Fields]
* SVC Cluster Name                               [B8_8G4]
* SVC Cluster Role                               [Master]                +
* HACMP site                                      svc_sitea              +
* SVC Cluster IP Address                         [10.12.5.55]
  SVC Cluster Second IP Address                  []
* Remote SVC Partner                             [B12_4F2]
```

Figure 5-16 Adding a SAN Volume Controller cluster

In this scenario, both SAN Volume Controller clusters have master roles. The same procedure that is shown in Figure 5-16 is repeated for the SAN Volume Controller cluster B12_4F2 at site svc_siteb. After you complete the SAN Volume Controller cluster definition, Example 5-12 shows its configuration.

Example 5-12 SAN Volume Controller cluster definition

```
[svcxd_a1][/]> cllssvc -a
#SVCNAME ROLE SITENAME IPADDR IPADDR2 RPARTNER
B8_8G4 Master svc_sitea 10.12.5.55  B12_4F2
B12_4F2 Master svc_siteb 10.114.63.250  B8_8G4
```

5.3.12 Adding the SVC PPRC relationships

In this section, you define the VDisks to be part of the mirror relationships between the sites.

Important: It is critical to know the hdisk or vpath and its corresponding VDisk to use at each site to define the correct relationships. For more information, see “Identifying VDisk SAN Volume Controller to hdisk AIX client” on page 159.

To add the SAN Volume Controller relationships:

1. Run the **smitty svcpprc_def** command.
2. Select **SVC PPRC Relationships Definition** → **Add an SVC PPRC Relationship**.
3. Complete the following parameters (Figure 5-17):
   a. In the Relationship Name parameter, enter the name that is used by both SAN Volume Controller and PowerHA for configuration of SVC PPRC relationships. Use no more than 20 alphanumeric characters and underscores.
   b. In the Master VDisk Info parameter, enter the name in the `vdisk_name@svc_cluster` format for the master and auxiliary VDisk names. The master VDisk is the disk the primary site for the resource group that the SVC PPRC Relationship will be a part of.
   c. In the Auxiliary VDisk Info parameter, enter the auxiliary VDisk at the backup site for the resource group that the SVC PPRC relationship will be a part of.
4. Repeat this procedure for every disk as needed. In this scenario, we repeated the procedure six times to create relationship svc_disk3 up to svc_disk8. We use this procedure later for Metro Mirror svc_disk2, svc_disk3, svc_disk4, and svc_disk5 to mirror svc_sitea to svc_siteb. For Global Mirror, we use svc_disk6, svc_disk7, and svc_disk8 to mirror svc_siteb to svc_sitea.

```
                        Add an SVC PPRC Relationship

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

                                                        [Entry Fields]
* Relationship Name                                    [svc_disk2]
* Master VDisk Info                                   [svc_haxd0001@B8_8G4]
* Auxiliary VDisk Info
[haxd_svc_v0001@B12_4F2]
```

Figure 5-17 Adding an SVC PPRC relationship

After you complete the procedure, the SVC PPRC relationships are displayed as shown in Example 5-13.

Example 5-13 SVC PPRC relationship

```
[svcxd_a1][/]> cllsrelationship -a
relationship_name MasterVdisk_info AuxiliaryVdisk_info
svc_disk2       svc_haxd0001@B8_8G4 haxd_svc_v0001@B12_4F2
svc_disk3       svc_haxd0002@B8_8G4 haxd_svc_v0002@B12_4F2
svc_disk4       svc_haxd0003@B8_8G4 haxd_svc_v0003@B12_4F2
svc_disk5       svc_haxd0004@B8_8G4 haxd_svc_v0004@B12_4F2
```

```
svc_disk6       haxd_svc_v0005@B12_4F2 svc_haxd0005@B8_8G4
svc_disk7       haxd_svc_v0006@B12_4F2 svc_haxd0006@B8_8G4
svc_disk8       haxd_svc_v0007@B12_4F2 svc_haxd0007@B8_8G4
```

5.3.13 Adding the SVC PPRC replicated resource configuration

In this section, you define the replicated resources to be added in the resource groups. To add the SVC PPRC replicated resource configuration:

1. Run the **smitty svcpprc_def** command.
2. Select **SVC PPRC replicated Resource Configuration** → **Add an SVC PPRC Resource**.
3. In the Add an SVC PPRC Resource panel (Figure 5-18), complete the following parameters:
   a. In the SVC PPRC Consistency Group Name parameter, enter the name that is used by SAN Volume Controller and that is used in the resource group configuration. Do not use more than 20 alphanumeric characters and underscores.
   b. In the Master SVC Cluster Name parameter, enter the name of the master cluster, which is the SAN Volume Controller cluster that is connected to the PowerHA Primary Site.
   c. In the Auxiliary SVC Cluster Name parameter, enter the name of the SAN Volume Controller cluster that is connected to the PowerHA Backup/Recovery Site.
   d. In the List of Relationships parameter, enter the list of names of the SVC PPRC relationships.
   e. In the Copy Type parameter, enter either Metro or Global.
   f. In the HACMP Recovery Action parameter, select the action to be taken by PowerHA if a site failover occurs for a replicated pair. The options are Manual for Manual intervation required or Automated.

   Except for the SVC PPRC Consistency Group Name field, every field provides a list the option that was previously defined.
4. Repeat step 3 for each replicated relationship that was previously created. In this scenario, we repeat it one more time.

```
                        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                    [svc_metro]
* Master SVC Cluster Name                            [B8_8G4] +
* Auxiliary SVC Cluster Name                         [B12_4F2] +
* List of Relationships             [svc_disk2 svc_disk3 svc_disk4 svc_disk5] +
* Copy Type                                          [METRO] +
* HACMP Recovery Action                              [AUTO] +
```

Figure 5-18 Adding SVC PPRC replicated resource

After you complete the parameters, you see the SVC PPRC Resource listing as shown in Example 5-14.

Example 5-14 SPVC PPRC resources

```
[svcxd_a1][/]> cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8           GLOBAL AUTO
```

5.3.14 Creating volume groups in the svc_sitea

To configure a local cluster, you can create the volume group locally and then import it to the other system. However, we have two situations in this scenario:

- We have shared disks between nodes svcxd_a1 and svcxd_a2 provided by the SAN Volume Controller Cluster B8_8G4.
- We have shared disks between nodes svcxd_b1 and svcxd_b2 provided by the SAN Volume Controller Cluster B12_4F2.

First, we create the volume groups in the nodes from svc_sitea, and then, we import the volume groups to the nodes from svc_siteb. Temporary replication is required to the remote SAN Volume Controller Cluster.

C-SPOC: You cannot use C-SPOC to create a volume group for replicated resources.

You can use the **smit mkvg** fast path to create a volume group as desired. Make sure that the option Active volume group AUTOMATICALLY at system restart is set to no. Repeat this step as need for each volume group.

5.3.15 Creating the volumes groups

Create the siteametrovg and sitebmetrovg volume groups in the svcxd_a1 node (Example 5-15).

Example 5-15 Physical volumes

```
[svcxd_a1][/]> lspv
hdisk0          000fe4110889e1a9                    rootvg          active
hdisk1          000fe4112579eb7c                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg
hdisk3          000fe4112579ee11                    siteametrovg
hdisk4          000fe4112579ee4c                    siteametrovg
hdisk5          000fe4112579ee89                    siteametrovg
hdisk6          000fe4112579eec4                    sitebglobalvg
hdisk7          000fe4112579eefe                    sitebglobalvg
hdisk8          000fe4112579ef3d                    sitebglobalvg
hdisk9          000fe4112579ef78                    vg_hba
[svcxd_a1][/]> lsvg -l siteametrovg
siteametrovg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv00             jfs2log    1       1       1    open/syncd    N/A
lvsiteametro1       jfs2       4       4       1    open/syncd    /dev/siteametro1
lvsiteametro2       jfs2       4       4       1    open/syncd    /dev/siteametro2
lvsiteametro3       jfs2       4       4       1    open/syncd    /dev/siteametro3
```

```
[svcxd_a1][/]> lsvg -l sitebglobalvg
sitebglobalvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv01             jfs2log    1       1       1    closed/syncd  N/A
lvsitebglobal1      jfs2       4       4       1    closed/syncd
/dev/sitebglobal1
lvsitebglobal2      jfs2       4       4       1    closed/syncd
/dev/sitebglobal2
lvsitebglobal3      jfs2       4       4       1    closed/syncd
/dev/sitebglobal3
```

Major numbers: Determining a major number is required only when using NFS. However, for clusters, keep the major numbers the same.

In the svcxd_a2 node, run the command to import the volumes groups (Example 5-16).

Example 5-16 Importing the volume groups

```
[svcxd_a2][/]>importvg -V 38 -y siteametrovg hdisk2
[svcxd_a2][/]>importvg -V 39 -y sitebglobalvg hdisk6
```

After you import the volume groups, enter the **lspv** command to check the list of the physical volumes (Example 5-17).

Example 5-17 Physical volumes

```
[svcxd_a2][/]> lspv
hdisk0          000fe401088fc86b                    rootvg          active
hdisk1          00c7cd9e84341284                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg
hdisk3          000fe4112579ee11                    siteametrovg
hdisk4          000fe4112579ee4c                    siteametrovg
hdisk5          000fe4112579ee89                    siteametrovg
hdisk6          000fe4112579eec4                    sitebglobalvg
hdisk7          000fe4112579eefe                    sitebglobalvg
hdisk8          000fe4112579ef3d                    sitebglobalvg
hdisk9          000fe4112579ef78                    vg_hba
```

Ensure that AUTO VARYON on the volumes group is disabled. When you use enhanced concurrent volume groups, this setting is the default setting. If it is not set, run the **chvg -a n <vgname>** command for each volume group. Run the **varyonvg <vgname>** command to verify that all logical volumes and file systems exist and can be mounted (Example 5-18).

Example 5-18 Volume groups

```
[svcxd_a2][/]> lsvg -l siteametrovg
siteametrovg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv00             jfs2log    1       1       1    open/syncd    N/A
lvsiteametro1       jfs2       4       4       1    open/syncd    /dev/siteametro1
lvsiteametro2       jfs2       4       4       1    open/syncd    /dev/siteametro2
lvsiteametro3       jfs2       4       4       1    open/syncd    /dev/siteametro3
[svcxd_a2][/]> lsvg -l sitebglobalvg
sitebglobalvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
```

```
loglv01             jfs2log    1       1       1    closed/syncd  N/A
lvsitebglobal1      jfs2       4       4       1    closed/syncd  /dev/sitebglobal1
lvsitebglobal2      jfs2       4       4       1    closed/syncd  /dev/sitebglobal2
lvsitebglobal3      jfs2       4       4       1    closed/syncd  /dev/sitebgl
```

5.3.16 Creating temporary SVC PPRC relationships

Before you continue to the next step, ensure that the file systems are unmounted and that the volume group is varied off. In this step, you replicate the previously LVM-related information to the remote auxiliary VDisks. This step is crucial for importing the volume group information on the nodes from svc_siteb.

Reference: We set up all the temporary relationships from the SCV command line. For the correct command syntax, see the *SVC CLI Guide*, SC26-7903.

Creating the temporary relationships in this scenario requires running the command that is shown in Example 5-19.

Example 5-19 Creating the PPRC relationships

```
ssh admin@<master_SVC_cluster> svctask mkrcrelationship -master
<master_Vdisk_name> -aux <aux_Vdisk_name> -cluster <Aux_SVC_cluster> -name
<relationship_name>
```

In this scenario, to make the PPRC relationships, use the command that is shown in Example 5-20.

Example 5-20 Creating the PPRC relationships

```
ssh admin@B8_8G4 svctask mkrcrelationship -master svc_haxd0001 -aux haxd_svc_v0001
-cluster B12_4F2 -name temp1
```

The <master_SVC_cluster> value: The <*master_SVC_cluster*> can be a name that is resolvable in the `/etc/hosts` file.

After you create the relationships, monitor the creation by using the sample outputs that are taken before and after the relationship is in sync (Example 5-21).

Example 5-21 SAN Volume Controller relationships

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship temp1
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
primary master
consistency_group_id 0
consistency_group_name svc_metro
```

```
state inconsistent_copying
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro

[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship temp1
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
primary master
consistency_group_id 0
consistency_group_name svc_metro
state consistent_synchrnized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

After the copy completes with success, remove the SVC PPRC relationship (Example 5-22).

Example 5-22 Removing an SVC PPRC relationship

```
ssh admin@<master_SVC_Custer> svctask rmrcrelationship <relationship name>
```

In our case, this command translates to the scenario shown in Example 5-23.

Example 5-23 Removing an SVC PPRC relationship

```
[svcxd_a1][/]> ssh admin@B8_8G svctask rmrcrelationship temp1
```

Repeat the command that is shown in Example 5-22 for each relationship that is created.

5.3.17 Importing the volume groups to the remote site svc_siteb

After you remove the relationship by using the SMIT menus or the command line on the nodes from the site svc_siteb, import the volume groups that were previously created. In this scenario, volume groups siteametrovg and sitebglobalvg are imported on nodes svcxd_b1and svcxd_b2.

Important: Before you import the volume groups, check in the remote disk or system whether the PVID is present by running the **chdev -l hdisk# -a pv=yes** command. This PVID must match the hdisk of the opposite site as a true complete copy of the disk. The disk is *not* a shared disk.

Example 5-24 checks the PVID on the svcxd_b1 node.

Example 5-24 PVID

```
[svcxd_b1][/]> lspv
hdisk0          000fe401088fc86b                    rootvg          active
hdisk1          00c7cd9e84341284                    rootvg          active
hdisk2          000fe4112579ec10
hdisk3          000fe4112579ee11
hdisk4          000fe4112579ee4c
hdisk5          000fe4112579ee89
hdisk6          000fe4112579eec4
hdisk7          000fe4112579eefe
hdisk8          000fe4112579ef3d
hdisk9          00ca02ef25b24924                      vg_hbb
```

Import the volume groups on node svcxd_b1 as shown in Example 5-25.

Example 5-25 Importing the volume groups

```
[svcxd_b1][/]>importvg -V 38 -y siteametrovg hdisk2
[svcxd_b1][/]>importvg -V 39 -y sitebglobalvg hdisk6
```

After you import the volume groups, check them by using the **lspv** command (Example 5-26).

Example 5-26 Physical volumes

```
[svcxd_b1][/]> lspv
hdisk0          00ca02ef74a67ade                    rootvg          active
hdisk1          00ca02ef751884e8                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg
hdisk3          000fe4112579ee11                    siteametrovg
hdisk4          000fe4112579ee4c                    siteametrovg
hdisk5          000fe4112579ee89                    siteametrovg
hdisk6          000fe4112579eec4                    sitebglobalvg
hdisk7          000fe4112579eefe                    sitebglobalvg
hdisk8          000fe4112579ef3d                    sitebglobalvg
hdisk9          00ca02ef25b24924                    vg_hbb
```

Run the **lsvg** command on each volume group to check whether you can read them (Example 5-27).

Example 5-27 Volume groups read

```
[svcxd_b1][/]> lsvg -l siteametrovg
siteametrovg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv00             jfs2log    1       1       1    closed/syncd  N/A
lvsiteametro1       jfs2       4       4       1    closed/syncd  /dev/siteametro1
lvsiteametro2       jfs2       4       4       1    closed/syncd  /dev/siteametro2
lvsiteametro3       jfs2       4       4       1    closed/syncd  /dev/siteametro3
```

```
[svcxd_b1][/]> lsvg -l sitebglobalvg
sitebglobalvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv01             jfs2log    1       1       1    open/syncd    N/A
lvsitebglobal1      jfs2       4       4       1    open/syncd
/dev/sitebglobal1
lvsitebglobal2      jfs2       4       4       1    open/syncd
/dev/sitebglobal2
lvsitebglobal3      jfs2       4       4       1    open/syncd   /dev/sitebglobal3
```

Before you import these volume groups to the node svcxd_b2, vary off all of the volume groups that you imported in the last procedure. On node svcxd_b2, we checked the PVID and imported volumes groups as shown in Example 5-28.

Example 5-28 Volume groups

```
[svcxd_b2][/]> lsvg -l siteametrovg
siteametrovg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv00             jfs2log    1       1       1    closed/syncd  N/A
lvsiteametro1       jfs2       4       4       1    closed/syncd  /dev/siteametro1
lvsiteametro2       jfs2       4       4       1    closed/syncd  /dev/siteametro2
lvsiteametro3       jfs2       4       4       1    closed/syncd  /dev/siteametro3

[svcxd_b2][/]> lsvg -l sitebglobalvg
sitebglobalvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
loglv01             jfs2log    1       1       1    closed/syncd  N/A
lvsitebglobal1      jfs2       4       4       1    closed/syncd
/dev/sitebglobal1
lvsitebglobal2      jfs2       4       4       1    closed/syncd
/dev/sitebglobal2
lvsitebglobal3      jfs2       4       4       1    closed/syncd
/dev/sitebglobal3
```

5.3.18 Creating the resource groups

To add a resource group, run the **smitty hacmp** command. Then, select **Extended Configuration → Extended Resource Configuration → Add a Resource Group**. In the SMIT menu (Figure 5-19 on page 180), extra resource group fields are available when sites are defined to the cluster.

```
                        Add a Resource Group (extended)

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

                                                      [Entry Fields]
* Resource Group Name                                [RG_sitea]

  Inter-Site Management Policy                       [Prefer Primary Site] +
* Participating Nodes from Primary Site              [svcxd_a1 svcxd_a2] +
  Participating Nodes from Secondary Site            [svcxd_b2 svcxd_b1] +

  Startup Policy                  Online On Home Node Only+
  Failover Policy                Failover To Next Priority Node In The List+
  Failback Policy                Neve Failback+
```

Figure 5-19 Adding a resource group

The Inter-Site Management Policy parameter has the following options:

Ignore If you select this option, the resource group will not have ONLINE SECONDARY instances. Use this option if you use cross-site LVM mirroring. You can also use it with PowerHA Enterprise Edition for Metro Mirror.

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 instance 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 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 or failback.

In the Participating Nodes from Primary Site parameter, specify the nodes in order for the primary site.

In the Participating Nodes from Secondary Site parameter, specify the nodes in order for the secondary site.

Important: The resource groups that include PPRC replicated resources have the following considerations:

- Inter-site management policy of online both sides is not supported.
- Startup policies of online using distribution policy and online on all available nodes are not supported.
- Failover policy failover by using dynamic node priority is not supported.

Complete the fields as desired. Repeat this action for any additional resource groups that you need to add. In this scenario, we create two resource groups:

- One with nodes svcxd_a1 and svcxd_a2 as the primary site (svc_sitea) for Metro Mirror
- One with nodes svcxd_b1 svcxd_b2 as the secondary site (svc_siteb) for Global Mirror

Example 5-29 shows these two resource groups. Only the relevant fields are shown.

Example 5-29 Resources groups

```
Resource Group Name                                 RG_sitea
Participating Node Name(s)                svcxd_a1 svcxd_a2 svcxd_b2 svcxd_b1
Startup Policy                                      Online On Home Node Only
Failover Policy                       Failover To Next Priority Node In The List
Failback Policy                                     Never Failback
Site Relationship                                   Prefer Primary Site

Resource Group Name                                 RG_siteb
Participating Node Name(s)                svcxd_b1 svcxd_b2 svcxd_a2 svcxd_a1
Startup Policy                                      Online On Home Node Only
Failover Policy                       Failover To Next Priority Node In The List
Failback Policy                                     Never Failback
Site Relationship                                   Prefer Primary Site
```

5.3.19 Adding volumes groups, replicated resources, and services IP labels

To add the SAN Volume Controller cluster:

1. Run the **smitty hacmp** command.
2. Select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resource Group Configuration** → **Change/Show Resources and Attributes for a Resource Group**.
3. Choose one of the resource groups that you created previously, which is RG_sitea in this scenario. Then, press Enter.
4. Enter the desired resources.

   In addition to the normal resources of service IP and volume groups, the SVC PPRC Replicated Resources is also used to include the replicated resources that you created previously. For svc_sitea and siteametrovg, the replicated resource svc_metro is added as shown in Figure 5-20.
5. Press F4 to access the list of options for the field. Press F7 to select each option.
6. Repeat this step as needed. In this scenario, we repeat it one more time for RG_siteb to include the service IP label svcxd_a2_sv and svcxd_b2_sv, the volume group sitebglobalvg, and the replicated resource svc_global.

After you complete this panel, for RG_sitea or RG_siteb resource group in this example, the SMIT panel (Figure 5-20) is displayed.

```
      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]                                                 [Entry Fields]
  Resource Group Name                                 RG_sitea
  Inter-site Management Policy                        Prefer Primary Site
  Participating Nodes from Primary Site               svcxd_a1 svcxd_a2
  Participating Nodes from Secondary Site             svcxd_b2 svcxd_b1

  Startup Policy                                      Online On Home Node Only
  Failover Policy                                     Failover To Next Priority
Node In The List
  Failback Policy                                     Never Failback

  Service IP Labels/Addresses                        [svcxd_a1_sv svcxd_b1_sv]+
  Application Servers                                []

  Volume Groups                                      [siteametrovg ]         +

SVC PPRC Replicated Resources                       [svc_metro]             +
```

Figure 5-20 Add SAN Volume Controller replicated resources into resource group

Example 5-30 shows the details of the resource groups that are used in this scenario.

Example 5-30 Resources groups

```
Resource Group Name                                 RG_sitea
Participating Node Name(s)                          svcxd_a1 svcxd_a2 svcxd_b2
svcxd_b1
Startup Policy                                      Online On First Available Node
Failover Policy                                    Failover To Next Priority Node
In The List
Failback Policy                                     Never Failback
Site Relationship                                   Prefer Primary Site
Service IP Label                                    svcxd_a1_sv svcxd_b1_sv
Volume Groups                                       siteametrovg
SVC PPRC Replicated Resources                       svc_metro

Resource Group Name                                 RG_siteb
Participating Node Name(s)                          svcxd_b1 svcxd_b2 svcxd_a2
svcxd_a1
Startup Policy                                      Online On Home Node Only
Failover Policy                                    Failover To Next Priority Node
In The List
Failback Policy                                     Never Failback
Site Relationship                                   Prefer Primary Site
Service IP Label                                    svcxd_a2_sv svcxd_b2_sv
Volume Groups                                       sitebglobalvg
SVC PPRC Replicated Resources                       svc_global
```

5.3.20 Synchronizing the cluster

Now synchronize the cluster. Enter the **smitty hacmp** SMIT command. Then, select **Extended Configuration** → **Extended Verification and Synchronization**. After the synchronization is successfully completed, the cluster is ready to be started and tested.

Tip: Before cluster synchronization, you can run the **/usr/es/sbin/cluster/svcpprc/cl_verify_svcpprc_config** script to verify the SAN Volume Controller copy services. This command aids in troubleshooting if errors occur during cluster synchronization.

5.4 Adding and removing disks to PowerHA Enterprise Edition for SAN Volume Controller

Next, you add and remove disks for PowerHA Enterprise Edition for SAN Volume Controller.

5.4.1 Adding disks to PowerHA Enterprise Edition with SAN Volume Controller

In this scenario, you create one new VDisk in SAN Volume Controller cluster, which is B8_8G4 in this scenario, to be a *master VDisk*. You associate it with the nodes from site svc_sitea (svcxd_a1 and svcxd_a2), in this scenario. You also create one new VDisk in SAN Volume Controller Cluster, which is B12_4F2 in this scenario, to be *auxiliary VDisk*. You associate it with the nodes from svc_siteb (svcxd_b1 and svcxd_b2), in this scenario (Example 5-31).

Example 5-31 New disks that are created from SAN Volume Controller cluster

```
[svcxd_a1][/]> lspv
hdisk0          000fe4110889e1a9                    rootvg          active
hdisk1          000fe4112579eb7c                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg    concurrent
hdisk3          000fe4112579ee11                    siteametrovg    concurrent
hdisk4          000fe4112579ee4c                    siteametrovg    concurrent
hdisk5          000fe4112579ee89                    siteametrovg    concurrent
hdisk6          000fe4112579eec4                    sitebglobalvg   concurrent
hdisk7          000fe4112579eefe                    sitebglobalvg   concurrent
hdisk8          000fe4112579ef3d                    sitebglobalvg   concurrent
hdisk9          000fe4112579ef78                    vg_hba
hdisk10         none                                None

[svcxd_b1][/]> lspv
hdisk0          00ca02ef74a67ade                    rootvg          active
hdisk1          00ca02ef751884e8                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg    concurrent
hdisk3          000fe4112579ee11                    siteametrovg    concurrent
hdisk4          000fe4112579ee4c                    siteametrovg    concurrent
hdisk5          000fe4112579ee89                    siteametrovg    concurrent
hdisk6          000fe4112579eec4                    sitebglobalvg   concurrent
hdisk7          000fe4112579eefe                    sitebglobalvg   concurrent
hdisk8          000fe4112579ef3d                    sitebglobalvg   concurrent
hdisk9          00ca02ef25b24924                    vg_hbb
hdisk10         none                                None
```

This scenario simulates how to add one more disk in the sitemetrovg volume group and then how to add it to the replicated resource svc_metro in SVC PPRC (Example 5-32).

Example 5-32 SVC PPRC replicated resources

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8 GLOBAL          AUTO
```

First, place the PVID in the disk that was created to be the master VDisk in the nodes from the svc_sitea (Example 5-33).

Example 5-33 Adding the PVID to the new disk

```
[svcxd_a1][/]> lspv
hdisk0          000fe4110889e1a9                    rootvg          active
hdisk1          000fe4112579eb7c                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg    concurrent
hdisk3          000fe4112579ee11                    siteametrovg    concurrent
hdisk4          000fe4112579ee4c                    siteametrovg    concurrent
hdisk5          000fe4112579ee89                    siteametrovg    concurrent
hdisk6          000fe4112579eec4                    sitebglobalvg   concurrent
hdisk7          000fe4112579eefe                    sitebglobalvg   concurrent
hdisk8          000fe4112579ef3d                    sitebglobalvg   concurrent
hdisk9          000fe4112579ef78                    vg_hba
hdisk10         none                                None
[svcxd_a1][/]> chdev -l hdisk10 -a pv=yes
hdisk10 changed
[svcxd_a1][/]> lspv
hdisk0          000fe4110889e1a9                    rootvg          active
hdisk1          000fe4112579eb7c                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg    concurrent
hdisk3          000fe4112579ee11                    siteametrovg    concurrent
hdisk4          000fe4112579ee4c                    siteametrovg    concurrent
hdisk5          000fe4112579ee89                    siteametrovg    concurrent
hdisk6          000fe4112579eec4                    sitebglobalvg   concurrent
hdisk7          000fe4112579eefe                    sitebglobalvg   concurrent
hdisk8          000fe4112579ef3d                    sitebglobalvg   concurrent
hdisk9          000fe4112579ef78                    vg_hba
hdisk10         000fe41163c82dff                    None
```

You must use the same procedure that is shown in Example 5-32 for node svcxd_a2 because it shares the disks from svc_sitea.

Creating the SAN Volume Controller-PPRC relationship

Now, create a SAN Volume Controller-PPRC relationship from the master VDisk to the auxiliary VDisk between the SAN Volume Controller clusters (Example 5-34 on page 185).

Important: You must identify the affinity with the new VDisk and the hdisk. For more information, see “Identifying VDisk SAN Volume Controller to hdisk AIX client” on page 159.

Example 5-34 Adding the SVC PPRC relationship

```
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkrcrelationship -master svc_haxd0009 -aux
haxd_svc_v0009 -cluster B12_4F2 -name svc_disk10

[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk10
id 24
name svc_disk10
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 24
master_vdisk_name svc_haxd0009
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 8
aux_vdisk_name haxd_svc_v0009
primary master
consistency_group_id
consistency_group_name
state inconsistent_stopped
bg_copy_priority 50
progress 0
freeze_time
status online
sync
copy_type metro
```

Then, start the relationship copy (Example 5-35).

Example 5-35 Starting the relationship copy

```
[svcxd_a1][/]> ssh admin@B8_8G4 svctask startrcrelationship svc_disk10

[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk10
id 24
name svc_disk10
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 24
master_vdisk_name svc_haxd0009
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 8
aux_vdisk_name haxd_svc_v0009
primary master
consistency_group_id
consistency_group_name
state inconsistent_copying
bg_copy_priority 50
progress 1
freeze_time
status online
sync
copy_type metro
```

Example 5-36 shows the relationship after the synchronization is completed.

Example 5-36 Relationship synchronized

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk10
id 24
name svc_disk10
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 24
master_vdisk_name svc_haxd0009
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 8
aux_vdisk_name haxd_svc_v0009
primary master
consistency_group_id
consistency_group_name
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

Now, put this relationship in the svc_metro consistent group that you created. In Example 5-37, you can see how to identify a consistent group and how to add a relationship in a consistent group.

Example 5-37 Identifying a consistent group and adding a relationship in a consistent group

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp
id               name              master_cluster_id master_cluster_name aux_cluster_id   aux_cluster_name
primary        state          relationship_count copy_type
0                svc_metro         0000020064009B10  B8_8G4              0000020060A0469E B12_4F2
master         consistent_synchronized 5                 metro
1                svc_global        0000020060A0469E  B12_4F2             0000020064009B10 B8_8G4
master         consistent_synchronized 3                 global

[svcxd_a1][/]> ssh admin@B8_8G4 svctask chrcrelationship -consistgrp svc_metro svc_disk10

[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary master
state consistent_synchronized
relationship_count 5
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
```

```
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
RC_rel_id 24
RC_rel_name svc_disk10
```

Now, add the PVID in the disks from svc_siteb (Example 5-38).

Example 5-38 Adding the PVID in the new disk

```
svcxd_b1][/]> lspv
hdisk0          00ca02ef74a67ade                    rootvg          active
hdisk1          00ca02ef751884e8                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg    concurrent
hdisk3          000fe4112579ee11                    siteametrovg    concurrent
hdisk4          000fe4112579ee4c                    siteametrovg    concurrent
hdisk5          000fe4112579ee89                    siteametrovg    concurrent
hdisk6          000fe4112579eec4                    sitebglobalvg   concurrent
hdisk7          000fe4112579eefe                    sitebglobalvg   concurrent
hdisk8          000fe4112579ef3d                    sitebglobalvg   concurrent
hdisk9          00ca02ef25b24924                    vg_hbb
hdisk10         none                                None
[svcxd_b1][/]> chdev -l hdisk10 -a pv=yes
hdisk10 changed
[svcxd_b1][/]> lspv
hdisk0          00ca02ef74a67ade                    rootvg          active
hdisk1          00ca02ef751884e8                    rootvg          active
hdisk2          000fe4112579ec10                    siteametrovg    concurrent
hdisk3          000fe4112579ee11                    siteametrovg    concurrent
hdisk4          000fe4112579ee4c                    siteametrovg    concurrent
hdisk5          000fe4112579ee89                    siteametrovg    concurrent
hdisk6          000fe4112579eec4                    sitebglobalvg   concurrent
hdisk7          000fe4112579eefe                    sitebglobalvg   concurrent
hdisk8          000fe4112579ef3d                    sitebglobalvg   concurrent
hdisk9          00ca02ef25b24924                    vg_hbb
hdisk10         000fe41163c82dff                    None
```

You must repeat this procedure for the svcxd_b2 node because it shares the disks from svc_siteb.

Discovering HACMP-related information

After you configure the disk, run the **discover** command by using the SMIT panels. Enter the **smitty hacmp** command. Then, select **Extended Configuration** → **Discover HACMP-related Information from Configured Nodes**.

Adding disks to a volume group

To expand the siteametrovg volume group, enter the **smitty cl_admin** command. Then, select **Storage** → **Volume Groups** → **Set Characteristics of a Volume Group** → **Add a Volume to a Volume Group**.

In the next SMIT screen (Figure 5-21), select the volume group that you want to expand, which is sitemetrovg in this case.

```
       Select the Volume Group that will hold the new Logical Volume       ||
| Move cursor to desired item and press Enter. Use arrow keys to scroll.   |
|                                                                          |
|   #Volume Group    Resource Group             Node List                  |
|    siteametrovg    RG_sitea                   svcxd_a1,svcxd_a2,svcxd_b1 |
|    sitebglobalvg   RG_siteb                   svcxd_a1,svcxd_a2,svcxd_b1 |
|    siteametrovg    RG_sitea                   svcxd_a2,svcxd_b1,svcxd_b2 |
|    sitebglobalvg   RG_siteb                   svcxd_a2,svcxd_b1,svcxd_b2 |
|    vg_hbb          <Not in a Resource Group>  svcxd_b1,svcxd_b2          |
```

Figure 5-21 Selecting the volume group to be expanded

In the Physical Volume Names panel (Figure 5-22), select the desired disk.

```
+--------------------------------------------------------------------------+
|                          Physical Volume Names                           |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   000fe41163c82dff ( hdisk10 on all selected nodes )                     |
|                                                                          |
| F1=Help                 F2=Refresh              F3=Cancel                |
| F8=Image                F10=Exit                Enter=Do                 |
| /=Find                  n=Find Next                                      |
+--------------------------------------------------------------------------+
```

Figure 5-22 Selecting the desired physical volume

Check the options that are selected and press Enter (Figure 5-23).

```
                        Add a Volume to a Volume Group

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

                                                        [Entry Fields]
  VOLUME GROUP name                                   siteametrovg
  Resource Group Name                                 RG_sitea
  Node List
svcxd_a1,svcxd_a2,svcxd_b1,svcxd_b2
  Reference node                                      svcxd_a1
  VOLUME names                                        hdisk10
```

Figure 5-23 Checking the options that are selected

After you complete the addition of the disk to the volume group, you see the result in both nodes (Example 5-39).

Example 5-39 Checking the disks in the siteametrovg volume group

```
[svcxd_a1][/]> lsvg -p siteametrovg
siteametrovg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk2            active            1437        1036        288..00..173..287..288
hdisk3            active            1437        1433        288..283..287..287..288
hdisk4            active            1437        1433        288..283..287..287..288
hdisk5            active            1437        1433        288..283..287..287..288
hdisk10           active            637         637         128..127..127..127..128

[svcxd_b1][/]> lsvg -p siteametrovg
siteametrovg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk2            active            1437        1036        288..00..173..287..288
hdisk3            active            1437        1433        288..283..287..287..288
hdisk4            active            1437        1433        288..283..287..287..288
hdisk5            active            1437        1433        288..283..287..287..288
hdisk10           active            637         637         128..127..127..127..128
```

Changing the SVC PPRC replicated resource

After you extend the volume group, add the new SVC PPRC relationship that you created to be managed by PowerHA.

To add the new SVC PPRC relationship:

1. Enter the **smitty hacmp** command.
2. Select **Extended Configuration → Extended Resource Configuration → Extended Resource Configuration → Configure SVC PPRC-Replicated Resources → SVC PPRC Relationships Definition → Add an SVC PPRC Relationship**.
3. In the Add an SVC PPRC Relationship panel (Figure 5-24), enter the required information about the SVC PPRC relationship that you created previously.

```
                    Add an SVC PPRC Relationship

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

                                                      [Entry Fields]
* Relationship Name                                  [svc_disk10]
* Master VDisk Info                                  [svc_haxd0009@B8_8G4]
* Auxiliary VDisk Info                               [haxd_svc_v0009@B12_4F2]
```

Figure 5-24 Adding a PPRC relationship

4. Add this new SVC PPRC relationship to the SVC PPRC replicated resources. In this case, we add it to the replicated resource svc_metro, which we used for the disks from the siteametrovg volume group.

To change an SVC PPRC resource:

a. Run the **smitty hacmp** command.

b. Select **Extended Configuration → Extended Resource Configuration → Extended Resource Configuration → Configure SVC PPRC-Replicated Resources → SVC PPRC-Replicated Resource Configuration → Change/Show an SVC PPRC Resource**.

c. In the SMIT panel, select the resource name (Figure 5-25).

```
+--------------------------------------------------------------------------+
|             Select the SVC PPRC Resource Name to Change/Show             |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   svc_metro                                                              |
|   svc_global                                                             |
|                                                                          |
| F1=Help                 F2=Refresh              F3=Cancel                |
| F8=Image                F10=Exit                Enter=Do                 |
| /=Find                  n=Find Next                                      |
+--------------------------------------------------------------------------+
```

Figure 5-25 Selecting the SAN Volume Controller resource

d. In the Change/Show SVC PPRC Resource panel (Figure 5-26), add the svc_disk10 relationship.

```
                            Change / Show SVC PPRC Resource

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

                                                      [Entry Fields]
  SVC PPRC Consistency Group Name                    svc_metro
  New SVC PPRC Consistency Group Name                []
* Master SVC Cluster Name                            [B8_8G4] +
* Auxiliary SVC Cluster Name                         [B12_4F2] +
* List of Relationships [svc_disk2 svc_disk3 svc_disk4 svc_disk5 svc_disk10]                +
* Copy Type                                          METRO +
* HACMP Recovery Action                              MANUAL +
```

Figure 5-26 Changing the SVC PPRC resource

Example 5-40 shows the created relationships.

Example 5-40 SVC PPRC relationship

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllsrelationship -a
relationship_name MasterVdisk_info AuxiliaryVdisk_info
svc_disk2       svc_haxd0001@B8_8G4 haxd_svc_v0001@B12_4F2
svc_disk3       svc_haxd0002@B8_8G4 haxd_svc_v0002@B12_4F2
svc_disk4       svc_haxd0003@B8_8G4 haxd_svc_v0003@B12_4F2
svc_disk5       svc_haxd0004@B8_8G4 haxd_svc_v0004@B12_4F2
svc_disk6       haxd_svc_v0005@B12_4F2 svc_haxd0005@B8_8G4
svc_disk7       haxd_svc_v0006@B12_4F2 svc_haxd0006@B8_8G4
```

```
svc_disk8       haxd_svc_v0007@B12_4F2 svc_haxd0007@B8_8G4
svc_disk10      svc_haxd0009@B8_8G4 haxd_svc_v0009@B12_4F2
```

Upon completion, you see the SVC PPRC resources as shown in Example 5-41.

Example 5-41 SVC PPRC resources

```
               [svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro      B8_8G4         B12_4F2        svc_disk2 svc_disk3 svc_disk4 svc_disk5 svc_disk10 METRO
MANUAL
svc_global      B12_4F2          B8_8G4          svc_disk6 svc_disk7 svc_disk8                GLOBAL AUTO
```

Synchronizing the cluster

Synchronize the cluster. Enter the **smitty hacmp** command. Then, select **Extended Configuration → Extended Verification and Synchronization**.

5.4.2 Removing disks from PowerHA Enterprise Edition with SAN Volume Controller

To remove a disk in PowerHA with SAN Volume Controller, follow the next steps:

1. In PowerHA
   a. Remove the disk from the volume group.
   b. Remove the SVC PPRC relationship definition in the SVC PPRC Resource.
   c. Remove the SVC PPRC relationship.
   d. Synchronize the cluster.
2. In the SAN Volume Controller Cluster
   a. Remove the SVC PPRC relationship definition from the consistent group.
   b. Remove the SVC PPRC relationship definition.
   c. Remove the VDisks definitions from both SAN Volume Controller clusters.

5.5 Testing PowerHA Enterprise Edition with SAN Volume Controller

This section explains the testing to perform with PowerHA Enterprise Edition with SAN Volume Controller. This section includes the following test scenarios:

- Failure site (soft and hard)
- Storage loss
- Convert replication mode (metro versus global)
- Loss of replication links (auto versus manual)

5.5.1 Monitoring the SVC PPRC relationship

By using the SAN Volume Controller console (Figure 5-27 on page 192), you can check the relationships for consistency group svc_metro that the Primary column is marked as Master. In this case, SVC PPRC replication goes from SAN Volume Controller cluster B8_8G4 to SAN Volume Controller cluster B12_4F2 because the master cluster is the SAN Volume Controller

cluster B8_8G4. The consistency group svc_global in the Primary column is marked as Master. In this case, SVC PPRC replication goes from SAN Volume Controller cluster B12_4F2 to the B8_8G4 SAN Volume Controller cluster because their master cluster is the SAN Volume Controller cluster B12_4F2. In the State column, you can check the status of the copy between the SAN Volume Controller clusters.

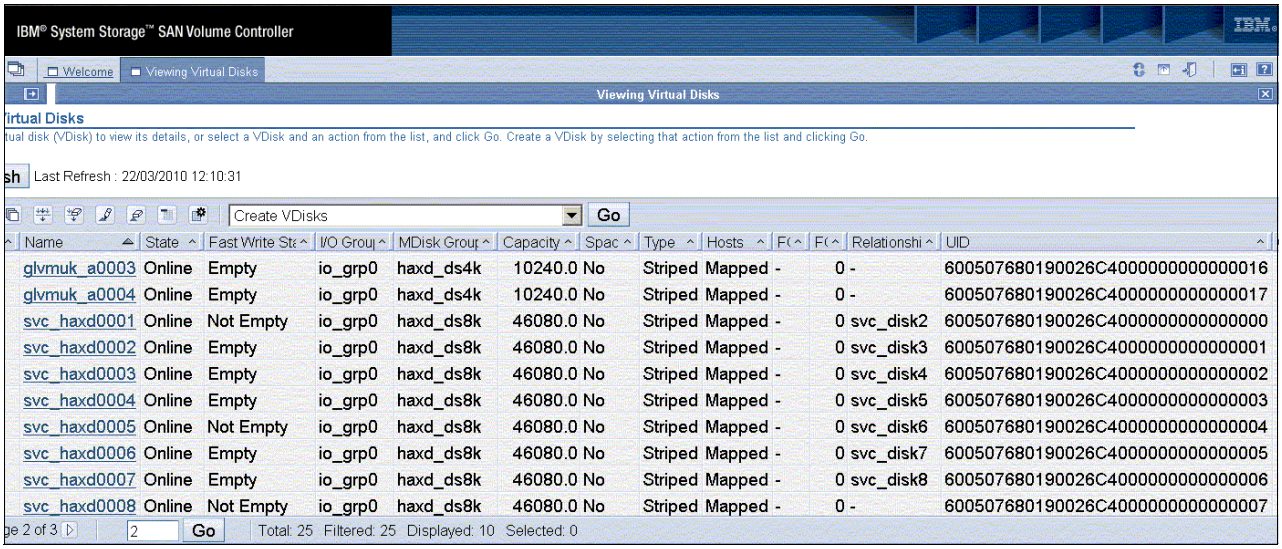


Figure 5-27 SVC PPRC relationship

Example 5-42 shows how to check the status of the svc_disk2 relationship by using the command line.

Example 5-42 SVC PPRC relationship of svc_disk2

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk2
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
primary master
consistency_group_id 0
consistency_group_name svc_metro
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

Verify the status of the SVC PPRC relationship by using the information in the following sections:

master_cluster_name
The name of the master SAN Volume Controller cluster B8_8G4.

aux_cluster_name The name of the auxiliary SAN Volume Controller cluster B12_4F2.

primary The way of the SVC PPRC relationship, in this case master to auxiliary.

consistency_group_name
The name of the consistency group that is included.

state The status of the copy between both SAN Volume Controller clusters.

5.5.2 Site failure (soft and hard)

This section provides the following examples of site failures:

- Soft test through a C-SPOC operation
- Hard test through a **halt -q** command

Soft test

Test the PowerHA Enterprise Edition with SAN Volume Controller by using the C-SPOC operation. As shown in Example 5-43, resource group RG_sitea is online on node svcxd_a1 on site svc_sitea. Resource group RG_siteb is online on node svcxd_b1 on site svc_siteb.

Example 5-43 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Resource group RG_sitea manages the SVC PPRC replicated resource svc_metro. Resource group RG_siteb manages the SVC PPRC replicated resource svc_global. Example 5-44 shows the SVC PPRC replicated resources.

Example 5-44 SVC PPRC replicated resource

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro      B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global     B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8           GLOBAL AUTO
```

In the first test, we move the resource group RG_sitea to the node svcxd_a2 on the same site svc_sitea. To move the resource group:

1. Enter the **smitty cl_admin** command.
2. Select **Resource Groups and Applications → Move a Resource Group to Another Node / Site → Move Resource Groups to Another Node**.
3. For Resource Group(s) to be Moved, select **RG_sitea,** and for Destination Node select **svcxd_a2** (Figure 5-28).

```
                    Move Resource Group(s) to Another Node

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

                                                      [Entry Fields]
  Resource Group(s) to be Moved                      RG_sitea
  Destination Node                                   svcxd_a2
```

Figure 5-28 Moving the resource group to another node

4. Check that RG_sitea is online on the svcxd_a2 node (Example 5-45).

Example 5-45 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                      svcxd_a1@svc_s
               ONLINE                       svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-46 shows that the status of the relationship did not change because the node from svc_sitea shares the disks.

Example 5-46 SVC PPRC relationship status

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk2
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
primary master
consistency_group_id 0
```

```
consistency_group_name svc_metro
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

5. Move RG_sitea resource group to the svcxd_b1 node on the svc_siteb site (Figure 5-29).

```
                    Move Resource Group(s) to Another Node

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

                                                      [Entry Fields]
  Resource Group(s) to be Moved                      RG_sitea
  Destination Node                                   svcxd_b1
```

Figure 5-29 Moving the resource group

6. Check that the RG_sitea is online on node svcxd_b1 (Example 5-47).

Example 5-47 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE SECONDARY             svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE                       svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-48 shows that the status of the relationship changed. Now the SAN Volume Controller cluster B8_8G4 works as auxiliary for B12_4F2 because the Primary field is now aux. The field state is still in the consistent_synchronized state.

Example 5-48 SVC PPRC relationship status

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk2
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
```

```
aux_vdisk_name haxd_svc_v0001
primary aux
consistency_group_id 0
consistency_group_name svc_metro
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

7. Move the resource group RG_sitea to node svcxd_b2 on site svc_siteb (Figure 5-30).

```
       Move Secondary Instance(s) of Resource Group(s) to Another Node

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

                                                      [Entry Fields]
  Resource Group(s) to be Moved                      RG_sitea
  Destination Node                                   svcxd_b2
```

Figure 5-30 Moving the resource group to another node

Example 5-49 shows that RG_sitea is online in the node svcxd_b2 on site svc_siteb.

Example 5-49 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE SECONDARY            svcxd_a1@svc_s
               OFFLINE                     svcxd_a2@svc_s
               ONLINE                      svcxd_b2@svc_s
               OFFLINE                     svcxd_b1@svc_s

RG_siteb       ONLINE                      svcxd_b1@svc_s
               OFFLINE                     svcxd_b2@svc_s
               ONLINE SECONDARY            svcxd_a2@svc_s
               OFFLINE                     svcxd_a1@svc_s
```

Example 5-50 shows that the status of the relationship did not change, as the nodes from svc_siteb share disks.

Example 5-50 SVC PPRC relationship status

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk2
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
```

```
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
primary aux
consistency_group_id 0
consistency_group_name svc_metro
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

8. Return the resource group RG_sitea to node svcxd_a1 on site svc_sitea (Figure 5-31).

```
                    Move Resource Group(s) to Another Node

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

                                                        [Entry Fields]
  Resource Group(s) to be Moved                        RG_sitea
  Destination Node                                     svcxd_a1
```

Figure 5-31 Moving the resource group back

Example 5-51 shows that RG_sitea is online in node svcxd_a1 on site svc_sitea.

Example 5-51 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-52 shows that the status of the relationship changed. Now the SAN Volume Controller cluster B12_4F2 works as an auxiliary of B8_8G4 because the Primary field is now master. The State field still is in the consistent_synchronized state.

Example 5-52 SVC PPRC relationship status

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcrelationship svc_disk2
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
```

```
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
primary master
consistency_group_id 0
consistency_group_name svc_metro
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
copy_type metro
```

Hard test

Test PowerHA for SAN Volume Controller by using the **halt -q** command to simulate node and site problems. Example 5-53 shows that the resource group RG_siteb is online on node svcxd_b1 on site svc_siteb. This resource group manages the svc_global replicated resource for SVC PPRC.

Example 5-53 Resource group status

```
[svcxd_b1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-54 shows the status of the svc_global consistency group that is managed from RG_siteb.

Example 5-54 SVC PPRC consistency group status

```
[svcxd_b1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
```

```
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Status: You can check the status of the relationship by using the command line of the SVC PPRC consistency group or by using the SVC PPRC relationship. For more information, see the *SVC CLI Guide*, SC26-7903.

In the first test, issue the **halt -q** command in node svcxd_b1 (Example 5-55).

Example 5-55 Simulating failure of a node

```
[svcxd_b1][/]> halt -q
....Halt completed....
```

Example 5-56 shows that RG_siteb is online in the svcxd_b2 node on the svc_siteb site.

Example 5-56 Resource group status

```
[svcxd_b2][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       OFFLINE                      svcxd_b1@svc_s
               ONLINE                       svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-57 shows that the status of the consistency group did not change because the nodes from svc_siteb share disks.

Example 5-57 SVC PPRC consistency group status

```
[svcxd_b2][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
status
```

```
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Enter the **halt -q** command on the svcxd_b2 node, and simulate a failure of the svc_siteb site (Example 5-58).

Example 5-58 Simulating a failure of a node

```
[svcxd_b2][/]> halt -q
....Halt completed....
```

Example 5-59 shows that RG_siteb is online on the svcxd_a2 node on the svc_sitea site.

Example 5-59 Resource group status

```
[svcxd_a2][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       OFFLINE                      svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE                       svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-60 shows that the status of the consistency group changed. Now the SAN Volume Controller cluster B12_4F2 works as an auxiliary of B8_8G4 because the primary field is now `aux`. The state field still is in the `consistent_synchronized` state.

Example 5-60 SVC PPRC consistency group status

```
[svcxd_a2][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary aux
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
```

```
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Enter the **halt -q** command on the svcxd_a2 node, and simulate a failure of this node (Example 5-61).

Example 5-61 Simulating failure of a node

```
[svcxd_b2][/]> halt -q
....Halt completed....
```

Example 5-62 shows that RG_siteb is online on the svcxd_a1 node on the svc_sitea site.

Example 5-62 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                     svcxd_a1@svc_s
               OFFLINE                    svcxd_a2@svc_s
               OFFLINE                    svcxd_b2@svc_s
               OFFLINE                    svcxd_b1@svc_s

RG_siteb       OFFLINE                    svcxd_b1@svc_s
               OFFLINE                    svcxd_b2@svc_s
               OFFLINE                    svcxd_a2@svc_s
               ONLINE                     svcxd_a1@svc_s
```

Example 5-63 shows that the status of the consistency group did not change because the nodes from svc_sitea share disks.

Example 5-63 SVC PPRC consistency group status

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary aux
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
```

```
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Enter the **halt -q** command on the svcxd_a1 node, and simulate a failure of the svc_sitea site (Example 5-64). In this scenario, we test both resource groups.

Example 5-64 Simulating the failure of a node

```
[svcxd_a1][/]> halt -q
....Halt completed....
```

Example 5-65 shows that RG_siteb is online on the svcxd_b1 node on the svc_siteb site, and RG_sitea is online on the svcxd_b2 node on the svc_siteb site.

Example 5-65 Resource group status

```
[svcxd_b1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                      svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE                       svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-66 shows that the status of the consistency group changed. Now the SAN Volume Controller cluster B8_8G42 works as an auxiliary of B12_4F2 because the primary field is now `master`. The state field is still in the `consistent_synchronized` state.

Example 5-66 SVC PPRC consistency group status

```
[svcxd_b1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
```

```
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Example 5-67 shows that the status of the consistency group changed. Now the SAN Volume Controller cluster B8_8G42 works as an auxiliary of B12_4F2 because the primary field is now `master`. The state field is still in the `consistent_synchronized` state.

Example 5-67 SVC PPRC consistency group status

```
[svcxd_b2][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary aux
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

5.5.3 Storage loss

This section provides an example of storage loss. To simulate the storage loss in this scenario, remove the VDisks map from the nodes of the svc_sitea. Remove the VDisk that is used from the siteametrovg volume group that is managed by the RG_sitea resource group. Example 5-68 shows the hdisks that are used in siteametrovg.

Example 5-68 Disks listing from siteametrovg

```
[svcxd_a1][/]> lsvg -p siteametrovg
siteametrovg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk2            active            1437        1036        288..00..173..287..288
hdisk3            active            1437        1433       288..283..287..287..288
hdisk4            active            1437        1433       288..283..287..287..288
hdisk5            active            1437        1433       288..283..287..287..288
```

Example 5-69 shows that RG_sitea is online on the svcxd_a1 node on the svc_sitea site.

Example 5-69 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

As shown in Example 5-70, remove the VDisk maps from the disks that are used in the siteametrovg volume group from the svcxd_a1 and svcxd_b1 nodes.

Example 5-70 Removing the VDisk mapping

```
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_1_A1 svc_haxd0001
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_2_A2 svc_haxd0001
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_1_A1 svc_haxd0002
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_2_A2 svc_haxd0002
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_1_A1 svc_haxd0003
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_2_A2 svc_haxd0003
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_1_A1 svc_haxd0004
[svcxd_a1][/]>ssh admin@B8_8G4 svctask rmvdiskhostmap -host SVC_550_2_A2 svc_haxd0005
```

As shown in Example 5-71, you now see the AIX errorlog messages about a physical disk that is missing, an lvm I/O error, and quorum lost. The `hacmp.out` file shows messages about a volume group failure.

Example 5-71 AIX errorlog and hacmp.log outputs

```
From hacmp.out:

HACMP Event Preamble
-----------------------------------------------------------------------------

Enqueued rg_move release event for resource group 'RG_sitea'.

Reason for recovery of Primary instance of Resource group 'RG_sitea' from
TEMP_ERROR state on node 'svcxd_a1' was 'Volume group failure'.

Enqueued rg_move acquire event for resource group 'RG_sitea'.

Cluster Resource State Change Complete Event has been enqueued.
-----------------------------------------------------------------------------
Mar 18 10:53:49 EVENT START: resource_state_change svcxd_a1

From errorlog:
LABEL:          LVM_IO_FAIL
Date/Time:       Thu Mar 18 10:53:46 2010
```

```
LABEL:          LVM_SA_QUORCLOSE
Date/Time:       Thu Mar 18 10:53:46 2010

LABEL:          LVM_SA_PVMISS
Date/Time:       Thu Mar 18 10:53:46 2010
```

Example 5-72 shows that RG_sitea has moved and is now online on the svcxd_b1 node on the svc_siteb site.

Example 5-72 Resource group status

```
[svcxd_b1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE SECONDARY             svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE                       svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE SECONDARY             svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Example 5-73 shows that the status of the consistency group changed. Now the SAN Volume Controller cluster B8_8G42 works as an auxiliary of B12_4F2 because the primary field is now `aux`. The state field is still in an `consistent_synchronized` state because, in this test, we remove the map from the VDisk to the hosts, and the relationship still exists.

Example 5-73 SVC PPRC consistency group status

```
[svcxd_b1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary aux
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

As shown in Example 5-74, add the VDisk maps to the disks used in the siteametrovg volume group on the svcxd_a1 and svcxd_b1 nodes.

Example 5-74 Adding VDisk mapping

```
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -host SVC_55O_1_A1 svc_haxd0001
Virtual Disk to Host map, id [0], successfully created
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -host SVC_55O_1_A1 svc_haxd0002
Virtual Disk to Host map, id [1], successfully created
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -host SVC_55O_1_A1 svc_haxd0003
Virtual Disk to Host map, id [2], successfully created
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -host SVC_55O_1_A1 svc_haxd0004
Virtual Disk to Host map, id [3], successfully created

[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -force -host SVC_55O_2_A2 svc_haxd0001
Virtual Disk to Host map, id [0], successfully created
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -force -host SVC_55O_2_A2 svc_haxd0002
Virtual Disk to Host map, id [1], successfully created
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -force -host SVC_55O_2_A2 svc_haxd0003
Virtual Disk to Host map, id [2], successfully created
[svcxd_a1][/]> ssh admin@B8_8G4 svctask mkvdiskhostmap -force -host SVC_55O_2_A2 svc_haxd0004
Virtual Disk to Host map, id [3], successfully created
```

Example 5-75 shows that, after you run the **cfgmgr** command, you can access the disk again from the SAN Volume Controller Cluster in the node on the svc_sitea site.

Example 5-75 Disks

```
[svcxd_a1][/]> cfgmgr
[svcxd_a1][/]> lsdev -Cc disk
hdisk0 Available 23-T1-01 MPIO FC 2145
hdisk1 Available 23-T1-01 MPIO FC 2145
hdisk2 Available 23-T1-01 MPIO FC 2145
hdisk3 Available 23-T1-01 MPIO FC 2145
hdisk4 Available 33-T1-01 MPIO FC 2145
hdisk5 Available 33-T1-01 MPIO FC 2145
hdisk6 Available 33-T1-01 MPIO FC 2145
hdisk7 Available 33-T1-01 MPIO FC 2145
hdisk8 Available 33-T1-01 MPIO FC 2145
hdisk9 Available 23-T1-01 MPIO FC 2145

[svcxd_a2][/]> cfgmgr
[svcxd_a2][/]> lsdev -Cc disk
hdisk0 Available 14-T1-01 MPIO FC 2145
hdisk1 Available 14-T1-01 MPIO FC 2145
hdisk2 Available 14-T1-01 MPIO FC 2145
hdisk3 Available 14-T1-01 MPIO FC 2145
hdisk4 Available 14-T1-01 MPIO FC 2145
hdisk5 Available 14-T1-01 MPIO FC 2145
hdisk6 Available 14-T1-01 MPIO FC 2145
hdisk7 Available 14-T1-01 MPIO FC 2145
hdisk8 Available 14-T1-01 MPIO FC 2145
hdisk9 Available 14-T1-01 MPIO FC 2145
```

Example 5-76 shows the status of the resource group after we move RG_sitea to the svcxd_a1 node on the svc_sitea site.

Example 5-76 Resource group status

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE SECONDARY             svcxd_a1@svc_s
```

Example 5-77 shows that the status of the consistency group changed. Now the SAN Volume Controller cluster B12 _4F2 works as an auxiliary of B8_8G4 because the primary field is now `master`. The state field still is in the `consistent_synchronized` state.

Example 5-77 SVC PPRC consistency group status

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary master
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

5.5.4 Convert replication mode

This section provides an example of how to convert SVC PPRC replication mode in a PowerHA for SAN Volume Controller environment. It includes the following tests scenarios:

- Convert Global to Metro
- Convert Metro to Global

We use the svc_global SVC PPRC Resource as an example in this test scenario to convert the SVC PPRC replication mode. Example 5-78 lists the SVC PPRC resources.

Example 5-78 SVC PPRC resources

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8           GLOBAL AUTO
```

Converting Global to Metro

In this first example, we change the consistency group svc_global from global to metro SVC PPRC replication mode. Then, we check the status of the consistency group svc_global. Example 5-79 shows the current status of the copy_type field (global) and the state field (consistent_synchronized) from the svc_global consistency group.

Example 5-79 Consistency group status

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

To change an SVC PPRC resource, enter the **smitty hacmp** command. Select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure SVC PPRC-Replicated Resources → SVC PPRC-Replicated Resource Configuration → Change/Show an SVC PPRC Resource → select the SVC PPRC Resource** (Figure 5-32).

```
+--------------------------------------------------------------------------+
|             Select the SVC PPRC Resource Name to Change/Show             |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   svc_metro                                                              |
|   svc_global                                                             |
|                                                                          |
| F1=Help                 F2=Refresh              F3=Cancel                |
| F8=Image                F10=Exit                Enter=Do                 |
| /=Find                  n=Find Next                                      |
+--------------------------------------------------------------------------+
```

Figure 5-32 SVC PPRC resources

In the Change/Show SVC PPRC Resource panel (Figure 5-33), change the Copy Type field to METRO.

```
                        Change / Show SVC PPRC Resource

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

                                                      [Entry Fields]
  SVC PPRC Consistency Group Name                     svc_global
  New SVC PPRC Consistency Group Name                 []
* Master SVC Cluster Name                             [B12_4F2]                +
* Auxiliary SVC Cluster Name                          [B8_8G4]                 +
* List of Relationships                             [svc_disk6 svc_disk7 svc_disk8] +
* Copy Type                                          METRO                    +
* HACMP Recovery Action                               AUTO                    +
```

Figure 5-33 Changing the copy type in an SVC PPRC resource

After you change the SVC PPRC resource, list the resource (Example 5-80).

Example 5-80 SVC PPRC resources

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8           METRO AUTO
```

Now, synchronize the cluster.

Tip: In this example, we use the **cl_verify_svcpprc_config** utility to verify the cluster (Example 5-81 on page 210). PowerHA synchronization processes always call this utility when they use SVC PPRC resources.

Example 5-81 Verifying the SVC PPRC configuration

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/utils]> ./cl_verify_svcpprc_config
Verifying HACMP-SVCPPRC configuration...
cl_verify_svcpprc_config: Checking available nodes
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_a1
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_a2
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_b1
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_b2
cl_verify_svcpprc_config: Checking available SVCs
cl_verify_svcpprc_config: Checking license, release level and disk map for SVC B8_8G4 at 10.12.5.55
cl_verify_svcpprc_config: Checking license, release level and disk map for SVC B12_4F2 at 10.114.63.250
cl_verify_svcpprc_config: Checking consistency groups
cl_verify_svcpprc_config: Checking consistency group svc_metro
cl_verify_svcpprc_config: Checking consistency group svc_global
cl_verify_svcpprc_config: Checking resource groups.
cl_verify_svcpprc_config: Checking resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a1 for resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a2 for resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b2 for resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b1 for resource group RG_sitea
cl_verify_svcpprc_config: Checking volume group siteametrovg in group RG_sitea on site svc_sitea
cl_verify_svcpprc_config: Checking volume group siteametrovg in group RG_sitea on site svc_siteb
cl_verify_svcpprc_config: Checking resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b1 for resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b2 for resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a2 for resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a1 for resource group RG_siteb
cl_verify_svcpprc_config: Checking volume group sitebglobalvg in group RG_siteb on site svc_sitea
cl_verify_svcpprc_config: Checking volume group sitebglobalvg in group RG_siteb on site svc_siteb
cl_verify_svcpprc_config: Verifying consistency groups against the SVC configuration
cl_verify_svcpprc_config: Establishing consistency group svc_metro
cl_verify_svcpprc_config: WARNING: Consistency Group svc_metro already exists
cl_verify_svcpprc_config: Verifying relationships for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk2 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk2 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk3 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk3 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk4 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk4 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk5 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk5 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Establishing consistency group svc_global
cl_verify_svcpprc_config: WARNING: Consistency Group svc_global already exists
cl_verify_svcpprc_config: Verifying relationships for consistency group svc_global
cl_verify_svcpprc_config: Verifying relationship svc_disk6 in consistency group svc_global
cl_verify_svcpprc_config: Relationship svc_disk6 already exists for consistency group svc_global
cl_verify_svcpprc_config: Verifying relationship svc_disk7 in consistency group svc_global
cl_verify_svcpprc_config: Relationship svc_disk7 already exists for consistency group svc_global
cl_verify_svcpprc_config: Verifying relationship svc_disk8 in consistency group svc_global
cl_verify_svcpprc_config: Relationship svc_disk8 already exists for consistency group svc_global
HACMP-SVCPPRC configuration verified successfully.  Status=0
```

After you synchronize the cluster, check the status of the svc_global consistency group (Example 5-82).

Example 5-82 The consistency group

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state inconsistent_stopped
relationship_count 3
freeze_time
status
sync
copy_type metro
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

The copy type field changed to `metro`, and the state is `inconsistent_stopped`. This change means that the copy of the master cluster is accessible for read and write I/O, but the copy of the aux cluster is not accessible. Start a copy process to make a consistent copy. Example 5-83 shows how to start the copy process of the svc_global consistency group.

Example 5-83 Starting and monitoring a consistency group

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/utils]> ssh admin@B8_8G4 svctask startrcconsistgrp svc_global

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/utils]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state inconsistent_copying
relationship_count 3
freeze_time
status
sync
copy_type metro
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

After the copy process finishes, check the status of the svc_global consistency group (Example 5-84).

Example 5-84 Consistency group status

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/utils]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type metro
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Converting the consistency group from metro mode to global mode

In this second example, we change the svc_global consistency group from metro mode to global replication mode for SVC PPRC. As shown in Example 5-84 on page 212, check the status of svc_global. Notice that the status of the copy_type field is `metro` and the state field is `consistent_synchronized`.

First, change the SVC PPRC resource. Enter the **`smitty hacmp`** command. Then, select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resources Configuration** → **Configure SVC PPRC-Replicated Resources** → **SVC PPRC-Replicated Resource Configuration** → **Change/Show an SVC PPRC Resource**. For the SVC PPRC Resource, select **svc_global** → **Change / Show SVC PPRC Resource** (Figure 5-34).

```
                          Change / Show SVC PPRC Resource

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

                                                  [Entry Fields]
  SVC PPRC Consistency Group Name                 svc_global
  New SVC PPRC Consistency Group Name             []
* Master SVC Cluster Name                         [B12_4F2] +
* Auxiliary SVC Cluster Name                      [B8_8G4] +
* List of Relationships                           [svc_disk6 svc_disk7 svc_disk8] +
* Copy Type                                       GLOBAL+
* HACMP Recovery Action                           AUTO +
```

Figure 5-34 Changing the SVC PPRC resource

After you change the SVC PPRC resource, list the resource as shown in Example 5-85.

Example 5-85 SVC PPRC resources

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8           GLOBAL AUTO
```

Now synchronize the cluster. After you synchronize the cluster, check the status of the svc_global consistency group (Example 5-86).

Example 5-86 Consistency group status

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state inconsistent_stopped
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

As performed in the first test, start a copy process to make a consistent copy. Example 5-87 shows the process for copying and monitoring the svc_global consistency group.

Example 5-87 Starting and monitoring a consistency group

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state inconsistent_copying
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
```

```
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

After you complete the copy, check status of the svc_global consistency group (Example 5-88).

Example 5-88 Consistency group status after the copy

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

5.5.5 Lost of the replication links (auto versus manual)

This section provides examples of the loss of the replication links between the SAN Volume Controller clusters. This section includes the following tests:

- Loss of the replication links in the SVC PPRC resource with manual recovery action
- Loss of the replication links in the SVC PPRC resource with auto recovery action

In this scenario, we configure the svc_metro consistency group as the MANUAL recovery action and svc_global as the AUTO recovery action (Example 5-89).

Example 5-89 SVC PPRC resources

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc -a
svcpprc_consistencygrp MasterCluster   AuxiliaryCluster relationships   CopyType        RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4 svc_disk5 METRO MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8           GLOBAL AUTO
```

Loss of replication links in SVC PPRC resource with manual recovery

In this scenario, you test the manual recovery action that is configured in the svc_metro SVC PPRC resource. This resource is managed by the RG_sitea resource group. In our test, as shown in Figure 5-35, the RG_sitea resource group is on the svcxd_a1 node on the svc_sitea site.

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                      svcxd_a1@svc_s
               OFFLINE                     svcxd_a2@svc_s
               ONLINE SECONDARY            svcxd_b2@svc_s
               OFFLINE                     svcxd_b1@svc_s

RG_siteb       ONLINE                      svcxd_b1@svc_s
               OFFLINE                     svcxd_b2@svc_s
               ONLINE SECONDARY            svcxd_a2@svc_s
               OFFLINE                     svcxd_a1@svc_s
```

Figure 5-35 Resource Group status

Before you start this test, check that the svc_metro consistency group is Consistent Synchronized between SAN Volume Controller cluster B8_8G4 and B12_4F2 (Figure 5-36).

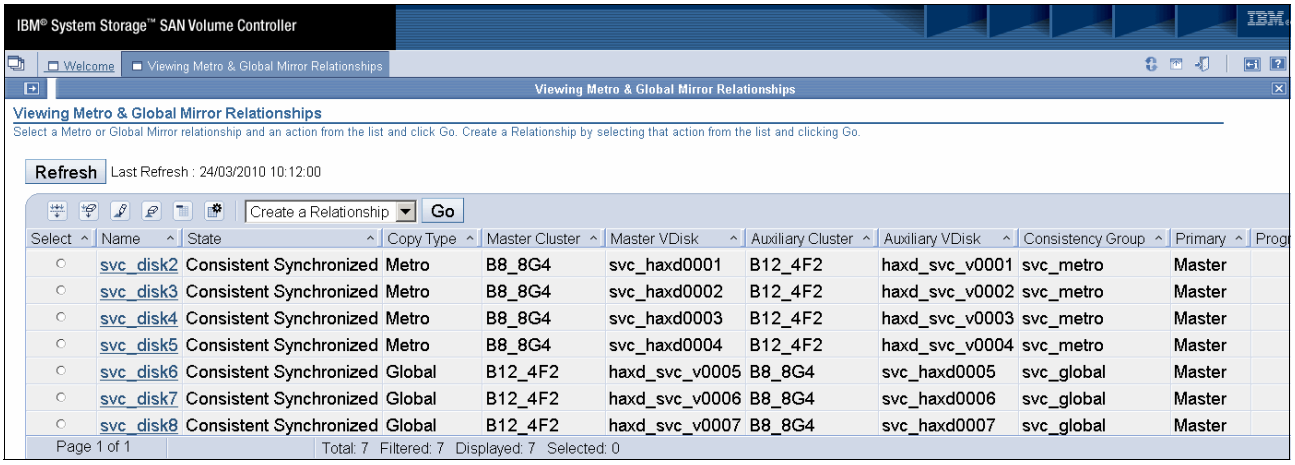


Figure 5-36 Viewing the SVC PPRC relationships

After stopping the replication links between the SAN Volume Controller clusters B8_8G4 and B12_4F2, the state of the consistency group changed. Example 5-90 shows the `idling_disconnected` state in the svc_metro consistency group on the B8_8G4 cluster for SAN Volume Controller. In this state, the VDisks in this half of the consistency group are all operating in the primary role B8_8G4 and can accept read or write I/O operations.

Example 5-90 SVC PPRC consistency group

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
```

```
aux_cluster_name
primary master
state idling_disconnected
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

Example 5-91 shows the `consistent_disconnected` state for the svc_metro consistency group on the B12_4F2 cluster for SAN Volume Controller. In this state, the VDisks in this half of the consistency group are all operating in the secondary B12_4F2 role and can accept read I/O operations, but not write I/O operations.

Example 5-91 SVC PPRC consistency group

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary master
state consistent_disconnected
relationship_count 4
freeze_time 2010/03/23/15/25/00
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

Reference: For more information about the SVC PPRC states, see 5.6, “Troubleshooting PowerHA Enterprise Edition for SAN Volume Controller” on page 230.

If you prefer, you can check the state of the consistency group relationship in the SAN Volume Controller console. Figure 5-37 shows the cluster B8_G4 for SAN Volume Controller for the svc_metro consistency group in the `idling_disconnected` state.

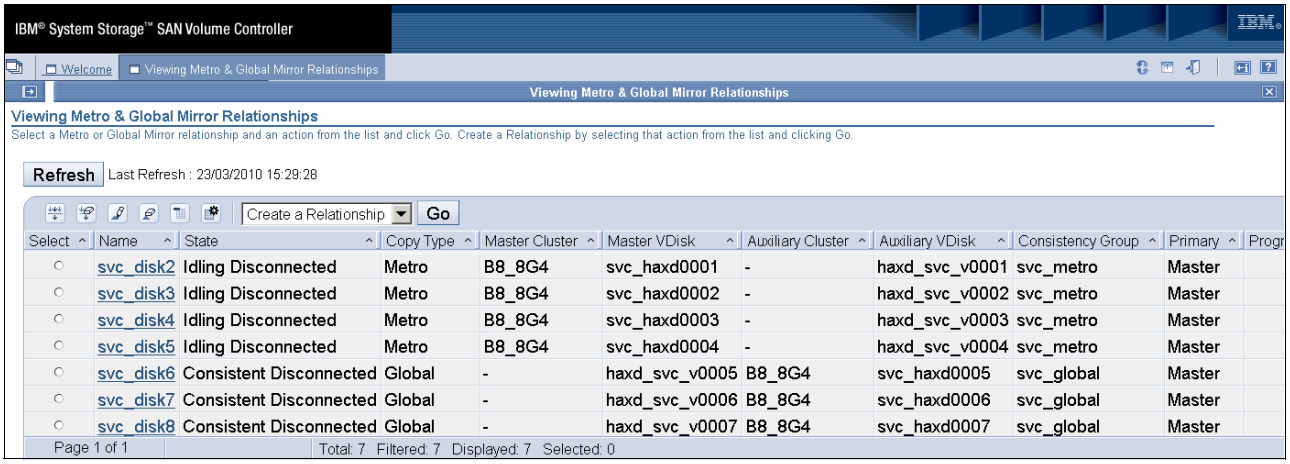


Figure 5-37 Viewing the SVC PPRC relationships

In Figure 5-38, we see in SAN Volume Controller cluster B12_4F2 for the consistency group svc_metro the consistent_disconnected state.

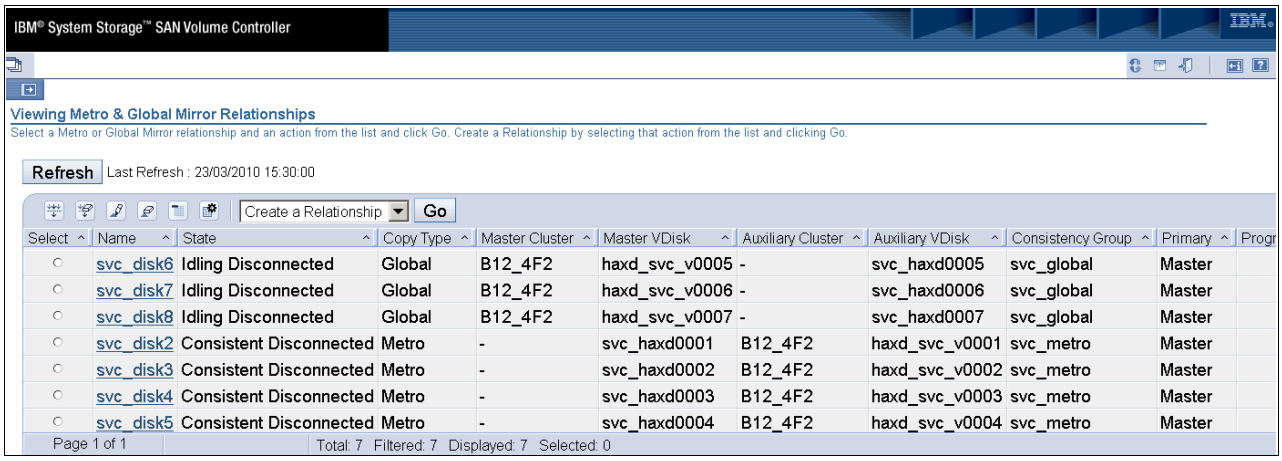


Figure 5-38 Viewing the SVC PPRC relationships

Now, to simulate a site failure on nodes from site svc_sitea, run the **`reboot -q`** command (Example 5-92).

Example 5-92 Simulating the failure of a node

```
[svcxd_a1][/]> reboot -q

[svcxd_a2][/]> reboot -q
```

After the failure, RG_sitea resource group is in the ERROR state at the svcxd_b1 node on the svc_siteb site (Figure 5-39).

```
[svcxd_b2][/var/hacmp/log]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                      svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ERROR                        svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Figure 5-39 Resource group status

The RG_sitea resource group cannot bring ONLINE on the svcxd-b1 site because of the consistent_disconnected state of the svc_metro consistency group in the SAN Volume Controller cluster B12_4F2.

In the hacmp.out file from the svcxd_b1 node, you can check the following instructions about the SVC PPRC resource in MANUAL Recovery Action (Example 5-93).

Example 5-93 The hacmp.out file from the svcxd_b1 node

```
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 admin@10.12.5.55 svctask stoprcconsistgrp -access svc_metro

        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 admin@10.12.5.55 svctask stoprcconsistgrp -access svc_metro

        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 admin@10.12.5.55 svcinfo lsrcconsistgrp -delim : svc_metro
             ssh admin@10.114.63.250 svcinfo lsrcconsistgrp -delim : svc_metro

    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

B8_8G4 Master svc_sitea 10.12.5.55  B12_4F2
B12_4F2 Master svc_siteb 10.114.63.250  B8_8G4

        so if you want the RG's online on the secondary site you would pick 10.114.63.250 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 admin@10.12.5.55 svctask startrcconsistgrp -force -primary [ aux | master ]
svc_metro
        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 admin@10.12.5.55 svcinfo lsrcconsistgrp -delim : svc_metro
             ssh admin@10.114.63.250 svcinfo lsrcconsistgrp -delim : svc_metro
             Sample output from one of the above commands:
               id:255
               name:FVT_CG1
               master_cluster_id:0000020064805 6E6
               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.
```

Now, run the **stoprcconsistgrp -access** command to enable I/O to the backup VDisks in the SAN Volume Controller Cluster B12_4F2. Then, you can see the `idling_disconnected` state for the svc_metro consistency group (Example 5-94).

Example 5-94 Enabling I/O to back up SAN Volume Controller cluster VDisk

```
[svcxd_b2][/]> ssh admin@B12_4F2 svctask stoprcconsistgrp -access svc_metro

[svcxd_b2][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary
state idling_disconnected
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

After you change the svc_metro consistency group to the `idling_disconnected` state, bring the resource group online by manually using C-SPOC. In this scenario, we changed the state of the consistency group, and the resource group is brought online automatically (Figure 5-40).

```
[svcxd_b2][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                      svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               ONLINE                       svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Figure 5-40 Resource group status

After you reconnect all physical PPRC links, the state of the svc_metro consistency group changes to `idling` in both SAN Volume Controller clusters (Example 5-95). In this state, the master disks and the auxiliary disks are operating in the primary role. Then, both are accessible for write I/O. In this state, the relationship or consistency group accepts a **`Start`** command.

Example 5-95 SVC PPRC consistency group

```
[svcxd_b1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary
state idling
relationship_count 4
freeze_time
status
sync out_of_sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5

[svcxd_b1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
```

```
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary
state idling
relationship_count 4
freeze_time
status
sync out_of_sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

Next, restart the svc_metro consistency group now that the RG_sitea resource group is online on the svc_siteb and accessing VDisks from SAN Volume Controller cluster B12_4F2.

For the svc_metro consistency group, the SAN Volume Controller cluster B12_4F2 is `aux`, and SAN Volume Controller cluster B8_8G4 is `master`. Restart the copy from the SAN Volume Controller cluster B12_4F2 to the SAN Volume Controller cluster B8_8G4 (Example 5-96).

Example 5-96 Restarting the consistency group svc_metro

```
[svcxd_b1][/]>  ssh admin@B8_8G4 svctask startrcconsistgrp -force -primary aux svc_metro

[svcxd_b1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary aux
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

Now after you check the `consistent_synchronized` state of the svc_metro consistency group, and the cluster services are running on the nodes from the svc_sitea site, use C-SPOC to move the RG_sitea resource group to the svcxd_a1 node (Figure 5-41).

```
[svcxd_b1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Figure 5-41 Resource froup status

Now the resource group is moved to the svc_sitea site, and the svc_metro consistency group is changed from primary `aux` to primary `master` (that is, from B8_8G4 to B12_4F2) (Example 5-97).

Example 5-97 SVC PPRC consistency group

```
[svcxd_b1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary master
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

Loss of replication links in SVC PPRC resource with auto recovery

Before you test this scenario, when you select the SVC PPRC Resource with auto recovery action, you see the message alert that is shown in Figure 5-42.

```
Recovery Action is currently set to "AUTO" and corresponds to the
Enter "y" to continue or "n" to abort [n]:n a site failure.  However, it must be recognized
that HACMP 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 HACMP/XD that
will indeed preclude a partitioned cluster.)  Additionally, if the disk subsystem supports it,
there should be Copy Services Servers at each site.
```

Figure 5-42 Selecting auto recovery action

Important: To avoid a partitioned cluster, have multiple xd_ip networks provide independent heartbeat paths and, if it is possible, have different vendors provide each network.

In this scenario, we test the AUTO recovery action that is configured in the svc_global SVC PPRC resource. This resource is managed by the RG_siteb resource group. In this test, as shown in Figure 5-43, the RG_siteb resource group is on the svcxd_b1 node on the svc_siteb site.

```
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                      svcxd_a1@svc_s
               OFFLINE                     svcxd_a2@svc_s
               OFFLINE                     svcxd_b2@svc_s
               OFFLINE                     svcxd_b1@svc_s

RG_siteb       ONLINE                      svcxd_b1@svc_s
               OFFLINE                     svcxd_b2@svc_s
               OFFLINE                     svcxd_a2@svc_s
               OFFLINE                     svcxd_a1@svc_s
```

Figure 5-43 Resource group status

After stopping the replication links between the SAN Volume Controller B8_8G4 and B12_4F2 clusters, the consistency group state changed. Example 5-98 shows the `idling_disconnected` state in the svc_global consistency group on the SAN Volume Controller cluster B12_4F2. In this state, the VDisks in this half of the consistency group are all operating in the primary role B12_4F2 and can accept read or write I/O operations.

Example 5-98 SVC PPRC consistency group

```
[svcxd_b1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name
primary master
state idling_disconnected
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Example 5-99 shows the `consistent_disconnected` state for the svc_global consistency group on the SAN Volume Controller cluster B8_8G4. In this state, the VDisks in this half of the consistency group are all operating in the secondary B8_8G4 role and can accept read I/O operations but not write I/O operations.

Example 5-99 SVC PPRC consistency group

```
[svcxd_b1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_disconnected
relationship_count 3
freeze_time 2010/03/23/19/43/16
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

To simulate a site failure in the nodes from the svc_siteb site, run the **reboot -q** command (Example 5-100).

Example 5-100 Simulating failure of a node

```
[svcxd_b1][/]> reboot -q

[svcxd_b2][/]> reboot -q
```

You can see that RG_sitea resource group is in the ONLINE state on the svcxd_a2 node on the svc_sitea site (Figure 5-44).

```
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       OFFLINE                      svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               ONLINE                       svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Figure 5-44 Resource group status

In the SAN Volume Controller cluster, the svc_global consistency group automatically changed the state from consistent_disconnected to idling_disconnected with the cluster in the SAN Volume Controller Cluster B8_8G4 (Example 5-101).

Example 5-101 SVC PPRC consistency group

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary
state idling_disconnected
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8

[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
```

```
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name
primary master
state idling_disconnected
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

After reconnecting all the physical PPRC links, the state of the svc_global consistency group changes to `idling` in both SAN Volume Controller clusters (Example 5-102).

Example 5-102 SVC PPRC consistency group status

```
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary
state idling
relationship_count 3
freeze_time
status
sync out_of_sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8

[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary
state idling
relationship_count 3
freeze_time
```

```
status
sync out_of_sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

Next restart the svc_global consistency group in the correct direction now that the RG_siteb resource group is online on svc_sitea and accessing VDisks from SAN Volume Controller cluster B8_8G4.

For the svc_global consistency group, the SAN Volume Controller cluster B88G4 is `aux` and SAN Volume Controller cluster B12_4F2 is `master`. Restart the copy from SAN Volume Controller cluster B8_8G4 to SAN Volume Controller cluster B12_4F2 (Example 5-103).

Example 5-103 Restarting consistency group svc_metro

```
[svcxd_a1][/]> ssh admin@B12_4F2 svctask startrcconsistgrp -force -primary aux svc_global

[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary aux
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

After you check the `consistent_synchronized` state of the svc_global consistency group and that the cluster services are running in the nodes from the svc_siteb site, by using C-SPOC, move the resource group RG_siteb to the node svcxd_b1 (Figure 5-45).

```
[svcxd_a1][/]> clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                       svcxd_a1@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_b1@svc_s

RG_siteb       ONLINE                       svcxd_b1@svc_s
               OFFLINE                      svcxd_b2@svc_s
               OFFLINE                      svcxd_a2@svc_s
               OFFLINE                      svcxd_a1@svc_s
```

Figure 5-45 Resource group status

After the resource group moves to site svc_siteb, check that the svc_global consistency group changes from primary `aux` to primary `master` (that is, from B12_4F2 to B8_8G4) (Example 5-104).

Example 5-104 SVC PPRC consistency group status

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_global
id 1
name svc_global
master_cluster_id 0000020060A0469E
master_cluster_name B12_4F2
aux_cluster_id 0000020064009B10
aux_cluster_name B8_8G4
primary master
state consistent_synchronized
relationship_count 3
freeze_time
status
sync
copy_type global
RC_rel_id 4
RC_rel_name svc_disk6
RC_rel_id 5
RC_rel_name svc_disk7
RC_rel_id 6
RC_rel_name svc_disk8
```

5.6 Troubleshooting PowerHA Enterprise Edition for SAN Volume Controller

The following topics help you to troubleshoot SVC PPRC clusters.

SVC PPRC states

The following PPRC volume states are possible for either Consistency Groups or PPRC Relationships:

- inconsistent_stopped

  In this state, the primary is accessible for read and write i/o, but the secondary is not accessible for either. A copy process needs to be started to make the secondary consistent.

- inconsistent_copying

  In this state, the primary is accessible for read and write I/O, but the secondary is not accessible for either. This state is entered after a start command is issued to an InconsistendStopped 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 copies data from the primary to the secondary virtual disk.

- consistent_stopped

  In this state, the secondary VDisk contains a consistent image, but it might be out-of-date regarding the primary VDisk.

- consistent_synchronized

  In this state, the primary VDisk is accessible for read and write I/O. Writes that are sent to the primary VDisk are sent to both primary and secondary VDisks. Good completion must be received for both writes, the write must be failed to the host, or a state transition out of ConsistentSychronized must occur before a write is completed to the host.

- idling

  Both master and auxiliary disks are operating in the primary role. Then, both are accessible for write I/O. In this state the relationship or consistency group accepts a **`Start`** command. The remote copy maintains a record of regions on each disk, which received write I/O when idling. This record is used to determine which areas must be copied after a **`Start`** command.

- idling_disconnected

  The virtual disks in this half of the relationship or consistency group are all in the primary 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 transition to a connected state.

- inconsistent_disconnected

  The virtual disks in this half of the relationship or consistency group are all in the secondary 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 secondary role and accept read I/O but not write I/O. This state is entered from ConsistentSychronized or ConsistentStopped when the secondary side of a relationship becomes disconnected.

- empty

  This state applies only to consistency groups. A consistency group has no relationship and 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.

Viewing the state of the relationship and consistency groups

To view the state of all relationship groups that PowerHA manages, by using the configured resource group, run the command that is shown in Example 5-105.

Example 5-105 SVC PPRC relationships

```
ssh admin@<SVC Cluster host or IP address> svcinfo lsrcrelationship

[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcrelationship
id               name             master_cluster_id master_cluster_name master_vdisk_id master_vdisk_name
aux_cluster_id   aux_cluster_name  aux_vdisk_id   aux_vdisk_name    primary        consistency_group_id
consistency_group_name state               bg_copy_priority progress       copy_type
0                svc_disk2         0000020064009B10  B8_8G4              0               svc_haxd0001
0000020060A0469E B12_4F2           0              haxd_svc_v0001    master         0
svc_metro              consistent_synchronized 50                              metro
1                svc_disk3         0000020064009B10  B8_8G4              1               svc_haxd0002
0000020060A0469E B12_4F2           1              haxd_svc_v0002    master         0
svc_metro              consistent_synchronized 50                              metro
2                svc_disk4         0000020064009B10  B8_8G4              2               svc_haxd0003
0000020060A0469E B12_4F2           2              haxd_svc_v0003    master         0
svc_metro              consistent_synchronized 50                              metro
3                svc_disk5         0000020064009B10  B8_8G4              3               svc_haxd0004
0000020060A0469E B12_4F2           3              haxd_svc_v0004    master         0
svc_metro              consistent_synchronized 50                              metro
4                svc_disk6         0000020060A0469E  B12_4F2             4               haxd_svc_v0005
0000020064009B10 B8_8G4            4              svc_haxd0005      master         1
svc_global             consistent_synchronized 50                              global
5                svc_disk7         0000020060A0469E  B12_4F2             5               haxd_svc_v0006
0000020064009B10 B8_8G4            5              svc_haxd0006      master         1
svc_global             consistent_synchronized 50                              global
6                svc_disk8         0000020060A0469E  B12_4F2             6               haxd_svc_v0007
0000020064009B10 B8_8G4            6              svc_haxd0007      master         1
svc_global             consistent_synchronized 50                              global
```

To view the state of specific relationship groups that PowerHA manages, by using the configured resource group, run the command that is shown in Example 5-106.

Example 5-106 SVC PPRC relationship

```
ssh admin@<svc host> svcinfo lsrcrelationship <relationship name>

[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcrelationship svc_disk2
id 0
name svc_disk2
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
master_vdisk_id 0
master_vdisk_name svc_haxd0001
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
aux_vdisk_id 0
aux_vdisk_name haxd_svc_v0001
```

```
primary master
consistency_group_id 0
consistency_group_name svc_metro
state consistent_synchronized
bg_copy_priority 50
progress
freeze_time
status online
sync
```

To view the state of all consistency groups that PowerHA manages, by using the configured resource group, run the command that is shown in Example 5-107.

Example 5-107 SVC PPRC consistency group status

```
ssh admin@<svc cluster> svcinfo lsrcconsistgrp

[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp
id               name              master_cluster_id master_cluster_name aux_cluster_id
aux_cluster_name  primary        state          relationship_count copy_type
0                svc_metro         0000020064009B10  B8_8G4              0000020060A0469E
B12_4F2           master         consistent_synchronized 4                  metro
1             svc_global      0000020060A0469E  B12_4F2           0000020064009B10 B8_8G4
master         consistent_synchronized 3                  global
```

To view the state of specific relationship groups that PowerHA manages, by using the configured resource group, run the command that is shown in Example 5-108.

Example 5-108 SVC PPRC consistency group

```
ssh admin@<svc cluster> svcinfo lsrcconsistgrp <consistency group name>
[svcxd_a1][/]> ssh admin@B8_8G4 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B8_8G4
aux_cluster_id 0000020060A0469E
aux_cluster_name B12_4F2
primary master
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

Viewing SVC PPRC replicated resources configurations

You can use commands to view information about SVC PPRC replicated resources that are related to the SAN Volume Controller cluster and that are managed by PowerHA Enterprise Edition for SAN Volume Controller. These commands are stored in the `/usr/es/sbin/cluster/svcpprc/cmds` directory.

The **cllssvc** command shows the SAN Volume Controller cluster information:

```
cllssvc [-n < svccluster_name >] [-c]
```

The command lists information about all SAN Volume Controller clusters in the PowerHA configuration or a specific SAN Volume Controller cluster. If no SAN Volume Controller is specified, all SAN Volume Controller clusters that are defined must be listed. If a specific SAN Volume Controller cluster is provided by using the **-n** flag, information about this SAN Volume Controller is displayed. The **-c** flag displays information in a colon-delimited format. Example 5-109 shows the results of running the **cllssvc** command.

Example 5-109 Results of the cllssvc command

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvc
B8_8G4
B12_4F2

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvc -a
#SVCNAME ROLE SITENAME IPADDR IPADDR2 RPARTNER
B8_8G4 Master svc_sitea 10.12.5.55  B12_4F2
B12_4F2 Master svc_siteb 10.114.63.250  B8_8G4

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvc -n B8_8G4
#SVCNAME ROLE SITENAME IPADDR IPADDR2 RPARTNER
B8_8G4 Master svc_sitea 10.12.5.55  B12_4F2
```

The **cllsrelationship** command shows information about all SVC PPRC relationships or a specific PPRC relationship.

```
cllsrelationship [-n <relationship_name>] [-c] [-a] [-h]
```

If no resource name is specified, the names of all PPRC resources that are defined are listed. If the **-a** flag is provided, full information about all PPRC relationships is displayed. If a specific relationship is provided by using the **-n** flag, only information about this relationship is displayed. The **-c** flag displays information in a colon-delimited format. The **-h** flag turns off the display of column headers. Example 5-110 shows the results of running the **cllsrelationship** command.

Example 5-110 Results of the cllsrelationship command

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllsrelationship
svc_disk2
svc_disk3
svc_disk4
svc_disk5
svc_disk6
svc_disk7
svc_disk8

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllsrelationship -a
relationship_name MasterVdisk_info AuxiliaryVdisk_info
svc_disk2       svc_haxd0001@B8_8G4 haxd_svc_v0001@B12_4F2
```

```
svc_disk3       svc_haxd0002@B8_8G4 haxd_svc_v0002@B12_4F2
svc_disk4       svc_haxd0003@B8_8G4 haxd_svc_v0003@B12_4F2
svc_disk5       svc_haxd0004@B8_8G4 haxd_svc_v0004@B12_4F2
svc_disk6       haxd_svc_v0005@B12_4F2 svc_haxd0005@B8_8G4
svc_disk7       haxd_svc_v0006@B12_4F2 svc_haxd0006@B8_8G4
svc_disk8       haxd_svc_v0007@B12_4F2 svc_haxd0007@B8_8G4

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllsrelationship -n svc_disk2
relationship_name MasterVdisk_info AuxiliaryVdisk_info
svc_disk2       svc_haxd0001@B8_8G4 haxd_svc_v0001@B12_4F2
```

The **cllssvcpprc** shows information about all SVC PPRC resources or a specific SVC PPRC resource.

```
cllssvcpprc [-n < svcpprc_consistencygrp >] [-c] [-a] [-h]
```

If no resource name is specified, the names of all PPRC resources that are defined are listed. If the -a flag is provided, full information about all PPRC resources is displayed. If a specific resource is provided by using the **-n** flag, only information about this resource is displayed. The **-c** flag displays information in a colon-delimited format. The **-h** flag turns off the display of column headers. Example 5-111 shows the results of running the **cllssvcpprc** command.

Example 5-111 Results of the cllssvcpprc command

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> ./cllssvcpprc
svc_metro
svc_global

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> cllssvcpprc -a
svcpprc_consistencygrp MasterCluster  AuxiliaryCluster relationships  CopyType
RecoveryAction
svc_metro       B8_8G4          B12_4F2         svc_disk2 svc_disk3 svc_disk4
svc_disk5 METRO           MANUAL
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8
GLOBAL          AUTO

[svcxd_a1][/usr/es/sbin/cluster/svcpprc/cmds]> cllssvcpprc -n svc_global
svcpprc_consistencygrp MasterCluster  AuxiliaryCluster relationships  CopyType
RecoveryAction
svc_global      B12_4F2         B8_8G4          svc_disk6 svc_disk7 svc_disk8
GLOBAL          AUTO
```

The **cl_verify_svcpprc_config** command verifies the SAN Volume Controller definition in the PowerHA configuration. This command is stored in the /usr/es/sbin/cluster/svcpprc/utils directory (Example 5-112).

Example 5-112 Results of the cl_verify_svcpprc-config command

```
[svcxd_a1][/usr/es/sbin/cluster/svcpprc/utils]> ./cl_verify_svcpprc_config
Verifying HACMP-SVCPPRC configuration...
cl_verify_svcpprc_config: Checking available nodes
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_a1
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_a2
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_b1
cl_verify_svcpprc_config: Retrieving disk information from node svcxd_b2
cl_verify_svcpprc_config: Checking available SVCs
cl_verify_svcpprc_config: Checking license, release level and disk map for SVC B8_8G4 at 10.12.5.55
```

```
cl_verify_svcpprc_config: Checking license, release level and disk map for SVC B12_4F2 at 10.114.63.250
cl_verify_svcpprc_config: Checking consistency groups
cl_verify_svcpprc_config: Checking consistency group svc_metro
cl_verify_svcpprc_config: Checking consistency group svc_global
cl_verify_svcpprc_config: Checking resource groups.
cl_verify_svcpprc_config: Checking resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a1 for resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a2 for resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b2 for resource group RG_sitea
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b1 for resource group RG_sitea
cl_verify_svcpprc_config: Checking volume group siteametrovg in group RG_sitea on site svc_sitea
cl_verify_svcpprc_config: Checking volume group siteametrovg in group RG_sitea on site svc_siteb
cl_verify_svcpprc_config: Checking resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b1 for resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_b2 for resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a2 for resource group RG_siteb
cl_verify_svcpprc_config: Checking SVC virtual disks on node svcxd_a1 for resource group RG_siteb
cl_verify_svcpprc_config: Checking volume group sitebglobalvg in group RG_siteb on site svc_sitea
cl_verify_svcpprc_config: Checking volume group sitebglobalvg in group RG_siteb on site svc_siteb
cl_verify_svcpprc_config: Verifying consistency groups against the SVC configuration
cl_verify_svcpprc_config: Establishing consistency group svc_metro
cl_verify_svcpprc_config: WARNING: Consistency Group svc_metro already exists
cl_verify_svcpprc_config: Verifying relationships for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk2 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk2 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk3 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk3 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk4 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk4 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Verifying relationship svc_disk5 in consistency group svc_metro
cl_verify_svcpprc_config: Relationship svc_disk5 already exists for consistency group svc_metro
cl_verify_svcpprc_config: Establishing consistency group svc_global
cl_verify_svcpprc_config: WARNING: Consistency Group svc_global already exists
cl_verify_svcpprc_config: Verifying relationships for consistency group svc_global
cl_verify_svcpprc_config: Verifying relationship svc_disk6 in consistency group svc_global
cl_verify_svcpprc_config: Relationship svc_disk6 already exists for consistency group svc_global
cl_verify_svcpprc_config: Verifying relationship svc_disk7 in consistency group svc_global
cl_verify_svcpprc_config: Relationship svc_disk7 already exists for consistency group svc_global
cl_verify_svcpprc_config: Verifying relationship svc_disk8 in consistency group svc_global
cl_verify_svcpprc_config: Relationship svc_disk8 already exists for consistency group svc_global
HACMP-SVCPPRC configuration verified successfully.  Status=0
```

After the successful verification of the SAN Volume Controller configurations, it establishes all the SAN Volume Controller relationships that re defined to PowerHA on the SAN Volume Controller clusters and adds them to the corresponding consistency groups.

6

Configuring PowerHA SystemMirror Enterprise Edition with ESS/DS Metro Mirror

This chapter explains how to install, configure, and use the IBM PowerHA SystemMirror Enterprise Edition Metro Mirroring with disk system command-line interface (DSCLI) management. This procedure is accomplished by using Peer-to-Peer Remote Copy (PPRC) to maintain copies of data between sites on a set of paired disks. The PowerHA resource group contains information about the volume pairs and the paths to communicate between sites. PowerHA uses **`dscli`** commands to associate the PPRC-mirrored volumes to the active site. Metro Mirroring supports only synchronous writes between the sites.

This chapter provides examples of implementing the setup, testing various failover scenarios, and the addition and removal of disks from the cluster.

The chapter includes the following sections:

- Planning
- Software requirements
- Considerations and restrictions
- Environment example
- Installing and configuring Metro Mirroring
- Test scenarios
- Adding and removing LUNs
- Commands for troubleshooting or gathering information
- PowerHA Enterprise Edition: SPPRC DSCLI security enhancements

6.1 Planning

Before you configure the DSCLI metro mirroring environment, you must decide the following items:

- Nodes and sites to used
- Networks, including XD_ip networks between sites and local networks
- The Copy Services Servers (CSS) to use from both sites
- The disks to use
- The vpaths to use, including volume IDS
- The PPRC replicated resources
- The port pairs for the PPRC paths
- The volume pairs
- The volume groups that are managed by the PPRC replicated resources and which resource groups they will be a part of

Reference: For more planning information, see the *HACMP for AIX 6.1 Planning and Administration Guide,* SC23-4863.

6.2 Software requirements

For DSCLI PPRC, the following PowerHA file sets are required:

- The base PowerHA file sets
- `cluster.es.pprc.cmds`
- `cluster.es.pprc.rte`
- `cluster.es.spprc.cmds`
- `cluster.es.spprc.rte`
- `cluster.msg.en_US.pprc`
- Other language message sets if required

XD release notes: For complete information about the software requirements, see the XD release notes in `/usr/es/sbin/cluster/release_notes_xd`.

The following file sets are required:

- AIX 5.3 TL09 RSCT V2.4.12
- AIX 6.1 TL02 RSCT V2.5.5
- **`dscli`** software

  See the *Command-Line Interface User's Guide for the DS6000 series and DS8000 series.*
- Java 1.4.1

  This PowerHA APAR IZ74478 removes the restriction of this Java 1.4.1 version requirement.

PowerHA 6.1 requirements: These requirements are for PowerHA 6.1 only because this IBM Redbooks publication focuses on PowerHA SystemMirror Enterprise Edition 6.1.

6.3 Considerations and restrictions

For information about IBM System Storage models and PowerHA support, go to the IBM Disk Storage Systems website at:

http://www.ibm.com/servers/storage/disk/index.html

Volume group considerations

You can find a complete list of considerations in the *HACMP for AIX 6.1 Metro Mirror: Planning and Administration Guide*, SC23-4863. The following considerations are of importance:

- A volume group must have the same volume major number across all cluster nodes.
- Resource groups to be managed by PowerHA cannot contain volume groups with both PPRC-protected and non-PPRC-protected disks. See the following examples:
  - 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.
- Only nonconcurrent volume groups can be used with PPRC.
- Resource groups cannot manage both DSCLI and Direct management (ESS CLI)-managed PPRC resources simultaneously.
- Only the Synchronous PPRC (Metro Mirror) function for IBM TotalStorage Copy Services is supported (no Global Copy or Global Mirror).
- C-SPOC operations on nodes at the same site as the source volumes successfully perform all tasks supported in HACMP.
- C-SPOC operations do not succeed on nodes at the remote site (that contain the target volumes) for the following LVM operations:
  - Creating or extending a volume group
  - Operations that require nodes at the target site to write to the target volumes, such as changing file system size, changing mount point, and adding LVM mirrors.

    They cause an error message in C-SPOC. However, nodes on the same site as the source volumes can successfully perform these tasks. The changes are then propagated to the other site by using a lazy update.
- For C-SPOC operations to work on all other LVM operations, perform all C-SPOC operations when the cluster is active on all PowerHA nodes.

6.4 Environment example

Figure 6-1 and Figure 6-2 on page 241 show implementation examples of our testing environment.

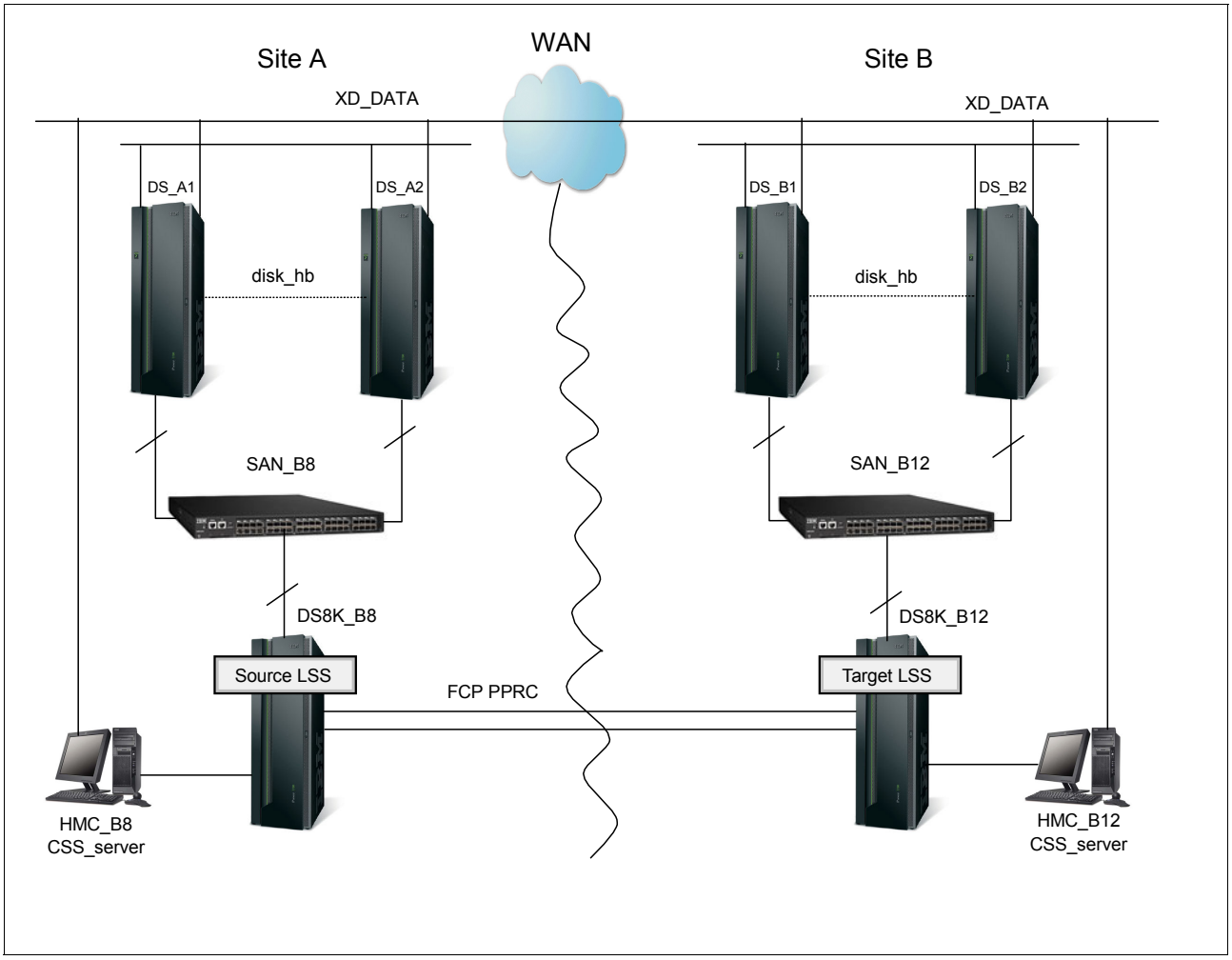


Figure 6-1 Nodes, disks, sites, and networks

CSSs per site: The environment in Figure 6-1 shows one CSS (Storage HMC) per site. However, you can also configure two CSSs per site for redundancy.

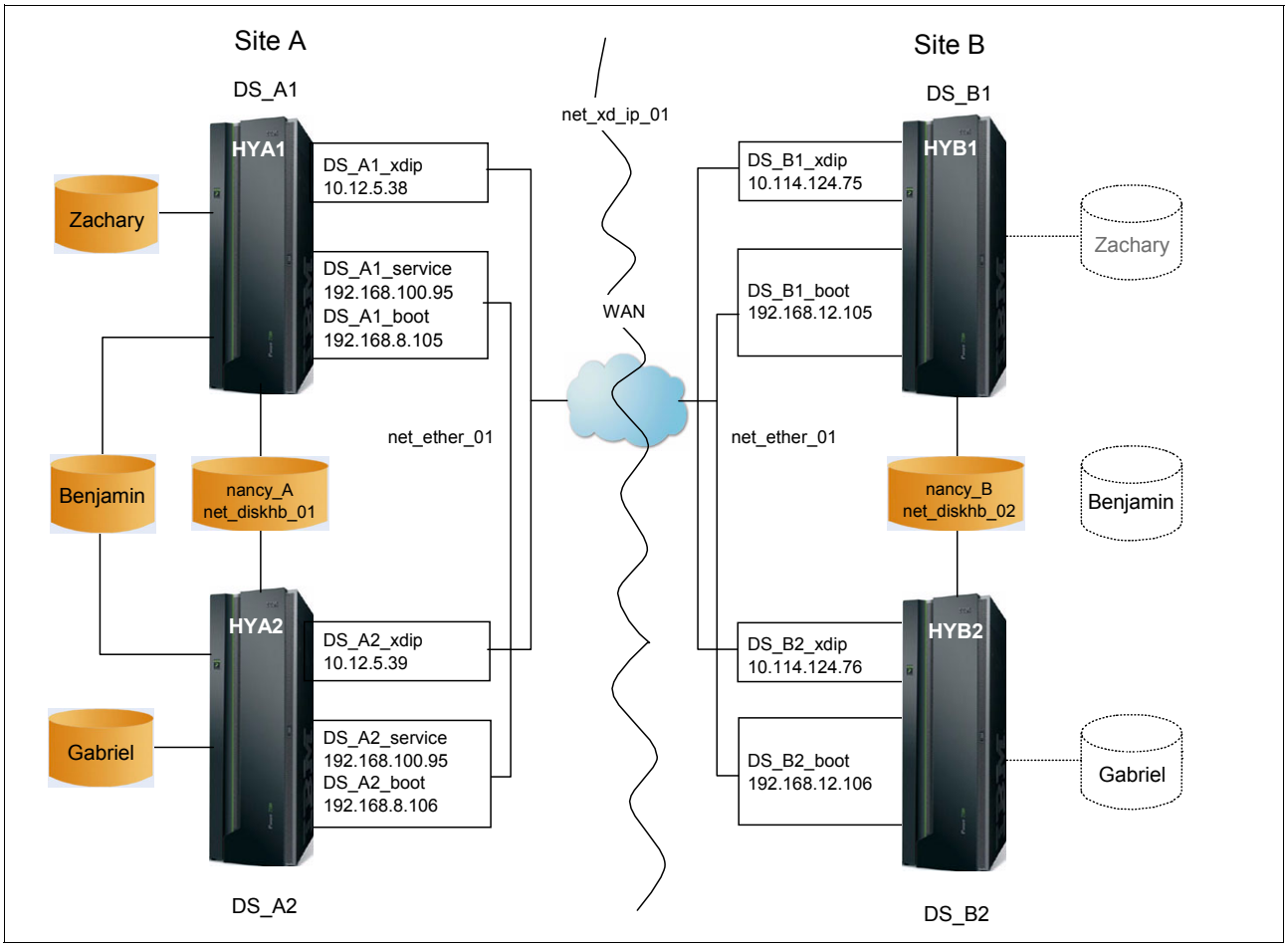


Figure 6-2 Detailed volume groups, disks, and network relationships

The following configuration information is used when testing our scenarios for this chapter.

6.4.1 Volume information

The volume configuration has the following details:

- Asite
  - HMC/CSS server: HMC_DS8K_B8 10.12.6.17
  - DS Server ID: IBM.2107-75BALB1
- Bsite
  - HMC/CSS server: HMC_DS8K_B12 10.114.232.150
  - DS Server ID: IBM.2107-7585461
- Asite Bsite disk relationship:
  - Pri-Sec Port Pair IDs I0002->I0102
  - LSS Pair: **80 20** PPRC Volume Pairs:

    8001 ->2001
    8002 ->2002
    8003 ->2003
    8004 ->2004

– Pri-Sec Port Pair IDs: **I0133->I0132**

– LSS Pair: **81 30** PPRC Volume Pairs:

8101 ->3001
8102 ->3002
8103 ->3003
8104 ->3004

6.4.2 Topology (cltopinfo command)

The topology configuration has the following details:

- NODE DS_A1:
  - Network net_XD_ip_01

    DS_A1_xdip: 10.12.5.38
  - Network net_diskhb_01

    hbdiska1: /dev/hdisk8
  - Network net_diskhb_02
  - Network net_ether_01
    - DS_A1_service: 192.168.100.55
    - DS_B2_service: 10.10.12.116
    - DS_B1_service: 10.10.12.115
    - DS_A2_service: 192.168.100.95
    - DS_A1_boot: 192.168.8.105
- NODE DS_A2:
  - Network net_XD_ip_01

    DS_A2_xdip: 10.12.5.39
  - Network net_diskhb_01

    hbdiska2: /dev/hdisk8
  - Network net_diskhb_02
  - Network net_ether_01
    - DS_A1_service: 192.168.100.55
    - DS_B2_service: 10.10.12.116
    - DS_B1_service: 10.10.12.115
    - DS_A2_service: 192.168.100.95
    - DS_A2_boot: 192.168.8.106
- NODE DS_B1:
  - Network net_XD_ip_01

    DS_B1_xdip: 10.114.124.75
  - Network net_diskhb_01
  - Network net_diskhb_02

    hbdiskb1: /dev/hdisk9
  - Network net_ether_01
    - DS_A1_service: 192.168.100.55
    - DS_B2_service: 10.10.12.116
    - DS_B1_service: 10.10.12.115

- DS_A2_service: 192.168.100.95
- DS_B1_boot: 192.168.12.105

- NODE DS_B2:
  - Network net_XD_ip_01

    DS_B2_xdip: 10.114.124.76
  - Network net_diskhb_01
  - Network net_diskhb_02

    hbdiskb2: /dev/hdisk9
  - Network net_ether_01
    - DS_A1_service: 192.168.100.55
    - DS_B2_service: 10.10.12.116
    - DS_B1_service: 10.10.12.115
    - DS_A2_service: 192.168.100.95
    - DS_B2_boot: 192.168.12.106

6.4.3 Volume groups

The configuration has the following volume groups:

- zachary
- gabriel
- benjamin

6.4.4 Resource groups

The resource group has the following configuration:

- Resource group Zachary
  - Startup policy: Online on home node only
  - Failover policy: Failover to next priority node in the list
  - Failback policy: Never failback
  - Participating nodes: DS_A1 DS_A2 DS_B1 DS_B2
  - Service IP label: DS_A1_service
  - Service IP label: DS_B1_service
- Resource group Gabriel
  - Startup policy: Online on home node only
  - Failover policy: Failover to next priority node in the list
  - Failback policy: Never failback
  - Participating nodes: DS_A1 DS_A2 DS_B1 DS_B2
- Resource group Benjamin
  - Startup policy: Online on home node only
  - Failover policy: Failover to next priority node in the list
  - Failback policy: Never failback
  - Participating nodes: DS_A1 DS_A2 DS_B1 DS_B2
  - Service IP label: DS_A2_service
  - Service IP label: DS_B2_service

6.5 Installing and configuring Metro Mirroring

To learn how to install and configure PPRC Metro Mirror by using the **dscli**, see the *HACMP for AIX 6.1 Metro Mirror: Planning and Administration Guide,* SC23-4863. This section provides further examples.

6.5.1 Installing the software

Install the required software that is listed in 6.2, “Software requirements” on page 238. After installing the DSCLI, the **dscli** commands are in the `/opt/ibm/dscli/dscli` directory.

6.5.2 Setting up the disks and volume groups

Disks: In this scenario, the disks that are used are assigned to the node. You can see the disks by using the **lspv** command.

To set up the disks and volume groups:

1. Determine the disks and their associated volume ID:

   ```
   pcmpath query essmap
   ```

   Look at the last four digits of the LUN SN to get the ID and the first two digits of those last four digits to get the LSS. For example, 75BALB18001 is ID 8001, and the LSS is 80 (Figure 6-3).

```
Site A
hdisk2      path0     21-T1-01[FC] fscsi0   75BALB1 8001  IBM 2107-900
hdisk3      path0     21-T1-01[FC] fscsi0   75BALB1 8002  IBM 2107-900
hdisk4      path0     21-T1-01[FC] fscsi0   75BALB1 8003 IBM 2107-900
hdisk9      path3     31-T1-01[FC] fscsi1   75BALB1 8104  IBM 2107-900

Site B
hdisk2      path3     03-08-02[FC] fscsi0   7585461 2001  IBM 2107-900
hdisk3      path0     03-08-02[FC] fscsi0   7585461 2002  IBM 2107-900
hdisk4      path3     03-08-02[FC] fscsi0   7585461 2003  IBM 2107-900
hdisk9      path3     03-08-02[FC] fscsi0   7585461 3004  IBM 2107-900
```

Figure 6-3 Results of the pcmpath query essmap command

Alternatively, you can also use the **lscfg -vl <hdisk>** command to see this information (Figure 6-4).

```
Site A
> lscfg -vl hdisk3 |grep Serial
Serial Number...............75BALB18002

Site B
> lscfg -vl hdisk3 | grep Serial
Serial Number...............75854612002
```

Figure 6-4 Results of the lscfg command to see the hdisk serial number

2. Modify the default `dscli.profile` file on both sites for simplification of the **`dscli`** commands.

> **Attention:** If the default `dscli.profile` file has the user name and password uncommented, PPRC verification fails with a message similar to the following example:
>
> `spprc_verify_config[1621] dspmsg -s 7 pprc.cat 999 spprc_verify_config: ERROR 10.12.6.17~~IBM.2107-75BALB1 does not match the HMC storage id for PPRC replicated resource.`
>
> Rename the default `/opt/ibm/dscli/profile/dscli.profile` file if you entered a user name and a password after you run the **`rmpprc`** command.

   a. Edit the `/opt/ibm/dscli/profile/dscli.profile` file or create your own. If you create your own profile, when you run the **`dscli`** command, use the **`-cfg <profilename>`** option.

   b. Uncomment and change the hmc1, hmc2, username, password, devid, and remotedevid (Example 6-1).

Example 6-1 Changing the dscli profile fields

```
hmc1:   10.12.6.17
username: dinoadm
password: ds8kitso
devid:         IBM.2107-75BALB1
remotedevid:   IBM.2107-7585461
```

3. Get the worldwide node name on the remote site:

   `dscli lssi`

   Figure 6-5 shows the results of running this command.

```
root@DS_575_1_B1 /opt/ibm/dscli/profile > dscli -cfg default.dscli.profile
dscli> lssi
Name  ID               Storage Unit     Model WWNN             State  ESSNet
==============================================================================
ess11 IBM.2107-7585461 IBM.2107-7585460 922   5005076303FFC4D2 Online Enabled
dscli>
```

Figure 6-5 Results of the dscli lssi command

4. Get the local and attached ports associated with the WWNN. Use the **`dscli lsavailpprcport`** command on the local site. Use the remote WWNN and the targetLSS and sourceLSS (Figure 6-6).

```
root@DS_550_1_A1 /opt/ibm/dscli/profile > dscli -cfg dscli.profile.admin
dscli> lsavailpprcport -remotewwnn 5005076303FFC4D2 80:20
Local Port Attached Port Type
=============================
I0002      I0102         FCP
I0002      I0132         FCP
I0133      I0102         FCP
I0133      I0132         FCP
```

Figure 6-6 dscli lsavailpprcport by using dscli.profile.admin

5. Create a temporary PPRC relationship between the sites to mirror the data to the remote secondary site.

   a. Use the **mkpprcpath** command to create the path (Figure 6-7).

```
/opt/ibm/dscli/dscli mkpprcpath -srclss 80 -tgtlss 20 -remotewwnn
5005076303FFC4D2 I0002:I0102
/opt/ibm/dscli/dscli mkpprcpath -srclss 81 -tgtlss 30 -remotewwnn
5005076303FFC4D2 I0133:I0132
```

Figure 6-7 Results of the dscli command mkpprcpath by using dscli.profile.

   b. Use the **mkpprc** command to map the PPRC. In our example in Figure 6-8, we have eight disks.

```
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8001:2001
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8002:2002
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8003:2003
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8004:2004
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8101:3001
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8102:3002
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8103:3003
/opt/ibm/dscli/dscli mkpprc -type mmir -mode full 8104:3004
```

Figure 6-8 Results of the dscli command mkpprc by using dscli.profile

   c. Use the **lspprc** command to monitor the replication process and status. The Full Duplex state indicates that it is replicated. Copy Pending means that it is still mirroring data. The **dscli lspprc** command with the **-l** option shows the tracks that are not in sync (Figure 6-9).

```
/opt/ibm/dscli/dscli lspprc -l 8001-8004:2001-2004 8101-8104:3001-3004

ID        State        Reason Type OutOf Sync Tracks
8001:2001 Full Duplex  - Metro Mirror 0
8002:2002 Copy Pending - Metro Mirror 2145
8003:2003 Full Duplex  - Metro Mirror 0
8004:2004 Full Duplex  - Metro Mirror 0
8101:3001 Copy Pending - Metro Mirror 376177
8102:3002 Copy Pending - Metro Mirror 564342
8103:3003 Copy Pending - Metro Mirror 490387
8104:3004 Copy Pending - Metro Mirror 759424
```

Figure 6-9 Results of the dscli command lspprc by using dscli.profile

The dscli lspprc command: The **dscli lspprc** command in Figure 6-9 shows Full Duplex when the PPRC instance is accessible for read and write operations.

6. Create on the primary site the volume groups, logical volumes, and file systems to be mirrored on the PPRC disks. All volume information is copied between the PPRC disk pairs, including the pvid. The disks must have the same volume group name on both sites. Run the following command on the other node at the primary site:

```
importvg -y <volumegroupname> <hdisk>
```

7. Remove the PPRC relationship after replication completes according to the **lspprc** command shown in Figure 6-10. If the `dscli.profile` file was modified with a user name and password, comment them or rename the profile. Type `y` when prompted to remove the remote copy and volume pair relationship.

```
/opt/ibm/dscli/dscli rmpprc 8001:2001
/opt/ibm/dscli/dscli rmpprc 8002:2002
/opt/ibm/dscli/dscli rmpprc 8003:2003
/opt/ibm/dscli/dscli rmpprc 8004:2004
/opt/ibm/dscli/dscli rmpprc 8101:3001
/opt/ibm/dscli/dscli rmpprc 8102:3002
/opt/ibm/dscli/dscli rmpprc 8103:3003
/opt/ibm/dscli/dscli rmpprc 8104:3004
```

Figure 6-10 Results of the dscli command rmpprc with dscli.profile to remove the pprc relationship

8. On the remote site, import the volume groups on both nodes. Remember to use the same volume group name:

```
importvg -y <volumegroupname> <hdisk>
```

6.5.3 Configuring the PPRC replicated resources

The PPRC is created in PowerHA to describe the relationship between the sites. PowerHA uses this information to set up the source and target relationship for mirroring the data. Start configuring the PPRC resources:

1. Run the **smitty hacmp** command and select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resources Configuration** → **PPRC-Managed Replicated Resources Configuration**.
2. Select **DSCLI-Managed PPRC Replicated Resource Configuration** (Figure 6-11).

```
 DSCLI-Managed PPRC Replicated Resource Configuration

Move cursor to desired item and press Enter.

  Copy Services Server Configuration
  DSS Disk Subsystem Configuration
  DSCLI-Managed PPRC Replicated Resource Configuration
  PPRC Consistency Groups Configuration
  Verify PPRC Configuration
  SNMP Trap Configuration
```

Figure 6-11 PPRC replicated resource configuration menu

3. In SMIT, select **Copy Services Server Configuration → Add a Copy Services Server** (Figure 6-12 and Figure 6-13 on page 248).

Hint: We configured two Copy Services Servers (CSSs) per site for redundancy.

```
                        Add a Copy Services Server

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

                                                      [Entry Fields]
* CSS Subsystem Name                                 [HMC_DS8K_B8]
* CSS Site Name                                       Asite +
* CSS IP Address                                     [10.12.6.17]
* CSS User ID                                        [dinoadm]
* CSS Password                                       [ds8kitso]
```

Figure 6-12 Add a Copy Services Server panel for Asite

```
                        Add a Copy Services Server

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

                                                      [Entry Fields]
* CSS Subsystem Name                                 [HMC_DS8K_B12]
* CSS Site Name                                       Bsite +
* CSS IP Address                                     [10.114.232.150]
* CSS User ID                                        [itso_user]
* CSS Password                                       [ds8kitso]
```

Figure 6-13 Add a Copy Services Server panel for Bsite

4. In SMIT, select **DSS Disk Subsystem Configuration → Add an ESS Disk Subsystem** (Figure 6-14 and Figure 6-15 on page 249).

```
                        Add an ESS Disk Subsystem

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

                                                      [Entry Fields]
* ESS Subsystem Name                                 [HMC_DS8K_B8]
* ESS Site Name                                       Asite +
  ESS Cluster1 IP Address                            [10.12.6.17]
  ESS Cluster2 IP Address                            []
* ESS User ID                                        [dinoadm]
* ESS Password                                       [ds8kitso]
* Full ESS Storage ID                                [IBM.2107-75BALB1] +
* List of CS Servers                                  HMC_DS8K_B8 +
```

Figure 6-14 Add an ESS Disk Subsystem panel for Asite

IP address: IP address 10.12.6.17 is the IP address of the CSS that was configured in Figure 6-12 on page 248.

```
                          Add an ESS Disk Subsystem

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

                                                      [Entry Fields]
* ESS Subsystem Name                                 [HMC_DS8K_B12]
* ESS Site Name                                       Bsite +
  ESS Cluster1 IP Address                            [10.114.232.150]
  ESS Cluster2 IP Address                            []
* ESS User ID                                        [itso_user]
* ESS Password                                       [ds8kitso]
* Full ESS Storage ID                                [IBM.2107-7585461] +
* List of CS Servers                                  HMC_DS8K_B12 +
```

Figure 6-15 Add an ESS Disk Subsystem panel for Bsite

IP address: IP address 10.114.232.150 is the IP address of the CSS that was configured in Figure 6-13 on page 248.

5. In SMIT, select **DSCLI-Managed PPRC Replicated Resource Configuration** → **Add a PPRC Resource** (Figure 6-16, Figure 6-17 on page 250, and Figure 6-18 on page 250).

```
                              Add a PPRC Resource

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

                                                      [Entry Fields]
* PPRC Resource Name                                 [Zachary_pprc]
* HACMP Sites                                        [Asite Bsite] +
* PPRC Volume Pairs                                  [8001->2001]
* ESS Pair [HMC_DS8K_B8 HMC_DS8K_B12]        +
* LSS Pair                                           [80 20] +
* PPRC Type                                           mmir +
* Pri-Sec Port Pair IDs                              [I0002->I0102]
* Sec-Pri Port Pair IDs                              [I0102->I0002]
* PPRC Link Type                                       FCP +
* PPRC Critical Mode                                   OFF +
* PPRC Recovery Action                                 MANUAL +
* Volume Group                                       [zachary]
```

Figure 6-16 Add a PPRC Resource panel for Zachary_pprc

```
                              Add a PPRC Resource

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

                                                      [Entry Fields]
* PPRC Resource Name                                 [Gabriel_pprc]
* HACMP Sites                                        [Asite Bsite] +
* PPRC Volume Pairs                                  [8003->2003 8004->2004]
* ESS Pair [HMC_DS8K_B8 HMC_DS8K_B12]+
* LSS Pair                                           [80 20] +
* PPRC Type                                           mmir +
* Pri-Sec Port Pair IDs                              [I0002->I0102]
* Sec-Pri Port Pair IDs                              [I0102->I0002]
* PPRC Link Type                                      FCP +
* PPRC Critical Mode                                  OFF +
* PPRC Recovery Action                                AUTOMATED +
* Volume Group                                       [gabriel]
```

Figure 6-17 Add a PPRC Resource panel for Gabriel_pprc

```
                              Add a PPRC Resource

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

                                                      [Entry Fields]
* PPRC Resource Name                                 [Benjamin_pprc]
* HACMP Sites                                        [Asite Bsite] +
* PPRC Volume Pairs                                  [8101->3001 8102->3002]
* ESS Pair [HMC_DS8K_B8 HMC_DS8K_B12]+
* LSS Pair                                           [81 30] +
* PPRC Type                                           mmir +
* Pri-Sec Port Pair IDs                              [I0133->I0132]
* Sec-Pri Port Pair IDs                              [I0132->I0133]
* PPRC Link Type                                      FCP +
* PPRC Critical Mode                                  OFF +
* PPRC Recovery Action                                AUTOMATED +
* Volume Group                                       [benjamin]
```

Figure 6-18 Add a PPRC Resource panel for Benjamin_pprc

6.5.4 Verifying the DSCLI managed configuration

Use SMIT to verify the dscli configuration. Run the **smitty hacmp** command. Then, select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → PPRC-Managed Replicated Resources Configuration → DSCLI-Managed PPRC Replicated Resource Configuration → Verify PPRC Configuration**.

6.5.5 Configuring the resource groups

This example uses three resource groups. Each group contains one PPRC resource. Use SMIT to add the resource groups. Run the **smitty hacmp** command. Then, select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resource Group Configuration → Add a Resource Group**.

- The Zachary resource contains the PPRC resource Zachary_pprc. The volume group zachary is to start with the primary site Asite on node DS_A1 (Figure 6-19).

```
 Resource Group Name                                 Zachary
Inter-site Management Policy                        Online On Either Site +
* Participating Nodes from Primary Site              [DS_A1 DS_A2] +
  Participating Nodes from Secondary Site            [DS_B1 DS_B2] +

  Startup Policy Online On Home NodeOnly         +
  Failover Policy Failover To Next Priority Node > +
  Failback Policy                                     Never Failback +
```

Figure 6-19 Configuring resource group Zachary

- The Gabriel resource contains the PPRC resource Gabriel_pprc. The volume group gabriel is to start with the primary site Asite on node DS_A2 (Figure 6-20).

```
 Resource Group Name                                 Gabriel
Inter-site Management Policy                        Online On Either Site +
* Participating Nodes from Primary Site              [DS_A2 DS_A1] +
  Participating Nodes from Secondary Site            [DS_B2 DS_B1] +

  Startup Policy Online On Home Node Only         +
  Failover Policy Failover To Next Priority Node > +
  Failback Policy                                     Never Failback +
```

Figure 6-20 Configure resource group Gabriel

- The Benjamin resource contains the PPRC resource Benjamin_pprc. The volume group benjamin is to start with the primary site Bsite on node DS_B2 (Figure 6-21).

```
 Resource Group Name                                 Benjamin
  New Resource Group Name                            []
  Inter-site Management Policy                        Prefer Primary Site +
* Participating Nodes from Primary Site              [DS_B2 DS_B1] +
  Participating Nodes from Secondary Site            [DS_A2 DS_A1] +

  Startup Policy Online On Home Node Only         +
  Failover Policy Failover To Next Priority Node > +
  Failback Policy                                     Never Failback +
```

Figure 6-21 Configuring resource group Benjamin

6.5.6 Configuring the resources and attributes for the resource group

Run the **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resource Group Configuration** → **Change/Show Resources and Attributes for a Resource Group** to update the resource groups with the extra required information such as volume groups, service IP addresses, and pprc resources (Figure 6-22, Figure 6-23, and Figure 6-24).

```
Service IP Labels/Addresses [DS_A1_service DS_B1_service]     +
Volume Groups                                          [zachary] +
PPRC Replicated Resources                              [Zachary_pprc] +
```

Figure 6-22 Changes for the Zachary resource group

```
Volume Groups                                          [gabriel ] +
PPRC Replicated Resources                              [Gabriel_pprc] +
```

Figure 6-23 Changes for the Gabriel resource group

```
Service IP Labels/Addresses [DS_A2_service DS_B2_service]     +
Volume Groups                                          [benjamin ] +
PPRC Replicated Resources                              [Benjamin_pprc] +
```

Figure 6-24 Changes for the Benjamin resource group

6.5.7 Synchronizing the cluster

In SMIT, select **Extended Configuration** → **Extended Verification and Synchronization**. Select **Both** for Verify, Synchronize, or Both.

6.6 Test scenarios

This section demonstrates the state of the system based on various test scenarios. The following examples are shown:

- Moving a resource group from one site to another by using the C-SPOC commands
- Halting both nodes on one site
- The loss of local disk storage on one site
- The loss of the PPRC connection between sites
- The loss of the XD_ip network
- Total site failure, loss of PPRC, and XD_ip networks

6.6.1 Moving a resource group to another site

Moving a RESOURCE GROUP is often desired when the site needs to be brought down for maintenance. The C-SPOC SMIT menus are the easiest and most common way to move the RESOURCE GROUP between sites. In this example, we move the resource zachary from Asite to Bsite.

1. Use the **/usr/sbin/cluster/utilities/clRGinfo** command to ensure that the resource that you move is shown as ONLINE SECONDARY on a node on the other site (Example 6-2).

Example 6-2 Results of the /usr/sbin/cluster/utilities/clRGinfo command

```
/ > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zachary        ONLINE                       DS_A1@Asite
               OFFLINE                      DS_A2@Asite
               OFFLINE                      DS_B1@Bsite
               ONLINE SECONDARY             DS_B2@Bsite

Gabriel        ONLINE                       DS_A2@Asite
               OFFLINE                      DS_A1@Asite
               ONLINE SECONDARY             DS_B2@Bsite
               OFFLINE                      DS_B1@Bsite

Benjamin       ONLINE                       DS_B2@Bsite
               OFFLINE                      DS_B1@Bsite
               ONLINE SECONDARY             DS_A2@Asite
               OFFLINE                      DS_A1@Asite
```

2. Use SMIT CSPOC to move the resource. Enter **smitty cspoc**. Then, select **Resource Groups and Applications → Move a Resource Group to Another Node / Site → Move Resource Groups to Another Site**. Select the ONLINE resource that you want to move. In our example, Zachary is seen ONLINE on DS_A1 at the Asite and the ONLINE_SECONDARY is DS_B2 at the Bsite (Example 6-3).

Example 6-3 Moving a resource group to another site

```
Select Resource Group(s)
| Move cursor to desired item and press Enter.                              |
|                                                                           |
|   #                                                                       |
|   # Resource Group                State              Node(s) / Site       |
|   #                                                                       |
|   Zachary                         ONLINE             DS_A1 / Asite        |
|   Zachary                         ONLINE SECONDARY   DS_B2 / Bsite        |
|   Gabriel                         ONLINE             DS_A2 / Asite        |
|   Gabriel                         ONLINE SECONDARY   DS_B2 / Bsite        |
|   Benjamin                        ONLINE             DS_B2 / Bsite        |
|   Benjamin                        ONLINE SECONDARY   DS_A2 / Asite        |
```

While the move is running, the **clRGinfo** command shows an intermediate state of ACQUIRING (Example 6-4).

Example 6-4 Results of the clRGinfo command, intermediate state when moving a resource group

```
Zachary        ONLINE SECONDARY             DS_A1@Asite
               OFFLINE                      DS_A2@Asite
               OFFLINE                      DS_B1@Bsite
               ACQUIRING                    DS_B2@Bsite
```

Upon completion of the resource move, a **clRGinfo** shows the resource now ONLINE on the Bsite and that a node at the Asite is now ONLINE_SECONDARY (Example 6-5).

Example 6-5 Results of the clRGinfo command, resource moved to another site

```
Zachary         ONLINE SECONDARY            DS_A1@Asite
                OFFLINE                     DS_A2@Asite
                OFFLINE                     DS_B1@Bsite
                ONLINE                      DS_B2@Bsite
```

6.6.2 Loss of both nodes at one site

To see what happens when both nodes at a site crash at the same time, perhaps because of a power loss, run the **halt -q** command. In this example, the PPRC disks still have power, and the pprc links between the sites are still functional.

1. Observe where the resource groups are currently ONLINE. In this example, Zachary is on DS_A1 and Gabriel is on DS_A1 (Example 6-6).

Example 6-6 Results of the clRGinfo command before doing a halt on Asite nodes

```
clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
Zachary        ONLINE                      DS_A1@Asite
               OFFLINE                     DS_A2@Asite
               ONLINE SECONDARY            DS_B1@Bsite
               OFFLINE                     DS_B2@Bsite

Gabriel        ONLINE                      DS_A2@Asite
               OFFLINE                     DS_A1@Asite
               ONLINE SECONDARY            DS_B2@Bsite
               OFFLINE                     DS_B1@Bsite

Benjamin       ONLINE                      DS_B2@Bsite
               OFFLINE                     DS_B1@Bsite
               ONLINE SECONDARY            DS_A2@Asite
               OFFLINE                     DS_A1@Asite
```

2. Enter the **halt -q** command on both nodes DS_A1 and DS_A2 simultaneously.

   The resource groups move to the ONLINE SECONDARY nodes when the nodes at the Asite are unresponsive. The PPRC disks are placed into a suspended state on the remote site that is taking over the resources (Example 6-7).

Example 6-7 Results of the dscli lspprc 2001-30ff command on Bsite

```
2001:8001 Suspended   Host Source Metro Mirror 20        60
Disabled      Invalid
2003:8003 Suspended   Host Source Metro Mirror 20        60
Disabled      Invalid
2004:8004 Suspended   Host Source Metro Mirror 20        60
Disabled      Invaid
```

When the Asite nodes are restarted and PowerHA is started, the Bsite can take ownership of the disks, as shown in Example 6-8 by the output of the **dscli lspprc 2001-30ff** command.

Example 6-8 Results of the dscli lpprc command on Bsite

```
dscli -cfg default.dscli.profile lspprc 2001-2004:8001-8004 3001-3004:8101-8104
2001:8001 Full Duplex -      Metro Mirror 20       60            Disabled
Invalid
2003:8003 Full Duplex -      Metro Mirror 20       60            Disabled
Invalid
2004:8004 Full Duplex -      Metro Mirror 20       60            Disabled
Invalid
3001:8101 Full Duplex -      Metro Mirror 30       60            Disabled
Invalid
3002:8102 Full Duplex -      Metro Mirror 30       60            Disabled
Invalid

clRGinfo shows the Asite now as ONLINE SECONDARY
Zachary        ONLINE SECONDARY            DS_A1@Asite
               OFFLINE                     DS_A2@Asite
               ONLINE                      DS_B1@Bsite
               OFFLINE                     DS_B2@Bsite

Gabriel        OFFLINE                     DS_A2@Asite
               ONLINE SECONDARY            DS_A1@Asite
               ONLINE                      DS_B2@Bsite
               OFFLINE                     DS_B1@Bsite

Benjamin       ONLINE                      DS_B2@Bsite
               OFFLINE                     DS_B1@Bsite
               OFFLINE                     DS_A2@Asite
               ONLINE SECONDARY            DS_A1@Asite
```

3. Safely move the resource groups back to the primary Asite.

6.6.3 Loss of local disk storage on one site

The loss of local disks that are defined in the resource groups causes the resource group to fail over to another node or site.

1. Observe where the resource groups are and the disks' status before you remove the disks (Example 6-9).

Example 6-9 Results of the clRGinfo and lspprc commands before local disk loss

```
Zachary        ONLINE                      DS_A1@Asite
               OFFLINE                     DS_A2@Asite
               ONLINE SECONDARY            DS_B1@Bsite
               OFFLINE                     DS_B2@Bsite

Gabriel        ONLINE                      DS_A2@Asite
               OFFLINE                     DS_A1@Asite
               OFFLINE                     DS_B2@Bsite
               ONLINE SECONDARY            DS_B1@Bsite

Benjamin       ONLINE SECONDARY            DS_B2@Bsite
```

```
                OFFLINE                     DS_B1@Bsite
                OFFLINE                     DS_A2@Asite
                ONLINE                      DS_A1@Asite

dscli> lspprc -dev IBM.2107-75BALB1 -remotedev IBM.2107-7585461 8001-81ff
8001:2001 Full Duplex -      Metro Mirror 80       60            Disabled
Invalid
8003:2003 Full Duplex -      Metro Mirror 80       60            Disabled
Invalid
8004:2004 Full Duplex -      Metro Mirror 80       60            Disabled
Invalid
8101:3001 Full Duplex -      Metro Mirror 81       60            Disabled
Invalid
8102:3002 Full Duplex -      Metro Mirror 81       60            Disabled
Invalid
```

2. Remove disks by using the **dscli chvolgrp** command. Use the **lsvolgrp** and **showvolgrp** commands to see the volume group information. In Example 6-10, the A002, A102, A003, and A103 volumes are used for the rootvg.

Example 6-10 The dscli commands lsvolgrp, showvolgrp, and chvolgrp

```
dscli> lsvolgrp
haxd_ds8k_a1        V6  SCSI Mask
haxd_ds8k_a2        V7  SCSI Mask

dscli> showvolgrp V6
Name haxd_ds8k_a1
ID   V6
Type SCSI Mask
Vols 8001 8002 8003 8004 8101 8102 8103 8104 A002 A102

dscli> showvolgrp V7
Name haxd_ds8k_a2
ID   V7
Type SCSI Mask
Vols 8001 8002 8003 8004 8101 8102 8103 8104 A003 A103


dscli> chvolgrp -action remove -volume 8001-8004,8101-8104 V6
dscli> chvolgrp -action remove -volume 8001-8004,8101-8104 V7
```

3. Observe the resource groups that failed over because of a loss of disks (Example 6-11).

Example 6-11 Results of the clRGinfo command showing the resources that moved

```
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
Zachary        ONLINE SECONDARY            DS_A1@Asite
               OFFLINE                     DS_A2@Asite
               ONLINE                      DS_B1@Bsite
               OFFLINE                     DS_B2@Bsite

Gabriel        ONLINE SECONDARY            DS_A2@Asite
               OFFLINE                     DS_A1@Asite
               ONLINE                      DS_B2@Bsite
```

```
                OFFLINE                       DS_B1@Bsite

Benjamin        ONLINE                        DS_B2@Bsite
                OFFLINE                       DS_B1@Bsite
                ONLINE SECONDARY              DS_A2@Asite
                OFFLINE                       DS_A1@Asite
```

Failover: In Example 6-11 on page 256, if PowerHA is unable to read the disks (the disks are missing), it performs a failover.

From the Bsite, the disks are now available for writes with a Full Duplex state (Example 6-12).

Example 6-12 The dscli lspprc on Bsite

```
dscli> lspprc -dev IBM.2107-7585461 -remotedev IBM.2107-75BALB1 2000-30ff
2001:8001 Full Duplex -      Metro Mirror 0
2003:8003 Full Duplex -      Metro Mirror 0
2004:8004 Full Duplex -      Metro Mirror 0
3001:8101 Full Duplex -      Metro Mirror 0
```

4. To return the disks for use, enter the **dscli chvolgrp** command (Example 6-13).

Example 6-13 The dscli chvolgrp command to add volumes back

```
dscli> chvolgrp -action add -volume 8001-8004,8101-8104 V6
dscli> chvolgrp -action add -volume 8001-8004,8101-8104 V7
```

6.6.4 Loss of the PPRC connection between sites

The loss of the PPRC communication should not cause a failover. The site that has the PPRC disks marks them as suspended and continues to use them. When the links are re-established between the sites, the volumes resynchronize and are put back into the Full Duplex state.

6.6.5 Loss of all XD_ip networks between sites

A loss of just the XD_ip networks causes a failover to the remote site, but the primary site also keeps the resource groups ONLINE, which is a split partition, with resource groups now ONLINE on both sites.

Create a failure of the XD_ip network on the Asite by running the **ifconfig down detach** command on the XD_ip adapters. You must run this command on the primary site so that the remote site sees a loss off communication on the network. A cable pull would also work. Using the **ifconfig down** command on the XD_ip network on the remote site is just an adapter down with no attempt to fail over the resources.

1. Observe where the resource groups are currently ONLINE. In Example 6-14, Zachary is on DS_A1, Gabriel is on DS_A1, and Benjamin is on DS_A1. All resources are on Asite DS_A1 for the simplicity of this test.

Example 6-14 clRGinfo before stopping XD_ip networks

```
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zachary        ONLINE                       DS_A1@Asite
               OFFLINE                      DS_A2@Asite
               ONLINE SECONDARY             DS_B1@Bsite
               OFFLINE                      DS_B2@Bsite

Gabriel        OFFLINE                      DS_A2@Asite
               ONLINE                       DS_A1@Asite
               OFFLINE                      DS_B2@Bsite
               ONLINE SECONDARY             DS_B1@Bsite

Benjamin       OFFLINE                      DS_B2@Bsite
               ONLINE SECONDARY             DS_B1@Bsite
               OFFLINE                      DS_A2@Asite
               ONLINE                       DS_A1@Asite
```

2. With the XD_ip on DS_A1 and on DS_A2 as en1, on both DS_A1 and DS_A2, enter the **ifconfig en1 down detach** command.
3. Observe that from Asite the resources are all ONLINE on Asite and the disks are in suspended state (Example 6-15 and Example 6-16 on page 258).

Example 6-15 clRGinfo command issued on DS_A1

```
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zachary        ONLINE                       DS_A1@Asite
               OFFLINE                      DS_A2@Asite
               OFFLINE                      DS_B1@Bsite
               OFFLINE                      DS_B2@Bsite

Gabriel        OFFLINE                      DS_A2@Asite
               ONLINE                       DS_A1@Asite
               OFFLINE                      DS_B2@Bsite
               OFFLINE                      DS_B1@Bsite

Benjamin       OFFLINE                      DS_B2@Bsite
               OFFLINE                      DS_B1@Bsite
               OFFLINE                      DS_A2@Asite
               ONLINE                       DS_A1@Asite
```

Example 6-16 dscli lspprc command issued to CSS server on Asite

```
dscli>  lspprc -dev IBM.2107-75BALB1 -remotedev IBM.2107-7585461 8001-81f
ID        State       Reason                     Type         SourceLSS Timeout
==============================================================================
8001:2001 Suspended   Internal Conditions Target Metro Mirror 80        60
8003:2003 Suspended   Internal Conditions Target Metro Mirror 80        60
```

```
8004:2004 Suspended   Internal Conditions Target Metro Mirror 80        60
8101:3001 Suspended   Internal Conditions Target Metro Mirror 81        60
```

4. Observe that on the Bsite, the resources are all on the Bsite and the disks are all suspended (Example 6-17 and Example 6-18).

Example 6-17 clRGinfo command issued on DS_B1

```
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
Zachary        OFFLINE                     DS_A1@Asite
               OFFLINE                     DS_A2@Asite
               ONLINE                      DS_B1@Bsite
               OFFLINE                     DS_B2@Bsite

Gabriel        OFFLINE                     DS_A2@Asite
               OFFLINE                     DS_A1@Asite
               ONLINE                      DS_B2@Bsite
               OFFLINE                     DS_B1@Bsite

Benjamin       ONLINE                      DS_B2@Bsite
               OFFLINE                     DS_B1@Bsite
               OFFLINE                     DS_A2@Asite
               OFFLINE                     DS_A1@Asite
```

Example 6-18 dscli lspprc command issued to CSS server on Bsite

```
dscli> lspprc -dev IBM.2107-7585461 -remotedev IBM.2107-75BALB1 2000-30ff
ID        State     Reason      Type         SourceLSS Timeout (secs)
==============================================================================
2001:8001 Suspended Host Source Metro Mirror 20        60
2003:8003 Suspended Host Source Metro Mirror 20        60
2004:8004 Suspended Host Source Metro Mirror 20        60
3001:8101 Suspended Host Source Metro Mirror 30        60
3002:8102 Suspended Host Source Metro Mirror 30        60
```

We now have a split partition. When the XD_ip network is restored, the Bsite crashes because of a group services domain merge (GS_DOM_MERGE_ER in errpt).

5. Power on the remote site nodes and start PowerHA on them. If resource groups do not come online or are getting errors, check the output of the **lspprc** command on both sites disks. Data divergence is likely, in which case manual intervention is required to decide which data to use. Depending on the state of the PPRC volumes, you might need to enter the **dscli failbackpprc** or **dscli resumepprc** command on one of the sites.

6.6.6 Total site failure

A *total site failure* is the loss of disks and nodes at one site. The remote site loses communication to the local site through the PPRC and XD_ip connections. A failover of all resources occurs to the remote site.

6.7 Adding and removing LUNs

Adding and removing disks or LUNs from DS PPRC is not dynamic, but you can do it with the cluster services running. The changes to the disks and volume groups are updated on the other nodes as the resource is moved to them. This task is considered a lazy update.

Important: C-SPOC operations do not succeed on nodes at the remote site (that contain the target volumes) for the following LVM operations:

- 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 CSPOC.

However, nodes on the same site as the source volumes can successfully perform these tasks. The changes are then propagated to the other site by using a lazy update. For C-SPOC operations to work on all other LVM operations, we performed all C-SPOC operations when the cluster was active on all PowerHA nodes.

6.7.1 Adding a disk or LUN

To add a disk or LUN:

1. After the LUN is added to the nodes, verify that it is added by entering the **pcmpath query essmap** command to see the mapping. Example 6-19 shows that hdisk3 is added to the volume group zachary.

Example 6-19 pcmpath query essmap |grep hdisk3

```
> lspv
hdisk0          000fe41108faf388                    rootvg          active
hdisk1          000fe41168e0be14                    None
hdisk2          000fe4112498bb13                    zachary         active
hdisk3          000fe4112498bb7a                    None
> pcmpath query essmap |grep hdisk3
hdisk3      path0     21-T1-01[FC] fscsi0   75BALB18002  IBM 2107-900 51.2GB
80      2   0000   0e   Y   R1-B1-H1-ZC    2   RAID5
hdisk3      path1     31-T1-01[FC] fscsi1   75BALB18002  IBM 2107-900 51.2GB
80      2   0000   0e   Y   R1-B2-H3-ZD  133   RAID5
```

2. Ensure that the resource group is ONLINE on the primary site.
3. Add the volume pair of disks to the PPRC resource. To change the PPRC resource, run the **smitty hacmp** command. Select **Extended Configuration** → **Extended Resource Configuration** → **HADSCLI-Managed PPRC Replicated Resource ConfigurationCMP Extended Resources Configuration** → **PPRC-Managed Replicated Resources Configuration** → **DSCLI-Managed PPRC Replicated Resource Configuration** → **Change/Show a PPRC Resource**.

4. Select the PPRC resource to change (Figure 6-25).

```
                          Change / Show PPRC Resource

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

                                                        [Entry Fields]
  PPRC Resource Name                                    Zachary_pprc
  New PPRC Resource Name                               []
* HACMP Sites                                          [Asite Bsite] +
  List of Volume Pairs                                 [8001->2001 8002->2002]
* ESS Pair                                             [HMC_DS8K_B8 HMC_DS8K_B12+
* LSS Pair                                             [80 20] +
* PPRC Type                                            mmir +
* PRI-SEC PortPair IDs                                [I0002->I0102]
* SEC-PRI PortPair IDs                                [I0102->I0002]
* PPRC Link Type                                       FCP +
* PPRC Critical Mode                                   OFF +
* PPRC Recovery Action                                 MANUAL +
  Volume Group                                         [zachary]
```

Figure 6-25 Adding 8002-2002 volume pairs

5. Use SMIT to verify the PPRC configuration.
6. Verify and synchronize the cluster. Enter the **smitty hacmp** command, and select **Extended Configuration** → **Extended Verification and Synchronization**.
7. Use **dscli** to establish the PPRC connection by entering the **dscli mkpprc** command (Example 6-20).

Example 6-20 dscli mkpprc command to create the pprc connection

```
dscli -cfg dscli.profile.admin mkpprc -dev IBM.2107-75BALB1 -remotedev
IBM.2107-7585461  -type mmir -mode full 8002:2002
```

8. Wait for the replication to finish for the new PPRC.
9. Use the **dscli lpprc** command to ensure that it is in Full Duplex state before you continue to the next step (Figure 6-26).

```
dscli lpprc showing still in process of copying:

8001:2001 Full Duplex  -      Metro Mirror 80        60
8002:2002 Copy Pending -      Metro Mirror 80        60
8003:2003 Full Duplex  -      Metro Mirror 80        60
8004:2004 Full Duplex  -      Metro Mirror 80        60
8101:3001 Full Duplex  -      Metro Mirror 81        60
8102:3002 Full Duplex  -      Metro Mirror 81        60
```

Figure 6-26 dscli lsppc showing copy pending

10. Add the physical volume to the cluster. Run the **smitty cspoc** command. Select **Storage** → **Volume Groups** → **Set Characteristics of a Volume Group**.

11. In the Set Characteristics of Volume Group panel (Figure 6-27), select **Add a Volume to a Volume Group**.

```
                     Set Characteristics of a Volume Group

Move cursor to desired item and press Enter.

  Add a Volume to a Volume Group
  Change/Show characteristics of a Volume Group
  Remove a Volume from a Volume Group
  Enable/Disable a Volume Group for Cross-Site LVM Mirroring Verification




  +--------------------------------------------------------------------------+
  |                          Physical Volume Names                           |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  |   000fe4112498bb7a ( hdisk3 on all selected nodes )                      |
  |   000fe41124997862 ( hdisk7 on all selected nodes )                      |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
  | F8=Image                F10=Exit                Enter=Do                 |
F1| /=Find                  n=Find Next                                      |
F9+--------------------------------------------------------------------------+
```

Figure 6-27 Adding hdisk3 to volume group zachary

12. In the Add a Volume to a Volume Group panel (Figure 6-28), select the volume group to which to add the disk. Then, select the hdisk to add to the volume group.

```
                        Add a Volume to a Volume Group

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

                                                        [Entry Fields]
  VOLUME GROUP name                                   zachary
  Resource Group Name                                 Zachary
  Node List                                           DS_A1,DS_A2,DS_B1,DS_>
  Reference node                                      DS_A1
  VOLUME names                                        hdisk3
```

Figure 6-28 Adding a volume to volume group

Attention: When you use CSPOC to modify a volume group, expect to see errors when you contact the remote site. From the output of the SMIT command in Figure 6-28, we notice the following messages:

```
cl_extendvg: Error executing clupdatevg zachary 000fe4112498bb13 on node DS_B1.
cl_extendvg: Error executing clupdatevg zachary 000fe4112498bb13 on node DS_B2.
```

13. Move the resource to the other site to do a lazy update of the volume group. Run the **smitty cspoc** command. Select **Resource Groups and Applications** → **Move a Resource Group to Another Node / Site** → **Move Resource Groups to Another Site**. Select the resource group to move to the other site.

6.7.2 Removing a disk or LUN

To remove a disk or LUN:

1. Use the **lspv -p** command to observe usage on the disks in the volume group. In this sample scenario, we remove hdisk3 from the volume group zachary.
2. Use the **migratepv** command to migrate all logical volumes from the disk to be removed (the source) to the disk that is to remain (the target) (Example 6-21).

Example 6-21 The migratepv command

```
root@DS_550_1_A1 / > migratepv hdisk3 hdisk2
```

3. Remove the disk from the volume group by using the SMIT CSPOC. Run the **smitty cspoc** command, select **Storage** → **Volume Groups** → **Set Characteristics of a Volume Group** → **Remove a Volume from a Volume Group**, and select the volume group and then the disk to remove (Figure 6-29).

```
                    Remove a Volume from a Volume Group

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

                                                      [Entry Fields]
  VOLUME GROUP name                                  zachary
  Resource Group Name                                Zachary
  Node List                                          DS_A1,DS_A2,DS_B1,DS_B2
  VOLUME names                                       hdisk3
  Reference node                                     DS_A1
  FORCE deallocation of all partitions on            no +
    this physical volume?
```

Figure 6-29 hdisk3 to be removed from zachary volume group

Attention: Expect an error from the execution of the **cspoc** command in Figure 6-29 when it tries to reduce the volume group on the other site:

```
cl_reducevg: Error executing clupdatevg zachary 000fe4112498bb13 on node DS_B2.
cl_reducevg: Error executing clupdatevg zachary 000fe4112498bb13 on node DS_B1.
```

This behavior is normal and expected.

4. Move the resource to the other site so that the volume group is updated. Move the resource group to the other node at the remote site to see the change.

5. Determine the volume pair to delete from the PPRC resource. The **pcmpath query essmap** command is used to see the disk-to-volume mapping (Example 6-22).

Example 6-22 pcmpath query output to find volume

```
root@DS_550_1_A1 / > pcmpath query essmap
hdisk3      path0     21-T1-01[FC] fscsi0   75BALB18002  IBM 2107-900 51.2GB

The above output has been truncated and the 8002 is highlighted to show the
volume to be removed from the pprc resource
```

6. Remove the PPRC relationship. Enter the **dscli rmpprc** command.
7. Change the PPRC resource. Run the **smitty hacmp** command. Select **Extended Configuration → Extended Resource Configuration → HADSCLI-Managed PPRC Replicated Resource ConfigurationCMP Extended Resources Configuration → PPRC-Managed Replicated Resources Configuration → DSCLI-Managed PPRC Replicated Resource Configuration**. Select **Change/Show a PPRC Resource**.
8. In the Change/Show PPRC Resource panel, select the PPRC to change (Figure 6-30).

```
                        Change / Show PPRC Resource

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

                                                      [Entry Fields]
  PPRC Resource Name                                  Zachary_pprc
  New PPRC Resource Name                             []
* HACMP Sites                                        [Asite Bsite] +
  List of Volume Pairs                               [8001->2001 8002->2002]
* ESS Pair                                           [HMC_DS8K_B8 HMC_DS8K_B12+
* LSS Pair                                           [80 20] +
* PPRC Type                                          mmir +
* PRI-SEC PortPair IDs                              [I0002->I0102]
* SEC-PRI PortPair IDs                              [I0102->I0002]
* PPRC Link Type                                     FCP +
* PPRC Critical Mode                                 OFF +
* PPRC Recovery Action                               MANUAL +
  Volume Group                                       [zachary]
```

Figure 6-30 Removing 8002-2002 volume pair

9. Verify the PPRC replicated resource. Run the **smitty hacmp** command. Select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → PPRC-Managed Replicated Resources Configuration → DSCLI-Managed PPRC Replicated Resource Configuration → Verify PPRC Configuration**.
10. Remove the PPRC pair. On the node where the resource group is ONLINE, use the **dscli rmpprc** command (for example, **dscli rmpprc 8002:2002**).
11. Verify and synchronize the cluster. Run the **smitty hacmp** command. Then, select **Extended Configuration → Extended Verification and Synchronization**.

6.8 Commands for troubleshooting or gathering information

The following command can be helpful when troubleshooting or gathering information about your system:

- **dscli** commands (/opt/ibm/dscli/dscli):
  - **lspprc**
  - **lskey**
  - **lssi**
  - **lsavailpprcport**
  - **lspprcpath**
  - **mkesconpprcpath**
  - **mkpprcpath**
  - **rmpprcpath**
  - **mkpprc**
  - **rmpprc**
  - **failbackpprc**
  - **failoverpprc**
  - **freezepprc**
  - **pausepprc**
  - **resumepprc**
  - **unfreezepprc**
- New PPRC commands (/usr/es/sbin/cluster/pprc/spprc/cmds/) (Figure 6-31):
  - **cllsspprc -a**
  - **cllsdss -a**
  - **cllscss -a**

```
 / > /usr/es/sbin/cluster/pprc/spprc/cmds/cllsdss -a
ESS name       Sitename       ESS IP address1     Username   Password   ESS IP address2     ESS ID
CSS LIST
HMC_DS8K_B8     Asite           10.12.6.17            dinoadm    ds8kitso
IBM.2107-75BALB1 HMC_DS8K_B8
HMC_DS8K_B12    Bsite           10.114.232.150        itso_user  ds8kitso
IBM.2107-7585461 HMC_DS8K_B12


/ >   /usr/es/sbin/cluster/pprc/spprc/cmds/cllsspprc -a
Resource        Sitelist      Volume Pairs         ESS Pair             LSS Pair      PPRC Type   PriSec
Port IDs SecPri Port IDs Link Type  Crit Mode  Recovery Action   Volume Group
Zachary_pprc    Asite Bsite   8001->2001 8002->2002 HMC_DS8K_B8 HMC_DS8K_B12 80 20    mmir
 I0002->I0102    I0102->I0002     FCP        OFF        AUTOMATED        zachary
Gabriel_pprc    Asite Bsite   8003->2003 8004->2004 HMC_DS8K_B8 HMC_DS8K_B12 80 20    mmir
 I0002->I0102    I0102->I0002     FCP        OFF        AUTOMATED        gabriel
Benjamin_pprc   Asite Bsite   8101->3001            HMC_DS8K_B8 HMC_DS8K_B12 81 30    mmir
 I0133->I0132    I0132->I0133     FCP        OFF        AUTOMATED        benjamin


/ > /usr/es/sbin/cluster/pprc/spprc/cmds/cllscss -a
CSS Name        CLI Type        Sitename        CSS IP Address       Username   Password
HMC_DS8K_B8     DSCLI           Asite           10.12.6.17            dinoadm    ds8kitso
HMC_DS8K_B12    DSCLI           Bsite           10.114.232.150        itso_user  ds8kitso
```

Figure 6-31 Results for cllsdss, cllsspprc, and cllscss

- **pcmpath query essmap** shows the disk-to-volume ID association (Figure 6-32).

```
root@DS_550_1_A1 / > pcmpath query essmap
Disk        Path  P     Location   adapter      LUN SN       Type       Size   LSS    Vol   Rank  C/A   S   Connection  port RaidMode
-------     ----- -    ----------  --------    --------  ------------  ----   ---    ---   ----  ---   -   ----------- ---- --------
hdisk0      path0     21-T1-01[FC] fscsi0   75BALB1A002  IBM 2107-900 10.2GB    a0      2   0000   0e   Y   R1-B1-H1-ZC    2   RAID5
hdisk0      path1     31-T1-01[FC] fscsi1   75BALB1A002  IBM 2107-900 10.2GB    a0      2   0000   0e   Y   R1-B2-H3-ZD  133   RAID5
hdisk1      path0     21-T1-01[FC] fscsi0   75BALB1A102  IBM 2107-900 10.2GB    a1      2   ffff   17   Y   R1-B1-H1-ZC    2   RAID5
hdisk1      path1     31-T1-01[FC] fscsi1   75BALB1A102  IBM 2107-900 10.2GB    a1      2   ffff   17   Y   R1-B2-H3-ZD  133   RAID5
hdisk2      path0     21-T1-01[FC] fscsi0   75BALB18001  IBM 2107-900 51.2GB    80      1   0000   0e   Y   R1-B1-H1-ZC    2   RAID5
hdisk2      path1     31-T1-01[FC] fscsi1   75BALB18001  IBM 2107-900 51.2GB    80      1   0000   0e   Y   R1-B2-H3-ZD  133   RAID5
hdisk3      path0     21-T1-01[FC] fscsi0   75BALB18002  IBM 2107-900 51.2GB    80      2   0000   0e   Y   R1-B1-H1-ZC    2   RAID5
hdisk3      path1     31-T1-01[FC] fscsi1   75BALB18002  IBM 2107-900 51.2GB    80      2   0000   0e   Y   R1-B2-H3-ZD  133   RAID5
hdisk4      path0     21-T1-01[FC] fscsi0   75BALB18003  IBM 2107-900 51.2GB    80      3   0000   0e   Y   R1-B1-H1-ZC    2   RAID5
hdisk4      path1     31-T1-01[FC] fscsi1   75BALB18003  IBM 2107-900 51.2GB    80      3   0000   0e   Y   R1-B2-H3-ZD  133   RAID5
hdisk5      path0     21-T1-01[FC] fscsi0   75BALB18004  IBM 2107-900 51.2GB    80      4   0000   0e   Y   R1-B1-H1-ZC    2   RAID5
hdisk5      path1     31-T1-01[FC] fscsi1   75BALB18004  IBM 2107-900 51.2GB    80      4   0000   0e   Y   R1-B2-H3-ZD  133   RAID5
hdisk6      path0     21-T1-01[FC] fscsi0   75BALB18101  IBM 2107-900 51.2GB    81      1   ffff   17   Y   R1-B1-H1-ZC    2   RAID5
hdisk6      path1     31-T1-01[FC] fscsi1   75BALB18101  IBM 2107-900 51.2GB    81      1   ffff   17   Y   R1-B2-H3-ZD  133   RAID5
hdisk7      path0     21-T1-01[FC] fscsi0   75BALB18102  IBM 2107-900 51.2GB    81      2   ffff   17   Y   R1-B1-H1-ZC    2   RAID5
hdisk7      path1     31-T1-01[FC] fscsi1   75BALB18102  IBM 2107-900 51.2GB    81      2   ffff   17   Y   R1-B2-H3-ZD  133   RAID5
hdisk8      path0     21-T1-01[FC] fscsi0   75BALB18103  IBM 2107-900 51.2GB    81      3   ffff   17   Y   R1-B1-H1-ZC    2   RAID5
hdisk8      path1     31-T1-01[FC] fscsi1   75BALB18103  IBM 2107-900 51.2GB    81      3   ffff   17   Y   R1-B2-H3-ZD  133   RAID5
hdisk9      path2     21-T1-01[FC] fscsi0   75BALB18104  IBM 2107-900 51.2GB    81      4   ffff   17   Y   R1-B1-H1-ZC    2   RAID5
hdisk9      path3     31-T1-01[FC] fscsi1   75BALB18104  IBM 2107-900 51.2GB    81      4   ffff   17   Y   R1-B2-H3-ZD  133   RAID5
```

Figure 6-32 Results of the pcmath query essmap command

6.9 PowerHA Enterprise Edition: SPPRC DSCLI security enhancements

Security is enhanced by creating an encrypted password file for use with the **dscli** commands. This file is created based on the entries in the HACMP configuration. The first time that sync and verify are run the encrypted `passwd` file is created in the `/var/hacmp/logs/pprc/run/security` directory (or the directory that you specified for log file redirection). This directory has permissions of 600, and the files are added as a file collection that gets pushed to all the other nodes. The file gets created only one time on the first node.

Keep in mind the following maintenance considerations:

- If the HMC/SMC password is changed, you must change the password in the HACMP configuration and remove the encrypted `passwd` file. If you do not do this step, the next time that verify is run, the wrong `passwd` file is used.
- If you have more than one `passwd` file and one of them is removed, the file can be removed from all the other nodes, causing the commands to fail. Do not remove any files from `../pprc/run/security` unless you changed your passwords, removed all files, and re-created them all to ensure that you have good `passwd` files.

7

Configuring PowerHA SystemMirror Enterprise Edition with SRDF replication

This chapter provides a practical description of the PowerHA Enterprise Edition configuration in an environment that uses EMC nationSymmetrix Remote Data Facility (SRDF) replication.

The chapter includes the following sections:

- General considerations
- Planning
- Environment description
- Installation and configuration
- Test scenarios
- Maintaining the cluster configuration with SRDF replicated resources
- Troubleshooting PowerHA Enterprise Edition SRDF managed replicated resources
- Commands for managing the SRDF environment

7.1 General considerations

The IBM PowerHA SystemMirror Enterprise Edition (formerly HACMP/XD) support several extended distance and disaster recovery storage-based data replication facilities. Such facilities include EMC Symmetrix Remote Data Facility.

The PowerHA Enterprise Edition (PowerHA Enterprise Edition) support for the EMC SRDF solution combines EMC SRDF and PowerHA Enterprise Edition Version 6.1 facilities. It provides a fully automated, highly available disaster recovery and management solution in a cluster with EMC Symmetrix storage.

Based on the integration of EMC SRDF with PowerHA, the PowerHA solution provide the following functions:

- Management of EMC SRDF automatic failover of SRDF-protected volume pairs between nodes within a site
- Management of EMC SRDF automatic failover of SRDF-protected volume pairs between sites
- Automatic failover/reintegration of server nodes that are attached to SRDF protected disk volume pairs within and between sites
- Support for selected inter-site management policies
- Support for cluster verification and synchronization
- Limited support for the PowerHA C-SPOC
- Flexible user-customizable resource group policies

EMC SRDF provides the following integration features with the PowerHA Enterprise Edition:

- When a production site failure occurs, restarts the application with the copy of the data at the secondary site
- Manages the states of the SRDF relationships so that the customer knows when the data is consistent and can control on which site the applications are run
- Manages both synchronous (SRDF/S) and asynchronous (SRDF/A) mirroring
- Supports SRDF consistency Groups (SRDF/CG)

7.1.1 Operational considerations for the current release of integration

The following considerations apply for the PowerHA and SRDF integration features:

- The EMC Symmetrix storage must be directly connected to the cluster node. Storage configurations that use an indirect connection to the AIX host system through a host computer are not supported.
- PowerHA does not trap SNMP notification events for the Symmetrix storage. If the SRDF link goes down when the cluster is up, and later the link is repaired, you must manually resynchronize the pairs.
- Pair creation must be done outside of the cluster control. You must create the pairs before you start the cluster services.
- PowerHA Enterprise Edition and SRDF do not correct the pairs or resynchronize the pairs if the pair states are in an invalid state. A pair in an invalid state might lead to data corruption if the PowerHA cluster tries to recover from this state.

- Resource groups that are managed by PowerHA cannot contain volume groups with disks that are SRDF-protected and disks that are not SRDF-protected.
- C-SPOC cannot perform the following LVM operations on nodes at the remote site (that contain the target volumes):
  - The creation of a volume group.
  - Operations that require nodes at the target site to read from the target volumes. Such operations cause an error message in CSPOC and include such functions as changing file system size, changing the mount point, and adding LVM mirrors. However, nodes on the same site as the source volumes can successfully perform these tasks, and the changes are then propagated to the other site by using a lazy update.

C-SPOC operations: For C-SPOC operations to work on all other LVM operations, perform all C-SPOC operations with the SRDF pairs in synchronized or consistent states, and the cluster must be ACTIVE on all nodes.

- SRDF has the following functional considerations:
  - Multihop configurations are not supported.
  - Mirroring to BCV devices is not supported.
  - Concurrent RDF configurations are not supported.

7.1.2 Documentation resources

For information about the installation and configuration of the SRDF integration feature, see the release notes file in `/usr/es/sbin/cluster/release_notes_xd` for the version of PowerHA Enterprise Edition that you implemented.

See the following documentation resources:

- *HACMP for AIX 6.1 Administration Guide*, SC23-4862.
- HACMP manuals on AIX 6.1 in the AIX 6.1 Information Center:

  `http://publib.boulder.ibm.com/infocenter/aix/v6r1/index.jsp`
- Storage options from EMC:

  `http://www.emc.com/products/category/storage.htm`
- EMC SRDF technology:

  `http://www.emc.com/products/detail/software/srdf.htm`
- EMC online service management:

  `http://powerlink.emc.com`

7.2 Planning

This section provides a description of the planning considerations when planning to integrate IBM PowerHA with EMC SRDF.

7.2.1 Hardware considerations

When you plan for PowerHA Enterprise Edition integration support with EMC SRDF, consider that only specific EMC Symmetrix storage subsystems are supported. For the specific storage devices and microcode levels that are supported, see EMC Release for Announcement (RFA). In the current release of the integration feature, the following Symmetrix models are supported:

- DMX-3 with Enginuity software 5772.103.93 and subsequent fix packages
- DMX-4 with Enginuity software 5773.134.94 and subsequent fix packages
- V-Max with Enginuity software 5874.188.156 and subsequent fix packages

Before you install the cluster software, verify that the systems and the storage resources are prepared as follows:

- Power Systems servers and Symmetrix storage are connected in both sites.
- Fabrics are operational and the zoning configuration is in place. Besides the normal attachment of the hosts to the storage subsystem using the storage area network (SAN) in each site, you must configure your SAN to enable communication between the RF ports on the storage subsystems for SRDF replication. See *EMC Symmetrix Remote Data Facility (SRDF) - Connectivity Guide*, P/N 300-003-885.
- Volumes on each storage are defined and mapped to the AIX hosts in each site.
- Gatekeeper disks are defined and attached on all hosts. A gatekeeper disk (GK) is a small device (usually under 10 MB) defined on the storage subsystem, which allows SYMCLI commands to retrieve configuration and status information from the Symmetrix array without interfering with normal Symmetrix operations. For more information about managing the gatekeeper devices, see the *EMC Solutions Enabler - Installation Guide*, P/N 300-008-918, document for your version of SYMCLI.

7.2.2 Software prerequisites

Installation and configuration of the PowerHA cluster using EMC SRDF requires the following prerequisites:

- One of the following versions of AIX operating system on the cluster nodes:
  - AIX Version 6.1 with 6100-02 Technology Level or later
  - IBM AIX 5L™ Version 5.3 with 5300-09 Technology Level or later
- PowerHA Enterprise Edition Version 6.1 (**`cluster.es.server.rte`** 6.1) or later.
- Java must be installed on each of the cluster nodes. IBM Java v1.4 is preferred.

  Important: If you want to use a non-IBM version of Java, you must create a link from the standard Java location to your version (that is, create a symbolic link from `/usr/java14/jre/bin/java` to the Java executable file). IBM has certified only the IBM version of Java.

- Solutions Enabler for AIX software (SYMCLI software for AIX). The file set level for SYMCLI.SYMCLI.rte must be 7.0.0.0 or later.
- PowerPath and Disk Driver attachment V5.3 and subsequent fix packages.

For documentation and software downloads that are related to the Powerpath product and for SRDF product documentation, see the EMC Powerlink site at:

`https://powerlink.emc.com`

7.3 Environment description

Our environments consist of two sites with two nodes inside each site (Figure 7-1).

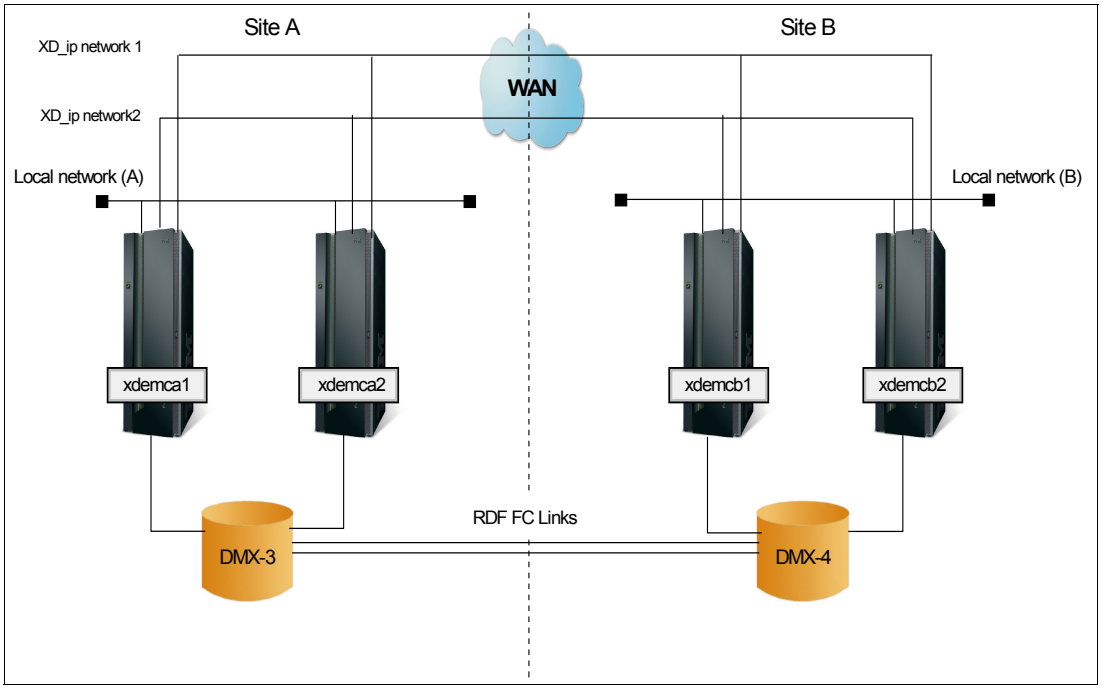


Figure 7-1 General overview of the PowerHA environment with SRDF replication

Each site has an EMC storage subsystem that is attached to the local nodes. Direct Matrix Architecture (DMX) storage subsystems are connected to each other using redundant Fibre Channel (FC) link connections.

Two TCP/IP links (networks) are dedicated for XD heartbeat between the sites. One Ethernet adapter per node is dedicated for an XD_ip network. In our case, the same IP subnet is used for the XD_ip network, but in a production environment, separate IP segments might be used on each site that is being routed between each other. For a production environment configure multiple communication paths between the sites, including any additional non-IP network in case of routing or other TCP/IP issues between the sites. In addition to the XD_ip networks, each site has local subnets for client access.

For our scenario, we create first a resource group, having the primary site A containing a group of SRDF relationships that operate in synchronous mode (SRDF/S). Later, we add a second resource group to the cluster configuration, having the primary site B with a group of SRDF relations that operate in asynchronous mode (SRDF/A). We use the capability of PowerHA Enterprise Edition to dynamically integrate a second site-related resource group in an SRDF environment without disrupting the existing one. PowerHA Enterprise Edition supports coexistence between SRDF replicated resources that operate in different modes.

Consider that our configuration is used only for describing the installation and configuration steps and for testing several scenarios. In a production environment, whether to use synchronous or asynchronous replication is determined by several factors such as distance between sites, communication bandwidth, and application I/O pattern. For general considerations that are related to replication options, see 1.4.1, “Synchronous versus asynchronous replication” on page 22.

For illustration purposes, the resource groups that we create contain only IP labels, volume groups, file systems, and SRDF replicated resources. Further details are provided within the next sections.

In our environment, we use the following versions of the software on the cluster nodes:

- AIX V6.1 TL04 SP02
- PowerHA Enterprise Edition V6.1 SP01
- EMC disk attachment drivers 5.3.0.2
- EMC PowerPath 5.3
- SYMCLI V7.0.0.0

Two EMC Symmetrix storage subsystems are used for the SRDF relationships:

- DMX-3, mcode Version 5772.103.93
- DMX-4, mcode Version 5773.134.94

7.4 Installation and configuration

The following sections explain the steps to install the PowerHA Enterprise Edition 6.1 software, the prerequisites, and the configuration of the SRDF integration in PowerHA Enterprise Edition.

7.4.1 Installing and configuring the prerequisite software

This section explains how to install and configure the prerequisite software:

- EMC disk drivers
- PowerPath software
- SYMCLI packages

At the end of the installation, we run the discovery process to identify the Symmetrix storage subsystems that are attached in our environment.

To install the prerequisite software on AIX for all nodes in the cluster:

1. Install the EMC drivers that are required to access the LUNs on the DMX arrays. To install the following file sets, use the AIX **`installp`** command:
   - Symmetrix disk attachment drivers:
     - `EMC.Symmetrix.aix.rte`
     - `EMC.Symmetrix.fcp.rte`

File sets: When you install the device drivers for the Symmetrix storage, you can choose the `EMC.Symmetrix.aix.rte` file set and only *one* of the following file sets:

- `EMC.Symmetrix.fcp.rte` (EMC Symmetrix FCP support software)
- `EMC.Symmetrix.fcp.MPIO.rte` (EMC Symmetrix FCP MPIO support software)
- `EMC.Symmetrix.fcp.PowerMPIO.rte` (EMC Symmetrix FCP PowerPath MPIO support software)

For the current release of the SRDF integration feature, use EMC Symmetrix FCP or PowerPath MPIO support software.

– PowerPath multipathing software and utilities:

  - EMCpower.base
  - EMCpower.consistency_grp
  - EMCpower.encryption
  - EMCpower.migration_enabler
  - EMCpower.mpx

Register the Powerpath software license by applying the available license keys by using the **emcpreg -install** command. Use the **powermt check_registration** command to verify the installation keys. For information about the licensing features for Powerpath software, contact the EMC representative.

If the Symmetrix disks are already configured on the system:

a. Remove their existing definition:

```
rmdev -dl <hdisk#>
```

b. Reconfigure the AIX disk definition in the Object Data Manager (ODM):

```
/usr/lpp/Symmetrix/bin/emc_cfgmgr
```

c. Initialize the PowerPath devices:

```
powermt config
```

For more information about Powerpath installation and configuration, see *EMC PowerPath for AIX - Installation and Administration Guide,* P/N 300-008-341.

2. Install the Solutions Enabler Software (SYMCLI). In our environment, we use a shared Network File System (NFS) directory that is mounted on all hosts, containing the SYMCLI 7.0 installation media, and we run the installer script in silent mode on each cluster node (Example 7-1).

Example 7-1 Installing the SYMCLI packages in silent mode

```
root@xdemca1:/mnt/EMC/SE7000>./se7000_install.sh -install -silent

 #------------------------------------------------------------------------------
 #                            EMC Installation Manager
 #------------------------------------------------------------------------------
 Copyright 2009, EMC Corporation
 All rights reserved.

 The terms of your use of this software are governed by the
 applicable contract.

 Solutions Enabler Native Installer[RT] Kit Location : /mnt/EMC/SE7000

 Checking for OS version compatibility......

.........
```

You can use **se7000_install.sh -check** to verify the installation.

For more information about SYMCLI V7.0, see *EMC Solutions Enabler Version 7.0 - Installation Guide,* P/N 300-008-918.

3. At the end of the installation process, run the discovery process to update the SYMAPI local database with the discovered Symmetrix storage. Verify that the storage is attached to the nodes in each site. Also, verify that you have applied the license key for the

BASE/Symmetrix feature before you run the discovery by listing the contents of the license file. For our installation, the default location is `/var/symapi/config/symapi_licenses.dat`.

Our PowerHA configuration with SRDF uses the following licensed features:

- BASE / Symmetrix
- SRDF / Symmetrix
- SRDFA / Symmetrix
- SRDF/CG / Symmetrix

For more information about the Symmetrix license features, contact an EMC representative.

Run **`symcfg discover`** to start the discovery process. At the end of the process, verify the discovered devices by running the **`symcfg list`** command (Example 7-2).

Example 7-2 Symmetrix arrays (site A view)

```
root@xdemca1:/>symcfg list

                                S Y M M E T R I X

                                          Mcode    Cache      Num Phys  Num Symm
    SymmID       Attachment  Model        Version  Size (MB)  Devices   Devices

    00019010 0304 Local      DMX3-24      5772       32768       294      4063
    00019010 1983 Remote     DMX4-24      5773       65536         0      1743
```

Running the **`symcfg list`** command from local node xdemca1 shows that the Symmetrix system model DMX3-24 is locally attached to our node and that the Symmetrix system DMX4-24 is remotely attached to the DMX3. Note the Symmetrix ID for both local and remote systems. Repeat the discovery process on each node of the cluster. Example 7-3 shows a similar view of the storage devices from site B.

Example 7-3 Symmetrix arrays (site B view)

```
root@xdemcb1:/>symcfg list

                                S Y M M E T R I X

                                          Mcode    Cache      Num Phys  Num Symm
    SymmID       Attachment  Model        Version  Size (MB)  Devices   Devices

    00019010 1983 Local      DMX4-24      5773       65536       259      1743
    00019010 0304 Remote     DMX3-24      5772       32768         0      4063
```

In the output of Example 7-3, note the change between the local and the remote storage subsystem when you run the list command from site B.

Fibre Channel suggestions for the SRDF environment

Use the following additional settings for a cluster that uses the Symmetrix storage and the SRDF integration features:

- Set the `fc_err_recov` attribute for Fibre Channel ports to fast_fail (Example 7-4).

Example 7-4 Setting Fibre Channel port to fast_fail

```
# lsattr -El fscsi0
attach       switch    How this adapter is CONNECTED         False
```

```
dyntrk       no        Dynamic Tracking of FC Devices        True
fc_err_recov fast_fail FC Fabric Event Error RECOVERY Policy True
scsi_id      0x70400   Adapter SCSI ID                       False
sw_fc_class  3         FC Class for Fabric                   True
```

- Configure the `gkselect` file after you install the SYMCLI package to limit the use of gatekeepers and the number of communication paths. Use the **`symdev list pd`** command and choose 5 - 6 gatekeepers. Limiting the number of GK paths reduces the amount of time that is spent querying paths when communication is lost, resulting in long wait times that lead to cluster config_too_long errors. Add the selected devices to the `/var/symapi/config/gkselect` configuration file on each cluster node (Example 7-5).

Example 7-5 A gkselect file

```
# cat /var/symapi/config/gkselect
/dev/rhdiskpower10
/dev/rhdiskpower11
/dev/rhdiskpower12
/dev/rhdiskpower13
/dev/rhdiskpower14
/dev/rhdiskpower15
```

- To automate the lock recovery for SYMCLI if a node crash occurs, consider adding the following options in the `/var/symapi/config/daemon_options` configuration file on all of the cluster nodes:

```
storapid:internode_lock_recovery=enable
storapid:internode_lock_recovery_heartbeat_interval=10
storapid:internode_lock_information_export=enable
```

Enable also the RDF daemon on all the cluster nodes by setting the following option in the `/var/symapi/config/options` configuration file:

```
SYMAPI_USE_RDFD=ENABLE
```

After you update the option files, restart all daemons for the changes to take effect:

```
/usr/symcli/storbin/stordaemon shutdown all
/usr/symcli/storbin/stordaemon start storapid
/usr/symcli/storbin/stordaemon start storrdfd
```

7.4.2 Installing the cluster software

Before you install the cluster file sets, install the AIX prerequisite file sets. For more information, see the *HACMP for AIX Installation Guide*, SC23-4862, at:

`.ibm.com/infocenter/clresctr/vxrx/index.jsp`

For the SRDF integration with PowerHA, you must install the following specific file sets:

- `cluster.es.sr.cmds`: ES HACMP, EMC SRDF commands
- `cluster.es.sr.rte`: ES HACMP, EMC SRDF run time

Use the AIX **`installp`** command or the SMIT menus to install the cluster software from the installation media. For our environment, we install the following cluster packages:

- `cluster.adt.es`
- `cluster.doc.en_US.es`
- `cluster.es.client`
- `cluster.es.cspoc`

- cluster.es.server
- cluster.es.sr
- cluster.es.worksheets
- cluster.license
- cluster.man.en_US.es
- cluster.xd.license

7.4.3 Defining the cluster topology

This section explains the steps that are performed to set up the cluster topology. We provide the relevant System Management Interface Tool (SMIT) panel examples for defining the cluster topology for our environment. For information about the PowerHA topology tasks, SMIT menus, and paths that are needed to run a particular configuration task, see the *HACMP for AIX Administration Guide*, SC23-4862.

First, we prepare the network configuration as illustrated in Figure 7-2.

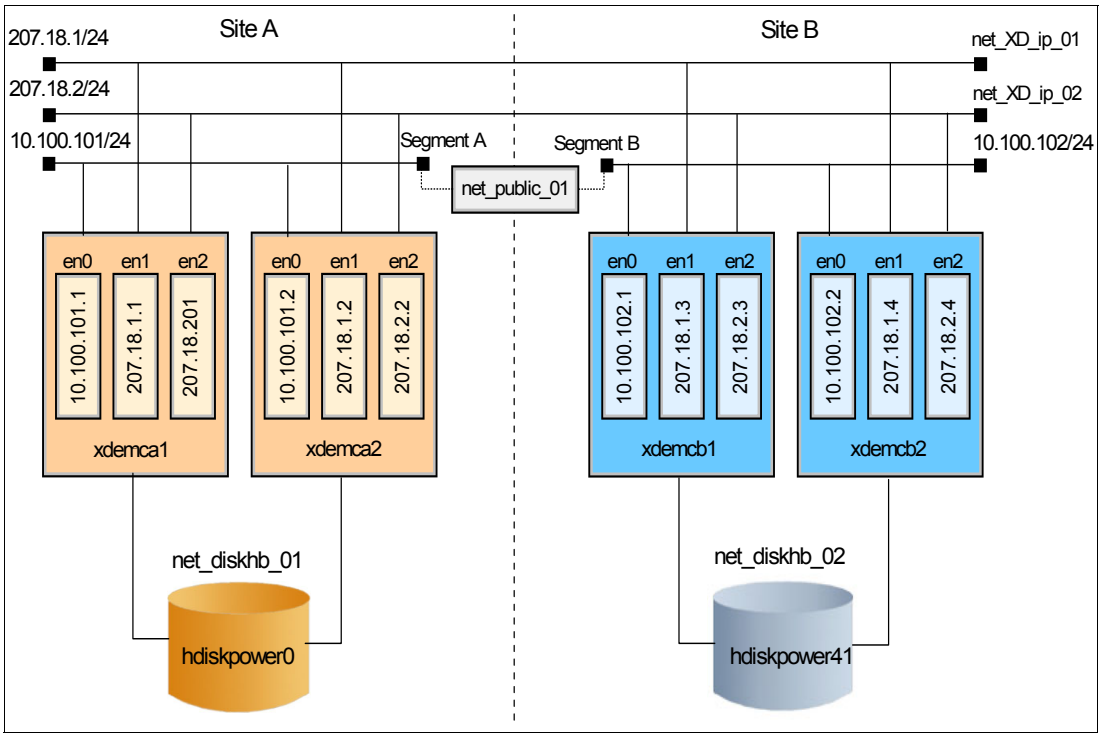


Figure 7-2 Cluster topology overview

The cluster configuration has the following IP networks:

- net_XD_ip_01 and net_XD_ip_02 are XD_ip-type networks, and they are used for cluster inter-site heartbeat.
- net_public_01 is an ether-type network that is used for the client communication with the nodes at each site. This network contains two IP segments, one in each site. There is no routing between these segments. Each is used within a site to hold the service IP addresses specific to that site when the resource group is activated.

The IP interface names that are used for the cluster communication interfaces and service IP labels are defined in the /etc/hosts file on each node (Example 7-6).

Example 7-6 IP interface names and service IP labels

```
# XD_IP SiteA
207.18.1.1      xdemca1_xdip1
207.18.2.1      xdemca1_xdip2
207.18.1.2      xdemca2_xdip1
207.18.2.2      xdemca2_xdip2

# XD_IP SiteB
207.18.1.3      xdemcb1_xdip1
207.18.2.3      xdemcb1_xdip2
207.18.1.4      xdemcb2_xdip1
207.18.2.4      xdemcb2_xdip2

#Boot IP addresses
10.100.101.1 xdemca1_boot
10.100.101.2 xdemca2_boot
10.100.102.1 xdemcb1_boot
10.100.102.2 xdemcb2_boot

# Service IP addresses
10.10.10.1 xdemcaa_sv
10.10.11.1 xdemcab_sv
10.10.11.2 xdemcbb_sv
10.10.10.2 xdemcba_sv
```

In addition to the IP networks, there are two disk heartbeat networks local in each site:

- net_diskhb_01
- net_diskhb_02

To define the cluster topology:

1. Define the cluster, which is xdemc in this example (Figure 7-3). By using SMIT menus, enter the **smitty hacmp** command. Select **Extended Configuration** → **Extended Topology Configuration** → **Configure an HACMP Cluster** → **Add/Change/Show an HACMP Cluster**.

```
                    Add/Change/Show an HACMP Cluster

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

                                                        [Entry Fields]
* Cluster Name                                        [xdemc]

  NOTE: HACMP must be RESTARTED
  on all nodes in order for change to take effect
```

Figure 7-3 Adding the cluster definition

2. Add the nodes to the cluster definition. Add both the local and the remote nodes. Ensure that there is a communication path between all nodes in both sites (Figure 7-4). We use IP addresses of the interfaces for inter-site communication.

```
                        Add a Node to the HACMP Cluster

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

                                                      [Entry Fields]
* Node Name                                          [xdemca1]
  Communication Path to Node                         [xdemca1_xdip1]       +
```

Figure 7-4 Adding the node definition to the cluster

3. Define the sites. In our environment, we use the site that is named siteA, which contains the xdemca1 and xdemca2 nodes (Figure 7-5).

```
                                  Add a Site

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

                                         [Entry Fields]
  * Site Name                           [siteA]                        +
  * Site Nodes                          xdemca1 xdemca2                +
```

Figure 7-5 Defining the cluster site for location A

A similar operation is performed for siteB, which contains the xdemcb1 and xdemcb2 nodes (Figure 7-6).

```
                                  Add a Site

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

                                         [Entry Fields]
  * Site Name                           [siteB]                        +
  * Site Nodes                          xdemcb1 xdemcb2                +
```

Figure 7-6 Defining the cluster site for location B

Run the cluster discovery process to collect information, including their IP and disk configuration, which can be used later to facilitate the topology definition.

4. Configure the environment with three IP networks:
   – A public network, which includes both IP segments of the sites (Figure 7-7).

```
               Add an IP-Based Network to the HACMP Cluster

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

                                                      [Entry Fields]
* Network Name                                       [net_public_01]
* Network Type                                        ether
* Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.255.0]
* Enable IP Address Takeover via IP Aliases          [Yes]                +
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 7-7 Defining the public network

   We enable IP address takeover (IPAT) by using aliasing for this network. Later, we define the service IP labels configurable on multiple nodes with two IP addresses, each bound to a site.

   – Two XD_ip networks that are used for PowerHA heartbeating between sites. Figure 7-8 shows an example definition of a XD_ip network that is used for our environment.

```
             Add an IP-Based Network to the HACMP Cluster

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

                                                      [Entry Fields]
* Network Name                                       [net_XD_ip_01]
* Network Type                                        XD_ip
* Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.255.0]
* Enable IP Address Takeover via IP Aliases          [No]                 +
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 7-8 Defining the XD_ip network

We disable the IPAT on this network (XD_ip). Later, we define service IP labels node-bounded for this network.

5. Define the cluster communication interfaces for each network:
   – The client access network. We add the boot interfaces from both local and remote nodes on the net_public_01 network (Figure 7-9).

```
 # Network / Node                                                          |
|   #       Interface         IP Label IP Address                         |
|                                                                         |
|   # net_public_01 / xdemca1                                             |
| >         en0               xdemca1_boot                    10.100.101 |
|                                                                         |
|   # net_public_01 / xdemca2                                             |
| >         en0               xdemca2_boot                    10.100.101 |
|                                                                         |
|   # net_public_01 / xdemcb1                                             |
| >         en0               xdemcb1_boot                    10.100.102 |
|                                                                         |
|   # net_public_01 / xdemcb2                                             |
| >         en0               xdemcb2_boot                    10.100.102 |
|                                                                         |
```

Figure 7-9 Adding the communication interfaces for the public network

   Now, add both IP segments in sites A and B (10.100.101/24 and 10.100.102/24) to the same cluster network.
   – The XD_ip network. We define on each node the associated communication interface. See Figure 7-10 for an example of adding a communication interface for a node to an XD_ip network. We apply the same operation for all cluster nodes' XD_ip interfaces.

```
                    Add a Communication Interface

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

                                                      [Entry Fields]
* IP Label/Address                                   [xdemca1_xdip1]       +
* Network Type                                       XD_ip
* Network Name                                       net_XD_ip_01
* Node Name                                          [xdemca1]             +
  Network Interface                                  []
```

Figure 7-10 Adding the communication interfaces for the XD_ip network

For each communication interface in the XD_ip networks, define a node-bounded service IP address. Define the node-bound IP address. Run the **smitty hacmp** command. Select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure HACMP Service IP Labels/Addresses → Add a Service IP Label/Address → Bound to a Single Node**. Select the node name and the IP interface name on that node for the XD_ip network.

6. Add the non-IP communication devices. We use a disk heartbeat network at each site (Figure 7-2 on page 276). We use a dedicated heartbeat disk, part of an enhanced-concurrent volume group, defined locally in each site. To facilitate adding communication devices into the cluster definition, we run the discovery process on each node of the cluster. Figure 7-11 shows we add the hdiskpower0 associated communication devices on nodes xdemca1 and xdemca2 at site A.

```
 Configure HACMP Communication Interfaces/Devices

+--------------------------------------------------------------------------+
|  Select Point-to-Point Pair of Discovered Communication Devices to Add   |
|                                                                          |
| Move cursor to desired item and press F7.                                |
|     ONE OR MORE items can be selected.                                   |
| Press Enter AFTER making all selections.                                 |
|                                                                          |
| [TOP]                                                                    |
|   # Node                              Device   Pvid                      |
| >   xdemca1                           hdiskpower0 00cfb52d62799c35       |
| >   xdemca2                           hdiskpower0 00cfb52d62799c35       |
|     xdemcb1                           hdiskpower41 00cfb52d62799c35      |
|     xdemcb2                           hdiskpower41 00cfb52d62799c35      |
|                                                                          |
| F1=Help                F2=Refresh             F3=Cancel                  |
| F7=Select              F8=Image               F10=Exit                   |
| Enter=Do               /=Find                 n=Find Next                |
+--------------------------------------------------------------------------+
```

Figure 7-11 Adding the communication devices for nodes at site A

At the end of the task the cluster network definition for disk heartbeat is automatically created.

7. Verify and synchronize the cluster definition. Example 7-7 shows the cluster topology that is defined up to this point after you synchronize the cluster definitions across the nodes by running the **cltopinfo** command.

Example 7-7 Cluster topology definition

```
root@xdemca1:/>cltopinfo
Cluster Name: xdemc
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 5 network(s) defined

NODE xdemca1:
        Network net_XD_ip_01
                xdemca1_xdip1   207.18.1.1
        Network net_XD_ip_02
                xdemca1_xdip2   207.18.2.1
        Network net_diskhb_01
                xdemca1_hdiskpower0_01  /dev/hdiskpower0
        Network net_diskhb_02
        Network net_public_01
```

```
                xdemca1_boot    10.100.101.1

NODE xdemca2:
        Network net_XD_ip_01
                xdemca2_xdip1   207.18.1.2
        Network net_XD_ip_02
                xdemca2_xdip2   207.18.2.2
        Network net_diskhb_01
                xdemca2_hdiskpower0_01  /dev/hdiskpower0
        Network net_diskhb_02
        Network net_public_01
                xdemca2_boot    10.100.101.2

NODE xdemcb1:
        Network net_XD_ip_01
                xdemcb1_xdip1   207.18.1.3
        Network net_XD_ip_02
                xdemcb1_xdip2   207.18.2.3
        Network net_diskhb_01
        Network net_diskhb_02
                xdemcb1_hdiskpower41_01 /dev/hdiskpower41
        Network net_public_01
                xdemcb1_boot    10.100.102.1

NODE xdemcb2:
        Network net_XD_ip_01
                xdemcb2_xdip1   207.18.1.4
        Network net_XD_ip_02
                xdemcb2_xdip2   207.18.2.4
        Network net_diskhb_01
        Network net_diskhb_02
                xdemcb2_hdiskpower41_01 /dev/hdiskpower41
        Network net_public_01
                xdemcb2_boot    10.100.102.2
```

7.4.4 Defining the SRDF configuration

Figure 7-12 illustrates our SRDF storage configuration.

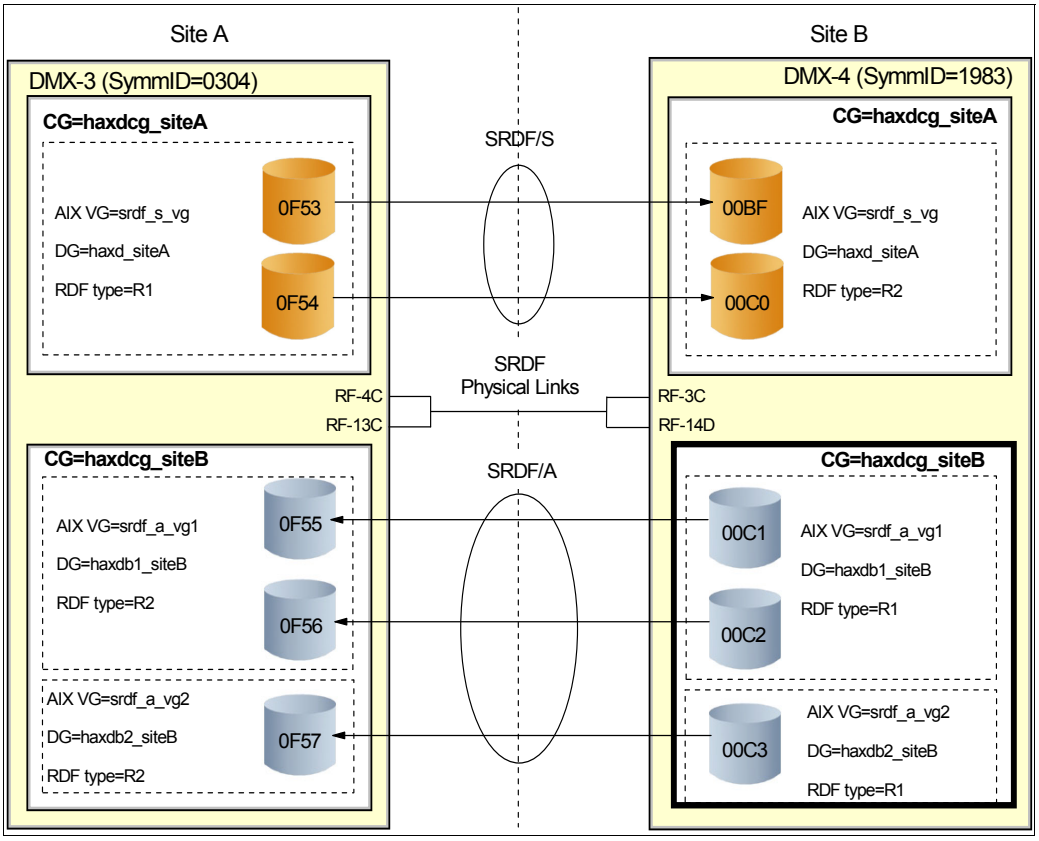


Figure 7-12 Diagram of the storage configuration for SRDF replication

Table 7-1 shows the disk association with the volume groups for our environment. We define the following volume groups:

- srdf_s_vg, enhanced concurrent capable, which is associated with the resource group with primary site A
- srdf_a_vg1, srdf_a_vg2: standard AIX volume groups that are associated with the resource group with primary site B

Table 7-1 Volumes and volume groups that are defined on the Symmetrix storage subsystems

Site	Node	Symmetrix ID (SymmID)	Volume ID (Sym)	AIX Disk	AIX volume group
siteA	xdemca1/ xdemca2	0304	0F52	hdiskpower0	diskhb_a_vg
			0F53 0F54	hdiskpower1 hdiskpower2	srdf_s_vg
			0F55 0F56	hdiskpower3 hdiskpower4	srdf_a_vg1
			0F57	hdiskpower5	srdf_a_vg2

Site	Node	Symmetrix ID (SymmID)	Volume ID (Sym)	AIX Disk	AIX volume group
siteB	xdemcb1/ xdemcb2	1983	00BE	hdiskpower41	diskhb_b_vg
			00BF 00C0	hdiskpower42 hdiskpower43	srdf_s_vg
			00C1 00C2	hdiskpower44 hdiskpower45	srdf_a_vg1
			00C3	hdiskpower46	srdf_a_vg2

AIX definitions practice: In our environment, the nodes in the same site have the same mapping of Symmetrix devices to the PowerPath hdiskpower# devices in AIX. Although this configuration is not required, keep the AIX disk definitions the same on all nodes of the cluster for facilitating the configuration and management operations of the disk volumes with the cluster.

Table 7-2 details the configuration for the device groups and the composite groups that are used in our environment.

Table 7-2 Definitions of SRDF pairs, device, and composite groups

Composite group	Device group	RDF1 members	RDF1 group name (no.)	RDF2 members	RDF2 group name (no.)
haxdcg_siteA	haxd_siteA	0F53	haxd1_rdfg(41)	00BF	haxd1_rdfg(41)
		0F54	haxd1_rdfg(41)	00C0	haxd1_rdfg(41)
haxdcg_siteB	haxd1_siteB	00C1	haxd2_rdfg(42)	0F55	haxd2_rdfg(42)
		00C2	haxd2_rdfg(42)	0F56	haxd2_rdfg(42)
	haxd2_siteB	00C3	haxd2_rdfg(42)	0F57	haxd2_rdfg(42)

We create two dynamic RDF groups:

- haxd1_rdfg(41) for the SRDF relations A → B
- haxd2_rdfg for the SRDF relations B → A containing the corresponding device pairs.

RDF1 and RDF2 represent the roles of the devices in the replication relationship:

- Source (R1 type)
- Target (R2 type)

The devices part of the SRDF pairs in the same storage subsystem are grouped into device groups (DGs). A device group can contain only storage volumes in the same storage subsystem. DG is a logical entity that is used on each host for managing the SRDF relationships. The DG configuration is stored in the SYMAPI database, local on each node in a default configuration.

The DGs defined can be further grouped in a composite group (CG). A CG can include devices spanned across multiple storage subsystems. They are also used for managing the SRDF relationships. You can enable particular functions at the CG level such as using consistency to preserve the dependent writes on the remote storage for a group of SRDF device pairs.

To create the first SRDF replicated resource with primary site A associated with the composite group haxdcg_siteA:

1. Create an RDF group that contains the ports in the local and remote storage subsystems that are used for data replication. To list all defined RDF groups in your configuration, enter the **symcfg list -ra all** command. For our configuration, we allocate a new RDF group haxd1_rdfg with number 41, not used until now (Example 7-8).

   Example 7-8 Creating a dynamic RDF group

```
root@xdemca1:/> symrdf addgrp -label haxd1_rdfg -rdfg 41 -sid 0304 -dir 04C,13C
-remote_rdfg 41 -remote_sid 1983 -remote_dir 03C,14D

Execute a Dynamic RDF Addgrp operation for group
'haxd1_rdfg' on Symm: 000190100304 (y/[n]) ? y

 Successfully Added Dynamic RDF Group 'haxd1_rdfg' for Symm: 000190100304
```

2. Verify the RDF group definition by using the **symcfg list -ra all** command. You can reduce the scope of the list command to a single Symmetrix storage by using the -sid <SymmID> option.
3. Create the replication relationship between the disk pairs. First, create a file that contains the disk pairs. The pairs are delimited by blank spaces (Example 7-9).

   Example 7-9 File contains the SRDF disks pairs

```
root@xdemca1:/>cat /tmp/sitea_disks.txt
0F53 00BF
0F54 00C0
```

4. Establish the SRDF relationship with synchronous operating mode by using the file that was previously created as input (Example 7-10).

   Example 7-10 Creating the SRDF relationships

```
root@xdemca1:/>symrdf -file /tmp/sitea_disks.txt createpair -type R1 -sid 0304
-rdfg 41 -establish -rdf_mode sync

Execute an RDF 'Create Pair' operation for device file
'/tmp/sitea_disks.txt' (y/[n]) ? y

An RDF 'Create Pair' operation execution is in progress for device
file '/tmp/sitea_disks.txt'. Please wait...

    Create RDF Pair in (0304,041)...................................Started.
    Create RDF Pair in (0304,041)...................................Done.
    Mark target device(s) in (0304,041) for full copy from source....Started.
    Devices: 0F53-0F54 in (0304,041)............................... Marked.
    Mark target device(s) in (0304,041) for full copy from source....Done.
    Merge track tables between source and target in (0304,041).......Started.
    Devices: 0F53-0F54 in (0304,041)............................... Merged.
    Merge track tables between source and target in (0304,041).......Done.
    Resume RDF link(s) for device(s) in (0304,041)..................Started.
    Resume RDF link(s) for device(s) in (0304,041)..................Done.

The RDF 'Create Pair' operation successfully executed for device
file '/tmp/sitea_disks.txt'.
```

5. Verify the status of the synchronization process for the SRDF pairs by using the **verify** command:

```
symrdf -file /tmp/sitea_disks.txt -sid 0304 -rdfg 41 verify -synchronized -i 10
```

6. Using the interval flag (**-i**), continue to run the command at the specified interval (seconds) until all the devices in the file list are synchronized.

7. Check the status of the SRDF pairs by using the **symrdf list pd** command. Example 7-11 shows the status of all defined SRDF relationships that contain any of the disks that are mapped on host xdemca1 by running the specified command. You can see in the example that the devices on the storage at site A are type R1, corresponding to their source roles in the SRDF/S relationships.

Example 7-11 List of the SRDF relationships (site A view)

```
root@xdemca1:/>symrdf list pd

Symmetrix ID: 000190100304

                            Local Device View
-------------------------------------------------------------------------
                   STATUS     MODES                     RDF  S T A T E S
Sym        RDF     ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

0F53 00BF   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
0F54 00C0   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized

Total                                -------- --------
  Track(s)                                  0        0
  MB(s)                                   0.0      0.0


Legend for MODES:

 M(ode of Operation)   : A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)              : X = Enabled, . = Disabled
 A(daptive Copy)       : D = Disk Mode, W = WP Mode, . = ACp off
 (Mirror) T(ype)       : 1 = R1, 2 = R2
 (Consistency) E(xempt): X = Enabled, . = Disabled, M = Mixed, - = N/A
```

8. For a similar view, run the same command on a node at site B (Example 7-12). You can see in the output example that the type of the devices on the storage at site B is R2, corresponding to their target states in the SRDF/S relationships.

Example 7-12 List of the SRDF relationships (site B view)

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
-------------------------------------------------------------------------
                   STATUS     MODES                     RDF  S T A T E S
Sym        RDF     ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
```

```
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C0 0F54   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized

Total                                -------- --------
  Track(s)                                  0        0
  MB(s)                                   0.0      0.0


Legend for MODES:

 M(ode of Operation)   : A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)              : X = Enabled, . = Disabled
 A(daptive Copy)       : D = Disk Mode, W = WP Mode, . = ACp off
 (Mirror) T(ype)       : 1 = R1, 2 = R2
 (Consistency) E(xempt): X = Enabled, . = Disabled, M = Mixed, - = N/A
```

Tip: You can use the **symrdf -sid <SymmID> -rdfg <rdfg_ID> list** command to list the status of the SRDF pairs that are part of a specific RDF(RA) group.

9. Create the device group definition on a node. In Example 7-13, we create the device group haxd_siteA.

Example 7-13 Defining the device group haxd_siteA on node xdemca1

```
root@xdemca1:/>symdg create haxd_siteA -type RDF1
root@xdemca1:/>symdg list

                        D E V I C E      G R O U P S

                                                          Number of
  Name               Type     Valid  Symmetrix ID  Devs GKs  BCVs  VDEVs  TGTs

  haxd_siteA         RDF1     N/A    N/A            0     0     0      0     0
```

10. Add the volumes on the storage at site A to the device group definition (Example 7-14).

Example 7-14 Associating volumes to the device group haxd_siteA

```
root@xdemca1:/>symld -sid 000190100304 -g haxd_siteA addall -RANGE 0F53:0F54
root@xdemca1:/>symdg list

                        D E V I C E      G R O U P S

                                                          Number of
  Name               Type     Valid  Symmetrix ID  Devs GKs  BCVs  VDEVs  TGTs

  haxd_siteA         RDF1     Yes    000190100304     2  0     0      0     0
```

11. Export the device definition from the xdemca1 node and import the definitions on all nodes.

The sycfg sync command: The storage device roles change when the SRDF relationships are created (from REGULAR to RDF1or RDF2). Therefore, run the **symcfg sync** command before the **import** command on all cluster nodes other than the initiating node of the SRDF relations to update the local database with the current device status.

12. To import the definition file on a node other than the one that was created, copy the file on the target node. We create two files:
    - The `Import` file for the nodes in the same site

      In our environment, we use this file to import the definition of the device group on the xdemca2 node. Example 7-15 shows how we use export/import DG commands in our environment.

      Example 7-15 Exporting and importing the DG definitions for a node in the same site

```
On node xdemca1:
root@xdemca1:/>symdg exportall -f /tmp/dg_local.cfg
root@xdemca1:/>cat /tmp/dg_local.cfg
<haxd_siteA>
1 000190100304
S F53 DEV001
S F54 DEV002

We copy the /tmp/dg_local.cfg on node xdemca2 in the same directory and run
the import command
On xdemca2:
root@xdemca2:/>symdg importall -f /tmp/dg_local.cfg

Creating device group 'haxd_siteA'
Adding STD device F53 as DEV001...
Adding STD device F54 as DEV002...
```

    - The `Import` file for the nodes in the remote site

      In Example 7-16, we perform the export/import operations by using the xdemcb1 node at the remote site as the import target. The same import operation command applies for the xdemcb2 node. The **export** command uses the -rdf flag to generate a file that contains the corresponding R2 devices on the storage in the remote site.

      Example 7-16 Exporting and importing the DG definitions for a node in the remote site

```
On node xdemca1:

root@xdemca1:/>symdg exportall -f /tmp/dg_remote.cfg -rdf
root@xdemca1:/>cat /tmp/dg_remote.cfg
<haxd_siteA>
2 000190101983
S 0BF DEV001
S 0C0 DEV002

Copy the file /tmp/dg_remote.cfg on the target node xdemcb1 and import de DG
definition on the node.

On node xdemcb1:

root@xdemcb1:/>cat /tmp/dg_remote.cfg
```

```
<haxd_siteA>
2 000190101983
S OBF DEV001
S OCO DEV002

root@xdemcb1:/>symdg importall -f /tmp/dg_remote.cfg

Creating device group 'haxd_siteA'
Adding STD device OBF as DEV001...
Adding STD device OCO as DEV002...
```

13. Verify the DG definition on a node by using the **symdg list** or **symdg show <DG name>** commands. See an output example of the **symdg show** command in 7.8.1, “SYMCLI commands for SRDF environment” on page 330.

14. Add the SRDF replicated resource definition in the cluster configuration. The cluster script now creates a CG and synchronizes the definition across all cluster nodes. To succeed, you must have propagated the device group definition on all cluster nodes.

    To create the PowerHA SRDF replicated resource, run the **smitty hacmp** command. Then, select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resources Configuration** → **Configure EMC SRDF Replicated Resources** → **Add EMC SRDF Replicated Resource** (Figure 7-13).

```
                    Add an EMC SRDF Replicated Resource

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

                                                      [Entry Fields]
* EMC SRDF Composite Group Name                     [haxdcg_siteA]        +
* EMC SRDF Mode                                     SYNC                  +
  Device Groups                                      haxd_siteA            +
* Recovery Action                                    AUTO                  +
* Consistency Enabled                                YES                   +
```

Figure 7-13 Adding an SRDF replicate resource

15. In the Add an EMC SRDF Replicated Resource panel, define the following options in the menu:

    – EMC SRDF Composite Group Name

      Specify a name for the composite group to be defined. In our scenario, we use `haxdcg_siteA` for the CG name that is associated with the resource group with the primary site A.

    – EMC SRDF Mode

      Choose between `SYNC` and `ASYNC`, depending on the SRDF type of relationship that you created. In our case, we set this parameter to `SYNC`, which corresponds with the synchronous relationships that were already defined in Example 7-8 on page 285.

    – Device Groups

      Specify the device groups that are associated with the composite group definition. In our case, we include `DG haxd_siteA`, which were already defined on all nodes of the cluster.

– Recovery Action

  Specify whether the resource will be automatically processed or manually processed if a failover condition across sites applies. Consider that the MANUAL option for processing the replicated resource is also correlated with the state of the SRDF relations at the time of failover of the resource group that contains the replicated resource to the secondary site. For more information about auto versus manual option settings, see 5.5.5, “Lost of the replication links (auto versus manual)” on page 214.

– Consistency Enabled

  Specify whether consistency will be applied for the relationships in the composite group. Use of this option requires an SRDF/CG license feature that is applied on all cluster nodes. When you enable consistency, the SRDF/CG feature prevents a dependent-write I/O operation from reaching the secondary site if a previous I/O write was not completed on both sides. For more information about the SRDF/CG feature, see *EMC Symmetrix Remote Data Facility (SRDF) - Product Guide*, P/N 300-001-165. In our scenario, we enable the consistency for the current composite group.

SRDF: When you define the SRDF replicated resource, the cluster script verifies the SRDF environment and propagates the CG definitions across the cluster nodes before running the cluster synchronization. You must run the cluster verification and synchronization task to synchronize the SRDF cluster resource definition across the nodes.

Later, we add the second replicated resource to the cluster configuration on top of the existing configuration.

7.4.5 Importing the volume group definition in the remote site

For our environment, the volume group configuration that is defined on nodes at site A has an already defined enhanced concurrent volume group named srdf_s_vg containing the /data1 file system (Example 7-17).

Example 7-17 Displaying the srdf_s_vg concurrent volume group

```
root@xdemca1:/>lsvg -l srdf_s_vg
srdf_s_vg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
srdf_s_loglv        jfs2log    1       1       1    open/syncd    N/A
srdf_s_lv           jfs2       1000    1000    2    open/syncd    /data1
```

Now, the definition of the volume group exists only on the nodes at site A, which are the xdemca1 node and the xdemca2 node. To integrate the volume group in the cluster configuration, import the volume group on the nodes at the remote site. To import the volume group, split the SRDF relationships. We perform this task on the CG level (Example 7-18).

Example 7-18 Splitting the SRDF relations

```
root@xdemca1:/>symrdf -cg haxdcg_siteA split -force

Execute an RDF 'Split' operation for composite
group 'haxdcg_siteA' (y/[n]) ? y

An RDF 'Split' operation execution is
in progress for composite group 'haxdcg_siteA'. Please wait...
```

```
    Suspend RDF link(s) for device(s) in (0304,041).................Done.
    Read/Write Enable device(s) in (0304,041) on RA at target (R2)...Done.

The RDF 'Split' operation successfully executed for
composite group 'haxdcg_siteA'.
```

In the next step, we import the srdf_s_vg volume group from the disks that are defined on the storage at site B. Ensure that the disks have the PVIDs available on the nodes at site B before you run the **importvg** command. You can activate the PVID of a disk by running the **chdev -l <hdisk#> -a pv=yes** command. Verify that the PVIDs on the secondary disks match the existing PVIDs of the disks at the primary site. Example 7-19 provides an example of importing the volume group on xdemcb2 node at site B.

Example 7-19 Importing the volume group on a remote node

```
root@xdemcb2:/>lspv
.........

hdiskpower42    none                                None
hdiskpower43    none                                None

root@xdemcb2:/> chdev -l hdiskpower42 -a pv=yes
root@xdemcb2:/>chdev -l hdiskpower43 -a pv=yes


root@xdemcb2:/>lspv
.........
hdiskpower42    00cfb52d62712021                    None
hdiskpower43    00cfb52d6278a02c                    None
.........
root@xdemcb2:/>importvg -y srdf_s_vg hdiskpower42
srdf_s_vg
0516-783 importvg: This imported volume group is concurrent capable.
        Therefore, the volume group must be varied on manually.
```

Change the SRDF relationships back to their normal status of replication by running the **symrdf establish** operation on the haxdcg_siteA composite group (Example 7-20).

Example 7-20 Resuming the SRDF relationships in CG haxdcg_siteA

```
root@xdemca1:/>symrdf -cg haxdcg_siteA establish

Execute an RDF 'Incremental Establish' operation for composite
group 'haxdcg_siteA' (y/[n]) ? y

An RDF 'Incremental Establish' operation execution is
in progress for composite group 'haxdcg_siteA'. Please wait...

    Write Disable device(s) in (0304,041) on RA at target (R2).......Done.
    Suspend RDF link(s) for device(s) in (0304,041).................Done.
    Resume RDF link(s) for device(s) in (0304,041)..................Started.
    Merge track tables between source and target in (0304,041).......Started.
    Devices: 0F53-0F54 in (0304,041)............................... Merged.
    Merge track tables between source and target in (0304,041).......Done.
    Resume RDF link(s) for device(s) in (0304,041).................Done.
```

```
The RDF 'Incremental Establish' operation successfully initiated for
composite group 'haxdcg_siteA'.
```

7.4.6 Defining the cluster resources

Figure 7-14 illustrates our final cluster resource group configuration. It contains two resource groups:

- RG_sitea, with the inter-site policy: prefer primary site A
- RG_siteb, with the inter-site policy: prefer primary site B

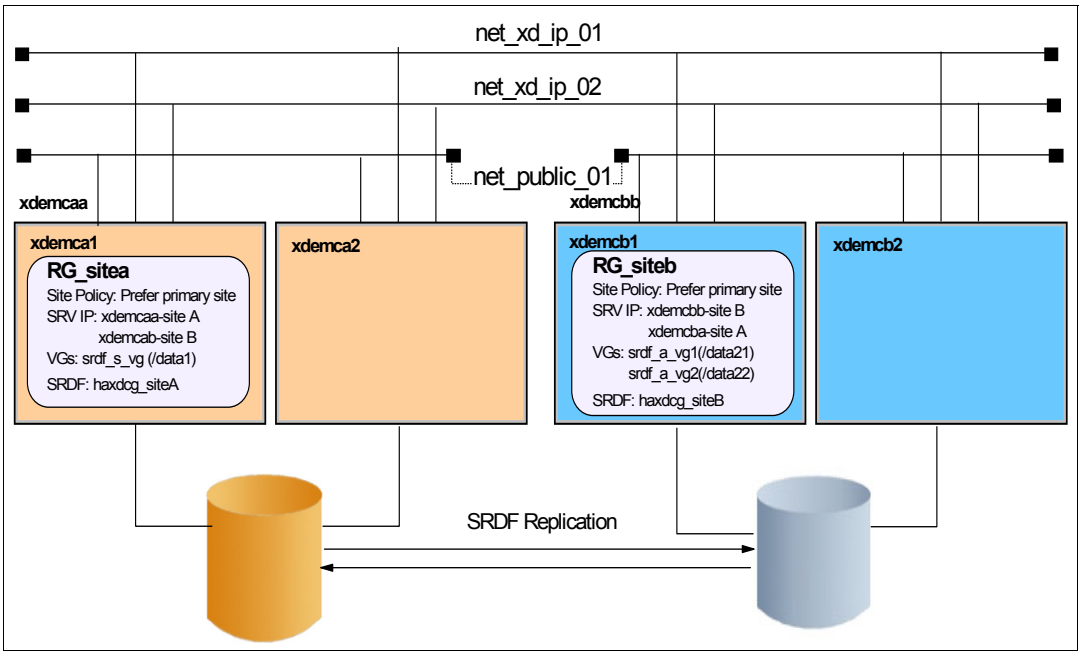


Figure 7-14 Resource group configuration

Each resource group contains:

- The service IP labels specific to each site
- Volume groups that contain jfs2 file systems
- The associated SRDF replicated resource that contains the device pairs that are already described in Figure 7-12 on page 283

First, define the RG_sitea resource group, and later we add the second resource group RG_siteb on top of the existing configuration.

Define the following resources at site A, part of the RG_sitea resource group:

- The service IP labels

  The resource group contains two service IP labels:

  - xdemcaa_sv, used when the resource group is active at site A
  - xdemcab_sv, used when the resource group is active at site B

Figure 7-15 shows an example of defining the service IP label configurable on multiple nodes that are used at site A.

```
 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.

                                                      [Entry Fields]
* IP Label/Address                                    xdemcaa_sv           +
  Netmask(IPv4)/Prefix Length(IPv6)                   []
* Network Name                                        net_public_01
  Alternate HW Address to accompany IP Label/Address  []
  Associated Site                                     siteA                +
```

Figure 7-15 Defining a service IP label site specific

A similar operation is performed to define the service IP label *xdemcab* specific to site B.

- The srdf_s_vg volume group and the associated /data1 file system

  These resources are added in the resource group definition after the RG_sitea resource group is created.

- The SRDF replicated resource, previously created in Figure 7-13 on page 289

Create the RG_sitea resource group with the primary site A with the default values of the failover/failback policy parameters (Figure 7-16).

```
                        Add a Resource Group (extended)

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

                                                   [Entry Fields]
* Resource Group Name                             [RG_sitea]

  Inter-Site Management Policy                    [Prefer Primary Site]    +
* Participating Nodes from Primary Site           [xdemca1 xdemca2]        +
  Participating Nodes from Secondary Site         [xdemcb1 xdemcb2]        +

  Startup Policy               Online On Home Node Only                    +
  Failover Policy              Failover To Next Priority Node In The List  +
  Failback Policy              Failback To Higher Priority Node In The List +
```

Figure 7-16 Creating the resource group RG_sitea

Add the resources to the resource group definition. Figure 7-17 shows the SRDF replicated resource option in the configuration menu.

```
        Change/Show All Resources and Attributes for a Resource Group

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

                                                      [Entry Fields]
  Resource Group Name                                 RG_sitea
  Inter-site Management Policy                        Prefer Primary Site
  Participating Nodes from Primary Site               xdemca1 xdemca2
  Participating Nodes from Secondary Site             xdemcb1 xdemcb2

  Startup Policy                   Online On Home Node Only
  Failover Policy                  Failover To Next Priority Node In The List
  Failback Policy                  Failback To Higher Priority Node In The List
  Failback Timer Policy (empty is immediate)          []                        +

  Service IP Labels/Addresses                         [xdemcaa_sv xdemcab_sv]   +
  Application Servers                                 []                        +

  Volume Groups                                       [srdf_s_vg ]              +
  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                                      []                        +

  Fast Connect Services                               []                        +
  Communication Links                                 []                        +

  Primary Workload Manager Class                      []                        +
  Secondary Workload Manager Class                    []                        +

  Miscellaneous Data                                  []
  WPAR Name                                           []                        +
  EMC SRDF Replicated Resources                       [haxdcg_siteA]            +
```

Figure 7-17 Adding the SRDF replicated resource to the resource group

Verify and synchronize the cluster definition across all nodes.

Verification and synchronization: During the cluster verification and synchronization, we observed that the verification scripts for the SRDF configuration verify the presence of the **emcpowerreset** utility that is used for breaking the SCSI reservation during the failover process. Currently this utility is included in the EMC Symmetrix disk attachment drivers (**EMC.Symmetrix.aix.rte**). The verification scripts also populate the HACMP custom disks method in case it is not already populated.

After synchronizing, the cluster configuration starts the cluster services on the nodes and checks the resource group status. Example 7-21 shows the final status of the RG_sitea after starting the cluster services and resource group acquirement in our environment.

Example 7-21 Resource group status after starting the cluster services

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
```

Verify in the clRGinfo output for the states of the resource group on both sites:

- ONLINE: It is associated with the active state of the resource group on a node at a site. That node activates the resources in the resource group and starts any defined application servers. Example 7-22 shows the status of the IP addresses and file systems on the xdemca1 node, currently the node with primary online state of the RG_sitea resource group.

  Example 7-22 Status of the resources on node xdemca1

```
IP Addresses:

root@xdemca1:/>netstat -i
Name  Mtu   Network     Address           ZoneID Ipkts Ierrs Opkts Oerrs  Coll
en0   1500  link#2      0.9.6b.dd.e.e4       -  5081578   0  1986826   3  0
en0   1500  10.100.101  xdemca1_boot         -  5081578   0  1986826   3  0
en0   1500  10.10.10    xdemcaa_sv           -  5081578   0  1986826   3  0
en1   1500  link#3      0.9.6b.dd.e.e5       -  953097    0  947163    3  0
en1   1500  207.18.1    xdemca1_xdip1        -  953097    0  947163    3  0
en2   1500  link#4      0.11.25.cd.36.60     -  396595    0  380680    3  0
en2   1500  207.18.2    xdemca1_xdip2        -  396595    0  380680    3  0
lo0   16896 link#1                          -  3238556   0  3240273   0  0
lo0   16896 127         loopback             -  3238556   0  3240273   0  0
lo0   16896 ::1                            0  3238556   0  32402730  0  0

Volume group status:

root@xdemca1:/>lsvg srdf_s_vg
VOLUME GROUP:       srdf_s_vg               VG IDENTIFIER:
00cfb52d00004c000000012743c2926b
VG STATE:           active                  PP SIZE:        8 megabyte(s)
VG PERMISSION:      read/write              TOTAL PPs:      1078 (8624
megabytes)
MAX LVs:            256                     FREE PPs:       77 (616 megabytes)
```

```
LVs:                2                       USED PPs:       1001 (8008
megabytes)
OPEN LVs:           2                       QUORUM:         2 (Enabled)
TOTAL PVs:          2                       VG DESCRIPTORS: 3
STALE PVs:          0                       STALE PPs:      0
ACTIVE PVs:         2                       AUTO ON:        no
Concurrent:         Enhanced-Capable        Auto-Concurrent: Disabled
VG Mode:            Concurrent
Node ID:            1                       Active Nodes:       2
MAX PPs per VG:     32512
MAX PPs per PV:     1016                    MAX PVs:        32
LTG size (Dynamic): 256 kilobyte(s)         AUTO SYNC:      no
HOT SPARE:          no                      BB POLICY:      relocatable

Check the file system /data1 is mounted:

root@xdemca1:/>df -m
Filesystem    MB blocks      Free %Used    Iused %Iused Mounted on
/dev/hd4         384.00     15.86   96%    15098    77% /
/dev/hd2        3968.00    619.90   85%    52355    26% /usr
/dev/hd9var      640.00    109.86   83%     7712    23% /var
/dev/hd3        3072.00   2907.02    6%      881     1% /tmp
/dev/hd1          64.00     63.65    1%        5     1% /home
/dev/hd11admin   128.00    127.64    1%        5     1% /admin
/proc               -         -    -         -     -  /proc
/dev/hd10opt    1536.00    954.80   38%    15768     7% /opt
/dev/livedump    256.00    249.28    3%       26     1% /var/adm/ras/livedump
/dev/srdf_s_lv   8000.00   7848.32    2%        7     1% /data1
```

- ONLINE SECONDARY: This state is associated with the resource group status at the secondary site. When the cluster services are started on the secondary nodes, the resource group is acquired in this state, indicating that the remote site is up and ready to acquire the resource group in a failover case.

Compare the resource group status with the actual status of the SRDF relationships by running the **symrdf list pd** command (Example 7-23).

Example 7-23 Status of the SRDF relationships

```
root@xdemca1:/>symrdf list pd

Symmetrix ID: 000190100304

                            Local Device View
---------------------------------------------------------------------------
                    STATUS     MODES                      RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

0F53 00BF   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
0F54 00C0   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
...........
```

In the output of Example 7-23 observe that during normal operation, the RG_sitea is acquired at site A. The SRDF status of the pairs shows the volumes at site A as primary volumes (R1 type), while a view from site B shows the volumes at site B as secondary (R2 type).

7.4.7 Adding the second resource group to the existing configuration

For our scenario, we use the environment that was already created that contains an active resource group RG_sitea at site A, and we dynamically add a second resource group RG_siteb with primary site B. The operation that is described in this section can also be considered for a scenario in which you add a second resource group at the same primary site.

During this setup, we show the relevant steps. For more details, also verify the steps that are performed when you define the RG_sitea resource group.

With the cluster topology configuration in place, complete the following steps. Unless specified, all the commands are run at site B, on the xdemcb1 node.

1. Define the SRDF environment for the SRDF/A relations:
   a. Create an RDF group haxd2_rdfg(42) on both storage subsystems:

      ```
      symrdf addgrp -label haxd2_rdfg -rdfg 42 -sid 1983 -dir 03C,14D -remote_rdfg
      42 -remote_sid 0304 -remote_dir 04C,13C
      ```

   b. Establish the SRDF/A relationships. We put the disks pairs for the site B resource group (Table 7-2 on page 284) in the `/tmp/srdfb_disks.txt` file:

      ```
      symrdf -file /tmp/disks_siteb.txt createpair -type R1 -sid 1983 -rdfg 42
      -establish -rdf_mode async
      ```

   c. Check the SRDF relationship status for all the pairs that are attached to the cluster nodes defined so far. Example 7-24 shows the status of both SRDF/S and SRDF/A relationships after the synchronization process finishes. Note the difference between the final states of the relationships for the sync and async cases.

Example 7-24 Verifying the status of the SRDF relationships

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                                Local Device View
-------------------------------------------------------------------------------
                     STATUS     MODES                      RDF  S T A T E S
Sym        RDF       ---------  -----  R1 Inv   R2 Inv -----------------------
Dev  RDev  Typ:G     SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C0 0F54   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C1 0F55   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C2 0F56   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C3 0F57   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
........
```

d. Define the new device groups for the resource group at site B and add the volumes to the device groups.

   - Group haxd1_siteB:

     ```
     symdg create haxd1_siteB -type RDF1
     symld -sid 1983 -g haxd1_siteB addall -RANGE 00C1:00C2
     ```

   - Group haxd2_siteB:

     ```
     symdg create haxd2_siteB -type RDF1
     symld -sid 1983 -g haxd2_siteB add dev 00C3
     ```

e. Propagate the DG definitions across the cluster nodes. Since the SRDF configuration already contains the haxd_siteA device group, we use the **export** and **import** commands that are applied only to the newly created device groups:

   - For the nodes on the same site, on xdemcb1 node export the DGs definitions:

     ```
     symdg export haxd1_siteB -f /tmp/haxd1_siteB.cfg
     symdg export haxd2_siteB -f /tmp/haxd2_siteB.cfg
     ```

     Copy the files on the xdemcb2 node and run on the target node:

     ```
     symcfg sync
     symdg import haxd1_siteB -f /tmp/haxd1_siteB.cfg
     symdg import haxd2_siteB -f /tmp/haxd2_siteB.cfg
     ```

   - For the nodes in the remote site (xdemca1 and xdemca2): On node xdemcb1 export the DGs definitions for the remote site by using the **-rdf** option:

     ```
     symdg export haxd1_siteB -f /tmp/haxd1_siteB_remote.cfg -rdf
     symdg export haxd2_siteB -f /tmp/haxd2_siteB_remote.cfg -rdf
     ```

     Copy the files on the remote nodes xdemca1 and xdemca2 and run on each of them:

     ```
     symcfg sync
     symdg import haxd1_siteB -f /tmp/haxd1_siteB_remote.cfg
     symdg import haxd2_siteB -f /tmp/haxd2_siteB_remote.cfg
     ```

An alternative method is available for propagating the device group definitions across the local and remote nodes. That is, run the device group creation and the disk association commands that are used on the initiating node (in our case xdemcb1) on each node of the cluster by using the corresponding devices at each site. For more information about the device group and composite group operations, see *EMC Solutions Enabler Symmetrix SRDF Family CLI - Product Guide,* P/N 300-000-877.

2. Define the SRDF replicated resource in the PowerHA Enterprise Edition configuration. Now, we create the haxdcg_siteB composite group that contains both device groups haxd1_siteB and haxd2_siteB using the SMIT menus. On the xdemcb1 node, we access the SRDF replicated resource main menu by using the **smitty cl_srdf_def** fast path command and choose the **Add EMC SRDF Replicated Resource** option (Figure 7-18).

```
                    Add an EMC SRDF Replicated Resource

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                [Entry Fields]
* EMC SRDF Composite Group Name               [haxdcg_siteB]               +
* EMC SRDF Mode                               ASYNC                        +
  Device Groups                                haxd1_siteB haxd2_siteB      +
* Recovery Action                              MANUAL                       +
* Consistency Enabled                          YES                          +
```

Figure 7-18 Adding an SRDF replicated resource

The SRDF replicated resource that is associated with the resource group of site B has the following specific settings:

- EMC SRDF Composite Group Name: haxdcg_siteB

  This parameter represents the name of the composite group that is being created.

- EMC SRDF Mode: ASYNC

  This parameter is correlated with the asynchronous mode of operation of the disk pairs in the haxdcg_siteB composite group.

- Device groups: haxd1_siteB and haxd2_siteB

  Here the device groups are specified as part of the haxdcg_siteB composite group.

- Recovery action: MANUAL

  When the resource recovery action is set to manual, if a site failover occurs, user intervention is required to manually fail over the SRDF relationships to the secondary location and make the storage volumes available to the recovery nodes.

- Consistency enabled: YES

  When consistency is enabled on the composite group, the data on the remote volumes (part of the CG) is protected against a partial failure of the SRDF relationships. Therefore, the data consistency is ensured on the recovery site.

  Consistency: Consistency must be enabled when you use the ASYNC mode of operation.

3. Define the service IP labels for the resource group of site B. In our environment we use two service labels on the existing net_public_01: xdemcbb cluster network when the resource group is ONLINE at site B, and xdemcba when the resource group is ONLINE at site A. For defining the service labels, we use the same SMIT menu as when we defined the service IP labels for RG_sitea (Figure 7-15 on page 293).

4. Import the volume group definitions of site B on the nodes at site A. At site B, we defined two standard volume groups, srdf_a_vg1 and srdf_a_vg2, each containing a `jfs2` file system (Example 7-25).

Example 7-25 Volume group configuration on the nodes at site B

```
root@xdemcb1:/>lsvg -l srdf_a_vg1
srdf_a_vg1:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
srdfa1loglv         jfs2log    1       1       1    open/syncd    N/A
srdfa1lv            jfs2       1000    1000    2    open/syncd    /data21
root@xdemcb1:/>lsvg -l srdf_a_vg2
srdf_a_vg2:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
srdfa2loglv         jfs2log    1       1       1    open/syncd    N/A
srdfa2lv            jfs2       500     500     1    open/syncd    /data22
```

For importing the volume group definitions on the remote site, we use the same method that is described in 7.4.5, “Importing the volume group definition in the remote site” on page 290.

To split the relationships, enter:

```
symrdf -cg haxdcg_siteB split -force -nop
```

Import the volume groups on the target nodes at site A (see also the hdiskpower# mapping to the volume groups in the shaded box on page 284):

```
importvg -y srdf_a_vg1 hdiskpower3
importvg -y srdf_a_vg2 hdiskpower5
```

Re-establish the relationships:

```
symrdf -cg haxdcg_siteB establish -nop
```

The **-nop** option (no prompt) is used in this case to skip the confirmation prompt when the command is run.

5. Add the resource group definition in the cluster configuration. For site B, we use the default inter-site policy: prefer primary site (Figure 7-19).

```
                    Add a Resource Group (extended)

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

                                                [Entry Fields]
* Resource Group Name                           [RG_siteb]

  Inter-Site Management Policy                  [ignore]                  +
* Participating Nodes from Primary Site         [xdemcb1 xdemcb2]         +
  Participating Nodes from Secondary Site       [xdemca1 xdemca2]         +

  Startup Policy              Online On Home Node Only                    +
  Failover Policy             Failover To Next Priority Node In The List  +
  Failback Policy             Failback To Higher Priority Node In The List +
```

Figure 7-19 Adding the resource group RG_siteb

6. Change the attributes for the resource group to include the IP addresses, volume groups, and SRDF replicated resources (Figure 7-20).

```
       Change/Show All Resources and Attributes for a Resource Group

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

                                                      [Entry Fields]
  Resource Group Name                                 RG_siteb
  Inter-site Management Policy                        Prefer Primary Site
  Participating Nodes from Primary Site               xdemcb1 xdemcb2
  Participating Nodes from Secondary Site             xdemca1 xdemca2

  Startup Policy               Online On Home Node Only
  Failover Policy              Failover To Next Priority Node In The List
  Failback Policy              Failback To Higher Priority Node In The List
  Failback Timer Policy (empty is immediate)          []                   +

  Service IP Labels/Addresses                        [xdemcbb_sv xdemcba_sv] +
  Application Servers                                []                     +

  Volume Groups                                      [srdf_a_vg1 srdf_a_vg2 ]+
  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                                      []                    +

  Fast Connect Services                               []                    +
  Communication Links                                 []                    +

  Primary Workload Manager Class                      []                    +
  Secondary Workload Manager Class                    []                    +

  Miscellaneous Data                                  []
  WPAR Name                                           []                    +
  EMC SRDF Replicated Resources                      [haxdcg_siteB]       +
```

Figure 7-20 Adding the resources to the RG_siteb resource group

7. Synchronize the cluster configuration while the cluster services are started on all nodes. We check the resource groups status by using **clRGinfo**. See the final state of the resource groups in the cluster (Example 7-26).

Example 7-26 Resource group status

```
root@xdemca2:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

You can compare the status of the resource group with the status of the SRDF relationships by querying the actual state of the SRDF relationships. We use the **symrdf list pd** command to show all the existing relationships for the volumes that are attached on a host (Example 7-27).

Example 7-27 Listing the SRDF relations for the volumes that are attached on a host

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
---------------------------------------------------------------------------
                     STATUS     MODES                     RDF  S T A T E S
Sym        RDF       ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G     SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41    RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C0 0F54   R2:41    RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C1 0F55   R1:42    RW RW RW   A..1-       0        0 RW  WD   Consistent
00C2 0F56   R1:42    RW RW RW   A..1-       0        0 RW  WD   Consistent
00C3 0F57   R1:42    RW RW RW   A..1-       0        0 RW  WD   Consistent
.........
```

The command was issued from a node at site B. When the RG_siteb resource group is active on site B, the 00C1-00C3 devices that are part of the resource group have the R1 role (source volumes). While for the RG_sitea resource group, which is active at site A, the 00BF-00C0 volumes on storage at site B have the R2 role (target volumes).

7.5 Test scenarios

In this section, we test the PowerHA Enterprise Edition cluster with the SRDF integration features. Carefully plan the testing of the PowerHA Enterprise Edition cluster because it usually requires a longer time to be performed, longer outages, and it might affect the data

availability at both sites. In many environments, you can consider periodic site failover tests as a good practice to verify the readiness for a real disaster case.

Beyond the local high-availability configuration, PowerHA Enterprise Edition scenarios involve new factors:

- Sites and the node associations with sites
- IP and non-IP communication links between sites
- Data replication between storage subsystems

For our scenario, we do not perform local HA tests or redundancy tests, such as bringing down one of the two XD_ip networks because we focus on an extended distance scenario.

Losing the XD_ip communication links can be considered a particular failure case. Configure redundant IP/non-IP communication paths to avoid isolation of the sites. Losing all the communication paths between sites leads to a partitioned state of the cluster and to data divergence between sites if the replication links are also unavailable. For considerations related to the partitioned cluster case, see 9.3, “Partitioned cluster considerations” on page 457.

Another particular case is the loss of the replication paths between the storage subsystems while the cluster is running on both sites. To avoid this situation, configure redundant communication links for the SRDF replication. You must manually recover the status of the SRDF pairs after the storage links are operational.

Important: The PowerHA software does not monitor the SRDF link status. If the SRDF link goes down when the cluster is up, and later the link is repaired, you must manually resynchronize the pairs.

In the following sections, we describe the following test cases:

- Graceful failover scenario
- Total site failure in two scenarios:
  - Fully automated
  - With manual intervention
- Loss of storage access in a site

For all the tests in our environment, we start from the same initial state. Each resource group is active in the associated primary site, RG_sitea at site A and RG_siteb at site B (Example 7-28).

Example 7-28 Resource group states on each node

```
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

7.5.1 Graceful site failover

A graceful site failover operation is performed during a planned outage at the primary site. We perform the graceful failover operation between sites by running an RG move operation against the RG_sitea resource group in primary site A. The cluster manager performs the following operations:

- Release the primary online instance of RG_sitea at site A.
- Release the secondary online instance of RG_sitea at site B.
- Acquire RG_sitea in the secondary online state at site A.
- Acquire RG_sitea in the online primary state at site B.

We run the RG move command by using SMIT menus. Run the **smitty hacmp** command. Then, select **System Management (C-SPOC)** → **Resource Group and Applications** → **Move a Resource Group to Another Node / Site** → **Move Resource Groups to Another Site**. Select the ONLINE instance of the RG_sitea to be moved to the remote site.

Figure 7-21 shows the SMIT panel for the move operation.

```
 Move Resource Group(s) to Another Site

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

                                                      [Entry Fields]
  Resource Group(s) to be Moved                      RG_sitea
  Destination Site                                   siteB
```

Figure 7-21 RG move to another site using SMIT menus

The resource group RG_sitea is moved to site B, and Example 7-29 shows the final status.

Example 7-29 Final state of the resource group in the cluster

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               OFFLINE                        xdemcb1@siteB
               ONLINE                         xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

In Example 7-30, you can see the status of the SRDF relationships from both sites after the failover operation is completed.

Example 7-30 Status of SRDF relations after resource group movement

```
Site A view:

                              Local Device View
-------------------------------------------------------------------------
```

```
                    STATUS     MODES                     RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

0F53 00BF   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
0F54 00C0   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
…..

Site B view:

                            Local Device View
------------------------------------------------------------------------------
                    STATUS     MODES                     RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
00C0 0F54   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
....
```

In the output in Example 7-30, the R1/R2 roles of the volumes are swapped after the resource group movement in comparison to their initial value. The replication is established now from site B to site A. In this SRDF state, volumes at site B (R1 type) can be used for read/write operations. Their corresponding targets (R2 type) at site A cannot be updated (WD state - write denied) and can be opened only for read operations.

To fail back the resource group to the original site A, perform the same RG move operation of the RG_sitea resource group, now active at site B to site A.

7.5.2 Total site failure

The following scenarios are tested:

- A fully automated site failover by using the RG_sitea resource group that contains an SRDF/S replicated resource with the AUTO option
- A manual failover by using the RG_siteb resource group that contains an SRDF/A replicated resource with the MANUAL option

Fully automated scenario

A site failure condition is triggered by the cluster software when there are no available nodes to maintain the resource group active in the primary site.

To simulate a site failure in our environment for the RG_sitea resource group at site A:

1. Set the ports offline on the storage at site A that is used for SRDF communication. Run the following commands on a host at site A:

   ```
   symcfg -RA 13C -sid 0304 offline
   symcfg -RA 04C -sid 0304 offline
   ```

2. Halt the xdemca1and xdemca2 nodes on site A by using the **`halt -q`** command.

The cluster detects that no available nodes are at the primary site for activating RG_sitea. Therefore, a failover process takes place. At the end, the resource group is acquired on site B

on the highest priority node in the list (xdemcb1). Example 7-31 shows the **clRGinfo** command output. As shown in Example 7-31, there is no ONLINE SECONDARY state for the RG_sitea and RG_siteb resource groups because there are no available nodes at site A.

Example 7-31 Status of the resource group after failover

```
root@xdemcb1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                        xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               OFFLINE                        xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

Example 7-32 shows the SRDF relationship status after the failover.

Example 7-32 SRDF relationship status after site failover event

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
---------------------------------------------------------------------------
                     STATUS     MODES                     RDF  S T A T E S
Sym        RDF       ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G     SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41    RW RW NR   S..2-   13474        0 RW  NA   Partitioned
00C0 0F54   R2:41    RW RW NR   S..2-   66574        0 RW  NA   Partitioned
00C1 0F55   R1:42    RW RW RW   A..1-       0        0 RW  NA   TransIdle
00C2 0F56   R1:42    RW RW RW   A..1-       0        0 RW  NA   TransIdle
00C3 0F57   R1:42    RW RW RW   A..1-       0        0 RW  NA   TransIdle

Total                                -------- --------
  Track(s)                             80048        0
  MB(s)                               5003.0      0.0


Legend for MODES:

 M(ode of Operation)   : A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)              : X = Enabled, . = Disabled
 A(daptive Copy)       : D = Disk Mode, W = WP Mode, . = ACp off
 (Mirror) T(ype)       : 1 = R1, 2 = R2
 (Consistency) E(xempt): X = Enabled, . = Disabled, M = Mixed, - = N/A
```

Because the storage connections are being lost, the status of the SRDF/S relations became *Partitioned*, whereas the status of SRDF/A relations became *TransIdle*. For volume pairs that

are associated with the RG_sitea resource group, R1 and R2 roles did not swap, but the devices at site B are now enabled for read/write operations.

Failback to the initial state after resuming the primary site

After we resume operations at site A and reestablish the RDF links, the status of the relationships changed for both SRDF/S and SRDF/A (Example 7-33). In our environment we enable the RDF ports by using the following command on the nodes at site A:

```
symcfg -RA <PortNo> -sid 0304 online
```

Example 7-33 SRDF status of pairs after the RDF links are operational

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
-------------------------------------------------------------------------
                    STATUS     MODES                     RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41   RW RW NR   S..2-   13474        0 RW  RW   Split
00C0 0F54   R2:41   RW RW NR   S..2-   66574        0 RW  RW   Split
00C1 0F55   R1:42   RW RW NR   A..1-       0        1 RW  WD   Suspended
00C2 0F56   R1:42   RW RW NR   A..1-       0    16388 RW  WD   Suspended
00C3 0F57   R1:42   RW RW NR   A..1-       0    16389 RW  WD   Suspended

Total                                -------- --------
  Track(s)                              80048    32778
  MB(s)                                5003.0   2048.6


Legend for MODES:

 M(ode of Operation)   : A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)              : X = Enabled, . = Disabled
 A(daptive Copy)       : D = Disk Mode, W = WP Mode, . = ACp off
 (Mirror) T(ype)       : 1 = R1, 2 = R2
 (Consistency) E(xempt): X = Enabled, . = Disabled, M = Mixed, - = N/A
```

In Example 7-33, observe that SRDF/S pairs changed to Split state, while the SRDF/A pairs changed to Suspended.

Important: Do not initiate the failback of the resource group after a failover until you check that the RDF links are up and verify the state of the SRDF pairs.

At this time, you need to update the R1 volumes on site A associated with the RG_sitea resource group with the R2 data on site B, modified while RG_sitea was active at site B. You need to perform this operation manually before the failback operation is initiated by the cluster. If you start the cluster failback operations before you update the R1 volumes while the SRDF pairs are in this state, you establish normal R1 → R2 replication and overwrite the R2 volumes at site B with the existing image at site A.

We perform the following operations on a node at site B to update the R1 volumes with the latest data version on R2 volumes. (For each change in the SDRF state of the CG pairs, a list status is provided by using the **symrdf list pd** command.)

1. Fail over the SRDF relationship on the haxdcg_siteA composite group (Example 7-34).

   ```
   symrdf -cg haxdcg_siteA failover -force
   ```

   Example 7-34 Failover output of the SRDF relationship

   ```
   Symmetrix ID: 000190101983

                                 Local Device View
   ---------------------------------------------------------------------------
                       STATUS     MODES                     RDF  S T A T E S
   Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
   Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
   ---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

   00BF 0F53   R2:41   RW RW NR   S..2-       3        0 RW  WD   Failed Over
   00C0 0F54   R2:41   RW RW NR   S..2-   16390        0 RW  WD   Failed Over
   00C1 0F55   R1:42   RW RW NR   A..1-       0        1 RW  WD   Suspended
   00C2 0F56   R1:42   RW RW NR   A..1-       0        0 RW  WD   Suspended
   00C3 0F57   R1:42   RW RW NR   A..1-       0        1 RW  WD   Suspended
   ```

2. Disable the consistency on the haxdcg_siteA composite group:

   ```
   symcg -cg haxdcg_siteA disable -force
   ```

3. Update the R1 volumes at site A with the image of the R2 volumes at site B (Example 7-35):

   ```
   symrdf -cg haxdcg_siteA update -force
   ```

   Example 7-35 Updating the pairs

   ```
   Symmetrix ID: 000190101983

                                 Local Device View
   ---------------------------------------------------------------------------
                       STATUS     MODES                     RDF  S T A T E S
   Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
   Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
   ---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

   00BF 0F53   R2:41   RW RW RW   S..2-       1        0 RW  WD   R1 Updated
   00C0 0F54   R2:41   RW RW RW   S..2-       0        0 RW  WD   R1 Updated
   00C1 0F55   R1:42   RW RW NR   A..1-       0        1 RW  WD   Suspended
   00C2 0F56   R1:42   RW RW NR   A..1-       0        0 RW  WD   Suspended
   00C3 0F57   R1:42   RW RW NR   A..1-       0        1 RW  WD   Suspended
   ```

4. Enable back the consistency on the haxdcg_siteA composite group:

   ```
   symcg -cg haxdcg_siteA enable
   ```

5. Initiate the failback operation of the resource group to the primary site. In our environment, the resource group policy for RG_sitea is set to Prefer Primary Site. Therefore, an automatic failback takes places after you start the cluster services on the nodes at site A.

We finally start the cluster services on the xdemca1 and xdemca2 nodes by using **smitty clstart**. Figure 7-22 shows the SMIT panel.

```
                        Start Cluster Services

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

                                                      [Entry Fields]
* Start now, on system restart or both                now                    +
  Start Cluster Services on these nodes               [xdemca1,xdemca2]      +
* Manage Resource Groups                              Automatically          +
  BROADCAST message at startup?                       false                  +
  Startup Cluster Information Daemon?                 true                   +
  Ignore verification errors?                         false                  +
  Automatically correct errors found during           Interactively          +
  cluster start?
```

Figure 7-22 Starting the cluster services on the nodes at site A

After the failback operation is completed, the cluster re-establishes the R1 → R2 relationships to their original state. The RG_sitea is acquired on the highest priority node at site A (xdemca1). Example 7-36 shows the **clRGinfo** command output.

Example 7-36 Resource group status after failback to primary site A

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

When the primary site comes back online, both RG_sitea and RG_siteb resource group states are changed. RG_sitea is primary online at site A and secondary online at site B, whereas RG_siteb, previously primary online at site B, acquires the secondary online state at site A.

Both SRDF relationships are recovered from their previous states (Example 7-36 on page 309). Their final status is Synchronized for SRDF/S pairs and Consistent for SRDF/A pairs. Example 7-37 shows the new status of the pairs that are listed from site B.

Example 7-37 SRDF pair status after starting the cluster services at site A and failback of RG_sitea

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
```

```
--------------------------------------------------------------------------------
                     STATUS     MODES                    RDF  S T A T E S
Sym        RDF       ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G     SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41    RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C0 0F54   R2:41    RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C1 0F55   R1:42    RW RW RW   A..1-       0        0 RW  WD   Consistent
00C2 0F56   R1:42    RW RW RW   A..1-       0        0 RW  WD   Consistent
00C3 0F57   R1:42    RW RW RW   A..1-       0        0 RW  WD   Consistent
.........
```

Manual failover of the SRDF resources

For this test, we use an SRDF replicated resource with the recovery action parameter set to MANUAL. In our environment, the RG_siteb resource group contains the replicated resource haxdcg_siteB with the manual option. You can check the replicated resource groups by using the **cllssr** command (Example 7-38).

Example 7-38 Output of the cllssr showing the replicated resource groups status

```
root@xdemcb1:/>/usr/es/sbin/cluster/sr/cmds/cllssr -a
SRDFCgName   SRDFMode DeviceGroups              RecoveryAction ConsistencyEnabled
haxdcg_siteA SYNC     haxd_siteA               AUTO           YES
haxdcg_siteB ASYNC    haxd1_siteB haxd2_siteB  MANUAL         YES
```

To simulate a manual failover of the SRDF resources:

1. Simulate the total site failure condition in the same way as in the fully automated scenario that was applied for the nodes and the storage at site B. See “Fully automated scenario” on page 305.

   After the cluster detects no available nodes at site B to acquire the resource group RG_siteb, a site failover event takes place. Now, the cluster tries to acquire the RG_siteb resource group at site A.

2. Use the MANUAL option is used on the SRDF replicated resource. Therefore, the cluster notifies the user for the required actions that need to be performed on the SRDF replicated resource and does not activate the RG_siteb resource group at the secondary site A. Example 7-39 shows the message that is posted in the `hacmp.out` file.

   Example 7-39 hacmp.out file message

   ```
   RECOMMENDED USER ACTIONS:

   We are at this stage because EMC SRDF links are in a state which cannot be
   automatically rectified by HACMP. HACMP will not be knowing whether the
   primary (RDF1) or the secondary (RDF2) has the latest consistent data.
   It is the responsibility of the user to ascertain for the correct data
   and synchronize the same to the mirrored storage.

   The current state  is not handled by HACMP with Manual Recovery Action.
   User should manually correct "TransIdle" state and restart HACMP.

   STEP 1:Verify the current state of EMC SRDF pairs within the Composite Group.

   STEP 2:Identify the storage(RDF1 or RDF2) that has the latest consistent data.
   ```

```
STEP 3:Synchronize the data to the corresponding mirrored storage.

STEP 4:Please consult the EMC Storage documentation for instructions in
this regard.

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 node.

                        END RECOMMENDED USER ACTIONS:
```

In our test environment, we observed that the resource group processing ended in an ERROR state. Example 7-40 shows the final status of the RG_siteb resource group.

Example 7-40 Resource group status after site B failed

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               OFFLINE                        xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       OFFLINE                        xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               ERROR                          xdemca2@siteA
```

3. Check the status of the SRDF pairs. We observed that part of the RG_siteb resource group pairs are in TransIdle state with the WD access mode (Example 7-41).

Example 7-41 SRDF pair status after site B failure (site A view)

```
root@xdemca1:/>symrdf list pd

Symmetrix ID: 000190100304

                            Local Device View
---------------------------------------------------------------------------
                    STATUS     MODES                    RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

0F53 00BF   R1:41   RW RW NR   S..1-       0        1 RW  NA   Partitioned
0F54 00C0   R1:41   RW RW NR   S..1-       0        0 RW  NA   Partitioned
0F55 00C1   R2:42   RW WD RW   A..2-       0        0 WD  NA   TransIdle
0F56 00C2   R2:42   RW WD RW   A..2-       0        0 WD  NA   TransIdle
0F57 00C3   R2:42   RW WD RW   A..2-       0        0 WD  NA   TransIdle
....
```

4. Follow the user actions that are provided in the `hacmp.out` file. Because site B failed, there are no active links between the DMX storage subsystems now. Perform a manual failover of the SRDF pairs in the composite group to activate the disks in the secondary site. Example 7-42 shows the details of performing these operations in our environment.

Example 7-42 Failover of the SRDF/A pairs on the secondary site A

```
root@xdemca1:/>symrdf -cg haxdcg_siteB failover -immediate -force

Execute an RDF 'Failover' operation for composite
group 'haxdcg_siteB' (y/[n]) ? y

An RDF 'Failover' operation execution is
in progress for composite group 'haxdcg_siteB'. Please wait...

    Read/Write Enable device(s) in (0304,042) on RA at target (R2)...Done.

The RDF 'Failover' operation successfully executed for
composite group 'haxdcg_siteB'.

====================================
The final status of the disk pairs:
====================================

root@xdemca1:/>symrdf list pd

Symmetrix ID: 000190100304

                            Local Device View
-------------------------------------------------------------------------------
                    STATUS     MODES                      RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

0F53 00BF   R1:41   RW RW NR   S..1-       0        2 RW  NA   Partitioned
0F54 00C0   R1:41   RW RW NR   S..1-       0        0 RW  NA   Partitioned
0F55 00C1   R2:42   RW RW NR   S..2-       0        0 RW  NA   Partitioned
0F56 00C2   R2:42   RW RW NR   S..2-       0        0 RW  NA   Partitioned
0F57 00C3   R2:42   RW RW NR   S..2-       0        0 RW  NA   Partitioned
```

5. After changing the state of the disks in the secondary site, bring the resource group online.
6. Run the **varyonvg** command on xdemca1 node and on the volume groups that are associated with the RG_siteb, srdf_a_vg1, and srdf_a_vg2 resource groups:

```
varyonvg srdf_a_vg1
varyonvg srdf_a_vg2
```

7. Activate the RG_siteb resource group at site A by using the SMIT menu by running the **smitty hacmp** command and selecting **System Management (C-SPOC)** → **Resource Group and Applications** → **Bring a Resource Group Online**. Select the RG_siteb target resource group that is associated with the primary online state from the list and then the target node for the resource group activation, **xdemca1**.

Figure 7-23 shows the final SMIT panel.

```
                        Bring a Resource Group Online

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

                                                        [Entry Fields]
  Resource Group to Bring Online                       RG_siteb
  Node on Which to Bring Resource Group Online         xdemca1
```

Figure 7-23 Bring online resource group RG_siteb in the secondary site A

Example 7-43 shows the final state of the resource groups.

Example 7-43 Resource group status after bringing RG_sitea online at site A

```
root@xdemca1:/var/hacmp/log>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               OFFLINE                        xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       OFFLINE                        xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

Failback to the primary site after site failover

After the primary site B of the RG_siteb resource group is back online and the RDF links are operational, we check the SRDF pairs status by running the **`symrdf list pd`** command (Example 7-44).

Example 7-44 Results of the symrdf list pd showing the SRDF pair status

```
Symmetrix ID: 000190101983

                            Local Device View
---------------------------------------------------------------------------
                    STATUS     MODES                    RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41   RW WD NR   S..2-       0        0 WD  RW   Suspended
00C0 0F54   R2:41   RW WD NR   S..2-       0        0 WD  RW   Suspended
00C1 0F55   R1:42   RW RW NR   A..1-       0        0 RW  RW   Split
00C2 0F56   R1:42   RW RW NR   A..1-       0        0 RW  RW   Split
00C3 0F57   R1:42   RW RW NR   A..1-       0        0 RW  RW   Split
```

After the RDF links became operational, the status of the pairs changed as follows:

- The SRDF/S pairs are in Suspended state. The primary online instance of the RG_sitea resource group was active at site A during the test and the pairs are re-established after the RG_sitea resource group is acquired on site B in the secondary online state.
- The SRDF/A pairs are in Split state. Read/write operations are allowed at both ends.

You can now re-establish the SRDF/A pairs, updating the storage at site B with the content of the volumes at site A (siteA → siteB replication). We run the update operation of the R1 devices (site B) from R2 (site A) and re-establish the original state of the CG SRDF pairs with the following sequence of tasks. Unless specified, the operations are run on the xdemca1 node.

1. Run the failover operation on the haxdcg_siteB composite group:

   ```
   symrdf -cg haxdcg_siteB failover -immediate -force -symforce
   ```

   Example 7-45 shows the state of the pairs after the failover command.

   Example 7-45 Pairs status after failover

   ```
   Symmetrix ID: 000190100304

                                Local Device View
   ---------------------------------------------------------------------------
                         STATUS     MODES                     RDF  S T A T E S
   Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
   Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
   ---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

   0F53 00BF   R1:41   RW RW RW   S..1-       0        0 RW  WD   Suspended
   0F54 00C0   R1:41   RW RW RW   S..1-       0        0 RW  WD   Suspended
   0F55 00C1   R2:42   RW RW NR   A..2-       3        0 RW  WD   Failed Over
   0F56 00C2   R2:42   RW RW NR   A..2-       6        0 RW  WD   Failed Over
   0F57 00C3   R2:42   RW RW NR   A..2-       6        0 RW  WD   Failed Over
   ```

2. Update the R1 volumes at site B with the latest version of the data on the R2 volumes at site A:

   ```
   symrdf -cg haxdcg_siteB update -force
   ```

   Example 7-46 shows the state of the pairs after you run the update command and the R1 update is completed.

   Example 7-46 Output of the state of the pairs after running the update command

   ```
   Symmetrix ID: 000190100304

                                Local Device View
   ---------------------------------------------------------------------------
                         STATUS     MODES                     RDF  S T A T E S
   Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
   Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
   ---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

   0F53 00BF   R1:41   RW RW RW   S..1-       0        0 RW  WD   Suspended
   0F54 00C0   R1:41   RW RW RW   S..1-       0        0 RW  WD   Suspended
   0F55 00C1   R2:42   RW RW RW   A..2-       0        0 RW  WD   R1 Updated
   0F56 00C2   R2:42   RW RW RW   A..2-       0        0 RW  WD   R1 Updated
   0F57 00C3   R2:42   RW RW RW   A..2-       0        0 RW  WD   R1 Updated
   ```

The volumes of site B that are part of RG_siteb are now updated from their corresponding pairs at site A.

3. Bring offline the RG_siteb resource group at site A to fail back the SRDF relationship to its original state from site B to site A. Use the cluster C-SPOC SMIT menus. Run the **smitty hacmp** command. Then, select **System Management (C-SPOC)** → **Resource Group and Applications** → **Bring a Resource Group Offline**.
4. After you bring the RG_siteb resource group offline at site A, fail back the SRDF pairs to their original state (R1 → R2):

   ```
   symrdf -cg haxdcg_siteB failback -force
   ```

   Example 7-47 shows the status of the disk pairs after the failback operation.

Example 7-47 Status of the pairs after failback

```
root@xdemcb2:/>symrdf list pd

Symmetrix ID: 000190101983

                                Local Device View
-------------------------------------------------------------------------------
                     STATUS     MODES                     RDF  S T A T E S
Sym        RDF       ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G     SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41   RW WD RW   S..2-       0        0 WD  RW   Suspended
00C0 0F54   R2:41   RW WD RW   S..2-       0        0 WD  RW   Suspended
00C1 0F55   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C2 0F56   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C3 0F57   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
```

5. Start the cluster services on the xdemcb1 and xdemcb2 nodes. When you start the cluster services on the nodes at site B, the RG_sitea is acquired in the secondary online state. As a result, SRDF/S pairs are automatically re-established.

   Because of the previous manual activation/deactivation of the RG_siteb resource group, the cluster does not automatically acquire the resource group. Manually bring the resource group back online on a node at site B.
6. Activate the RG_siteb at site B by using the SMIT C-SPOC menus. Run the **smitty hacmp** command. Select **System Management (C-SPOC)** → **Resource Group and Applications** → **Bring a Resource Group Online**. Activate the RG_siteb resource group on the xdemcb1 node. Example 7-48 shows the final resource group status.

Example 7-48 Final status of the resource groups after bringing the resource group online

```
root@xdemcb2:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               OFFLINE                        xdemcb1@siteB
               ONLINE SECONDARY               xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
```

```
                ONLINE SECONDARY            xdemca1@siteA
                OFFLINE                     xdemca2@siteA
```

Example 7-49 shows the final status of the SRDF pairs.

Example 7-49 SRDF pair status

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
-------------------------------------------------------------------------
                    STATUS     MODES                     RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv -----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- -------- ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C0 0F54   R2:41   RW WD RW   S..2-       0        0 WD  RW   Synchronized
00C1 0F55   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C2 0F56   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C3 0F57   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
```

7.5.3 Storage access failure

We assume an initial state of the resource groups with each one active in its primary site, and perform the test on the RG_sitea resource group active at site A.

We simulate the storage loss condition in our environment for site A by setting offline all the HBAs connected to the storage devices on both nodes at site A. We run on both xdemca1 and xdemca2 nodes the command:

```
powermt disable hba=<hba#>
```

The result is a resource group failover first to the second node in the priority list, xdemca2. Since the storage connection is also disabled for the xdemca2 node, the resource group fails over to the secondary site. Example 7-50 shows an intermediate state of the RG_sitea resource group trying to activate on the xdemca2 node, but failing.

Example 7-50 RG_sitea resource group that is being acquired on xdemca2 node

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                        xdemca1@siteA
               ACQUIRING                      xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

```
........

root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       OFFLINE                        xdemca1@siteA
               TEMPORARY ERROR                xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

The RG_sitea resource group finally fails over to the secondary site. After the failover, Example 7-51 shows the final state of the resource group.

Example 7-51 The state of the resource groups after the failover

```
root@xdemcb1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB

RG_siteb       ONLINE                         xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
               ONLINE SECONDARY               xdemca1@siteA
               OFFLINE                        xdemca2@siteA
```

Example 7-52 shows the state of the SRDF relationship. We run the command to check the SRDF state of the volume on a host at site B. Only the storage connection at site B is available now.

Example 7-52 State of SRDF relationship after failover

```
root@xdemcb1:/>symrdf list pd

Symmetrix ID: 000190101983

                            Local Device View
-----------------------------------------------------------------------------
                    STATUS     MODES                     RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

00BF 0F53   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
00C0 0F54   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
```

```
00C1 0F55   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C2 0F56   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
00C3 0F57   R1:42   RW RW RW   A..1-       0        0 RW  WD   Consistent
```

Observe again the swap of R1/R2 roles for the volume pairs that are associated with the RG_sitea resource group because of the failover process to site B, while the RDF links are operational.

Recovery actions after the storage connection is re-enabled

Enable the storage connection on nodes at site A using the **powermt enable hba=<hba#>** command. After you re-enable the storage connection for the nodes at site A, check the status of the FC adapters by using the **powermt display** command (Example 7-53).

Example 7-53 Display the HBA status on a node

```
root@xdemca1:/>powermt display
Symmetrix logical device count=58
CLARiiON logical device count=0
Hitachi logical device count=0
Invista logical device count=0
HP xp logical device count=0
Ess logical device count=0
HP HSx logical device count=0
================================================================================
----- Host Bus Adapters ---------  ------ I/O Paths -----  ------ Stats ------
###  HW Path                       Summary   Total   Dead  IO/Sec Q-IOs Errors
================================================================================
   0 fscsi0                        optimal     116      0       -     0      0
   1 fscsi1                        optimal     116      0       -     0      0
```

Both adapters are now enabled and paths to the storage volumes are available again. The resource group is now active at site B. To fall back the resource group to the primary site A, perform a resource group move operation from site B to site A. We perform the resource group movement on a cluster node using the SMIT C-SPOC menu (Figure 7-24).

```
                    Move Resource Group(s) to Another Site

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

                                                      [Entry Fields]
  Resource Group(s) to be Moved                      RG_sitea
  Destination Site                                   siteA
```

Figure 7-24 Performing resource group move from site B to site A

Example 7-54 shows the final state of the resource groups.

Example 7-54 Final state of the resource groups after RG_sitea failback

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                      Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                     xdemca1@siteA
```

```
                OFFLINE                 xdemca2@siteA
                ONLINE SECONDARY        xdemcb1@siteB
                OFFLINE                 xdemcb2@siteB

RG_siteb        ONLINE                  xdemcb1@siteB
                OFFLINE                 xdemcb2@siteB
                ONLINE SECONDARY        xdemca1@siteA
                OFFLINE                 xdemca2@siteA
```

7.6 Maintaining the cluster configuration with SRDF replicated resources

This section describes how to maintain the cluster configuration with SRDF replicated resources.

7.6.1 Changing an SRDF replicated resource

You can change the SRDF replicated resource by using the SMIT menus. Run the **smitty hacmp** command. Then, select **Extended Configuration → Extended Resource Configuration → Configure EMC SRDF Replicated Resources → Change/Show EMC SRDF replicated resource**. Select the name of the resource that you want to modify (Figure 7-25).

```
                    Change/Show EMC SRDF Replicated Resource

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

                                                       [Entry Fields]
  EMC SRDF Composite Group Name                       haxdcg_siteA
  New EMC SRDF Composite Group Name                   []                      +
* EMC SRDF Mode                                       SYNC                    +
  Device Groups                                        haxd_siteA              +
* Recovery Action                                      AUTO                    +
* Consistency Enabled                                  YES                     +
```

Figure 7-25 Change/Show SRDF Replicated Resource

You can dynamically modify the following attributes of the SRDF replicated resource:

New EMC SRDF Composite Group Name
You can specify a new name for the CG that you already selected.

EMC SRDF Mode Specify SYNC or ASYNC mode of operation for the pairs in the CG.

Device Groups You can change the device group part of the CG selected.

Recovery Action You can specify AUTO or MANUAL for the recovery action on SRDF volumes during the failover process.

Consistency Enabled You can select YES or NO to enable or disable the consistency on the specified CG.

Consistency: When you use an asynchronous mode of operation with the SRDF relationships, you must enable consistency on the composite group. We implemented consistency for synchronous operating mode.

When you change the composite group options by using the SMIT menus, the cluster scripts perform the required modifications on the composite group configuration and propagate the new composite group definition on the cluster nodes. You must synchronize the cluster after you making these changes to distribute the cluster definition of the SRDF replicated resource on all cluster nodes.

However, to manually synchronize the composite group definition on the cluster nodes:

1. Export the CG definition from the initiating node. We consider the node that is performing the initial CG change to be the initiating node:
   - For the local import on nodes at the same site:

     ```
     symcg export <CG_name> -f <file_import_local>
     ```
   - For import on nodes at a remote site:

     ```
     symcg export <CG_name> -f <file_import_remote> -rdf
     ```
2. Delete the existing definition on the target node, if applicable:

   ```
   symcg delete <CG_name> -force
   ```
3. Import the definition on the rest of the nodes in the cluster:
   - Nodes at the same site

     ```
     symcg import <CG_name> -f <file_import_local> -rdf_consistency
     ```
   - Nodes at remote site

     ```
     symcg import <CG_name> -f <file_import_remote> -rdf_consistency
     ```

7.6.2 Removing an SRDF replicated resource

Before you remove an SRDF replicated resource from the cluster configuration, remove the SRDF resource from the resource group definition. Run the **`smitty hacmp`** command. Select **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resource Group Configuration** → **Change/Show Resources and Attributes for a Resource Group**.

To access the SMIT menu for removing a SRDF-replicated resource run the **smitty hacmp** command. Select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure EMC SRDF Replicated Resources → Remove EMC SRDF Replicated Resource**. Select the replicated resource group that you want to delete (Figure 7-26).

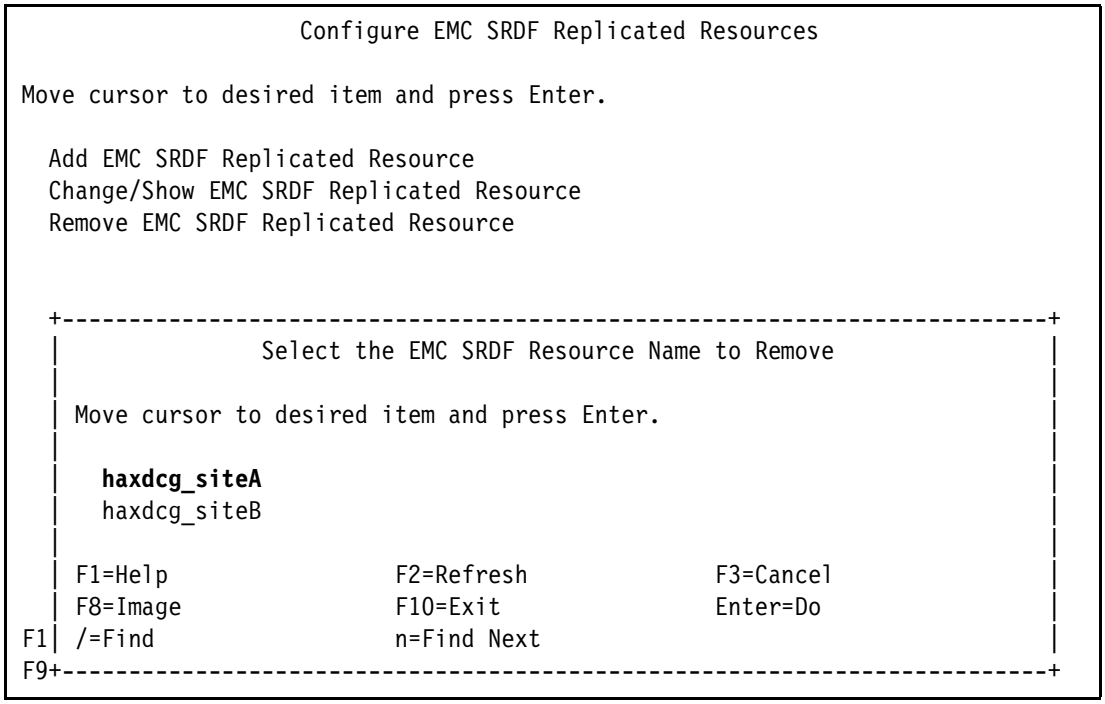


Figure 7-26 Deleting an SRDF replicated resource

Definitions: When you delete the SRDF replicated resource, only the cluster definition is removed. The device group and composite group definitions remain.

See 7.8.2, “Deleting existing device group and composite group definitions” on page 338, for an example on how to remove CG and DG definitions from the cluster nodes.

7.6.3 Changing between SRDF/S and SRDF/A operation modes

You can change the operating mode of the SRDF relationships between SYNC and ASYNC operating modes on a composite group level. The operation is dynamic and it does not require a resource group reactivation.

In our example, we try to change the resource group RG_sitea from synchronous mode to asynchronous mode. Use **smitty cl_srdf_def** to access the SMIT SRDF replicated resource. On the Change/Show EMC SRDF Replicated Resource menu, select the CG group name, and change the EMC SRDF Mode option from SYNC to ASYNC or vice versa (Figure 7-27).

```
                  Change/Show EMC SRDF Replicated Resource

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

                                                      [Entry Fields]
  EMC SRDF Composite Group Name                      haxdcg_siteA
  New EMC SRDF Composite Group Name                  []                      +
* EMC SRDF Mode                                      ASYNC                   +
  Device Groups                                       haxd_siteA              +
* Recovery Action                                     AUTO                    +
* Consistency Enabled                                 YES                     +
```

Figure 7-27 Changing the SRDF operation mode on CG haxdcg_siteA from SYNC to ASYNC

When you change the SRDF resource, the cluster runs the SRDF configuration scripts, which covert the relationships in the composite group from SYNC mode to ASYNC mode of operation. Example 7-55 shows the final status of the SRDF relations.

Example 7-55 Status of the SRDF relations after changing from SYNC to ASYNC mode of operation

```
root@xdemca2:/>symrdf -cg haxdcg_siteA query

Composite Group Name      :  haxdcg_siteA
Composite Group Type      :  RDF1
Number of Symmetrix Units :  1
Number of RDF (RA) Groups :  1
RDF Consistency Mode      :  MSC


Symmetrix ID                     : 000190100304    (Microcode Version: 5772)
Remote Symmetrix ID              : 000190101983    (Microcode Version: 5773)
RDF (RA) Group Number            :  41 (28)

        Source (R1) View             Target (R2) View       MODES STATES
--------------------------------  ------------------------ ----- ------ --------
              ST                LI      ST                       C  S
  Standard     A                 N       A                       o  u
Logical  Sym   T  R1 Inv  R2 Inv  K       T  R1 Inv  R2 Inv      n  s   RDF Pair
Device   Dev   E  Tracks Tracks   S Dev   E  Tracks  Tracks MDAE  s  p   STATE
-------------------------------- -- ------------------------ ----- ------ ---------
DEV001   0F53 RW       0       0 RW 00BF WD       0       0 A..-  X  - Consistent
DEV002   0F54 RW       0       0 RW 00C0 WD       0       0 A..-  X  - Consistent

Total            ------- -------            ------- -------
  Track(s)             0       0                  0       0
  MBs                0.0     0.0                0.0     0.0

Legend for MODES:
```

```
 M(ode of Operation)   : A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)              : X = Enabled, . = Disabled
 A(daptive Copy)       : D = Disk Mode, W = WP Mode, . = ACp off
 (Consistency) E(xempt): X = Enabled, . = Disabled, M = Mixed, - = N/A

Legend for STATES:

 Cons(istency State)   : X = Enabled, . = Disabled, M = Mixed, - = N/A
 Susp(end State)       : X = Online, . = Offline, P = Offline Pending, - = N/A
```

Run the cluster synchronization operation from the node where you performed the SRDF resource change to distribute the configuration of the SRDF replicated resource to all nodes of the cluster.

7.6.4 Adding volumes to the cluster configuration

This section shows how to extend an existing volume group in an PowerHA Enterprise Edition environment with SRDF replication by using PowerHA C-SPOC.

To add new volumes to a cluster environment with SRDF replication, first prepare the AIX configuration and establish the SRDF replication for the new pair of disks.

Adding disks: When you add disks to the cluster configuration, you are not allowed to mix volume groups that contain SRDF-protected disks with volume groups that contain non-SRDF-protected disks within the same resource group.

In the following scenario, we use a disk pair that is already configured on the hosts on both sites (Example 7-56), and we add the disk to the existing srdf_s_vg volume group at site A. In our scenario, we create a pair of disks in a SRDF/S relationship and add it to the existing device group (haxd_siteA) and composite haxdcg_siteA group.

Example 7-56 Additional disks prepared for adding to the cluster configuration

```
Command: sympd list

Site A - xdemca1 and xdemca2

Symmetrix ID: 000190100304

        Device Name           Directors                  Device
--------------------------- ------------ -------------------------------------
                                                                          Cap
Physical                Sym  SA :P DA :IT  Config        Attribute    Sts   (MB)
--------------------------- ------------ -------------------------------------

...
/dev/rhdiskpower6      0F58 07B:1 16D:C4  RAID-5        N/Grp'd      RW    4314
...


Site B - xdemcb1 and xdemcb2

Symmetrix ID: 000190101983
```

```
       Device Name           Directors                  Device
--------------------------- ------------ ------------------------------------
                                                                          Cap
Physical               Sym  SA :P DA :IT  Config        Attribute    Sts   (MB)
--------------------------- ------------ ------------------------------------
...
/dev/rhdiskpower47     00C4 13B:1 01B:CB  RAID-5        N/Grp'd      RW    4314
...
```

To extend the srdf_s_vg volume group and the /data1 file system from this volume group, and then to synchronize the volume group definition in the cluster:

1. Add the volumes by using C-SPOC. Disks need a PVID defined on all hosts. To create a PVID for the new volume on all hosts at site A (Example 7-57):

   ```
   chdev -l <dev> -a pv=yes
   ```

   Example 7-57 Creating a PVID on the new disk at site A

   ```
   root@xdemca1:/>lspv | grep hdiskpower6
   hdiskpower6     none                                None
   root@xdemca1:/>chdev -l hdiskpower6 -a pv=yes
   hdiskpower6 changed
   root@xdemca1:/>lspv | grep hdiskpower6
   hdiskpower6     00cfb52d65afdcd1                    None
   ```

2. Create the SRDF relationship between the volumes. In Example 7-58, we use the RDF group 41 containing the volumes of the target volume group for the storage at site A (SymID=0304). Use the **symdg show** command on the device group where you want to add the new disk to find the RFD group ID during disk pair creation. In Example 7-58, we add a new pair of disks in a SRDF/S relationship.

   Example 7-58 Establish the SRDF relationship between the disk pairs

   ```
   root@xdemca1:/>cat /tmp/diskadd.txt
   0F58 00C4

   root@xdemca1:/>symrdf -file /tmp/diskadd.txt createpair -type R1 -sid 0304
   -rdfg 41 -establish -rdf_mode sync -nop

   An RDF 'Create Pair' operation execution is in progress for device
   file '/tmp/diskadd.txt'. Please wait...

       Create RDF Pair in (0304,041)....................................Started.
       Create RDF Pair in (0304,041)....................................Done.
       Mark target device(s) in (0304,041) for full copy from source....Started.
       Devices: 0F58-0F58 in (0304,041)............................... Marked.
       Mark target device(s) in (0304,041) for full copy from source....Done.
       Merge track tables between source and target in (0304,041).......Started.
       Devices: 0F58-0F58 in (0304,041)............................... Merged.
       Merge track tables between source and target in (0304,041).......Done.
       Resume RDF link(s) for device(s) in (0304,041)...................Started.
       Resume RDF link(s) for device(s) in (0304,041)...................Done.

   The RDF 'Create Pair' operation successfully executed for device
   file '/tmp/diskadd.txt'.
   ```

3. Check the status of the replication by using the **symrdf list pd** command.
4. Activate the PVID on the disk at the remote site. We use the **chdev** command to accomplish this task (Example 7-59). We can run the command without splitting the SRDF pair because the target disk at site B can be opened for read operations.

Example 7-59 Activate the PVID of the disk at the remote site

```
root@xdemcb1:/>lspv | grep hdiskpower47
hdiskpower47    none                                None
root@xdemcb1:/>chdev -l hdiskpower47 -a pv=yes
hdiskpower47 changed
root@xdemcb1:/>lspv | grep hdiskpower47
hdiskpower47    00cfb52d65afdcd1                    None
```

The PVID of the disk is the same as the PVID of the source disk at site A.

5. Add the device in the device group that contains the srdf_s_vg target volume group. In Example 7-60, we add the 0F58 (hdiskpower6) volume to the DG haxd_siteA already defined.

Example 7-60 Adding a disk to the device group on the xdemca1 node

```
root@xdemca1:/>symdg list

                        D E V I C E      G R O U P S

                                                                    Number of
  Name                Type     Valid  Symmetrix ID  Devs   GKs  BCVs  VDEVs  TGTs

  haxd_siteA          RDF1     Yes    000190100304     2     1     0      0     0
  haxd1_siteB         RDF2     Yes    000190100304     2     1     0      0     0
  haxd2_siteB         RDF2     Yes    000190100304     1     1     0      0     0

root@xdemca1:/>symld -sid 0304 -g haxd_siteA add dev 0F58

root@xdemca1:/>symdg list

                        D E V I C E      G R O U P S

                                                                    Number of
  Name                Type     Valid  Symmetrix ID  Devs   GKs  BCVs  VDEVs  TGTs

  haxd_siteA          RDF1     Yes    000190100304     3     1     0      0     0
  haxd1_siteB         RDF2     Yes    000190100304     2     1     0      0     0
  haxd2_siteB         RDF2     Yes    000190100304     1     1     0      0     0
```

6. Propagate the DG update on the cluster nodes. You can use local and remote commands to run at each site (Example 7-61).

   Important: Before you add the disk to the device group definition on a node in the cluster, run the **symcfg sync** command to update the SYMAPI local database with the newly discovered devices and their SRDF status.

   Example 7-61 Adding the disk definition in the DGs on all cluster nodes

```
On nodes at site A:
symld -sid 0304 -g haxd_siteA add dev 0F58

On nodes at site B:
symld -sid 1983 -g haxd_siteA add dev 00C4
```

7. Add the disk to the CG defined at site A. On each node of the cluster, add the disk in the haxdcg_siteA composite group (Example 7-62).

   Example 7-62 Modifying the composite group definitions on all nodes

```
Disable the consistency on the composite group:
root@xdemca1:/>symcg -cg haxdcg_siteA disable -force

Execute a consistency 'Disable' operation for composite
group 'haxdcg_siteA' (y/[n]) ? y

A consistency 'Disable' operation execution is
in progress for composite group 'haxdcg_siteA'. Please wait...

The consistency 'Disable' operation successfully executed for
composite group 'haxdcg_siteA'.

RUN on nodes at site A:
symcfg sync
symcg -cg haxdcg_siteA add dev F58

RUN on site nodes at site B:
symcfg sync
symcg -cg haxdcg_siteA add dev 0C4

Enable the consistency on the composite group:
root@xdemca1:/>symcg -cg haxdcg_siteA enable

Execute a consistency 'Enable' operation for composite
group 'haxdcg_siteA' (y/[n]) ? y

A consistency 'Enable' operation execution is
in progress for composite group 'haxdcg_siteA'. Please wait...


The consistency 'Enable' operation successfully executed for
composite group 'haxdcg_siteA'.
```

   We disabled the consistency on the CG before we added the volume in the composite group.

8. Add the volume to the volume group on the xdemca1 node where the RG_sitea resource group is active, and extend the existing /data1 file system with 4 GB:

   ```
   extendvg srdf_s_vg hdiskpower6
   ```

   Check the srdf_s_vg volume group physical volumes (Example 7-63).

Example 7-63 srdf_s_vg volume group physical volumes

```
root@xdemca1:/>lsvg -p srdf_s_vg
srdf_s_vg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdiskpower1       active            539         77          00..00..00..00..77
hdiskpower2       active            539         0           00..00..00..00..00
hdiskpower6       active            539         539
108..108..107..108..108
```

After you extend the `/data1` file system by using the **`chfs -a size=+4G /data1`** command, physical partitions are allocated from the new hdiskpower6 disk (Example 7-64).

Example 7-64 Physical partitions allocated from the new hdiskpower6 disk

```
root@xdemca1:/>lsvg -p srdf_s_vg
srdf_s_vg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdiskpower1       active            539         0           00..00..00..00..00
hdiskpower2       active            539         0           00..00..00..00..00
hdiskpower6       active            539         104         00..00..00..00..104
```

9. Synchronize the volume group definition across the cluster nodes by using C-SPOC. Run the **`smitty cl_vg`** command, and select **Synchronize a Volume Group Definition** (Figure 7-28).

Important: Perform the C-SPOC operation only when the cluster is active on all PowerHA nodes and the underlying SRDF pairs are in a synchronized state.

```
                                Volume Groups

Move cursor to desired item and press Enter.

  List All Volume Groups
  Create a Volume Group
  Create a Volume Group with Data Path Devices

  Set Characteristics of a Volume Group
  Enable a Volume Group for Fast Disk Takeover or Concurrent Access

  Import a Volume Group
  Mirror a Volume Group
+--------------------------------------------------------------------------+
|                    Select the Volume Group to Update                     |
|                                                                          |
| Move cursor to desired item and press Enter. Use arrow keys to scroll.   |
|                                                                          |
|   #Volume Group    Resource Group             Node List                  |
|    srdf_a_vg1      RG_siteb                   xdemca1,xdemca2,xdemcb1,xd |
|    srdf_a_vg2      RG_siteb                   xdemca1,xdemca2,xdemcb1,xd |
|    srdf_s_vg       RG_sitea                   xdemca1,xdemca2,xdemcb1,xd |
|    diskhb_b_vg     <Not in a Resource Group>  xdemcb1,xdemcb2            |
|    srdf_a_vg1      RG_siteb                   xdemcb1,xdemcb2            |
|    srdf_a_vg2      RG_siteb                   xdemcb1,xdemcb2            |
|    srdf_s_vg       RG_sitea                   xdemcb1,xdemcb2            |
|                                                                          |
| F1=Help                 F2=Refresh              F3=Cancel                |
| F8=Image                F10=Exit                Enter=Do                 |
| /=Find                  n=Find Next                                      |
+--------------------------------------------------------------------------+
```

Figure 7-28 Synchronizing the volume group definition by using C-SPOC

Expected error: When using C-SPOC to modify a volume group that contains an SRDF replicated resource, expect to see the following error:

```
cl_extendvg: Error executing clupdatevg srdf_s_vg 00cfb52d62712021 on node
xdemcb1

cl_extendvg: Error executing clupdatevg srdf_s_vg 00cfb52d62712021 on node
xdemcb2
```

The updated definition of the volume group for the nodes in the remote site can be performed later by the cluster lazy update process during a site failover, or you can perform a resource group move to the other site.

7.7 Troubleshooting PowerHA Enterprise Edition SRDF managed replicated resources

All PowerHA SRDF operations are performed on a composite group and not on individual device groups. The composite group that is enabled for consistency is the consistency group. The consistency groups operate in unison to preserve the integrity and dependent write consistency of a database that is distributed across multiple arrays. In the Change/Show EMC SRDF replicated resource panel, if the SRDF consistency group field is set to Yes, then consistency is enabled for the SRDF replicated resource, which is defined as the composite group on the EMC storage. Otherwise, consistency is not enabled for the resource.

Table 7-3 lists SRDF states and the recovery actions that are attempted for the SRDF relationships inside a composite group.

Table 7-3 Valid states for the SRDF relationships in a composite group

State	Description	Action
SyncInProg	A synchronization is currently in progress between the R1 and the R2. There are existing invalid tracks between the two pairs and the logical link between both sides of an RDF pair is up.	State is recovered. Final state is Synchronized/Consistent.
Synchronized	The R1 and the R2 are currently in a synchronized state. The same content exists on the R2 as the R1. There are no invalid tracks between the two pairs. This state is applicable only to synchronous mirroring.	No need for any action.
Split	The R1 and the R2 are currently Ready to their hosts, but the link is Not Ready or Write Disabled.	State recovered only with AUTO Recovery Action. Otherwise, no recovery is attempted and RG goes to ERROR state.
FailedOver	The R1 is Not Ready or Write Disabled and operations have been failed over to the R2.	State is recovered. Final state is Synchronized/Consistent.
R1 Updated	The R1 is Not Ready or Write Disabled to the host, there are no local invalid tracks on the R1 side, and the link is Ready or Write Disabled.	State is recovered. Final state is Synchronized/Consistent.
R1 UpdInProg	The R1 is Not Ready or Write Disabled to the host, there are invalid local (R1) tracks on the source side, and the link is Ready or Write Disabled.	State is recovered. Final state Synchronized/Consistent.
Suspended	The RDF links are suspended and are Not Ready or Write Disabled. If the R1 is Ready while the links are suspended, any I/O accumulates as invalid tracks owed to the R2.	State is recovered. Final state is Synchronized/Consistent.
Partitioned	The SYMAPI is unable to communicate through the corresponding RDF path to the remote Symmetrix.	State recovered only with AUTO Recovery Action. Otherwise, no recovery is attempted and RG goes to ERROR state.
Mixed	A composite SYMAPI device group RDF pair state. Different SRDF pair states exist within a device group.	State will *not* be recovered. RG goes to ERROR state.
Invalid	This is the default state when no other SRDF state applies. The combination of R1, R2, and RDF link states and statuses do not match any other pair state. This state can occur if there is a problem at the disk director level.	State will *not* be recovered. RG goes to ERROR state.

State	Description	Action
Consistent	The R2 SRDF/A capable devices are in a consistent state. Consistent state signifies the normal state of operation for device pairs that are operating in asynchronous mode. This state is applicable only to asynchronous mirroring.	No action required.
Transmit idle	The SRDF/A session cannot push data in the transmit cycle across the link because the link is down. This state is applicable only to asynchronous mirroring.	State recovered only with AUTO Recovery Action. Otherwise, no recovery is attempted and RG goes to ERROR state.

7.8 Commands for managing the SRDF environment

This section contains information about commands to help manage an SRDF environment.

7.8.1 SYMCLI commands for SRDF environment

The SYMCLI commands can help manage the SRDF environment.

- Example 7-65 shows how to check the communication link between the local and the remote Symmetrix storage subsystems.

Example 7-65 Checking the communication link

```
# symcfg list

                                S Y M M E T R I X

                                          Mcode    Cache      Num Phys  Num Symm
    SymmID       Attachment  Model       Version  Size (MB)  Devices   Devices

    000190100304 Local       DMX3-24     5772       32768       294      4063
    00019010 1983 Remote     DMX4-24     5773       65536         0

# symrdf ping -sid 1983

Successfully pinged (Remotely) Symmetrix ID: 000190101983
```

- Example 7-66 shows the status of the remote link directors on both local and remote storage.

Example 7-66 Displaying the status of the remote link directors

```
# symcfg -RA all list


Symmetrix ID: 000190100304 (Local)

                S Y M M E T R I X   R D F   D I R E C T O R S


                                          Remote        Local    Remote
Ident  Symb  Num  Slot  Type       Attr  SymmID        RA Grp   RA Grp  Status

RF-4C   04C   36     4  RDF-BI-DIR   -   000190101983 200 (C7) 200 (C7) Online
```

```
                                        -   000190101983 201 (C8) 201 (C8)
                                        -   000190101983 202 (C9) 202 (C9)
                                        -   000190101983  41 (28)  41 (28)
                                        -   000190101983  42 (29)  42 (29)
                                        -   000190101983  43 (2A)  43 (2A)
                                        -   000190101983  45 (2C)  45 (2C)
                                        -   000190101983  74 (49)  74 (49)
                                        -   000190101983 205 (CC) 205 (CC)
                                        -   000190101983 210 (D1) 210 (D1)
                                        -   000190101983 203 (CA) 203 (CA)
                                        -   000190101983 204 (CB) 204 (CB)
                                        -   000190101983  42 (29)  42 (29)
                                        -   000190101983  41 (28)  41 (28)
                                        -   000190101983  43 (2A)  43 (2A)
RF-13C  13C   45    13  RDF-BI-DIR     -   000190101983 200 (C7) 200 (C7) Online
                                        -   000190101983 201 (C8) 201 (C8)
                                        -   000190101983 202 (C9) 202 (C9)
                                        -   000190101983  41 (28)  41 (28)
                                        -   000190101983  42 (29)  42 (29)
                                        -   000190101983  43 (2A)  43 (2A)
                                        -   000190101983 203 (CA) 203 (CA)
                                        -   000190101983 204 (CB) 204 (CB)
                                        -   000190101983  74 (49)  74 (49)
                                        -   000190101983  45 (2C)  45 (2C)
                                        -   000190101983 205 (CC) 205 (CC)
                                        -   000190101983  41 (28)  41 (28)
                                        -   000190101983  42 (29)  42 (29)
                                        -   000190101983  43 (2A)  43 (2A)
                                        -   -             211 (D2)   -
RF-4D   04D   52     4  RDF-BI-DIR     -   -               1 (00)   -      Offline
                                        -   -              10 (09)   -
                                        -   -              11 (0A)   -
                                        -   -              12 (0B)   -
                                        -   -              13 (0C)   -
                                        -   -              14 (0D)   -
                                        -   -              17 (10)   -
                                        -   -              18 (11)   -
                                        -   -              22 (15)   -
                                        -   -              23 (16)   -
                                        -   -              25 (18)   -
                                        -   -              26 (19)   -
                                        -   -              27 (1A)   -
                                        -   -              40 (27)   -
                                        -   -              46 (2D)   -
                                        -   -             101 (64)   -
                                        -   -             102 (65)   -
                                        -   -             103 (66)   -
                                        -   -             104 (67)   -
                                        -   -             105 (68)   -
                                        -   -             106 (69)   -
                                        -   -             107 (6A)   -
RF-13D  13D   61    13  RDF-BI-DIR     -   -               1 (00)   -      Offline
                                        -   -              10 (09)   -
                                        -   -              11 (0A)   -
                                        -   -              12 (0B)   -
```

```
                                        -   -            13 (0C)   -
                                        -   -            14 (0D)   -
                                        -   -            17 (10)   -
                                        -   -            19 (12)   -
                                        -   -            22 (15)   -
                                        -   -            23 (16)   -
                                        -   -            25 (18)   -
                                        -   -            26 (19)   -
                                        -   -            28 (1B)   -
                                        -   -            40 (27)   -
                                        -   -            46 (2D)   -
                                        -   -           101 (64)   -
                                        -   -           102 (65)   -
                                        -   -           103 (66)   -
                                        -   -           104 (67)   -
                                        -   -           105 (68)   -
                                        -   -           106 (69)   -
                                        -   -           107 (6A)   -

Symmetrix ID: 000190101983 (Remote)

                S Y M M E T R I X   R D F   D I R E C T O R S


                                                 Remote        Local    Remote
Ident  Symb  Num  Slot  Type       Attr  SymmID        RA Grp   RA Grp  Status

RF-3C  03C   35     3  RDF-BI-DIR   -   000190100304 200 (C7) 200 (C7) Online
                                        -   000190100304 200 (C7) 200 (C7)
                                        -   000190100304 201 (C8) 201 (C8)
                                        -   000190100304 201 (C8) 201 (C8)
                                        -   000190100304 202 (C9) 202 (C9)
                                        -   000190100304 202 (C9) 202 (C9)
                                        -   000190100304  41 (28)  41 (28)
                                        -   000190100304  41 (28)  41 (28)
                                        -   000190100304  42 (29)  42 (29)
                                        -   000190100304  42 (29)  42 (29)
                                        -   000190100304  43 (2A)  43 (2A)
                                        -   000190100304  43 (2A)  43 (2A)
                                        -   000190100304 203 (CA) 203 (CA)
                                        -   000190100304  45 (2C)  45 (2C)
                                        -   000190100304 204 (CB) 204 (CB)
                                        -   000190100304  74 (49)  74 (49)
                                        -   000190100304  74 (49)  74 (49)
                                        -   000190100304 205 (CC) 205 (CC)
                                        -   000190100304  45 (2C)  45 (2C)
                                        -   000190100304 205 (CC) 205 (CC)
                                        -   000190100304 210 (D1) 210 (D1)
                                        -   000190100304 203 (CA) 203 (CA)
                                        -   000190100304 204 (CB) 204 (CB)
RF-14C 14C   46    14  RDF-BI-DIR   -   -             45 (2C)   -      Offline
                                        -   -             74 (49)   -
                                        -   -            200 (C7)   -
                                        -   -            201 (C8)   -
                                        -   -            202 (C9)   -
```

```
                                        -   -            203 (CA)   -
                                        -   -            204 (CB)   -
                                        -   -            205 (CC)   -
                                        -   -            211 (D2)   -
RF-3D   03D   51     3  RDF-BI-DIR     -   -              1 (00)   -      Offline
                                        -   -             10 (09)   -
                                        -   -             11 (0A)   -
                                        -   -             12 (0B)   -
                                        -   -             13 (0C)   -
                                        -   -             14 (0D)   -
                                        -   -             17 (10)   -
                                        -   -             19 (12)   -
                                        -   -             22 (15)   -
                                        -   -             23 (16)   -
                                        -   -             25 (18)   -
                                        -   -             26 (19)   -
                                        -   -             28 (1B)   -
                                        -   -             40 (27)   -
                                        -   -             46 (2D)   -
                                        -   -            101 (64)   -
                                        -   -            102 (65)   -
                                        -   -            103 (66)   -
                                        -   -            104 (67)   -
                                        -   -            105 (68)   -
                                        -   -            106 (69)   -
                                        -   -            107 (6A)   -
RF-14D  14D   62    14  RDF-BI-DIR     -   000190100304  41 (28)  41 (28) Online
                                        -   000190100304  42 (29)  42 (29)
                                        -   000190100304  41 (28)  41 (28)
                                        -   000190100304  42 (29)  42 (29)
                                        -   000190100304  43 (2A)  43 (2A)
                                        -   000190100304  43 (2A)  43 (2A)
                                        -   -              1 (00)   -
                                        -   -             10 (09)   -
                                        -   -             11 (0A)   -
                                        -   -             12 (0B)   -
                                        -   -             13 (0C)   -
                                        -   -             14 (0D)   -
                                        -   -             17 (10)   -
                                        -   -             18 (11)   -
                                        -   -             22 (15)   -
                                        -   -             23 (16)   -
                                        -   -             25 (18)   -
                                        -   -             26 (19)   -
                                        -   -             27 (1A)   -
                                        -   -             40 (27)   -
                                        -   -             46 (2D)   -
                                        -   -            101 (64)   -
                                        -   -            102 (65)   -
                                        -   -            103 (66)   -
                                        -   -            104 (67)   -
                                        -   -            105 (68)   -
                                        -   -            106 (69)   -
                                        -   -            107 (6A)   -
```

- Example 7-67 shows I/O statistics for the RDF ports on the local storage (for 5-second time intervals).

Example 7-67 IO statistics for the RDF ports

```
symstat -RA all -i 5

         DIRECTOR     IO/sec     Cache Requests/sec   % RW   KB/sec
22:05:28              Remote    READ   WRITE      RW  Hits   rcvd+sent

22:05:33 RF-4C            14       0       0       0   N/A           0
         RF-13C           17       0       4       4     0         299
                        ------  ------  ------  ------   ---  ----------
Total                     31       0       4       4     0         299


22:05:39

22:05:44 RF-4C           421       0     404     404     0       25883
         RF-13C          673       0     655     655     0       41871
                        ------  ------  ------  ------   ---  ----------
Total                   1094       0    1059    1059     0       67754
```

- Device group and composite group example outputs: Example 7-68 shows detailed information about a device group.

Example 7-68 Sample symdg show output

```
root@xdemca1:/>symdg show haxd_siteA

Group Name:  haxd_siteA

    Group Type                                   : RDF1     (RDFA)
    Device Group in GNS                          : No
    Valid                                        : Yes
    Symmetrix ID                                 : 000190100304
    Group Creation Time                          : Tue Mar  9 09:28:26 2010
    Vendor ID                                    : EMC Corp
    Application ID                               : SYMCLI

    Number of STD Devices in Group               :    2
    Number of Associated GK's                    :    1
    Number of Locally-associated BCV's           :    0
    Number of Locally-associated VDEV's          :    0
    Number of Locally-associated TGT's           :    0
    Number of Remotely-associated VDEV's(STD RDF):    0
    Number of Remotely-associated BCV's (STD RDF):    0
    Number of Remotely-associated TGT's(TGT RDF) :    0
    Number of Remotely-associated BCV's (BCV RDF):    0
    Number of Remotely-assoc'd RBCV's (RBCV RDF) :    0
    Number of Remotely-assoc'd BCV's (Hop-2 BCV) :    0
    Number of Remotely-assoc'd VDEV's(Hop-2 VDEV):    0
    Number of Remotely-assoc'd TGT's (Hop-2 TGT) :    0

    Standard (STD) Devices (2):
        {
```

```
    ---------------------------------------------------------------------
                                                Sym                 Cap
    LdevName              PdevName              Dev  Att. Sts       (MB)
    ---------------------------------------------------------------------
    DEV001                /dev/rhdiskpower1     0F53      RW        4314
    DEV002                /dev/rhdiskpower2     0F54      RW        4314
    }

Associated GateKeeper Devices (1):
    {
    ---------------------------------------------------------------------
                                                Sym                 Cap
    LdevName              PdevName              Dev       Sts       (MB)
    ---------------------------------------------------------------------
    N/A                   /dev/rhdiskpower10    0F9A      RW           6
    }

Device Group RDF Information
    {
    RDF Type                                : R1
    RDF (RA) Group Number                   : 41              (28)

    Remote Symmetrix ID                     : 000190101983

    R2 Device Is Larger Than The R1 Device : False

    Paired with a Diskless Device           : False
    Paired with a Concurrent Device         : False
    Paired with a Cascaded Device           : False

    RDF Pair Configuration                  : Normal
    RDF STAR Mode                           : False

    RDF Mode                                : Synchronous
    RDF Adaptive Copy                       : Disabled
    RDF Adaptive Copy Write Pending State   : N/A
    RDF Adaptive Copy Skew (Tracks)         : 32767

    RDF Device Domino                       : Disabled

    RDF Link Configuration                  : Fibre
    RDF Link Domino                         : Disabled
    Prevent Automatic RDF Link Recovery     : Enabled
    Prevent RAs Online Upon Power ON        : Enabled

    Device RDF Status                       : Ready           (RW)

    Device RA Status                        : Ready           (RW)
    Device Link Status                      : Ready           (RW)
    Time of Last Device Link Status Change : N/A

    Device Suspend State                    : N/A
    Device Consistency State                : Disabled
    Device Consistency Exempt State         : N/A
    RDF R2 Not Ready If Invalid             : Disabled
```

```
Device RDF State                      : Ready           (RW)
Remote Device RDF State               : Write Disabled  (WD)

RDF Pair State (  R1 <===> R2 )       : Synchronized

Number of R1 Invalid Tracks           : 0
Number of R2 Invalid Tracks           : 0

RDFA Information:
    {
    Session Number                    : 40
    Cycle Number                      : 0
    Number of Devices in the Session  : 2
    Session Status                    : Inactive
    Consistency Exempt Devices        : No

    Session Consistency State         : N/A
    Minimum Cycle Time                : 00:00:30
    Average Cycle Time                : 00:00:00
    Duration of Last cycle            : 00:00:00
    Session Priority                  : 33

    Tracks not Committed to the R2 Side: 0
    Time that R2 is behind R1         : 00:00:00
    R2 Image Capture Time             : N/A
    R2 Data is Consistent             : N/A
    R1 Side Percent Cache In Use      :  0
    R2 Side Percent Cache In Use      :  0

    Transmit Idle Time                : 00:00:00
    R1 Side DSE Used Tracks           :  0
    R2 Side DSE Used Tracks           :  0
    }
}
```

- Example 7-69 shows detailed information about a CG.

Example 7-69 List the detailed configuration of a composite group

```
root@xdemca1:/>symcg show haxdcg_siteA

Composite Group Name:  haxdcg_siteA

  Composite Group Type                          : RDF1
  Valid                                         : Yes
  CG in PowerPath                               : No
  CG in GNS                                     : No
  RDF Consistency Protection Allowed            : No
  RDF Consistency Mode                          : NONE
  Concurrent RDF                                : No
  Cascaded RDF                                  : No

  Number of RDF (RA) Groups                     :    1
  Number of STD Devices                         :    2
  Number of CRDF STD Devices                    :    0
```

```
  Number of BCV's (Locally-associated)              :    0
  Number of VDEV's (Locally-associated)             :    0
  Number of TGT's Locally-associated                :    0
  Number of CRDF TGT Devices                        :    0
  Number of RVDEV's (Remotely-associated VDEV)      :    0
  Number of RBCV's (Remotely-associated STD-RDF)    :    0
  Number of BRBCV's (Remotely-associated BCV-RDF)   :    0
  Number of RRBCV's (Remotely-associated RBCV)      :    0
  Number of RTGT's (Remotely-associated)            :    0
  Number of Hop2 BCV's (Remotely-assoc'ed Hop2 BCV) :    0
  Number of Hop2 VDEV's (Remotely-assoc'ed Hop2 VDEV):   0
  Number of Hop2 TGT's (Remotely-assoc'ed Hop2 TGT) :    0

  Number of Symmetrix Units (1):
    {

    1) Symmetrix ID                                 : 000190100304
       Microcode Version                            : 5772
       Number of STD Devices                        :    2
       Number of CRDF STD Devices                   :    0
       Number of BCV's (Locally-associated)         :    0
       Number of VDEV's (Locally-associated)        :    0
       Number of TGT's Locally-associated           :    0
       Number of CRDF TGT Devices                   :    0
       Number of RVDEV's (Remotely-associated VDEV) :    0
       Number of RBCV's (Remotely-associated STD_RDF) :  0
       Number of BRBCV's (Remotely-associated BCV-RDF):  0
       Number of RTGT's (Remotely-associated)       :    0
       Number of RRBCV's (Remotely-associated RBCV) :    0
       Number of Hop2BCV's (Remotely-assoc'ed Hop2BCV):  0
       Number of Hop2VDEVs(Remotely-assoc'ed Hop2VDEV):  0
       Number of Hop2TGT's (Remotely-assoc'ed Hop2TGT):  0

       Number of RDF (RA) Groups (1):
         {

         1) RDF (RA) Group Number  :  41           (28)
            Remote Symmetrix ID    : 000190101983
            Microcode Version      : 5773
            Recovery RA Group      : N/A           (N/A)
            RA Group Name          : N/A

            STD Devices (2):
              {

--------------------------------------------------------------------------
                                                    Sym  Device
Flags     Cap
              LdevName    PdevName                  Dev  Config        Sts
CSRT     (MB)

--------------------------------------------------------------------------
              DEV001      /dev/rhdiskpower1         0F53 RDF1+R-5      RW
.--1     4314
```

```
              DEV002      /dev/rhdiskpower2         0F54 RDF1+R-5      RW
.--1     4314
              }
          }
     }

Legend:
  RDFA Flags:
       C(onsistency)   : X = Enabled, . = Disabled, - = N/A
       (RDFA) S(tatus) : A = Active, I = Inactive, - = N/A
       R(DFA Mode)     : S = Single-session mode, M = MSC mode, - = N/A
       (Mirror) T(ype) : 1 = R1, 2 = R2, - = N/A
```

7.8.2 Deleting existing device group and composite group definitions

Example 7-70 shows how to delete a configuration by using a demoCG composite group with its consistency enabled and containing two device groups, RDFdemo1 and RDFdemo2.

Example 7-70 Deleting composite group definitions

```
# symrdf -cg demoCG split -nop -force         ->  split de pairs and stop the
                                                  replication;
# symcg -cg demoCG disable –force             ->  need to disable consistency on
                                                  the CG in order to delete the
                                                  pairs
# symrdf deletepair -cg demoCG                ->  delete the SRDF pairs
# symcg delete demoCG -force                  ->  delete the CG
# symdg delete RDFdemo1 –force                ->  delete DG1
# symdg delete RDFdemo2 -force                ->  delete DG2
#symrdf removegrp -sid 000190100304 -rdfg 23 ->  delete RDF dynamic group no 23,
                                                  containing the previous SRDF
                                                  relations
```

8

Configuring PowerHA SystemMirror Enterprise Edition with Geographic Logical Volume Manager

The IBM PowerHA SystemMirror Enterprise Edition with the GLVM provides disaster recovery and data mirroring capability for the data at geographically separated sites. It protects the data against total site failure by remote mirroring and supports unlimited distance between participating sites.

GLVM is base technology available with AIX using IP-based data mirroring between sites and is integrated with AIX standard Logical Volume Manager (LVM). GLVM was introduced with the HACMP/XD 5.2 release and integrated with it for automated high availability.

This solution increases data availability by providing continuing service during hardware or software outages (or both), planned or unplanned, for a two-site cluster. The distance between sites might be unlimited, and both sites might access the mirrored volume groups serially over IP-based networks.

By using this solution, your business application can continue running at the takeover system at a remote site while the failed system is being recovered from a disaster or a planned outage.

The software takes advantage of the following software components to reduce downtime and recovery time during disaster recovery:

- AIX LVM and GLVM
- TCP/IP subsystem
- PowerHA SystemMirror Enterprise Edition for AIX

The chapter includes the following sections:

- Planning the implementation of PowerHA Enterprise Edition with GLVM
- Installing and configuring PowerHA for GLVM
- Configuration wizard for GLVM

- Configuring a 4-node, 2-site PowerHA for GLVM
- Configuring a 3-node, 2-site PowerHA for GLVM
- Performance with aio_cache
- Monitoring
- Test scenarios
- Performing management operations on the cluster
- Migration from HAGEO (AIX 5.3) to GLVM (AIX 6.1)
- Data divergence in PowerHA for GLVM

8.1 Planning the implementation of PowerHA Enterprise Edition with GLVM

Before you set up PowerHA with GLVM, familiarize yourself with the planning steps to implement PowerHA. For information about planning a PowerHA cluster, see the *HACMP Planning Guide*, SC23-4861.

To plan the implementation of geographically mirrored volume groups in a PowerHA cluster, you must complete the planning tasks for implementing GLVM and the planning tasks for PowerHA integration with GLVM. See the planning worksheet in the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

The actions that you take for planning the environment depend on the initial infrastructure installation that you implemented as in the following examples:

- If you have Power Systems without PowerHA cluster configuration, plan for geographically mirrored volume groups, install PowerHA Enterprise Edition for GLVM, and integrate the PowerHA clusters onboth sites.
- Alternatively, you might have a PowerHA cluster that is already configured and you want to implement a disaster recovery solution between two sites that use the mirroring function. In this case, plan the geographically mirrored volume groups while keeping in mind your already-configured cluster resources, networks, and sites. You might need to add a second site or nodes to your cluster. You also might need to extend the existing volume groups. After you plan and configure the geographically mirrored volume groups, add them to the resource groups in the PowerHA cluster.

8.1.1 Requirements and considerations

PowerHA Enterprise Edition for GLVM has the following requirements and considerations:

- Plan and configure only two sites, one local and one remote. Site names must correspond with the site names in PowerHA.
- Up to four IP-based networks for XD_data can be configured in PowerHA.
- Fast disk takeover and disk heartbeating are not supported on remote disks that are part of a geographically mirrored volume group.
- Disk heartbeating over disks within the same site is supported. You can configure disk heartbeating for disks within the same site that belong to the geographically mirrored volume group that is also enhanced concurrently.
- You cannot create dependent resource groups for the resource group that contains geographically mirrored volume groups and has non-concurrent inter-site management policy.
- The rootvg volume group cannot be geographically mirrored.

- Use non-concurrent or enhanced concurrent mode volume groups (only for sync mode). For enhanced concurrent volume groups that you also want to make geographically mirrored, ensure that PowerHA cluster services are running before you add and manage RPVs.
- For asynchronous mirroring, volume groups must be in a scalable format, and enhanced concurrent mode volume groups cannot be used.
- You must disable the auto-on and bad-block relocation options of the volume group.
- Create mirror pools for use with asynchronous mirroring. Mirror pools are required for using asynchronous mirroring but are optional when using synchronous mirroring.
- Set the inter-disk allocation policy for logical volumes in AIX to super strict.
- Carefully consider quorum and forced varyon issues when you plan the geographically mirrored volume groups. For more information, see “Quorum and forced varyon in the Data divergence” in the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.
- In PowerHA Enterprise Edition for GLVM, the C-SPOC utility does not allow managing the geographically mirrored volume groups from a single node.

Software requirements

The IBM PowerHA Enterprise Edition for GLVM requires specific versions of AIX and RSCT. The RSCT file sets are shipped with the PowerHA installation media and are installed automatically. The PowerHA Enterprise Edition software uses 1 MB of disk space. Ensure that the `/usr` file system has 1 MB of free disk space for the upgrade.

Reference: AIX 6.1 Technology Level 2 SP3 must be installed for asynchronous mirroring to use mirror pools. For more information see Chapter 2, “Infrastructure considerations” on page 39, or the support flash site at:

`http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FLASH10673`

Required PTFs: Remember to install the latest PTFs of PowerHA for GLVM. At the time this book was written, the following PTFs were required to allow GLVM to function correctly:

- IZ69484: HA/XD GLVM does not function with two nodes at a site.
- IZ69964: POWERHA/XD with GLVM single-adapter network problems.
- IZ69945: POWERHA/XD with GLVM single-adapter network rg move problems.

Planning for geographically mirrored volume groups

You must complete the following tasks when planning for the PowerHA for GLVM:

1. Identify the RPV sites and nodes that belong to each site.
2. Identify the volume groups that you plan to configure as geographically mirrored volume groups.
3. Plan remote physical volumes (RPVs). Plan the remote physical volume clients and servers for each physical volume in a volume group.
4. Plan connections for the GLVM mirroring function between sites, and plan to ensure security of the connections.

Planning for geographically mirrored volume groups in the cluster

When you plan to integrate geographically mirrored volume groups into the cluster, you must perform the following tasks:

1. Plan the PowerHA sites. The RPV server's site names and PowerHA site names must match.
2. Plan the PowerHA networks. Define the networks that are used by PowerHA for GLVM to the cluster configuration.
3. Identify the volume groups to be included in specified PowerHA resource groups.
4. Plan which resource groups will contain the geographically mirrored volume groups.

Planning for asynchronous mirroring

The asynchronous mirroring function is split into several functional areas. Figure 8-1 shows asynchronous mirroring at the high level that is based on a simple two-node stand-alone GLVM configuration. Node A is at the production site, and node B is at the disaster recovery site (Figure 8-1).

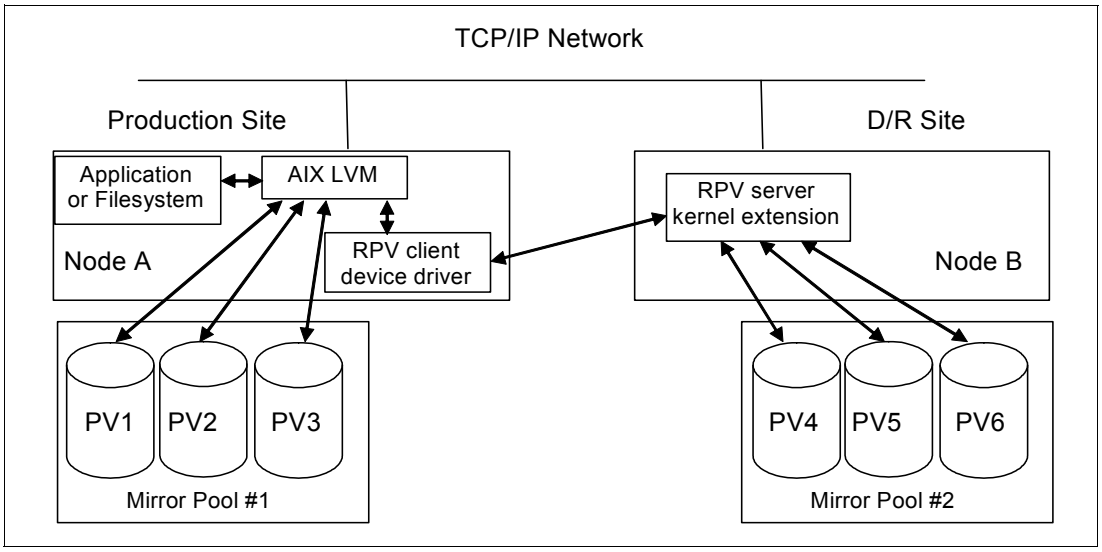


Figure 8-1 Asynchronous mirroring at a high level

Mirror pool consideration

Mirror pools provide a way for you to group disks together within a volume group. GLVM requires that volume groups use super strict mirror pools. Super strict mirror pools must follow these rules:

- Local and remote disks cannot belong to the same mirror pool.
- No more than three mirror pools per volume group.
- Each mirror pool must contain at least one copy of each logical volume.

  There are exceptions. When you create a logical volume, you must configure it so that each mirror pool gets a copy. However, if you create a mirror pool in a volume group where logical volumes exist, logical volume copies are not automatically created in the new mirror pool. You must create them by running the **mirrorvg** or **mklvcopy** command. Asynchronous GLVM mirroring requires a new type of logical volume for caching of asynchronous write requests. Do not mirror this logical volume across sites. Super strict mirror pools handle this new aio_cache logical volume type as a special case.

Additionally, mirror pools provide extra benefits to the asynchronous mirroring function:

- Synchronous or asynchronous is an attribute of a mirror pool. Rather than having to configure individual RPV devices, mirror pools provide a convenient way for users to manage asynchronous mirroring at a higher level.
- The decision of whether to mirror synchronously or asynchronously is made at the mirror pool level.

Therefore, you can decide to mirror from the production site to the disaster recovery site asynchronously and then to mirror from the disaster recovery site back to the production site synchronously. You can accomplish this task by configuring the mirror pool that contains the disaster recovery site disks as asynchronous when you configure the mirror pool that contains the production site disks as synchronous.

8.1.2 Asynchronous mirroring practices

To make asynchronous GLVM configurations as highly available and efficient as possible, you must understand your options.

Protecting against disk failure

To protect against disk failure:

- Asynchronous mirroring cannot offer the same level of protection against disk and disk adapter failures as synchronous mirroring because the remote-site disks are back level.
- Losing the production site disks that contain the cache logical volume might require a full resynchronization of the data to the remote site. LVM allows up to three mirror copies. Therefore, adding another mirror copy at the production site can offer better protection, which means that mirroring from the disaster recovery site to the production site must be done synchronously.
- The best protection can be achieved by using a disk subsystem that has built-in data mirroring or RAID capabilities at each site.

Protecting against site failure

To protect against site failure:

- Having only one node at the production site means that a node crash results in a site failure.
- You can achieve better protection by having two nodes at the production site so that a node crash results in failover to standby node at the same site.
- Reduces the likelihood of having to use an earlier level of data at the remote site.

For more information, see the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

8.2 Installing and configuring PowerHA for GLVM

This section includes the following sections:

- Installation components and prerequisites for implementing PowerHA Enterprise Edition for GLVM
- Configuring geographically mirrored volume groups
- Integrating geographically mirrored volume groups into a PowerHA cluster

8.2.1 Installation components and prerequisites for implementing PowerHA Enterprise Edition for GLVM

This section provides information about the installation prerequisites for PowerHA for GLVM.

Installation components

The PowerHA Enterprise Edition software for GLVM comprises several components. The file sets listed in Table 8-1 are required to configure PowerHA for GLVM.

Table 8-1 Installation components

Component	Description	File sets

GLVM	Builds upon AIX to provide the geographic mirroring and storage management functions	`glvm.rpv.client`
		`glvm.rpv.server`
		`glvm.rpv.util`
		`glvm.rpv.msg.en_US`

PowerHA Enterprise Edition for GLVM	Provides integration of the PowerHA high-availability function with the GLVM	`cluster.xd.glvm`
		`cluster.xd.license`
		`cluster.doc.en_US.glvm`

Installation prerequisites

Before you install the PowerHA Enterprise Edition for GLVM, remember to install the necessary software in the cluster nodes. The software has the following prerequisites:

- The latest versions of PowerHA, AIX, and RSCT.
- The PowerHA Enterprise Edition software, which uses 1 MB of disk space. Ensure that the `/usr` file system has 1 MB of free disk space for installation.

8.2.2 Configuring geographically mirrored volume groups

You configure geographically mirrored volume groups, including their corresponding logical volumes and RPVs. By configuring these entities, you can have a copy of the data for your application that is mirrored at a remote site with PowerHA Enterprise Edition support.

Configuration prerequisites

Before you configure a GLVM environment, complete these steps:

1. Install AIX on all nodes in the cluster.
2. Select site names for each site.
3. Consult the planning worksheets for the service IP addresses on the local site. These addresses serve as the service IP addresses for the RPV clients (that reside on the local site), and you will make the addresses known to the remote site.
4. Similarly, consult the planning worksheets for the service IP addresses on the remote site. These addresses serve as the service IP labels/addresses for the RPV servers (that reside on the remote site). You make the addresses known to the local site.
5. By using the standard AIX LVM SMIT panels, configure volume groups, logical volumes, and file systems for which you plan to configure geographic mirror with the GLVM utilities. Ensure that either standard or enhanced concurrent mode LVM volume groups exist for the data that you plan to be geographically mirrored.
6. For all logical volumes that are planned to be geographically mirrored, ensure that the inter-disk allocation policy is set to super strict.

7. Create volume groups and mirror pools for use with asynchronous mirroring. Mirror pools are required for using asynchronous mirroring but are optional when you use synchronous mirroring.

Cluster services: If you want to turn your existing enhanced volume groups into geographically mirrored volume groups by adding RPVs, PowerHA cluster services must be running on the nodes. Although you can create standard volume groups in AIX and add RPVs to these groups by using the GLVM utilities in SMIT, to add RPVs to enhanced concurrent volume groups, PowerHA cluster services must be running.

Configuration for standard volume groups

If you choose to have standard volume groups, you can choose either of the following actions:

- Configure the groups in AIX SMIT, add RPVs by using the GLVM utilities in SMIT, configure the PowerHA cluster, and add geographically mirrored volume groups to the resource groups.
- Configure a cluster without sites, add another site, and then add RPVs to existing volume groups, making them geographically mirrored volume groups.

For information about configuring remove physical volume servers and clients, see the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

Configuring an RPV client/server for standard volume group at node on each site

To configure a standard geographically mirrored volume by using the GLVM functions:

1. Configure an RPV server site name.

   On all nodes at the remote site, configure an RPV server site name. A site name is an attribute that is assigned to all RPV servers configured on the remote site. In a PowerHA cluster, the RPV server site name must match the site name in PowerHA.

2. Configure an RPV server.

   On one node at the remote site, configure an RPV server for the physical volumes that you plan to configure as geographically mirrored volumes. You specify that the RPV server should start immediately, which indicates that the RPV server is in the available state. Set the Configure Automatically at System Restart? field to No.

3. Configure an RPV client.

   On the local site, configure an RPV client for each RPV server. Configuring an RPV client for the previously configured RPV server establishes the physical volume as an RPV. You also specify for the RPV client to start immediately, which indicates that the RPV client is the available state.

4. Add RPVs to the volume group.

   On one participating node at the local site, add an already-configured RPV to the volume group. When at a later stage you add a mirror copy, the volume group becomes a geographically mirrored logical volume. Perform these tasks by using the GLVM utilities SMIT interface.

Extending the geographically mirrored standard volume group to other nodes in the cluster

After you complete the client/server pairs for a standard volume group at a node on each site, extend the configuration to other nodes in the cluster:

1. Configure an RPV server in the defined state on each node at the remote site.
2. Configure an RPV client in the defined state on each node at the local site.

3. To make the mirroring function work both ways in a PowerHA cluster, configure the RPV client on each node at the remote site and configure an RPV server on each node at the local site. You must understand that RPV servers and clients must be configured at both sites.
4. Vary off the volume groups and update the volume group definitions on the nodes at the other site by running the **importvg** command.

For more information, see the *PowerHA for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

Configuring enhanced concurrent mode volume groups

If you choose enhanced concurrent volume groups that will also be geographically mirrored with GLVM, for certain tasks, you must configure a PowerHA cluster and run cluster services. You can create an enhanced concurrent volume group in AIX, or alternatively, extend an existing enhanced concurrent volume group to another site.

Configuring RPV client/server pairs for an enhanced concurrent volume group on a node at each site

If you configured a PowerHA cluster that spans two sites, configure an enhanced concurrent mode volume groups that is also geographically mirrored:

1. While the cluster services are running, configure an enhanced concurrent volume group on the disks at the local site by using the AIX LVM utilities.
2. Using the GLVM utilities in SMIT, on the node at the local site, add RPVs to the volume group.
   a. On the local site, add the RPV clients to the volume group.
   b. On the local site, use the GLVM utilities to add a volume group mirror copy to the RPVs.
   c. Import the volume group to all the remaining nodes at the local site.
   d. On the local site, vary off the volume group and configure all the RPV clients into the defined state.
   e. On the remote site, configure all the RPV servers into the defined state.
3. By using the GLVM utilities in SMIT, on the node at the remote site, add RPVs to allow the disks at the local sited to be accessed from the remote site.
4. Import the volume group to all of the nodes at the remote site.
5. Vary off the volume groups on all nodes at the remote site.
6. Add the enhanced concurrent volume group to a PowerHA resource group.

Extending an existing enhanced concurrent volume group to nodes at the remote site

When starting with an existing enhanced concurrent volume group (configured within just one site), extend this volume group to the other site:

1. Extend the cluster to contain another site. Add a site name, and then add the nodes at the second site to the PowerHA cluster.
2. Add RPVs to the volume group to allow the disks at the remote site to be accessed from the local site:
   a. On the local site, add the RPVs to the volume group.
   b. On the local site, use GLVM utilities to add a volume group mirror copy to the RPVs.
   c. On the local site, vary off the volume group and configure all RPV clients into the defined state.
   d. On the remote site, configure all the RPV servers into the defined state.

3. Add RPVs to allow the disks at the local site to be accessed from the remote site.
4. Import the volume group to all of the nodes at the remote site.
5. Vary off the volume group on all nodes at the remote site.
6. Add the enhanced concurrent geographically mirrored volume group to a PowerHA resource group.

Configuring asynchronous mirroring

Before you configure asynchronous mirroring, read these general considerations for asynchronous mirroring:

- All disks in all mirror pools must be accessible to be configured for asynchronous mirroring.
- After a mirror pool is configured for asynchronous mirroring, certain active disks are needed from each mirror pool to convert the mirror pool from asynchronous mirroring to synchronous mirroring. If you want to remove one or more mirror pools from a site that is down, disable asynchronous mirroring using the **chmp** command with the **-S** and **-f** flags.
- Asynchronous mirroring is supported only on nonconcurrent scalable volume groups with mirror pools set to be super strict.
- You must disable the auto-on and bad-block relocation options of the volume group.
- The volume group must be varied on to make mirror pool changes.
- You cannot remove or reduce an aio_cache type logical volume when it is part of the asynchronous mirroring setup.
- The rootvg cannot be configured for asynchronous mirroring.

Setting up asynchronous mirroring for a mirror pool

To set up asynchronous mirroring for a mirror pool:

1. If your volume group is in ordinary or big VG format, convert it to a scalable VG format before you configure it for asynchronous mirroring, because asynchronous mirroring requires the volume group to be in scalable VG format. You can change your volume group to scalable VG format. First, vary off the volume group. Then enter the following **chvg** command:

   ```
   chvg -G datavg
   ```

2. If you have not already done so, prevent the volume group from being varied online automatically during system startup:

   ```
   chvg -a n datavg
   ```

3. Because asynchronous mirroring requires the bad block relocation policy to be turned off for the volume group, turn off bad block relocation:

   ```
   chvg -b n datavg
   ```

4. Because asynchronous mirroring requires the volume group to be configured to use super strict mirror pools, configure the super strict mirror pools:

   ```
   chvg -M s datavg
   ```

   Asynchronous mirroring also requires the local and remote disks to belong to separate mirror pools.

5. Define a mirror pool for the local disks at the production site. You need to choose a name for the mirror pool. This name needs to be unique only within the volume group. Therefore, you might choose to use the same name for more than one volume group. For example, you might want to use the site name to make it more meaningful. You can define the mirror pool for the local disks by using the **chpv** command:

   ```
   chvg -p Pougkeepsie hdisk10 hdisk11
   ```

6. Configure your existing logical volumes to belong to the mirror pool that you just created. You also need to turn off bad block relocation for each logical volume:

   ```
   chlv -m copy1=Pougkeepsie -b n dataloglv
   chlv -m copy1=Pougkeepsie n datafslv
   ```

7. Confirm that you have an ordinary volume group that is configured for super strict mirror pools and that all local disks at the production site belong to the Asite mirror pool:

   ```
   lsmp -A datavg
   ```

8. Define the RPV clients and RPV servers that are required for your cluster. On each node, define an RPV server for each local disk that belongs to the volume group, and define an RPV client for each remote disk that belongs to the volume group. Then, the RPV client/server pairs that you create enable LVM to access remote physical volumes as though they were ordinary local disks.

9. After you create the RPV clients and servers, add the remote physical volumes, which are identified by their ROV client names to the volume group. You can define the mirror pool for the remote physical volumes in the same step by using the **extendvg** command:

   ```
   extendvg -p Austin datavg hdisk31 hdisk32
   ```

   In this example command, the remote physical volumes hdisk31 and hdisk32 belong to a mirror pool named Austin.

10. After you add the remote physical volumes to your volume group, add mirror copies of your logical volumes to them. You can use the GLVM utilities in SMIT or the **mirrorvg** command:

    ```
    mirrorvg -c 2 -p copy2=Austin datavg
    ```

    In this example command, a second mirror copy of each logical volume is created on the disks that reside in the Austin mirror pool.

11. Because asynchronous mirroring requires a logical volume of type aio_cache to serve as the cache device, create this logical volume. Use the usual steps to create a logical volume, except specify `aio_cache` as the logical volume type. Also, the logical volume must be on the disks in the opposite site's mirror pool. You can perform this step with AIX LVM in the SMIT utilities or using the **mklv** command:

    ```
    mklv -y datacachelv1 -t aio_cache -p copy1=Pougkeepsie -b n datavg 100
    ```

    In this example command, the cache logical volume is in the Poughkeepsie mirror pool. It is used for caching during asynchronous mirroring to the disks in the Austin mirror pool.

12. Configure a mirror pool to use asynchronous mirroring. You can use GLVM utilities in SMIT or the **chmp** command:

    ```
    chmp -A -m Austin datavg
    ```

    In this example command, the Austin mirror pool is configured to use asynchronous mirroring. The **chmp** command automatically determines that the datacachelv1 logical volume is the cache device to use because it is resides in the opposite site's mirror pool.

13. Optional: Configure asynchronous mirroring for the Poughkeepsie mirror pool by creating a logical volume in the Austin mirror pool to serve as the cache device:

    ```
    mklv -y datacachelv2 -t aio_cache -p copy1=Austin -b n datavg 100
    ```

14. Configure the Poughkeepsie mirror pool to use asynchronous mirroring by using the same procedure that you followed for the Austin mirror pool:

    ```
    chmp -A -m Poughkeepsie datavg
    ```

15. Confirm that asynchronous mirroring is configured by running the **lsmp** command again:

    ```
    lsmp -A datavg
    ```

Now when the volume group is varied online at the Poughkeepsie site, the local disks in the Poughkeepsie mirror pool are updated synchronously, just as though they were an ordinary volume group. The remote disks in the Austin mirror pool are updated asynchronously using the cache device on the local Poughkeepsie disks. Likewise, when the volume group is varied online at the Austin site, the local disks in the Austin mirror pool are updated synchronously,. Also the remote disks in the Poughkeepsie mirror pool are updated asynchronously by using the cache device on the local Austin disks.

If you already have a GLVM volume group that is configured for synchronous mirroring, you might decide to reconfigure it for asynchronous mirroring, too. To reconfigure a GLVM volume group for asynchronous mirroring, see 8.10.1, “Converting synchronous GMVGs to asynchronous GMVGs” on page 412.

8.2.3 Integrating geographically mirrored volume groups into a PowerHA cluster

To set up a PowerHA Enterprise Edition for GLVM configuration, you can following several steps:

- Configure only two HACMP sites, one local and one remote site. For instructions about adding sites to the cluster, see “Configuring sites” in the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.
- Configure XD-type networks and any site-specific networks if planned.
- Prepare the geographically mirrored volume groups for HACMP cluster verification.
- Configure resource groups in PowerHA/XD for GLVM.
- Add the resource groups to the cluster.
- Verify and synchronize the GLVM configuration in PowerHA

For more information, see 8.4, “Configuring a 4-node, 2-site PowerHA for GLVM” on page 354.

8.3 Configuration wizard for GLVM

Configuring GLVM mirroring and PowerHA, which uses GLVM resources, can be a difficult process. By using the new configuration wizard, you can quickly and easily configure a two-site, two-node cluster that uses GLVM for data replication between sites.

The configuration wizard discovers most of the required information. Therefore, you need only to answer a few simple questions, and the wizard completes the configuration process. This feature is part of the package included with the PowerHA Enterprise Edition.

8.3.1 Prerequisites

Note the following prerequisites:

- Install the following file sets on the local node and remote node in addition to base PowerHA file sets:
  - `bos.rte` 6.1.2.0 or later
  - `cluster.xd.glvm` 6.1.0.0 or later
  - `glvm.rpv.client` 6.1.0.0 or later
  - `glvm.rpv.server` 6.1.0.0 or later

- The volume group must be varied on the local node and set to be the scalable type.
- The volume group must not already include any RPV disks.
- One or more logical volumes must be already defined.
- All IP network interfaces must be connected and configured.
- The application must be installed and configured.
- The application's service IP label must be added to `/etc/hosts` on all nodes.
- Persistent IP addresses are required if the XD_data network has multiple interfaces.

Note: Historically, PowerHA required a XD_data network with persistent IP for mirroring. From PowerHA 6.1, GLVM can now use boot IP for mirroring and the XD_data network can be configured with a single interface. Persistent IP is automatically configured if the local and remote nodes have preconfigured IP aliases and the user has not supplied explicit IP through SMIT.

8.3.2 Considerations

The GLVM wizard has a few considerations:

- The GLVM wizard cannot be used where a pre-existing GLVM configuration is incomplete or broken.
- The GLVM wizard requires one-to-one mapping of disks for mirroring. For example, for mirroring a 50-GB disk on the local node, the remote node must have a disk that is 50 GB or more.
- Although GLVM and PowerHA support IPv6, IPv6 cannot be configured directly with the wizard.
- The GLVM wizard creates a synchronous geographically mirrored volume group by default. If you want to create an asynchronous geographically mirrored volume group, you can convert the volume group after the cluster configuration. For more information, see 8.10.1, "Converting synchronous GMVGs to asynchronous GMVGs" on page 412.

8.3.3 Starting with the GLVM wizard

After all IP networks, volume groups, and applications are configured, you can start the GLVM wizard. To access the GLVM Cluster Configuration assist from the SMIT interface run the **`smitty hacmp`** command and select **Initialization and Standard Configuration** → **Configuration Assistants** → **GLVM Cluster Configuration Assistant** or use the short fastpath **`smitty cl_2siteglvm_configassist`**.

Figure 8-2 shows the GLVM Cluster Configuration Assistant Menu.

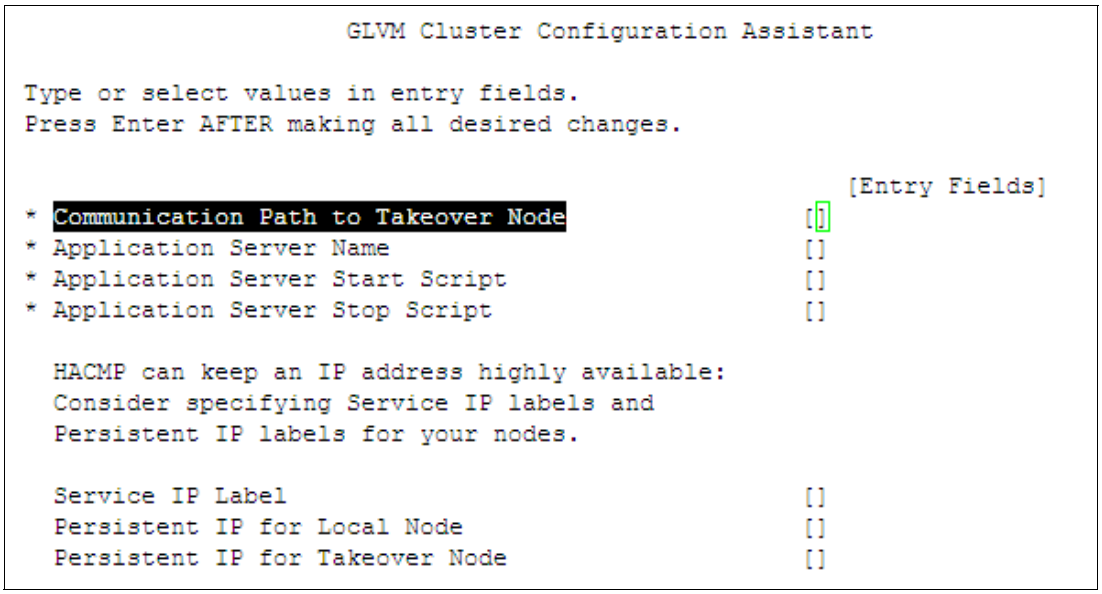


Figure 8-2 GLVM Cluster Configuration Assistant menu

The GLVM Cluster Configuration Assistant creates the following PowerHA configuration:

- Cluster: <*user supplied application name*>_cluster
- Nodes: one per site, use host name for node name
- Sites: siteA and siteB

  If the RPV server site name is already configured at each site, the GLVM wizard configures the PowerHA sites as the defined RPV server site name.
- XD_data network: single XD_data network

  An IP-Alias is enabled and it includes all inter-connected network interfaces.
- Application server
- One or more service IPs when provided
- One Resource Group: <*user supplied application name*>_group

  Inter-site management policy is set to prefer primary site.
- Persistent IP address for each node (optional for single interface networks).
- Create the RPV server, clients, and associated GMVGs with a PowerHA resource group. Geographically mirrored volume groups will be set as synchronous mirroring by default.

Note: After intermediate failure of the GLVM wizard, you can remove the PowerHA configuration, but the GMVG/VG configuration remains. Therefore, you manually remove the GLVM/VG configuration before performing a basic GLVM and PowerHA configuration with the GLVM Cluster Configuration Assistant.

All output from the GLVM Configuration Assistant is logged to `/var/hacmp/log/cl2siteconfig_assist.log` (default). The log file can be redirected by using standard SMIT panels smitty cm_hacmp_log_viewing_and_management_menu_dmn.

8.3.4 Configuring GLVM and PowerHA by using the GLVM wizard

To configure GLVM and PowerHA by using the GLVM wizard:

1. Configure the volume group and one or more logical volumes. The volume group must be of the scalable VG format type, and autovaryon is turned off (Example 8-1).

Example 8-1 Creating a volume group and logical volume

```
root@GLVM_A1 / > mkvg -S -n -y yurivg hdisk1
yurivg
root@GLVM_A1 / >   mklv -t jfs2 -y lvyuri yurivg 10
lvyuri
```

2. Configure all IP network interfaces and add them to the `/etc/hosts` file on all nodes (Example 8-2).

Example 8-2 /etc/hosts files and IP network interfaces configured

```
root@GLVM_A1 / > cat /etc/hosts
.....
#XD_Data
10.10.101.107 GLVM_A1 GLVM_A1_XD1
10.10.201.107 GLVM_B1 GLVM_B1_XD1
10.10.102.107 GLVM_A1_XD2
10.10.202.107 GLVM_B1_XD2

#Persistent Network
10.10.100.107 GLVM_A1_per
10.10.200.107 GLVM_B1_per

#Service Network
10.10.10.10 Yuri_svc

root@GLVM_A1 / > ifconfig -a
en1:
flags=1e080863,c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64B
IT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 10.10.101.107 netmask 0xffffff00 broadcast 10.10.101.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
en2:
flags=1e080863,c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64B
IT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 10.10.102.107 netmask 0xffffff00 broadcast 10.10.102.255
         tcp_sendspace 131072 tcp_recvspace 65536 rfc1323 0
```

3. Install the application server and its start and stop scripts (Example 8-3).

Example 8-3 Configured application server

```
Application server: YuriApp
Start script: /usr/es/sbin/cluster/yuri.start.sh
Stop script: /usr/es/sbin/cluster/yuri.stop.sh
```

4. Start the GLVM Cluster Configuration Assistant. Run **smitty cl_2siteglvm_configassist** and enter the information that you previously configured (Figure 8-3).

```
                    GLVM Cluster Configuration Assistant

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* Communication Path to Takeover Node                [GLVM_B1] +
* Application Server Name                            [YuriApp]
* Application Server Start Script [/usr/es/sbin/cluster/yuri.start.sh]
* Application Server Stop Script [/usr/es/sbin/cluster/yuri.stop.sh]

HACMP can keep an IP address highly available:
  Consider specifying Service IP labels and
  Persistent IP labels for your nodes.

  Service IP Label                                  [Yuri_svc] +
  Persistent IP for Local Node                      [GLVM_A1_per] +
  Persistent IP for Takeover Node                   [GLVM_B1_per] +
```

Figure 8-3 Example of GLVM Cluster Configuration Assistant

Figure 8-4 shows the result of the PowerHA configuration using the GLVM wizard with the following information:

- Cluster: YuriApp_cluster
- Nodes: GLVM_A1 and GLVM_B1
- Sites: siteA and siteB
- XD_data network: net_XD_data_01
- Application server: YuriApp
- Service IP address: Yuri_svc
- One resource group: YuriApp_group
- Persistent IP address for each node: GLVM_A1_XDT_per, GLVM_A2_XDT_per
- RPV server and client in each node that is configured and added to geographically mirrored volume group: yurivg
- Mirror pools for each site that is also created for geographically mirrored volume group: yurivg

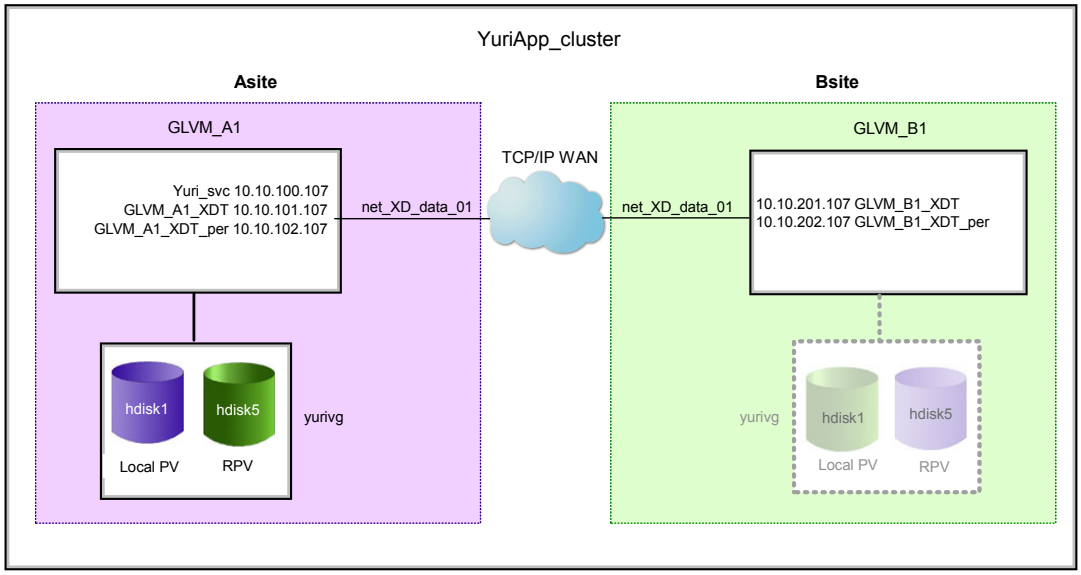


Figure 8-4 A GLVM and PowerHA configuration using GLVM wizard

8.4 Configuring a 4-node, 2-site PowerHA for GLVM

You can configure geographically mirrored volume groups and integrate them into a 4-node, 2-site cluster configuration.

In Figure 8-5, each site has a storage subsystem that is attached to the local nodes within each site. We use private communication lines for the XD_data type networks between sites. We define the XD_ip type network for the communication links between sites in the separate IP segments from XD_data networks for heartbeating. You might consider configuring non-IP networks, such as an XD_rs232 network.

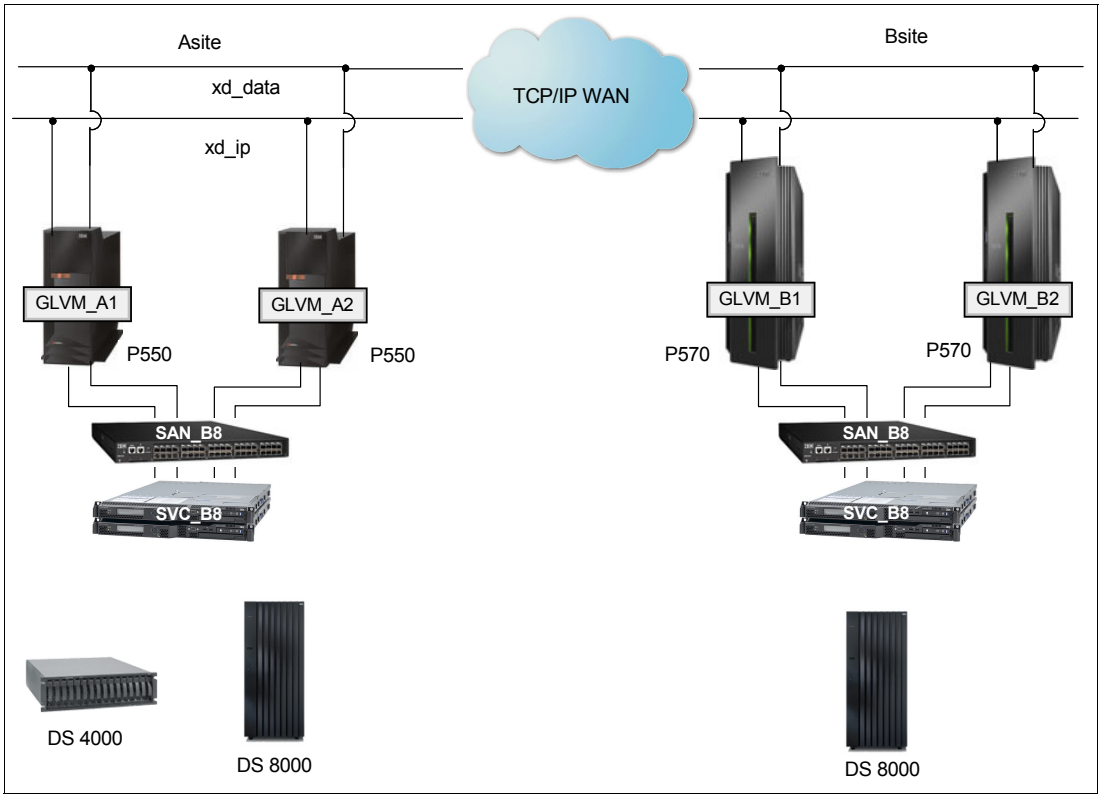


Figure 8-5 General overview of the PowerHA environment for GLVM

Although not shown in Figure 8-5, each site has its own local Ethernet type network, non-routed between sites for the client access to the cluster nodes. Have non-IP networks between the nodes in one site to prevent false failover.

For the TCP/IP WAN, we installed a Linux system and added a routing table to enable the network connection between nodes in each site. In addition, to establish latency between sites, we installed a WAN simulator.

In our scenario in Figure 8-6, we create two private communication lines for the XD_data networks between sites and two Ethernet interfaces on each node for connection to each private network. These networks are used for the geographically mirrored volume groups.

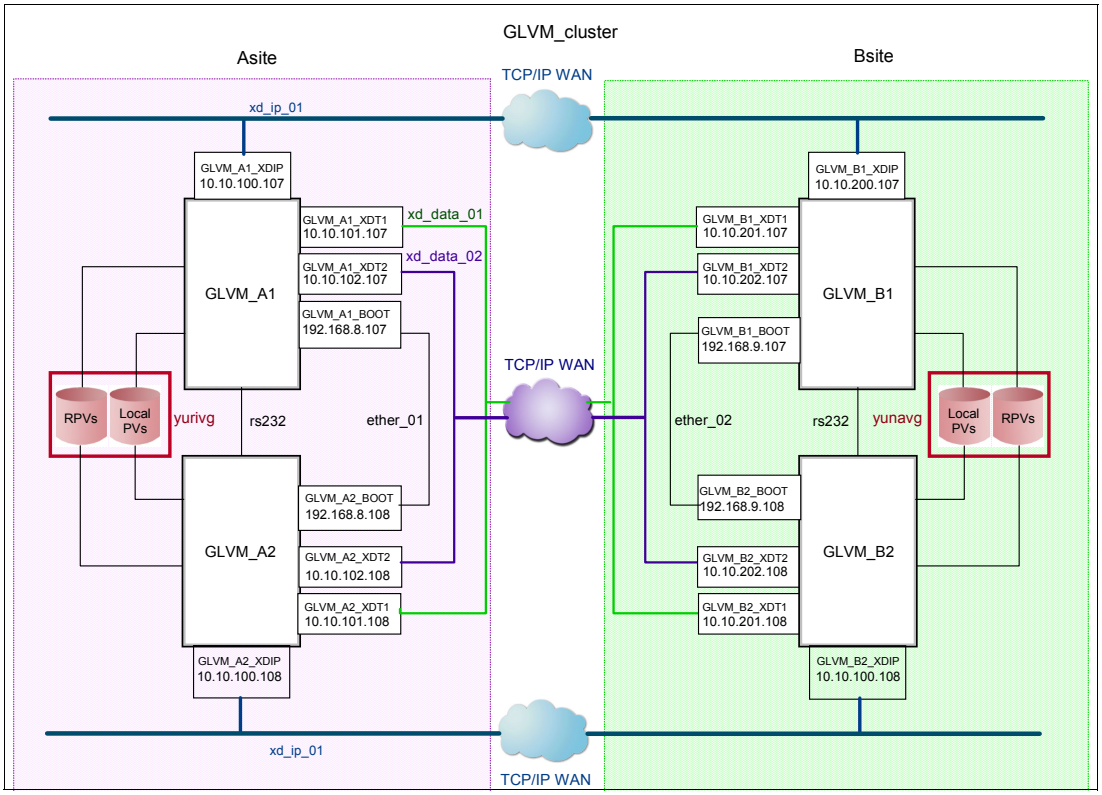


Figure 8-6 Four-node, 2-site PowerHA for GLVM cluster configuration

We define one XD_ip network for the communication links between sites in a separate IP segment from the XD_data networks. Each site has its own local network, non-routed between sites for the client access to the cluster nodes. Although we did not configure non-IP networks within a site, in a production environment, have non-IP networks. For more information, see the *HACMP for AIX: Planning and Administration Guide*, SC23-4862.

We create first a resource group primarily at Asite that contains a GMVG, yurivg. Later, we add a second resource group to the cluster configuration, primarily at Bsite with another GMVG, yunavg. The cluster has the following information:

- Cluster name: GLVM_cluster
- Site and participating nodes:
  - Asite: GLVM_A1 and GLVM_A2
  - Bsite: GLVM_B1 and GLVM_B2
- Topology
  - net_XD_data_01: GLVM_A1_XDT1, GLVM_A2_XDT1, GLVM_B1_XDT1, and GLVM_B2_XDT1
  - net_XD_data_02: GLVM_A1_XDT2, GLVM_A2_XDT2, GLVM_B1_XDT2, and GLVM_B2_XDT2
  - net_XD_ip_01: GLVM_A1_XDIP, GLVM_A2_XDIP, GLVM_B1_XDIP, and GLVM_B2_XDIP

– net_ether_01: GLVM_A1_boot, and GLVM_A2_boot
– net_ether_02: GLVM_B1_boot and GLVM_B2_boot

Table 8-2 lists the resource groups and their resources.

Table 8-2 Resource group in 4-node, 2-site PowerHA for GLVM cluster configuration

Resource group name	Application server	Service IP	Volume group
GLVM_A1_RG		GLVM_A1	
GLVM_A2_RG		GLVM_A2	
GLVM_B1_RG		GLVM_B1	
GLVM_B2_RG		GLVM_B2	
YuriRG	YuriApp		yurivg
YunaRG	YunaApp		yunavg

8.4.1 Configuring a geographically mirrored volume group

In this step, you configure geographically mirrored volume groups, their corresponding logical volumes, and RPVs. Figure 8-7 shows the configuration of a geographically mirrored volume group, yurivg, primarily on Asite. yurivg consists of the local physical volumes hdisk1 and hdisk2 and remote physical volumes hdisk5 and hdisk6.

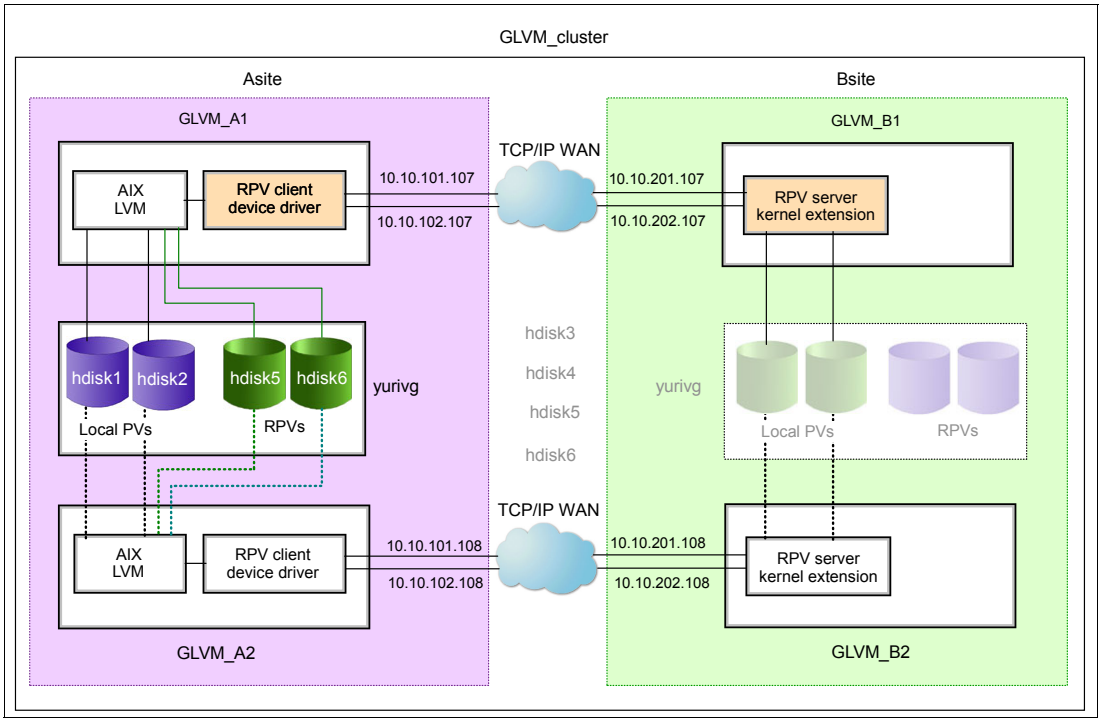


Figure 8-7 Remote physical volume configuration for yurivg activated in GLVM_A1 (Asite)

To configure a geographically mirrored volume group:

1. Configure the volume group yurivg and the logical volume lvyuri on the GLVM_A1 node at the local site (Example 8-4).

   The volume group must be of type scalable VG and autovaryon is turned off. It is a good idea for the logical volume to turn off bad block relocation and set superstrict for the geographically mirrored volume groups.

   Example 8-4 Create a volume group, logical volume, and file system

```
root@GLVM_A1 / > mkvg -S -n -y yurivg -V 50 hdisk1 hdisk2
yurivg
root@GLVM_A1 / > mklv -t jfs2 -b n -s s -u 1024 -y lvyuri yurivg  10
lvyuri
root@GLVM_A1 / > crfs -v jfs2 -d lvyuri -m /yuri -A no
File system created successfully.
163628 kilobytes total disk space.
New File System size is 327680
root@GLVM_A1 / > chlv -s s -u  1024 -b n  loglv00
```

2. Configure an RPV server site name Bsite on the GLVM_B1 and GLVM_B2 nodes at the remote site (Figure 8-8). Run the **smitty rpvserver** command. Then, select **Remote Physical Volume Server Site Name Configuration → Remote Physical Volume Server Site Name Configuration → Define / Change / Show Remote Physical Volume Server Site Name**.

```
        Define / Change / Show Remote Physical Volume Server Site Name

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* Remote Physical Volume Server Site Name            [Bsite]
```

   Figure 8-8 Define RPV server site name on Bsite

3. Check the available physical volumes and configure RPV servers. Run the **smitty rpvserver** command and select **Remote Physical Volume Server Site Name Configuration → Add Remote Physical Volume Servers**.

   If you have multiple RPV client IP addresses, place the IP addresses, separated by commas, in the IP address of the remote physical volume client. For example, the RPV client IP address is 10.10.101.107, 10.10.102.107 for GLVM_B1 node.

   Specify that the RPV server start immediately, which indicates that the RPV server is in the available state. Set the Configure Automatically at System Restart field to No (Example 8-5 and Figure 8-9 on page 358).

   Example 8-5 RPV server configuration in GLVM_B1 node for yurivg

```
root@GLVM_B1 / > lspv
hdisk0          00c1f170e170c9cd                    rootvg          active
hdisk1          00c1f1702fab9d25                    None
hdisk2          00c1f1702fab9da9                    None
hdisk3          00c0f6a02fae3172                    None
hdisk4          00c0f6a02fae31dd                    None
```

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
Physical Volume Identifiers                    00c0f6a02fae3172
00c0f6a02fae31dd
* Remote Physical Volume Client Internet Address
[10.10.101.107,10.10.102.107] +
  Configure Automatically at System Restart?          [no] +
  Start New Devices Immediately?                      [yes] +
```

Figure 8-9 Add Remote Physical Volume Servers panel

4. Configure RPV clients on GLVM_A1, which creates RPV clients hdisk5 and hdisk6 on the local node. Run the **smitty rpvclient** command. Select **Add Remote Physical Volume Clients**. Then, enter the remote physical volume server internet address of GLVM_B1 and the remote physical volume local internet address of GLVM_A1.

   Specify that the RPV client start immediately, which indicates that the RPV client in the available state (Figure 8-10 and Example 8-6).

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                      [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.201.107
  Remote Physical Volume Local Internet Address       10.10.101.107
  Physical Volume Identifiers
00c0f6a02fae3172 0000000000000000 00c0f6a02fae31dd 0000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +
```

Figure 8-10 Add Remote Physical Volume Clients panel

Example 8-6 RPV client configuration in GLVM_A1 node for yurivg

```
root@GLVM_A1 / > lsdev -Cc disk
hdisk0 Available          Virtual SCSI Disk Drive
hdisk1 Available 32-T1-01 MPIO FC 2145
hdisk2 Available 32-T1-01 MPIO FC 2145
hdisk3 Available 22-T1-01 MPIO FC 2145
hdisk4 Available 22-T1-01 MPIO FC 2145
hdisk5 Available          Remote Physical Volume Client
hdisk6 Available          Remote Physical Volume Client
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg          active
hdisk2          000fe4112f998235                    yurivg          active
hdisk3          000fe4012f9a9f43                    None
hdisk4          000fe4012f9a9fcf                    None
hdisk5          00c0f6a02fae3172                    None
hdisk6          00c0f6a02fae31dd                    None
```

Multiple RPV server-client networks: If there are multiple RPV server-client networks (Figure 8-11 and Example 8-7), add them through GLVM utilities or by using the **chdev** command. Run the **smitty glvm_utils** command, and select **Remote Physical Volume Clients** → **Change Multiple Remote Physical Volume Clients**.

```
/usr/sbin/chdev -l hdiskX -a server_addr='server_ip -a local_addr=local_ip
-a server_addr2=server_ip2 -a local_addr2=local_ip2
```

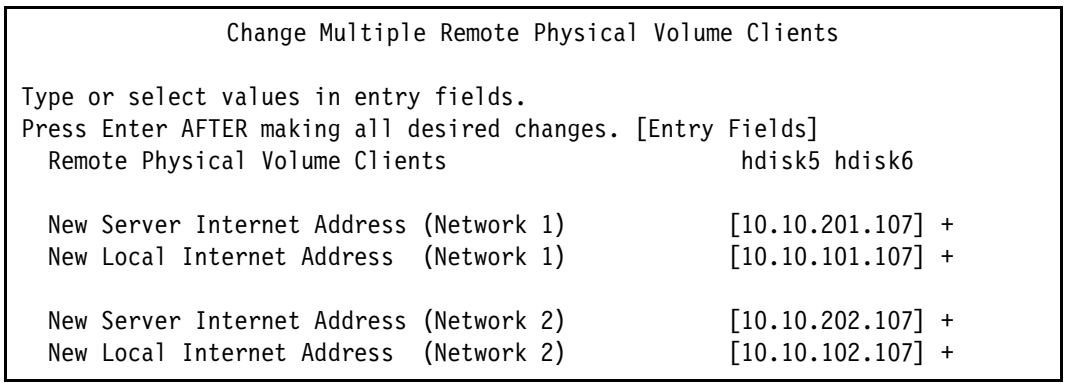


Figure 8-11 Change Multiple Remote Physical Volume Clients

Example 8-7 Configuring multiple networks on RPV clients using the chdev command

```
root@GLVM_A1 / > chdev -l hdisk5 -a local_addr=10.10.101.107 -a
local_addr2=10.10.102.107 -a server_addr=10.10.201.107 -a
server_addr2=10.10.202.107
root@GLVM_A1 / > chdev -l hdisk6 -a local_addr=10.10.101.107 -a
local_addr2=10.10.102.107 -a server_addr=10.10.201.107 -a
server_addr2=10.10.202.107
```

5. As performed in step 3, configure RPV servers on GLVM_B2 (Example 8-8 and Figure 8-12 on page 360).

Example 8-8 Configuring RPV servers on GLVM_B2 node

```
root@GLVM_B2 / > lspv
hdisk0          00c0f6a0684f5ab8                    rootvg          active
hdisk1          00c1f1702fab9d25                    None
hdisk2          00c1f1702fab9da9                    None
hdisk3          00c0f6a02fae3172                    None
hdisk4          00c0f6a02fae31dd                    None
hdisk9          00c0f6a064f9deea                    None
```

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Physical Volume Identifiers               00c0f6a02fae3172
00c0f6a02fae31dd
* Remote Physical Volume Client Internet Address
[10.10.101.108,10.10.102.108] +
  Configure Automatically at System Restart?         [no] +
  Start New Devices Immediately?                     [yes] +
```

Figure 8-12 Configuring RPV servers on GLVM_B2 node

6. As performed in step 4, configure RPV clients on GLVM_A2 (Figure 8-13).

```
                    Add Remote Physical Volume Clients

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

                                                      [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.201.108
  Remote Physical Volume Local Internet Address       10.10.101.108
  Physical Volume Identifiers
00c0f6a02fae31720000000000000000 00c0f6a02fae31dd0000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

Change Multiple Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
  Remote Physical Volume Clients                      hdisk5 hdisk6
  New Server Internet Address (Network 1)            [10.10.201.108] +
  New Local Internet Address  (Network 1)            [10.10.101.108] +

  New Server Internet Address (Network 2)            [10.10.202.108] +
  New Local Internet Address  (Network 2)            [10.10.102.108] +
```

Figure 8-13 Configure RPV clients on GLVM_A2 node and add more networks

7. Add remote physical volumes to the volume group (Figure 8-14). You can use GLVM_utilities or the **extendvg** command. Run the **smitty glvm_utils** command and select **Geographically Mirrored Volume Groups → Manage Legacy Geographically Mirrored Volume Groups → Add Remote Physical Volumes to a Volume Group**.

```
                Add Remote Physical Volumes to a Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* VOLUME GROUP name                                      yurivg
  Force                                                 [no] +
* REMOTE PHYSICAL VOLUMES Name                          hdisk5 hdisk6
```

Figure 8-14 Add RPVs on the volume group

Alternatively, enter the following command:

```
root@GLVM_A1 / > extendvg  yurivg hdisk5 hdisk6
```

8. Add a remote site mirror copy to the volume group (Figure 8-15). You can use GLVM_utilities or the **mirrorvg** command. Run the **smitty glvm_utils** command and select **Geographically Mirrored Volume Groups → Manage Legacy Geographically Mirrored Volume Groups → Add a Remote Site Mirror Copy to a Volume Group**.

```
               Add a Remote Site Mirror Copy to a Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* VOLUME GROUP name                                      yurivg
  Mirror Sync Mode                                     [Foreground] +
* REMOTE PHYSICAL VOLUME names                          hdisk5 hdisk6
  Number of COPIES of each logical partition            2 +
  Keep Quorum Checking On?                              no +
```

Figure 8-15 Add a remote site mirror copy to the yurivg volume group

9. When volume group mirroring is finished, vary off the volume group on the GLVM_A1 node and import it to the GLVM_A2 node (Example 8-9). For the shared volume group, have the same volume group major number within the cluster. For more information, see *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

Example 8-9 Import volume group yurivg in GLVM_A2

```
root@GLVM_A2 / > importvg -y yurivg -V 50 000fe4112f99817c
yurivg
```

8.4.2 Extending the geographically mirrored standard volume group to other nodes in the cluster

You must configure RPV servers and clients at both sites for the mirroring function to work both ways in the cluster. Figure 8-16 on page 362 shows the remote physical volume configuration of yurivg is varied on GLVM_B1, enabling the mirroring function to work in both directions.

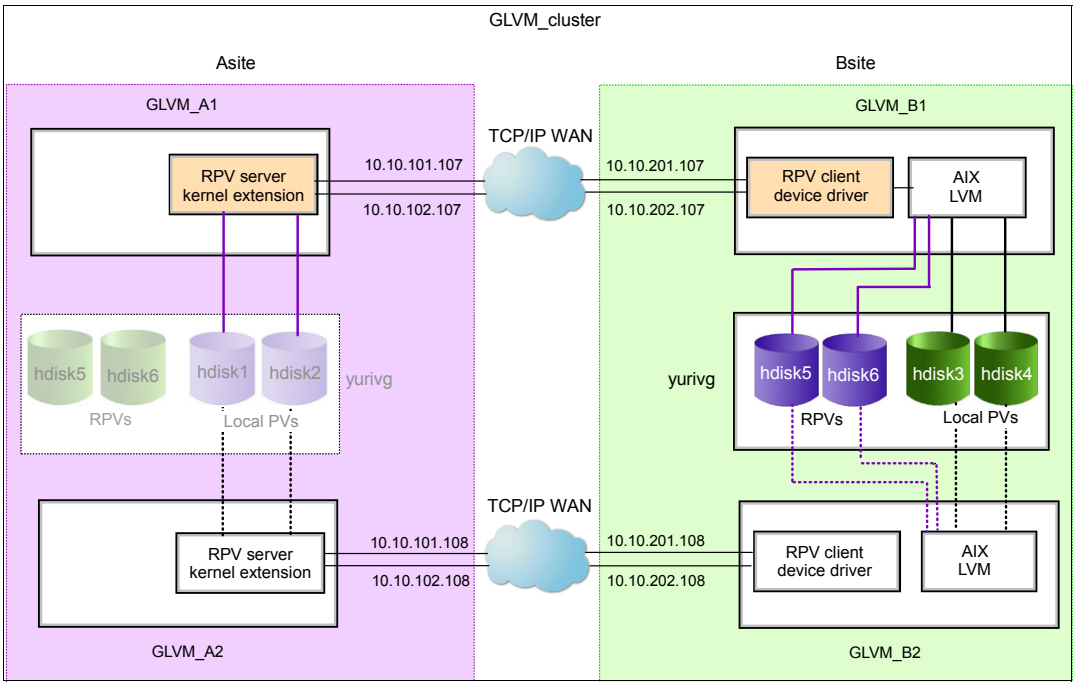


Figure 8-16 Remote physical volume configuration for yurivg activated in GLVM_B1(Bsite)

To configure RPV servers and clients at both sites for the mirroring function to work both ways in the cluster:

1. Configure RPV servers in the defined state (Figure 8-17) on the GLVM_B1 and GLVM_B2 nodes at the remote site by using GLVM utilities or the **rmdev** command. Run the **smitty rpvserver** command, and select **Remove Remote Physical Volume Servers**.

```
                    Remove Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Servers                        rpvserver0 rpvserver1
  Keep definitions in database?                        [yes] +

or

root@GLVM_B1 / > rmdev -l rpvserver0
rpvserver0 Defined
root@GLVM_B1 / > rmdev -l rpvserver1
rpvserver1 Defined

root@GLVM_B2 / > rmdev -l rpvserver0
rpvserver0 Defined
root@GLVM_B2 / > rmdev -l rpvserver1
rpvserver1 Defined
```

Figure 8-17 Configure RPV servers in the defined state

2. Configure RPV clients in the defined state on the GLVM_A1 and GLVM_A2 nodes at the local site by using GLVM utilities or the **rmdev** command (Figure 8-18). Run the **smitty rpvclient** command, and select **Remove Remote Physical Volume Clients**.

```
                    Remove Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Clients                       hdisk5 hdisk6
  Keep definitions in database?                       [yes] +

or

root@GLVM_A1 / > rmdev -l hdisk5
hdisk5 Defined
root@GLVM_A1 / > rmdev -l hdisk6
hdisk6 Defined

root@GLVM_A2 / > rmdev -l hdisk5
hdisk5 Defined
root@GLVM_A2 / > rmdev -l hdisk6
hdisk6 Defined
```

Figure 8-18 Configure RPV clients in the defined state

3. Configure an RPV server site named Asite on the GLVM_A1 and GLVM_A2 nodes at the local site (Figure 8-19).

```
        Define / Change / Show Remote Physical Volume Server Site Name

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* Remote Physical Volume Server Site Name            [Asite]
```

Figure 8-19 Define RPV server site name on Asite

4. Configure RPV servers on GLVM_A1 and GLVM_A2 at the local site (Example 8-10, Figure 8-20 on page 364, and Figure 8-21 on page 364).

Example 8-10 Configuring RPV servers on GLVM_A1

```
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg
hdisk2          000fe4112f998235                    yurivg
hdisk3          000fe4012f9a9f43                    None
hdisk4          000fe4012f9a9fcf                    None
```

```
                    Add Remote Physical Volume Servers

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

                                                        [Entry Fields]
  Physical Volume Identifiers                 000fe4112f99817c
000fe4112f998235
* Remote Physical Volume Client Internet Address
[10.10.201.107,10.10.202.107] +
  Configure Automatically at System Restart?          [no] +
  Start New Devices Immediately?                      [yes] +
```

Figure 8-20 Configuring RPV servers on GLVM_A1

```
                      Add Remote Physical Volume Servers

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

                                                  [Entry Fields]
  Physical Volume Identifiers                  000fe4112f99817c 000fe4112f998235
* Remote Physical Volume Client Internet Address [10.10.201.107,10.10.202.107] +
  Configure Automatically at System Restart?       [no] +
  Start New Devices Immediately?                   [yes] +
```

Figure 8-21 Configuring RPV servers on GLVM_A2

5. Configure RPV clients on the GLVM_B1 and GLVM_B2 nodes at the remote site and configure multiple networks on RPV clients by using the **chdev** command or the GLVM_utilities (Figure 8-22 and Figure 8-23 on page 365).

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.101.107
  Remote Physical Volume Local Internet Address       10.10.201.107
  Physical Volume Identifiers
000fe4112f99817c0000000000000000 000fe4112f9982350000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

root@GLVM_B1 / > chdev -l hdisk5 -a local_addr=10.10.201.107 -a
local_addr2=10.10.202.107 -a server_addr=10.10.101.107 -a
server_addr2=10.10.102.107
root@GLVM_B1 / > chdev -l hdisk6 -a local_addr=10.10.201.107 -a
local_addr2=10.10.202.107 -a server_addr=10.10.101.107 -a
server_addr2=10.10.102.107
```

Figure 8-22 Configure RPV clients on GLVM_B1

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.101.108
  Remote Physical Volume Local Internet Address       10.10.201.108
  Physical Volume Identifiers
000fe4112f99817c0000000000000000 000fe4112f998235000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +


root@GLVM_B2 / > chdev -l hdisk5 -a local_addr=10.10.201.108 -a
local_addr2=10.10.202.108 -a server_addr=10.10.101.108 -a
server_addr2=10.10.102.108
root@GLVM_B2 / > chdev -l hdisk6 -a local_addr=10.10.201.108 -a
local_addr2=10.10.202.108 -a server_addr=10.10.101.108 -a
server_addr2=10.10.102.108
```

Figure 8-23 Configure RPV clients on GLVM_B2

6. Vary off the volume group at the local site (Asite) and update the volume groups' definitions on the nodes at remove site (Bsite) by running the **importvg** command (Example 8-11).

Example 8-11 Import volume group yurivg in Bsite

```
root@GLVM_B1 / > importvg -y yurivg -V 50 000fe4112f99817c
yurivg
root@GLVM_B1 / > varyoffvg yurivg

root@GLVM_B2 / > importvg -y yurivg -V 50 000fe4112f99817c
yurivg
root@GLVM_B2 / > varyoffvg yurivg
```

We have another volume group, yunavg, at the remote site (Bsite) to mutually take over the local site (Asite) in case of site failure (Figure 8-24). To configure the volume group as a geographically mirrored volume group, go back to 8.2.2, “Configuring geographically mirrored volume groups” on page 344, and repeat step 1 and step 2.

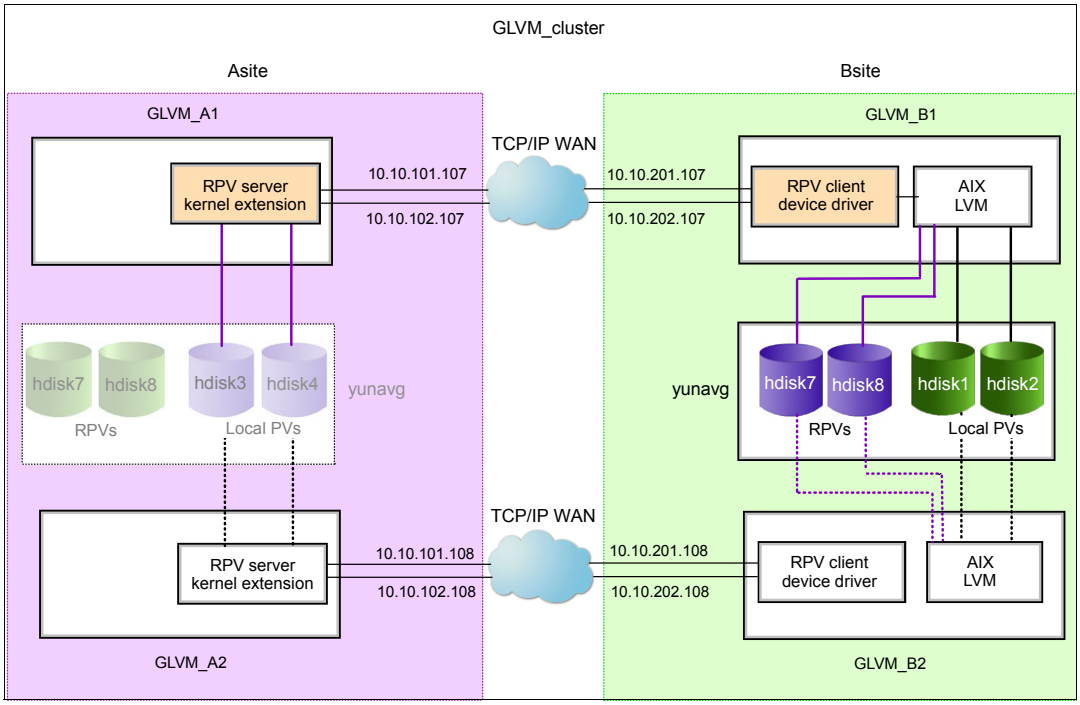


Figure 8-24 Remote physical volume configuration for yunavg activated in GLVM_B1 (Bsite)

8.4.3 Configuring the cluster, the nodes, and the sites

For this procedure, you must be familiar with configuring PowerHA. For more information, see the *HACMP for AIX 6.1 Planning Guide*, SC23-4861, and the *HACMP for AIX 6.1 Administration Guide*, SC23-4862.

To configure the cluster, nodes, and sites:

1. Configure the cluster and nodes:
   - Cluster Name: GLVM_cluster
   - Nodes: GLVM_A1, GLVM_A2, GLVM_B1, GLVM_B2
2. Configure the sites. Site names must match the site names defined for RPVs (Figure 8-25). Run the **`smitty hacmp`** command and select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Sites**.

```
                                Add Site

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
* Site Name                                            [Asite] +
* Site Nodes                                          GLVM_A1 GLVM_A2 +
```

Figure 8-25 Site configuration

8.4.4 Configuring XD-type networks and communication interfaces

The following steps explain how to configure two XD_data networks and one XD_ip network. It is assumed that all IP addresses have been configured and that an entry has been added in the `/etc/hosts` file for all nodes (Example 8-12).

Example 8-12 XD-type network entries on /etc/hosts file

```
#XD_IP Network
10.10.100.107 GLVM_A1_XDIP
10.10.100.108 GLVM_A2_XDIP
10.10.200.107 GLVM_B1_XDIP
10.10.200.108 GLVM_B2_XDIP

#XD_IP Data1
10.10.101.107 GLVM_A1_XDT1
10.10.101.108 GLVM_A2_XDT1
10.10.201.107 GLVM_B1_XDT1
10.10.201.108 GLVM_B2_XDT1

#XD_IP Data2
10.10.102.107 GLVM_A1_XDT2
10.10.102.108 GLVM_A2_XDT2
10.10.202.107 GLVM_B1_XDT2
10.10.202.108 GLVM_B2_XDT2
```

Configuring the XD-type networks

To configure XD-type networks:

1. Run the **`smitty hacmp`** command, select **Extended Configuration** → **Extended Topology Configuration** → **Configure HACMP Networks** → **Add a Network to the HACMP Cluster**, and press Enter.
2. From the predefined IP-base Network Types list, select the **XD_data** network type and add it to the PowerHA cluster (Figure 8-26). Make sure that you have created two XD_data networks.

```
                Add an IP-Based Network to the HACMP Cluster

Type or select values in entry fields.
Press Enter AFTER making all desired changes.
                                                      [Entry Fields]
* Network Name                                       [net_XD_data_02]
* Network Type                                        XD_data
* Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.255.0]
* Enable IP Address Takeover via IP Aliases          [Yes] +
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 8-26 Adding a second XD_data network for the cluster

3. Go to step 2, select the **XD_ip** network type, and enter the information (Figure 8-27).

```
                Add an IP-Based Network to the HACMP Cluster

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

                                                        [Entry Fields]
* Network Name                                         [net_XD_ip_01]
* Network Type                                          XD_ip
* Netmask(IPv4)/Prefix Length(IPv6)                    [255.255.255.0]
* Enable IP Address Takeover via IP Aliases            [Yes] +
  IP Address Offset for Heartbeating over IP Aliases   []
```

Figure 8-27 Adding a XD_ip network for the cluster

Adding communication interfaces or devices for XD-type networks

To add a communication interface or devices for XD-type networks:

1. Run the **smitty hacmp** command and select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices**.
2. Select **Add Pre-defined Communication Interfaces and Devices, communication interfaces appropriate networks net_XD_data_01** (Figure 8-28 on page 368).

```
                        Add a Communication Interface

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* IP Label/Address                                    [GLVM_A1_XDT] +
* Network Type                                         XD_data
* Network Name                                         net_XD_data_01
* Node Name                                           [GLVM_A1] +
  Network Interface                                   []
```

Figure 8-28 Add a communication interface to the defined net_XD_data_01 network

3. Repeat these steps for all nodes and for each network. You can verify them by using SMIT. Select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Change/Show Communication Interfaces/Devices**. Alternatively, use the **cllsif** cluster utility (Example 8-13).

Example 8-13 Configured XD-type networks on the GLVM_cluster

```
root@GLVM_A1 / > cllsif
Adapter              Type       Network    Net Type   Attribute  Node       IP Address       Hardware
Address Interface Name   Global Name      Netmask          Alias for HB Prefix Length
GLVM_A1_XDT1        boot      net_XD_data_01 XD_data    public     GLVM_A1    10.10.101.107
en1                         255.255.255.0               24
GLVM_A1_XDT2        boot      net_XD_data_02 XD_data    public     GLVM_A1    10.10.102.107
en2                         255.255.255.0               24
GLVM_A1_XDIP        boot       net_XD_ip_01 XD_ip      public     GLVM_A1    10.10.100.107
en0                         255.255.255.0               24
GLVM_A2_XDT1        boot      net_XD_data_01 XD_data    public     GLVM_A2    10.10.101.108
en1                         255.255.255.0               24
```

```
GLVM_A2_XDT2        boot      net_XD_data_02 XD_data    public     GLVM_A2    10.10.102.108
en3                           255.255.255.0               24
GLVM_A2_XDIP        boot       net_XD_ip_01 XD_ip      public     GLVM_A2    10.10.100.108
en0                           255.255.255.0               24
GLVM_B1_XDT1        boot      net_XD_data_01 XD_data    public     GLVM_B1    10.10.201.107
en2                           255.255.255.0               24
GLVM_B1_XDT2        boot      net_XD_data_02 XD_data    public     GLVM_B1    10.10.202.107
en3                           255.255.255.0               24
GLVM_B1_XDIP        boot       net_XD_ip_01 XD_ip      public     GLVM_B1    10.10.200.107
en1                           255.255.255.0               24
GLVM_B2_XDT1        boot      net_XD_data_01 XD_data    public     GLVM_B2    10.10.201.108
en2                           255.255.255.0               24
GLVM_B2_XDT2        boot      net_XD_data_02 XD_data    public     GLVM_B2    10.10.202.108
en3                           255.255.255.0               24
GLVM_B2_XDIP        boot       net_XD_ip_01 XD_ip      public     GLVM_B2    10.10.200.108
en1                           255.255.255.0               24
```

8.5 Configuring site-specific networks

Next you configure site-specific networks that are net_ether_01 for GLVM_A1 and GLVM_A2 and net_ether_02 for GLVM_B1 and GLVM_B2, and site-specific service IP and persistent IP addresses. Service IP addresses are site-specific, which means that they do not take over across sites, but can take over to the node within the site.

8.5.1 Configuring the ether-type networks

To configure the ether-type networks:

1. Run the **smitty hacmp** command and select **Extended Configuration → Extended Topology Configuration → Configure HACMP Networks → Add a Network to the HACMP Cluster**.
2. From the Predefined IP-base Network Types list, select the **ether** network type and add it to the PowerHA cluster (Figure 8-29). Create two ether-type networks for both sites.

```
                Add an IP-Based Network to the HACMP Cluster

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* Network Name                                      [net_ether_01]
* Network Type                                       ether
* Netmask(IPv4)/Prefix Length(IPv6)                 [255.255.255.0]
* Enable IP Address Takeover via IP Aliases         [Yes] +
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 8-29 Adding an ether-type network

Adding the communication interfaces or devices for the ether networks

To add the communication interfaces or devices for the defined ether networks:

1. Run **smitty hacmp** and select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices**.

2. Select **Add Pre-defined Communication Interfaces and Devices, communication interfaces appropriate networks net_ether_01**.
3. Repeat these steps for all nodes and for each network interface. You can verify by running the **smitty hacmp** command. Select **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Change/Show Communication Interfaces/Devices**. Alternatively, use the **cllsif** cluster utility (Example 8-14).

Example 8-14 Configured ether-type networks on the GLVM_cluster

```
root@GLVM_A1 > cllsif | grep ether
GLVM_A1_boot   boot    net_ether_01 ether  public   GLVM_A1    192.168.8.107    en3     255.255.255.0    24
GLVM_A2_boot   boot    net_ether_01 ether  public   GLVM_A2    192.168.8.108    en2     255.255.255.0    24
GLVM_B1_boot   boot    net_ether_02 ether  public   GLVM_B1    192.168.9.107    en0     255.255.255.0    24
GLVM_B2_boot   boot    net_ether_02 ether  public   GLVM_B2    192.168.9.108    en0     255.255.255.0    24
```

8.5.2 Configuring the persistent IP addresses for each node

To configure the persistent IP addresses for each node:

1. Run **smitty hacmp** and select **Extended Configuration → Extended Topology Configuration → Configure HACMP Persistent Node IP Label/Addresses> Add a Persistent Node IP Label/Address**.
2. Select the node, enter the information that is shown in Figure 8-30, and press Enter.

```
                    Add a Persistent Node IP Label/Address

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* Node Name                                          GLVM_A1
* Network Name                                      [net_ether_01] +
* Node IP Label/Address                             [GLVM_A1_pers] +
  Netmask(IPv4)/Prefix Length(IPv6)                 []
```

Figure 8-30 Add a persistent node IP label/address for GLVM_A1

3. Repeat these steps to add a persistent IP label/address to each node.

Configuring the service IP addresses for each node

To configure:

1. Run **smitty hacmp**. Select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure HACMP Service IP Labels/Addresses → Add a Service IP Label/Address**.
2. Select **Configurable on Multiple Nodes** and **ether-type network**. The associated site is set to ignore (Figure 8-31 on page 371) because the service IP addresses are site-specific, which means that they do not take over across sites but can take over to the node within the site.

```
   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. [Entry Fields]
* IP Label/Address                                    GLVM_A1+
  Netmask(IPv4)/Prefix Length(IPv6)                    []
* Network Name                                         net_ether_01
  Alternate Hardware Address to accompany IP Label/A []
  ddress
  Associated Site                                      ignore +
```

Figure 8-31 Add a service IP label/address for GLVM_A1

3. Repeat these steps to add a service IP label/address for each node.

8.5.3 Configuring resource groups in PowerHA for GLVM

Before you complete the following steps, the geographically mirrored volume groups must be configured and imported in every node at each site. Also, you must have configured the application servers.

Before you integrate geographically mirrored volume groups into resource groups, check that all volume groups are varied off and that the RPV servers and clients are all in the defined state. For more information, see the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

To configure the resource group:

1. Run **`smitty hacmp`**. Select **Extended Configuration** → **Extended Topology Configuration** → **Add a Resource Group (extended)** (Figure 8-32).

```
                        Add a Resource Group (extended)

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
* Resource Group Name                                [GLVM_A1_RG]
  Inter-Site Management Policy                       [ignore] +
* Participating Nodes from Primary Site              [GLVM_A1 GLVM_A2] +
  Participating Nodes from Secondary Site            [] +

  Startup Policy                                      Online On Home Node Only +
  Failover Policy                     Failover To Next PriorityNode In The List +
  Failback Policy                  Failback To Higher PriorityNode In The List +
```

Figure 8-32 Adding a resource group

2. Repeat step 1 to add more resource groups (Table 8-3 on page 372). YuriRG and YunaRG are the resource groups that take over the remote site in case of site failure. You must set the inter-site policy and select the participating nodes from the secondary site.

Table 8-3 Resource groups names and roles

Resource group name	Intersite policy	Participating nodes from primary site	Participating nodes from secondary site
GLVM_A1_RG	Ignore	GLVM_A1 GLVM_A2	NA
GLVM_A2_RG	Ignore	GLVM_A2 GLVM_A1	NA
GLVM_B1_RG	Ignore	GLVM_B1 GLVM_B2	NA
GLVM_B2_RG	Ignore	GLVM_B2 GLVM_B1	NA
YuriRG	Prefer primary site	GLVM_A1 GLVM_A2	GLVM_B1 GLVM_B2
YunaRG	Prefer primary site	GLVM_B1 GLVM_B2	GLVM_A1 GLVM_A2

For all resource groups, the startup policy is Online On Home Node Only, the failover policy is Fall over To Next Priority Node, and the failback policy is Failback To Higher Priority Node.

The inter-site policy for the resource groups that contains the geographically mirrored volume group is set to prefer primary site. It might not be set to ignore or to online on both sites.

Add a resource into the resource groups and integrate the geographically mirrored volume group into the PowerHA resource group:

1. Run `smitty hacmp`. Select **Extended Configuration → Extended Topology Configuration → Configure HACMP Persistent Node IP Label/Addresses> Change/Show Resources and Attributes for a Resource Group**.
2. Select the resource group and enter the information that is listed in Table 8-4.

Table 8-4 List of resource group names mapped to application servers and service IP addresses

Resource group name	Application server	Service IP	Volume group
GLVM_A1_RG		GLVM_A1	
GLVM_A2_RG		GLVM_A2	
GLVM_B1_RG		GLVM_B1	
GLVM_B2_RG		GLVM_B2	
YuriRG	YuriApp		yurivg
YunaRG	YunaApp		yunavg

3. Set the “Use forced varyon of volume groups, if necessary” field to `true` (Figure 8-33), although the default is false. This setting allows PowerHA Enterprise Edition for GLVM to vary on the volume group if the quorum is disabled and the remote site fails.

```
            Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes [Entry Fields]
  Resource Group Name                                  YuriRG
  Inter-site Management Policy                          Prefer Primary Site
  Participating Nodes from Primary Site                  GLVM_A1 GLVM_A2
  Participating Nodes from Secondary Site               GLVM_B1 GLVM_B2
  Startup Policy                                       Online On Home Node Only
  Failover Policy                        Failover To Next Priority Node In The List
  Failback Policy                        Failback To Higher Priority Node In The
List
  Failback Timer Policy (empty is immediate)              [] +

  Service IP Labels/Addresses                            [] +
  Application Servers                                     [YuriApp] +
  Volume Groups                                         [yurivg ] +
  Use forced varyon of volume groups, if necessary          true +
  Automatically Import Volume Groups                       false +
  Allow varyon with missing data updates?                   true +
      (Asynchronous GLVM Mirroring Only)
  Default choice for data divergence recovery                [] +
      (Asynchronous GLVM Mirroring Only)
```

Figure 8-33 Integrating the geographically mirrored volume yurivg into the resource group

Note: If you have configured the asynchronous geographically mirrored volume group, set the default choice for data divergence recovery to preserve when recovering from data divergency. Allow varyon with missing data updates to allow or prevent data divergence when a non-grace failover occurs.

4. Repeat the necessary steps to complete each resource group configuration.

8.5.4 Verifying and synchronizing the GLVM configuration

Synchronize the configuration changes that you completed to the other cluster nodes (Figure 8-34).

```
                    HACMP Verification and Synchronization

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* Verify, Synchronize or Both                        [Both] +
* Automatically correct errors found during          [Interactively] +
  verification?
* Force synchronization if verification fails?       [No] +
* Verify changes only?                               [No] +
* Logging                                            [Standard] +
```

Figure 8-34 PowerHA verification and synchronization of GLVM_cluster

Automatic corrective action: When you configure geographically mirrored volume groups, certain automatic corrective actions of the PowerHA cluster verification utility, such as automatically importing or exporting a volume group, do not work. Therefore, run the automatic corrective action of the cluster verification utility in Interactive mode.

8.6 Configuring a 3-node, 2-site PowerHA for GLVM

This section describes an implementation scenario of a three-node cluster in two sites with PowerHA and GLVM.

8.6.1 Creating a GLVM cluster between two sites

In this scenario, we create a GLVM cluster between two sites, where the first site is in NY and the second site in UK. We have two nodes at site NY connected to shared external storage IBM DS8000 and one node in the UK, which only has internal disks. The physical volumes hdisk1 - hdisk4 at site NY are mirrored with the physical volumes hdisk1 - hdisk4 at the UK site. Using the RPV server and RPV client configuration, hdisk1 - hdisk4 at site NY are presented as hdisk5 - hdisk8 at the UK site. The reverse is also true. Figure 8-35 shows the logical diagram of our testing environment.

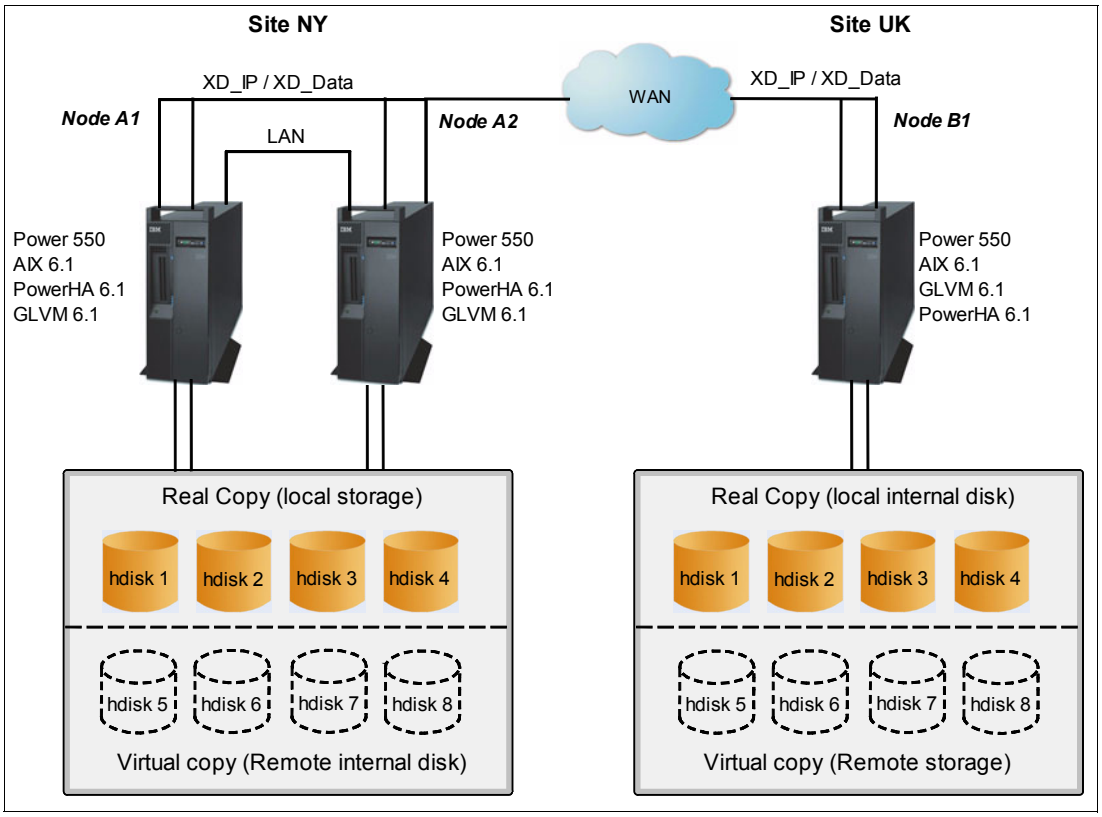


Figure 8-35 Logical diagram of test environment

Creating the base cluster

Creating the base cluster by configuring the cluster topology:

1. Use the **smitty hacmp** menu, and select **Extended Topology Configuration**.
2. Add a cluster. The name of our cluster is GLVM_NY_UK.
3. Add cluster nodes. We have three nodes in our cluster. Two nodes reside in NY, and one node resides in the UK. The first node in NY is named GLVM_NY_A1, the second node is named GLVM_NY_A2, and the node in the UK is named GLVM_UK_B1.
4. Add a cluster site. We create a two-site cluster that consists of the NY site and the UK site. The NY site has two nodes, namely the GLVM_NY_A1 and GLVM_NY_A2 nodes, where the UK site has one node, GLVM_UK_B1.
5. Add cluster networks. Our cluster has only one network connection from the NY site to the UK site. This setup is not recommended for configuration in a production implementation because the network becomes a single point of failure. It is best to have redundant networks for XD_data and for XD_ip.

   In our case, we create three networks that consist of one XD_data network, one XD_ip network, and one IP persistent network. But only one network is used in our configuration because of a network infrastructure limitation.
6. Add communication interfaces to our networks. As described in the previous step, we have only one network connection from the NY site to the UK site. In this condition, we create only three communication interfaces (Table 8-5).

Table 8-5 Communication interfaces

Communication interface name	IP address
GLVM_NY_A1 XD_data	10.12.5.50
GLVM_NY_A2 XD_data	10.12.5.53
GLVM_UK_B1 XD_data	10.137.62.112

Example 8-15 shows the result of creating our cluster topology.

Example 8-15 Cluster topology

```
aixod17.lpar.co.uk:/ # cltopinfo
Cluster Name: GLVM_NY_UK
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
There are 3 node(s) and 3 network(s) defined
NODE GLVM_NY_A1:
        Network net_XD_data_01
                GLVMUK_A1_XDIP  10.12.5.50
        Network net_XD_ip_01
        Network net_ether_02
NODE GLVM_NY_A2:
        Network net_XD_data_01
                GLVMUK_A2_XDIP  10.12.5.53
        Network net_XD_ip_01
        Network net_ether_02
NODE GLVM_UK_B1:
        Network net_XD_data_01
                GLVMUK_B1_XDIP  10.137.62.212
```

```
Network net_XD_ip_01
Network net_ether_02
```

Creating resource groups

In our scenario, we create two resource groups:

- Zhifa_rg
- Rebecca_rg

Zhifa_rg has a primary node, GLVM_NY_A1, whereas Rebecca_rg has a primary node, GLVM_NY_A2. Figure 8-36 shows detail of these resource groups.

```
                        Add a Resource Group (extended)

Resource Group 1:
[Entry Fields]
* Resource Group Name                                [Zhifa_rg]

  Inter-Site Management Policy                       [Prefer Primary Site]   +
* Participating Nodes from Primary Site              [GLVM_NY_A1 GLVM_NY_A2] +
  Participating Nodes from Secondary Site            [GLVM_UK_B1]            +

  Startup Policy                                      Online On Home Node O> +
  Failover Policy                                     Failover To Next Prio> +
  Failback Policy                                     Failback To Higher Pr> +

Resource Group 2:
* Resource Group Name                                [Rebecca_rg]

  Inter-Site Management Policy                       [Prefer Primary Site]   +
* Participating Nodes from Primary Site              [GLVM_NY_A2 GLVM_NY_A1] +
  Participating Nodes from Secondary Site            [GLVM_UK_B1]            +

  Startup Policy                                      Online On Home Node O> +
  Failover Policy                                     Failover To Next Prio> +
  Failback Policy                                     Failback To Higher Pr> +
```

Figure 8-36 Cluster resource group

We now need to configure GLVM, which will be integrated with our cluster resource group.

Configuring GLVM

In this section, we briefly describe the steps for configuring GLVM in a three-node cluster configuration. For information about how to configure GLVM, see 8.3.4, “Configuring GLVM and PowerHA by using the GLVM wizard” on page 352.

In our environment, node1 and node2 have four units of shared disk that reside in an IBM DS8000, whereas node3 at the UK site only has internal disks. Example 8-16 shows disk configuration for these nodes before configuring GLVM.

Example 8-16 Disk configuration before configuring GLVM

```
root@GLVMUK_57O_1_A1 / > lspv
hdisk0          00c1f170e170ae72                    rootvg          active
hdisk1          00c1f1703fd37be3                    None
```

```
hdisk2          00c1f1703fd38b28                    None
hdisk3          00c7cd9e843445ad                    None
hdisk4          00c1f1703fd39c15                    None

root@GLVMUK_570_2_A2 / > lspv
hdisk0          00c0f6a02e4d55c0                    rootvg          active
hdisk1          00c0f6a04e161d77                    None
hdisk2          00c1f1703fd37be3                    None
hdisk3          00c1f1703fd38b28                    None
hdisk4          00c7cd9e843445ad                    None
hdisk5          00c1f1703fd39c15                    None

aixod17.lpar.co.uk:/ # lspv
hdisk0          0003ee875d250553                    rootvg          active
hdisk1          0003ee876c83ee54                    None
hdisk2          0003ee876c83fa0e                    None
hdisk3          0003ee876c840e63                    None
hdisk4          0003ee876c841919                    None
```

Example 8-16 on page 376 shows that hdisk1 - hdisk4 in the GLVM_UK_A1 node are the same disks as hdisk2 - hdisk5 in the GLVM_NY_A2 node. These four disks are configured as GLVM disks at the NY site. Those disks have their corresponding GLVM disks in the GLVM_UK_B1 node, namely hdisk1 - hdisk4.

To configure GLVM:

1. Create the RPV server at the UK site.

   These steps make the hard disk from the GLVM_UK_B1 node accessible in the GLVM_NY_A1 and GLVM_NY_A2 nodes. We create the access by using **smitty glvm_utils**, and Example 8-17 shows the result.

   Example 8-17 RPV server in the GLVM_UK_B1 node

   ```
   aixod17.lpar.co.uk:/ # lsdev -t rpvstype
   rpvserver0 Available  Remote Physical Volume Server
   rpvserver1 Available  Remote Physical Volume Server
   rpvserver2 Available  Remote Physical Volume Server
   rpvserver3 Available  Remote Physical Volume Server
   ```

2. Create the RPV client at the NY site.

   This step acknowledges the remote disk from the GLVM_UK_B1 node and makes it available in the local node. We run the **smitty rpvclient** command in both of the nodes at the UK site, and Example 8-18 shows the result.

   Example 8-18 RPV client configuration in the GLVM_NY_A1 node

   ```
   root@GLVMUK_570_1_A1 / > lsdev -t rpvclient
   hdisk5 Available  Remote Physical Volume Client
   hdisk6 Available  Remote Physical Volume Client
   hdisk7 Available  Remote Physical Volume Client
   hdisk8 Available  Remote Physical Volume Client

   root@GLVMUK_570_1_A1 / > lsdev -Cc disk
   hdisk0 Available          Virtual SCSI Disk Drive
   hdisk1 Available 02-01-02 MPIO FC 2145
   hdisk2 Available 02-01-02 MPIO FC 2145
   ```

```
hdisk3 Available 02-01-02 MPIO FC 2145
hdisk4 Available 02-01-02 MPIO FC 2145
hdisk5 Available          Remote Physical Volume Client
hdisk6 Available          Remote Physical Volume Client
hdisk7 Available          Remote Physical Volume Client
hdisk8 Available          Remote Physical Volume Client
```

Check also the configuration in the GLVM_NY_A2 node, and the output must also have four rpvclients.

3. Create the RPV server at the NY site.

   Create the RPV server at the NY site in both nodes, beginning with creating it in the GLVM_NY_A1 node. Example 8-19 shows the result.

Example 8-19 RPV server configuration in the GLVM_NY_A1 node

```
root@GLVMUK_570_1_A1 / > lsdev -t rpvstype
rpvserver0 Available  Remote Physical Volume Server
rpvserver1 Available  Remote Physical Volume Server
rpvserver2 Available  Remote Physical Volume Server
rpvserver3 Available  Remote Physical Volume Server
```

Then, create the RPV server in the GLVM_NY_A2 node. Before you create it, bring the RPV server in the GLVM_NY_A1 node into the define state because both nodes use the same hard disk for their RPV server configurations. Example 8-20 shows the result of creating the RPV server in the GLVM_NY_A2 node.

Example 8-20 RPV server configuration in the GLVM_NY_A2 node

```
root@GLVMUK_570_2_A2 / > lsdev -t rpvstype
rpvserver0 Available  Remote Physical Volume Server
rpvserver1 Available  Remote Physical Volume Server
rpvserver2 Available  Remote Physical Volume Server
rpvserver3 Available  Remote Physical Volume Server
```

4. Create the RPV client at the UK site.

   Even though two nodes provide the RPV server at the NY site, create the RPV client only once in the GLVM_UK_B1 node because both of these RPV servers use the same physical disk. Example 8-21 shows the result of creating the RPV client in the GLVM_NY_B1 node.

Example 8-21 RPV client configuration in the GLVM_UK_B1 node

```
aixod17.lpar.co.uk:/ # lsdev -t rpvclient
hdisk5 Available  Remote Physical Volume Client
hdisk6 Available  Remote Physical Volume Client
hdisk7 Available  Remote Physical Volume Client
hdisk8 Available  Remote Physical Volume Client

aixod17.lpar.co.uk:/ # lsdev -Cc disk
hdisk0 Available  Virtual SCSI Disk Drive
hdisk1 Available  Virtual SCSI Disk Drive
hdisk2 Available  Virtual SCSI Disk Drive
hdisk3 Available  Virtual SCSI Disk Drive
hdisk4 Available  Virtual SCSI Disk Drive
hdisk5 Available  Remote Physical Volume Client
hdisk6 Available  Remote Physical Volume Client
```

```
hdisk7 Available  Remote Physical Volume Client
hdisk8 Available  Remote Physical Volume Client
```

Creating the volume group

To create the volume group, we use the regular SMIT lvm menus. Begin by running **smitty mkvg** and choose the preferred volume group type. We chose to create a scalable volume group that later will be required for configuring the mirror pool disk. We create two volume groups:

- ny1vg
- ny2vg

Volume group ny1vg uses two hdisk units from the local disk and two hdisk units from the remote disk. Volume group ny2vg also uses two hdisk units from local disk and two hdisk units from the remote disk.

When creating a volume group, certain parameters must follow certain rules:

- Asynchronous mirroring requires the volume group be in scalable VG format.
- Set the Activate volume group AUTOMATICALLY at system restart option to `no`.
- Turn off the bad block relocation policy for the volume group as required by asynchronous mirroring.
- Disable quorum for the volume groups.
- Set the “Enable strict mirror pools need” option to `yes`.

Example 8-22 shows the disk configuration after you create the volume group.

Example 8-22 Physical volumes and volume group configuration

```
root@GLVMUK_570_1_A1 / > lspv
hdisk0          00c1f170e170ae72                    rootvg          active
hdisk1          00c1f1703fd37be3                    ny1vg
hdisk2          00c1f1703fd38b28                    ny1vg
hdisk3          00c7cd9e843445ad                    ny2vg
hdisk4          00c1f1703fd39c15                    ny2vg
hdisk5          0003ee876c83ee54                    ny1vg
hdisk6          0003ee876c83fa0e                    ny1vg
hdisk7          0003ee876c840e63                    ny2vg
hdisk8          0003ee876c841919                    ny2vg
```

Vary off the volume group in the GLVM_UK_A1 node and then, import it in all other nodes.

Configuring the mirror pool

This section briefly explains the mirror pool configuration. For information about the definition, command, and steps to configure mirror pool disks, see 4.1.3, “Mirror pool disk” on page 125.

In our configuration, we have two volume groups, ny1vg and ny2vg. Each volume group consists of two disk units at the NY site and two disk units at the UK site. We create two mirror pools in each volume group. Example 8-23 shows the configuration.

Example 8-23 Mirror pool configuration

```
root@GLVMUK_570_1_A1 / > lsmp -A ny1vg
VOLUME GROUP:       ny1vg               Mirror Pool Super Strict: yes
```

```
MIRROR POOL:        ny1_mp2             Mirroring Mode:           SYNC
MIRROR POOL:        ny1_mp1             Mirroring Mode:           SYNC
root@GLVMUK_570_1_A1 / > lsmp -A ny2vg
VOLUME GROUP:       ny2vg               Mirror Pool Super Strict: yes

MIRROR POOL:        ny2_mp1             Mirroring Mode:           SYNC
MIRROR POOL:        ny2_mp2             Mirroring Mode:           SYNC

root@GLVMUK_570_1_A1 / > lsvg -P ny1vg
Physical Volume   Mirror Pool
hdisk1            ny1_mp1
hdisk2            ny1_mp1
hdisk5            ny1_mp2
hdisk6            ny1_mp2

root@GLVMUK_570_1_A1 / > lsvg -P ny2vg
Physical Volume   Mirror Pool
hdisk3            ny2_mp1
hdisk4            ny2_mp1
hdisk7            ny2_mp2
hdisk8            ny2_mp2
```

In Example 8-23 on page 379, we have the ny1_mp1 and ny1_mp2 mirror pools in the ny1vg volume group and the ny2_mp1 and ny2_mp2 mirror pools in the ny2vg volume group.

Creating logical volumes

For this GLVM cluster, we use the asynchronous method replication. As explained in 4.1.3, “Mirror pool disk” on page 125, for asynchronous replication, we need to create the aio_cache logical volume, not the mirror, and reside in one mirror pool only. We also create one logical volume, jfs2, for the data in each volume group.

To create the logical volume:

1. Run **smitty mklv** and choose a volume group where the logical volume will reside.
2. Input the parameter for creating a logical volume in a mirror pool configuration with asynchronous replication. Several parameters need to be input:
   - Logical volume type

     This parameter is related to the type of logical volume. We can choose jfs, jfs2, sysdump, paging, jfslog, jfs2log, boot, and aio_cache by pressing F4.
   - Number of copies of each logical partition

     This parameter must be filled with the same mirror pool number on that volume group. The reason is that each mirror pool has the copy of every logical volume on that volume group.
   - Enable BAD BLOCK relocation

     This parameter must be set to no. AIX detects bad blocks and then relocates the data on the disk drives. Whenever a problem is on the disks, the data relocation takes place automatically for data protection. In a mirror pool environment, this feature must be set to no because we have a set group of disks, and each group of disks must have every copy of the logical volume.

– Mirror Pool for First Copy

  This parameter indicates the mirror pool name where the first copy of the logical volume resides.

– Mirror Pool for Second Copy

  This parameter indicates the mirror pool name where the first copy of the logical volume resides.

– Mirror Pool for Third Copy

  This parameter indicates the mirror pool name where the first copy of the logical volume resides.

3. Repeat this step for any logical volumes and each volume group as needed.
4. If you are using file systems, manually create the jfslog LV for each volume group to specify a unique name to each. We created ny1log and ny2log. Upon creation, initialize the jfslog LV by running:

   ```
   logform /dev/<loglvname>
   ```

Example 8-24 shows the details of the logical volume that was created.

Example 8-24 Logical volume configuration

```
root@GLVMUK_570_1_A1 / > lsvg -l ny1vg
ny1vg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
ny1_lv              jfs2       4       8       2    closed/syncd  N/A
ny1_aiocache        aio_cache  2       2       1    closed/syncd  N/A
ny1log              jfs2log    1       2       2    closed/syncd  N/A
ny1_2_aiocache      aio_cache  2       2       1    closed/syncd  N/A

root@GLVMUK_570_1_A1 / > lsvg -l ny2vg
ny2vg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
ny2_lv              jfs2       4       8       2    closed/syncd  N/A
ny2_aiocache        aio_cache  2       2       1    closed/syncd  N/A
ny2log              jfs2log    1       2       2    closed/syncd  N/A
ny2_2_aiocache      aio_cache  2       2       1    closed/syncd  N/A
root@GLVMUK_570_1_A1 / >
```

Converting from synchronous to asynchronous

So far, we created a mirror pool configuration, and the replication method is still using synchronize replication. To it to asynchronous replication, run the SMIT **`smitty glvm_utils`** command. Then, select **Geographically Mirrored Volume Groups → Manage Geographically Mirrored Volume Groups with Mirror Pools → Configure Mirroring Properties of a Mirror Pool → Convert to Asynchronous Mirroring for a Mirror Pool**.

Figure 8-37 shows the results.

```
             Convert to Asynchronous Mirroring for a Mirror Pool

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

                                                      [Entry Fields]
* VOLUME GROUP Name                                   ny1vg
* MIRROR POOL Name                                    ny1_mp2
  Logical Volume for I/O Cache                       [ny1_aiocache]          +
  I/O Cache High Water Mark Value                    [100]                    #
```

Figure 8-37 Converting synchronous to asynchronous mirror pool.

We convert all the mirror pool disks. Example 8-25 shows the status of all logical volumes after the conversion.

Example 8-25 Mirror pool status after converting to asynchronous

```
root@GLVMUK_570_1_A1 / > lsmp -A ny1vg
MIRROR POOL:        ny1_mp2             Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: active              ASYNC CACHE LV:           ny1_aiocache
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        no
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

MIRROR POOL:        ny1_mp1             Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: inactive            ASYNC CACHE LV:           ny1_2_aiocache
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        yes
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

root@GLVMUK_570_1_A1 / > lsmp -A ny2vg
VOLUME GROUP:       ny2vg               Mirror Pool Super Strict: yes

MIRROR POOL:        ny2_mp1             Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: inactive            ASYNC CACHE LV:           ny2_2_aiocache
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        yes
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

MIRROR POOL:        ny2_mp2             Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: active              ASYNC CACHE LV:           ny2_aiocache
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        no
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no
```

Integrating the cluster resource group with GLVM

We created the Zhifa_rg and Rebecca_rg resource groups in previous steps. These resource groups are defined in the primary site and the secondary site with no volume group integrated.

We intend to include the ny1vg and ny2vg volume groups in these resource groups. We run the **smitty hacmp** command, and then integrate the volume group (Figure 8-38).

```
        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                               Zhifa_rg
  Inter-site Management Policy                      Prefer Primary Site
  Participating Nodes from Primary Site             GLVM_NY_A1 GLVM_NY_A2
  Participating Nodes from Secondary Site           GLVM_UK_B1

  Startup Policy                                    Online On Home Node O>
  Failover Policy                                   Failover To Next Prio>
  Failback Policy                                   Failback To Higher Pr>
  Failback Timer Policy (empty is immediate)        []                     +

  Service IP Labels/Addresses                       []                     +
  Application Servers                               [Zhifa_app]            +

  Volume Groups                                     [ny1vg]                +
  Use forced varyon of volume groups, if necessary   false                 +
  Automatically Import Volume Groups                 false                 +
  Default choice for data divergence recovery        ignore                +
```

Figure 8-38 Integrated a volume group into a resource group

We also run the command in Figure 8-38 to include the ny2vg volume group into the Rebecca_rg resource group.

Verification and synchronization

The last step in configuring the cluster is the cluster verification and synchronization. We run the **smitty hacmp** command and select **Extended Configuration → Extended Verification and Synchronization**.

The result must be OK.

8.6.2 Testing the cluster

In our cluster, we perform two tests: node failure and site failure.

Test 1: Node failure

In this test, we simulate one node at the NY site suddenly going down. We expect the resource group with the volume group and application to be taken over by the second node at the NY site.

We completed this test by using the **halt -q** AIX command in the GLVM_NY_A1 node. This command stops the processor and powers off the server, similarly to when we remove the power cable from the server.

The status of the resource group before and after the node failure can be checked by using the **/usr/es/sbin/cluster/utilities/clRGinfo** command. We check before and after the testing. Example 8-26 shows the result.

Example 8-26 Resource group status before and after the testing node failure

```
Before testing:
aixod17.lpar.co.uk:/ # clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zhifa_rg       ONLINE                       GLVM_NY_A1@NY
               OFFLINE                      GLVM_NY_A2@NY
               ONLINE SECONDARY             GLVM_UK_B1@UK

Rebecca_rg     ONLINE                       GLVM_NY_A2@NY
               OFFLINE                      GLVM_NY_A1@NY
               ONLINE SECONDARY             GLVM_UK_B1@UK

After testing:
aixod17.lpar.co.uk:/ # clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zhifa_rg       OFFLINE                      GLVM_NY_A1@NY
               ONLINE                       GLVM_NY_A2@NY
               ONLINE SECONDARY             GLVM_UK_B1@UK

Rebecca_rg     ONLINE                       GLVM_NY_A2@NY
               OFFLINE                      GLVM_NY_A1@NY
               ONLINE SECONDARY             GLVM_UK_B1@UK
```

Example 8-26 shows that when the GLVM_NY_A1 node failed, the resource group Zhifa_rg is automatically taken over by node GLVM_NY_A2. When the failed node, GLVM_NY_A1, is up and joins the cluster again, the resource group Zhifa_rg is reacquired by node GLVM_NY_A1.

Test 2: Site failure

In this test, we simulate the server at the UK site being down. We expect all the resource groups to be taken over by the GLVM_UK_B1 node in the UK site.

We perform this test by running **halt -q** at the same time in the GLVM_NY_A1 and GLVM_NY_A2 nodes, and then check the resource group status before and after the testing. Example 8-27 shows the result.

Example 8-27 Resource group status before and after the testing node failure

```
Before testing:
aixod17.lpar.co.uk:/ # clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zhifa_rg       ONLINE                       GLVM_NY_A1@NY
               OFFLINE                      GLVM_NY_A2@NY
               ONLINE SECONDARY             GLVM_UK_B1@UK

Rebecca_rg     ONLINE                       GLVM_NY_A2@NY
```

```
               OFFLINE                      GLVM_NY_A1@NY
               ONLINE SECONDARY             GLVM_UK_B1@UK
```

After testing:

```
aixod17.lpar.co.uk:/ # clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
Zhifa_rg       OFFLINE                      GLVM_NY_A1@NY
               OFFLINE                      GLVM_NY_A2@NY
               ONLINE                       GLVM_UK_B1@UK

Rebecca_rg     OFFLINE                      GLVM_NY_A2@NY
               OFFLINE                      GLVM_NY_A1@NY
               ONLINE                       GLVM_UK_B1@UK
```

Example 8-27 on page 384 shows that when the GLVM_NY_A1 and GLVM_NY_A2 nodes are down, the Zhifa_rg and Rebecca_rg resource groups are taken over automatically by the GLVM_NY_A2 node. It takes around 5 minutes for the GLVM_UK_B2 node to detect the site failure in the NY site.

8.7 Performance with aio_cache

Using the asynchronous GLVM requires a logical volume of the aio_cache type to buffer the writes until mirrored to the remote site. When the aio_cache fills up, GLVM reverts to synchronous mode. The larger the aio_cache is, the more data can be buffered for the disk writes to the remote rpvserver. Also consider that the larger the aio_cache, the more data that can be lost if a site failure occurs.

When choosing the size of the aio_cache, choose an aio_cache size as small as possible, and then monitor whether it is getting filled too quickly. You can increase the size of aio_cache if it is filling too rapidly.

The **rpvstat** command can be used to monitor the aio_cache usage. Use **rpvstat -C** to monitor and **rpvstat -r** to reset the counters (Example 8-28).

Example 8-28 rpvstat -C command

```
root@GLVM_A1 / > rpvstat -C

Remote Physical Volume Statistics:

                                 Max    Pending    Total  Max
                    Total Async  Cache  Cache      Cache  Cache   Cache Free
GMVG Name        ax Writes       Util % Writes     Wait % Wait    Space KB
---------------- -- ------------ ------ ---------- ------ ------- ----------
yurivg            A          2563   0.32          0   0.00       0     326655
```

The **chmp** command can be used to raise and lower the high water mark, which is the percent of aio_cache that will be used. The default is 100%. To change the percentage of the I/O cache size used, enter:

```
chmp -h <percent> -m <mirrorpool> <volumegroup>
```

You must the **-h** option when making the volume group asynchronous if you want it less than 100% because it cannot be used to decrease the size after it is created. Only a dynamic increase is supported (Figure 8-39). The rpvclient detects the high water mark change on these two conditions:

- The mirror pool is changed from asynchronous to synchronous and back to asynchronous.
- The rpvclients are stopped and started, which is accomplished by bringing the resource group offline and then online to see the change.

```
root@GLVM_A1 / > lsmp -A yurivg
VOLUME GROUP:       yurivg              Mirror Pool Super Strict: yes

MIRROR POOL:        Asite               Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: inactive            ASYNC CACHE LV:           datacachelv2
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        yes
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

MIRROR POOL:        Bsite               Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: active              ASYNC CACHE LV:           datacachelv1
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        no
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

root@GLVM_A1 / > chmp -S -m Bsite yurivg

root@GLVM_A1 / > chmp -A -c datacachelv1 -h 3 -m Bsite yurivg

root@GLVM_A1 / > lsmp -A yurivg
VOLUME GROUP:       yurivg              Mirror Pool Super Strict: yes

MIRROR POOL:        Asite               Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: inactive            ASYNC CACHE LV:           datacachelv2
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        yes
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

MIRROR POOL:        Bsite               Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: active              ASYNC CACHE LV:           datacachelv1
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        no
ASYNC CACHE HWM:    3                   ASYNC DATA DIVERGED:      no

root@GLVM_A1 / > rpvstat -C

Remote Physical Volume Statistics:

                                  Max    Pending    Total  Max
                    Total Async   Cache  Cache      Cache  Cache   Cache Free
GMVG Name        ax Writes        Util % Writes     Wait % Wait    Space KB
---------------- -- ------------- ------ ---------- ------ ------- ----------
yurivg            A              0   0.00          0   0.00       0       8703
root@GLVM_A1 /yuri > rpvstat -C

Remote Physical Volume Statistics:

                                  Max    Pending    Total  Max
                    Total Async   Cache  Cache      Cache  Cache   Cache Free
GMVG Name        ax Writes        Util % Writes     Wait % Wait    Space KB
---------------- -- ------------- ------ ---------- ------ ------- ----------
yurivg            A              1  10.63          0   0.00       0       8694
```

Figure 8-39 Changing the cache size percentage

8.8 Monitoring

To assist you in monitoring the state of your GLVM environment, including the RPVs and GMVGs that you have configured, GLVM includes two tools:

- **rpvstat**
- **gmvstat**

These commands provide real-time status information about RPVs and GMVGs.

8.8.1 The rpvstat command

The **rpvstat** command provides status monitoring for RPV Clients. It displays the following information for one or more RPV clients:

- RPV client name
- Connection status
- Total number of completed reads
- Total number of KBs read
- Total number of read errors
- Total number of pending reads
- Total number of pending KBs to read
- Total number of completed writes
- Total number of KBs written
- Total number of write errors
- Total number of pending writes
- Total number of pending KBs to write

The **rpvstat** command can optionally display its I/O-related statistics on a per-network basis. The network summary option shows more information about network throughput in KBps.

The **rpvstat** command can also display the highest recorded values for the pending statistics, which includes the following historical high water mark numbers:

- Maximum number of pending reads per device and network (high water mark)
- Maximum number of pending KBs to read per device and network (high water mark)
- Maximum number of pending writes per device and network (high water mark)
- Maximum number of pending KBs to write per device and network (high water mark)

  These statistics are reported on a separate display and include the additional statistic of the number of I/O operations that were tried again (combination of both reads and writes).

The **rpvstat** command allows information to be displayed for all RPV clients on the system or for a subset of RPV clients that is specified by the RPV client name on the command line. The **rpvstat** command also allows the information to be monitored (redisplayed at user-specified intervals).

The **rpvstat** command interacts with the RPV client pseudo device driver to retrieve the information that the client displays.

The rpvstat command man page

The **rpvstat** command man page provides reference information for the **rpvstat** command. The purpose of this command is to display RPV client statistics.

Syntax

The **rpvstat** command has the following syntax:

```
rpvstat -h
rpvstat [-n] [-t] [-i Interval [-c Count] [-d]] [rpvclient_name . . .]
rpvstat -N [-t] [-I Interval [-c Count] [-d]]
rpvstat -m [-n] [-t] [rpvclient_name . . .]
rpvstat -R [-r][rpvclient_name . . .]
rpvstat -r [-R] [rpv-device(s)...]
rpvstat -A [-t] [-i Interval [-d] [-c Count] ] [rpv-device(s)...] |
rpvstat -C [-t] [-i Interval [-d] [-c Count] ] [rpv-device(s)...]
```

Description

The **rpvstat** command displays statistical the following information available from the RPV client device:

- RPV client name
- Connection status
- Total number of completed reads
- Total number of KBs read
- Total number of read errors
- Total number of pending reads
- Total number of pending KBs to read
- Total number of completed writes
- Total number of KBs written
- Total number of write errors
- Total number of pending writes
- Total number of pending KBs to write
- Statistics for asynchronous I/O
- Statistics for asynchronous I/O cache

The read/write errors are displayed together. These counters indicate the number of I/O errors returned to the application.

The **rpvstat** command can optionally display its I/O-related statistics on a per-network basis. A network summary option of the command displays more I information about the network throughput in KBps. The throughput is calculated per interval time that is specified by the user while in monitoring mode.

The **rpvstat** command can also display the highest recorded values for the pending statistics, which includes the following historical high water mark numbers:

- Maximum number of pending reads per network
- Maximum number of pending KBs to read per network
- Maximum number of pending writes per network
- Maximum number of pending KBs to write per network

These statistics are reported on a separate display and include the additional statistic of the number of I/O operations that are tried again (both reads and writes). This count records the number of I/Os that are tried again that have occurred on this network or device. This information can be used as an indicator of a marginal or failing network.

You can also display the statistics for asynchronous mirroring. The **rpvstat** command prints overall asynchronous statistics using the –A option. To display statistics per device, specify the list of devices. You can display the asynchronous I/O cache information by using the **-C** option (Table 8-6).

Table 8-6 Flags

Flags	Description
-h	Display command syntax and usage.
-R	Resets counters in the RPV clients (requires root privilege).
-t	Includes date and time in display.
-n	Displays statistics for individual mirroring networks.
-N	Displays summary statistics by mirroring network, including throughput rate for each network.
-i Interval	Automatically shows the status every <Interval> seconds. The value of the <*Interval*> parameter must be an integer greater than zero and less than or equal to 3600. If the <*Interval*> parameter is not specified, the status information is displayed once.
-c Count	Shows information at the indicated interval <Count> times. The value of the <Count> parameter must be an integer greater than zero and less than or equal to 999999. If the <Interval> parameter is specified, but the <Count> parameter is not, it is displayed indefinitely.
-m	Displays historical maximum pending values (high water mark values) and accumulated retry count.
-d	Displays applicable monitored statistics as delta amounts from prior value.
-A	Displays the statistics for asynchronous I/O.
-C	Display the statistics for asynchronous I/O cache.
-r	Resets counters for the asynchronous I/O cache information. You can specify the -R and -r options together to reset all counters. (This needs root privilege.)

You can use other flags to get more statistical information:

- In monitor mode (-i) if the -d option is also specified, certain statistics (completed reads, completed writes, completed KB read, completed KB written, and errors) are represented as delta amounts from their previously displayed values. These statistics are prefixed with a plus sign (+) on the second and succeeding displays. A delta value is not displayed under certain circumstances, such as when an error is detected in the previous iteration or when a configuration change is made between iterations.
- When a list of RPV client devices is not explicitly listed on the command line, the list of all available RPV clients is generated at command initiation. In monitor mode, this list of RPV clients to display is not refreshed on each display loop. This means that any additional RPV clients added or deleted are not recognized until the command is started again.
- The -i interval is the time, in seconds, between each successive gathering and display of RPV statistics in monitor mode. This interval is not a precise measure of the elapsed time between each successive updated display. The **rpvstat** command obtains information that it displays by calling system services and has no control over the amount of time that these services take to complete their processing. Larger numbers of RPVs result in the **rpvstat** command taking longer to gather information and elongates the time between successive displays in monitor mode, sometimes taking much longer than the -i interval between displays.

The -A option prints the following statistical information for one or more asynchronous devices:

- Asynchronous device name.
- Asynchronous status. The status is printed as a single character.
  - **A** Device is fully configured for asynchronous I/O and can accept asynchronous I/Os.
  - **I** Asynchronous configuration is incomplete.
  - **U** The device is not configured with asynchronous configuration. Therefore, it acts as a synchronous device. All statistics are printed as zero.
  - **X** Device status cannot be retrieved. All the remaining statistics are printed as zero.
- Total number of asynchronous remote writes completed. The writes are mirrored and complete.
- Total asynchronous remote writes that are completed in KB. The writes are mirrored and complete.
- Total number of asynchronous writes that are pending to mirror. The writes are in the cache. These writes are complete as far as LVM is concerned, but are not yet mirrored.
- Total asynchronous writes that are pending to mirror in KB. The writes are in the cache. These writes are complete as per LVM is concerned, but are not yet mirrored.
- Total number of writes whose response is pending. These writes are in the pending queue and are not yet written to cache.
- Total asynchronous write responses that are pending in KB. These writes are in the pending queue and are not yet written to cache.

The -C option prints the following statistical information about the asynchronous I/O cache. The VG name is extracted from the ODM:

- Volume group name.
- Asynchronous status. The status is printed as a single character:
  - **A** Device is fully configured for asynchronous I/O and can accept asynchronous I/Os.
  - **I** Asynchronous configuration is incomplete.
  - **U** The device is not configured with asynchronous configuration and, therefore, acts as an asynchronous device. All statistics are printed as zero.
  - **X** Device status cannot be retrieved. All the remaining statistics are printed as zero.
- Total asynchronous writes.
- Maximum cache utilization in percent.
- Number of pending asynchronous writes that are waiting for the cache flush after the cache hits the high water mark.
- Percentage of writes that are waiting for the cache flush after the cache hits the high water mark limit.
- Maximum time waited after the cache hits high water mark in seconds.
- Current free space in cache in KB.

Notes

Note the following information:

- The count of reads and writes is accumulated on a per-buffer basis. For example, an application I/O passes a vector of buffers in a single read or write call. Instead of counting that read or write as a single I/O, it is counted as the number of buffers in the vector.

- The count of completed and pending I/O KB is truncated. Any fractional amount of a KB is dropped in the output display.
- The cx field in the display output shows the connection status and can have the following values:

  A number The count of active network connections between the RPV client and its RPV server.

  Y Indicates that the connection represented by the IP address is available and functioning.

  N Indicates that the connection represented by the IP address is not available.

  X Indicates that the required information cannot be retrieved from the device driver for the following reasons:

  - The device driver is not loaded.
  - The device is not in the available state.
  - The device has been deleted.

Exit status

This command returns the following exit values:

0 No errors.
>0 An error occurred.

Examples

This section lists sample commands that are used to display the information that is requested:

- To display statistical information for all RPV clients, use the following command:

  ```
  rpvstat
  ```

- To display statistical information for RPV client hdisk14, use the following command:

  ```
  rpvstat hdisk14
  ```

- To reset the statistical counters in RPV client hdisk23, use the following command:

  ```
  rpvstat -R hdisk23
  ```

- To display statistical information for RPV client hdisk14 and repeat the display every 30 seconds for 12 times, use the following command:

  ```
  rpvstat hdisk14 -i 30 -c 12
  ```

- To display statistical information for all RPV clients and include information by mirroring network, enter the following command:

  ```
  rpvstat -n
  ```

- To display statistical information for all mirroring networks, use the following command:

  ```
  rpvstat -N
  ```

- To display statistical information about maximum pending values for all RPV clients, use the following command:

  ```
  rpvstat -m
  ```

Files

The `/usr/sbin/rpvstat` contains the **`rpvstat`** command.

Related information

See the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

Sample display output for the rpvstat command

The following samples display various outputs for the **rpvstat** command.

- Example 1: Run the **rpvstat** command with no options to display all RPV client on the local node, along with accumulated statistics (Figure 8-40).

```
rpvstat
                                Comp Reads  Comp Writes  Comp KBRead  Comp KBWrite Errors
RPV Client          cx          Pend Reads  Pend Writes  Pend KBRead  Pend KBWrite
------------------ -- ---------- ---------- ------------ ------------ ------
hdisk144            4          100         22482004     1540         22834857     384
                               2           39           5            1740
hdisk158            3          34          488700       888          8336500      122
                               0           10           0            5120
hdisk202            X          0           0            0            0            0
                               0           0            0            0
```

Figure 8-40 Accumulated statistics

- Example 2: Run the **rpvstat** command with no options, but specify a single RPV client to display accumulated statics for that particular RPV client only (Figure 8-41).

```
rpvstat hdisk158
                                Comp Reads  Comp Writes  Comp KBRead  Comp KBWrite Errors
RPV Client          cx          Pend Reads  Pend Writes  Pend KBRead  Pend KBWrite
------------------ -- ---------- ---------- ------------ ------------ ------
hdisk158            3          34          488700       888          8336500      122
                               0           10           0            5120
```

Figure 8-41 Accumulated statistics

- Example 3: Run the **rpvstat** command with the -n option to show accumulated RPV client statistics for each currently defined network (Figure 8-42).

```
rpvstat -n
                                Comp Reads  Comp Writes  Comp KBRead  Comp KBWrite Errors
RPV Client          cx          Pend Reads  Pend Writes  Pend KBRead  Pend KBWrite
------------------ -- ---------- ---------- ------------ ------------ ------
hdisk144            4          100         22482004     1540         22834857     384
                               2           39           5            1740
 103.17.133.102     Y          58          5620504      768          5100384      81
                               1           11           2            18038
 103.17.133.104     Y          10          5620500      420          4892247      101
                               1           15           1            20448
 103.17.133.202     Y          30          5620598      210          5822041      98
                               0           10           1            16384
 103.17.133.204     Y          2           5620502      142          7020185      104
                               0           3            1            16870

hdisk158            3          34          488700       888          8336500      122
                               0           10           0            5120
 103.17.133.102     Y          12          122175       313          2084100      28
                               0           2            0            1280
 103.17.133.104     N          4           122174       170          2000311      32
                               0           2            0            1288
 103.17.133.202     Y          8           122176       276          2118732      30
                               0           3            0            1284
 103.17.133.204     Y          10          122175       129          2133357      32
                               0           3            0            1268
```

Figure 8-42 Output of the rspvstat command with the -n option

- Example 4: Run the **rpvstat** command specifying a single RPV client, a monitor interval of 30 seconds with 3 repeats, and a display of the date and time for each interval. When running in monitor mode with the -d option, several of the repeated statistics show only the delta from their previous value, as indicated by the preceding plus sign (+) (Figure 8-43):

  -i Interval.
  -c Repeat.
  -d Deltas indicated by plus sign.
  -t Displays the date and time for each interval.

```
rpvstat –t -i 30 -c 3 -d hdisk158

Remote Physical Volume Statistics                    Thu Jan 4 18:11:35 2007

                        Comp Reads Comp Writes  Comp KBRead Comp KBWrite Errors
RPV Client         cx   Pend Reads Pend Writes  Pend KBRead Pend KBWrite
------------------ -- ------------ ------------ ------------ ------------ ------
hdisk158            3           34       488700          888      8336500    122
                                 0           10            0         5120

Remote Physical Volume Statistics                    Thu Jan 4 18:12:05 2007

                        Comp Reads Comp Writes  Comp KBRead Comp KBWrite Errors
RPV Client         cx   Pend Reads Pend Writes  Pend KBRead Pend KBWrite
------------------ -- ------------ ------------ ------------ ------------ ------
hdisk158            3           +0          +10           +0        +5120     +0
                                 0           8             0         4096

Remote Physical Volume Statistics                    Thu Jan 4 18:12:35 2007

                        Comp Reads Comp Writes  Comp KBRead Comp KBWrite Errors
RPV Client         cx   Pend Reads Pend Writes  Pend KBRead Pend KBWrite
------------------ -- ------------ ------------ ------------ ------------ ------
hdisk158            3           +1         +100           +5       +51200     +0
                                 0           4             0         2048
```

Figure 8-43 Accumulated statistics

- Example 5: Run the **rpvstat** command with the -N option to display summary statistics for each mirroring network. Monitor every 30 seconds for a total of two repeats. This can be used to detect errors on a particular network (Figure 8-44).

```
rpvstat -N -i 30 -c 2 –d

Remote Physical Volume Statistics

                        Comp Reads Comp Writes  Comp KBRead Comp KBWrite Errors
RPV Client Network      Pend Reads Pend Writes  Pend KBRead Pend KBWrite KB/sec
------------------     ------------ ------------ ------------ ------------ ------
103.17.133.102                 10       122175         4645     20841000     28
                                0            2         1280         1280      -
103.17.133.104                 20       122174        12456      2000311     32
                                0            2            0         1288      -

Remote Physical Volume Statistics

                        Comp Reads Comp Writes  Comp KBRead Comp KBWrite Errors
RPV Client Network      Pend Reads Pend Writes  Pend KBRead Pend KBWrite KB/sec
------------------     ------------ ------------ ------------ ------------ ------
103.17.133.102                 +1         +100          +23        +2356     +0
                                0            7            0           20     79
103.17.133.104                 +0          +22           +0          +58     +0
                                0            2            0           13      1
```

Figure 8-44 Accumulated statistics

- Example 6: Run the **rpvstat** command with the **-m** option to display the maximum pending statistics (high water marks). This option shows the high water mark statistics first by RPV device (for all networks), and then by network (for all devices) (Figure 8-45).

```
rpvstat -m
                                  Maximum     Maximum     Maximum      Maximum  Total
RPV Client          cx Pend Reads Pend Writes Pend KBRead Pend KBWrite Retries
------------------ -- ---------- ----------- ----------- ------------ -------
hdisk144            4         28          56         568         9154      38
hdisk158            3         20         131          98         4817      14
hdisk173            0          9          27          31          990     212
hdisk202            X          0           0           0            0       0

Network Summary:
 103.17.133.102                10          71         220         3312      10
 103.17.133.104                15          51         382         6231     201
 103.17.133.202                14          23          98          754      34
 103.17.133.204                 9          11          31          321      12
```

Figure 8-45 Statistics

- Example 7: The **rpvstat** command with the -A option displays asynchronous I/O statistics (Figure 8-46).

```
rpvstat -A

Remote Physical Volume Statistics:

                 Completd Completed   Cached   Cached      Pending  Pending
                 Async    Async       Async    Async       Async    Async
RPV Client    ax Writes   KB Writes   Writes   KB Writes   Writes   KB Writes
------------ -- -------- ----------- -------- ----------- -------- -----------
hdisk10       A        0           0        0           0        0           0
hdisk9        A      230         115       10           5        0           0
hdisk8        A        2           8        0           0        0           0
hdisk7        A       29         116        0           0        0           0
hdisk6        A        0           0        0           0        0           0
```

Figure 8-46 Accumulated statistics

8.8.2 The gmvstat command

The **gmvgstat** command provides status monitoring for GMVGs to display the information for one or more GMVGs:

- GMVG name
- Number of physical volumes (PVs) in the GMVG on the local system
- Number of remote physical volumes (RPVs) in the GMVG represented on the local system but physically on a remote system
- Total number of volumes (PVs + RPVs)
- Number of stale volumes
- Total number of physical partitions (PPs) in the VG
- Number of stale PPs in the VG
- Synchronization state of the GMVG - percentage of PPs synchronized (that is, not stale)

The **gmvgstat** command allows this information to be displayed for all GMVGs on the system or for a subset of GMVGs specified by the GMVG name on the command line. The command display optionally includes the associated rpvstat display output for RPV clients that are associated with the specified GMVGs.

The gmvgstat command man page

The **gmvgstat** command man page provides reference information for the **gmvgstat** command. The purpose of this command is to display GMVG statistics.

Syntax

This command has the following syntax:

```
gmvgstat [-h] | [-r] [-t] [-i Interval [-c Count] [-w]]
[gmvg_name . . .]
```

Description

The **gmvgstat** command shows status information for one or more GMVGs:

- Number of physical volumes
- Number of remote physical volumes
- Total number of volumes (PVs and RPVs)
- Number of stale volumes
- Total number of physical partitions (PPs)
- Number of stale PPs
- Percentage GMVG is synchronized

The **gmvgstat** command can optionally be started in monitor mode by specifying the **-i** and **-c** flags.

If one or more GMVG names are supplied on the command line, the **gmvgstat** command verifies that each listed GMVG name is a valid, available, online GMVG. In monitor mode, the user-supplied list of GMVGs is verified during each loop.

If no GMVG names are supplied on the command line, the **gmvgstat** command reports information about all valid, available, online GMVGs. In monitor mode, the list of GMVGs to report on is regenerated during each loop. The **gmvgstat** command has the following flags:

-h Displays command syntax and help.

-r Includes information for each individual RPV client that is associated with the displayed GMVGs.

-t Displays header with date and time.

-i interval Automatically shows the status every <Interval> seconds. The value of the <Interval> parameter must be an integer greater than zero and less than or equal to 3600. If the <Interval> parameter is not specified, this parameter shows the status information once.

The **-i** interval is the time, in seconds, between each successive gathering and display of GMVG statistics in monitor mode. This interval is not a precise measure of the elapsed time between each successive updated display. The **gmvgstat** command obtains certain information that it displays by calling other commands and has no control over the amount of time that these commands take to complete their processing. Larger numbers of GMVGs result in the **gmvgtstat** command taking longer to gather information and elongates the time between successive displays in monitor mode. In some cases, an underlying command can take excessively long to complete and results in the **gmvgstat** command taking much longer than the **-i** interval between displays.

-c Count Shows information at the indicated interval <Count> times. The value of the <Count> parameter must be an integer greater than zero and less than or equal to 999999. If the <Interval> parameter is specified, but the <Count> parameter is not, it is shown indefinitely.

-w Clears the panel between each redisplay.

Operands

This operand displays the `gmvg_name`, which is name of one or more GMVGs for which to display information. If no GMVG names are specified, information for all valid, available, online GMVGs is displayed.

Exit status

The Exist status command returns the following exit values:

0 No errors.
>0 An error occurred.

Examples

This section lists sample commands used to display the information requested:

- To display statistical information for all GMVGs use the command:

  ```
  gmvgstat
  ```

- To display statistical information for the GMVG named red_gmvg7 use the command:

  ```
  gmvgstat red_gmvg7
  ```

- To display statistical information for the GMVG named red_gmvg7 with statistics for all the RPVs associated with that volume group, use the command:

  ```
  gmvgstat -r red_gmvg7
  ```

- To display information for GMVG red_gmvg7 that is automatically redisplayed every 10 seconds, enter the following command:

  ```
  gmvgstat red_gmvg7 -i 10
  ```

- To display information for GMVG red_gmvg7 that is automatically redisplayed every 10 seconds for 20 intervals and clears the screen between each redisplay, enter the following command:

  ```
  gmvgstat red_gmvg7 -i 10 -c 20 -w
  ```

Files

The `/usr/sbin/gmvgstat` contains the **gmvgstat** command.

Related information

See the *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338.

Sample display output for gmvgstat command

These samples display various output for the **gmvgstat** command.

- Example 1: Run the **gmvgstat** command with no options to display each GMVG on the local node along with associated statistics (Figure 8-47).

```
gmvgstat
GMVG Name         PVs  RPVs  Tot Vols   St Vols   Total PPs   Stale PPs  Sync
---------------  ----  ----  --------  --------  ----------  ----------  ----
red_gmvg7          60    63       123         0    29846140           0  100%
blue_gmvg23         5     5        10         1        5926         384   93%
green_gmvg19      152   152       304       152       91504       45752   50%
```

Figure 8-47 Statistics

- Example 2: Run the **gmvgstat** command with no options, but specifying the GMVG blue_gmvg23 to display statistics for that particular GMVG only (Figure 8-48).

```
gmvgstat blue_gmvg23
GMVG Name         PVs  RPVs  Tot Vols   St Vols   Total PPs   Stale PPs  Sync
---------------  ----  ----  --------  --------  ----------  ----------  ----
blue_gmvg23         5     5        10         1         235          10   95%
```

Figure 8-48 Statistics

- Example 3: Run the **gmvgstat** command with the -t and -r options, specifying the GMVG blue_gmvg23 to display statistics for the specified GMVG followed by statistics for each RPV included in blue_gmvg23 (from the **rpvstat** command) (Figure 8-49).

```
gmvgstat -t -r blue_gmvg23
Geographically Mirrored Volume Group Information          02:23:57 PM 16 Feb 2007
-----------------------------------------------          c689n02.ppd.pok.ibm.com
                                                                           siteA
GMVG Name         PVs  RPVs  Tot Vols   St Vols   Total PPs   Stale PPs  Sync
---------------  ----  ----  --------  --------  ----------  ----------  ----
blue_gmvg23         3     3         6         1         235          10   95%

Remote Physical Volume Statistics:

                        Comp Reads  Comp Writes  Comp KBRead Comp KBWrite Errors
RPV Client        cx    Pend Reads  Pend Writes  Pend KBRead Pend KBWrite
------------------ -- ------------ ------------ ------------ ------------ ------
hdisk144           4           100     22482004         1540     22834857    384
                               2           39            5         1740
hdisk158           3            34       488700          888       8336500    122
                               0           10            0         5120
hdisk202           X             0            0            0            0      0
                               0            0            0            0
```

Figure 8-49 Statistics

8.8.3 SMIT interfaces for GLVM status monitoring tools

The **rpvstat** and **gmvgstat** commands can also be run from SMIT. The descriptions of their SMIT interfaces are provided in the following sections.

The main entry point for the GLVM Status Monitor commands is on the Geographical Logical Volume Manager Utilities panel (fastpath: **glvm_utils**). The SMIT menu item shows the status monitors (Figure 8-50).

```
                Geographic Logical Volume Manager Utilities

Move cursor to desired item and press Enter.

  Geographically Mirrored Volume Groups
  Geographically Mirrored Logical Volumes
  Remote Physical Volume Clients
  Remote Physical Volume Servers
  Status Monitors
  GLVM Cluster Configuration Assistant
```

Figure 8-50 Geographical Logical Volume Manager Utilities panel

Selecting **Status Monitors** brings up the next panel (fastpath: **glvmmonitors**) (Figure 8-51).

```
                                Status Monitors

Move cursor to desired item and press Enter.

  Remote Physical Volume Client Statistics
  Geographically Mirrored Volume Group Status
  Asynchronous Mirroring Statistics
```

Figure 8-51 Status Monitors panel

Select **Remote Physical Volume Client Status** to display the SMIT interface for the **rpvstat** command for physical volumes. Select **Geographically Mirrored Volume Group Status** to display the SMIT interface for the **gmvgstat** command. Select **Asynchronous Mirroring Statistics** to display the SMIT interface for the **rpvstat** command for asynchronous monitoring.

SMIT interface for rpvstat

The following SMIT interfaces show how to use the **rpvstat** command. Select **Remote Physical Volume Client Statistics** from the Status Monitors panel to display the main SMIT interface panel for the **rpvstat** command (fastpath: **rpvstat**) (Figure 8-52).

```
                    Remote Physical Volume Client Statistics

Move cursor to desired item and press Enter.

  Display Remote Physical Volume Client Statistics
  Display Remote Physical Volume Client Statistics by Network
  Display Remote Physical Volume Client Maximum Pending Statistics
  Reset Remote Physical Volume Client Statistics
```

Figure 8-52 Remote Physical Volume Client Statistics panel

Select **Display Remote Physical Volume Client Statistics** to display the following panel (fastpath: **rpvstat_dialog**) (Figure 8-53).

```
                Display Remote Physical Volume Client Statistics

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Specific RPV(s) to display (leave blank for all)   [] +
  Display Statistics for Individual Mirroring Networ  no +
  ks?
  Include date and time in display?                   no +
  Monitor interval in seconds                        [] #
  Number of times to monitor                         [] #
  Display applicable monitored statistics as delta    no +
   values?
```

Figure 8-53 Display Remote Physical Volume Client Statistics panel

The **rpvstat** command has the following fields:

Specific RPVs to display (Leave blank for all.)
You can leave this field empty to display information for all RPV clients. Pressing PF4 provides a list from which you can use PF7 to select one or more RPV clients for which to display information. Or you can manually enter the names of one or more RPV clients in the entry field.

Display statistics for individual mirroring networks?
You can alternately select no or yes.

Include date and time in display?
You can alternately select no or yes.

Monitor interval in seconds
This field is optional and requires an integer value greater than zero and less than or equal to 3600.

Number of times to monitor
This field is optional, but requires that the monitor interval in seconds field have a value. The Number of times to monitor field requires an integer value greater than 0 and less than or equal to 999999.

Display applicable monitored statistics as delta values?
You can alternately select no or yes. This field applies only if the monitor interval in seconds field has a value.

After the fields are completed, run the **rpvstat** command to display statistical information for all of the indicated RPV clients.

Select **Display Remote Physical Volume Client Statistics by Network** on the previous Remote Physical Volume Client Statistics panel to display the following panel (fastpath: rpvstat_net_dialog) (Figure 8-54).

```
          Display Remote Physical Volume Client Statistics by Network

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
  Include date and time in display?                    no +
  Monitor interval in seconds                         [] #
  Number of times to monitor                          [] #
  Display applicable monitored statistics as delta     no +
    values?
```

Figure 8-54 Display Remote Physical Volume Client Statistics by Network panel

You can select from the following options for the field values to display RPV client statistics by network:

Include date and time in display?
You can alternately select no or yes.

Monitor interval in seconds
This field is optional. This field requires an integer value greater than zero and less than or equal to 3600.

Number of times to monitor
This field is optional, but requires the Include date and time in display

field to have a value. The Monitor interval in seconds field requires an integer value greater than 0 and less than or equal to 999999.

Display applicable monitored statistics as delta values?
You can alternately select no or yes. This field applies only if the Monitor interval in seconds field has a value.

After the fields are completed, run the **rpvstat -N** command to display statistical information for all of the remote mirroring networks. Select **Display Remote Physical Volume Client Maximum Pending Statistics** on the previous Remote Physical Volume Client Statistics panel to display the panel (fastpath: **rpvstat_pending_dialog**) shown in Figure 8-55.

```
        Display Remote Physical Volume Client Maximum Pending Statistics

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
  Specific RPV(s) to display (leave blank for all)   [] +
  Display Statistics for Individual Mirroring Networ  no +
  ks?
  Include date and time in display?                   no +
```

Figure 8-55 Display Remote Physical Volume Client Maximum Pending Statistics panel

You can choose from the following options to display RPV client maximum pending statistics:

Specific RPVs to display (Leave blank for all.)
You can leave this field blank to display statistics for all RPV clients. Pressing PF4 provides a list from which you can use PF7 to select one or more RPV clients. You can also manually enter the names of one or more RPV clients in this field.

Display statistics for individual mirroring networks
You can alternately select no or yes.

Include date and time in display?
You can alternately select no or yes.

After the fields are completed, press Enter to run the **rpvstatus -m** command to display high water mark statistical information for pending statistics.

Select **Reset RPV Client statistics** on the previous Remote Physical Volume Client Statistics panel to display the panel (fastpath: **rpvstat_reset_dialog**) (Figure 8-56).

```
                Reset Remote Physical Volume Client Statistics

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Specific RPV(s) to reset (leave blank for all)     []
```

Figure 8-56 Reset Remote Physical Volume Client Statistics panel

Run the **rpvstat -R** command to reset the statistical counters in the indicated RPV clients.

SMIT interface for rpvstat for asynchronous mirroring

The following SMIT interfaces use the **rpvstat** command for asynchronous mirroring statistics. Selecting **Asynchronous Mirroring Statistics** brings up the panel in Figure 8-57.

```
                        Asynchronous Mirroring Statistics

Move cursor to desired item and press Enter.

  Display Asynchronous I/O Progress Statistics
  Display Asynchronous I/O Cache Statistics
  Reset Asynchronous I/O Statistics
```

Figure 8-57 Asynchronous Mirroring Statistics panel

Select **Display Asynchronous I/O Progress Statistics**, and the Display Asynchronous I/O Progress Statistics panel (Figure 8-58) opens.

```
                  Display Asynchronous I/O Progress Statistics

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Specific RPV(s) to display (leave blank for all)   [] +
  Include date and time in display?                   no +
  Monitor interval in seconds                        [] #
  Number of times to monitor                         [] #
  Display applicable monitored statistics as delta    no +
   values?
```

Figure 8-58 Display Asynchronous I/O Progress Statistics panel

You can choose from the following options to display asynchronous I/O progress statistics:

Specific RPVs to display (Leave blank for all.)
: You can leave this field empty to display information for all RPV clients. Pressing PF4 provides a list from which you can use PF7 to select one or more RPV clients for which to display information. Or, you can manually enter the names of one or more RPV clients in the entry field.

Include date and time in display?
: You can alternately select no or yes.

Monitor interval in seconds
: This field is optional and requires an integer value greater than zero and less than or equal to 3600.

Number of times to monitor
: This field is optional, but requires that the Monitor interval in seconds field has a value. The Number of times to monitor field requires an integer value greater than 0 and less than or equal to 999999.

Display applicable monitored statistics as delta values?
: You can alternately select no or yes. This field applies only if the Monitor interval in seconds field has a value.

Selecting **Display Asynchronous I/O Cache Statistics** brings up the panel in Figure 8-59.

```
                  Display Asynchronous I/O Cache Statistics

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Specific RPV(s) to display (leave blank for all)   [] +
  Include date and time in display?                   no +
  Monitor interval in seconds                        [] #
 Number of times to monitor              []
# Display applicable monitored statistics as delta    no +
   values?
```

Figure 8-59 Display Asynchronous I/O Cache Statistics panel

You can choose from the following options to display asynchronous I/O cache statistics:

Include date and time in display?
You can alternately select no or yes.

Monitor interval in seconds
This field is optional and requires an integer value greater than zero and less than or equal to 3600.

Number of times to monitor
This field is optional, but requires that the Monitor interval in seconds field has a value. The Number of times to monitor field requires an integer value greater than 0 and less than or equal to 999999.

Display applicable monitored statistics as delta values?
You can alternately select no or yes. This field applies only if the Monitor interval in seconds field has a value.

Selecting **Reset Asynchronous IO Statistics** runs the **`rpvstat -A -R`** and **`rpvstat -C -R`** commands to reset all statistics.

SMIT interface for gmvgstat

The following SMIT interfaces show how to use the **`gmvgstat`** command.

Select **Geographically Mirrored Volume Group Status** from the new Status Monitors panel to display the next panel (fastpath: **`gmvgstat`**) (Figure 8-60).

```
                Geographically Mirrored Volume Group Status

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Specific GMVG(s) to display (leave blank for all)  [] +
  Include associated RPV Client statistics?           no +
  Include header with display?                        no +
  Monitor interval in seconds                        [] #
  Number of times to monitor                         [] #
```

Figure 8-60 Geographically Mirrored Volume Group Status panel

The following options are available to display the status of a geographically mirrored volume group:

Specific RPVs to display (Leave blank for all.)
You can leave this field empty to display information for all valid, available, online GMVGs. Pressing PF4 provides a list from which you can use PF7 to select one or more GMVGs for which to display information. You can also manually enter the names of one or more GMVGs in this field.

Include associated RPV Client statistics?
You can alternately select no or yes.

Include header with display?
You can alternately select no or yes.

Monitor interval in seconds
This field is optional and requires an integer value greater than zero and less than or equal to 3600.

Number of times to monitor
This field is optional, but requires a value in the Monitor interval in seconds field. The Number of times to monitor field requires an integer value greater than 0 and less than or equal to 999999.

After the fields are complete, run the **gmvgstat** command to display statistical information for all of the indicated GMVGs.

8.9 Test scenarios

This section describes test scenarios.

8.9.1 Graceful site failover

In this section, we move the resource group yuriRG primary ONLINE at Asite to Bsite. This operation usually is performed during a planned outage at the primary site. Example 8-29 shows the current state of the cluster resource groups. The cluster manager performs the following sequence of operations:

1. Releases the primary online resource group in the primary site
2. Releases the secondary online resource group on the secondary site
3. Acquires the resource group in online primary state in the secondary site
4. Acquires the resource group in secondary online state in the primary site, if available

Example 8-29 Initial resource group status in the GLVM cluster

```
root@GLVM_A1 / > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
GLVM_A1_RG     ONLINE                       GLVM_A1
               OFFLINE                      GLVM_A2

GLVM_A2_RG     ONLINE                       GLVM_A2
               OFFLINE                      GLVM_A1

GLVM_B1_RG     ONLINE                       GLVM_B1
```

```
              OFFLINE                    GLVM_B2

GLVM_B2_RG    ONLINE                     GLVM_B2
              OFFLINE                    GLVM_B1

YuriRG        ONLINE                     GLVM_A1@Asite
              OFFLINE                    GLVM_A2@Asite
              ONLINE SECONDARY           GLVM_B1@Bsite
              OFFLINE                    GLVM_B2@Bsite

YunaRG        ONLINE                     GLVM_B1@Bsite
              OFFLINE                    GLVM_B2@Bsite
              ONLINE SECONDARY           GLVM_A1@Asite
              OFFLINE                    GLVM_A2@Asite
```

Run the **smitty hacmp** command. Then, select **System Management (C-SPOC)** → **Resource Groups and Applications** → **Move a Resource Group to Another Node / Site** → **Move Resource Group(s) to Another Site** (Example 8-30).

Example 8-30 Move YuriRG to Bsite

```
Move a Resource Group to Another Node / Site

Move cursor to desired item and press Enter.

  Move Resource Groups to Another Node
  Move Resource Groups to Another Site

+--------------------------------------------------------------------------+
Select Resource Group(s)
Move cursor to desired item and press Enter.                               |
#
# Resource Group                 State               Node(s) / Site
#
YuriRG                           ONLINE              GLVM_A1 / Asite
YuriRG                           ONLINE SECONDARY    GLVM_B1 / Bsite
YunaRG                           ONLINE              GLVM_B1 / Bsite
YunaRG                           ONLINE SECONDARY    GLVM_A1 / Asite
#
# Resource groups in node or site collocation configuration:
# Resource Group(s)                          State    Node / Site
#
+--------------------------------------------------------------------------+

Move Resource Group(s) to Another Site

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

                                                      [Entry Fields]
  Resource Group(s) to be Moved                      YuriRG
  Destination Site                                   Bsite
 Name of site containing data to be preserved
+
```

```
  by data divergence recovery processing
  (Asynchronous GLVM Mirroring only)
```

Check the resource group status after the resource group move. In Example 8-31, YuriRG has been moved to GLVM_B2 and GLVM_A1 becomes the ONLINE SECONDARY state.

Example 8-31 Resource group status after RG move in the GLVM cluster

```
root@GLVM_A1 / > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
GLVM_A1_RG     ONLINE                       GLVM_A1
               OFFLINE                      GLVM_A2

GLVM_A2_RG     ONLINE                       GLVM_A2
               OFFLINE                      GLVM_A1

GLVM_B1_RG     ONLINE                       GLVM_B1
               OFFLINE                      GLVM_B2

GLVM_B2_RG     ONLINE                       GLVM_B2
               OFFLINE                      GLVM_B1

YuriRG         ONLINE SECONDARY             GLVM_A1@Asite
               OFFLINE                      GLVM_A2@Asite
               OFFLINE                      GLVM_B1@Bsite
               ONLINE                       GLVM_B2@Bsite

YunaRG         ONLINE                       GLVM_B1@Bsite
               OFFLINE                      GLVM_B2@Bsite
               ONLINE SECONDARY             GLVM_A1@Asite
               OFFLINE                      GLVM_A2@Asite
```

8.9.2 Total site loss

In this section, we test a total site loss scenario. We initiate the scenario by running the **halt** command, expecting an unplanned outage at the primary site (Figure 8-61).

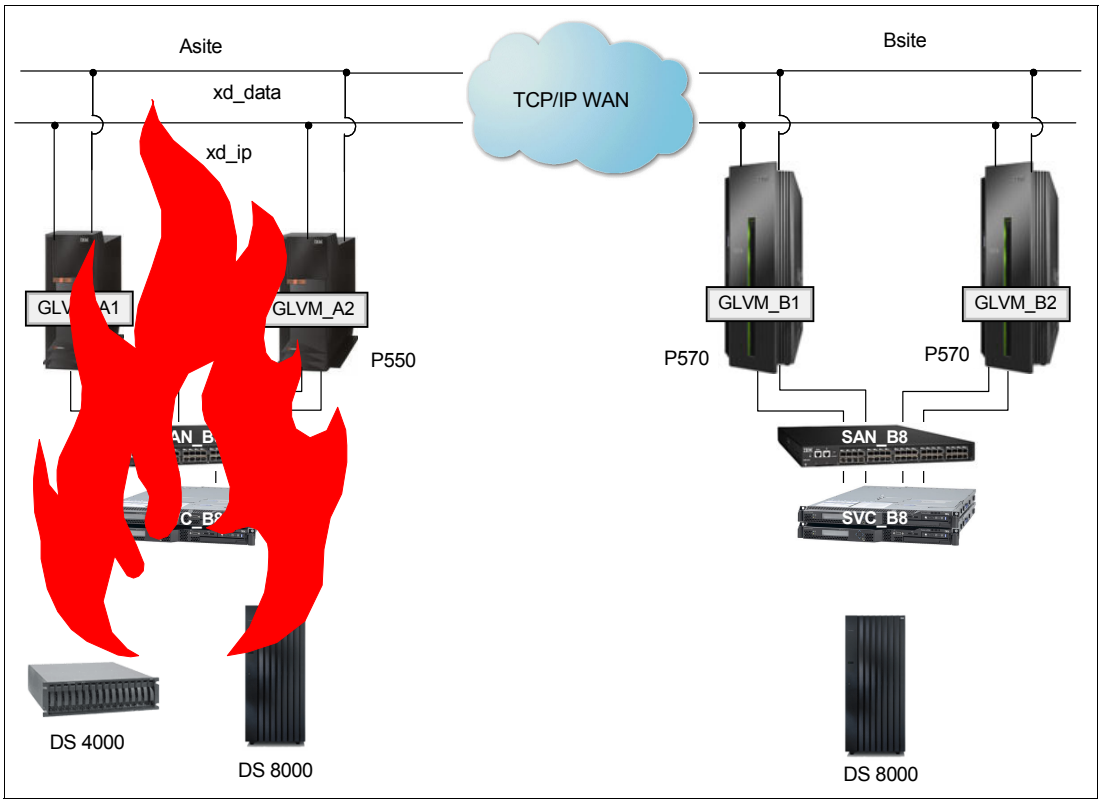


Figure 8-61 Total site (Asite) loss in the GLVM_cluster

Example 8-32 shows the current state of the cluster resource groups. We run the **halt** command in each node, GLVM_A1 and GLVM_A2.

Example 8-32 Initial resource group status in the GLVM cluster

```
root@GLVM_A1 / > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
GLVM_A1_RG     ONLINE                       GLVM_A1
               OFFLINE                      GLVM_A2

GLVM_A2_RG     ONLINE                       GLVM_A2
               OFFLINE                      GLVM_A1

GLVM_B1_RG     ONLINE                       GLVM_B1
               OFFLINE                      GLVM_B2

GLVM_B2_RG     ONLINE                       GLVM_B2
               OFFLINE                      GLVM_B1

YuriRG         ONLINE                       GLVM_A1@Asite
               OFFLINE                      GLVM_A2@Asite
```

```
                ONLINE SECONDARY            GLVM_B1@Bsite
                OFFLINE                     GLVM_B2@Bsite

YunaRG          ONLINE                      GLVM_B1@Bsite
                OFFLINE                     GLVM_B2@Bsite
                ONLINE SECONDARY            GLVM_A1@Asite
                OFFLINE                     GLVM_A2@Asite
```

When the cluster manager detects a site failure, the following procedure is performed:

1. Detects that the primary site is down
2. Releases the secondary online resource group on the secondary site
3. Acquires the resource group in the online primary state at the secondary site

After a site failure, YuriRG is in ONLINE state at the GLVM_B1 node. The GLVM_A1_RG and GLVM_A2_RG are in OFFLINE status because the resource groups are set to ignore inter-site management policy, which means that the resource groups are only in the site not taking over the remote site (Example 8-33).

Example 8-33 Resource group status after total site loss in Asite

```
root@GLVM_B1 / > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
GLVM_A1_RG     OFFLINE                     GLVM_A1
               OFFLINE                     GLVM_A2

GLVM_A2_RG     OFFLINE                     GLVM_A2
               OFFLINE                     GLVM_A1

GLVM_B1_RG     ONLINE                      GLVM_B1
               OFFLINE                     GLVM_B2

GLVM_B2_RG     ONLINE                      GLVM_B2
               OFFLINE                     GLVM_B1

YuriRG         OFFLINE                     GLVM_A1@Asite
               OFFLINE                     GLVM_A2@Asite
               ONLINE                      GLVM_B1@Bsite
               OFFLINE                     GLVM_B2@Bsite

YunaRG         ONLINE                      GLVM_B1@Bsite
               OFFLINE                     GLVM_B2@Bsite
               OFFLINE                     GLVM_A1@Asite
               OFFLINE                     GLVM_A2@Asite
```

8.9.3 XD_data network loss

In this section, we demonstrate all XD_data network loss between sites by pulling out the network cables between sites (Figure 8-62).

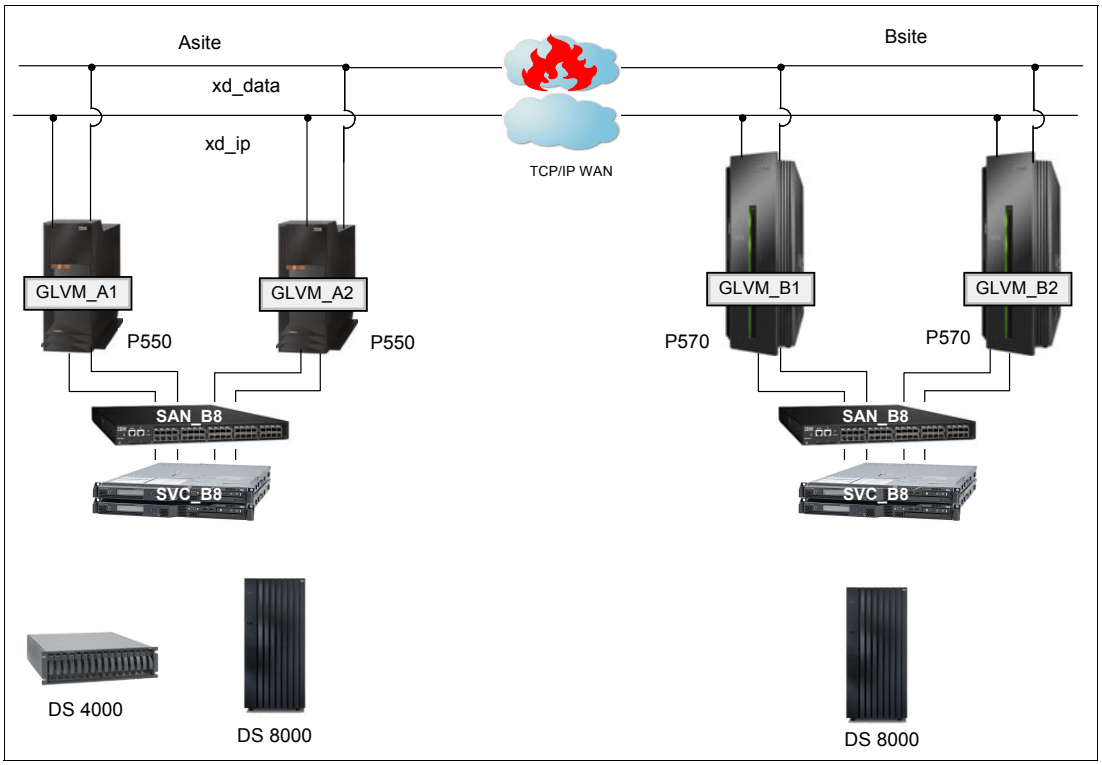


Figure 8-62 XD_data network loss

When we lose all XD_data networks, the cluster manager performs a site isolation. You notice the error messages in the errpt log shown in Example 8-34. The RPV disks are in the missing state. However, since XD_ip Network is still alive, the resource group and application are still ONLINE on both sites.

Example 8-34 errpt after losing XD_data networks

```
root@GLVM_A1 > errpt
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
F7DDA124   0323140510 U H LVDD           PHYSICAL VOLUME DECLARED MISSING
52715FA5   0323140510 U H LVDD           FAILED TO WRITE VOLUME GROUP STATUS AREA
E86653C3   0323140510 P H LVDD           I/O ERROR DETECTED BY LVM
F7DDA124   0323140510 U H LVDD           PHYSICAL VOLUME DECLARED MISSING
52715FA5   0323140510 U H LVDD           FAILED TO WRITE VOLUME GROUP STATUS AREA
E86653C3   0323140510 P H LVDD           I/O ERROR DETECTED BY LVM
D1E21BA3   0323140510 I S errdemon       LOG FILE EXPANDED TO REQUESTED SIZE
EAA3D429   0323140510 U S LVDD           PHYSICAL PARTITION MARKED STALE
E86653C3   0323140510 P H LVDD           I/O ERROR DETECTED BY LVM
....
root@GLVM_A1 > lsvg -p yurivg
yurivg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk1            active            1275        1265 255..245..255..255..255
hdisk2            active            1275        1274 255..254..255..255..255
hdisk5            missing           1275        1265 255..245..255..255..255
```

```
hdisk6            missing           1275        1274 255..254..255..255..255
root@GLVM_A1 > lsvg -l yurivg
yurivg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lvyuri              jfs2       10      20      2    open/stale    /yuri
loglv00             jfs2log    1       2       2    open/stale    N/A
....
root@GLVM_A1 / > clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
GLVM_A1_RG     ONLINE                       GLVM_A1
               OFFLINE                      GLVM_A2

GLVM_A2_RG     ONLINE                       GLVM_A2
               OFFLINE                      GLVM_A1

GLVM_B1_RG     ONLINE                       GLVM_B1
               OFFLINE                      GLVM_B2

GLVM_B2_RG     ONLINE                       GLVM_B2
               OFFLINE                      GLVM_B1

YuriRG         ONLINE                       GLVM_A1@Asite
               OFFLINE                      GLVM_A2@Asite
               ONLINE SECONDARY             GLVM_B1@Bsite
               OFFLINE                      GLVM_B2@Bsite

YunaRG         ONLINE                       GLVM_B1@Bsite
               OFFLINE                      GLVM_B2@Bsite
               ONLINE SECONDARY             GLVM_A1@Asite
               OFFLINE                      GLVM_A2@Asite
```

When the XD_data network is up again, PowerHA merges the sites automatically, brings the RPV disks to the active state, and synchronizes the data.

8.9.4 Storage loss at one site

In this section, we initiate a storage loss at one site by changing the zoning configuration (Figure 8-63). Unlike the PowerHA Enterprise Edition solution, the GLVM cluster does not fail over to the remote site because the cluster still has a good copy of the data on the RPV disks that are physically on the remote site and the quorum has been turned off.

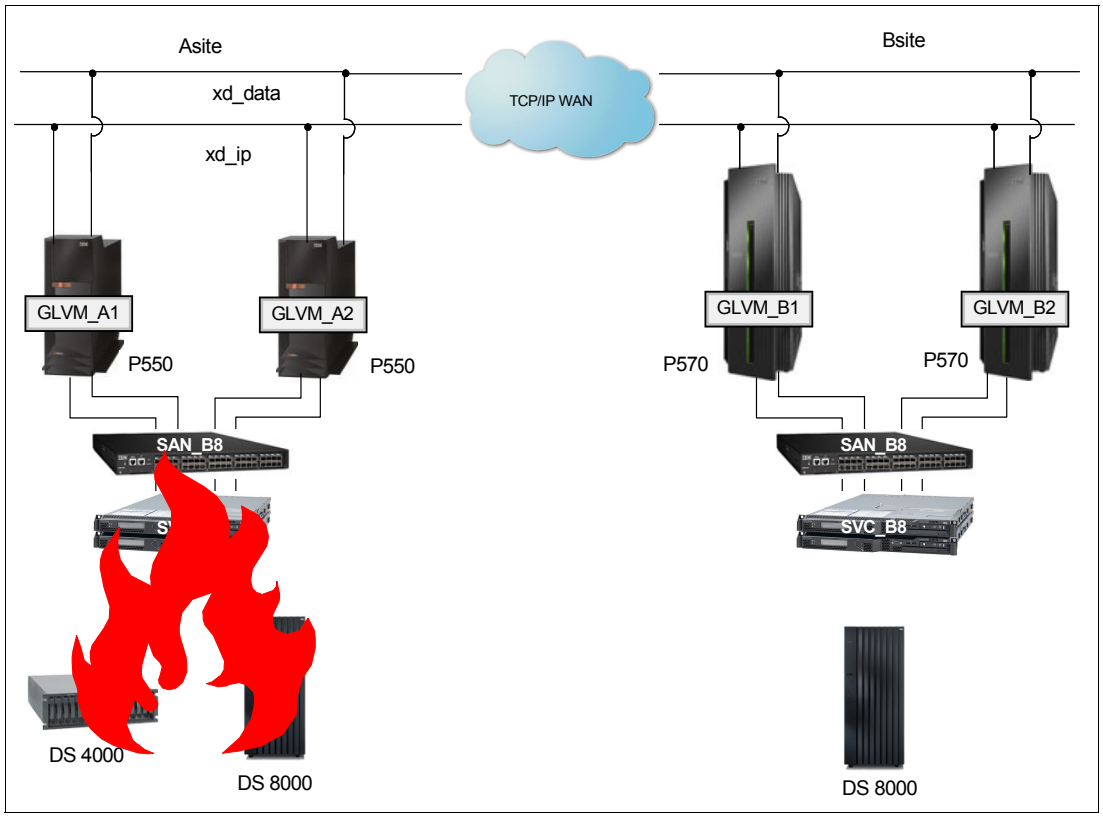


Figure 8-63 Storage loss at local site

After the local storage loss at Asite (Example 8-35), the local PVs are in the missing state and the logical volumes are now stale. At the Bsite, the RPVs are in the missing state, and the logical volumes are in the stale state.

Example 8-35 VG status after local storage loss in Asite

```
root@GLVM_A1 / > lsvg -p yurivg
yurivg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk1            missing           1275        1265 255..245..255..255..255
hdisk2            missing           1275        1274 255..254..255..255..255
hdisk5            active            1275        1265 255..245..255..255..255
hdisk6            active            1275        1274 255..254..255..255..255
root@GLVM_A1 / > lsvg -l yurivg
yurivg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lvyuri              jfs2       10      20      2    open/stale    /yuri
loglv00             jfs2log    1       2       2    open/stale    N/A
....
root@GLVM_B1 / > lsvg -p yunavg
yunavg:
```

```
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk1            active            1275        1265 255..245..255..255..255
hdisk2            active            1275        1274 255..254..255..255..255
hdisk7            missing           1275        1265 255..245..255..255..255
hdisk8            missing           1275        1274 255..254..255..255..255
root@GLVM_B1 / > lsvg -l yunavg
yunavg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lvyuna              jfs2       10      20      2    open/stale    /yuna
loglv01             jfs2log    1       2       2    open/stale    N/A
```

When the local storage is brought backup online, run the **varyonvg** command to enable the local physical volume to set it to the active state, and run the **syncvg** for the data synchronization (Example 8-36).

Example 8-36 Recover from local storage lost

```
root@GLVM_A1 / > varyonvg yurivg
root@GLVM_A1 / > lsvg -p yurivg
yurivg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk1            active            1275        1265 255..245..255..255..255
hdisk2            active            1275        1274 255..254..255..255..255
hdisk5            active            1275        1265 255..245..255..255..255
hdisk6            active            1275        1274 255..254..255..255..255
root@GLVM_A1 / > lsvg -l yurivg
yurivg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lvyuri              jfs2       10      20      2    open/stale    /yuri
loglv00             jfs2log    1       2       2    open/stale    N/A
root@GLVM_A1 / > syncvg -v yurivg
0516-1296 lresynclv: Unable to completely resynchronize volume.
        The logical volume has bad-block relocation policy turned off.
        This may have caused the command to fail.
0516-934 /usr/sbin/syncvg: Unable to synchronize logical volume lvyuri.
0516-932 /usr/sbin/syncvg: Unable to synchronize volume group yurivg.
0516-1296 lresynclv: Unable to completely resynchronize volume.
        The logical volume has bad-block relocation policy turned off.
        This may have caused the command to fail.
0516-934 /usr/sbin/syncvg: Unable to synchronize logical volume lvyuri.
0516-932 /usr/sbin/syncvg: Unable to synchronize volume group yurivg.
root@GLVM_A1 / > lsvg -l yurivg
yurivg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lvyuri              jfs2       10      20      2    open/syncd    /yuri
loglv00             jfs2log    1       2       2    open/syncd    N/A
```

However, the volume group is asynchronously mirrored. Vary off and vary on the volume group with the **varyonvg** command with the -k [rem / loc] option:

```
varyonvg -k rem yurivg
```

Alternatively, you can bring the resource group that has the geographically mirrored volume group offline and online again.

8.10 Performing management operations on the cluster

This section demonstrates how to perform management operations in the cluster.

8.10.1 Converting synchronous GMVGs to asynchronous GMVGs

If you already have an existing GLVM volume group that is configured for synchronous mirroring, you might decide to reconfigure it for asynchronous mirroring also. In our example, the existing volume group yurivg is reconfigured for asynchronous mirroring as follows.

Note: Asynchronous GMVGs require volume groups with strict mirror pools enabled, and the local physical volumes and GLVM remote physical volumes that are configured into separate mirror pools. Therefore, when a site is lost, a full copy of the data is available. For more information, see 8.2.2, “Configuring geographically mirrored volume groups” on page 344.

1. Bring the resource group offline that has the volume group yurivg. Run **smitty hacmp** command and select **System Management (C-SPOC)** → **Resource Groups and Applications** → **Bring a Resource Group Offline**. Select both the **ONLINE** and the **ONLINE SECONDARY** states of YuriRG (Example 8-37).

Example 8-37 Bringing a resource group offline

```
Bring a Resource Group Offline

Move cursor to desired item and press Enter.
Resource Groups and Applications

+--------------------------------------------------------------------------+
Select Resource Group(s)
Move cursor to desired item and press Enter.                               |
#
# Resource Group                  State                Node(s) / Site
#
YuriRG                            ONLINE               GLVM_A1 / Asite
YuriRG                            ONLINE SECONDARY     GLVM_B1 / Bsite
YunaRG                            ONLINE               GLVM_B1 / Bsite
YunaRG                            ONLINE SECONDARY     GLVM_A1 / Asite
#
# Resource groups in node or site collocation configuration:
# Resource Group(s)                           State    Node / Site
#
+--------------------------------------------------------------------------+
```

2. Configure related RPV servers and RPV clients into the available state on all nodes at each site (Example 8-38 on page 413).

RPV servers and clients that are related to yurivg: You must configure all RPV servers and clients that are related to yurivg into the available state on all nodes throughout the cluster. Otherwise, the import volume group in step 11 does not work.

Example 8-38 Configuring RPV clients and its server for yurivg

```
root@GLVM_A1 / > lsdev -Cc disk
hdisk0 Available          Virtual SCSI Disk Drive
hdisk1 Available 32-T1-01 MPIO FC 2145
hdisk2 Available 32-T1-01 MPIO FC 2145
hdisk3 Available 22-T1-01 MPIO FC 2145
hdisk4 Available 22-T1-01 MPIO FC 2145
hdisk5 Defined            Remote Physical Volume Client
hdisk6 Defined            Remote Physical Volume Client
hdisk7 Defined            Remote Physical Volume Client
hdisk8 Defined            Remote Physical Volume Client
root@GLVM_A1 / > mkdev -l hdisk5
hdisk5 Available
root@GLVM_A1 / > mkdev -l hdisk6
hdisk6 Available
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg
hdisk2          000fe4112f998235                    yurivg
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    yunavg
hdisk5          00c0f6a02fae3172                    yurivg
hdisk6          00c0f6a02fae31dd                    yurivg
......
root@GLVM_B1 / > mkdev -l rpvserver0
rpvserver0 Available
root@GLVM_B1 / > mkdev -l rpvserver1
rpvserver1 Available
root@GLVM_B1 / > lsattr -El rpvserver0
auto_online n                                 Configure at System Boot   True
client_addr 10.10.101.108                     Client IP Address          True
client_addr 10.10.101.107                     Client IP Address          True
client_addr 10.10.102.107                     Client IP Address          True
client_addr 10.10.102.108                     Client IP Address          True
rpvs_pvid   00c0f6a02fae31720000000000000000 Physical Volume Identifier True
root@GLVM_B1 / > lsattr -El rpvserver1
auto_online n                                 Configure at System Boot   True
client_addr 10.10.101.108                     Client IP Address          True
client_addr 10.10.101.107                     Client IP Address          True
client_addr 10.10.102.107                     Client IP Address          True
client_addr 10.10.102.108                     Client IP Address          True
rpvs_pvid   00c0f6a02fae31dd00000000000000000 Physical Volume Identifier True
```

3. Convert the volume group to scalable VG format by using the **chvg -G** command. The volume group must be varied off before you run the command.

   ```
   root@GLVM_A1 / > chvg -G yurivg
   ```

4. Turn off bad block relocation. Asynchronous mirroring requires that the bad block relocation policy be turned off for the volume group. You can turn off bad block relocation by using the **chvg -b n** command.

   The volume group must be varied on, and the logical volumes must be in the closed state to turn off the bad block relocation:

   ```
   root@GLVM_A1 / > varyonvg yurivg
   root@GLVM_A1 / > chvg -b n yurivg
   ```

5. Configure the volume group to use super strict mirror pools. Asynchronous mirroring requires the volume group to be configured to use super strict mirror pools. You can configure super strict mirror pools by using the **chvg -M s** command:

```
root@GLVM_A1 / > chvg -M s yurivg
```

6. Configure a mirror pool of disks at each site. Asynchronous mirroring also requires the local and remote disks to belong to separate mirror pools. Set a mirror pool for the local disks and remote disks:

```
root@GLVM_A1 / > lsvg -p yurivg
yurivg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk1            active            1275        1265 255..245..255..255..255
hdisk2            active            1275        1274 255..254..255..255..255
hdisk5            active            1275        1265 255..245..255..255..255
hdisk6            active            1275        1274 255..254..255..255..255
root@GLVM_A1 / > chpv -p Asite hdisk1 hdisk2
root@GLVM_A1 / > chpv -p Bsite hdisk5 hdisk6
```

7. Configure the logical volumes to belong to the new mirror pool with bad block relocation turned off. A second mirror copy of the logical volume is on the Bsite mirror pool:

```
root@GLVM_A1 / > chlv -m copy1=Asite -m copy2=Bsite -b n lvyuri
root@GLVM_A1 / > chlv -m copy1=Asite -m copy2=Bsite -b n loglv00
```

8. Create a logical volume of type aio_cache for each mirror pool. Asynchronous mirroring requires a logical volume of type aio_cache to serve as the cache device.

```
root@GLVM_A1 /> mklv -y datacachelv1 -t aio_cache -p copy1=Asite -s s -u 1024
-b n yurivg 10
datacachelv1
root@GLVM_A1 /> mklv -y datacachelv2 -t aio_cache -p copy1=Bsite -s s -u 1024
-b n yurivg 10
datacachelv2
```

   The datacachelv1 that is in the Asite mirror pool is used for cache during asynchronous mirroring to the disks in the Bsite mirror pool. The datacachelv2 that is in the Bsite mirror pool is used for cache during asynchronous mirroring to the disks in the Asite mirror pool.

9. Convert to asynchronous mirroring for mirror pools:

```
root@GLVM_A1 / > chmp -A -m Asite yurivg
root@GLVM_A1 / > chmp -A -m Bsite yurivg
```

   In this example, the Asite and Bsite mirror pools are configured to use asynchronous mirroring. The **chmp** command automatically determines the datacachelv1 logical volume for the cache device to use for Bsite because it is in the opposite site's mirror pool. You can verify asynchronous mirroring configuration with the **lsmp** command (Example 8-39).

Example 8-39 Asynchronous mirroring configuration in yurivg

```
root@GLVM_A1 / > lsmp -A yurivg
VOLUME GROUP:       yurivg              Mirror Pool Super Strict: yes

MIRROR POOL:        Asite               Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: inactive            ASYNC CACHE LV:           datacachelv2
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        yes
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no

MIRROR POOL:        Bsite               Mirroring Mode:           ASYNC
ASYNC MIRROR STATE: active              ASYNC CACHE LV:           datacachelv1
```

```
ASYNC CACHE VALID:  yes                 ASYNC CACHE EMPTY:        no
ASYNC CACHE HWM:    100                 ASYNC DATA DIVERGED:      no
```

10. For all other nodes, export the existing yurivg and import again (Example 8-40). All RPV servers and clients must be in the available state.

Example 8-40 import asynchronous volume group in other nodes

```
root@GLVM_A1 / > varyoffvg yurivg
root@GLVM_B1 / > exportvg yurivg
root@GLVM_B1 / >  importvg -f -y yurivg -V 50 000fe4112f99817c
yurivg
root@GLVM_B1 / > lsvg -l yurivg
yurivg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
lvyuri              jfs2       10      20      2    closed/syncd  /yuri
loglv00             jfs2log    1       2       2    closed/syncd  N/A
datacachelv1        aio_cache  10      10      1    closed/syncd  N/A
datacachelv2        aio_cache  10      10      1    open/syncd    N/A
```

11. Change the attribute of the resource group to preserve the asynchronous geographically mirrored volume group from data divergency. If this value is not set, manually select a site for recovery after data divergence (Figure 8-64).

```
        Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Resource Group Name                                  YuriRG
  Inter-site Management Policy                         Prefer Primary Site
  Participating Nodes from Primary Site                GLVM_A1 GLVM_A2
  Participating Nodes from Secondary Site              GLVM_B1 GLVM_B2
Startup Policy                                        Online On Home Node Only
Failover Policy                                       Failover To Next
Priority Node In The List
Failback Policy                                       Failback To Higher
Priority Node In The List
Failback Timer Policy (empty is immediate)           [] +

  Service IP Labels/Addresses                         [] +
  Application Servers                                 [YuriApp] +
  Volume Groups                                       [yurivg ] +
  Use forced varyon of volume groups, if necessary     true +
  Automatically Import Volume Groups                   false +
  Allow varyon with missing data updates?              true +
      (Asynchronous GLVM Mirroring Only)
  Default choice for data divergence recovery          Bsite +
      (Asynchronous GLVM Mirroring Only)
```

Figure 8-64 Configure attribute of YuriRG for data divergency

Data divergence: The allow varyon with missing data updates determines whether HACMP event processing automatically allows data divergence to occur during a non-graceful site failover to the remote site. It is set to enable fully automatic site failover by default. The default choice for the Data divergence recovery entry field provides an opportunity to configure your most likely decision for data divergence recovery in advance. For more information, see 8.12, “Data divergence in PowerHA for GLVM” on page 429.

12. Run verification and synchronization and bring the resource group ONLINE at the primary site and ONLINE SECONDARY at the secondary site (Example 8-41).

Example 8-41 Bringing the resource group YuriRG online

```
Resource Groups and Applications
Bring a Resource Group Online
+--------------------------------------------------------------------------+
Select a Resource Group
Move cursor to desired item and press Enter.
YuriRG                             OFFLINE
YuriRG                           OFFLINE_SECONDARY
+--------------------------------------------------------------------------+

....

Bring a Resource Group Online

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Resource Group to Bring Online                      YuriRG
  Node on Which to Bring Resource Group Online        GLVM_B1
  Name of site containing data to be preserved +
  by data divergence recovery processing
  (Asynchronous GLVM Mirroring only)
```

8.10.2 Adding physical volumes to a running cluster

The current *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide*, SA23-1338, does not mention adding RPV disks to a running cluster. That guide explains only how to make changes to the cluster resources and GMVGs when the cluster is stopped. This section explains how to add RPV disks to a running cluster.

Procedure 1 adds disk to a site that holds the primary instance of the resource group (for example, where the volume group is varied online and the applications are running). Procedure 2 adds disks at the remote site (the site where the volume group is offline or the secondary instance of the resource group). If you want to add a disk at both sites, follow procedure 1 and procedure 2, one after the other.

In our example, we add a physical volume and its RPV clients into yunavg varied on Bsite.

Procedure 1

To add disks at the primary site (where the volume group is online and the applications are running):

1. On the node that has the volume group varied on, identify a new physical disk that will be added to the volume group. This disk must be accessible by all the nodes at that site

and check that all disks have PVIDs assigned. If not, assign the PVIDs by using **chdev -a pv=yes -l** (for example, **chdev -a pv=yes -l hdisk3**, where hdisk3 is a new disk added to the system).

In Example 8-42, we add a local physical volume hdisk2 (00c1f1702fab9da9) on Bsite to yunavg and configure a remote physical volume from hdisk4 (00c1f1702fab9da9) on Asite and add it to the volume group for geographic mirroring.

Example 8-42 Check physical disks to add into yunavg in Bsite

```
root@GLVM_B1 / > lspv
hdisk0          00c1f170e170c9cd                    rootvg          active
hdisk1          00c1f1702fab9d25                    yunavg          active
hdisk2          00c1f1702fab9da9                    None
hdisk3          00c0f6a02fae3172                    yurivg
hdisk4          00c0f6a02fae31dd                    yurivg
hdisk7          000fe4012f9a9f43                    yunavg          active
root@GLVM_B2 / > lspv
hdisk0          00c0f6a0684f5ab8                    rootvg          active
hdisk1          00c1f1702fab9d25                    yunavg
hdisk2          00c1f1702fab9da9                    None
hdisk3          00c0f6a02fae3172                    yurivg
hdisk4          00c0f6a02fae31dd                    yurivg
hdisk7          000fe4012f9a9f43                    yunavg
```

2. On the node that has the volume group varied on (holding the primary instance of the resource group), create an RPV server instance for this disk (Figure 8-65).

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Physical Volume Identifiers                          00c1f1702fab9da9
* Remote Physical Volume Client Internet Address
[10.10.101.107,10.10.102.108] +
  Configure Automatically at System Restart?         [no] +
  Start New Devices Immediately?                     [yes] +
```

Figure 8-65 Create RPV server for hdisk2 00c1f1702fab9da9 in GLVM_B1 node

3. Create an RPV server instance on all remaining nodes of this site (if any) by using the disk that is identified in Figure 8-66.

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Physical Volume Identifiers                          00c1f1702fab9da9
* Remote Physical Volume Client Internet Address
[10.10.101.108,10.10.102.108] +
  Configure Automatically at System Restart?         [no] +
  Start New Devices Immediately?                     [yes] +
```

Figure 8-66 Creating an RPV server for hdisk2 00c1f1702fab9da9in GLVM_B2 node

4. On the remote site node (holding the secondary instance of the RG), create an RPV client for this disk (Figure 8-67).

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.201.107
  Remote Physical Volume Local Internet Address       10.10.101.107
  Physical Volume Identifiers 00c1f1702fab9da90000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

root@GLVM_A1 / > chdev -l hdisk8 -a local_addr2=10.10.102.107 -a
server_addr2=10.10.202.107
hdisk8 changed
```

Figure 8-67 Create RPV client in GLVM_A1 node

5. Similarly, create an RPV client instance on all remaining nodes on the remote site (if any) (Figure 8-68).

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.201.108
  Remote Physical Volume Local Internet Address       10.10.101.108
  Physical Volume Identifiers 00c1f1702fab9da90000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

root@GLVM_A2 / > chdev -l hdisk8 -a local_addr2=10.10.102.108 -a
server_addr2=10.10.202.108
hdisk8 changed
```

Figure 8-68 Create RPV client in GLVM_A2

6. On the node with the volume group varied on, run the **extendvg** command from the command line to include the local physical volume in the volume group, or use the AIX SMIT menu for volume groups to include the local physical disk (Example 8-43).

Example 8-43 Extend volume group in GLVM_B1

```
root@GLVM_B1 / > extendvg yunavg hdisk2
root@GLVM_B1 / > lspv
hdisk0          00c1f170e170c9cd                    rootvg          active
hdisk1          00c1f1702fab9d25                    yunavg          active
hdisk2          00c1f1702fab9da9                    yunavg          active
hdisk3          00c0f6a02fae3172                    yurivg
hdisk4          00c0f6a02fae31dd                    yurivg
hdisk7          000fe4012f9a9f43                    yunavg          active
```

Procedure 2

To add disks to the backup site (where the volume group is offline and data is being mirrored remotely):

1. On one of the remote site nodes (holding the secondary instance of the resource group), identify a new physical disk to be added to the volume group. This disk must be accessible by all the nodes at that site. For example, on the remote node, GLVM_A1, a new physical disk, hdisk4, has PVID 00c1f1702fab9da9, and in GLVM_A2 as well (Example 8-44).

Example 8-44 Check physical disks to add into yunavg in Asite

```
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg          active
hdisk2          000fe4112f998235                    yurivg          active
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    None
hdisk5          00c0f6a02fae3172                    yurivg          active
hdisk6          00c0f6a02fae31dd                    yurivg          active
hdisk7          00c1f1702fab9d25                    yunavg
root@GLVM_A2 / > lspv
hdisk1          000fe401fd2e6dc4                    rootvg          active
hdisk0          000fe4112f99817c                    yurivg
hdisk2          000fe4112f998235                    yurivg
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    None
hdisk7          00c1f1702fab9d25                    yunavg
```

2. Create an RPV server instance for this disk (Figure 8-69).

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Physical Volume Identifiers                          000fe4012f9a9fcf
* Remote Physical Volume Client Internet Address
[10.10.201.107,10.10.202.107] +
  Configure Automatically at System Restart?         [no] +
  Start New Devices Immediately?                     [yes] +
```

Figure 8-69 Create RPV server for hdisk4 000fe4012f9a9fcf in GLVM_A1

3. Similarly, create an RPV server instance on all other nodes of the remote site (Figure 8-70).

```
                    Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Physical Volume Identifiers                          000fe4012f9a9fcf
* Remote Physical Volume Client Internet Address
[10.10.201.108,10.10.202.108] +
  Configure Automatically at System Restart?         [no] +
  Start New Devices Immediately?                     [yes] +
```

Figure 8-70 Create RPV server for hdisk4 000fe4012f9a9fcf in GLVM_A2

4. On the node that has the volume group varied on (primary instance of the resource group), create an RPV client instance for this disk (Figure 8-71).

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.101.107
  Remote Physical Volume Local Internet Address       10.10.201.107
  Physical Volume Identifiers 000fe4012f9a9fcf0000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

root@GLVM_B1 / > chdev -l hdisk8 -a local_addr2=10.10.202.107 -a
server_addr2=10.10.102.107
hdisk8 changed
```

Figure 8-71 Creating an RPV client in GLVM_B1 node

5. Similarly, create an RPV client on all other nodes (if any) on this site (Example 8-72).

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Remote Physical Volume Server Internet Address      10.10.101.108
  Remote Physical Volume Local Internet Address       10.10.201.108
  Physical Volume Identifiers 000fe4012f9a9fcf0000000000000000
  I/O Timeout Interval (Seconds)                     [180] #
  Start New Devices Immediately?                     [yes] +

root@GLVM_B2 / > chdev -l hdisk8 -a local_addr2=10.10.202.108 -a
server_addr2=10.10.102.108
hdisk8 changed
```

Figure 8-72 Create RPV client in GLVM_B2

6. On the node that has the volume group varied on (holding the primary instance of the RG), extend the volume group to add the remote physical volume or disk to the volume group. Either use the **extendvg** command from the command line or use GLVM SMIT by running the **smitty glvm_utils** command and selecting **Geographic Logical Volume Groups** → **Add Remote Physical Volumes to a Volume Group** (Figure 8-73).

```
                Add Remote Physical Volumes to a Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* VOLUME GROUP name                                   yunavg
  Force                                              [no] +
* REMOTE PHYSICAL VOLUMES Name                       hdisk8
```

Figure 8-73 Extend volume group in GLVM_B1

Alternatively, enter the following command:

```
root@GLVM_B1 / > extendvg yunavg hdisk8
```

7. Run the **smitty hacmp** command and select **System Management → Storages → Volume Groups → Synchronize a Volume Group Definition**. This updates the volume group definition on all other nodes at each site. For example, after synchronizing a volume group definition from the GLVM_B1 node, the GLVM_A1 node is updated with newly extended yurivg (Example 8-45).

Example 8-45 Synchronize a volume group definition of yunavg

```
physical volume information in GLVM_A1 node before synchronization
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg          active
hdisk2          000fe4112f998235                    yurivg          active
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    None
hdisk5          00c0f6a02fae3172                    yurivg          active
hdisk6          00c0f6a02fae31dd                    yurivg          active
hdisk7          00c1f1702fab9d25                    yunavg
hdisk8          00c1f1702fab9da9                    None

...

physical volume information in GLVM_A1 node after synchronization
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg          active
hdisk2          000fe4112f998235                    yurivg          active
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    yunavg
hdisk5          00c0f6a02fae3172                    yurivg          active
hdisk6          00c0f6a02fae31dd                    yurivg          active
hdisk7          00c1f1702fab9d25                    yunavg
hdisk8          00c1f1702fab9da9                    yunavg
```

8.10.3 Removing physical volumes in a running cluster

Similarly, you can remove RPV disks in a running cluster. First, place all RPV servers and clients in an available state on all nodes for the physical disks that you want remove. Then, remove the disks by using AIX commands or SMIT GLVM utilities. After synchronization of the volume group, the definition of the volume group is updated on all other nodes at each site. In our example, we remove remote physical volumes from yunavg varied on Bsite (Example 8-46).

Example 8-46 Identify disks to remove in GLVM_B1

```
root@GLVM_B1 / > lsvg -p yunavg
yunavg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk1            active            1275        1265 255..245..255..255..255
hdisk2            active            1275        1265 255..245..255..255..255
hdisk7            active            1275        1265 255..245..255..255..255
hdisk8            active            1275        1265 255..245..255..255..255
```

To remove the physical volumes in a running cluster:

1. Configure all RPV clients and RPV servers that are related to the volume group yunavg in all nodes at each site (Example 8-47).

Example 8-47 Configuring RPV clients and servers of yunavg on GLVM_A2 for the available state

```
root@GLVM_A2 / > lspv
hdisk1          000fe401fd2e6dc4                    rootvg          active
hdisk0          000fe4112f99817c                    yurivg
hdisk2          000fe4112f998235                    yurivg
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    yunavg
root@GLVM_A2 / > mkdev -l hdisk7
hdisk7 Available
root@GLVM_A2 / > mkdev -l hdisk8
hdisk8 Available

root@GLVM_A2 / > lspv
hdisk1          000fe401fd2e6dc4                    rootvg          active
hdisk0          000fe4112f99817c                    yurivg
hdisk2          000fe4112f998235                    yurivg
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    yunavg
hdisk7          00c1f1702fab9d25                    yunavg
hdisk8          00c1f1702fab9da9                    yunavg
root@GLVM_A2 / > lsattr -El rpvserver2
auto_online n                                 Configure at System Boot   True
client_addr 10.10.201.108                     Client IP Address          True
client_addr 10.10.202.108                     Client IP Address          True
rpvs_pvid   00c1f1702fab9d250000000000000000 Physical Volume Identifier True
root@GLVM_A2 / > mkdev -l rpvserver2
rpvserver2 Available
root@GLVM_A2 / > lsattr -El rpvserver3
auto_online n                                 Configure at System Boot   True
client_addr 10.10.201.108                     Client IP Address          True
client_addr 10.10.202.108                     Client IP Address          True
rpvs_pvid   000fe4012f9a9fcf0000000000000000 Physical Volume Identifier True
root@GLVM_A2 / > mkdev -l rpvserver3
rpvserver3 Available
```

In the node that has the ONLINE state of the resource group YuriRG, all RPV clients are already in the available state. For example, GLVM_B1 has yurivg activated and RPV clients are active, enabling mirror to the remote copy.

In the node that has ONLINE SECONDARY of the resource group YuriRG, all RPV servers are already in the available state. For example, GLVM_A1 has all RPV servers in the available state.

2. Remove a remote physical volume by using the SMIT GLVM utilities (Figure 8-74).

```
              Remove Remote Physical Volumes From a Volume Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
* VOLUME GROUP name                                    yunavg
  FORCE deallocation of all partitions on this physi [no] +
  cal volume?
* Select One or more Remote Physical Volumes          hdisk8
```

Figure 8-74 Remove remote physical volume in GLVM_B1

Alternatively, enter the following command:

```
root@GLVM_B1 / > reducevg yunavg hdisk8
```

3. Run the **smitty hacmp** command. Select **System Management** → **Storages** → **Volume Groups** → **Synchronize a Volume Group Definition**. This updates the volume group definition on all other nodes at each site (Example 8-48).

Example 8-48 Synchronize a volume group definition of yunavg

```
physical volume information in GLVM_A1 node before synchronization
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg          active
hdisk2          000fe4112f998235                    yurivg          active
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    yunavg
hdisk5          00c0f6a02fae3172                    yurivg          active
hdisk6          00c0f6a02fae31dd                    yurivg          active
hdisk7          00c1f1702fab9d25                    yunavg
hdisk8          00c1f1702fab9da9                    yunavg

...

physical volume information in GLVM_A1 node after synchronization
root@GLVM_A1 / > lspv
hdisk0          000fe4111f25a1d1                    rootvg          active
hdisk1          000fe4112f99817c                    yurivg          active
hdisk2          000fe4112f998235                    yurivg          active
hdisk3          000fe4012f9a9f43                    yunavg
hdisk4          000fe4012f9a9fcf                    yunavg
hdisk5          00c0f6a02fae3172                    yurivg          active
hdisk6          00c0f6a02fae31dd                    yurivg          active
hdisk7          00c1f1702fab9d25                    yunavg
hdisk8          00c1f1702fab9da9                    None
```

8.11 Migration from HAGEO (AIX 5.3) to GLVM (AIX 6.1)

For those using HAGEO as their XD disaster recovery solution, a migration is required to get to GLVM. The IBM white paper, *Migration from HACMP/XD for HAGEO to PowerHA SystemMirror Enterprise Edition, GLVM Configuration*, was written by development and

explains the steps to migrate from HAGEO to GLVM. It addresses both a synchronous and an asynchronous HAGEO environment that is migrated to a synchronous or asynchronous GLVM environment. To download and read this paper, go to:

http://www.ibm.com/systems/power/software/availability/whitepapers/glvm.html

End of service: HAGEO is in end of service. For more information, go to:

http://www.ibm.com/software/support/systemsp/lifecycle/

The objective of the paper states:

> The white paper provides an overview of the HAGEO and GLVM technologies and compares their technical workings and their external interfaces. The requirements for both HAGEO and GLVM are reviewed. The different HAGEO mirroring configurations (sync, mwc, and async) are briefly reviewed and mapped to the mirroring configurations available in GLVM (sync and async). Migration prerequisites are presented along with the steps required before actually starting the migration process. The actual migration steps for two sample configurations (one with synchronous mirroring and another with asynchronous mirroring) are presented in detail.

The paper makes the following statements under the topic migration prerequisites:

> Migration from an HAGEO-based cluster to a GLVM-based cluster requires, for each PowerHA resource group, the removal of HAGEO and its artifacts followed by the configuration of GLVM. The application data can be retained at either site before introducing GLVM and remirroring the data to the other site. However, retain the data at the primary site (the production site where the application runs and is available most of the time) and remove it from the backup (recovery) site. PowerHA SystemMirror Enterprise Edition does not allow a resource group to simultaneously contain both HAGEO GMDs and GLVM GMVGs.

This means that the migration requires removing the HAGEO subsystem and then adding the new GLVM rpvservers and rpvclients. This method has the following drawbacks:

- When removing HAGEO, there is no longer a disaster recovery site in place.
- Adding GLVM requires a full synchronization of all the data across the XD_data network.
- If something goes wrong with the migration to GLVM, the back-out plan is to set up the HAGEO environment again, requiring full synchronization of the GMDS across the XD_data network.

An alternate solution is to set up GLVM while HAGEO is still in place. This method has its considerations as well:

- Since HAGEO is not supported on AIX 6.1, any asynchronous GMDS must first be migrated to synchronous GLVM.
- To have both GMDS and an rpvserver on the remote node, a second set of disks is required to contain the rpvservers. After the rpvservers are set up, the GMDS can be removed and the GMD disk space can be reused.

8.11.1 Overview of an alternative migration

As an alternative, use the following steps for migration:

1. Make backups of the PowerHA and HAGEO environments.
2. Create the rpvservers on the remote node on new disks and rpvclients on the local node. Add the rpvclient disks to the volume group to be mirrored.

3. Synchronize the logical volumes one at a time to minimize extra load on the XD_data network.
4. Complete the migration by removing the GMDS from the resource group, creating the rpvservers on the local site, and creating the rpvclients on the remote site.
5. Test the failover of the new GLVM environment.
6. Remove the GMD statemap logical volumes.
7. Follow PowerHA documentation to migrate to AIX 6.1 and asynchronous GLVM if required.

Tip: If a failure occurs and the resource group moves to the remote site, the GMDS and file systems will work as before. When it is time to move the resource group back to the primary site, make sure that the rpvclient disks are available by using the **mkdev -l <rpvdisk>** command. If you do not run this command, you see errors in errpt about stale partitions, LVM_SA_STALEPP.

You can still proceed if the rpvdisks are missing by issuing the **mkdev -l <rpvdisk>** and then **varyonvg -n <vg>**. This situation varies on the volume group, but does not sync the stale logical volumes. At a convenient time, continue the synchronization of the stale logical volumes with the **syncvg -l** command, monitoring the stale with the **lspv -p <rpvdisk>** command.

8.11.2 Adding GLVM to a running HAGEO environment

To add a GLVM to a running HAGEO environment:

1. Make a backup of the HACMP environment by using the **clsnapshot** command.
2. Make a backup of the HAGEO environment:

   ```
   /usr/sbin/gmd/geo_snapshot -t  -f 'oldhageo'
   ```

3. Define a remote physical volume server site name on one node on the remote site. Run the **smitty rpvserver** command and select **Remote Physical Volume Server Site Name Configuration → Define / Change / Show Remote Physical Volume Server Site Name**.
4. Create an rpvserver on one node on the remote site. Run the **smitty rpvserver** command, select **Remote Physical Volume Server Site Name Configuration → Add Remote Physical Volume Servers**, and select an *unused* disk that is not associated with any volume groups. This disk must be comparable to the disks used for GMD.
5. Create an rpvclient on a node on the local site. Run the **smitty rpvclient** command, select **Add Remote Physical Volume Clients**, and enter the XD_data IP of the rpvserver.
6. Add the RPV to the local volume group. Run the **smitty glvm_utils** command and select **Geographically Mirrored Volume Groups → Add Remote Physical Volumes to a Volume Group**.
7. Make sure that the logical volumes to be mirrored have the superstrict flag set (Example 8-49). This must be done for the statemaps also to get past error checking later when doing an HACMP verification and synchronization.

Example 8-49 Setting superstrict for logical volumes

```
> lslv datsm_data1
LOGICAL VOLUME:     datsm_data1            VOLUME GROUP:   geovg
LV IDENTIFIER:      0000858400004c0000000126e2b26aaf.3 PERMISSION:     read/write
VG STATE:           active/complete        LV STATE:       closed/syncd
```

```
TYPE:               statemap              WRITE VERIFY:   off
MAX LPs:            512                   PP SIZE:        16 megabyte(s)
COPIES:             2                     SCHED POLICY:   parallel
LPs:                1                     PPs:            2
STALE PPs:          0                     BB POLICY:      relocatable
INTER-POLICY:       minimum               RELOCATABLE:    yes
INTRA-POLICY:       middle                UPPER BOUND:    2
MOUNT POINT:        N/A                   LABEL:          None
MIRROR WRITE CONSISTENCY: on/ACTIVE
EACH LP COPY ON A SEPARATE PV ?: yes (superstrict)
Serialize IO ?:     NO
```

If `superstrict` is not on or the `Upper Bound` is not at least 2, change it by entering the following command:

```
chlv -s s -u 2 datsm_data1
```

8. Add a remote site mirror copy to the logical volume. HACMP requires all logical volumes on a volume group that is part of GLVM to be mirrored, which includes the statemap logical volumes. The statemaps logical volumes are removed later in the process. Run the **`smitty glvm_utils`** command. Select **Geographically Mirrored Logical Volumes** → **Add a Remote Site Mirror Copy to a Logical Volume**. Then, select the logical volume to be mirrored. Leave the option to SYNCHRONIZE set to no so that you can do it manually at a time when it does not affect the XD_DATA network.

   The **`lspv -p <rpvdisk>`** command shows where the stale partitions are (Example 8-50).

   Example 8-50 Checking for the stale partitions

```
> lspv -p hdisk9
hdisk9:
PP RANGE  STATE   REGION        LV NAME             TYPE       MOUNT POINT
  1-109   free    outer edge
110-129   used    outer middle  datlv_data1         jfs        N/A
130-130   used    outer middle  loglv_data1         jfslog     N/A
131-150   used    outer middle  datlv_data2         jfs        N/A
151-151   stale   outer middle  loglv_data2         jfslog     N/A
152-152   stale   outer middle  loglv00             jfs2log    N/A
153-154   stale   outer middle  fslv00              jfs2       N/A
155-217   free    outer middle
218-325   free    center
326-433   free    inner middle
434-542   free    inner edge
```

9. Synchronize the logical volumes one at time to minimize the data replication load on the XD network. Consider doing this step at off-peak times, depending on your network bandwidth. By using the **`syncvg -l <logical volume>`** command, you can synchronize just one logical volume copy (Example 8-51).

   Example 8-51 Sync one logical volume to the rpvserver

```
> syncvg -l loglv_data2

> lspv -p hdisk9
hdisk9:
PP RANGE  STATE   REGION        LV NAME             TYPE       MOUNT POINT
  1-109   free    outer edge
110-129   used    outer middle  datlv_data1         jfs        N/A
130-130   used    outer middle  loglv_data1         jfslog     N/A
```

```
131-150   used    outer middle  datlv_data2         jfs        N/A
151-151   used    outer middle  loglv_data2         jfslog     N/A
152-152   stale   outer middle  loglv00             jfs2log    N/A
153-154   stale   outer middle  fslv00              jfs2       N/A
155-217   free    outer middle
218-325   free    center
326-433   free    inner middle
434-542   free    inner edge
```

10. Stop the cluster services on all nodes when all the logical volumes are synchronized on the rpvclient.
11. Change `/etc/filesystems` to use the logical volumes instead of GMDS on all nodes in the cluster (Example 8-52).

Example 8-52 The /etc/filesystem using gmds and logical volumes

```
The before entry of one of the filesystems using gmds in /etc/filesystems
/geo_data2:
        dev             = /dev/gmd20
        vfs             = jfs
        log             = /dev/gmd21
        mount           = false
        options         = rw
        account         = false

What the entry looks like using the logical volumes
/geo_data2:
        dev             = /dev/datlv_data2
        vfs             = jfs
        log             = /dev/loglv_data2
        mount           = false
        options         = rw
        account         = false
```

12. Remove the GMDS from the resource groups and turn on the force flag for volume groups. Run the **smitty hacmp** command and select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resource Group Configuration → Change/Show Resources and Attributes for a Resource Group**. Blank out the Geomirror Devices and turn on the forced varyon of volume group (Example 8-53).

Example 8-53 SMIT Change/Show ALL Resources and Attributes panel

```
Use forced varyon of volume groups, if necessary    true
GeoMirror Devices                                   []
```

13. Verify and synchronize the cluster definition.
14. Create rpvservers on the local node and start the rpvservers. Create rpvservers to match the rpvservers that is created on the remote site.
15. Create rpvclients on the remote node and start the rpvclients. Create rpvclients to pair with the rpvservers on the local site. Example 8-54 shows the existing geovg volume group that is using GMDS.

Example 8-54 Disks, rpvserver, and rpvclient on the remote node

```
> lspv
hdisk0          none                                None
```

```
hdisk1          0000849208566aca                    rootvg          active
hdisk2          000105534fb8c100                    None
hdisk3          00011774996d81bb                    geovg           active
hdisk4          00011774996d9259                    None
hdisk5          00011774996da2f0                    None
hdisk6          00011774996db38c                    None
hdisk7          00011774996dd4bf                    None
hdisk8          00011774996de51f                    None
hdisk9          00011774996df55d                    None
hdisk10         00011774996d0c2f                    None

> lsrpvserver
  rpvserver0     00011774996d9259                hdisk4

> lsrpvclient
  hdisk10        00011774996d0c2f                siteA
```

16. On the remote node, varyoffvg the volume groups that are currently mirrored with GMDS.
17. Exportvg the GMD mirrored volume group (Example 8-55).

Example 8-55 geovg using gmd is exported

```
> varyoffvg geovg
> exportvg geovg
> lspv
hdisk0          none                                None
hdisk1          0000849208566aca                    rootvg          active
hdisk2          000105534fb8c100                    None
hdisk3          00011774996d81bb                    None
hdisk4          00011774996d9259                    None
hdisk5          00011774996da2f0                    None
hdisk6          00011774996db38c                    None
hdisk7          00011774996dd4bf                    None
hdisk8          00011774996de51f                    None
hdisk9          00011774996df55d                    None
hdisk10         00011774996d0c2f                    None
```

18. On all nodes, import the volume group by using the rpvserver and rpvclient disks (Example 8-56).

Example 8-56 The importvg and the vg varied on using the new disks

```
> importvg -y geovg hdisk4

> lspv
hdisk0          none                                None
hdisk1          0000849208566aca                    rootvg          active
hdisk2          000105534fb8c100                    None
hdisk3          00011774996d81bb                    None
hdisk4          00011774996d9259                    geovg           active
hdisk5          00011774996da2f0                    None
hdisk6          00011774996db38c                    None
hdisk7          00011774996dd4bf                    None
hdisk8          00011774996de51f                    None
```

```
hdisk9          00011774996df55d                    None
hdisk10         00011774996d0c2f                    geovg           active
```

19. Use the **varyoffvg** command to vary off the volume group.
20. Start the cluster.
21. The GMD definitions still exist and the disks on the remote site still have the old volume group information for GMDS. After the cluster is up and tested with GLVM, the GMDS can be removed. Use the **rmdev -l <gmd> -d** command or run the **smitty hageo** command and select **Configure GeoMirror Devices** → **Configure a GeoMirror Device** → **Remove a GeoMirror Device**. Remove the GMDS on all nodes in this fashion.
22. The disk on the remote node was already exported. It is now safe to use that disk for something else. In our example, hdisk3, which was the old geovg, can now be used for another purpose.
23. Remove the GMD statemap logical volumes that are in the new geovg, as they are no longer needed.
24. Migrate to AIX 6.1 as per IBM documentation.
25. Migrate to asynchronous GLVM as per IBM documentation.

8.12 Data divergence in PowerHA for GLVM

This section describes data divergence in PowerHA clusters with GLVM.

8.12.1 Quorum and forced varyon in geographically mirrored volume group

Allow a geographically mirrored volume group to be high availability, quorum disabled, and forced varyon. In rare cases, data integrity can be at risk when a copy of data in local disks is not synchronized with the copy on the remote disks. For example, the scenario can occur under the following conditions:

- All configured XD_data networks are down and the mirroring was stopped. Because quorum is disabled, the resource group is still online within the site.
- Because of a cluster condition (for example, a power outage at the primary site), the resource group fails over to another site. The volume group is forcefully varied on the remote site.
- After recovery from power outage, the node at the local site is started while XD_data networks are still down,

When the cluster services are started, PowerHA for GLVM determines whether any nodes at the local site have an XD_data network up and available and attempts to bring the resource group online. Although no nodes have the XD_data network available, PowerHA still attempts to acquire the resource group. If forced varyon is set to true, this setting might result in a situation in which the GMVG is forcefully varied on a node. The local disks will be available where data is not synchronized with the disks on the remote site.

To prevent data divergence, when the resource group is online on the node at the remote site:

- Do not start the node at the local site if all XD_data networks are down.
- Do not stop the cluster services on the higher priority node for the resource group at the local site if all XD_data are down.
- Do not stop and restart the cluster services on the higher priority node.

If one or more XD_data networks are available, PowerHA brings the resource group online and the geographically mirrored volume group is automatically synchronized from the copy on the remote site.

Note: If the forced varyon option is not set for the volume group and quorum is enabled, PowerHA for GLVM moves the resource group into an ERROR state. It does not attempt to activate it on a node where access to a single copy of the data exists. The cluster is protected from data divergence, and data integrity is secure. However, data availability is not always achievable automatically.

For more information, see the *HACMP for AIX 6.1 Geographic LVM planning and administration Guide*, SA23-1338.

8.12.2 Data divergence in asynchronous GMVGs

If the production site suddenly fails before the volume group can be varied offline, the physical volumes at the disaster recovery site are likely to be missing several of the updates that are still stored in the cache at the production site.

If this situation occurs, you must decide whether to allow data divergence to occur. You can wait until the production site is recovered, if that is possible, in which case you do not lose the data updates that are stored in the cache. Or you can activate the volume group at the disaster recovery site without the latest data updates, in which case you are likely to end up with data divergence, if the data at the production site has not been destroyed.

PowerHA for GLVM provides an attribute of a resource group. Allow varyon with missing data updates to allow two options:

- If it is defined as false, HACMP generates an event error when a normal **varyonvg** fails because it detected data divergence, and the resource group is in ERROR state, which means that manual intervention is required.
- If it is defined as true, HACMP event processing automatically allows data divergence to occur during a non-graceful site failover to the remote site. It allows you to run **varyonvg** with the **-d** flag to bring the volume group online after data divergence.

8.12.3 Recovering from data divergence for asynchronous GMVGs

When you are recovering your production site and data divergence has occurred, you can choose to allow data divergency recovery.

If you want to allow data divergence recovery to occur, you must indicate whether to preserve the data copy that resides at the local or the remote site. In this case, *local* refers to the local site, where you are running the recovery, and remote refers to the opposite site. The **varyonvg** command processing preserves the data at the site that you select and backs out the non-mirrored updates at the opposite site, effectively throwing them away, to merge the volume group back together.

Deciding which copy to preserve often depends on which site has more non-mirrored updates. If your applications ran for several hours at the disaster recovery site, you might want to keep the data that is at the disaster recovery site and lose the non-mirrored updates that reside in the cache at the production site. Conversely, if you did not make any meaningful changes to your data at the disaster recovery site (perhaps you mounted your file systems, for example), keep the data at the production site. In this case, the data divergence recovery processing backs out the data updates that were made during the mounting of the file

systems. The decision of which copy of the data to keep must be made based on the circumstances, and only you can make this decision.

PowerHA for GLVM provides an attribute of resource group. The default choice for data divergence recovery allows you to configure the site name of the site with the version of the data that you are most likely to choose to preserve when recovering from data divergence. Otherwise, this field can be left blank to indicate that you do not want to choose a default. For example, if Asite is the production site and Bsite is the remote site, you might want to configure Bsite as the default choice. If a production site failure occurs, and you want to keep the data at Bsite after the production site is recovered, you do not need to do anything to tell HACMP to carry out your decision (Figure 8-75).

```
          Change/Show All Resources and Attributes for a Resource Group

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Resource Group Name                                  YuriRG
  Inter-site Management Policy                         Prefer Primary Site
  Participating Nodes from Primary Site                GLVM_A1 GLVM_A2
  Participating Nodes from Secondary Site              GLVM_B1 GLVM_B2
Startup Policy                                        Online On Home Node Only
Failover Policy                                       Failover To Next Priority
Node In The List
Failback Policy                                       Failback To Higher Priority
Node In The List
Failback Timer Policy (empty is immediate)           [] +

  Service IP Labels/Addresses                         [] +
  Application Servers                                 [YuriApp] +
  Volume Groups                                       [yurivg ] +
  Use forced varyon of volume groups, if necessary     true +
  Automatically Import Volume Groups                   false +
  Allow varyon with missing data updates?              true +
      (Asynchronous GLVM Mirroring Only)
  Default choice for data divergence recovery          Bsite +
      (Asynchronous GLVM Mirroring Only)
```

Figure 8-75 Configure attribute of YuriRG for data divergency

The secondary instance of the resource group must be online before the primary instance is brought online, if the secondary instance is the one that contains the data to be preserved.

8.12.4 Overriding the default data divergence recovery

If you need to override a resource group's default data divergence recovery attribute, start cluster services on the production site without allowing automatic resource group management. Then, you must manage the resource group manually, and then, you can specify a new data divergence recovery value that overrides the default value.

First, start the cluster without allowing the joining nodes to acquire any resources. Select **Manually** for the Manage Resource Groups field (Figure 8-76).

```
                          Start Cluster Services

Type or select values in entry fields.
Press Enter AFTER making all desired changes.[Entry Fields]
* Start now, on system restart or both                now                    +
  Start Cluster Services on these nodes              [GLVM_A1]              +
* Manage Resource Groups                              Manually               +
  BROADCAST message at startup?                       true                   +
  Startup Cluster Information Daemon?                 false                  +
  Ignore verification errors?                         false                  +
  Automatically correct errors found during           Interactively          +
```

Figure 8-76 Overriding default data divergence recovery (Part 1 of 2)

Next, after the production site nodes join the cluster, you must manage the resource groups manually. Most likely the primary instance of the resource group is already running at the disaster recovery site. You have a few choices:

- Move the primary instance of the resource group back to the production site. This step performs what is known as a *site failback* to return the cluster to how it was before the site failure, with the production site asynchronously mirroring to the disaster recovery site. PowerHA automatically takes the primary resource group instance offline and then brings the secondary instance online at the disaster recovery site before bringing the primary instance online at the production site.
- Keep the primary instance at the disaster recovery site and bring the secondary instance of the resource group online at the production site. Then, asynchronous data mirroring occurs in the opposite direction, from the disaster recovery site back to the production site.
- If you want to switch back to the production site's version of the data, while continuing to run at the disaster recovery site, take the resource group offline at the disaster recovery site. Then, you can start the secondary instance at the production site, followed by the primary instance at the disaster recovery site (Figure 8-77).

```
                        Bring a Resource Group Online

Type or select values in entry fields.
Press Enter AFTER making all desired changes. [Entry Fields]
  Resource Group to Bring Online                      YunaRG
  Node on Which to Bring Resource Group Online        GLVM_A1
  Name of site containing data to be preserved        [Asite] +
  by data divergence recovery processing
  (Asynchronous GLVM Mirroring only)
```

Figure 8-77 Overriding default data divergence recovery (Part 2 of 2)

You can either complete the name of the site that contains data to be preserved by data divergence recovery processing field with a site name or leave it blank. If you leave the field blank, this operation fails.

Part 4

Maintenance, management, and disaster recovery

This part provides feedback about the management and maintenance of the cluster nodes. This part includes the following chapters:

- Chapter 9, “Maintenance and management” on page 435.
- Chapter 10, “Disaster recovery with DS8700 Global Mirror” on page 469
- Chapter 11, “Disaster recovery by using Hitachi TrueCopy and Universal Replicator” on page 515

9

Maintenance and management

Previous chapters review the requirements and installation steps for configuring the replication options available in the PowerHA Enterprise Edition. This chapter provides feedback about the management and maintenance of the cluster nodes. It also reviews considerations specific to partitioned clusters and data divergence.

This chapter includes the following sections:

- Maintaining the servers
- Resource group management
- Partitioned cluster considerations

9.1 Maintaining the servers

Maintaining consistency between the AIX images on the nodes within the cluster an important consideration whenever a clustering solution is implemented. The node count increases when you implement a multisite environment and the same consistency rules apply as in a single-site cluster. Ultimately, all of the cluster members must run the same levels of AIX and PowerHA software.

An advantage of a clustered environment is that administrators can use the inherent cluster functions to orchestrate the movement of resources between the servers. If a set of software fixes requires a restart, a cluster move operation gracefully releases and reacquires the resources automatically. Furthermore, enhancements to the cluster software now enable the ability to upgrade the file sets in a non-disruptive fashion. These benefits are described in more detail in the following sections.

9.1.1 AIX maintenance

Maintaining the operating system images is a constant task for AIX system administrators. Many variables affect the levels of file sets to be installed on the servers at any time, such as qualified levels, application requirements, and license considerations. However, one clear advantage in a clustered environment is that for planned maintenance the cluster software facilitates the movement of the resources between the servers.

Because the cluster controls the start and stop of the applications and handles the steps that are required to set up the replication between the sites, it provides a reliable mechanism to implement planned updates.

The cluster orchestrates the release and reacquisition of the resources. For example, if we consider a scenario where a cluster has two nodes at the primary site and two at the remote site, we can begin our AIX updates on the nodes not hosting any resources. First, we must stop cluster services, apply the updates, and reboot the server. As soon as the server is back online, we restart cluster services and reintegrate the server into the cluster. We continue that same approach for the remaining nodes in the cluster until all nodes are at the target level.

A common question is: How long can a cluster run in mixed mode? That is how long can it run with one server that runs one version of AIX and the remaining nodes that run with an older release. The answer is that the cluster is designed to handle intercluster communications between nodes that run at different levels. If the release of the PowerHA software is supported on the target AIX level, the updates can be performed on one node at time. The amount of time that the AIX levels can be mixed between the cluster nodes is indefinite. However, as explained in other publications that address migrations, while the cluster nodes are in a mixed mode, avoid any changes to the cluster configurations or any DARE operations.

9.1.2 PowerHA cluster maintenance

The PowerHA file sets are installed independently from the AIX base images. In the PowerHA 5.4.0.1 release, the ability to perform cluster updates in a non-disruptive way was introduced. This ability is referred to as the non-disruptive upgrade option (NDU) and it is applicable to patches or when moving to a new release. The only deviation is if a new release introduced new a file set with new functions that required a reboot.

Nondisruptive option: Use the nondisruptive option only when you upgrade PowerHA (not AIX or RSCT).

To perform a non-disruptive update, you must stop cluster services on the nodes with the UNMANAGE option. This option leaves the application and resources in the resource group online. This option is available from the cluster **clstop** panels and replaces the older Forced Stop option in older versions.

To stop cluster services without stopping your applications:

1. Enter the fastpath **smitty cl_admin** or **smitty hacmp**.
2. Select **System Management (C-SPOC)** and press Enter.
3. Select **Manage HACMP Services** → **Stop Cluster Services** and press Enter.
4. Choose **Unmanage resource groups**.

To avoid the loss of an accurate cluster status when performing updates, stop the cluster by using the UNMANAGE option on one cluster node at a time. Figure 9-1 shows a sample of the unmanage option.

```
                              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                [svcxd_a1]              +
  BROADCAST cluster shutdown?                          true                    +
* Select an Action on Resource Groups                  Bring Resource Groups> +

  +--------------------------------------------------------------------------+
  |                    Select an Action on Resource Groups                   |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  |   Bring Resource Groups Offline                                          |
  |   Move Resource Groups                                                   |
  |   Unmanage Resource Groups                                               |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
F1| F8=Image                F10=Exit                Enter=Do                 |
F5| /=Find                  n=Find Next                                      |
F9+--------------------------------------------------------------------------+
```

Figure 9-1 RG UNMANAGE option for updating the cluster

Regardless of the type of resource group you have, if you stop cluster services on the node on which this group is active and do not stop the application that belongs to the resource group, PowerHA puts the group into an UNMANAGED state and keeps the application running according to your request.

The resource group that contains the application remains in the UNMANAGED state (until you instruct PowerHA to start managing it again) and the application continues to run. While in this condition, PowerHA and the RSCT services continue to run, providing services to ECM VGs that the application servers might be using.

You can instruct PowerHA to start managing the resources again either by restarting cluster services on the node or by using SMIT to move the resource group to a node that is actively managing its resource groups.

If you have instances of replicated resource groups that use the extended distance capabilities of the PowerHA Enterprise Edition, the UNMANAGED SECONDARY state is used for resource groups that were previously in the ONLINE SECONDARY state.

Attention: When you stop cluster services on a node and place resource groups in an UNMANAGED state, the cluster stops managing the resources on that node. PowerHA does not react to individual resource failures, application failures, or even if the node crashes.

Similar to how AIX operating system levels must be the same across the cluster nodes, the PowerHA levels must also be the same. Running in a mixed mode is acceptable for short periods of time while performing cluster updates.

9.1.3 Displaying the cluster configuration

Understanding how the cluster is configured is critical to maintain the environment effectively. Typically, if an administrator builds the environment, the administrator already has an understanding of how the environment is laid out. However, for a new administrator who logs on to a cluster node for the first time, review the output of the commands that are listed in Table 9-1. These commands are general for gathering basic information about the cluster configuration.

Table 9-1 PowerHA commands to display cluster configuration

Command	Description
`clRGinfo`	Displays the status and location of the RGs
`cltopinfo -m`	Displays the cluster uptime and any missed heartbeats for any interfaces
`cldump`	Displays the current cluster configuration
`clshowres`	Details the content of the RGs defined
`cllsnw`	Displays the networks that are defined in the cluster
`clshowsrv -v`	Status of the cluster daemons
`clstat`	Cluster status
`lssrc -ls clstrmgrES`	Status of cluster manager and dynamic node priority calculations
`lssrc -ls topsvcs`	Shows RSCT heartbeat rings and any missed heartbeat counts

The information that is displayed by using the general commands does not show information specific to the replicated resources. The naming schemes that are used in the environment can help determine the type of replication that is being used. However, reviewing the PowerHA Enterprise Edition packages that are installed and running commands specific to each replication type displays the information specific to the replicated resources.

The following tables identify several of the new commands that get added to the cluster when the replication packages for each solution are installed. Several of these commands are used by the cluster processing and might not be supported as stand-alone commands, but they are useful to quickly identify how the replicated resources are configured.

Table 9-2 displays several new commands that are appended by the SAN Volume Controller cluster file sets. For detailed results, see Chapter 5, “Configuring PowerHA SystemMirror Enterprise Edition with Metro Mirror and Global Mirror” on page 153.

Table 9-2 SVC replication cluster reference commands

Command	Description
/usr/es/sbin/cluster/svcpprc/cmds/**cllssvc**	Lists information about all SVC clusters in the HACMP configuration or a specific SVC cluster. If no SAN Volume Controller is specified, all SVC clusters that are defined are listed. If a specific SVC cluster is provided by using the **-n** flag, information about this SVC only is displayed. The **-c** flag displays information in a colon-delimited format.
/usr/es/sbin/cluster/svcpprc/cmds/**cllssvcpprc**	Lists information about all SVC PPRC resources or a specific SVC PPRC resource. If no resource name is specified, the names of all PPRC resources that are defined are listed. If the -a flag is provided, full information about all PPRC resources is displayed. If a specific resource is provided by using the **-n** flag, information about this resource only is displayed. The **-c** flag displays information in a colon-delimited format. The **-h** flag turns off the display of column headers.
/usr/es/sbin/cluster/svcpprc/cmds/**cllsrelationship**	Lists information about all SVC PPRC relationships or a specific PPRC relationship. If no resource name is specified, the names of all PPRC resources that are defined are listed. If the -a flag is provided, full information about all PPRC relationships is displayed. If a specific relationship is provided by using the **-n** flag, information about this relationship only is displayed. The **-c** flag displays information in a colon-delimited format. The **-h** flag turns off the display of column headers.

Table 9-3 lists several new commands that are available to display the devices that are used in the Metro Mirror replication. For detailed results, see Chapter 6, “Configuring PowerHA SystemMirror Enterprise Edition with ESS/DS Metro Mirror” on page 237.

Table 9-3 DS replication cluster reference commands

Command	Description
/usr/es/sbin/cluster/pprc/spprc/cmds/**cllscss**	Lists information about all consistency group relationships. If no resource name is specified, the names of all PPRC resources that are defined are listed. If the -a flag is provided, full information about all PPRC relationships is displayed. If a specific relationship is provided by using the **-n** flag, information about this relationship only is displayed.
/usr/es/sbin/cluster/pprc/spprc/cmds/**cllsspprc**	Lists information about all SPPRC relationships or a specific PPRC relationship. If no resource name is specified, the names of all PPRC resources that are defined are listed. If the -a flag is provided, full information about all PPRC relationships is displayed. If a specific relationship is provided by using the **-n** flag, information about this relationship only is displayed.
/usr/es/sbin/cluster/pprc/spprc/cmds/**cllsdss**	Lists information about all DS PPRC relationships or a specific PPRC relationship. If no resource name is specified, the names of all PPRC resources that are defined are listed. If the -a flag is provided, full information about all PPRC relationships is displayed. If a specific relationship is provided by using the **-n** flag, information about this relationship only is displayed.

Table 9-4 shows the additional EMC cluster list command that displays the configuration for the replicated resources. For detailed results, see Chapter 7, “Configuring PowerHA SystemMirror Enterprise Edition with SRDF replication” on page 267.

Table 9-4 EMC SRDF replication cluster reference commands

Command	Description
/usr/es/sbin/cluster/sr/cmds/**cllssr**	Displays the replicated resource configuration. If the -a flag is provided, full information about all SRDF relationships is displayed. If a specific relationship is provided by using the **-n** flag, information about this relationship only is displayed.

Table 9-5 list GLVM-specific commands that display the replicated resources that are part of the cluster configuration. For examples, see Chapter 8, “Configuring PowerHA SystemMirror Enterprise Edition with Geographic Logical Volume Manager” on page 339.

Table 9-5 GLVM replication cluster reference commands

Command	Description
/usr/sbin/**rpvstat**	Provides the status of one or more Local Remote Physical Volume (RPV) clients.
/usr/sbin/**gmvgstat**	Provides the status of one or more Locally active GMVGs.

9.1.4 Cluster reporting

For reporting purposes, consider using built-in functions, such as the files generated when you create a cluster snapshot (`.info` and `.odm` files) or the cluster report function available in the Online Planning Worksheets (OLPW) or in the WebSMIT GUI interface.

Figure 9-2 shows partial output of a `.info` file for one of our clusters. The file contains a collection of commands that show the cluster configuration and resources that are defined in each cluster node.

```
#cat /usr/es/sbin/cluster/snapshots/Prod_SVC_Cluster_03_22_2010.info

<VER
clsnapshot_version=1.2.2.111
cluster_name=svcxd
cluster_id=1163553413
cluster_version=11
cluster_release=6.1
snapshot_instance=25
</VER
.....
Cluster Name: svcxd
Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No
There were 4 networks defined: net_XD_ip_01, net_diskhb_01, net_diskhb_02,
net_e
ther_01
There are 4 nodes in this cluster

NODE svcxd_a1:
        This node has 5 service IP label(s):

        Service IP Label svcxd_a1_hdisk9_01:
                IP address:     /dev/hdisk9
                Hardware Address:
                Network:        net_diskhb_01
                Attribute:      serial
.......
```

Figure 9-2 Sample cluster .info report file

The <snapshot>.odm file that also gets generated is what the cluster uses to restore a configuration or to update during a snapshot migration. It provides a collection of cluster stanzas that contain all of the cluster definitions. Figure 9-3 shows partial output from a sample from one of our test clusters.

```
# cat /usr/es/sbin/cluster/snapshots/Prod_SVC_Cluster_03_22_2010.odm
.....
HACMPsvc:
        svccluster_name = "B8_8G4"
        svccluster_role = "Master"
        sitename = "svc_sitea"
        cluster_ip = "10.12.5.55"
        cluster_2nd_ip = ""
        r_partner = "B12_4F2"
        version = ""
        reserved = ""

HACMPsvc:
        svccluster_name = "B12_4F2"
        svccluster_role = "Master"
        sitename = "svc_siteb"
        cluster_ip = "10.114.63.250"
        cluster_2nd_ip = ""
        r_partner = "B8_8G4"
        version = ""
        reserved = ""

HACMPsvcpprc:
        svcpprc_consistencygrp = "svc_metro"
        MasterCluster = "B8_8G4"
        AuxiliaryCluster = "B12_4F2"
        relationships = "svc_disk2 svc_disk3 svc_disk4 svc_disk5"
        CopyType = "METRO"
        RecoveryAction = "MANUAL"

HACMPsvcpprc:
        svcpprc_consistencygrp = "svc_global"
        svccluster_name = "B12_4F2"
        svccluster_role = "Master"
        sitename = "svc_siteb"
        cluster_ip = "10.114.63.250"
        cluster_2nd_ip = ""
        r_partner = "B8_8G4"
        version = ""
        reserved = ""
.....
```

Figure 9-3 Sample cluster .odm configuration file

The cluster report file that is generated by importing a cluster definition file into the Online Planning Worksheets on your desktop can also provide a useful cluster reference. This same report is automatically generated when you use the WebSMIT PowerHA GUI management console. Figure 9-4 shows a sample html report file from one of our four node test clusters.

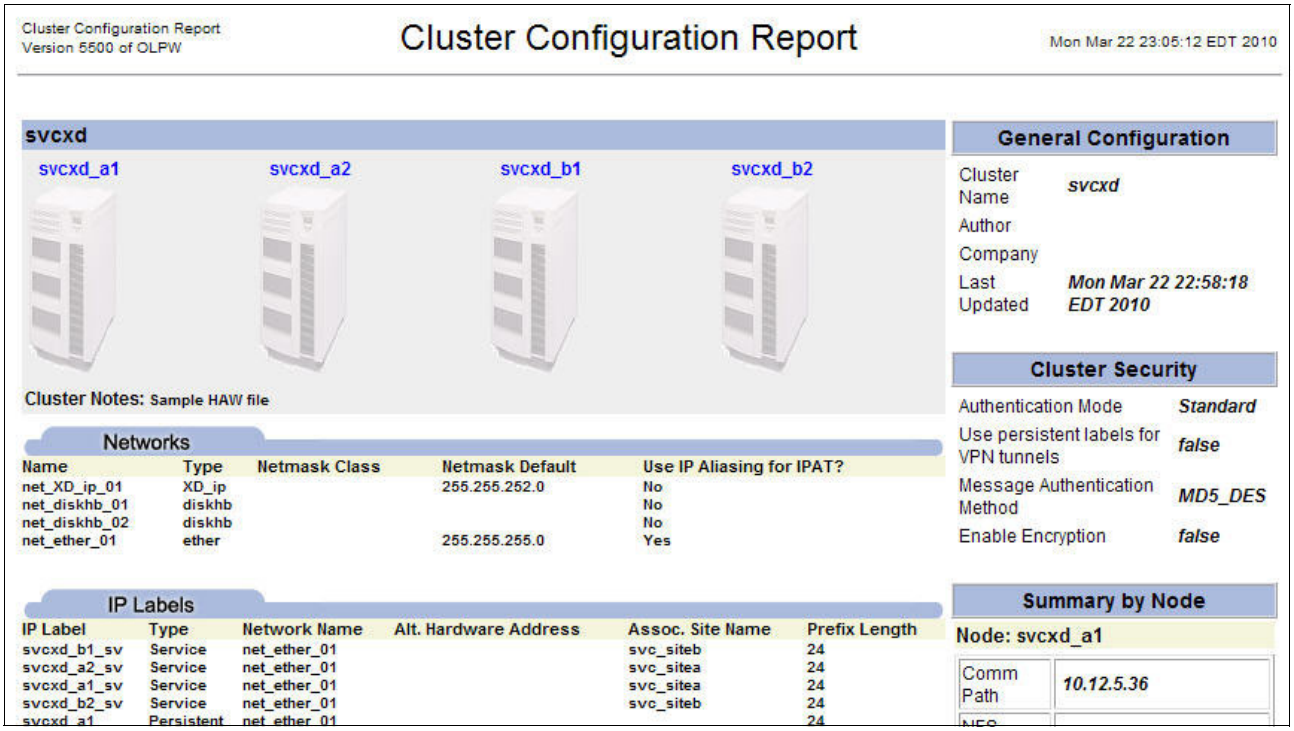


Figure 9-4 Sample cluster html report file

9.1.5 PowerHA Enterprise Edition and the cluster test tool

By using the cluster test tool utility, you can test a PowerHA cluster configuration and evaluate how a cluster operates under a set of specific circumstances. For example, such circumstances include when cluster services on a node fail or when a node loses connectivity to a cluster network. For a PowerHA Enterprise Edition environment, you might consider using the cluster test tool for the site-specific tests, such as graceful shutdown of a site with takeover.

By using the cluster test tool, you can test a PowerHA cluster in two ways:

- Automated testing (also known as *Automated Test Tool*): In this mode, the cluster test tool runs a series of predefined sets of tests on the cluster.
- Custom testing (also known as *Test Plan*): In this mode, you can create your own test plan or a custom testing routine that includes different tests available in the cluster test tool library.

Considerations about the cluster test tool for site-specific testing

The cluster test tool has certain considerations in a PowerHA Enterprise Edition environment that apply to both Metro Mirror and GLVM environments:

- Sites: You can perform general cluster testing for clusters that support sites, but not testing specific to any of the PowerHA Enterprise Edition integrated replication methods.

- Replicated resources: You can perform general cluster testing for clusters that include replicated resources, but not testing specific to replicated resources or any of the PowerHA Enterprise Edition products such as testing on storage replication relationships.

Site-specific tests

If sites are present in the cluster, the test tool runs the tests for them. The automated testing sequence that the cluster test tool uses contains two site-specific tests:

- auto_site

  This sequence of tests runs if you have any cluster configuration with sites. For example, this sequence is used for clusters with cross-site LVM mirroring configured that does not use XD_data networks. This sequence includes the following tests:

  SITE_DOWN_GRACEFUL
  Stops the cluster services on all nodes in a site while taking resources offline

  SITE_UP Restarts the cluster services on the nodes in a site

  SITE_DOWN_TAKEOVER
  Stops the cluster services on all nodes in a site and moves the resources to nodes at another site

  SITE_UP Restarts the cluster services on the nodes at a site

  RG_MOVE_SITE Moves a resource group to a node at another site

- auto_site_isolation

  This sequence of tests runs only if you configured sites and a XD-type network. This sequence includes the following tests:

  SITE_ISOLATION Isolates sites by failing XD_data networks

  SITE_MERGE Merges sites by bringing up XD_data networks

The support for the cluster test tool is included in the base PowerHA product. The following tests are available for the cluster test tool specific to PowerHA Enterprise Edition environments:

RG_MOVE_SITE Moves a resource group that is already online to an available node at a specific site.

SITE_ISOLATION Brings down all XD_data networks in the cluster at which the tool is running, causing a site isolation.

SITE_MERGE Brings up all XD_data networks in the cluster at which the tool is running, simulating a site merge.

Run the SITE_MERGE test after you run the SITE_ISOLATION test.

SITE_UP Starts cluster services on all nodes at the specified site that are currently stopped.

SITE_DOWN_TAKEOVER
Stops cluster services on all nodes at the specified site and moves the resources to nodes at another site by starting automatic rg_move events.

SITE_DOWN_GRACEFUL
Stops cluster services on all nodes at the specified site and takes the resources offline.

For more information about the cluster test tool, see the *HACMP for AIX 6.1 Administration Guide*, SC23-4862, which you can find in the Cluster Products Information Center at:

http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.hacmp.doc/hacmpbooks.html

9.1.6 Reporting a problem

If a problem occurs, collecting the appropriate information in a timely fashion is critical when performing problem determination. The 5.4.1 release of the PowerHA software introduced the first failure data capture (FFDC) functions, which ensures that you do not lose critical diagnostic data. To prevent the loss of the logs, the cluster automatically captures the local node's cluster snap data after you recover from a software or node failure. Only the most recent snap data for FFDC is retained. The FFDC data is saved in the `/tmp/ibmsupt/hacmp/ffdc.<DateTimeStamp>` directory.

To collect the logs and the cluster configuration, we use the *snap -e* option. This option is one of the primary sets of files that are collected by IBM AIX support when performing a detailed problem analysis. The **-e** option is specific to PowerHA and captures the system configuration and all of the cluster and RSCT logs from all clusters.

Any previous snaps can be cleared by using the -r flag. It might be worthwhile to save a copy of the compressed file in `/tmp/ibmsupt/< >` even after problem determination is performed.

For more snap usage options, run the **snap -?** command on the host.

9.2 Resource group management

A resource group (RG) is a container that is used by PowerHA to group a number of highly available resources. During normal operations, the resources are considered ONLINE or OFFLINE from the client's perspective. However, in the cluster processing, the resource groups have various states. This section identifies these states and the RG options specific to clusters configured for replication with the PowerHA Enterprise Edition.

9.2.1 Checking the status of the RGs with the clRGinfo utility

Identifying the state of the resource groups in the cluster typically provides a good indication of the state of the cluster resources. The status, however, does not indicate whether the application is truly running, but rather it tells only where it is hosted.

The **clRGinfo** command is shows you the state of the resource groups. You can use the **/usr/es/sbin/cluster/utilities/clRGinfo** command to monitor the resource group status and its location. This command also reports if a node temporarily has the highest priority for this instance.

Alternative: You can use the **clfindres** command instead of **clRGinfo**. The **clfindres** command is a link to clRGinfo. Only the root user can run the **clRGinfo** utility.

There are various flags available for the **clRGinfo** command (Example 9-1).

Example 9-1 clRGinfo command options

```
# clRGinfo -?
clRGinfo: illegal option -- ?
Usage: clRGinfo [-h] [-v] [-a] [-s|-c] [-t] [-p] [group name(s)]
```

For more information about each option, see the *HACMP for AIX 6.1 Administration Guide*, SC23-4862, in each corresponding release.

The sample output in Example 9-2 shows the status of the resources in the cluster that is used for the SAN Volume Controller Metro and Global Mirror test scenarios. An important observation from the sample output is how the use of the **-p** option provides details about the node with the highest priority and a timestamp from the most recent operation.

Example 9-2 clRGinfo -p sample output

```
# clRGinfo -p

Cluster Name: svcxd

Resource Group Name: RG_sitea
Secondary instance(s):
The following node temporarily has the highest priority for this instance:
svcxd_b2, user-requested rg_move performed on Fri Mar 19 19:47:57 2010

Node                          Primary State    Secondary State
----------------------------- ---------------
svcxd_a1@svc_sitea            ONLINE           OFFLINE
svcxd_a2@svc_sitea            OFFLINE          OFFLINE
svcxd_b2@svc_siteb            OFFLINE          ONLINE SECONDARY
svcxd_b1@svc_siteb            OFFLINE          OFFLINE

Resource Group Name: RG_siteb
Secondary instance(s):
The following node temporarily has the highest priority for this instancei:
svcxd_a2, user-requested rg_move performed on Fri Mar 19 19:48:47 2010

Node                          Primary State    Secondary State
----------------------------- ---------------
svcxd_b1@svc_siteb            ONLINE           OFFLINE
svcxd_b2@svc_siteb            OFFLINE          OFFLINE
svcxd_a2@svc_sitea            OFFLINE          ONLINE SECONDARY
svcxd_a1@svc_sitea            OFFLINE          OFFLINE
```

The output in Example 9-2 also shows the site-specific states for the two resource groups in the cluster. The secondary states are designed to attempt to identify where the resources are moved if a failover occurs across the sites.

Site failure: The cluster tries to use the location that is identified in the ONLINE SECONDARY output, but if a site failure occurs, the resources might not always end up on the expected target.

Table 9-6 lists the site-specific states.

Table 9-6 clRGinfo RG site-specific states

On the primary site	On the secondary site
ONLINE	ONLINE SECONDARY
OFFLINE	OFFLINE SECONDARY
ERROR	ERROR SECONDARY
UNMANAGED	UNMANAGED SECONDARY

Inter-site management policy

The option to define an inter-site management policy is available during the creation of any new resource group. There are associated implications for the different option available for this setting.

Figure 9-5 shows the sample output from the RG definition portion of the configuration through SMIT. Whenever sites are specified, determine the most appropriate inter-site management policy for your 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                                 []

  Inter-Site Management Policy                        [ignore]               +
* Participating Nodes from Primary Site               []                     +
  Participating Nodes from Secondary Site             []                     +

  Startup Policy                                       Online On Home Node O> +
  Failover Policy                                      Failover To Next Prio> +
  Failback Policy                                      Failback To Higher Pr> +

  +--------------------------------------------------------------------------+
  |                        Inter-Site Management Policy                      |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  |   ignore                                                                 |
  |   Prefer Primary Site                                                    |
  |   Online On Either Site                                                  |
  |   Online On Both Sites                                                   |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
F1| F8=Image                F10=Exit                Enter=Do                 |
F5| /=Find                  n=Find Next                                      |
F9+--------------------------------------------------------------------------+
```

Figure 9-5 RG site-specific options

The default option is set to Ignore, but the following options are available:

- Ignore: If you select this option, the resource group does not have ONLINE SECONDARY instances. Use this option if you use cross-site LVM mirroring. You can also use it with HACMP/XD for Metro Mirror.
- 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: 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 or failback.

The resource group moves to another site only if no node or condition exists under which it can be brought or kept ONLINE on the site where it is located. The site that owns the active resource group is called the primary site.

9.2.2 RG states

A resource group can be many states. Table 9-7 lists the possible states of a resource group. The RG state numbers are maintained internally and are not displayed by the **clRGinfo** command.

Table 9-7 PowerHA - resource group states

RG state number	Description
1	RG Invalid
2	RGOnline State
4	RG Offline State
8	RG Unknown State
16	RG Acquiring
32	RG Releasing
64	RG Error State
128	RG Temporary Error State
256	RG Online Secondary
512	RG Offline Secondary
1024	RG Acquiring Secondary
2048	RG should be Secondary online
4096	RG Releasing Secondary
8192	RG Releasing Peer State
16384	RG Error Secondary State
32768	RG Temp Error Secondary State

RG state number	Description
65536	RG Offline because of a failover
131072	RG Offline because of a parent offline
262144	RG Offline because of a lack of parent

Example 9-2 on page 446 shows output of the **clRGinfo** command.

9.2.3 RG site-specific dependencies

Resource group dependencies allow administrators to configure relationships between RGs that alter the default parallel processing and revert back to a sequential mode. This section reviews these options, but they are described in more detail in the base PowerHA documentation guides.

You can configure four types of dependencies between resource groups:

- Parent/child dependency
- Online on same node location dependency
- Online on different nodes location dependency
- Online on same site location dependency

Considerations for dependencies between resource groups

To obtain more granular control over the resource group movements, use the **clRGinfo -a** command to view what resource groups are going to be moved during the current cluster event. Also, use the output in the `hacmp.out` file. For more information, see the *HACMP for AIX 6.1 Administration Guide*, SC23-4862.

Dependencies between resource groups offer a predictable and reliable way of building clusters with multi-tiered applications. However, node_up processing in clusters with dependencies can take more time than in the clusters where the processing of resource groups on node_up is done in parallel. A resource group that depends on other resource groups cannot be started until other resource groups are started first. The config_too_long warning timer for node_up must be adjusted to be large enough to allow for this.

During verification, PowerHA verifies that your configuration is valid and that application monitoring is configured.

You can configure resource group dependencies in PowerHA Enterprise Edition clusters that use replicated resources for disaster recovery. However, you cannot have a combination of any non-concurrent startup policy and concurrent (online on both sites) inter-site management policy. You can have a concurrent startup policy that is combined with a non-concurrent inter-site management policy.

Configuring online on the same site dependency for RGs

The only site-specific dependency in the Enterprise Edition is the *online on same site*. When you configure two or more resource groups and establish a location dependency between them, they belong to a set for that particular dependency. This section explains the *online on same site* dependency set of resource groups.

The following rules and restrictions are applicable to the online on same site dependency set of resource groups:

- All resource groups in a same site dependency set must have the same intersite management policy, but might have different startup, failover, and failback policies. If failback timers are used, they must be identical for all resource groups in the set.
- All resource groups in the same site dependency set must be configured so that the nodes that can own the resource groups are assigned to the same primary and secondary sites.
- *Online using node distribution policy startup policy* is supported.
- Both concurrent and non-concurrent resource groups are allowed.
- You can have more than one same site dependency set in the cluster.
- All resource groups in the same site dependency set that are active (ONLINE) are required to be ONLINE on the same site, even though certain resource groups in the set may be OFFLINE or in an ERROR state.
- If you add a resource group included in a same node dependency set to a same site dependency set, you must add all the other resource groups in the same node dependency set to the same site dependency set.

Online on same site test scenario

To validate this option, we configured two new resource groups in one of our existing 4-node clusters. Each resource group was defined to be on the same site (Example 9-3).

Example 9-3 Resource groups that are created for the online on same site scenario

```
# clRGinfo -v
Cluster Name: xdemc

Resource Group Name: mikeRG1
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Failback Policy: Never Failback
Site Policy: Prefer Primary Site
Node                          Primary State    Secondary State
----------------------------- ---------------
xdemca1@siteA                     ONLINE           OFFLINE
xdemca2@siteA                     OFFLINE          OFFLINE
xdemcb1@siteB                     OFFLINE          ONLINE SECONDARY
xdemcb2@siteB                     OFFLINE          OFFLINE

Resource Group Name: mikeRG2
Startup Policy: Online On Home Node Only
Failover Policy: Failover To Next Priority Node In The List
Failback Policy: Never Failback
Site Policy: Prefer Primary Site
Node                          Primary State    Secondary State
----------------------------- ---------------
xdemca1@siteA                     ONLINE           OFFLINE
xdemca2@siteA                     OFFLINE          OFFLINE
xdemcb1@siteB                     OFFLINE          ONLINE SECONDARY
xdemcb2@siteB                     OFFLINE          OFFLINE
```

Next, we defined a site dependency between the two resource groups through the PowerHA panels. Figure 9-6 shows the PowerHA panel to define the dependency. Run the **smitty hacmp** command. Select **Extended Configuration** → **Extended Resource Configuration** → **Configure Resource Group Run-Time Policies** → **Configure Dependencies between Resource Groups** → **Configure Online on the Same Site Dependency** → **Add Online on the Same Site Dependency Between Resource Groups**. Then, press Enter.

```
         Add Online on the Same Site Dependency Between Resource Groups

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

                                                      [Entry Fields]
* Resource Groups to be Online on the same site      []                      +



  +--------------------------------------------------------------------------+
  |              Resource Groups to be Online on the same site               |
  |                                                                          |
  | Move cursor to desired item and press F7.                                |
  |     ONE OR MORE items can be selected.                                   |
  | Press Enter AFTER making all selections.                                 |
  |                                                                          |
  |   mikeRG1                                                                |
  |   mikeRG2                                                                |
  |                                                                          |
  | F1=Help                F2=Refresh             F3=Cancel                  |
F1| F7=Select              F8=Image               F10=Exit                   |
F5| Enter=Do               /=Find                 n=Find Next                |
F9+--------------------------------------------------------------------------+
```

Figure 9-6 Defining online on same site dependency SMIT panel

After you define the dependency, you push the update across all nodes by running a cluster synchronization. You confirm the dependency by running the following command:

```
# odmget HACMPrg_loc_dependency

HACMPrg_loc_dependency:
        id = 1
        set_id = 1
        group_name = "mikeRG1"
        priority = 0
        loc_dep_type = "SITECOLLOCATION"
        loc_dep_sub_type = "STRICT"

HACMPrg_loc_dependency:
        id = 2
        set_id = 1
        group_name = "mikeRG2"
        priority = 0
        loc_dep_type = "SITECOLLOCATION"
        loc_dep_sub_type = "STRICT"
```

With the relationship established, the cluster automatically groups them and requires you to perform administrative tasks against all RGs in the dependency set. Figure 9-7 shows the sample output when we tried to move only one of the resource groups to the remote site. With the dependency defined, the cluster provides only an option to move the set.

```
                Move a Resource Group to Another Node / Site

Mo+--------------------------------------------------------------------------+
  |                         Select a Resource Group                          |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  | [MORE...5]                                                               |
  |   RG_siteb                          ONLINE               xdemcb1 / siteB |
  |   RG_siteb                          ONLINE SECONDARY     xdemca1 / siteA |
  |   mikeRG1                           ONLINE SECONDARY     xdemcb1 / siteB |
  |   mikeRG2                           ONLINE SECONDARY     xdemcb1 / siteB |
  |                                                                          |
  |   #                                                                      |
  |   # Resource groups in node or site collocation configuration:           |
  |   # Resource Group(s)                           State    Node / Site     |
  |   #                                                                      |
  |   mikeRG1,mikeRG2                               ONLINE   xdemca1 / siteA |
```

Figure 9-7 rg_move of RGs using an online on same site dependency

Similarly, when we tried to remove only one of the resource groups, we received a message that indicates that the dependency needed to be removed first before the resource group can be deleted (Figure 9-8).

```
clrmgrp: ERROR: The resource group mikeRG2 is configured in the resource group
dependency configuration. Please delete the group from the depend ency
configuration prior to removing the group.
```

Figure 9-8 Message when trying to move one of the resource groups

9.2.4 Customizing inter-site RG recovery

When you install PowerHA and configure a new cluster with sites, the selective failover of resources included in the replicated resource groups is enabled by default. If necessary for recovery, PowerHA moves the resource group that contains the resources to the other site.

If a local network is down for the network that hosts a service IP address or if one of the volume groups in the RG has a quorum loss, the resource group is automatically relocated to the remote site.

If selective failover across sites is enabled, PowerHA tries to recover both the primary and the secondary instance of a resource group:

- If an acquisition failure occurs while the secondary instance of a resource group is being acquired, the cluster manager tries to recover the resource group's secondary instance, as it does for the primary instance. If no nodes are available for the acquisition, the resource group's secondary instance goes into global ERROR_SECONDARY state.

- If quorum loss is triggered and the resource group has its secondary instance online on the affected node, PowerHA tries to recover the secondary instance on another available node.
- If a local network_down occurs on an XD_data network, PowerHA moves the replicated resource groups that are ONLINE on the particular node to another available node on that site. This function of the primary instance is mirrored to the secondary instance so that secondary instances may be recovered via selective failover.

You can disable this default behavior by using the customized resource group recovery policy. Figure 9-9 shows the panel to disable the selective failover behavior between sites. Run the **smitty hacmp** command. Select **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Customize Resource Group and Resource Recovery → Customize Inter-Site Resource Group Recovery**. Then, press Enter.

```
                  Customize Inter-Site Resource Group Recovery

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

                                                      [Entry Fields]
  Action on Resource Group Failure                    failover              +
  Notify Method                                      []
```

Figure 9-9 Intersite RG recovery SMIT panel - default setting

If the notify method is specified, you can achieve similar behavior as when you define a HACMPpager event notification. The pager event notification uses *sendmail* to forward an email to the account specified.

Even if the selective failover function is disabled, PowerHA still moves the resource group if a node_down or node_up event occurs. Also, a user-designated rg_move operation continues to behave as expected with the setting disabled.

Tip: The PowerHA software documentation states that when you select this option, you can specify the notify method for the individual resource groups:

> "The resource groups that contain nodes from more than one site are listed. The resource groups include those with a site management policy of Ignore. These resource groups are not affected by this function even if you select one of them."

In our PowerHA 6.1 clusters, when this option is selected, it applies to all resource groups in the cluster. The granularity is not available for individual resource groups.

Disabling the selective failover test scenario

We tested this feature on our four-node cluster that is configured for SRDF replication between the sites. We began with the resource groups in the state that is shown in Example 9-4.

Example 9-4 Resource groups

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                        Node
-----------------------------------------------------------------------------
```

```
RG_sitea        ONLINE                      xdemca1@siteA
                OFFLINE                     xdemca2@siteA
                ONLINE SECONDARY            xdemcb1@siteB
                OFFLINE                     xdemcb2@siteB
```

For our first test, we failed all connections to the storage from each of the nodes at site A and watched the cluster lose quorum and attempt a move from node A1 to node A2 (Example 9-5).

Example 9-5 Failed connections to storage

```
# errpt
CAD234BE   0322123010 U H LVDD          QUORUM LOST, VOLUME GROUP CLOSING
52715FA5   0322123010 U H LVDD          FAILED TO WRITE VOLUME GROUP STATUS AREA
E86653C3   0322123010 P H LVDD          I/O ERROR DETECTED BY LVM
```

The cluster first released the resource group on node A1 and attempted to acquire it on node A2. After it detected that all connections were unavailable, it moved the resource group to the second site as expected (Example 9-6).

Example 9-6 Resource group moved to second site

```
root@xdemca1:/>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                       Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE SECONDARY            xdemca1@siteA
               OFFLINE                     xdemca2@siteA
               ONLINE                      xdemcb1@siteB
               OFFLINE                     xdemcb2@siteB
```

This test proved that selective failover on volume group loss worked. Next, we ran the same test with the selective failover option disabled:

1. Disable the selective failover setting by changing the default behavior to notify (Figure 9-10).

```
                Customize Inter-Site Resource Group Recovery

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

                                                      [Entry Fields]
  Action on Resource Group Failure                   notify                +
  Notify Method                                     [/usr/local/scripts/no>
```

Figure 9-10 Inter-site RG recovery SMIT panel - selective failover disabled

As shown in Example 9-7, the script was configured only to validate that the failover was started.

Example 9-7 Script validating the failover

```
root@xdemca1:/usr/local/scripts>ls -l
total 8
-rwxr-----    1 root     system           50 Mar 22 13:59 notify_failover.sh
```

```
cat notify_failover.sh
#!/bin/ksh
touch /usr/local/scripts/ran.sh
exit 0
```

2. Synchronize the change across all cluster nodes.

Attention: In our scenario, we only defined the **notify_failover.sh** script on the first node in the cluster. The cluster verification did not check or complain that it did not exist on all cluster nodes.

3. Disable all paths to the disks on both nodes at site A (Example 9-8).

Example 9-8 Disabling all paths to the disks on both nodes at site A

```
On Node A1
# powermt disable hba=0
# powermt disable hba=1

On Node A2
# powermt disable hba=0
# powermt disable hba=1

root@xdemca1:/usr/local/scripts>powermt display
Symmetrix logical device count=58
CLARiiON logical device count=0
Hitachi logical device count=0
Invista logical device count=0
HP xp logical device count=0
Ess logical device count=0
HP HSx logical device count=0
================================================================================
----- Host Bus Adapters ---------  ------ I/O Paths -----  ------ Stats ------
###  HW Path                       Summary   Total   Dead  IO/Sec Q-IOs Errors
================================================================================
   0 fscsi0                       failed      140    140       -     0      0
   1 fscsi1                       failed      140    140       -     0      0
```

4. Monitor the status of the resource group by using the repeated **clRGinfo** commands.

The results of our second test showed that with the selective failover option disabled the cluster first attempted an rg_move between the local nodes. After it attempted to acquire and release the resources on the second local node, it left the resources in the state that is shown in Example 9-9.

Example 9-9 Testing the selective failover option

```
root@xdemca1:/usr/local/scripts>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ERROR                          xdemca1@siteA
               ERROR                          xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               ERROR                          xdemcb2@siteB
```

While in the state shown in Example 9-9, we confirmed the status of the replicated LUNs. Since we only altered access to the storage from the host, the replicated LUNs remained in a consistent state (Example 9-10).

Example 9-10 State of the LUNs

```
root@xdemca1:/usr/local/scripts>symrdf list pd
Symmetrix ID: 000190100304

                            Local Device View
------------------------------------------------------------------------
                    STATUS     MODES                     RDF  S T A T E S
Sym        RDF      ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDATE  Tracks   Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- -------------

0F53 00BF   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized
0F54 00C0   R1:41   RW RW RW   S..1-       0        0 RW  WD   Synchronized

Total                                -------- --------
  Track(s)                                  0        0
  MB(s)                                   0.0      0.0

Legend for MODES:

 M(ode of Operation)   : A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
 D(omino)              : X = Enabled, . = Disabled
 A(daptive Copy)       : D = Disk Mode, W = WP Mode, . = ACp off
 (Mirror) T(ype)       : 1 = R1, 2 = R2
 (Consistency) E(xempt): X = Enabled, . = Disabled, M = Mixed, - = N/A
```

From the results of our tests, we concluded that disabling the selective failover function though the Inter-Site Resource Group Recovery panel works as expected.

To recover our environment, we brought the paths to the storage back by re-enabling the HBAs as follows:

```
# powermt enable hba=0
# powermt enable hba=1
```

We followed this step with a resource group operation to bring the resource group back online. After processing the resources, our resources were activated once again (Example 9-11).

Example 9-11 Activated resources

```
root@xdemca1:/usr/local/scripts>clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
RG_sitea       ONLINE                         xdemca1@siteA
               OFFLINE                        xdemca2@siteA
               ONLINE SECONDARY               xdemcb1@siteB
               OFFLINE                        xdemcb2@siteB
```

9.3 Partitioned cluster considerations

A split-brain, partitioned, or node isolation condition refers to a situation where more than one server activates the resources as if it is the primary node. Such a situation can occur when the communication is lost between the servers, and each site believes that it is the only one still online. In an enterprise cluster, the nodes at the recovery site activate the replicated volumes and bring the resources online. In certain situations, data divergence might occur, meaning that both sites might contain inconsistent copies and manual intervention is required to recover.

To protect against this condition, PowerHA uses Reliable Scalable Clustering Technologies (RSCT) to monitor and detect network failures. RSCT sends heartbeats over IP and non-IP networks. Using the information from RSCT, PowerHA handles three failure types:

- Network interface card (NIC) failure
- Network failure
- Node failure

The RSCT topology services daemon uses specially crafted packet transmission patterns to diagnose a failure by ruling out alternatives. Figure 9-11 illustrates this process.

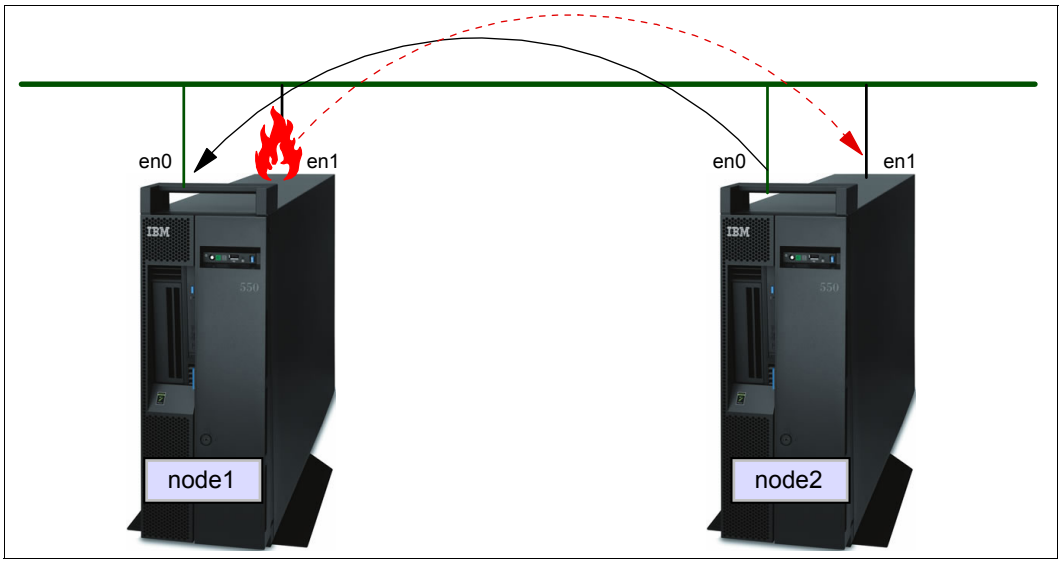


Figure 9-11 RSCT transmission patterns

In Figure 9-11, RSCT performs the following failure detection and diagnosis:

1. RSCT on node1 notices that heartbeat packets are no longer arriving from en1 and notifies node2 (which has also noticed that heartbeat packets are no longer arriving from en1).
2. RSCT on both nodes sends diagnostic packets between various combinations of NICs (including out through one NIC and back in by using another NIC on the same node).
3. The node realizes that all packets that involve node1's en1 are vanishing, but packets that involve node2's en1 are being received.
4. The diagnosis is that node1's en1 has failed.
5. After it determines that it is unable to communicate with the other node, RSCT concludes that the other node has failed and must be taken over.

9.3.1 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 acquire the data on its own storage copies. If this situation occurs, you risk losing access to the disks and potentially experience data divergence.

You can prevent partitioning between the sites in the following ways:

1. Define multiple IP networks between the sites.

   Having multiple network communication paths between the sites minimizes the risk that the clusters falsely attempt to activate the resources on the remote site. To achieve true redundancy, the networks must be backed by a separate network infrastructure. If that is not possible, there may not be any advantage to having a separate network defined. Instead, the multiple interfaces could be aggregated to form a single redundant logical interface, which adds redundancy to the communication interface.

2. Define a non-IP network between the sites.

   The use of alternate non-IP networks is also supported in the PowerHA Enterprise Edition. Suggestions for the devices that are required for an XD_rs232 network are explained more in 2.1, “Network considerations” on page 40. If the IP stack is not available, the extended serial network provides a means of communication.

3. Share small LUNs for disk heartbeating between the sites.

   In configurations where the SAN spans the two sites, leveraging SAN-based disk heartbeating is another option. When using disk replication, the heartbeats do not pass over any of the replicated LUNs. However, a small dedicated LUN from either subsystem might be mapped to both sites, and two disk heartbeat networks might be defined like in a cross-site LVM mirrored environment.

Discontinued modems: The use of modems between the two sites that use the dial back fail safe (DBFS) functionality has been discontinued.

The available XD network types each provide separate functions. However, they still behave the same as the local networks. The major difference is the slower failure detection rate cycle. The FDR attributes are compared in detail in 2.1, “Network considerations” on page 40.

XD_data A network that can be used for data replication only. This network supports adapter swap, but not failover to another node. RSCT heartbeat packets will be sent on this network. PowerHA 5.4.0 and later versions support EtherChanneling links and up to four XD_data networks.

XD_ip An IP-based network that is used for participation in RSCT protocols, heartbeating, and client communication. RSCT heartbeat packets are sent on this network.

XD_rs232 A network that can be used for serial communications of the same type as the RS232 network type, except that the heartbeat parameters have been modified for the greater distance. RSCT heartbeat packets are sent on this network.

Considering the limitations of an environment's infrastructure can help to identify single points of failure. For example, consider an environment where both network and I/O traffic pass over the same infrastructure (Figure 9-12).

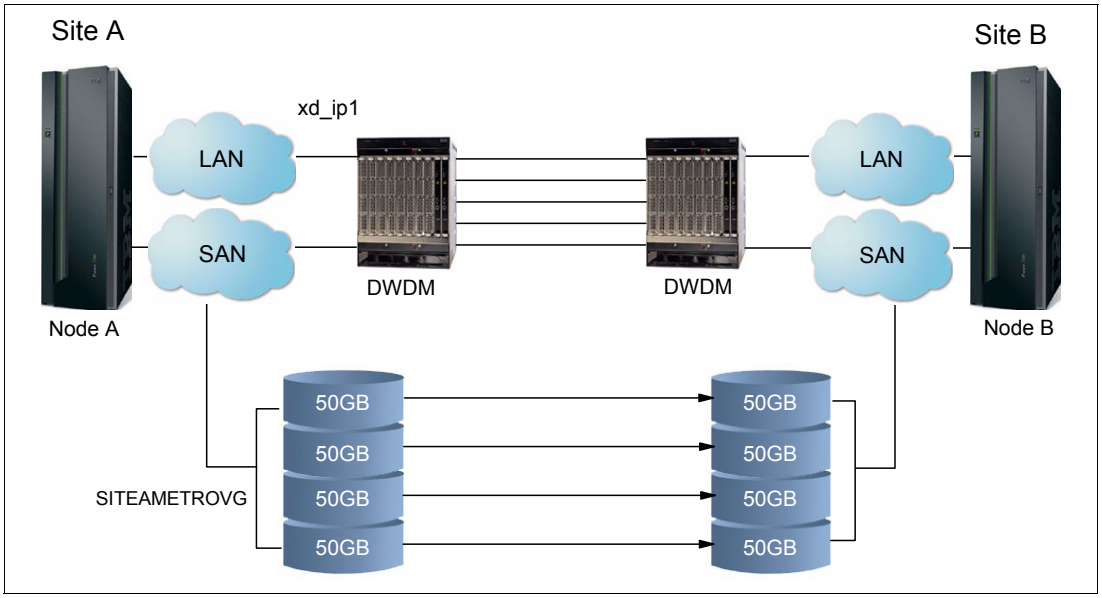


Figure 9-12 Single XD_IP network 4-node SAN Volume Controller Metro Mirror environment

In Figure 9-12, both Ethernet and fiber communications pass between the DWDM devices at each site. If these interconnects are the only communication links between the sites, a disruption to them causes cluster partitioning. If that situation occurs, the cluster nodes at each site assume primary roles for the applications.

Figure 9-13 shows the same environment, but includes additional networks based on the recommendations described previously.

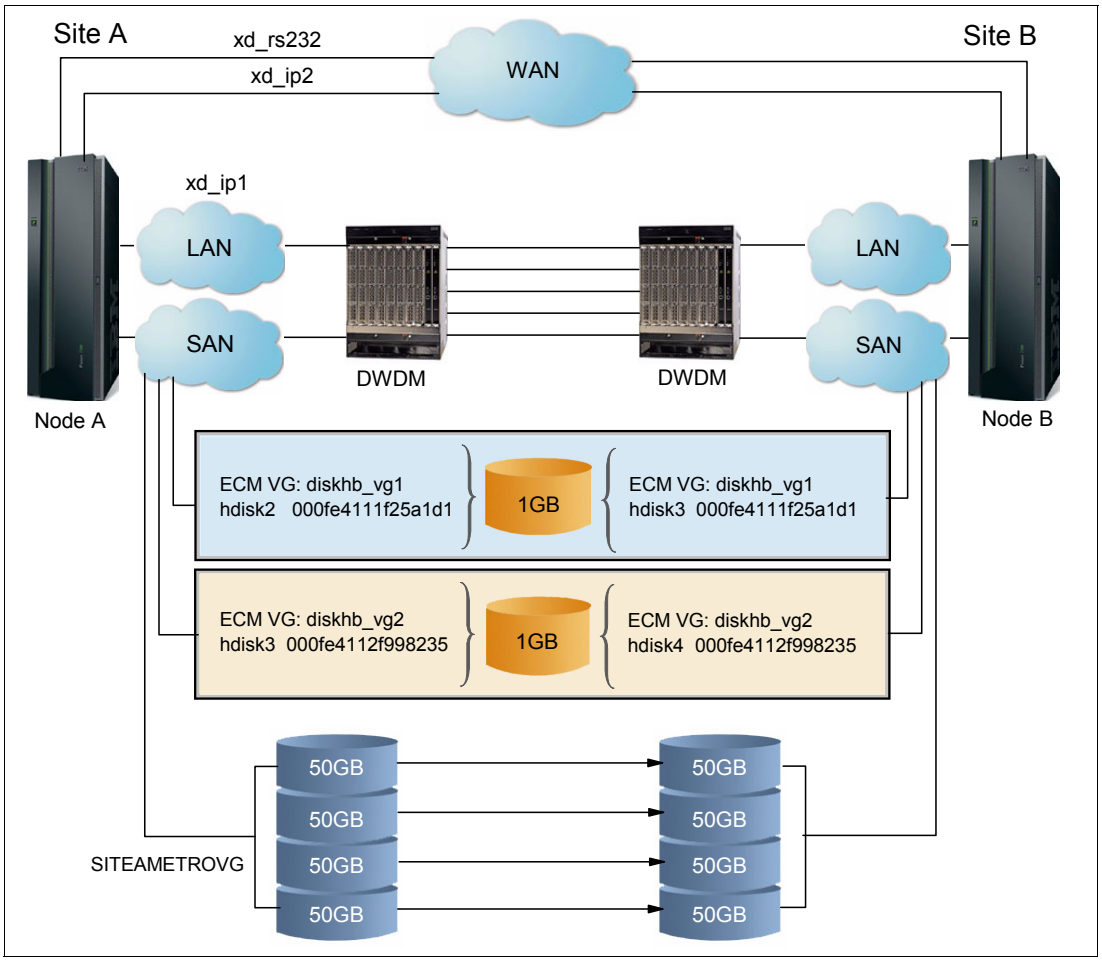


Figure 9-13 Redundant inter-site networks in a 4-node Metro Mirror environment

In addition to the XD_ip network that communicates over the DWDM infrastructure, we included a second XD_ip network that uses a separate network back bone. We also added an XD_rs232 network and two dedicated shared LUNs for heartbeat networks between the sites. Each one assume that the requirements and infrastructure are available to configure them. Each one provides protection against separate component failures and individually hardens the environment.

The second XD_ip network assumes that there are alternate switches that provide connectivity between the sites. The XD_rs232 network assumes that you also have the required hardware and infrastructure to support it. If neither option is available, multiple networks within the DWDM interconnects might be defined with the assumption that the physical links can be the single point of failure.

Sharing dedicated LUNs strictly for heartbeating is only possible if the inter-site connection provides an extended view of the SAN. If so, dedicated LUNs could be shared between the systems at each site. In our example, we used a LUN from each storage enclosure and defined two separate disk heartbeat networks. To accommodate this setup, the appropriate zoning and storage mappings must be in place. The heartbeating takes place independently from the replication of the LUNs set to copy data between the sites.

The heartbeat traffic for the disk heartbeat networks also passes over the DWDM connections. However, the added value is that there is still a heartbeating link if the physical interfaces that support the XD_ip network are offline. Our diagram only shows a single node in each site. Multisite implementations with additional nodes at each site need to balance between designing a ring topology for the disk heartbeat networks and whether to map a LUN from each storage subsystem to have even more redundant networks.

Environments that are dispersed in such a way that the SAN could not be extended ultimately rely on redundant IP-based networks on separately leased lines and XD_rs232 extended serial networks.

9.3.2 Expected behaviors in a partitioned cluster

A site-specific disaster might not always occur in an orderly fashion. Several conditions might occur during a disaster. The most ideal scenario is an event that causes the entire site to go down. One example of this situation is the complete loss of power. Neither the storage nor the servers at the primary site remain online and the resulting failover automatically acquires the resources.

Failures causing only the server or the storage subsystem to remain online introduce additional complexities. The cluster, however, is designed to accommodate them. If the server fails and the storage remains online, the cluster automatically fails over and redirects the flow of the replication. If the storage enclosure fails, the inherent selective failover behavior on volume group loss relocates the resource group to the second site.

The scenario that can put the cluster more at risk is if the communication across all of the links between the two sites suddenly ceases. The number of networks that span the sites and their corresponding failure detection rate settings determine how long it takes for the cluster to recognize a site as being down. For information about the FDR network settings, see 2.1, “Network considerations” on page 40.

Depending on the replication type that is being used, the cluster behaves differently during and after the acquisition of the resources at the remote site.

The scenarios that we reviewed for the partitioned cluster scenarios included our SAN Volume Controller Metro Mirror and Global Mirror Cluster, DS Metro Mirror Cluster, and EMC SRDF environments.

Scenario test results

In our environment, we tested bringing down the XD_ip links between the sites and partitioning our three 4-node PowerHA clusters. We repeated the tests in the SAN Volume Controller, DS, and EMC replication clusters that are used in the previous chapters.

To partition our clusters, first, we brought down the IP links between the sites. Rather than logging on to the network switch and disabling the ports, we brought down the XD_IP interfaces on the primary site by issuing:

```
# ifconfig en1 down detach
```

The resulting behavior in all clusters when the loss of heartbeat communication for the XD_ip network was acknowledged was for the acquisition of the resources at the remote site. No graceful release of resources took place, and the applications remained active at the primary site. After the failure detection rate cycle expired, we confirmed that the cluster brought the resources online on the remote site by using the **clRGinfo** command.

However, in the processing of the resources on site B, the cluster redirected the relationship between the source and target LUNs. The replicated LUNs at the second site were changed to a master role and the original site was changed to an auxiliary role. This change directly impacted access to the LUNs on the primary site. In our tests, we lost write access to the disks and several of our commands hung.

Example 9-12 shows output collected from the tests that were performed on our DS cluster while in a partitioned state.

Example 9-12 Output from the DS cluster in a partitioned state

```
On the primary node on Site A hosting the resources:
dscli> lspprc -dev IBM.2107-75BALB1 -remotedev IBM.2107-7585461 8001-81ff

ID        State       Reason                     Type         SourceLSS Timeout
==============================================================================
8001:2001 Suspended   Internal Conditions Target Metro Mirror 80        60
8003:2003 Suspended   Internal Conditions Target Metro Mirror 80        60
8004:2004 Suspended   Internal Conditions Target Metro Mirror 80        60
8101:3001 Suspended   Internal Conditions Target Metro Mirror 81        60

From the secondary node on Site B also now hosting the resources:
dscli> lspprc -dev IBM.2107-7585461 -remotedev IBM.2107-75BALB1 2000-30ff

ID        State     Reason      Type         SourceLSS Timeout (secs)
============================================================================
2001:8001 Suspended Host Source Metro Mirror 20        60
2003:8003 Suspended Host Source Metro Mirror 20        60
2004:8004 Suspended Host Source Metro Mirror 20        60
3001:8101 Suspended Host Source Metro Mirror 30        60
3002:8102 Suspended Host Source Metro Mirror 30        60
```

We did not have an active application or generate enough I/O to experience a system crash on our active nodes on site A. However, in a real-life scenario, a flood of LVM I/O failures because of the inability to write to disk might result in a crash.

Figure 9-14 shows the SAN Volume Controller resources online on both sites with the role of the source and target copies reversed.

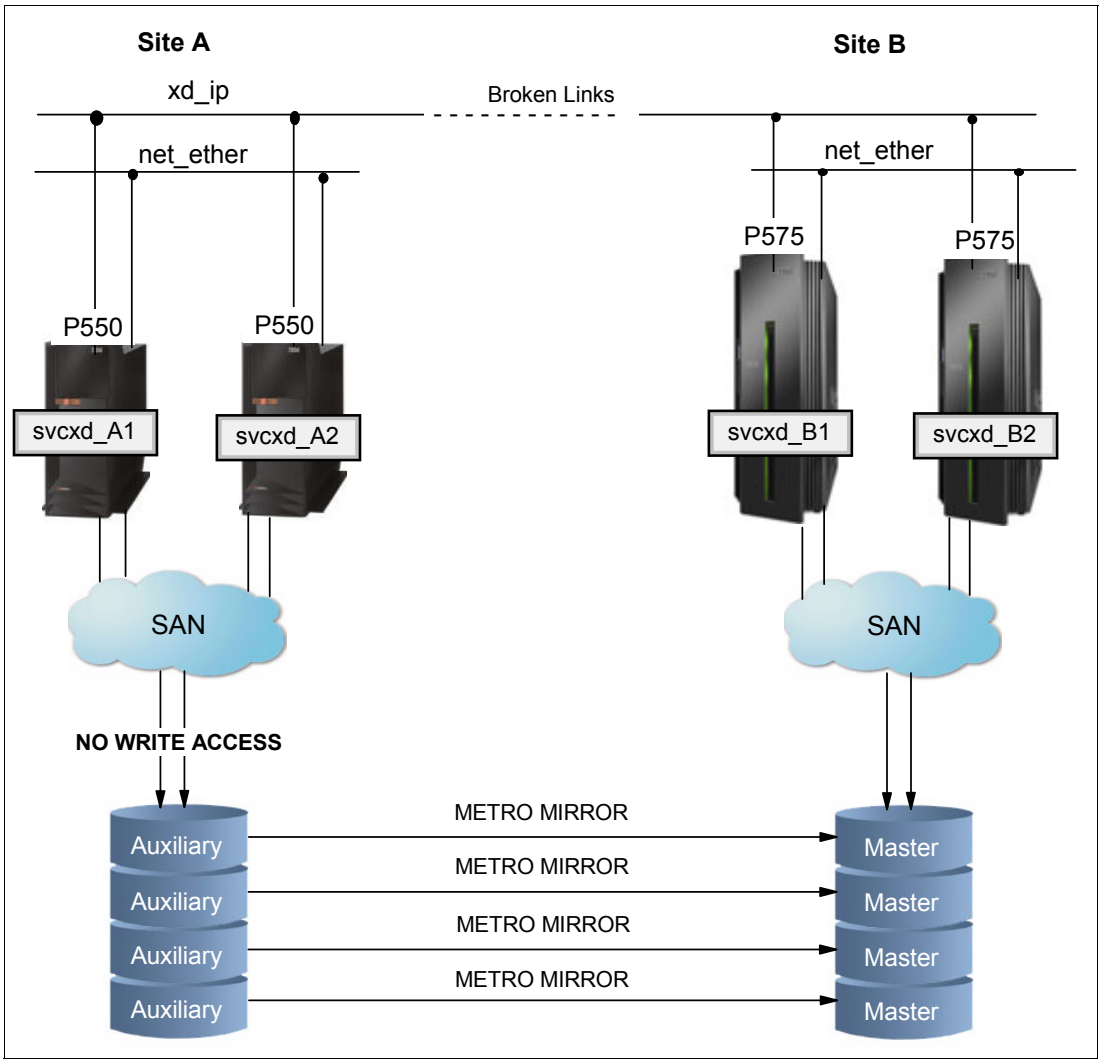


Figure 9-14 Partitioned SAN Volume Controller cluster scenario with resources active on site B

In a split scenario, we can actively run on the second site. Each of our sites is configured with the designated IPs in separate network segments. Therefore, upon failover, we did not experience duplicate IP address errors in the error log, which might be another consideration in a real customer scenario, depending on the setup.

One of the considerations for our testing was the effect of an intermittent problem between the sites. Therefore, the next step was to start the IP links again. After the IP links were started, the RSCT communication was re-established and the nodes on the remote site (site B) were halted. The halt was a result of the cluster that identifies that the resources were hosted on both sites. The RSCT group services experience a domain merge and log a GS_DOM_MER_ERR in the error report, which leads to the clstrmgrES core dumping and halts the nodes (Example 9-13 on page 464).

Example 9-13 RSCT error report message during domain merge

```
LABEL:          GS_DOM_MERGE_ER
IDENTIFIER:     9DEC29E1

Date/Time:       Wed Mar 24 19:48:06 2010
Sequence Number: 206835
Machine Id:      00CA02EF4C00
Node Id:         svcxd_b1
Class:           O
Type:            PERM
WPAR:            Global
Resource Name:   grpsvcs

Description
Group Services daemon exit to merge domains

Probable Causes
Network between two node groups has repaired

Failure Causes
Network communication has been blocked.
Topology Services has been partitioned.

        Recommended Actions
        Check the network connection.
Check the Topology Services.
Verify that Group Services daemon has been restarted
Call IBM Service if problem persists

Detail Data
DETECTING MODULE
RSCT,NS.C,1.107.1.49,4461
ERROR ID
6VbOvROqGee9/BU4/ckbQ7.....................
REFERENCE CODE

DIAGNOSTIC EXPLANATION
NS::Ack(): The master requests to dissolve my domain because of the merge with
other d
omain 1.24
```

These results are expected to protect data integrity by not ever allowing non-concurrent resources on both sites at the same time. However, a word of caution here is that we were left with the active nodes on site A with no write access to the LUNs. Effectively, the application services were now offline, and we were left with no choice but to recover the environment manually.

9.3.3 Recommendations for recovery

Depending on the types of failures that occur between the sites, the steps that are required to recover might be different. In a partitioned scenario that requires manual intervention, the primary objective is to identify the site that has the most current and consistent version of the data. This situation might not be identified easily, especially if it requires running queries from within the application.

To evaluate the situation and assess the damage, the status of the following items must be identified:

- State of the cluster nodes
- State of the heartbeat communication paths
- Consistency of the replicated volumes
- Status of the application data

The role of the planning team and the system architect is to design the environment so that it minimizes the risk of an intermittent network failure like the one described in the previous scenario. If the cluster becomes partitioned, you must quickly identify the component responsible for the failure. If network access is lost, attempting to reach the servers via other interfaces might be required. Logging on to the management consoles (HMC) might be required to determine whether the LPARs are still online. Bringing down all of the nodes in one of the sites may be the best plan to avoid a halt of certain servers if the heartbeat communication suddenly resumes.

A normal cluster shutdown can fail if access to the volumes is not available. A cluster stop starts the application stop scripts, which attempt to end the corresponding processes. However, cluster processing also attempts to unmount and vary off the resources, which might fail if access to the volumes is unavailable. In such a situation, performing a hard reset of the nodes is the best approach.

On the AIX hosts

A review of the AIX error report provides a good indication of whether access to separate components is unavailable. I/O errors, duplicate IP messages, and interface failure messages are easily identifiable as follows:

```
# errpt
CAD234BE   0322123010 U H LVDD           QUORUM LOST, VOLUME GROUP CLOSING
52715FA5   0322123010 U H LVDD           FAILED TO WRITE VOLUME GROUP STATUS AREA
E86653C3   0322123010 P H LVDD           I/O ERROR DETECTED BY LVM
```

Polling the status of the cluster resources shows you whether the cluster considers the resource group to be online, even if the application is not operational. To avoid this situation, have application monitors in place. If none are configured, polling the process table and checking the status of the specific applications also give an indication of the responsiveness on the system.

The system administrator already has a list of checks for the individual applications that are being hosted on the servers. Certain commands might hang when you try to poll the status if some of the backing resources are unavailable:

```
# ps -ef | grep ora_pmon_hatest
oracle  548872       1   0 17:19:23      -  0:00 ora_pmon_hatest
```

It may still be difficult to predict which servers to bring offline. Therefore, performing such actions as querying the disks and checking the status of the replicated resources is another critical part of troubleshooting (Example 9-14).

Example 9-14 Status of the replicated resources

```
# lquerypv -h /dev/hdisk4
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   000FE411 2579EE4C 00000000 00000000  |....%y.L........|
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  |................|

When not responsive the command may hang or not any information:
# lquerypv -h /dev/hdisk
   ....
```

The command may hang and eventually time out if access to the disks is not available. The method that is used to check the status of the replicated resources varies depending on the replication type that is being used.

Example 9-15 shows the status from the primary node of a SAN Volume Controller cluster in a partitioned scenario. In this output, the A1 node originally hosts the resources. After the remote site activates the resources, it changes the role of the source volumes from master to auxiliary. When this situation occurs, we are unable to query the disks.

Example 9-15 Primary node status on a SAN Volume Controller cluster for a partitioned scenario

```
[svcxd_a1][/]> ssh admin@B12_4F2 svcinfo lsrcconsistgrp svc_metro
id 0
name svc_metro
master_cluster_id 0000020064009B10
master_cluster_name B12_4F2
aux_cluster_id 0000020060A0469E
aux_cluster_name B8_8G4
primary aux
state consistent_synchronized
relationship_count 4
freeze_time
status
sync
copy_type metro
RC_rel_id 0
RC_rel_name svc_disk2
RC_rel_id 1
RC_rel_name svc_disk3
RC_rel_id 2
RC_rel_name svc_disk4
RC_rel_id 3
RC_rel_name svc_disk5
```

To determine the status of the network, we attempt to ping various routers and servers in the respective subnets. We also check the link status of the NICs by using the **ifconfig**, **entstat**, or **netstat** commands when using physical adapters on the host.

Example 9-16 shows the adapters that are active, but they are different if a problem occurs. When you use virtual Ethernet adapters from VIOS, troubleshooting might require a deeper investigation to identify the reason for the network loss.

Example 9-16 Adapter status on the host

```
# ifconfig -a

en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 192.168.8.103 netmask 0xffffff00 broadcast 192.168.8.255
        inet 192.168.100.173 netmask 0xffffff00 broadcast 192.168.100.255
        inet 192.168.100.54 netmask 0xffffff00 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>
        inet 10.12.5.36 netmask 0xfffffc00 broadcast 10.12.7.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0: flags=e08084b<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
        inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
        inet6 ::1/0
         tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
```

If a site split occurs, there might be situations in which a client considers bringing down the entire environment and manually activating the resources. The client might then perform manual integrity checks to better assess the situation. When the consistency of the data is confirmed, the cluster might be restarted in a phased approach. First, reactivate the cluster on the most appropriate site and then integrate the remaining cluster nodes that are based on feedback from the specific network, storage, and application administrators.

Summary

In conclusion, the use of the PowerHA Enterprise Edition with the various replication methods only provides as much resiliency as the infrastructure that supports it. Therefore, the significance of addressing potential single points of failure in a multisite cluster is even more apparent. Evaluate and plan the overall infrastructure in such a fashion that the risk of a partitioned cluster is minimized. The cluster software attempts to always protect the integrity of the data and efficiently manage the resources. Use such features as the custom application monitoring and pager notification methods if no other notification is already provided by separate external monitoring software. Going forward, the PowerHA SystemMirror Enterprise Edition plans to release additional enhancements and will continue to serve as the premier HA and disaster recovery solution on AIX for Power Systems.

10

Disaster recovery with DS8700 Global Mirror

This chapter explains how to configure disaster recovery based on IBM PowerHA SystemMirror for AIX Enterprise Edition by using IBM System Storage DS8700 Global Mirror as a replicated resource. This support was added in version 6.1 with service pack 3 (SP3).

This chapter includes the following sections:

- Planning for Global Mirror
- Installing the DSCLI client software
- Scenario description
- Configuring the Global Mirror resources
- Configuring AIX volume groups
- Configuring the cluster
- Failover testing
- LVM administration of DS8000 Global Mirror replicated resources

10.1 Planning for Global Mirror

Proper planning is crucial to the success of any disaster recovery solution. This topic outlines the basic requirements to implement Global Mirror and integrate it with the IBM PowerHA SystemMirror for AIX Enterprise Edition.

10.1.1 Software prerequisites

Global Mirror functions work with all the AIX levels that are supported by PowerHA SystemMirror Standard Edition. The following software is required for the configuration of the PowerHA SystemMirror for AIX Enterprise Edition for Global Mirror:

- The following base file sets for PowerHA SystemMirror for AIX Enterprise Edition 6.1:
  - `cluster.es.pprc.cmds`
  - `cluster.es.pprc.rte`
  - `cluster.es.spprc.cmds`
  - `cluster.es.spprc.rte`
  - `cluster.msg.en_US.pprc`

  PPRC and SPPRC file sets: The PPRC and SPPRC file sets are not required for Global Mirror support on PowerHA.

- The following additional file sets included in SP3 (must be installed separately and require the acceptance of licenses during the installation):
  - `cluster.es.genxd`

    ```
    cluster.es.genxd.cmds      6.1.0.0 Generic XD support - Commands
    cluster.es.genxd.rte       6.1.0.0 Generic XD support - Runtime
    ```

  - `cluster.msg.en_US.genxd`

    ```
    cluster.msg.en_US.genxd    6.1.0.0 Generic XD support - Messages
    ```

- AIX supported levels:
  - 5.3 TL9, RSCT 2.4.12.0 or later
  - 6.1 TL2 SP1, RSCT 2.5.4.0 or later
- The IBM DS8700 microcode bundle 75.1.145.0 or later
- DS8000 CLI (DSCLI) 6.5.1.203 or later client interface (must be installed on each PowerHA SystemMirror node):
  - Java 1.4.1 or later
  - APAR IZ74478, which removes the previous Java requirement
- The path name for the DSCLI client in the PATH for the root user on each PowerHA SystemMirror node (must be added)

10.1.2 Minimum DS8700 requirements

Before you implement PowerHA SystemMirror with Global Mirror, you must ensure that the following requirements are met:

- Collect the following information for all the HMCs in your environment:
  - IP addresses
  - Login names and passwords
  - Associations with storage units

- Verify that all the data volumes that must be mirrored are visible to all relevant AIX hosts.
- Verify that the DS8700 volumes are appropriately zoned so that the IBM FlashCopy® volumes are not visible to the PowerHA SystemMirror nodes.
- Ensure all Hardware Management Consoles (HMCs) are accessible by using the Internet Protocol network for all PowerHA SystemMirror nodes where you want to run Global Mirror.

10.1.3 Considerations

The PowerHA SystemMirror Enterprise Edition with DS8700 Global Mirror has the following considerations:

- The AIX Virtual SCSI is not supported in this initial release.
- No auto-recovery is available from a PPRC path or link failure.

  If the PPRC path or link between Global Mirror volumes breaks down, the PowerHA Enterprise Edition is unaware of it. (PowerHA does not process Simple Network Management Protocol (SNMP) for volumes that use DS8K Global Mirror technology for mirroring). In this case, the user must identify and correct the PPRC path failure. Depending on timing conditions, such an event can result in the corresponding Global Mirror session to go to a unrecoverable state. If this situation occurs, the user must manually stop and restart the corresponding Global Mirror Session (by using the **`rmgmir`** and **`mkgmir`** DSCLI commands) or an equivalent DS8700 interface.
- Cluster Single Point Of Control (C-SPOC) cannot perform some Logical Volume Manager (LVM) operations on nodes at the remote site that contain the target volumes.

  Operations that require nodes at the target site to read from the target volumes result in an error message in C-SPOC. Such operations include such functions as changing the file system size, changing the mount point, and adding LVM mirrors. However, nodes on the same site as the source volumes can successfully perform these tasks, and the changes can be propagated later to the other site by using a lazy update.

  Attention: For C-SPOC operations to work on all other LVM operations, you must perform all C-SPOC operations with the DS8700 Global Mirror volume pairs in a synchronized or consistent state. Alternatively, you must perform them in the *active* cluster on all nodes.

- The volume group names must be listed in the same order as the DS8700 mirror group names in the resource group.

10.2 Installing the DSCLI client software

You can download the latest version of the DS8000 DSCLI client software from the following web page:

`ftp://ftp.software.ibm.com/storage/ds8000/updates/DS8K_Customer_Download_Files/CLI`

Install the DS8000 DSCLI software on each PowerHA SystemMirror node. By default, the installation process installs the DSCLI in the `/opt/ibm/dscli` directory. Add the installation directory of the DSCLI into the PATH environment variable for the root user.

For more information about the DS8000 DSCLI, see the *IBM System Storage DS8000: Command-Line Interface User's Guide*, SC26-7916.

10.3 Scenario description

This scenario uses a three-node cluster named *Txrmnia*. Two nodes are in the primary site, *Texas*, and one node is in the site *Romania*. The `jordan` and `leeann` nodes are at the Texas site and the `robert` node is at the Romania site. The primary site, Texas, has both local automatic failover and remote recovery. Figure 10-1 provides a software and hardware overview of the tested configuration between the two sites.

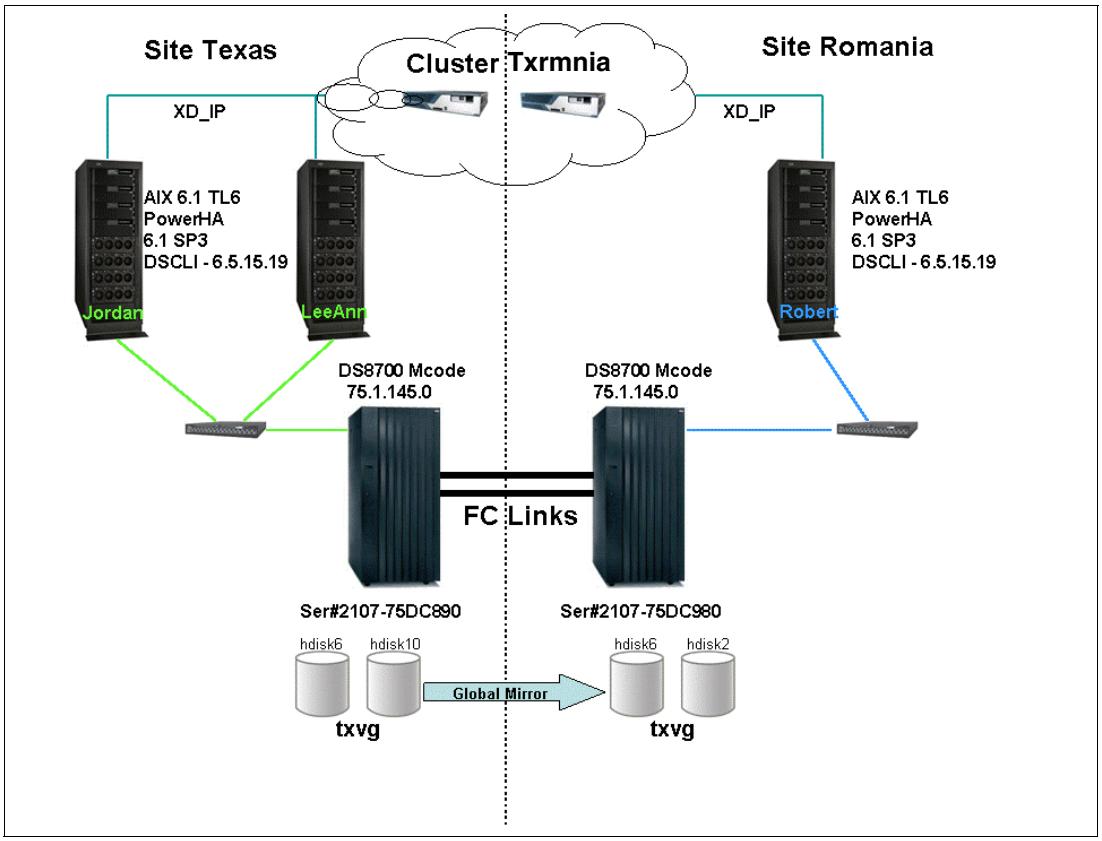


Figure 10-1 DS8700 Global Mirror test scenario

For this test, the resources are limited. Each system has a single IP, an XD_ip network, and single Fibre Channel (FC) host adapters. Ideally, redundancy might exist throughout the system, including in the local Ethernet networks, cross-site XD_ip networks, and FC connectivity. This scenario has a single resource group, `ds8kgmrg`, which consists of a service IP address (`service_1`), a volume group (`txvg`), and a DS8000 Global Mirror replicated resource (`texasmg`). To configure the cluster, see 10.6, “Configuring the cluster” on page 483.

10.4 Configuring the Global Mirror resources

This section explains how to perform the following tasks:

- Checking the prerequisites
- Identifying the source and target volumes
- Configuring the Global Mirror relationships

For each task, the DS8000 storage units are already added to the storage area network (SAN) fabric and zoned appropriately. Also, the volumes are already provisioned to the nodes.

For information about how to set up the storage units, see *IBM System Storage DS8700 Architecture and Implementation*, SG24-8786.

10.4.1 Checking the prerequisites

To check the prerequisites, follow these steps:

1. Ensure that the DSCLI installation path is in the PATH environment variable on all nodes.
2. Verify that you have the appropriate microcode version on each storage unit by running the **ver -lmc** command in a DSCLI session as shown in Example 10-1.

Example 10-1 Checking the microcode level

```
(0) root @ r9r4m21: : /
# dscli -cfg /opt/ibm/dscli/profile/dscli.profile.hmc1
Date/Time: October 6, 2010 2:15:33 PM CDT IBM DSCLI Version: 6.5.15.19  DS:
IBM.2107-75DC890

dscli> ver -lmc
Date/Time: October 6, 2010 2:15:41 PM CDT IBM DSCLI Version: 6.5.15.19 DS: -
Storage Image    LMC
=========================
IBM.2107-75DC890 5.5.1.490
dscli>
```

3. Check the code bundle level that corresponds to your LMC version on the “DS8700 Code Bundle Information” web page at:

   http://www.ibm.com/support/docview.wss?uid=ssg1S1003593

   The code bundle level must be at version 75.1.145.0 or later. Also on the same page, verify that your displayed DSCLI version corresponds to the installed code bundle level or a later level.

Example 10-2 shows the extra parameters that are inserted into the DSCLI configuration file for the storage unit in the primary site, `/opt/ibm/dscli/profile/dscli.profile.hmc1`. Adding these parameters helps to prevent from having to type them each time they are required.

Example 10-2 Editing the DSCLI configuration file

```
username: redbook
password: r3dbook
hmc1:   9.3.207.122
devid:          IBM.2107-75DC890
remotedevid:    IBM.2107-75DC980
```

10.4.2 Identifying the source and target volumes

Figure 10-2 on page 474 shows the volume allocation in DS8000 units for the scenario in this chapter. Global Copy source volumes are attached to both nodes in the primary site, Texas, and the corresponding Global Copy target volumes are attached to the node in the secondary site, Romania. The gray volumes, FlashCopy targets, are not shown to the hosts.

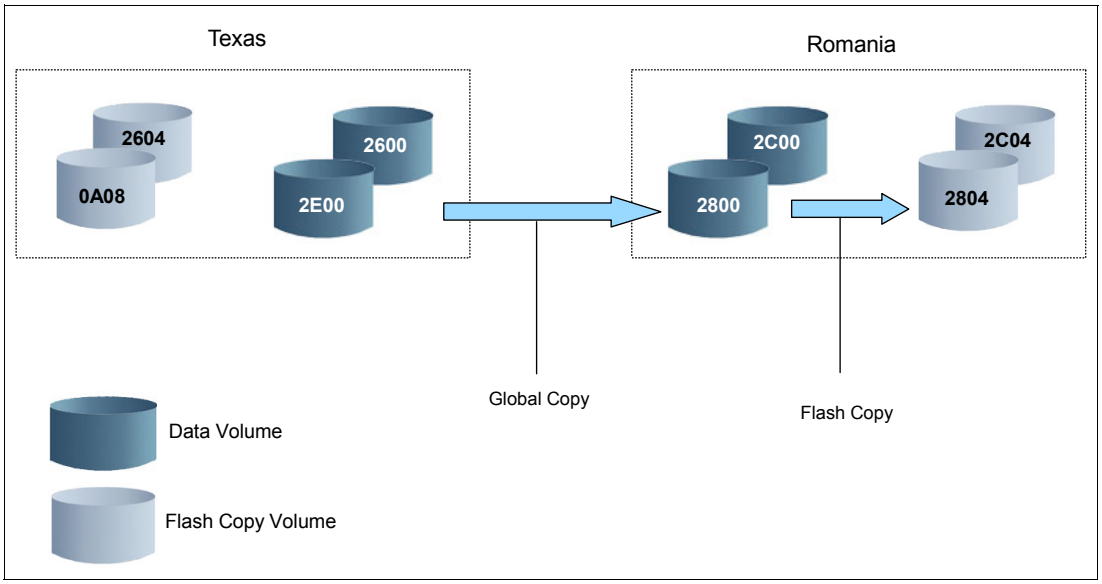


Figure 10-2 Volume allocation in DS8000 units

Table 10-1 shows the association between the source and target volumes of the replication relationship and between their logical subsystems (LSS, the two most significant digits of a volume identifier that are highlighted in bold in the table). Table 10-1 also indicates the mapping between the volumes in the DS8000 units and their disk names on the attached AIX hosts.

Table 10-1 AIX hdisk to DS8000 volume mapping

Site Texas		Site Romania	
AIX disk	**LSS/VOL ID**	**LSS/VOL ID**	**AIX disk**
hdisk10	**2E**00	**28**00	hdisk2
hdisk6	**26**00	**2C**00	hdisk6

You can easily obtain this mapping by using the **`lscfg -vl hdiskX | grep Serial`** command as shown in Example 10-3. The hdisk serial number is a concatenation of the storage image serial number and the ID of the volume at the storage level.

Example 10-3 The hdisk serial number in the lscfg command output

```
# lscfg -vl hdisk10 | grep Serial
        Serial Number...............75DC890 2E00
# lscfg -vl hdisk6 | grep Serial
        Serial Number...............75DC890 2600
```

Symmetrical configuration: In an actual environment (and different from this sample environment), to simplify the management of your Global Mirror environment, maintain a symmetrical configuration in terms of both physical and logical elements. With this type of configuration, you can keep the same AIX disk definitions on all nodes. It also helps you during configuration and management operations of the disk volumes within the cluster.

10.4.3 Configuring the Global Mirror relationships

In this section, you configure the Global Mirror replication relationships by performing the following tasks:

- Creating PPRC paths
- Creating Global Copy relationships
- Creating FlashCopy relationships
- Selecting an available Global Mirror session identifier
- Defining Global Mirror sessions for all involved LSSs
- Including all the source and target volumes in the Global Mirror session

Creating PPRC paths

In this task, the appropriate FC links were configured between the storage units. Example 10-4 shows the FC links that are available for the setup.

Example 10-4 Available FC links

```
dscli> lsavailpprcport -remotewwnn 5005076308FFC804 2e:28
Date/Time: October 5, 2010 5:48:09 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC890
Local Port Attached Port Type
============================
I0010      I0210         FCP
I0013      I0203         FCP
I0013      I0310         FCP
I0030      I0200         FCP
I0030      I0230         FCP
I0030      I0330         FCP
I0040      I0200         FCP
I0040      I0230         FCP
I0041      I0232         FCP
I0041      I0331         FCP
I0042      I0211         FCP
I0110      I0203         FCP
I0110      I0310         FCP
I0110      I0311         FCP
I0111      I0310         FCP
I0111      I0311         FCP
I0130      I0200         FCP
I0130      I0230         FCP
I0130      I0300         FCP
I0130      I0330         FCP
I0132      I0232         FCP
I0132      I0331         FCP
dscli>
```

To create PPRC paths:

1. Run the **lssi** command on the remote storage unit to obtain the remote wwnn parameter for the **lsavailpprcport** command. The last parameter is one possible pair of your source and target LSSs.
2. For redundancy and bandwidth, configure more FC links by using redundant SAN fabrics.

3. Among the multiple displayed links, choose two that have their ports on different adapters. Use them to create the PPRC path for the 2e:28 LSS pair (see Example 10-5).

Example 10-5 Creating pprc paths

```
dscli> mkpprcpath -remotewwnn 5005076308FFC804 -srclss 2e -tgtlss 28
I0030:I0230 I0110:I0203
Date/Time: October 5, 2010 5:55:46 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC890
CMUC00149I mkpprcpath: Remote Mirror and Copy path 2e:28 successfully
established.
dscli> lspprcpath 2e
Date/Time: October 5, 2010 5:56:13 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC890
Src Tgt State   SS   Port  Attached Port Tgt WWNN
=======================================================
2E  28  Success FF28 I0030 I0230         5005076308FFC804
2E  28  Success FF28 I0110 I0203         5005076308FFC804
dscli>
```

4. In a similar manner, configure one PPRC path for each other involved LSS pair.
5. Because the PPRC paths are unidirectional, create a second path, in the opposite direction, for each LSS pair. You use the same procedure, but work on the other storage unit (see Example 10-6). We select different FC links for this direction.

Example 10-6 Creating PPRC paths in opposite directions

```
dscli> mkpprcpath  -remotewwnn 5005076308FFC004 -srclss 28 -tgtlss 2e
I0311:I0111 I0300:I0130
Date/Time: October 5, 2010 5:57:02 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC980
CMUC00149I mkpprcpath: Remote Mirror and Copy path 28:2e successfully
established.
dscli>
```

Creating Global Copy relationships

Create Global Copy relationship between the source and target volumes and then check their status by using the commands that are shown in Example 10-7.

Example 10-7 Creating Global Copy relationships

```
dscli> mkpprc -type gcp 2e00:2800 2600:2c00
Date/Time: October 5, 2010 5:57:13 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship 2E00:2800 successfully created.
CMUC00153I mkpprc: Remote Mirror and Copy volume pair relationship 2600:2C00 successfully created.
dscli> lspprc 2e00:2800 2600:2c00
Date/Time: October 5, 2010 5:57:42 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=====================================================================================================
2600:2C00 Copy Pending -      Global Copy 26        60             Disabled      True
2E00:2800 Copy Pending -      Global Copy 2E        60             Disabled      True
dscli>
```

Creating FlashCopy relationships

Create FlashCopy relationships on both DS8000 storage units as shown in Example 10-8.

Example 10-8 Creating FlashCopy relationships

```
dscli> mkflash -tgtinhibit -nocp -record 2e00:0a08 2600:2604
Date/Time: October 5, 2010 4:17:13 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00137I mkflash: FlashCopy pair 2E00:0A08 successfully created.
CMUC00137I mkflash: FlashCopy pair 2600:2604 successfully created.
dscli> lsflash 2e00:0a08 2600:2604
Date/Time: October 5, 2010 4:17:31 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        SrcLSS SequenceNum Timeout ActiveCopy Recording Persistent Revertible
SourceWriteEnabled TargetWriteEnabled BackgroundCopy
==========================================================================================
2E00:0A08 0A    0         60      Disabled   Enabled   Enabled    Disabled   Enabled
Disabled           Disabled
2600:2604 26    0         60      Disabled   Enabled   Enabled    Disabled   Enabled
Disabled           Disabled
dscli>


dscli> mkflash -tgtinhibit -nocp -record 2800:2804 2c00:2c04
Date/Time: October 5, 2010 4:20:14 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
CMUC00137I mkflash: FlashCopy pair 2800:2804 successfully created.
CMUC00137I mkflash: FlashCopy pair 2C00:2C04 successfully created.
dscli> lsflash 2800:2804 2c00:2c04
Date/Time: October 5, 2010 4:20:38 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        SrcLSS SequenceNum Timeout ActiveCopy Recording Persistent Revertible
SourceWriteEnabled TargetWriteEnabled BackgroundCopy
==========================================================================================
2800:2804 28    0         60      Disabled   Enabled   Enabled    Disabled   Enabled
Disabled           Disabled
2C00:2C04 2C    0         60      Disabled   Enabled   Enabled    Disabled   Enabled
Disabled           Disabled
dscli>
```

Selecting an available Global Mirror session identifier

Example 10-9 lists the Global Mirror sessions that are already defined on each DS8000 storage unit. In this scenario, we chose 03 as the session identifier because it is free on both storage units.

Example 10-9 Sessions that are defined on both DS8000 storage units

```
dscli> lssession 00-ff
Date/Time: October 5, 2010 6:07:19 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
LSS ID Session Status Volume VolumeStatus PrimaryStatus        SecondaryStatus   FirstPassComplete
AllowCascading
=================================================================================================================
04     77      Normal 0400   Join Pending Primary Copy Pending Secondary Simplex True              Disable
0A     04      Normal 0A04   Join Pending Primary Suspended    Secondary Simplex False             Disable
16     05      Normal 1604   Join Pending Primary Suspended    Secondary Simplex False             Disable
16     05      Normal 1605   Join Pending Primary Suspended    Secondary Simplex False             Disable
18     02      Normal 1800   Join Pending Primary Suspended    Secondary Simplex False             Disable
1C     04      Normal 1C00   Join Pending Primary Suspended    Secondary Simplex False             Disable
1C     04      Normal 1C01   Join Pending Primary Suspended    Secondary Simplex False             Disable


dscli> lssession 00-ff
Date/Time: October 5, 2010 6:08:23 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
```

```
LSS ID Session Status Volume VolumeStatus PrimaryStatus   SecondaryStatus        FirstPassComplete
AllowCascading
=============================================================================================================
1A     20      Normal 1A00   Join Pending Primary Simplex Secondary Copy Pending True              Disable
1C     01      -      -      -            -               -                      -                 -
30     77      Normal 3000   Join Pending Primary Simplex Secondary Copy Pending True              Disable
dscli>
```

Defining Global Mirror sessions for all involved LSSs

Define the Global Mirror sessions for all the LSSs associated with source and target volumes as shown in Example 10-10. The same freely available session identifier, which is determined in “Selecting an available Global Mirror session identifier” on page 477, is used on both storage units.

Example 10-10 Defining the GM session for the source and target volumes

```
dscli> mksession -lss 2e 03
Date/Time: October 5, 2010 6:11:07 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00145I mksession: Session 03 opened successfully.
dscli> mksession -lss 26 03
Date/Time: October 5, 2010 6:11:25 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00145I mksession: Session 03 opened successfully.


dscli> mksession -lss 28 03
Date/Time: October 6, 2010 5:39:02 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
CMUC00145I mksession: Session 03 opened successfully.
dscli> mksession -lss 2c 03
Date/Time: October 6, 2010 5:39:15 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
CMUC00145I mksession: Session 03 opened successfully.
dscli>
```

Including all the source and target volumes in the Global Mirror session

Add the volumes in the Global Mirror sessions and verify their status by using the commands that are shown in Example 10-11.

Example 10-11 Adding source and target volumes to the Global Mirror sessions

```
dscli> chsession -lss 26 -action add -volume 2600 03
Date/Time: October 5, 2010 6:15:17 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00147I chsession: Session 03 successfully modified.
dscli> chsession -lss 2e -action add -volume 2e00 03
Date/Time: October 5, 2010 6:15:56 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00147I chsession: Session 03 successfully modified.
dscli> lssession 26 2e
Date/Time: October 5, 2010 6:16:21 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
LSS ID Session Status Volume VolumeStatus PrimaryStatus        SecondaryStatus   FirstPassComplete
AllowCascading
=============================================================================================================
26     03      Normal 2600   Join Pending Primary Copy Pending Secondary Simplex True              Disable
2E     03      Normal 2E00   Join Pending Primary Copy Pending Secondary Simplex True              Disable

dscli>
dscli> chsession -lss 2c -action add -volume 2c00 03
Date/Time: October 6, 2010 5:41:12 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
CMUC00147I chsession: Session 03 successfully modified.
dscli> chsession -lss 28 -action add -volume 2800 03
Date/Time: October 6, 2010 5:41:56 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
```

```
CMUC00147I chsession: Session 03 successfully modified.
dscli> lssession 28 2c
Date/Time: October 6, 2010 5:44:02 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
LSS ID Session Status Volume VolumeStatus PrimaryStatus     SecondaryStatus        FirstPassComplete
AllowCascading
=================================================================================================================
28     03      Normal 2800   Join Pending Primary Simplex Secondary Copy Pending True              Disable
2C     03      Normal 2C00   Join Pending Primary Simplex Secondary Copy Pending True              Disable
dscli>
```

10.5 Configuring AIX volume groups

In this scenario, you create a volume group and a file system on the hdisks that are associated with the DS8000 source volumes. These volumes are already identified in 10.4.2, “Identifying the source and target volumes” on page 473. They are hdisk6 and hdisk10 on the `jordan` node.

You must configure the volume groups and file systems on the cluster nodes. The application might need the same major number for the volume group on all nodes. Perform this configuration task because it might be useful later for additional configuration of the Network File System (NFS).

For the nodes on the primary site, you can use the standard procedure. You define the volume groups and file systems on one node and then import them to the other nodes. For the nodes on the secondary site, you must first suspend the replication on the involved target volumes.

10.5.1 Configuring volume groups and file systems on primary site

In this task, you create an AIX volume group on the hdisks that are associated with the DS8000 source volumes on the `jordan` node and import it on the `leeann` node:

1. Choose the next free major number on all cluster nodes by running the **`lvlstmajor`** command on each cluster node. The next common free major number on all systems is 50 as shown in Example 10-12.

Example 10-12 Running the lvlstmajor command on all cluster nodes

```
root@leeann: lvlstmajor
50...

root@robert: lvlstmajor
44..54,56...

root@jordan: # lvlstmajor
50...
```

2. Create a volume group, called `txvg`, and a file system, called /txro. These volumes are already identified in 10.4.2, “Identifying the source and target volumes” on page 473. They are hdisk6 and hdisk10 on the `jordan` node. Example 10-13 shows a list of commands to run on the `jordan` node.

Example 10-13 Creating txvg volume group on jordan

```
root@jordan: mkvg -V 50 -y txvg hdisk6 hdisk10
0516-1254 mkvg: Changing the PVID in the ODM.
txvg
root@jordan:chvg -a n xvg
root@jordan: mklv -e x -t jfs2 -y txlv txvg 250
txlv
root@jordan: mklv -e x -t jfs2log -y txloglv txvg 1
txloglv
root@jordan: crfs -v jfs2 -d /dev/txlv -a log=/dev/txloglv -m /txro -A no
File system created successfully.
1023764 kilobytes total disk space.
New File System size is 2048000
root@jordan: lsvg -p txvg
txvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk6            active            511         385
102..00..79..102..102
hdisk10           active            511         386
103..00..79..102..102
root@jordan:lspv|grep -e hdisk6 -e hdisk10
hdisk6          000a625afe2a4958                    txvg            active
hdisk10         000a624a833e440f                    txvg            active
root@jordan: varyoffvg txvg
root@jordan:
```

3. Import the volume group on the second node on the primary site, `leeann`, as shown in Example 10-14:

   a. Verify that the shared disks have the same PVID on both nodes.
   b. Run the **`rmdev -dl`** command for each hdisk.
   c. Run the **`cfgmgr`** program.
   d. Run the **`importvg`** command.

Example 10-14 Importing the txvg volume group on the leeann node

```
root@leean: rmdev -dl hdisk6
hdisk6 deleted
root@leean: rmdev -dl hdisk10
hdisk10 deleted
root@leean: cfgmgr
root@leean:lspv | grep -e hdisk6 -e hdisk10
hdisk6          000a625afe2a4958                    txvg
hdisk10         000a624a833e440f                    txvg
root@leean: importvg -V 51 -y txvg hdisk6
txvg
root@leean: lsvg -l txvg
txvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
txlv                jfs2       250     250     2    open/syncd    /txro
txloglv             jfs2log    1       1       1    open/syncd    N/A
```

```
root@leean: chvg -a n txvg
root@leean: varyoffvg txvg
```

10.5.2 Importing the volume groups in the remote site

To import the volume groups in the remote site, use the following steps. Example 10-15 shows the commands to run on the primary site.

1. Obtain a consistent replica of the data, on the primary site, by ensuring that the volume group is varied off as shown by the last command in Example 10-14.
2. Ensure that the Global Copy is in progress and that the Out of Sync count is 0.
3. Suspend the replication by using the **pausepprc** command.

Example 10-15 Pausing the Global Copy relationship on the primary site

```
dscli> lspprc -l 2600 2e00
Date/Time: October 6, 2010 3:40:56 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        Out Of Sync Tracks Tgt Read Src Cascade Tgt Cascade Date
Suspended SourceLSS Timeout (secs) Critical Mode First Pass Status Incremental Resync Tgt Write GMIR CG
PPRC CG  isTgtSE DisableAutoResync
==============================================================================================================
==============================================================================================================
2600:2C00 Copy Pending -      Global Copy 0                Disabled Disabled    Invalid     -
26        60             Disabled      True              Disabled           Disabled  N/A     Disabled
Unknown False
2E00:2800 Copy Pending -      Global Copy 0                Disabled Disabled    Invalid     -
2E        60             Disabled      True              Disabled           Disabled  N/A     Disabled
Unknown False
dscli> pausepprc 2600:2C00 2E00:2800
Date/Time: October 6, 2010 3:49:29 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00157I pausepprc: Remote Mirror and Copy volume pair 2600:2C00 relationship successfully paused.
CMUC00157I pausepprc: Remote Mirror and Copy volume pair 2E00:2800 relationship successfully paused.
dscli> lspprc -l 2600 2e00
Date/Time: October 6, 2010 3:49:41 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State     Reason      Type        Out Of Sync Tracks Tgt Read Src Cascade Tgt Cascade Date
Suspended SourceLSS Timeout (secs) Critical Mode First Pass Status Incremental Resync Tgt Write GMIR CG
PPRC CG  isTgtSE DisableAutoResync
==============================================================================================================
==============================================================================================================
2600:2C00 Suspended Host Source Global Copy 0                Disabled Disabled    Invalid     -
26        60             Disabled      True              Disabled           Disabled  N/A     Disabled
Unknown False
2E00:2800 Suspended Host Source Global Copy 0                Disabled Disabled    Invalid     -
2E        60             Disabled      True              Disabled           Disabled  N/A     Disabled
Unknown False
dscli>
```

4. To make the target volumes available to the attached hosts, use the **failoverpprc** command on the secondary site as shown in Example 10-16.

Example 10-16 The failoverpprc command on the secondary site storage unit

```
dscli> failoverpprc -type gcp 2C00:2600 2800:2E00
Date/Time: October 6, 2010 3:55:19 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
CMUC00196I failoverpprc: Remote Mirror and Copy pair 2C00:2600 successfully reversed.
CMUC00196I failoverpprc: Remote Mirror and Copy pair 2800:2E00 successfully reversed.
dscli> lspprc 2C00:2600 2800:2E00
Date/Time: October 6, 2010 3:55:35 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
```

```
ID        State     Reason      Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
==========================================================================================================
2800:2E00 Suspended Host Source Global Copy 28        60             Disabled      True
2C00:2600 Suspended Host Source Global Copy 2C        60             Disabled      True
dscli>
```

5. Refresh and check the PVIDs. Then, import and vary off the volume group as shown in Example 10-17.

Example 10-17 Importing the volume group txvg on the secondary site node, robert

```
root@robert: rmdev -dl hdisk2
hdisk2 deleted
root@robert: rmdev -dl hdisk6
hdisk6 deleted
root@robert: cfgmgr
root@robert: lspv |grep -e hdisk2 -e hdisk6
hdisk2          000a624a833e440f                    txvg
hdisk6          000a625afe2a4958                    txvg
root@robert: importvg -V 50 -y txvg hdisk2
txvg
root@robert: lsvg -l txvg
txvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
txlv                jfs2       250     250     2    closed/syncd  /txro
txloglv             jfs2log    1       1       1    closed/syncd  N/A
root@robert: varyoffvg txvg
```

6. Re-establish the Global Copy relationship as shown in Example 10-18.

Example 10-18 Re-establishing the initial Global Copy relationship

```
dscli> failbackpprc -type gcp 2600:2C00 2E00:2800
Date/Time: October 6, 2010 4:24:10 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
CMUC00197I failbackpprc: Remote Mirror and Copy pair 2600:2C00 successfully failed back.
CMUC00197I failbackpprc: Remote Mirror and Copy pair 2E00:2800 successfully failed back.
dscli> lspprc 2600:2C00 2E00:2800
Date/Time: October 6, 2010 4:24:41 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
==========================================================================================================
2600:2C00 Copy Pending -      Global Copy 26        60             Disabled      True
2E00:2800 Copy Pending -      Global Copy 2E        60             Disabled      True


dscli> lspprc 2800 2c00
Date/Time: October 6, 2010 4:24:57 AM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State               Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
==========================================================================================================
2600:2C00 Target Copy Pending -      Global Copy 26        unknown        Disabled      Invalid
2E00:2800 Target Copy Pending -      Global Copy 2E        unknown        Disabled      Invalid
dscli>
```

10.6 Configuring the cluster

To configure the cluster, you must complete all software prerequisites. Also, you must configure the `/etc/hosts` file properly, and verify that the `clcomdES` subsystem is running on each node.

To configure the cluster:

1. Add a cluster.
2. Add all three nodes.
3. Add both sites.
4. Add the XD_ip network.
5. Add the disk heartbeat network.
6. Add the base interfaces to XD_ip network.
7. Add the service IP address.
8. Add the DS8000 Global Mirror replicated resources.
9. Add a resource group.
10. Add a service IP, application server, volume group, and DS8000 Global Mirror Replicated Resource to the resource group.

10.6.1 Configuring the cluster topology

Configuring a cluster entails the following tasks:

- Adding a cluster
- Adding nodes
- Adding sites
- Adding networks
- Adding communication interfaces

Adding a cluster

To add a cluster:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select **Extended Configuration** → **Extended Topology Configuration** → **Configure an HACMP Cluster** → **Add/Change/Show an HACMP Cluster**.
3. Enter the cluster name, which is `Txrmnia` in this scenario, as shown in Figure 10-3. Press Enter.

```
                    Add/Change/Show an HACMP Cluster

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

                                                      [Entry Fields]
* Cluster Name                                        [Txrmnia]
```

Figure 10-3 Adding a cluster in the SMIT menu

The output is displayed in the SMIT Command Status window.

Adding nodes

To add the nodes:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration → Extended Topology Configuration → Configure HACMP Nodes → Add a Node to the HACMP Cluster**.
3. Enter the desired node name, which is `jordan` in this case, as shown in Figure 10-4. Press Enter. The SMIT Command Status window shows the output.

```
                        Add a Node to the HACMP Cluster

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

                                                      [Entry Fields]
* Node Name                                           [jordan]
  Communication Path to Node                          []                +
```

Figure 10-4 Add a Node SMIT menu

4. In this scenario, repeat these steps two more times to add the additional nodes of `leeann` and `robert`.

Adding sites

To add the nodes:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration → Extended Topology Configuration → Configure HACMP Sites → Add a Site**.
3. Enter the desired site name, which in this scenario is the Texas site with the nodes `jordan` and `leeann`, as shown in Figure 10-5. Press Enter. The SMIT Command Status window shows the output.

```
                                  Add a Site

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

                                                    [Entry Fields]
* Site Name                                        [Texas]              +
* Site Nodes                                        jordan leeann       +
```

Figure 10-5 Add a Site SMIT menu

4. In this scenario, repeat these steps to add the Romania site with the `robert` node.

Example 10-19 shows the site definitions. The dominance information is displayed, but not relevant until a resource group is defined later by using the nodes.

Example 10-19 cllssite information about site definitions

```
./cllssite
----------------------------------------------------
Sitename            Site Nodes          Dominance          Protection Type
----------------------------------------------------
Texas               jordan leeann                          NONE
Romania             robert                                 NONE
```

Adding networks

To add the nodes:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration → Extended Topology Configuration → Configure HACMP Networks → Add a Network to the HACMP Cluster**.
3. Choose the network type, which in this scenario is `XD_ip`.
4. Keep the default network name and press Enter (Figure 10-6).

```
                Add an IP-Based Network to the HACMP Cluster

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

                                                      [Entry Fields]
* Network Name                                       [net_XD_ip_01]
* Network Type                                        XD_ip
* Netmask(IPv4)/Prefix Length(IPv6)                  [255.255.255.0]
* Enable IP Address Takeover via IP Aliases          [Yes]                    +
  IP Address Offset for Heartbeating over IP Aliases []
```

Figure 10-6 Add an IP-Based Network SMIT menu

5. Repeat these steps but select a network type of `diskhb` for the disk heartbeat network and keep the default network name of `net_diskhb_01`.

Adding communication interfaces

To add the nodes:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration → Extended Topology Configuration → Configure HACMP Communication Interfaces/Devices → Add Communication Interfaces/Devices → Add Pre-defined Communication Interfaces and Devices → Communication Interfaces**.
3. Select the previously created network, which in this scenario is `net_XD_ip_01`.

4. Complete the SMIT menu fields. The first interface in this scenario is for `jordan` is shown in Figure 10-7. Press Enter. The SMIT Command Status window shows the output.

```
                        Add a Communication Interface

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

                                                      [Entry Fields]
* IP Label/Address                                   [jordan_base]         +
* Network Type                                        XD_ip
* Network Name                                        net_XD_ip_01
* Node Name                                          [jordan]              +
```

Figure 10-7 Add communication interface SMIT menu

5. Repeat these steps, and select **Communication Devices** to complete the disk heartbeat network.

The topology is now configured. Also, you can see all the interfaces and devices from the **`cllsif`** command output that is shown in Figure 10-8.

```
Adapter          Type       Network    Net Type   Attribute  Node       IP Address
jordan_base      boot       net_XD_ip_01 XD_ip      public     jordan     9.3.207.209
jordandhb        service    net_diskhb_01 diskhb     serial     jordan     /dev/hdisk8
leeann_base      boot       net_XD_ip_01 XD_ip      public     leeann     9.3.207.208
leeanndhb        service    net_diskhb_01 diskhb     serial     leeann     /dev/hdisk8
robert_base      boot       net_XD_ip_01 XD_ip      public     robert     9.3.207.207
```

Figure 10-8 Cluster interfaces and devices defined

10.6.2 Configuring cluster resources and resource group

The test scenario has only one resource group, which contains the resources of the service IP address, volume group, and DS8000 replicated resources. Configure the cluster resources and resource group as explained in the following sections.

Defining the service IP

Define the service IP by following these steps:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure HACMP Service IP Labels/Addresses → Add a Service IP Label/Address → Configurable on Multiple Nodes**.
3. Choose the `net_XD_ip_01` network and press Enter.
4. Choose the appropriate IP label or address. Press Enter. The SMIT Command Status window shows the output.

In this scenario, we added `serviceip_2`, as shown in Figure 10-9.

```
    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.

                                                      [Entry Fields]
* IP Label/Address                                    serviceip_2           +
  Netmask(IPv4)/Prefix Length(IPv6)                   []
* Network Name                                        net_XD_ip_01
  Alternate HW Address to accompany IP Label/Address  []
  Associated Site                                     ignore
```

Figure 10-9 Add a Service IP Label SMIT menu

In most true site scenarios, where each site is on different segments, it is common to create at least two service IP labels. You create one for each site by using the **Associated Site** option, which indicates the desire to have site-specific service IP labels. With this option, you can have a unique service IP label at each site. However, we do not use them in this test because we are on the same network segment.

Defining the DS8000 Global Mirror resources

To fully define the Global Mirror resources, follow these steps:

1. Add a storage agent or agents.
2. Add a storage system or systems.
3. Add a mirror group or groups.

Because these options are all new, define each one before you configure them:

Storage agent A generic name that is given by PowerHA SystemMirror for an entity such as the IBM DS8000 HMC. Storage agents typically provide a one-point coordination point and often use TCP/IP as their transport for communication. You must provide the IP address and authentication information that will be used to communicate with the HMC.

Storage system A generic name that is given by PowerHA SystemMirror for an entity such as a DS8700 Storage Unit. When you use Global Mirror, you must associate one storage agent with each storage system. You must provide the IBM DS8700 system identifier for the storage system. For example, `IBM.2107-75ABTV1` is a storage identifier for a DS8000 Storage System.

Mirror group A generic name that is given by PowerHA SystemMirror for a logical collection of volumes that must be mirrored to another storage system that are on a remote site. A Global Mirror session represents a mirror group.

Adding a storage agent

To add a storage agent, follow these steps:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **Extended Configuration** → **Extended Resource Configuration** → **HACMP Extended Resources Configuration** → **Configure DS8000 Global Mirror Resources** → **Configure Storage Agents** → **Add a Storage Agent**.

3. Complete the menu appropriately and press Enter. Figure 10-10 shows the configuration for this scenario. The SMIT Command Status window shows the output.

```
                              Add a Storage Agent

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

                                                      [Entry Fields]
* Storage Agent Name                                 [ds8khmc]
* IP Addresses                                       [9.3.207.122]
* User ID                                            [redbook]
* Password                                           [r3dbook]
```

Figure 10-10 Add a Storage Agent SMIT menu

It is possible to have multiple storage agents. However, this test scenario has only one storage agent that manages both storage units.

Important: The user ID and password are stored as flat text in the `HACMPxd_storage_agent.odm` file.

Adding a storage system

To add the storage systems, follow these steps:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure DS8000 Global Mirror Resources → Configure Storage Systems → Add a Storage System**.
3. Complete the menu appropriately and press Enter. Figure 10-11 shows the configuration for this scenario. The SMIT Command Status window shows the output.

```
                              Add a Storage System

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

                                                      [Entry Fields]
* Storage System Name                                [texasds8k]
* Storage Agent Name(s)                               ds8kmainhmc          +
* Site Association                                    Texas                +
* Vendor Specific Identification                     [IBM.2107-75DC890]    +
* WWNN                                               [5005076308FFC004]    +
```

Figure 10-11 Add a Storage System SMIT menu

4. Repeat these steps for the storage system at Romania site, and name it `romaniads8k`. Example 10-20 shows the configuration.

Example 10-20 Storage systems definitions

```
Storage System Name                          texasds8k
Storage Agent Name(s)                        ds8kmainhmc
Site Association                             Texas
Vendor Specific Identification               IBM.2107-75DC890
WWNN                                         5005076308FFC004

Storage System Name                          romaniads8k
Storage Agent Name(s)                        ds8kmainhmc
Site Association                             Romania
Vendor Specific Identification               IBM.2107-75DC980
WWNN                                         5005076308FFC804
```

Adding a mirror group

You are now ready to add the storage systems. To add a storage system:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Configuration → Configure DS8000 Global Mirror Resources → Configure Mirror Groups → Add a Mirror Group**.
3. Complete the menu appropriately and press Enter. Figure 10-12 show the configuration for this scenario. The SMIT Command Status window shows the output.

```
                              Add a Mirror Group

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

                                                    [Entry Fields]
* Mirror Group Name                                [texasmg]
* Storage System Name                               texasds8k romaniads8k  +
* Vendor Specific Identifier                       [03]       +
* Recovery Action                                   automatic              +
  Maximum Coordination Time                        [50]
  Maximum Drain Time                               [30]
  Consistency Group Interval Time                  [0]
```

Figure 10-12 Add a Mirror Group SMIT menu

Vendor Specific Identifier field: For the Vendor Specific Identifier field, provide only the Global Mirror session number.

Defining a resource group and Global Mirror resources

Now that you have all the components configured that are required for the DS8700 replicated resource, you can create a resource group and add your resources to it.

Adding a resource group

To add a resource group:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration → Extended Resource Configuration → HACMP Extended Resources Group Configuration → Add a Resource Group**.
3. Complete the menu appropriately and press Enter. Figure 10-13 shows the configuration in this scenario. Notice that for the Inter-Site Management Policy, we chose **Prefer Primary Site**. This option ensures that resource group starts automatically when the cluster is started in the primary Texas site. The SMIT Command Status window shows the output.

```
                          Add a Resource Group (extended)

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

                                                  [Entry Fields]
* Resource Group Name                            [ds8kgmrg]

  Inter-Site Management Policy                   [Prefer Primary Site]    +
* Participating Nodes from Primary Site          [jordan leeann]          +
  Participating Nodes from Secondary Site        [robert]                 +

  Startup Policy                                  Online On Home Node Only+
  Fallover Policy                                 Fallover To Next Priority Node > +
  Fallback Policy                                 Never Fallback
```

Figure 10-13 Add a Resource Group SMIT menu

Adding resources to a resource group

To add resources to a resource group:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration → Extended Resource Configuration → Change/Show Resources and Attributes for a Resource Group**.
3. Choose the resource group, which in this example is `ds8kgmrg`.
4. Complete the menu appropriately and press Enter. Figure 10-13 shows the configuration for this scenario. The SMIT Command Status window shows the output.

In this scenario, we added only a service IP label, the volume group, and the DS8000 Global Mirror Replicated Resources as shown in the streamlined **clshowres** command output in Example 10-21.

Volume group: The volume group names must be listed in the same order as the DS8700 mirror group names in the resource group.

Example 10-21 Resource group attributes and resources

```
Resource Group Name                             ds8kgmrg
Inter-site Management Policy                    Prefer Primary Site
Participating Nodes from Primary Site           jordan leeann
Participating Nodes from Secondary Site         robert
Startup Policy                                  Online On Home Node Only
```

```
Fallover Policy                                 Fallover To Next Priority Node
Fallback Policy                                 Never Fallback
Service IP Label                                serviceip_2
Volume Groups                                   txvg                    +
GENXD Replicated Resources                      texasmg                 +
```

DS8000 Global Mirror Replicated Resources field: In the SMIT menu for adding resources to the resource group, notice that the appropriate field is named *DS8000 Global Mirror Replicated Resources*. However, when you view the menu by using the **clshowres** command (Example 10-21 on page 490), the field is called *GENXD Replicated Resources*.

You can now synchronize the cluster, start the cluster, and begin testing it.

10.7 Failover testing

This section takes you through basic failover testing scenarios with the DS8000 Global Mirror replicated resources locally within the site and across sites. You must carefully plan the testing of a site cluster failover because more time is required to manipulate the secondary target LUNs at the recovery site. Also when you test the asynchronous replication, because of the nature of the asynchronous replication, it can also impact the data.

In these scenarios, redundancy tests, such as on IP networks that have only a single network, cannot be performed. Instead, you must configure redundant IP or non-IP communication paths to avoid isolation of the sites. The loss of all the communication paths between sites leads to a partitioned state of the cluster. Such a loss also leads to data divergence between sites if the replication links are also unavailable.

Another specific failure scenario is the loss of replication paths between the storage subsystems while the cluster is running on both sites. To avoid this type of loss, configure a redundant PPRC path or links for the replication. You must manually recover the status of the pairs after the storage links are operational again.

Important: If the PPRC path or link between Global Mirror volumes breaks down, the PowerHA Enterprise Edition is unaware. The reason is that PowerHA does not process SNMP for volumes that use DS8700 Global Mirror technology for mirroring. In such a case, you must identify and correct the PPRC path failure. Depending upon some timing conditions, such an event can result in the corresponding Global Mirror session that is going into an unrecoverable state. In this situation, you must manually stop and restart the corresponding Global Mirror session (by using the **rmgmir** and **mkgmir** DSCLI commands) or an equivalent DS8700 interface.

This topic takes you through the following tests:

- Graceful site failover
- Rolling site failure
- Site reintegration

Each test, other than the reintegration test, begins in the same initial state of the primary site that hosts the `ds8kgmrg` resource group on the primary node as shown in Example 10-22 on page 492. Before each test, we start copying data from another file system to the replicated file systems. After each test, we verify that the service IP address is online and that new data is in the file systems. We also had a script that inserted the current time and date, along with the local node name, into a file on each file system.

Example 10-22 Beginning of the test cluster resource group states

```
jordan# clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
ds8kgmrg       ONLINE                         jordan@Texas
               OFFLINE                        leeann@Texas
               ONLINE SECONDARY               robert@Romania
```

After each test, we show the Global Mirror states. Example 10-23 shows the normal running production status of the Global Mirror pairs from each site.

Example 10-23 Beginning states of the Global Mirror pairs

```
********************From node jordan at site Texas*****************************

dscli> lssession 26 2E
Date/Time: October 10, 2010 4:00:04 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
LSS ID Session Status         Volume VolumeStatus PrimaryStatus        SecondaryStatus   FirstPassComplete
AllowCascading
=================================================================================================================
==============
26     03      CG In Progress 2600   Active       Primary Copy Pending Secondary Simplex True
Disable
2E     03      CG In Progress 2E00   Active       Primary Copy Pending Secondary Simplex True
Disable

dscli> lspprc 2600 2E00
Date/Time: October 10, 2010 4:00:43 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================
2600:2C00 Copy Pending -      Global Copy 26        60             Disabled      True
2E00:2800 Copy Pending -      Global Copy 2E        60             Disabled      True

********************From remote node robert at site Romania*****************************

dscli> lssession 28 2c
Date/Time: October 10, 2010 3:54:58 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
LSS ID Session Status Volume VolumeStatus PrimaryStatus   SecondaryStatus        FirstPassComplete
AllowCascading
=================================================================================================================
======
28     03      Normal 2800   Join Pending Primary Simplex Secondary Copy Pending True              Disable
2C     03      Normal 2C00   Join Pending Primary Simplex Secondary Copy Pending True              Disable

dscli> lspprc 2800 2c00
Date/Time: October 10, 2010 3:55:48 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State               Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================
2600:2C00 Target Copy Pending -      Global Copy 26        unknown        Disabled      Invalid
2E00:2800 Target Copy Pending -      Global Copy 2E        unknown        Disabled      Invalid
```

10.7.1 Graceful site failover

Performing a controlled move of a production environment across sites is a basic test to ensure that the remote site can bring the production environment online. This test is done only during initial implementation testing or during a planned production outage of the site. In this test, we perform the graceful failover operation between sites by performing a resource group move.

In a true maintenance scenario, you might most likely perform a graceful site failover by stopping the cluster on the local standby node first. Then, you stop the cluster on the production node by using *Move Resource Group*.

Moving the resource group to another site: In this scenario, because we have only one node at the Romania site, we use the option to move the resource group to another site. If multiple remote nodes are members of the resource, use the option to move the resource group to another node instead.

During this move, the following operations are performed:

- Release the primary online instance of `ds8kgmrg` at the Texas site. This operation entails the following tasks:
  - Runs the application server stop.
  - Unmounts the file systems.
  - Varies off the volume group.
  - Removes the service IP address.
- Release the secondary online instance of `ds8kgmrg` at the Romania site.
- Acquire `ds8kgmrg` in the secondary online state at the Texas site.
- Acquire `ds8kgmrg` in the online primary state at the Romania site.

To move the resource group by using SMIT:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **System Management (C-SPOC) → Resource Groups and Applications → Move a Resource Group to Another Node / Site → Move Resource Groups to Another Site**.

3. Select the **ONLINE** instance of `ds8kgmrg` to be moved as shown in Figure 10-14.

```
                    Move a Resource Group to Another Node / Site

Move cursor to desired item and press Enter.

  Move Resource Groups to Another Node
  Move +--------------------------------------------------------------------------+
       |                          Select a Resource Group                         |
       |                                                                          |
       | Move cursor to desired item and press Enter. Use arrow keys to scroll.   |
       |                                                                          |
       |   #                                                                      |
       |   # Resource Group                  State                Node(s) / Site  |
       |   #                                                                      |
       |   ds8kgmrg                          ONLINE               jordan / Texas  |
       |   ds8kgmrg                          ONLINE SECONDARY     robert / Romani |
       |                                                                          |
       |   #                                                                      |
       |   # Resource groups in node or site collocation configuration:           |
       |   # Resource Group(s)                              State    Node / Site  |
       |   #                                                                      |
       |                                                                          |
       | F1=Help                F2=Refresh             F3=Cancel                  |
       | F8=Image               F10=Exit               Enter=Do                   |
F1=Help| /=Find                 n=Find Next                                       |
F9=Shel+--------------------------------------------------------------------------+
```

Figure 10-14 Selecting a resource group

4. Select the **Romania** site from the next menu as shown in Figure 10-15.

```
+--------------------------------------------------------------------------+
|                        Select a Destination Site                         |
|                                                                          |
| Move cursor to desired item and press Enter.                             |
|                                                                          |
|   # *Denotes Originally Configured Primary Site                          |
|   Romania                                                                |
|                                                                          |
| F1=Help                 F2=Refresh              F3=Cancel                |
| F8=Image                F10=Exit                Enter=Do                 |
| /=Find                  n=Find Next                                      |
+--------------------------------------------------------------------------+
```

Figure 10-15 Selecting a site for a resource group move

5. Verify the information in the final menu and press Enter.

Upon completion of the move, `ds8kgmrg` is online on the node `robert` as shown Example 10-24.

Attention: During our testing, a problem was encountered. After we performed the first resource group move between sites, we were unable to move it back because the pick list for destination site is empty. We could move it back by node. Later in our testing, the by-site option started working. However, it moved the resource group to the standby node at the primary site instead of the original primary node. If you encounter similar problems, contact IBM support.

Example 10-24 Resource group status after the site move to Romania

```
-------------------------------------------------------------------------
Group Name     State                        Node
-------------------------------------------------------------------------
ds8kgmrg       ONLINE SECONDARY             jordan@Texas
               OFFLINE                      leeann@Texas
               ONLINE                       robert@Romania
```

6. Repeat the resource group move to move it back to its original primary site, Texas, and node, `jordan`, to return to the original starting state. However, instead of using the option to move it another site, use the option to move it to another node.

Example 10-25 shows that the Global Mirror statuses are now swapped, and the local site is showing the LUNs now as the target volumes.

Example 10-25 Global Mirror status after the resource group move

```
******************From node jordan at site Texas***************************

dscli> lssession 26 2E
Date/Time: October 10, 2010 4:04:44 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
LSS ID Session Status Volume VolumeStatus PrimaryStatus   SecondaryStatus        FirstPassComplete
AllowCascading
=================================================================================================================
======
26     03      Normal 2600   Active       Primary Simplex Secondary Copy Pending True              Disable
2E     03      Normal 2E00   Active       Primary Simplex Secondary Copy Pending True              Disable

dscli> lspprc 2600 2E00
Date/Time: October 10, 2010 4:05:26 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State               Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================================
2800:2E00 Target Copy Pending -      Global Copy 28        unknown        Disabled      Invalid
2C00:2600 Target Copy Pending -      Global Copy 2C        unknown        Disabled      Invalid

******************From remote node robert at site Romania***************************

dscli> lssession 28 2C
Date/Time: October 10, 2010 3:59:25 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
LSS ID Session Status         Volume VolumeStatus PrimaryStatus        SecondaryStatus   FirstPassComplete
AllowCascading
=================================================================================================================
==============
28     03      CG In Progress 2800   Active       Primary Copy Pending Secondary Simplex True
Disable
2C     03      CG In Progress 2C00   Active       Primary Copy Pending Secondary Simplex True
Disable
```

```
dscli> lspprc 2800 2C00
Date/Time: October 10, 2010 3:59:35 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================
2800:2E00 Copy Pending -      Global Copy 28        60             Disabled      True
2C00:2600 Copy Pending -      Global Copy 2C        60             Disabled      True
```

10.7.2 Rolling site failure

This scenario entails performing a rolling site failure of the Texas site by using the following steps:

1. Halt the primary production node jordan at the Texas site.
2. Verify that the resource group ds8kgmrg is acquired locally by the node leeann.
3. Verify that the Global Mirror pairs are in the same status as before the system failure.
4. Halt the node leeann to produce a site down.
5. Verify that the resource group ds8kgmrg is acquired remotely by the robert node.
6. Verify that the Global Mirror pair states are changed.

Begin with all three nodes active in the cluster and the resource group online on the primary node as shown in Example 10-22 on page 492.

On the node jordan, we run the **reboot -q** command. The node leeann acquires the ds8kgmrg resource group as shown in Example 10-26.

Example 10-26 Local node failover within the site Texas

```
root@leeann: clRGinfo
-----------------------------------------------------------------------------
Group Name     State                         Node
-----------------------------------------------------------------------------
ds8kgmrg       OFFLINE                     jordan@Texas
               ONLINE                      leeann@Texas
               ONLINE SECONDARY            robert@Romania
```

Example 10-27 shows that the statuses are the same as when we started.

Example 10-27 Global Mirror pair status after a local failover

```
********************From node leeann at site Texas*****************************

dscli> lssession 26 2E
Date/Time: October 10, 2010 4:10:04 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
LSS ID Session Status         Volume VolumeStatus PrimaryStatus        SecondaryStatus   FirstPassComplete
AllowCascading
=================================================================================================================
==============
26     03      CG In Progress 2600   Active       Primary Copy Pending Secondary Simplex True
Disable
2E     03      CG In Progress 2E00   Active       Primary Copy Pending Secondary Simplex True
Disable

dscli> lspprc 2600 2E00
Date/Time: October 10, 2010 4:10:43 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================
2600:2C00 Copy Pending -      Global Copy 26        60             Disabled      True
2E00:2800 Copy Pending -      Global Copy 2E        60             Disabled      True
```

```
******************From remote node robert at site Romania**************************

dscli> lssession 28 2c
Date/Time: October 10, 2010 4:04:58 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
LSS ID Session Status Volume VolumeStatus PrimaryStatus   SecondaryStatus        FirstPassComplete
AllowCascading
=================================================================================================================
28     03      Normal 2800   Join Pending Primary Simplex Secondary Copy Pending True              Disable
2C     03      Normal 2C00   Join Pending Primary Simplex Secondary Copy Pending True              Disable

dscli> lspprc 2800 2c00
Date/Time: October 10, 2010 4:05:48 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State              Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================================
2600:2C00 Target Copy Pending -      Global Copy 26        unknown        Disabled      Invalid
2E00:2800 Target Copy Pending -      Global Copy 2E        unknown        Disabled      Invalid
```

Upon the cluster stabilization, we run the **reboot -q** command on the `leeann` node to start a `site_down` event. The `robert` node at the Romania site acquires the `ds8kgmrg` resource group as shown in Example 10-28.

Example 10-28 Hard failover between sites

```
root@robert: clRGinfo
-----------------------------------------------------------------------------
Group Name     State                          Node
-----------------------------------------------------------------------------
ds8kgmrg       OFFLINE                       jordan@Texas
               OFFLINE                       leeann@Texas
               ONLINE                        robert@Romania
```

You can also see that the replicated pairs are now in the *suspended* state at the remote site as shown in Example 10-29.

Example 10-29 Global Mirror pair status after site failover

```
******************From remote node robert at site Romania**************************
dscli> lssession 28 2c
Date/Time: October 10, 2010 4:17:28 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
LSS ID Session Status Volume VolumeStatus PrimaryStatus     SecondaryStatus   FirstPassComplete
AllowCascading
=================================================================================================================
28     03      Normal 2800   Join Pending Primary Suspended Secondary Simplex False             Disable
2C     03      Normal 2C00   Join Pending Primary Suspended Secondary Simplex False             Disable

dscli> lspprc 2800 2c00
Date/Time: October 10, 2010 4:17:55 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State     Reason      Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=================================================================================================================
2800:2E00 Suspended Host Source Global Copy 28        60             Disabled      False
2C00:2600 Suspended Host Source Global Copy 2C        60             Disabled      False
```

Important: Although the testing resulted in a `site_down` event, we never lost access to the primary storage subsystem. PowerHA does not check storage connectivity back to the primary site during this event. Before you move back to the primary site, re-establish the replicated pairs and get them all back in sync. If you replace the storage, you might also have to change the storage agent, storage subsystem, and mirror groups to ensure that the new configuration is correct for the cluster.

10.7.3 Site reintegration

Before you bring the primary site node back into the cluster, you must place the Global Mirror pairs back in sync:

Tip: Follow these steps “as is” because you can accomplish the same results by using various methods.

1. Verify that the Global Mirror statuses at the primary site are *suspended*.
2. Fail back PPRC from the secondary site.
3. Verify that the Global Mirror status at the primary site shows the target status.
4. Verify that out-of-sync tracks are 0.
5. Stop the cluster to ensure that the volume group I/O is stopped.
6. Fail over the PPRC on the primary site.
7. Fail back the PPRC on the primary site.
8. Start the cluster.

Failing back the PPRC pairs to the secondary site

To fail back the PPRC pairs to the secondary site:

1. Verify the current state of the Global Mirror pairs at the primary site from the `jordan` node. The pairs are suspended as shown in Example 10-30.

Example 10-30 Suspended pair status in Global Mirror on the primary site after node restart

```
******************From node jordan at site Texas**************************
dscli> lspprc 2600 2e00
Date/Time: October 10, 2010 4:27:48 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State     Reason      Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=============================================================================================
2600:2C00 Suspended Host Source Global Copy 26        60             Disabled      True
2E00:2800 Suspended Host Source Global Copy 2E        60             Disabled      True
```

2. On the remote node `robert`, fail back the PPRC pairs as shown in Example 10-31.

Example 10-31 Failing back PPRC pairs at the remote site

```
******************From node robert at site Romania**************************
dscli> failbackpprc -type gcp 2C00:2600 2800:2E00
Date/Time: October 10, 2010 4:22:09 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC980
CMUC00197I failbackpprc: Remote Mirror and Copy pair 2C00:2600 successfully failed back.
CMUC00197I failbackpprc: Remote Mirror and Copy pair 2800:2E00 successfully
```

3. After you run the fallback, check the status again of the pairs from the primary site to ensure that they are now shown as *Target* (Example 10-32).

Example 10-32 Verifying that the primary site LUNs are now target LUNs

```
******************From node jordan at site Texas**************************
dscli> lspprc 2600 2e00
Date/Time: October 10, 2010 4:44:21 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State               Reason Type        SourceLSS Timeout (secs) Critical Mode First
Pass Status
=============================================================================================
=========
2800:2E00 Target Copy Pending -      Global Copy 28        unknown        Disabled      Invalid
2C00:2600 Target Copy Pending -      Global Copy 2C        unknown        Disabled      Invalid
```

4. Monitor that the status of replication at the remote site by watching the *Out of Sync Tracks* field by using the **lspprc -l** command. After they are at *0*, as shown in Example 10-33, they are in sync. Then, you can stop the remote site in preparation to move production back to the primary site.

Example 10-33 Verifying that the Global Mirror pairs are back in sync

```
dscli> lspprc -l 2800 2c00
Date/Time: October 10, 2010 4:22:46 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State        Reason Type        Out Of Sync Tracks Tgt Read Src Cascade Tgt Cascade Date
Suspended SourceLSS
=============================================================================================================
============
2800:2E00 Copy Pending -      Global Copy 0                  Disabled Disabled    Invalid     -
28
2C00:2600 Copy Pending -      Global Copy 0                  Disabled Disabled    Invalid     -
2C        6
```

Failing over the PPRC pairs back to the primary site

To fail over the PPRC pairs back to the primary site:

1. Stop the cluster on node `robert` by using the **smitty clstop** command to bring the resource group down.
2. After the resources are offline, continue to fail over the PPRC on the primary site `jordan` node as shown Example 10-34.

Example 10-34 Failover PPRC pairs at local primary site

```
******************From node jordan at site Texas**************************
dscli> failoverpprc -type gcp  2600:2c00 2E00:2800
Date/Time: October 10, 2010 4:45:16 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC890
CMUC00196I failoverpprc: Remote Mirror and Copy pair 2600:2C00 successfully
reversed.
CMUC00196I failoverpprc: Remote Mirror and Copy pair 2E00:2800 successfully
reversed.
```

3. Again verify that the status is in the *suspended* state on the primary site and that the remote site shows the *copy* state as shown in Example 10-35.

Example 10-35 Global Mirror pairs that are suspended on the primary site

```
******************From node jordan at site Texas**************************
dscli> lspprc 2600 2E00
Date/Time: October 10, 2010 4:45:51 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State     Reason      Type        SourceLSS Timeout (secs) Critical Mode First Pass
Status
===============================================================================================
====
2600:2C00 Suspended Host Source Global Copy 26        60             Disabled      True
2E00:2800 Suspended Host Source Global Copy 2E        60             Disabled      True

******************From node robert at site Romania**************************
dscli> lspprc 2800 2c00
Date/Time: October 10, 2010 4:39:27 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass
Status
===============================================================================================
==
2800:2E00 Copy Pending -      Global Copy 28        60             Disabled      True
2C00:2600 Copy Pending -      Global Copy 2C        60             Disabled      True
```

Failing back the PPRC pairs to the primary site

You cannot complete the switchback to the primary site by failing back the Global Mirror pairs to the primary site by running the **failbackpprc** command as shown in Example 10-36.

Example 10-36 Failing back the PPRC pairs on the primary site

```
******************From node jordan at site Texas**************************
dscli> failbackpprc -type gcp  2600:2c00 2E00:2800
Date/Time: October 10, 2010 4:46:49 PM CDT IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75DC890
CMUC00197I failbackpprc: Remote Mirror and Copy pair 2600:2C00 successfully failed back.
CMUC00197I failbackpprc: Remote Mirror and Copy pair 2E00:2800 successfully failed back.
```

Verify the status of the pairs at each site as shown in Example 10-37.

Example 10-37 Global Mirror pairs failed back to the primary site

```
******************From node jordan at site Texas**************************
dscli> lspprc 2600 2e00
Date/Time: October 10, 2010 4:47:04 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
===============================================================================================
2600:2C00 Copy Pending -      Global Copy 26        60             Disabled      True
2E00:2800 Copy Pending -      Global Copy 2E        60             Disabled      True

******************From node robert at site Romania**************************
dscli> lspprc 2800 2c00
Date/Time: October 10, 2010 4:40:44 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State               Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=======================================================================================================
2600:2C00 Target Copy Pending -      Global Copy 26        unknown        Disabled      Invalid
2E00:2800 Target Copy Pending -      Global Copy 2E        unknown        Disabled      Invalid
```

Starting the cluster

To start the cluster:

1. Start all nodes in the cluster by using the **smitty clstart** command as shown Figure 10-16.

```
                              Start Cluster Services

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

                                                      [Entry Fields]
* Start now, on system restart or both                now                      +
  Start Cluster Services on these nodes               [jordan,leeann,robert]   +
* Manage Resource Groups                              Automatically            +
  BROADCAST message at startup?                       true                     +
  Startup Cluster Information Daemon?                 true                     +
  Ignore verification errors?                         false                    +
  Automatically correct errors found during           Interactively            +
  cluster start?
```

Figure 10-16 Restarting a cluster after a site failure

Upon startup of the primary node jordan, the resource group is automatically started on jordan and back to the original starting point as shown in Example 10-38.

Example 10-38 Resource group status after restart

```
-------------------------------------------------------------------------------
Group Name     State                          Node
-------------------------------------------------------------------------------
ds8kgmrg       ONLINE                         jordan@Texas
               OFFLINE                        leeann@Texas
               ONLINE SECONDARY               robert@Romania
```

2. Verify the pair and session status on each site as shown in Example 10-39.

Example 10-39 Global Mirror pairs back to normal

```
******************From node jordan at site Texas*****************************
dscli>lssession 26 2e
Date/Time: October 10, 2010 5:02:11 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
LSS ID Session Status         Volume VolumeStatus PrimaryStatus        SecondaryStatus   FirstPassComplete
AllowCascading
=================================================================================================================
==============
26     03      CG In Progress 2600   Active       Primary Copy Pending Secondary Simplex True
Disable
2E     03      CG In Progress 2E00   Active       Primary Copy Pending Secondary Simplex True
Disable


dscli> lspprc 2600 2e00
Date/Time: October 10, 2010 5:02:26 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC890
ID        State        Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=============================================================================================
2600:2C00 Copy Pending -      Global Copy 26        60             Disabled      True
2E00:2800 Copy Pending -      Global Copy 2E        60             Disabled      True
```

```
******************From node robert at site Romania**************************
dscli>lssession 28 2C
Date/Time: October 10, 2010 4:56:11 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
LSS ID Session Status Volume VolumeStatus PrimaryStatus   SecondaryStatus        FirstPassComplete
AllowCascading
=============================================================================================================
======
28     03      Normal 2800   Active       Primary Simplex Secondary Copy Pending True              Disable
2C     03      Normal 2C00   Active       Primary Simplex Secondary Copy Pending True              Disable

dscli> lspprc 2800 2c00
Date/Time: October 10, 2010 4:56:30 PM CDT IBM DSCLI Version: 6.5.15.19 DS: IBM.2107-75DC980
ID        State               Reason Type        SourceLSS Timeout (secs) Critical Mode First Pass Status
=============================================================================================================
2600:2C00 Target Copy Pending -      Global Copy 26        unknown        Disabled      Invalid
2E00:2800 Target Copy Pending -      Global Copy 2E        unknown        Disabled      Invalid
```

10.8 LVM administration of DS8000 Global Mirror replicated resources

This section provides the common scenarios for adding more storage to an existing Global Mirror replicated environment. These scenarios work primarily with the Texas site and the `ds8kgmrg` resource group. You perform the following tasks:

- Adding a Global Mirror pair to an existing volume group
- Adding a Global Mirror pair into a new volume group

Dynamically expanding a volume: This section does not provide information about dynamically expanding a volume because this option is not supported.

10.8.1 Adding a Global Mirror pair to an existing volume group

To add a Global Mirror pair to an existing volume group:

1. Assign a new LUN to each site, add the FlashCopy devices, and add the new pair into the existing session as explained in 10.4.3, “Configuring the Global Mirror relationships” on page 475. Table 10-2 summarizes the LUNs that are used from each site.

   Table 10-2 Summary of the LUNs used on each site

Texas		Romania	
AIX DISK	**LSS/VOL ID**	**AIX DISK**	**LSS/VOL ID**
hdisk11	2605	hdisk10	2C06

2. Define the new LUNs:
   a. Run the **cfgmgr** command on the primary node `jordan`.
   b. Assign the PVID on the node `jordan`.

      ```
      chdev -l hdisk11 -a pv=yes
      ```

   c. Configure disk and PVID on local node leeann by using the **cfgmgr** command.
   d. Verify that the PVID is displayed by running the **lspv** command.
   e. Pause the PPRC on the primary site.

f. Fail over the PPRC to the secondary site.
g. Configure the disk and PVID on the remote node `robert` with the **`cfgmgr`** command.
h. Verify that the PVID is displayed by running the **`lspv`** command.
i. Fail back the PPRC to the primary site.

3. Add the new disk into the volume group by using C-SPOC as follows:

> **Important:** C-SPOC cannot perform certain LVM operations on nodes at the remote site (that contain the target volumes). Such operations include operations that require nodes at the target site to read from the target volumes. These operations cause an error message in C-SPOC. These operations include such functions as changing file system size, changing mount point, and adding LVM mirrors. However, nodes on the same site as the source volumes can successfully perform these tasks. The changes can be propagated later to the other site by using a lazy update.
>
> For C-SPOC operations to work on all other LVM operations, perform all C-SPOC operations with the Global Mirror volume pairs in synchronized or consistent states or the ACTIVE cluster on all nodes.

a. From the command line, type the **`smitty cl_admin`** command.
b. In SMIT, select the path **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Add a Volume to a Volume Group**.
c. Select the **txvg** volume group from the menu.
d. Select the disk or disks by PVID as shown in Figure 10-17.

```
                    Set Characteristics of a Volume Group

Move cursor to desired item and press Enter.

  Add a Volume to a Volume Group
  Change/Show characteristics of a Volume Group
  Remove a Volume from a Volume Group
  Enable/Disable a Volume Group for Cross-Site LVM Mirroring Verification

  +--------------------------------------------------------------------------+
  |                          Physical Volume Names                           |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  |   000a624a987825c8 ( hdisk10 on node robert )                            |
  |   000a624a987825c8 ( hdisk11 on nodes jordan,leeann )                    |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
  | F8=Image                F10=Exit                Enter=Do                 |
F1| /=Find                  n=Find Next                                      |
F9+--------------------------------------------------------------------------+
```

Figure 10-17 Disk selection to add to the volume group

e. Verify the menu information, as shown in Figure 10-18, and press Enter.

```
                        Add a Volume to a Volume Group

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

                                                        [Entry Fields]
  VOLUME GROUP name                                   txvg
  Resource Group Name                                 ds8kgmrg
  Node List                                           jordan,leeann,robert
  Reference node                                      robert
  VOLUME names                                        hdisk10
```

Figure 10-18 Add a Volume C-SPOC SMIT menu

Upon completion of the C-SPOC operation, the local nodes were updated, but the remote node was not updated as shown in Example 10-40. This node was not updated because the target volumes are not readable until the relationship is swapped. You receive an error message from C-SPOC, as shown in the note after Example 10-40. However, the lazy update procedure at the time of failover pulls in the remaining volume group information.

Example 10-40 New disk added to volume group on all nodes

```
root@jordan: lspv |grep txvg
hdisk6          000a625afe2a4958                    txvg
hdisk10         000a624a833e440f                    txvg
hdisk11         000a624a987825c8                    txvg

root@leeann: lspv |grep txvg
hdisk6          000a625afe2a4958                    txvg
hdisk10         000a624a833e440f                    txvg
hdisk11         000a624a987825c8                    txvg

root@robert: lspv
hdisk2          000a624a833e440f                    txvg
hdisk6          000a625afe2a4958                    txvg
hdisk10         000a624a987825c8                    none
```

Attention: When you use C-SPOC to modify a volume group that contains a Global Mirror replicated resource, you can expect to see the following error message:

```
cl_extendvg: Error executing clupdatevg txvg 000a624a833e440f on node robert
```

You do not need to synchronize the cluster because all of these changes are made to an existing volume group. However, consider running a verification.

Adding a new logical volume

Again you use C-SPOC to add a new logical volume. As noted earlier, this process updates the local nodes within the site. For the remote site, when a failover occurs, the lazy update process updates the volume group information as needed. This process also adds a bit of extra time to the failover time.

To add a new logical volume:

1. From the command line, type the **smitty cl_admin** command.
2. In SMIT, select the path **System Management (C-SPOC)** → **Storage** → **Logical Volumes** → **Add a Logical Volume**.
3. Select the **txvg** volume group from the menu.
4. Select the newly added disk **hdisk11** as shown in Figure 10-19.

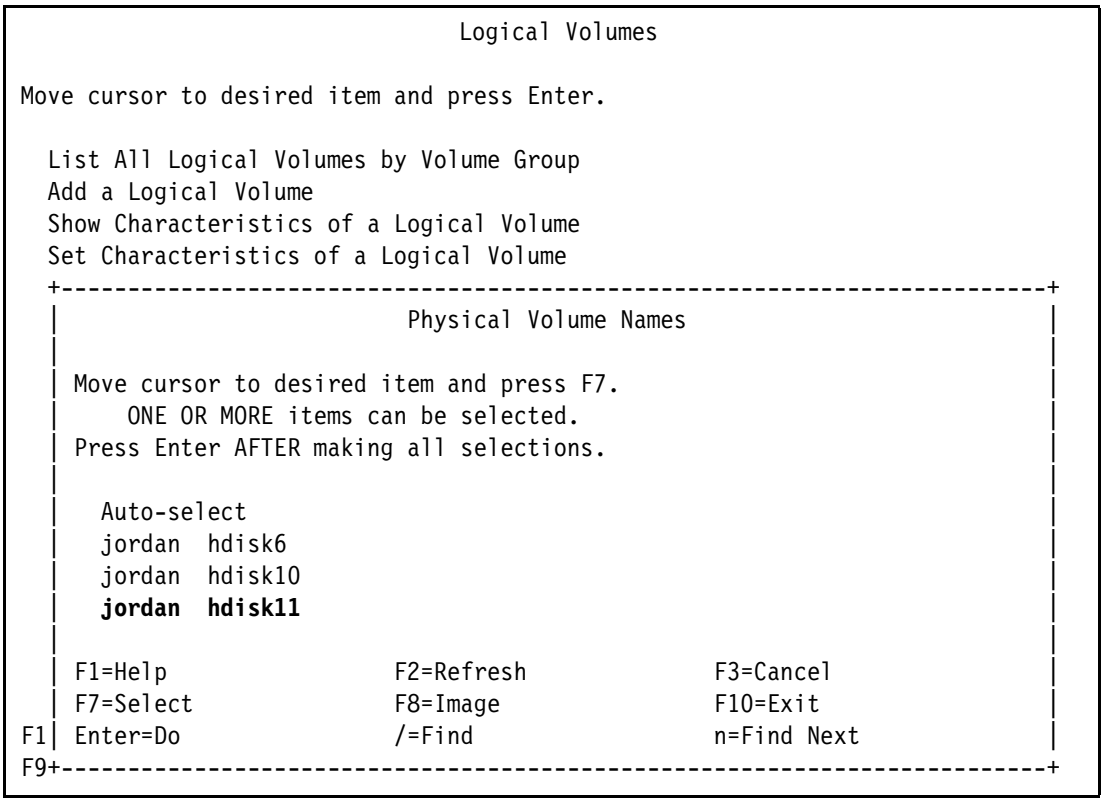


Figure 10-19 Choose disk for new logical volume creation

5. Complete the information in the final menu (Figure 10-20), and press Enter.

   We added a new logical volume, named `pattilv`, which consists of 100 logical partitions (LPARs) and selected `raw` for the type. We left all other values with their defaults.

```
                              Add a Logical Volume

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

[TOP]                                                   [Entry Fields]
  Resource Group Name                                  ds8kgmrg
  VOLUME GROUP name                                    txvg
  Node List                                            jordan,leeann,robert
  Reference node                                       jordan
* Number of LOGICAL PARTITIONS                        [100]                    #
  PHYSICAL VOLUME names                                hdisk11
  Logical volume NAME                                 [pattilv]
  Logical volume TYPE                                 [raw]                   +
  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                     1                      +
[MORE...15]
```

Figure 10-20 New logical volume C-SPOC SMIT menu

6. Upon completion of the C-SPOC operation, verify that the new logical volume is created locally on node `jordan` as shown in Example 10-41.

Example 10-41 Newly created logical volume

```
root@jordan:lsvg -l txvg
txvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
txlv                jfs2       250     150     3    open/syncd    /txro
txloglv             jfs2log    1       1       1    open/syncd    N/A
pattilv             raw        100     100     1    closed/syncd  N/A
```

Similar to when you create the volume group, you see an error message (Figure 10-21) about being unable to update the remote node.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

jordan: pattilv
cl_mklv: Error executing clupdatevg txvg 000a625afe2a4958 on node robert
```

Figure 10-21 C-SPOC normal error upon logical volume creation

Increasing the size of an existing file system

Again you use C-SPOC to perform this operation. As noted previously, this process updates the local nodes within the site. For the remote site, when a failover occurs, the lazy update process updates the volume group information as needed. This process also adds a bit of extra time to the failover time.

To increase the size of an existing file system, follow these steps:

1. From the command line, type the **`smitty cl_admin`** command.
2. In SMIT, select the path **System Management (C-SPOC) → Storage → File Systems → Change / Show Characteristics of a File System**.
3. Select the **txro** file system from the menu.
4. Complete the information in the final menu, and press Enter. In the example in Figure 10-22, notice that we change the size from 1024 MB to 1250 MB.

```
         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                                    txvg
  Resource Group Name                                  ds8kgmrg
* Node Names                                           robert,leeann,jordan

* File system name                                     /txro
  NEW mount point                                     [/txro]                  /
  SIZE of file system
          Unit Size                                    Megabytes               +
          Number of Units                             [1250]                   #
  Mount GROUP                                         []
  Mount AUTOMATICALLY at system restart?               no                      +
  PERMISSIONS                                          read/write              +
  Mount OPTIONS                                       []                       +
[MORE...7]
```

Figure 10-22 Changing the file system size on the final C-SPOC menu

5. Upon completion of the C-SPOC operation, verify that the new file system size locally on node `jordan` increased from 250 LPAR as shown in Example 10-41 on page 506 to 313 LPAR as shown Example 10-42.

Example 10-42 Newly increased file system size

```
root@jordan:lsvg -l txvg
txvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
txlv                jfs2       313     313     3    open/syncd    /txro
txloglv             jfs2log    1       1       1    open/syncd    N/A
pattilv             raw        100     100     1    closed/syncd  N/A
```

A cluster synchronization is not required because technically the resources did not change. All of the changes were made to an existing volume group that is already a resource in the resource group.

Testing the fallover after making the LVM changes

Because you do not know whether the cluster is going to work when you need it, repeat the steps from 10.7.2, “Rolling site failure” on page 496. The new logical volume `pattilv` and additional space on `/txro` show up on each node. However, a noticeable difference is on the site failover when the lazy update is performed to update the volume group changes.

10.8.2 Adding a Global Mirror pair into a new volume group

The steps to add a volume begin the same as the steps in 10.5, “Configuring AIX volume groups” on page 479. However, for completeness, this section provides an overview of the steps again and then provide details about the new LUNs to be used.

In this scenario, we reuse the LUNs from the previous section. We removed them from the volume group and removed the disks for all nodes except the main primary node `jordan`. In our process, we cleared the PVID and then assigned a new PVID for a clean start.

Table 10-3 provides a summary of the LUNs that we implemented in each site.

Table 10-3 Summary of the LUNs implemented in each site

Texas		Romania	
AIX dISK	**LSS/VOL ID**	**AIX dISK**	**LSS/VOL ID**
hdisk11	2605	hdisk10	2C06

Now continue with the following steps, which are the same as those steps for defining new LUNs:

1. Run the **cfgmgr** command on the primary node `jordan`.
2. Assign the PVID on the node `jordan`:

   ```
   chdev -l hdisk11 -a pv=yes
   ```

3. Configure the disk and PVID on the local node `leeann` by using the **cfgmgr** command.
4. Verify that PVID shows up by using the **lspv** command.
5. Pause the PPRC on the primary site.
6. Fail over the PPRC to the secondary site.
7. Fail back the PPRC to the secondary site.
8. Configure the disk and PVID on the remote node `robert` by using the **cfgmgr** command.
9. Verify that PVID shows up by using the **lspv** command.
10. Pause the PPRC on the secondary site.
11. Fail over the PPRC to the primary site.
12. Fail back the PPRC to the primary site.

The main difference between adding a volume group and extending an existing one is that, when you add a volume group, you must swap the pairs twice. When you extend an existing volume group, you can get away with only swapping once. This difference is similar to the original setup where we created all LVM components on the primary site and swap the PPRC pairs to the remote site to import the volume group and then swap it back.

You can avoid performing two swaps, as we showed, by not choosing to include the third node when creating the volume group. Then, you can swap the pairs, run **cfgmgr** on the new disk with the PVID, import the volume group, and swap the pairs back.

Creating a volume group

Create a volume group by using C-SPOC:

1. From the command line, type the **smitty cl_admin** command.
2. In SMIT, select the path **System Management (C-SPOC) → Storage → Volume Groups → Create a Volume to a Volume Group**.
3. Select the specific nodes. In this case, we chose all three nodes as shown in Figure 10-23.

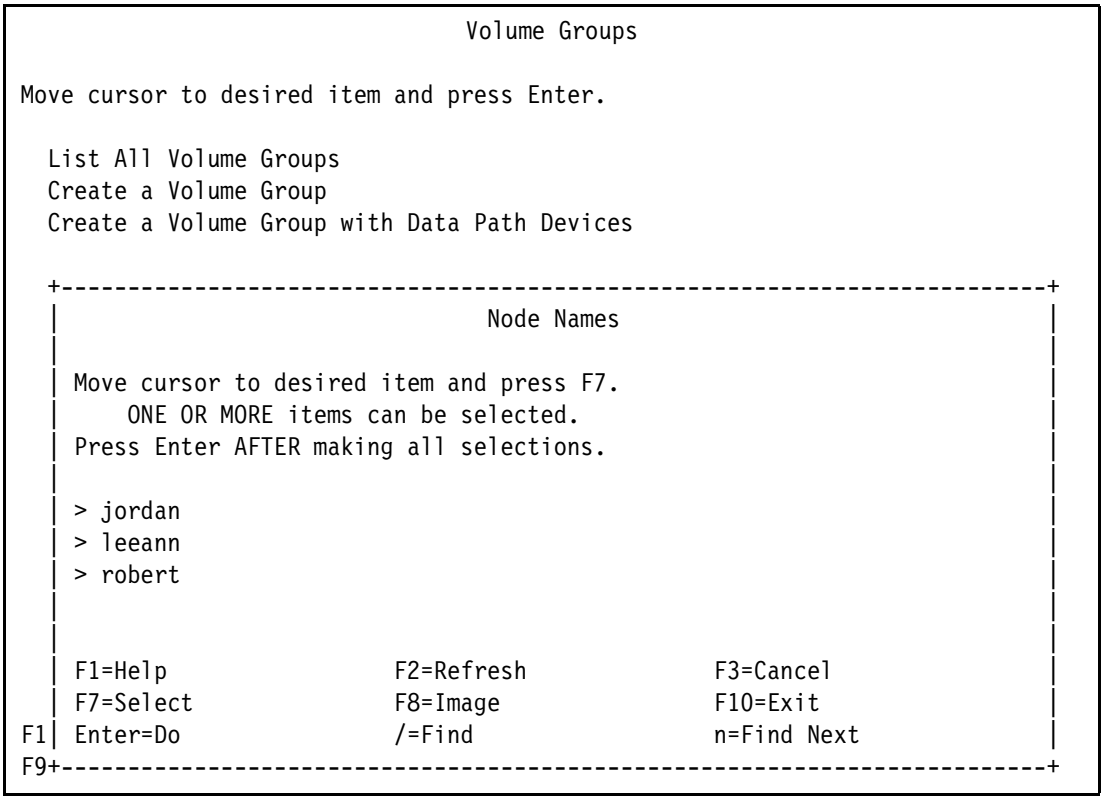


Figure 10-23 Adding a volume group node pick list

4. Select the disk or disks by PVID as shown in Figure 10-24.

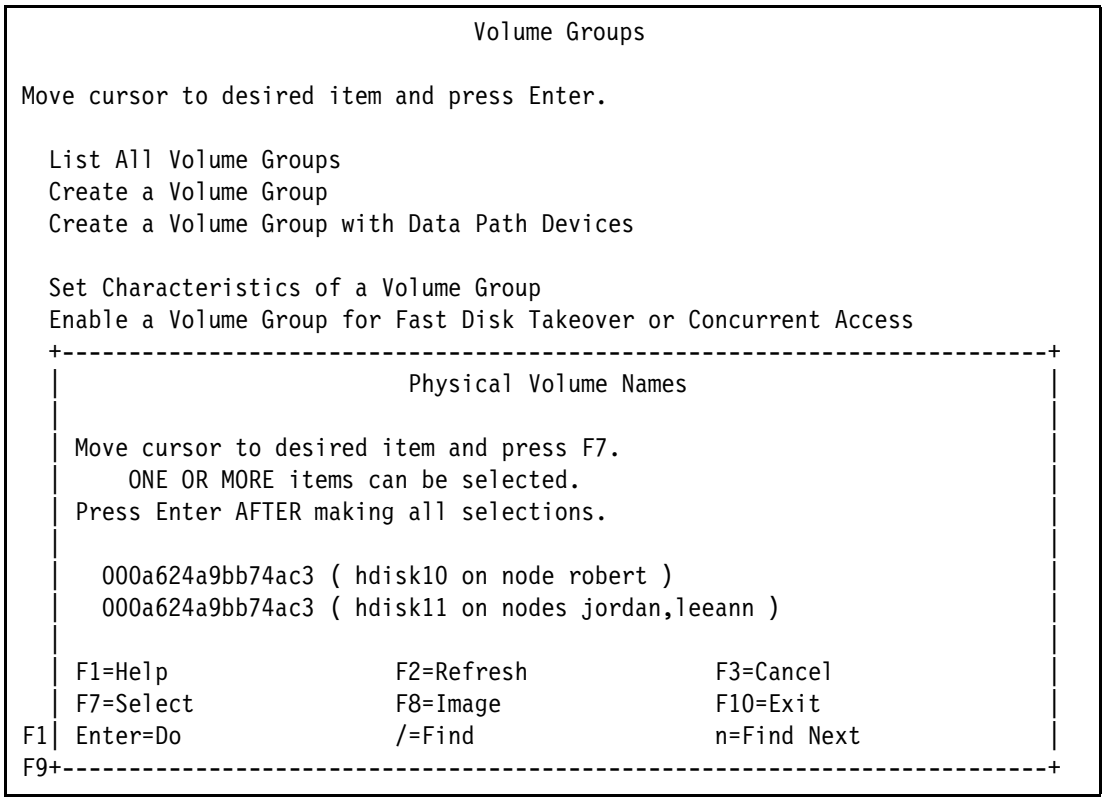


Figure 10-24 Selecting the disk or disks for the new volume group list

5. Select the volume group type. In this scenario, we select `scalable` as shown in Figure 10-25.

```
                                Volume Groups

Move cursor to desired item and press Enter.

  List All Volume Groups
  Create a Volume Group
  Create a Volume Group with Data Path Devices

  Set Characteristics of a Volume Group
  +--------------------------------------------------------------------------+
  |                            Volume Group Type                             |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  |   Legacy                                                                 |
  |   Original                                                               |
  |   Big                                                                    |
  |   Scalable                                                               |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
  | F8=Image                F10=Exit                Enter=Do                 |
F1| /=Find                  n=Find Next                                      |
F9+--------------------------------------------------------------------------+
```

Figure 10-25 Choosing the volume group type for the new volume group pick list

6. Select the resource group. We select `ds8kgmrg` as shown in Figure 10-26.

```
                        Create a Scalable Volume Group

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

[TOP]                                                  [Entry Fields]
  Node Names                                          jordan,leeann,robert
  Resource Group Name                                 [ds8kgmrg]              +
  PVID                                                 000a624a9bb74ac3
  VOLUME GROUP name                                   [princessvg]
  Physical partition SIZE in megabytes                 4                      +
  Volume group MAJOR NUMBER                           [51]                    #
  Enable Cross-Site LVM Mirroring Verification         false                  +
  Enable Fast Disk Takeover or Concurrent Access       Fast Disk Takeover or> +
  Volume Group Type                                    Scalable
  Maximum Physical Partitions in units of 1024         32                     +
  Maximum Number of Logical Volumes                    256                    +
```

Figure 10-26 Create a Scalable Volume Group (final) menu

7. Select a volume group name. We select **princessvg**. Then, press Enter.

Instead of using C-SPOC, you can perform the steps manually and then import the volume groups on each node as needed. However, remember to add the volume group into the resource group after you create it. With C-SPOC, you can automatically add it to the resource group while you are creating the volume group.

You can also use the C-SPOC CLI commands (Example 10-43). These commands are in the `/usr/es/sbin/cluster/cspoc` directory, and all begin with the `cli_` prefix. Similar to the SMIT menus, their operation output is also saved in the `cspoc.log` file.

Example 10-43 C-SPOC CLI commands

```
root@jordan: ls cli_*
cli_assign_pvids  cli_extendlv      cli_mkvg          cli_rmlv
cli_chfs          cli_extendvg      cli_on_cluster    cli_rmlvcopy
cli_chlv          cli_importvg      cli_on_node       cli_syncvg
cli_chvg          cli_mirrorvg      cli_reducevg      cli_unmirrorvg
cli_crfs          cli_mklv          cli_replacepv     cli_updatevg
cli_crlvfs        cli_mklvcopy      cli_rmfs
```

Upon completion of the C-SPOC operation, the local nodes are updated, but the remote node is not as shown in Example 10-44. The remote nodes are not updated because the target volumes are not readable until the relationship is swapped. You see an error message from C-SPOC as shown in the note after Example 10-44. After you create all LVM structures, you swap the pairs back to the remote node and import the new volume group and logical volume.

Example 10-44 New disk added to volume group on all nodes

```
root@jordan: lspv |grep princessvg
hdisk11         000a624a9bb74ac3                    princessvg

root@leeann: lspv |grep princessvg
hdisk11         000a624a9bb74ac3                    princessvg

root@robert: lspv |grep princessvg
```

Attention: When you use C-SPOC to add a volume group that contains a Global Mirror replicated resource, you might see the following error message:

```
cl_importvg: Error executing climportvg  -V 51 -c -y princessvg -Q
000a624a9bb74ac3 on node robert
```

While this message is normal, if you select any remote nodes, you can omit the remote nodes and then you do not see the error message. This step is allowed because you manually import it anyway.

When creating the volume group, it usually is automatically added to the resource group as shown in Example 10-45 on page 512. However, with the error message indicted in the previous attention box, it might not be automatically added. Therefore, double check that the volume group is added into the resource group before continuing. Otherwise, we do not have to change the resource group any further. The new LUN pairs are added to the same storage subsystems and the same session (3) that is already defined in the mirror group `texasmg`.

Example 10-45 New volume group added to existing resource group

```
Resource Group Name                              ds8kgmrg
Inter-site Management Policy                     Prefer Primary Site
Participating Nodes from Primary Site            jordan leeann
```

```
Participating Nodes from Secondary Site          robert
Startup Policy                                   Online On Home Node Only
Fallover Policy                                  Fallover To Next Priority Node
Fallback Policy                                  Never Fallback
Service IP Label                                 serviceip_2
Volume Groups                                    txvg princessvg +
GENXD Replicated Resources                       texasmg
```

Adding a logical volume on the new volume group

You repeat the steps in "Adding a new logical volume" on page 504 to create a new logical volume, named `princesslv`, on the newly created volume group, `princessvg`, as shown in Example 10-46.

Example 10-46 New logical volume on the newly added volume group

```
root@jordan: lsvg -l princessvg
princessvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
princesslv          raw        38      38      1    closed/syncd  N/A
```

Importing the new volume group to the remote site

To import the volume group, follow the steps in 10.5.2, "Importing the volume groups in the remote site" on page 481. As a review, we perform the following steps:

1. Vary off the volume group on the local site.
2. Pause the PPRC pairs on the local site.
3. Fail over the PPRC pairs on the remote site.
4. Fail back the PPRC pairs on the remote site.
5. Import the volume group.
6. Vary off the volume group on the remote site.
7. Pause the PPRC pairs on the remote site.
8. Fail over the PPRC pairs on the local site.
9. Fail back the PPRC pairs on the local site.

Synchronizing and verifying the cluster configuration

You now synchronize the resource group change to include the new volume group that was added. However, first run a verification only to check for errors. If you find errors, you must fix them manually because they are not automatically fixed in a running environment.

Then, synchronize and verify it:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **Extended Configuration** → **Extended Verification and Synchronization and Verification**.

3. Select the options as shown in Figure 10-27.

```
       HACMP Verification and Synchronization (Active Cluster Nodes Exist)

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

                                                      [Entry Fields]
* Verify changes only?                                [No]                    +
* Logging                                             [Standard]              +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
```

Figure 10-27 Extended Verification and Synchronization SMIT menu

4. Verify that the information is correct, and press Enter.

Upon completion, the cluster configuration is synchronize and can now be tested.

Testing the failover after adding a volume group

Because you do not know whether the cluster is going to work when needed, repeat the steps from 10.7.2, “Rolling site failure” on page 496. The new volume group `princessvg` and logical volume `princesslv` are showing up in each node.

11

Disaster recovery by using Hitachi TrueCopy and Universal Replicator

This chapter explains how to configure disaster recovery based on IBM PowerHA SystemMirror for AIX Enterprise Edition with Hitachi TrueCopy/Hitachi Universal Replicator (HUR) replication services. This support is added in version 6.1 with service pack 3 (SP3).

This chapter includes the following sections:

- Planning for TrueCopy/HUR management
- Overview of TrueCopy/HUR management
- Scenario description
- Configuring the TrueCopy/HUR resources
- Failover testing
- LVM administration of TrueCopy/HUR replicated pairs

11.1 Planning for TrueCopy/HUR management

Proper planning is crucial to the success of any implementation. Plan the storage deployment and replication necessary for your environment. This process is related to the applications and middleware that are being deployed in the environment, which can eventually be managed by PowerHA SystemMirror Enterprise Edition. This section lightly covers site, network, storage area network (SAN), and storage planning, which are all key factors. However, the primary focus of this section is the software prerequisites and support considerations.

11.1.1 Software prerequisites

The following software is required:

- One of the following AIX levels or later:
  - AIX 5.3 TL9 and RSCT 2.4.12.0
  - AIX 6.1 TL2 SP3 and RSCT 2.5.4.0
- Multipathing software
  - AIX MPIO
  - Hitachi Dynamic Link Manager (HDLM)
- PowerHA 6.1 Enterprise Edition with SP3

  The following additional file sets are included in SP3, must be installed separately, and require the acceptance of the license during the installation:

  - `cluster.es.tc`

    ```
    6.1.0.0 ES HACMP - Hitachi support - Runtime Commands
    6.1.0.0 ES HACMP - Hitachi support Commands
    ```

  - `cluster.msg.en_US.tc` (optional)

    ```
    6.1.0.0 HACMP Hitachi support Messages - U.S. English
    6.1.0.0 HACMP Hitachi Messages - U.S. English IBM-850
    6.1.0.0 HACMP Hitachi Messages – Japanese
    6.1.0.0 HACMP Hitachi Messages - Japanese IBM-eucJP
    ```

- Hitachi Command Control Interface (CCI) Version 01-23-03/06 or later
- USPV Microcode Level 60-06-05/00 or later

11.1.2 Minimum connectivity requirements for TrueCopy/HUR

For TrueCopy/HUR connectivity, you must have the following minimum requirements in place:

- Ensure connectivity from the local Universal Storage Platform VM (USP VM) to the AIX host ports.

  The external storage ports on the local USP VMs (Data Center 1 and Data Center 2) are zoned and cabled to their corresponding existing storage systems.
- Present both the primary and secondary source devices to the local USP VMs.

  Primary and secondary source volumes in the migration group are presented from the existing storage systems to the corresponding local USP VMs. This step is transparent to the servers in the migration set. No devices are imported or accessed by the local USP VMs at this stage.

- Establish replication connectivity between the target storage systems.

  TrueCopy initiator and MCU target ports are configured on the pair of target USP VMs, and an MCU/RCU pairing is established to validate the configuration.
- Ensure replication connectivity from the local USP VMs to the remote USP VM TrueCopy/HUR initiator. Also, ensure that MCU target ports are configured on the local and remote USP VMs. In addition, confirm that MCU and RCU pairing is established to validate the configuration.
- For HUR, configure Universal Replicator Journal Groups on local and remote USP VM storage systems.
- Configure the target devices.

  Logical devices on the target USP VM devices are formatted and presented to front-end ports or host storage domains. This way, device sizes, logical unit numbers, host modes, and presentation worldwide names (WWNs) are identical on the source and target storage systems. Devices are presented to host storage domains that correspond to both production and disaster recovery standby servers.
- Configure the target zoning.

  Zones are defined between servers in the migration group and the target storage system front-end ports, but new zones are not activated.

Ideally the connectivity is through redundant links, switches, and fabrics to the hosts and between the storage units themselves.

11.1.3 Considerations

Keep in mind the following considerations for mirroring PowerHA SystemMirror Enterprise Edition with TrueCopy/HUR:

- AIX Virtual SCSI is not supported in this initial release.
- Logical Unit Size Expansion (LUSE) for Hitachi is not supported.
- Only fence-level NEVER is supported for synchronous mirroring.
- Only HUR is supported for asynchronous mirroring.
- The dev_name must map to a logical device, and the dev_group must be defined in the HORCM_LDEV section of the `horcm.conf` file.
- The PowerHA SystemMirror Enterprise Edition TrueCopy/HUR solution uses dev_group for any basic operation, such as the **`pairresync`**, **`pairevtwait`**, or **`horctakeover`** operation. If several dev_names are in a dev_group, the dev_group must be enabled for consistency.
- PowerHA SystemMirror Enterprise Edition does not trap Simple Network Management Protocol (SNMP) notification events for TrueCopy/HUR storage. If a TrueCopy link goes down when the cluster is up and later the link is repaired, you must manually resynchronize the pairs.
- The creation of pairs is done outside the cluster control. You must create the pairs before you start the cluster services.
- Resource groups that are managed by PowerHA SystemMirror Enterprise Edition cannot contain volume groups with both TrueCopy/HUR-protected and non-TrueCopy/HUR-protected disks.
- All nodes in the PowerHA SystemMirror Enterprise Edition cluster must use same `horcm` instance.

- You cannot use Cluster Single Point Of Control (C-SPOC) for the following Logical Volume Manager (LVM) operations to configure nodes at the remote site that contain the target volume:
  - Creating a volume group
  - Operations that require nodes at the target site to write to the target volumes

    For example, changing the file system size, changing the mount point, or adding LVM mirrors cause an error message in C-SPOC. However, nodes on the same site as the source volumes can successfully perform these tasks. The changes are then propagated to the other site by using a lazy update.

C-SPOC on other LVM operations: For C-SPOC operations to work on all other LVM operations, perform all C-SPOC operations when the cluster is active on all PowerHA SystemMirror Enterprise Edition nodes and the underlying TrueCopy/HUR PAIRs are in a PAIR state.

11.2 Overview of TrueCopy/HUR management

Hitachi TrueCopy/HUR storage management uses Command Control Interface (CCI) operations from the AIX operating system and PowerHA SystemMirror Enterprise Edition environment. PowerHA SystemMirror Enterprise Edition uses these interfaces to discover and integrate the Hitachi Storage replicated storage into the framework of PowerHA SystemMirror Enterprise Edition. With this integration, you can manage high availability disaster recovery (HADR) for applications by using the mirrored storage.

Integration of TrueCopy/HUR and PowerHA SystemMirror Enterprise Edition provides the following benefits:

- Support for the Inter-site Management policy of Prefer Primary Site or Online on Either Site
- Flexible user-customizable resource group policies
- Support for cluster verification and synchronization
- Limited support for the C-SPOC in PowerHA SystemMirror Enterprise Edition
- Automatic failover and reintegration of server nodes that are attached to pairs of TrueCopy/HUR disk subsystem within sites and across sites
- Automatic management for TrueCopy/HUR links
- Management for switching the direction of the TrueCopy/HUR relationships when a site failure occurs. With this process, the backup site can take control of the managed resource groups in PowerHA SystemMirror Enterprise Edition from the primary site

11.2.1 Installing the Hitachi CCI software

Use the following steps as a guideline to help you install the Hitachi CCI on the AIX cluster nodes. You can also find this information in the `/usr/sbin/cluster/release_notes_xd` file. However, the release notes exist only if you already have the PowerHA SystemMirror Enterprise Edition software installed. Always consult the latest version of the *Hitachi Command Control Interface (CCI) User and Reference Guide*, MK-90RD011, which you can download from:

`http://communities.vmware.com/servlet/JiveServlet/download/1183307-19474`

If you are installing CCI from a CD, use the **RMinstsh** and **RMuninst** scripts on the CD to automatically install and uninstall the CCI software.

Important: You must install the Hitachi CCI software into the /HORCM/usr/bin directory. Otherwise, you must create a symbolic link to this directory.

For other media, use the instructions in the following sections.

Installing the Hitachi CCI software into a root directory

To install the Hitachi CCI software into the root directory, follow these steps:

1. Insert the installation medium into the proper I/O device.
2. Move to the current root directory:

   ```
   # cd /
   ```

3. Copy all files from the installation medium by using the **cpio** command:

   ```
   # cpio -idmu < /dev/XXXX XXXX = I/O device
   ```

   Preserve the directory structure (**d** flag) and file modification times (**m** flag), and copy unconditionally (**u** flag). For diskettes, load them sequentially, and repeat the command. An I/O device name of “floppy disk” designates a surface partition of the raw device file (unpartitioned raw device file).

4. Run the Hitachi Open Remote Copy Manager (HORCM) installation command:

   ```
   # /HORCM/horcminstall.sh
   ```

5. Verify installation of the proper version by using the **raidqry** command:

   ```
   # raidqry -h
   Model: RAID-Manager/AIX
   Ver&Rev: 01-23-03/06
   Usage: raidqry [options] for HORC
   ```

Installing the Hitachi CCI software into a nonroot directory

To install the Hitachi CCI software into a non-root directory:

1. Insert the installation medium, such as a CD, into the proper I/O device.
2. Move to the desired directory for CCI. The specified directory must be mounted by a partition of except root disk or an external disk.

   ```
   # cd /Specified Directory
   ```

3. Copy all files from the installation medium by using the **cpio** command:

   ```
   # cpio -idmu < /dev/XXXX XXXX = I/O device
   ```

   Preserve the directory structure (**d** flag) and file modification times (**m** flag), and copy unconditionally (**u** flag). For diskettes, load them sequentially, and repeat the command. An I/O device name of “floppy disk” designates a surface partition of the raw device file (unpartitioned raw device file).

4. Make a symbolic link to the /HORCM directory:

   ```
   # ln -s /Specified Directory/HORCM /HORCM
   ```

5. Run the HORCM installation command:

   ```
   # /HORCM/horcminstall.sh
   ```

6. Verify installation of the proper version by using the **raidqry** command:

```
# raidqry -h
Model: RAID-Manager/AIX
Ver&Rev: 01-23-03/06
Usage: raidqry [options] for HORC
```

Installing a newer version of the Hitachi CCI software

To install a newer version of the CCI software:

1. Confirm that HORCM is not running. If it is running, shut it down:

   ```
   One CCI instance: # horcmshutdown.sh
   Two CCI instances: # horcmshutdown.sh 0 1
   ```

   If Hitachi TrueCopy commands are running in the interactive mode, end the interactive mode and exit these commands by using the **-q** option.

2. Insert the installation medium, such as a CD, into the proper I/O device.

3. Move to the directory that contains the HORCM directory as in the following example for the root directory:

   ```
   # cd /
   ```

4. Copy all files from the installation medium by using the **cpio** command:

   ```
   # cpio -idmu < /dev/XXXX XXXX = I/O device
   ```

   Preserve the directory structure (**d** flag) and file modification times (**m** flag) and copy unconditionally (**u** flag). For diskettes, load them sequentially, and repeat the command. An I/O device name of “floppy disk” designates a surface partition of the raw device file (unpartitioned raw device file).

5. Run the HORCM installation command:

   ```
   # /HORCM/horcminstall.sh
   ```

6. Verify installation of the proper version by using the **raidqry** command:

   ```
   # raidqry -h
   Model: RAID-Manager/AIX
   Ver&Rev: 01-23-03/06
   Usage: raidqry [options] for HORC
   ```

11.2.2 Overview of the CCI instance

The CCI components on the storage system include the command device or devices and the Hitachi TrueCopy volumes, ShadowImage volumes, or both. Each CCI instance on a UNIX/PC server includes the following components:

- HORCM:
  - Log and trace files
  - A command server
  - Error monitoring and event reporting files
  - A configuration management feature
- Configuration definition file that is defined by the user
- The Hitachi TrueCopy user execution environment, ShadowImage user execution environment, or both, which contain the TrueCopy/ShadowImage commands, a command log, and a monitoring function.

11.2.3 Creating and editing the horcm.conf files

The configuration definition file is a text file that is created and edited by using any standard text editor, such as the **vi** editor. A sample configuration definition file, HORCM_CONF (/HORCM/etc/horcm.conf), is included with the CCI software. Use this file as the basis for creating your configuration definition files. The system administrator must copy the sample file, set the necessary parameters in the copied file, and place the copied file in the proper directory.

For detailed descriptions of the configuration definition files for sample CCI configurations, see the *Hitachi Command Control Interface (CCI) User and Reference Guide*, MK-90RD011, which you can download from:

http://communities.vmware.com/servlet/JiveServlet/download/1183307-19474

Important: Do not edit the configuration definition file while HORCM is running. Shut down HORCM, edit the configuration file as needed, and then restart HORCM.

You might have multiple CCI instances, each of which uses its own specific horcm#.conf file. For example, instance 0 might be horcm0.conf, instance 1 (Example 11-1) might be horcm1.conf, and so on. The test scenario that presented later in this chapter uses instance 2 and provides examples of the horcm2.conf file on each cluster node.

Example 11-1 The hormc.conf file

```
Example configuration files:

  horcm1.conf file on local node
  ------------------------------
   HORCM_MON
   #ip_address => Address of the local node
   #ip_address     service     poll(10ms)  timeout(10ms)
   10.15.11.194    horcm1      12000         3000

   HORCM_CMD
   #dev_name => hdisk of Command Device
   #UnitID 0 (Serial# eg. 45306)
   /dev/hdisk19

   HORCM_DEV
   #Map dev_grp to LDEV#
   #dev_group dev_name port# TargetID LU# MU#
   VG01       test01   CL1-B     1      5   0
   VG01       work01   CL1-B     1      24  0
   VG01       work02   CL1-B     1      25  0

   HORCM_INST
   #dev_group ip_address    service
   VG01       10.15.11.195   horcm1

 horcm1.conf file on remote node
 -------------------------------
   HORCM_MON
   #ip_address => Address of the local node
   #ip_address     service     poll(10ms)  timeout(10ms)
   10.15.11.195    horcm1      12000         3000
```

```
HORCM_CMD
#dev_name => hdisk of Command Device
#UnitID 0 (Serial# eg. 45306)
/dev/hdisk19

HORCM_DEV
#Map dev_grp to LDEV#
#dev_group dev_name port# TargetID LU# MU#
VG01       test01   CL1-B     1      5   0
VG01       work01   CL1-B     1      21  0
VG01       work02   CL1-B     1      22  0

HORCM_INST
#dev_group ip_address    service
VG01       10.15.11.194   horcm1

NOTE 1: For the horcm instance to use any available command device, in case one of
them fails, it is RECOMMENDED that, in your horcm file, under HORCM_CMD
section, the command device, is presented in the format below,
where 10133 is the serial # of the array:

         \\.\CMD-10133:/dev/hdisk/

For example:

        \\.\CMD-10133:/dev/rhdisk19 /dev/rhdisk20  ( note space in between).

NOTE 2:  The Device_File will show "-----" for the "pairdisplay -fd" command,
         which will also cause verification to fail, if the ShadowImage license
         has not been activated on the storage system and the MU# column is not
empty.
         It is therefore recommended that the MU# column be left blank if the
         ShadowImage license is NOT activated on the storage system.
```

Starting the HORCM instances

To start one instance of the CCI:

1. Modify the `/etc/services` file to register the port name/number (service) of the configuration definition file. Make the port name/number the same on all servers.

   ```
   horcm xxxxx/udp xxxxx = the port name/number of horcm.conf
   ```

2. Optional: If you want HORCM to start automatically each time the system starts, add `/etc/horcmstart.sh` to the system automatic startup file (such as the `/sbin/rc` file).
3. Run the **`horcmstart.sh`** script manually to start the CCI instance:

   ```
   # horcmstart.sh
   ```

4. Set the log directory (HORCC_LOG) in the command execution environment as needed.
5. Optional: If you want to perform Hitachi TrueCopy operations, do *not* set the `HORCC_MRCF` environment variable.
   - For the B shell:

     ```
     # HORCC_MRCF=1
     # export HORCC_MRCF
     ```

– For the C shell:

```
# setenv HORCC_MRCF 1
# pairdisplay -g xxxx xxxx = group name
```

To start two instances of the CCI, follow these steps:

1. Modify the `/etc/services` file to register the port name/number (service) of each configuration definition file. The port name/number must be different for each CCI instance.

   ```
   horcm0 xxxxx/udp xxxxx = the port name/number for horcm0.conf
   horcm1 yyyyy/udp yyyyy = the port name/number for horcm1.conf
   ```

2. If you want HORCM to start automatically each time the system starts, add `/etc/horcmstart.sh 0 1` to the system automatic startup file (such as the `/sbin/rc` file).
3. Run the **`horcmstart.sh`** script manually to start the CCI instances:

   ```
   # horcmstart.sh 0 1
   ```

4. Set an instance number to the environment that runs a command:

   For the B shell:

   ```
   # HORCMINST=X X = instance number = 0 or 1
   # export HORCMINST
   ```

   For the C shell:

   ```
   # setenv HORCMINST X
   ```

5. Set the log directory (HORCC_LOG) in the command execution environment as needed.
6. If you want to perform Hitachi TrueCopy operations, do **not** set the HORCC_MRCF environment variable.

   For B shell:

   ```
   # HORCC_MRCF=1
   # export HORCC_MRCF
   ```

   For C shell:

   ```
   # setenv HORCC_MRCF 1
   # pairdisplay -g xxxx xxxx = group name
   ```

11.3 Scenario description

This scenario uses four nodes, two in each of the two sites: *Austin* and *Miami*. Nodes `jessica` and `bina` are in the Austin site, and nodes `krod` and `maddi` are in the Miami site. Each site provides local automatic failover, along with remote recovery for the other site, which is often referred to as a *mutual takeover configuration*. Figure 11-1 on page 524 provides a software and hardware overview of the tested configuration between the two sites.

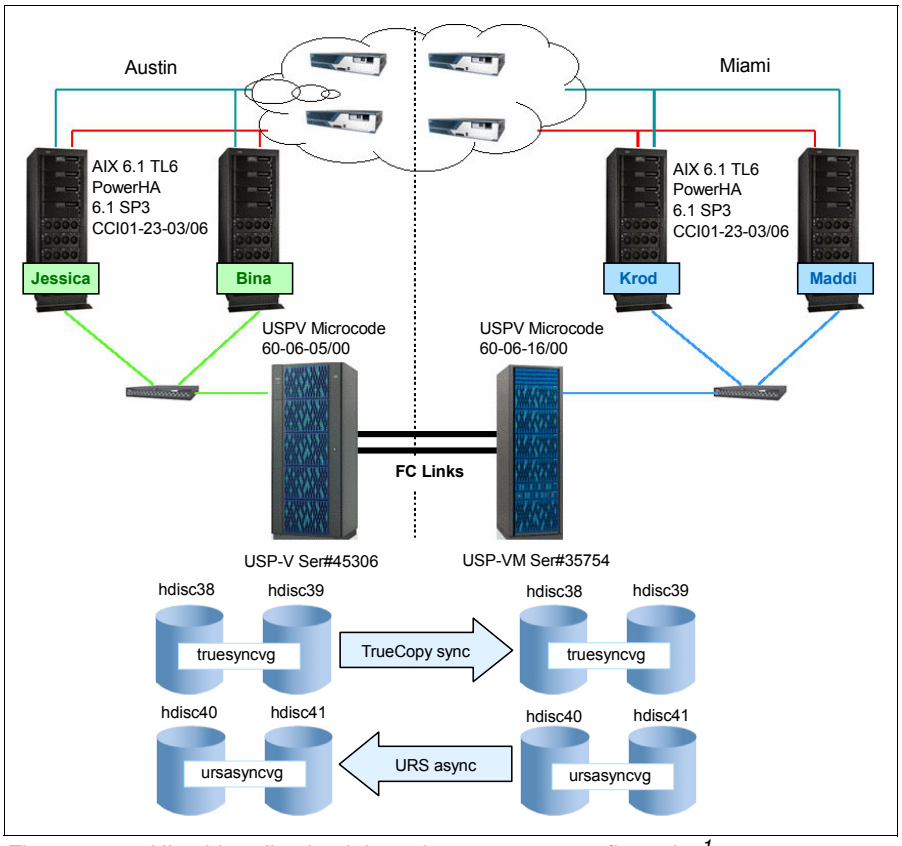


Figure 11-1 Hitachi replication lab environment test configuration[1]

Each site consists of two type Ethernet networks. In this case, both networks are used for a public Ethernet and for cross-site networks. Usually the cross-site network is on separate segments and is an `XD_ip` network. It is also common to use site-specific service IP labels. Example 11-2 shows the interlace list from the cluster topology.

Example 11-2 Test topology information

```
root@jessica: llsif
Adapter          Type       Network    Net Type   Attribute  Node       IP Address
jessica          boot       net_ether_02 ether      public    jessica    9.3.207.24
jessicaalt       boot       net_ether_03 ether      public    jessica    207.24.1.1
service_1        service    net_ether_03 ether      public    jessica    1.2.3.4
service_2        service    net_ether_03 ether      public    jessica    1.2.3.5
bina             boot       net_ether_02 ether      public    bina       9.3.207.77
bina   alt       boot       net_ether_03 ether      public    bina       207.24.1.2
service_1        service    net_ether_03 ether      public    bina       1.2.3.4
service_2        service    net_ether_03 ether      public    bina       1.2.3.5
krod             boot       net_ether_02 ether      public    krod       9.3.207.79
krod   alt       boot       net_ether_03 ether      public    krod       207.24.1.3
service_1        service    net_ether_03 ether      public    krod       1.2.3.4
service_2        service    net_ether_03 ether      public    krod       1.2.3.5
maddi            boot       net_ether_02 ether      public    maddi      9.3.207.78
maddi  alt       boot       net_ether_03 ether      public    maddi      207.24.1.4
service_1        service    net_ether_03 ether      public    maddi      1.2.3.4
service_2        service    net_ether_03 ether      public    maddi      1.2.3.5
```

[1] Courtesy of Hitachi Data Systems

In this scenario, each node or site has four unique disks that are defined through each of the two separate Hitachi storage units. The `jessica` and `bina` nodes at the Austin site have two disks, `hdisk38` and `hdisk3`. These disks are the primary source volumes that use TrueCopy synchronous replication for the `truesyncvg` volume group. The other two disks, `hdisk40` and `hdisk41`, are to be used as the target secondary volumes that use HUR for asynchronous replication from the Miami site for the `ursasyncvg` volume group.

The `krod` and `bina` nodes at the Miami site have two disks, `hdisk38` and `hdisk39`. These disks are the secondary target volumes for the TrueCopy synchronous replication of the `truesyncvg` volume group from the Austin site. The other two disks, `hdisk40` and `hdisk41`, are to be used as the primary source volumes for the `ursasyncvg` volume group that uses HUR for asynchronous replication.

11.4 Configuring the TrueCopy/HUR resources

This section explains how to perform the following tasks to configure the resources for TrueCopy/HUR:

- Assigning LUNs to the hosts (host groups)
- Creating replicated pairs
- Configuring an AIX disk and dev_group association

For each of these tasks, the Hitachi storage units were added to the SAN fabric and zoned appropriately. Also, the host groups were created for the appropriate node adapters, and the LUNs were created within the storage unit.

11.4.1 Assigning LUNs to the hosts (host groups)

In this task, you assign LUNs by using the Hitachi Storage Navigator. Although an overview of the steps is provided, always refer to the official Hitachi documentation for your version as needed.

To begin, the Hitachi USP-V storage unit is at the Austin site. The host group, `JessBina`, is assigned to port CL1-E on the Hitachi storage unit with the serial number 45306. Usually the host group is assigned to multiple ports for full multipath redundancy.

To assign the LUNs to the hosts:

1. Locate the free LUNs and assign them to the proper host group.
   a. Verify whether LUNs are currently assigned by checking the number of paths that are associated with the LUN. If the fields are blank, the LUN is unassigned.
   b. Assign the LUNs. To assign one LUN, click and drag it to a free LUN/LDEV location. To assign multiple LUNs, hold down the Shift key and click each LUN. Then, right-click the selected LUNs and drag them to a free location.

      This free location is indicated by a black and white disk image that also contains no information in the corresponding attribute columns of LDEV/UUID/Emulation as shown in Figure 11-2 on page 526.

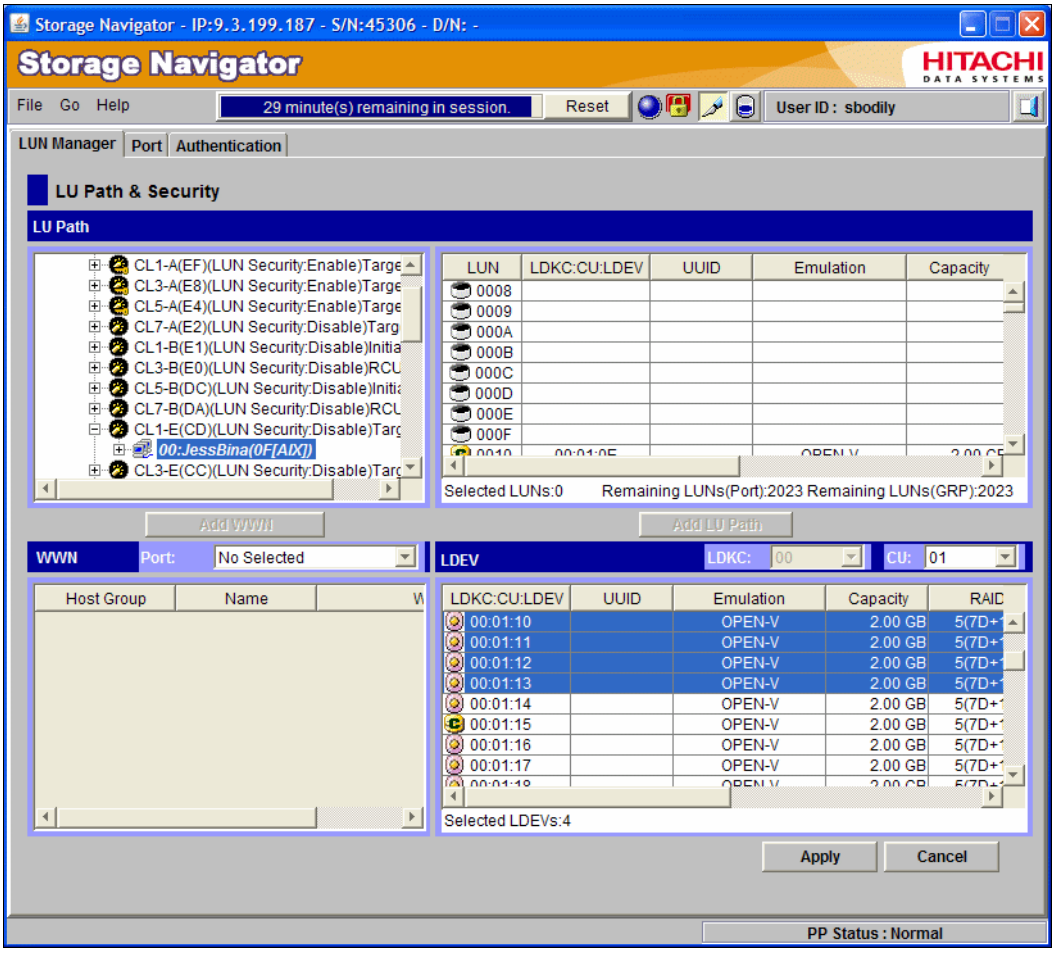


Figure 11-2 Assigning LUNs to the Austin site nodes[2]

2. In the path verification window (Figure 11-3), check the information and record the LUN number and LDEV numbers. You use this information later. However, you can also retrieve this information from the AIX system after the devices are configured by the host. Click **OK**.

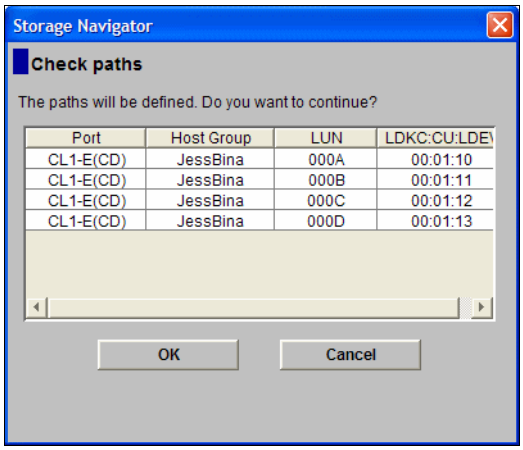


Figure 11-3 Checking the paths for the Austin LUNs[3]

[2] Courtesy of Hitachi Data Systems
[3] Courtesy of Hitachi Data Systems

3. Back on the **LUN Manager** tab (Figure 11-4), click **Apply** for these paths to become active and the assignment to be completed.

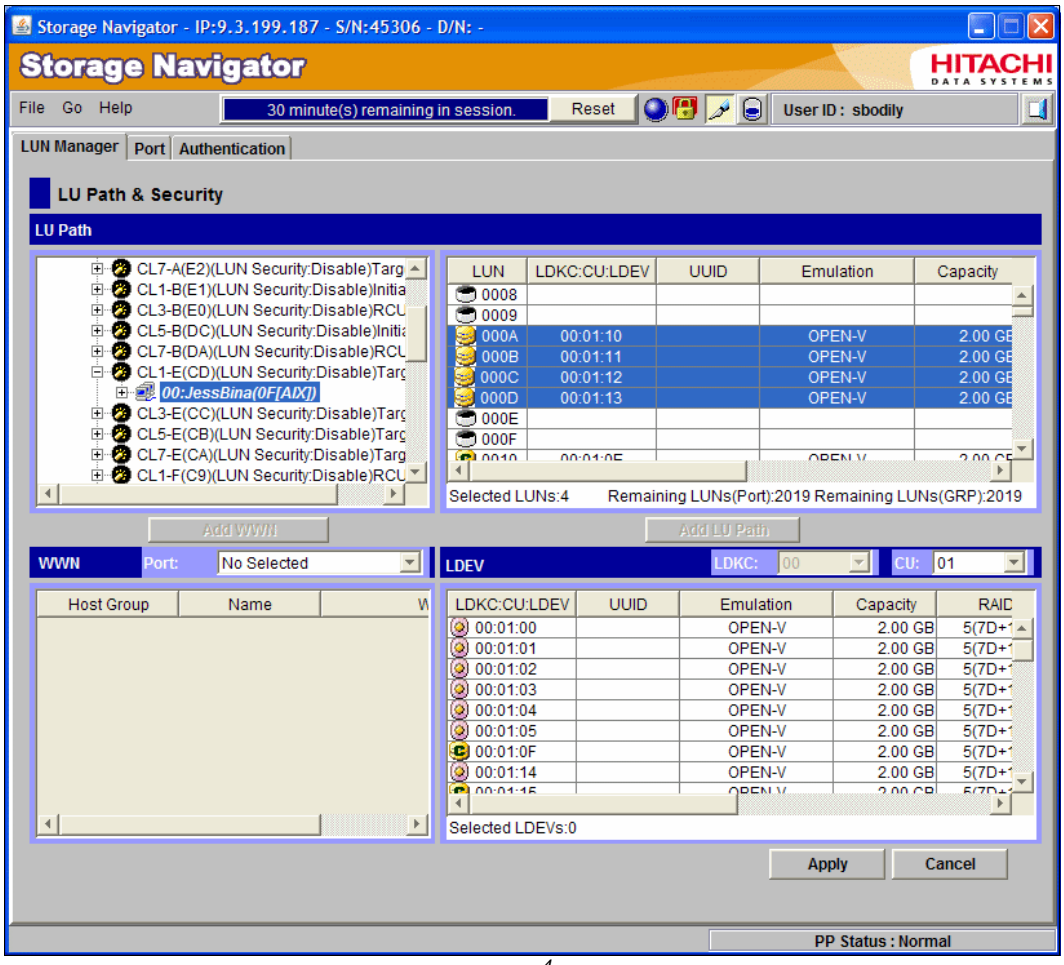


Figure 11-4 Applying LUN assignments for Austin[4]

You have completed assigning four more LUNs for the nodes at the Austin site. However, the lab environment already had several LUNs, including both command and journaling LUNs in the cluster nodes. These LUNs were added solely for this test scenario.

Important: If these LUNs are the first ones to be allocated to the hosts, you must also assign the command LUNs. See the appropriate Hitachi documentation as needed.

For the storage unit at the Miami site, repeat the steps that you performed for the Austin site. The host group, `KrodMaddi`, is assigned to port CL1-B on the Hitachi UPS-VM storage unit with the serial number 35764. Usually the host group is assigned to multiple ports for full multipath redundancy. Figure 11-5 on page 528 shows the result of these steps.

Again, record both the LUN numbers and LDEV numbers so that you can easily refer to them as needed when you create the replicated pairs. The numbers are also required when you add the LUNs into device groups in the appropriate `horcm.conf` file.

[4] Courtesy of Hitachi Data Systems

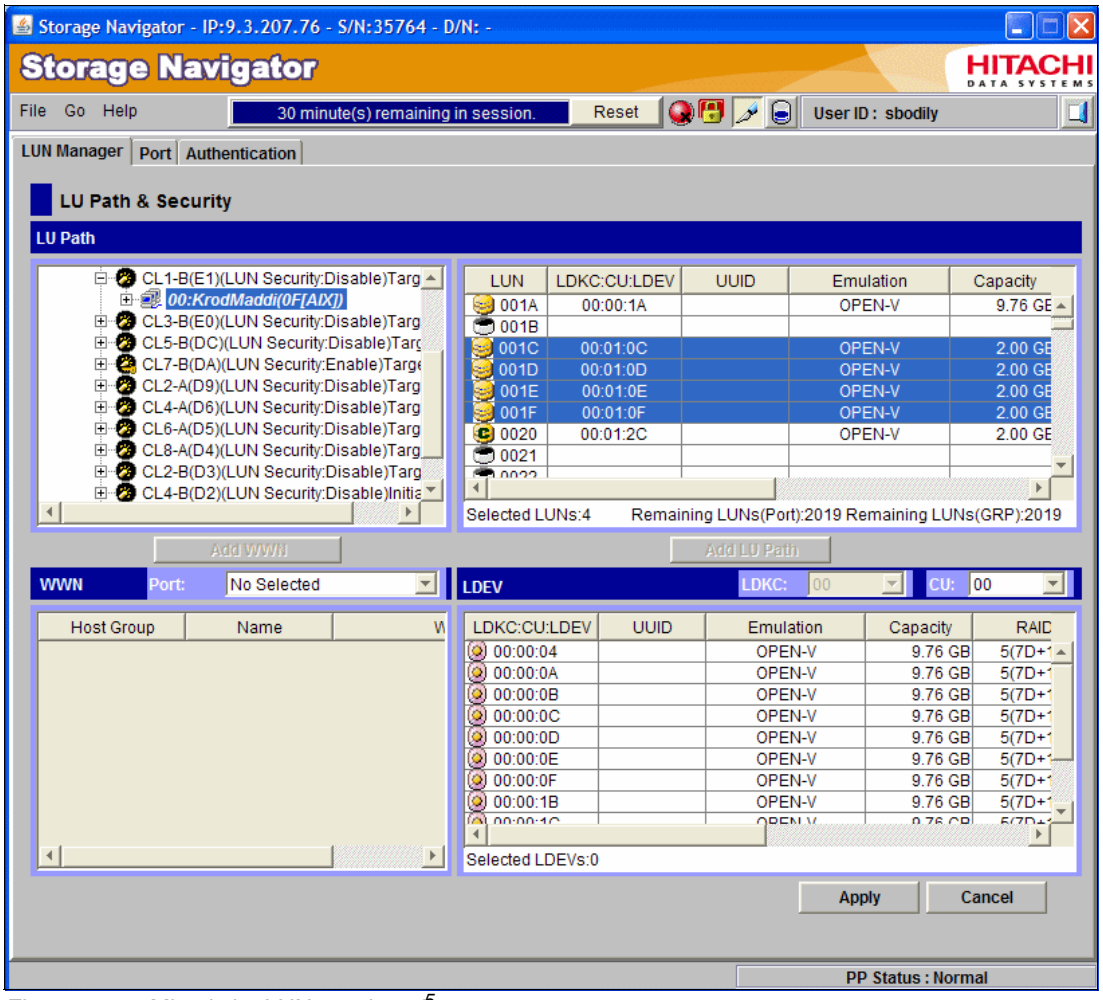


Figure 11-5 Miami site LUNs assigned[5]

11.4.2 Creating replicated pairs

PowerHA SystemMirror Enterprise Edition does not create replication pairs by using the Hitachi interfaces. You must use the Hitachi Storage interfaces to create the same replicated pairs before you use PowerHA SystemMirror Enterprise Edition to achieve an HADR solution. For information about setting up TrueCopy/HUR pairs, see the *Hitachi Command Control Interface (CCI) User and Reference Guide*, MK-90RD011, which you can download from:

`http://communities.vmware.com/servlet/JiveServlet/download/1183307-19474`

You must know which LUNs from each storage unit will be paired together. They must be the same size. In this case, all of the LUNs that are used are 2 GB in size. The pairing of LUNs also uses the LDEV numbers. The LDEV numbers are hexadecimal values that also show up as decimal values on the AIX host.

[5] Courtesy of Hitachi Data Systems

Table 11-1 translates the LDEV hex values of each LUN and its corresponding decimal value.

Table 11-1 LUN number to LDEV number comparison

Austin - 45306			Miami - 35764		
LUN number	**LDEV-HEX**	**LDEV-DEC number**	**LUN number**	**LDEV-HEX**	**LDEV-DEC number**
000A	00:01:10	272	001C	00:01:0C	268
000B	00:01:11	273	001D	00:01:0D	269
000C	00:01:12	274	001E	00:01:0E	271
000D	00:01:13	275	001F	00:01:0E	272

Although the pairing can be done by using the CCI, the example in this section shows how to create the replicated pairs through the Hitachi Storage Navigator. The appropriate commands are in the `/HORCM/usr/bin` directory. In this scenario, none of the devices were configured to the AIX cluster nodes.

Creating TrueCopy synchronous pairings

Beginning with the Austin Hitachi unit, create two synchronous TrueCopy replicated pairings.

1. From within Storage Navigator (Figure 11-6), select **Go** → **TrueCopy** → **Pair Operation**.

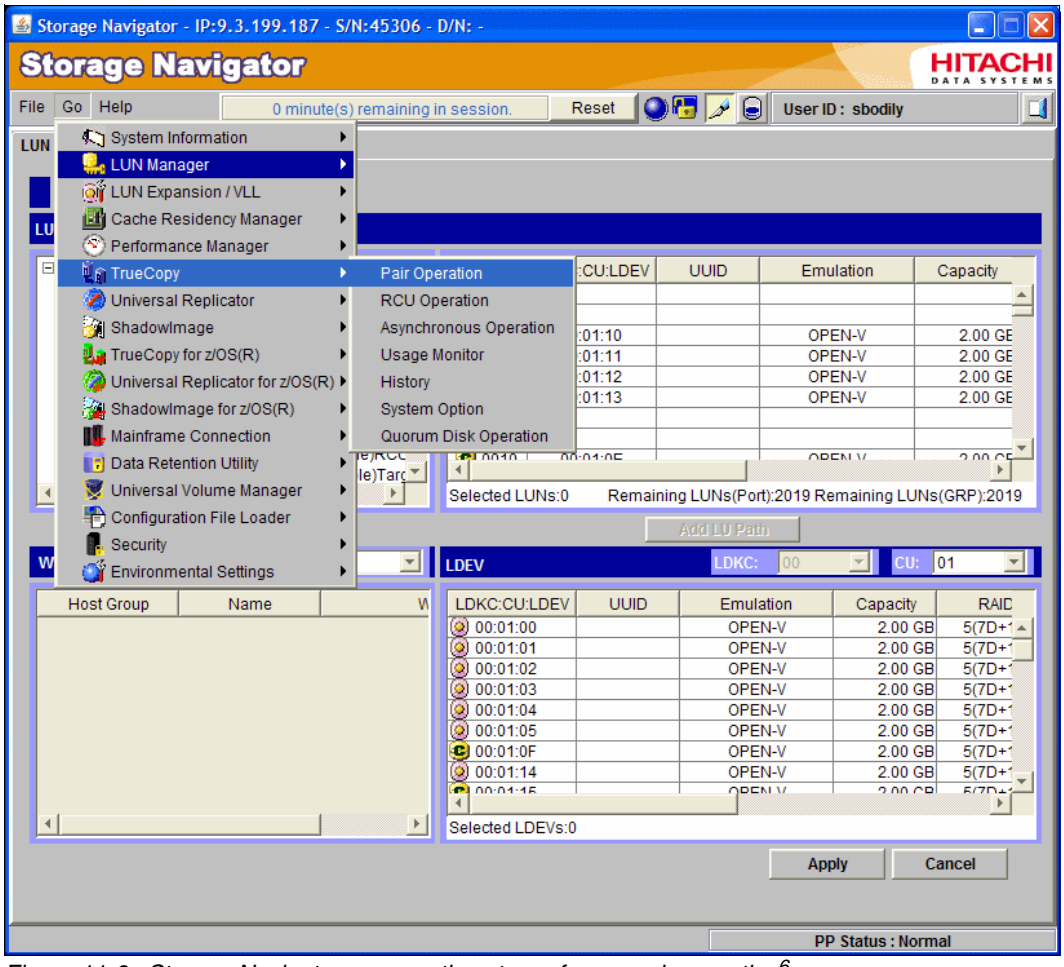


Figure 11-6 Storage Navigator menu options to perform a pair operation[6]

2. In the TrueCopy Pair Operation window (Figure 11-7), select the appropriate port, **CL-1E**, and find the specific LUNs to use (**00-00A** and **00-00B**).

   In this scenario, we predetermined that we want to pair these LUNs with 00-01C and 00-01D from the Miami Hitachi storage unit on port CL1-B. Notice in the occurrence of *SMPL* in the Status column next to the LUNs. SMPL indicates simplex, meaning that no mirroring is being used with that LUN.

3. Right-click the first Austin LUN (**00-00A**), and select **Paircreate** → **Synchronize** (Figure 11-7).

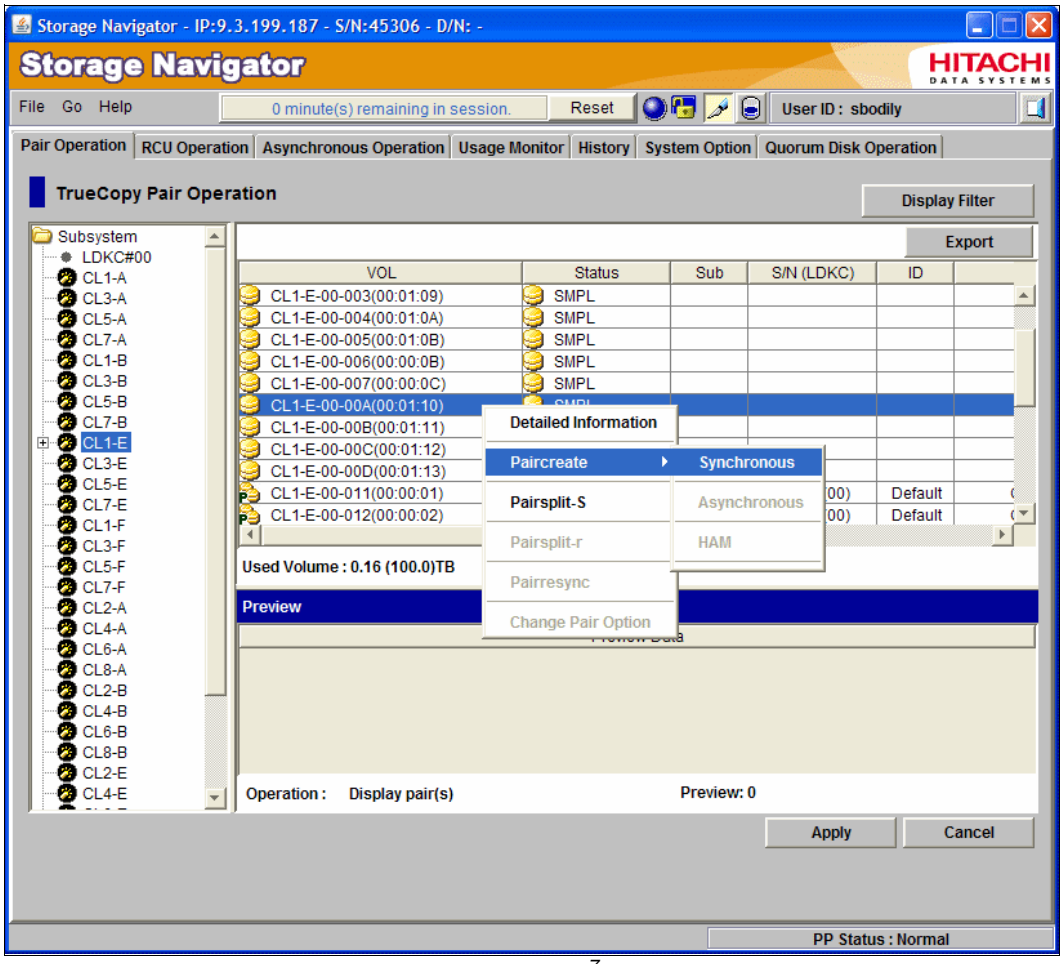


Figure 11-7 Creating a TrueCopy synchronous pairing[7]

[6] Courtesy of Hitachi Data Systems
[7] Courtesy of Hitachi Data Systems

4. In the full synchronous Paircreate menu (Figure 11-8), select the port and LUN that you previously created and recorded. Click **Set**.

   Because we have only one extra remote storage unit, the RCU field already shows the proper one for Miami.

5. Repeat step 4 for the second LUN pairing. Figure 11-8 shows details of the two pairings.

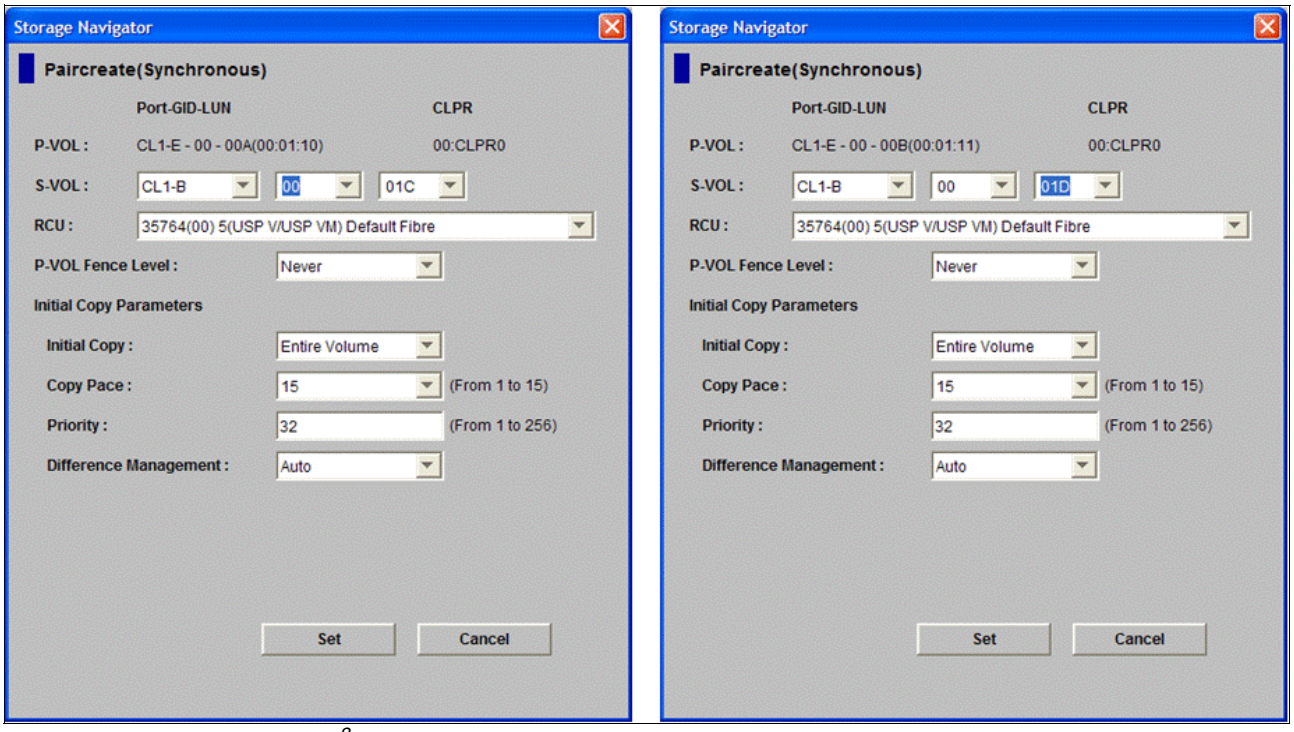


Figure 11-8 TrueCopy pairings[8]

[8] Courtesy of Hitachi Data Systems

6. After you complete the pairing selections, on the **Pair Operation** tab, verify that the information is correct, and click **Apply** to apply them all at one time.

   Figure 11-9 shows both of the source LUNs in the middle of the pane. It also shows an overview of which remote LUNs they are to be paired with.

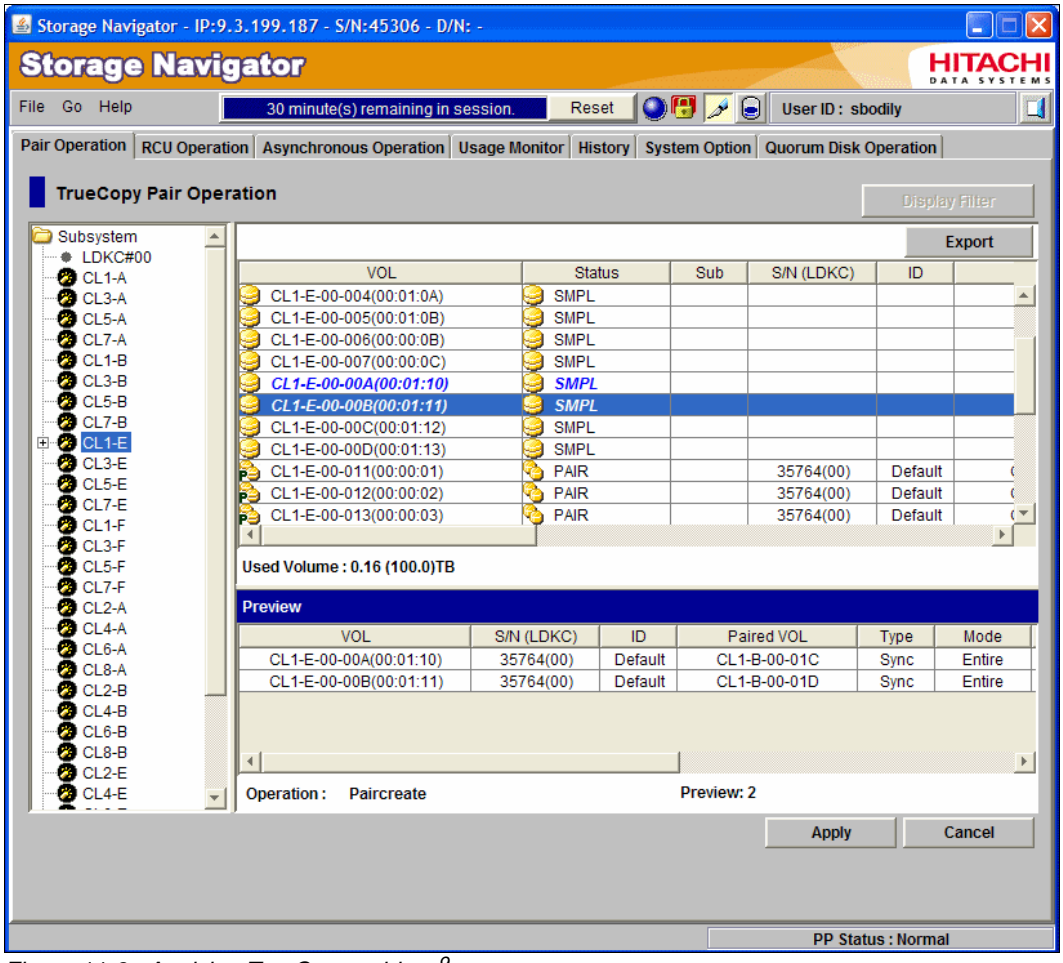


Figure 11-9 Applying TrueCopy pairings[9]

[9] Courtesy of Hitachi Data Systems

This step automatically starts copying the LUNs from the local Austin primary source to the remote Miami secondary source LUNs. You can also right-click a LUN and select **Detailed Information** as shown in Figure 11-10.

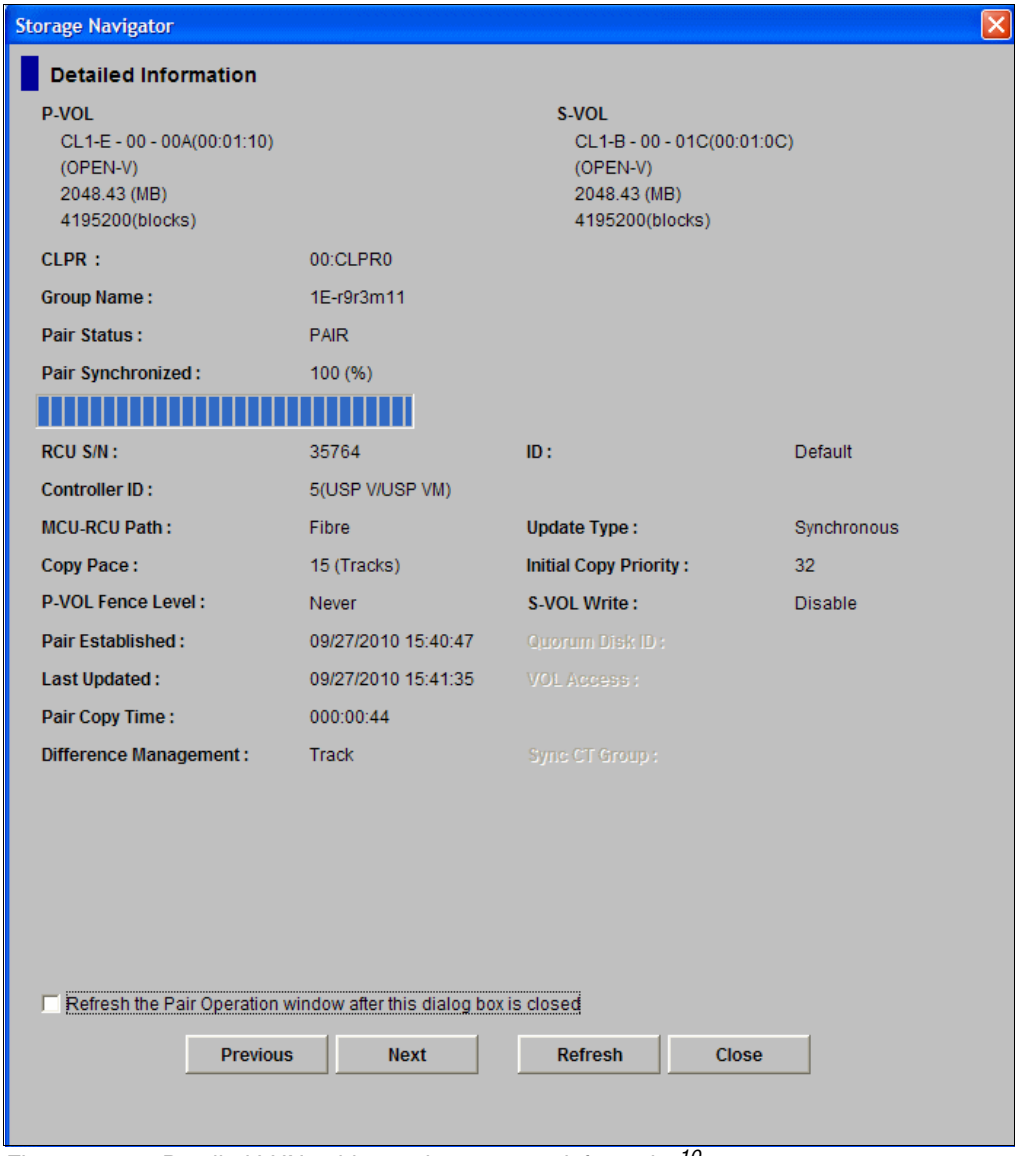


Figure 11-10 Detailed LUN pairing and copy status information[10]

[10] Courtesy of Hitachi Data Systems

After the copy is done, the status is displayed as *PAIR* as shown in Figure 11-11. You can also view this status from the management interface of either one of the storage units.

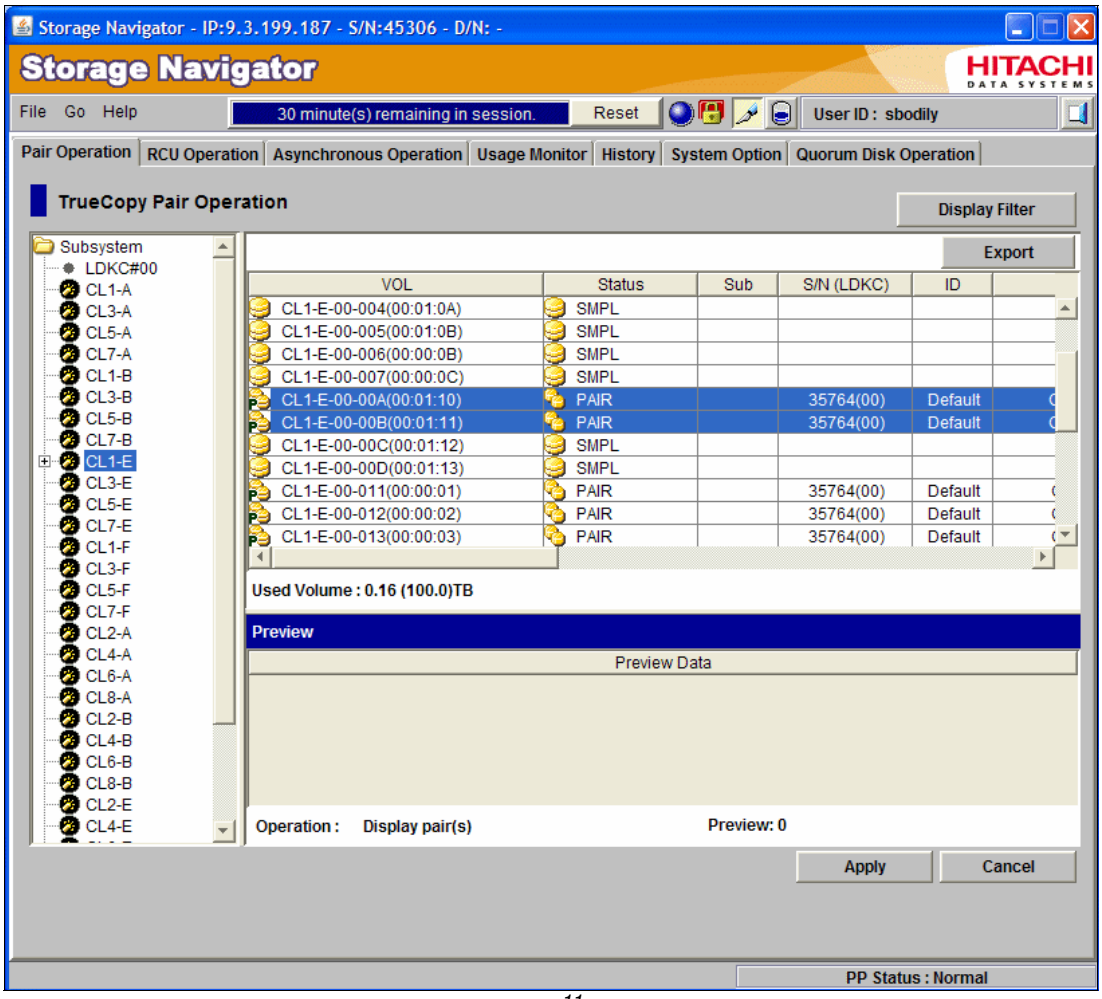


Figure 11-11 TrueCopy pairing and copy completed[11]

[11] Courtesy of Hitachi Data Systems

Creating a Universal Replicator asynchronous pairing

Now switch over to the Miami Hitachi storage unit to create the asynchronous replicated pairings.

1. From the Storage Navigator, select **Go** → **Universal Replicator** → **Pair Operation** (Figure 11-12).

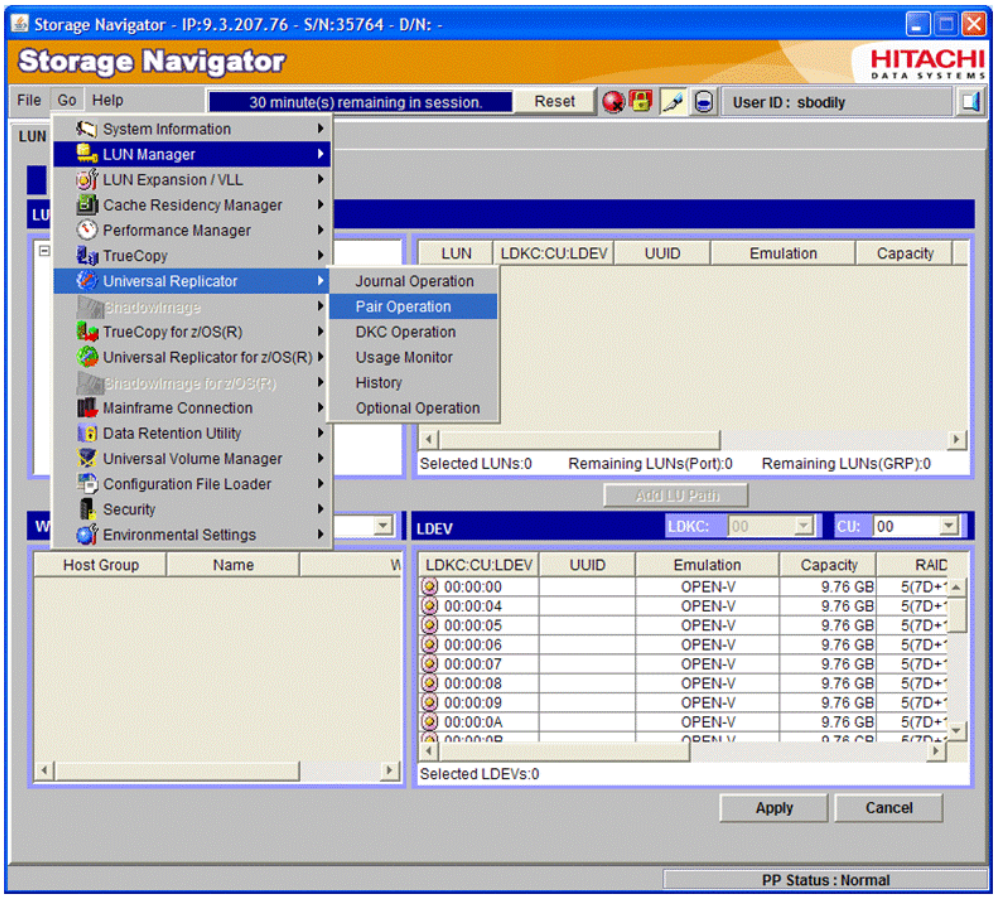


Figure 11-12 Menu selection to perform the pair operation[12]

[12] Courtesy of Hitachi Data Systems

2. In the Universal Replicator Pair Operation window (Figure 11-13), select the appropriate port **CL-1B** and find the specific LUNs that you want to use, which are **00-01E** and **00-01F** in this example). We already predetermined that we want to pair these LUNs with 00-0C and 00-00D from the Austin Hitachi storage unit on port CL1-E.

   Right-click one of the desired LUNs and select **Paircreate**.

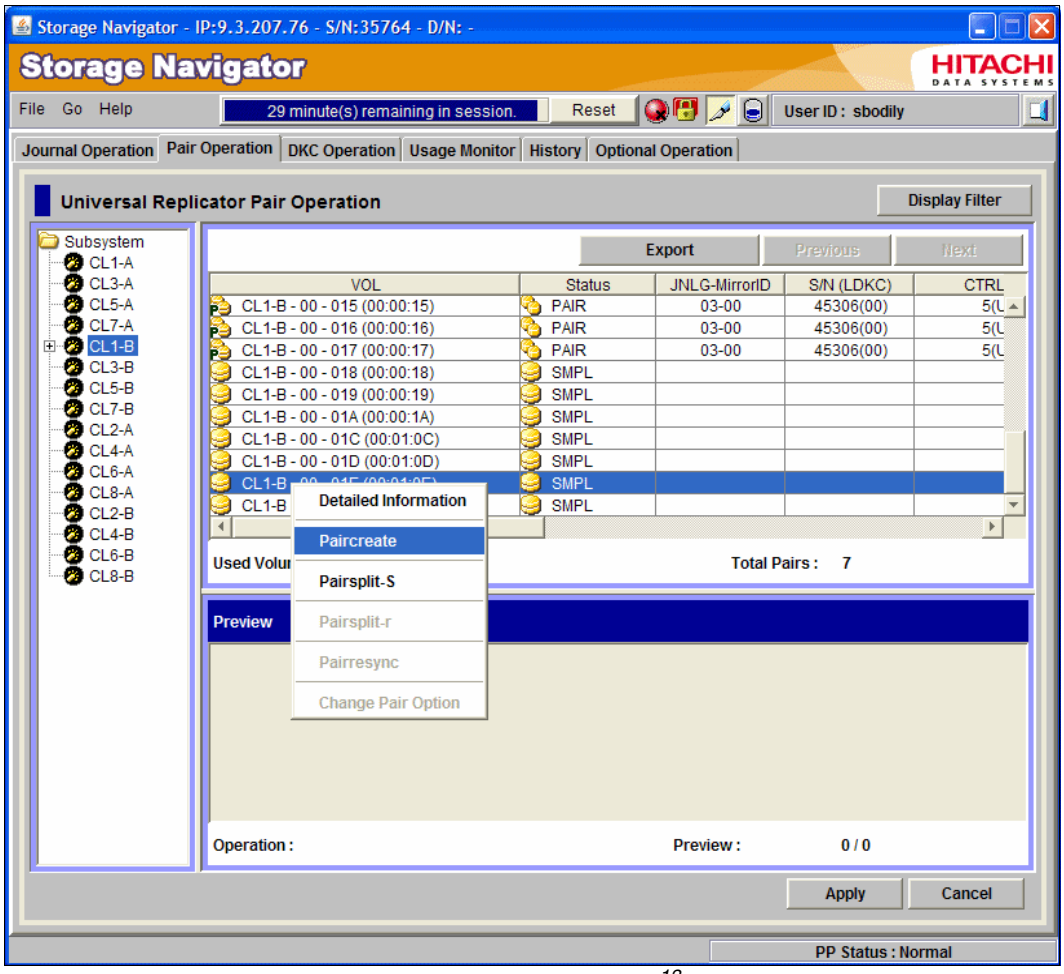


Figure 11-13 Selecting Paircreate in the Universal Replicator[13]

[13] Courtesy of Hitachi Data Systems

3. In the full synchronous Paircreate window, complete these steps:
   a. Select the proper port and LUN that you previously created and recorded.
   b. Because we have only one extra remote storage unit, the RCU field already shows the proper one for Austin.
   c. Unlike when using TrueCopy synchronous replication, when you use Universal Replicator, specify a *master journal volume* (M-JNL), a *remote journal volume* (R-JNL), and a *consistency (CT) group.*

      Important: If these LUNs are the first Universal Replicator LUNs to be allocated, you must also assign journaling groups and LUNs for both storage units. Refer to the appropriate Hitachi Universal Replicator documentation as needed.

      We chose ones that were previously created in the environment.
   d. Click **Set**
   e. Repeat these steps for the second LUN pairing.

   Figure 11-14 shows details of the two pairings.

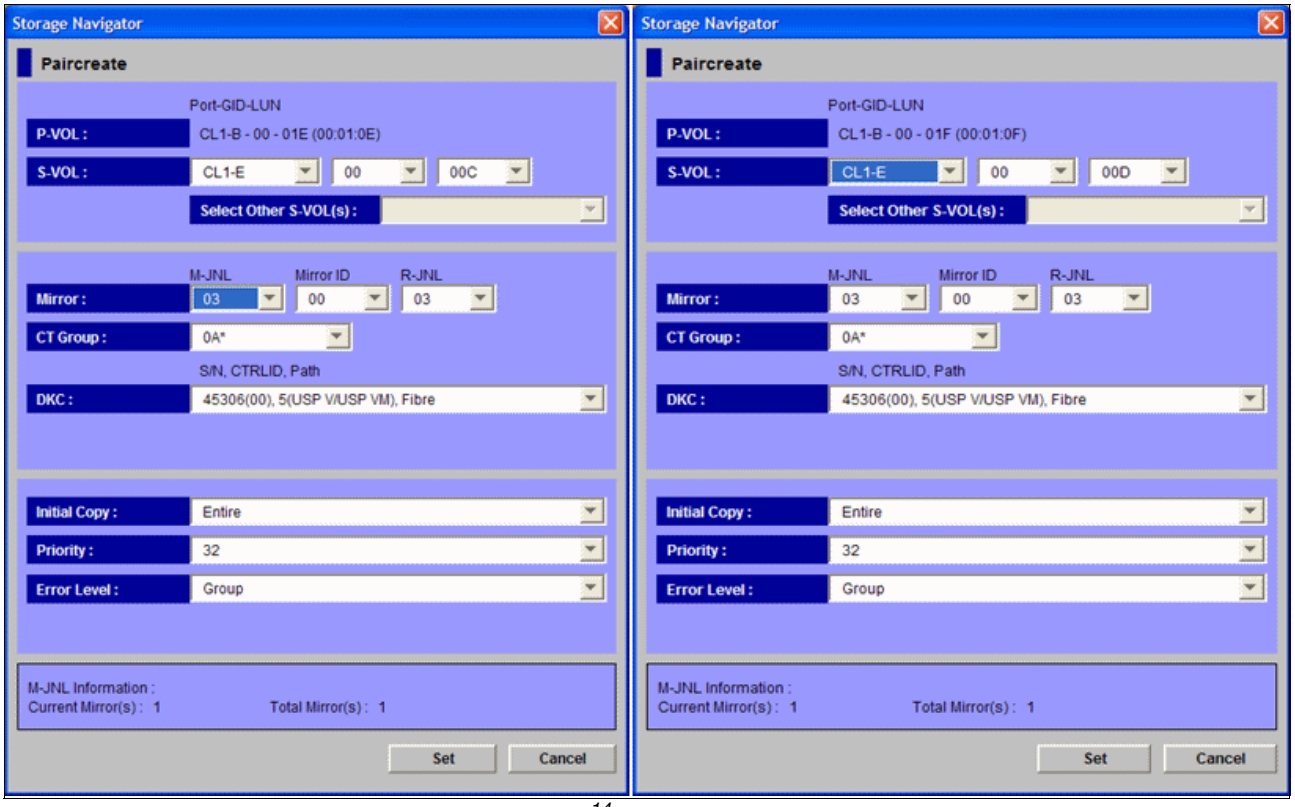


Figure 11-14 Paircreate details in Universal Replicator[14]

[14] Courtesy of Hitachi Data Systems

4. After you complete the pairing selections, on the **Pair Operation** tab, verify that the information is correct and click **Apply** to apply them all at one time.

   When the pairing is established, the copy automatically begins to synchronize with the remote LUNs at the Austin site. The status changes to *COPY,* as shown in Figure 11-15, until the pairs are in sync. After the pairs are synchronized, their status changes to *PAIR*.

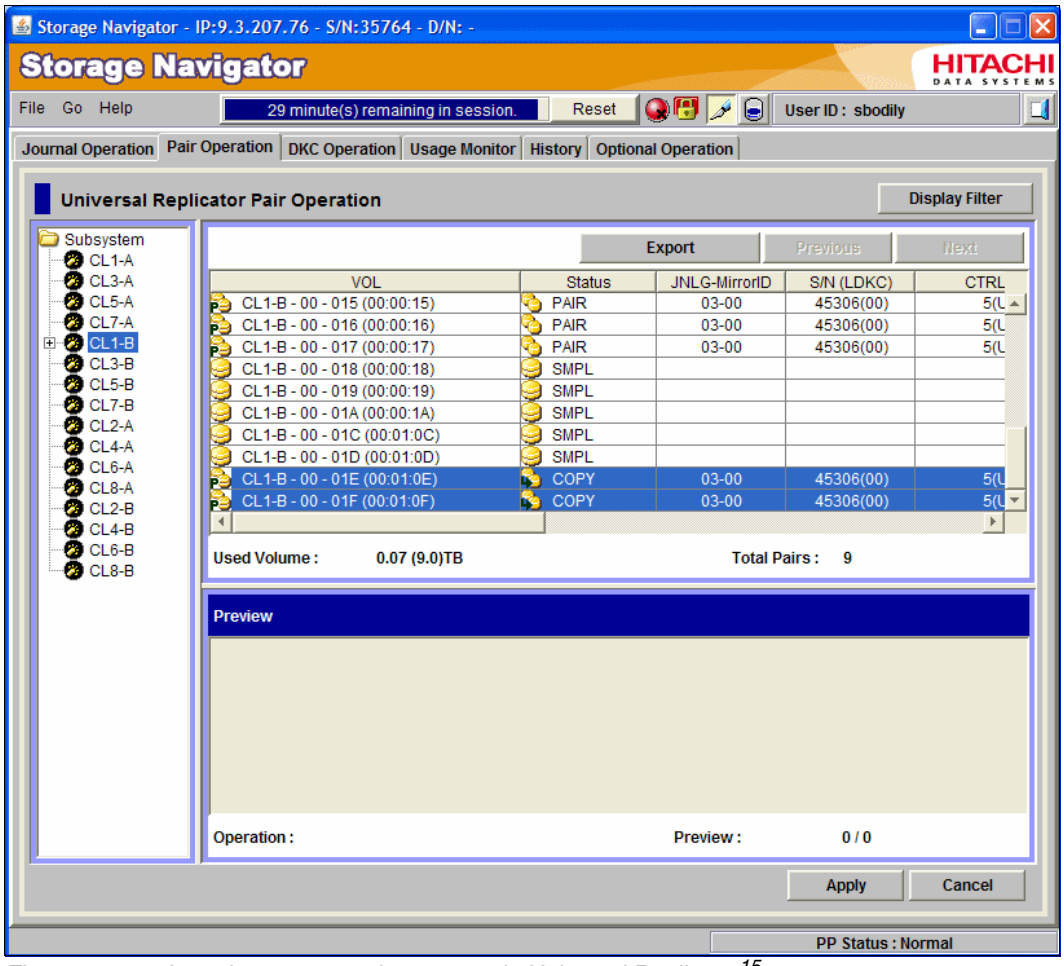


Figure 11-15 Asynchronous copy in progress in Universal Replicator[15]

[15] Courtesy of Hitachi Data Systems

5. Upon completion of the synchronization of the LUNs, configure the LUNs into the AIX cluster nodes. Figure 11-16 shows an overview of the Hitachi replicated environment.

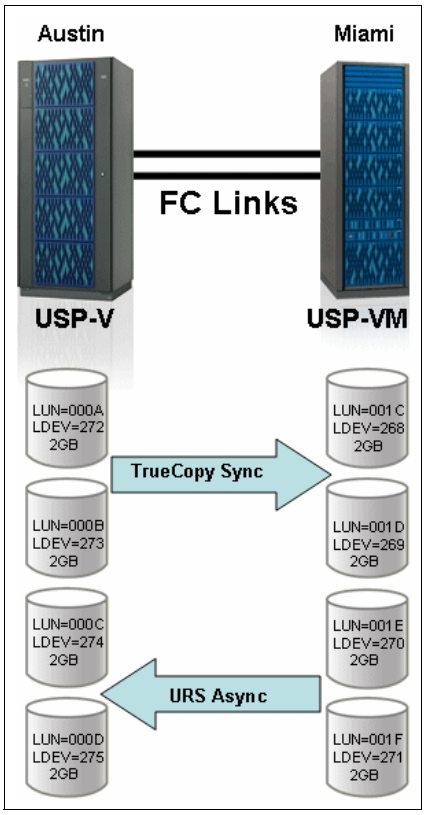


Figure 11-16 Replicated Hitachi LUN overview[16]

11.4.3 Configuring an AIX disk and dev_group association

Before you continue with the steps in this section, you must ensure that the Hitachi hdisks are made available to your nodes. You can run the **cfgmgr** command to configure the new hdisks. Also the CCI must already be installed on each cluster node. If you must install the CCI, see 11.2.1, “Installing the Hitachi CCI software” on page 518.

In the test environment, we already have hdisk0-37 on each of the four cluster nodes. After we run the **cfgmgr** command one each node, one at a time, we now have four more disks, hdisk38-hdisk41, as shown in Example 11-3.

Example 11-3 New Hitachi disks

```
root@jessica:
hdisk38         none                                None
hdisk39         none                                None
hdisk40         none                                None
hdisk41         none                                None
```

Although the LUN and LDEV numbers were written down during the initial LUN assignments, you must identify the correct LDEV numbers of the Hitachi disks and the corresponding AIX hdisks by performing the following steps:

[16] Courtesy of Hitachi Data Systems

1. On the PowerHA SystemMirror Enterprise Edition nodes, select the Hitachi disks and the disks that will be used in the TrueCopy/HUR relationships by running the **inqraid** command. Example 11-4 shows hdisk38-hdisk41, which are the Hitachi disks that we just added.

Example 11-4 Hitachi disks added

```
root@jessica:
# lsdev -Cc disk|grep hdisk|/HORCM/usr/bin/inqraid
hdisk38 -> [SQ] CL1-E Ser =   45306 LDEV = 272 [HITACHI ] [OPEN-V          ]
           HORC = P-VOL  HOMRCF[MU#0 = SMPL MU#1 = SMPL MU#2 = SMPL]
           RAID5[Group  1- 2] SSID = 0x0005
hdisk39 -> [SQ] CL1-E Ser =   45306 LDEV = 273 [HITACHI ] [OPEN-V          ]
           HORC = P-VOL  HOMRCF[MU#0 = SMPL MU#1 = SMPL MU#2 = SMPL]
           RAID5[Group  1- 2] SSID = 0x0005
hdisk40 -> [SQ] CL1-E Ser =   45306 LDEV = 274 [HITACHI ] [OPEN-V          ]
           HORC = S-VOL  HOMRCF[MU#0 = SMPL MU#1 = SMPL MU#2 = SMPL]
           RAID5[Group  1- 2] SSID = 0x0005 CTGID = 10
hdisk41 -> [SQ] CL1-E Ser =   45306 LDEV = 275 [HITACHI ] [OPEN-V          ]
           HORC = S-VOL  HOMRCF[MU#0 = SMPL MU#1 = SMPL MU#2 = SMPL]
           RAID5[Group  1- 2] SSID = 0x0005 CTGID = 10
```

2. Edit the HORCM LDEV section in the `horcm#.conf` file to identify the dev_group that will be managed by PowerHA SystemMirror Enterprise Edition. In this example, we use the `horcm2.conf` file.

   `Hdisk38` (ldev 272) and `hdisk39` (ldev 273) are the pair for the synchronous replicated resource group, which is primary at the Austin site. `Hdisk40` (ldev 275) and `hdisk41` (ldev276) are the pair for an asynchronous replicated resource, which is primary at the Miami site.

   Specify the device groups (dev_group) in the `horcm#.conf` file. We are using dev_group `htcdg01` with dev_names `htcd01` and `htcd02` for the synchronous replicated pairs. For the asynchronous pairs, we are using dev_group `hurdg01` and dev_names `hurd01` and `hurd02`. The device group names are needed later when you check that status of the replicated pairs and when you define the replicated pairs as a resource for PowerHA Enterprise Edition to control.

   Important: Do *not* edit the configuration definition file while HORCM is running. Shut down HORCM, edit the configuration file as needed, and then restart HORCM.

   Example 11-5 shows the `horcm2.conf` file from the `jessica` node, at the Austin site. Because two nodes are at the Austin site, the same updates were performed to the `/etc/horcm2.conf` file on the `bina` node. Notice that you can use either the decimal value of the LDEV or the hexadecimal value.

   We specifically did one pair each way just to show it and to demonstrate that it works. Although several groups were already defined, only those groups that are relevant to this scenario are shown.

Example 11-5 Horcm2.conf file used for the Austin site nodes

```
root@jessica:
/etc/horcm2.conf

HORCM_MON
#Address of local node...
#ip_address             service     poll(10ms)     timeout(10ms)
```

```
r9r3m11.austin.ibm.com  52323       1000           3000

HORCM_CMD
#hdisk of Command Device...
#dev_name               dev_name                dev_name
#UnitID 0 (Serial# 45306)
#/dev/rhdisk10
\\.\CMD-45306:/dev/rhdisk10 /dev/rhdisk14

HORCM_LDEV
#Map dev_grp to LDEV#...
#dev_group   dev_name    Serial#   CU:LDEV     MU#     siteA      siteB
#                                   (LDEV#)            hdisk   -> hdisk
#---------   ---------   -------   --------    ---   --------------------
htcdg01      htcd01       45306     272
htcdg01      htcd02       45306     273
hurdg01      hurd01       45306     01:12
hurdg01      hurd02       45306     01:13

# Address of remote node for each dev_grp...
HORCM_INST
#dev_group    ip_address service
htcdg01       maddi.austin.ibm.com  52323
hurdg01       maddi.austin.ibm.com  52323
```

For the `krod` and `maddi` nodes at the Miami site, the dev_groups, dev_names, and the LDEV numbers are the same. The difference is the specific serial number of the storage unit at that site. Also, the remote system or IP address for the appropriate system in the Austin site.

Example 11-6 shows the `horcm2.conf` file that we used for both nodes in the Miami site. Notice that, for the *ip_address* fields, fully qualified names are used instead of the IP address. If these names are resolvable, the format is still valid. However, the format uses the actual addresses as shown in Example 11-1 on page 521.

Example 11-6 The horcm2.conf file that is used for the nodes in the Miami site

```
root@krod:
horcm2.conf

HORCM_MON
#Address of local node...
#ip_address             service     poll(10ms)     timeout(10ms)
r9r3m13.austin.ibm.com  52323       1000           3000

HORCM_CMD
#hdisk of Command Device...
#dev_name               dev_name                dev_name
#UnitID 0 (Serial# 35764)
#/dev/rhdisk10
# /dev/hdisk19
\\.\CMD-45306:/dev/rhdisk11 /dev/rhdisk19

#HUR_GROUP    HUR_103_153  45306     01:53       0
htcdg01      htcd01       35764     268
htcdg01      htcd02       35764     269
hurdg01      hurd01       35764     01:0E
```

```
hurdg01      hurd02       35764     01:0F
# Address of remote node for each dev_grp...

HORCM_INST
#dev_group    ip_address service
htcdg01       bina.austin.ibm.com  52323
hurdg01       bina.austin.ibm.com  52323
```

3. Map the hdisks that are protected by TrueCopy to the TrueCopy device groups by using the **raidscan** command. In the following example, 2 is the HORCM instance number:

   ```
   lsdev -Cc disk|grep hdisk | /HORCM/usr/bin/raidscan -IH2 -find inst
   ```

   The **-find inst** option of the **raidscan** command registers the device file name (hdisk) to all mirror descriptors of the LDEV map table for HORCM. This option also allows the matching volumes on the `horcm.conf` file in protection mode and is started automatically by using the **/etc/horcmgr** command. Therefore, you do not need to use this option normally. This option is ended to avoid wasteful scanning when the registration is finished based on HORCM.

   Therefore, if HORCM no longer needs the registration, no further action is taken and it exits. You can use the **-find inst** option with the **-fx** option to view LDEV numbers in the hexadecimal format.

4. Verify that the PAIRs are established by running either the **pairvdisplay** command or the **pairvolchk** command against the device groups `htcdg01` and `hurdg01`.

   Example 11-7 shows how we use the **pairvdisplay** command. For device group `htcdg01`, the status of PAIR and fence of NEVER indicates that they are a synchronous pair. For device group `hurdg01`, the ASYNC fence option clearly indicates that it is in an asynchronous pair. Also, notice that the CTG field shows the consistency group number for the asynchronous pair that is managed by HUR.

Example 11-7 The pairdisplay command to verify that the pair status is synchronized

```
# pairdisplay -g htcdg01 -IH2 -fe
Group   PairVol(L/R) (Port#,TID, LU),Seq#,LDEV#.P/S,Status,Fence,Seq#,P-LDEV# M CTG JID AP
htcdg01 htcd01(L)   (CL1-E-0, 0, 10)45306   272.P-VOL PAIR NEVER ,35764   268 -   -   -  1
htcdg01 htcd01(R)   (CL1-B-0, 0, 28)35764   268.S-VOL PAIR NEVER ,-----   272 -   -   -  -
htcdg01 htcd02(L)   (CL1-E-0, 0, 11)45306   273.P-VOL PAIR NEVER ,35764   269 -   -   -  1
htcdg01 htcd02(R)   (CL1-B-0, 0, 29)35764   269.S-VOL PAIR NEVER ,-----   273 -   -   -  -

# pairdisplay -g hurdg01 -IH2 -fe
Group   PairVol(L/R) (Port#,TID, LU),Seq#,LDEV#.P/S,Status,Fence,Seq#,P-LDEV# M CTG JID AP
hurdg01 hurd01(L)   (CL1-E-0, 0, 12)45306   274.S-VOL PAIR ASYNC ,-----   270 -  10   3  1
hurdg01 hurd01(R)   (CL1-B-0, 0, 30)35764   270.P-VOL PAIR ASYNC ,45306   274 -  10   3  2
hurdg01 hurd02(L)   (CL1-E-0, 0, 13)45306   275.S-VOL PAIR ASYNC ,-----   271 -  10   3  1
hurdg01 hurd02(R)   (CL1-B-0, 0, 31)35764   271.P-VOL PAIR ASYNC ,45306   275 -  10   3  2
```

To show the output in Example 11-7, we removed the last three columns of the output because it was not relevant to what we are checking.

Unestablished pairs: If pairs are not yet established, the status is displayed as *SMPL*. To continue, you must create the pairs. For instructions about creating pairs from the command line, see the *Hitachi Command Control Interface (CCI) User and Reference Guide*, MK-90RD011, which you can download from:

http://communities.vmware.com/servlet/JiveServlet/download/1183307-19474

Otherwise, if you are using Storage Navigator, see 11.4.2, “Creating replicated pairs” on page 528.

Creating volume groups and file systems on replicated disks

After you identify the hdisks and dev_groups that will be managed by PowerHA SystemMirror Enterprise Edition, you must create the volume groups and file systems. To set up volume groups and file systems in the replicated disks, follow these steps:

1. On each of the four PowerHA SystemMirror Enterprise Edition cluster nodes, verify the next free major number by running the **`lvlstmajor`** command on each cluster node. Also, verify that the physical volume name for the file system can also be used across sites.

   In this scenario, we use the major numbers 56 for the `truesyncvg` volume group and 57 for the `ursasyncvg` volume group. We use these numbers later when we import the volume to the other cluster nodes. Although the major numbers are not required to match, it is a preferred practice.

   We create the `truesyncvg` scalable volume group on the `jessica` node where the primary LUNs are located. We also create the logical volumes, `jfslog`, and file systems as shown in Example 11-8.

Example 11-8 Details about the truesyncvg volume group

```
root@jessica:lsvg truesyncvg
VOLUME GROUP:       truesyncvg              VG IDENTIFIER:
00cb14ce00004c000000012b564c41b9
VG STATE:           active                  PP SIZE:        4 megabyte(s)
VG PERMISSION:      read/write              TOTAL PPs:      988 (3952 megabytes)
MAX LVs:            256                     FREE PPs:       737 (2948 megabytes)
LVs:                3                       USED PPs:       251 (1004 megabytes)
OPEN LVs:           3                       QUORUM:         2 (Enabled)
TOTAL PVs:          2                       VG DESCRIPTORS: 3
STALE PVs:          0                       STALE PPs:      0
ACTIVE PVs:         2                       AUTO ON:        no
MAX PPs per VG:     32768                   MAX PVs:        1024
LTG size (Dynamic): 256 kilobyte(s)         AUTO SYNC:      no
HOT SPARE:          no                      BB POLICY:      relocatable
PV RESTRICTION:     none

root@jessica:lsvg -l truesyncvg
lsvg -l truesyncvg
truesyncvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
oreolv              jfs2       125     125     1    closed/syncd  /oreofs
majorlv             jfs2       125     125     1    closed/syncd  /majorfs
truefsloglv         jfs2log    1       1       1    closed/syncd  N/A
```

We create the `ursasyncvg` big volume group on the `krod` node where the primary LUNs are located. We also create the logical volumes, `jfslog`, and file systems as shown in Example 11-9.

Example 11-9 Ursasyncvg volume group information

```
root@krod:lspv
hdisk40         00cb14ce5676ad24                    ursasyncvg      active
hdisk41         00cb14ce5676afcf                    ursasyncvg      active

root@krod:lsvg ursasyncvg
VOLUME GROUP:       ursasyncvg              VG IDENTIFIER:
00cb14ce00004c000000012b5676b11e
VG STATE:           active                  PP SIZE:        4 megabyte(s)
VG PERMISSION:      read/write              TOTAL PPs:      1018 (4072 megabytes)
MAX LVs:            512                     FREE PPs:       596 (2384 megabytes)
LVs:                3                       USED PPs:       422 (1688 megabytes)
OPEN LVs:           3                       QUORUM:         2 (Enabled)
TOTAL PVs:          2                       VG DESCRIPTORS: 3
STALE PVs:          0                       STALE PPs:      0
ACTIVE PVs:         2                       AUTO ON:        no
MAX PPs per VG:     130048
MAX PPs per PV:     1016                    MAX PVs:        128
LTG size (Dynamic): 256 kilobyte(s)         AUTO SYNC:      no
HOT SPARE:          no                      BB POLICY:      relocatable

root@krod:lsvg -l ursasyncvg
ursasyncvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
ursfsloglv          jfs2log    2       2       1    closed/syncd  N/A
hannahlv            jfs2       200     200     1    closed/syncd  /hannahfs
julielv             jfs2       220     220     1    closed/syncd  /juliefs
```

2. Vary off the newly created volume groups by running the **varyoffvg** command. To import the volume groups onto the other three systems, the pairs must be in sync.

   We run the **pairresync** command as shown in Example 11-10 on the local disks and make sure that they are in the PAIR state. This process verifies that the local disk information was copied to the remote storage. Notice that the command is being run on the respective node that contains the primary source LUNs and where the volume groups are created.

   Example 11-10 Pairresync command

   ```
   #root@jessica:pairresync -g htcdg01 -IH2

   #root@krod:pairresync -g hurdg01 -IH2
   ```

   Verify that the pairs are in sync with the **pairdisplay** command as shown in Example 11-7 on page 542.

3. Split the pair relationship so that the remote systems can import the volume groups as needed on each node. Run the **pairsplit** command against the device group as shown in Example 11-11.

Example 11-11 The pairsplit command to suspend replication

```
root@jessica: pairsplit -g htcdg01 -IH2

root@krod: pairsplit -g hurdg01 -IH2
```

To verify that the pairs are split, check the status by using the **pairdisplay** command. Example 11-12 shows that the pairs are in a suspended state.

Example 11-12 Pairdisplay shows pairs that are suspended

```
root@jessica: pairdisplay -g htcdg01 -IH2 -fe
Group   PairVol(L/R) (Port#,TID, LU),Seq#,LDEV#.P/S,Status,Fence,Seq#,P-LDEV# M CTG JID AP
htcdg01 htcd01(L)   (CL1-E-0, 0,  10)45306   272.P-VOL PSUS NEVER ,35764   268 -   -   -  1
htcdg01 htcd01(R)   (CL1-B-0, 0,  28)35764   268.S-VOL SSUS NEVER ,-----   272 -   -   -  -
htcdg01 htcd02(L)   (CL1-E-0, 0,  11)45306   273.P-VOL PSUS NEVER ,35764   269 -   -   -  1
htcdg01 htcd02(R)   (CL1-B-0, 0,  29)35764   269.S-VOL SSUS NEVER ,-----   273 -   -   -  -

root@krod: pairdisplay -g hurdg01 -IH2 -fe
Group   PairVol(L/R) (Port#,TID, LU),Seq#,LDEV#.P/S,Status,Fence,Seq#,P-LDEV# M CTG JID AP
hurdg01 hurd01(L)   (CL1-B-0, 0,  30)35764   270.P-VOL PSUS ASYNC ,45306   274 -  10   3  2
hurdg01 hurd01(R)   (CL1-E-0, 0,  12)45306   274.S-VOL SSUS ASYNC ,-----   270 -  10   3  1
hurdg01 hurd02(L)   (CL1-B-0, 0,  31)35764   271.P-VOL PSUS ASYNC ,45306   275 -  10   3  2
hurdg01 hurd02(R)   (CL1-E-0, 0,  13)45306   275.S-VOL SSUS ASYNC ,-----   271 -  10   3  1
```

4. To import the volume groups on the remaining nodes, ensure that the PVID is present on the disks by using one of the following options:
   - Run the **rmdev -dl** command for each hdisk and then run the **cfgmgr** command.
   - Run the appropriate **chdev** command against each disk to pull in the PVID.

   As shown in Example 11-13, we use the **chdev** command on each of the three additional nodes.

Example 11-13 The chdev command to acquire the PVIDs

```
root@jessica: chdev -l hdisk40 -a pv=yes
root@jessica: chdev -l hdisk41 -a pv=yes

root@bina: chdev -l hdisk38 -a pv=yes
root@bina: chdev -l hdisk39 -a pv=yes
root@bina: chdev -l hdisk40 -a pv=yes
root@bina: chdev -l hdisk41 -a pv=yes

root@krod: chdev -l hdisk38 -a pv=yes
root@krod: chdev -l hdisk39 -a pv=yes

root@maddi: chdev -l hdisk38 -a pv=yes
root@maddi: chdev -l hdisk39 -a pv=yes
root@maddi: chdev -l hdisk40 -a pv=yes
root@maddi: chdev -l hdisk41 -a pv=yes
```

5. Verify that the PVIDs are correctly showing on each system by running the **lspv** command as shown in Example 11-14. Because all four of the nodes have the exact hdisk numbering, we show the output only from one node, the `bina` node.

Example 11-14 LSPV listing to verify PVIDs are present

```
bina@root: lspv
hdisk38         00cb14ce564c3f44                    none
hdisk39         00cb14ce564c40fb                    none
hdisk40         00cb14ce5676ad24                    none
hdisk41         00cb14ce5676afcf                    none
```

6. Import the volume groups on each node as needed by using the **importvg** command. Specify the major number that you used earlier.
7. Disable both the auto varyon and quorum settings of the volume groups by using the **chvg** command.
8. Vary off the volume group as shown in Example 11-15.

Attention: PowerHA SystemMirror Enterprise Edition attempts to automatically set the AUTO VARYON to NO during verification, except in the case of remote TrueCopy/HUR.

Example 11-15 Importing the replicated volume groups

```
root@jessica: importvg -y ursasyncvg -V 57 hdisk40
root@jessica: chvg -a n -Q n ursasyncvg
root@jessica: varyoffvg ursasyncvg

root@bina: importvg -y truesyncvg -V 56 hdisk38
root@bina: importvg -y ursasyncvg -V 57 hdisk40
root@bina: chvg -a n -Q n truesyncvg
root@bina: chvg -a n -Q n ursasyncvg
root@bina: varyoffvg truesyncvg
root@bina: varyoffvg ursasyncvg

root@krod: importvg -y truesyncvg -V 56 hdisk38
root@krod: chvg -a n -Q n truesyncvg
root@krod: varyoffvg truesyncvg

root@maddi: importvg -y truesyncvg -V 56 hdisk38
root@maddi: importvg -y ursasyncvg -V 57 hdisk40
root@maddi: chvg -a n -Q n truesyncvg
root@maddi: chvg -a n -Q n ursasyncvg
root@maddi: varyoffvg truesyncvg
root@maddi: varyoffvg ursasyncvg
```

9. Re-establish the pairs that you split in step 3 on page 545 by running the **pairresync** command again as shown in Example 11-10 on page 544.
10. Verify again if they are in sync by using the **pairdisplay** command as shown in Example 11-7 on page 542.

11.4.4 Defining TrueCopy/HUR managed replicated resource to PowerHA

To add a replicated resource to be controlled by PowerHA consists of two specific steps per device group, and four steps overall:

- Adding TrueCopy/HUR replicated resources
- Adding the TrueCopy/HUR replicated resources to a resource group
- Verifying the TrueCopy/HUR configuration
- Synchronizing the cluster configuration

In these steps, the cluster topology was configured, including all four nodes, both sites, and networks.

Adding TrueCopy/HUR replicated resources

To define a TrueCopy replicated resource:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **Extended Configuration** → **Extended Resource Configuration** → **TrueCopy Replicated Resources** → **Add Hitachi TrueCopy/HUR Replicated Resource**.
3. In the Ad Hitachi TrueCopy/HUR Replication Resource panel, press Enter.
4. Complete the available fields appropriately and press Enter.

In this configuration, we created two replicated resources. One resource is for the synchronous device group, htcdg01, named *trulee.* The second resource is for the asynchronous device group, hurdg01, named *ursasyncRR*. Figure 11-17 shows both of the replicated resources.

```
               Add a HITACHI TRUECOPY(R)/HUR Replicated Resource

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

                                                    [Entry Fields]
* TRUECOPY(R)/HUR Resource Name                    [truelee]
* TRUECOPY(R)/HUR Mode                             SYNC                  +
* Device Groups                                   [htcdg01]              +
* Recovery Action                                  AUTO                  +
* Horcm Instance                                  [horcm2]
* Horctakeover Timeout Value                      [300]                  #
* Pairevtwait Timeout Value                       [3600]                 #

               Add a HITACHI TRUECOPY(R)/HUR Replicated Resource

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

                                                    [Entry Fields]
* TRUECOPY(R)/HUR Resource Name                    [ursasyncRR]
* TRUECOPY(R)/HUR Mode                             ASYNC                 +
* Device Groups                                   [hurdg01]              +
* Recovery Action                                  AUTO                  +
* Horcm Instance                                  [horcm2]
* Horctakeover Timeout Value                      [300]                  #
* Pairevtwait Timeout Value                       [3600]                 #
```

Figure 11-17 TrueCopy/HUR replicated resource definitions

For a complete list of all of defined TrueCopy/HUR replicated resources, run the **cllstc** command, which is in the /usr/es/sbin/cluster/tc/cmds directory. Example 11-16 shows the output of the **cllstc** command.

Example 11-16 The cllstc command to list the TrueCopy/HUR replicated resources

```
root@jessica: cllstc -a
Name      CopyMode DeviceGrps   RecoveryAction  HorcmInstance  HorcTimeOut   PairevtTimeout
truelee       SYNC     htcdg01      AUTO            horcm2          300           3600
ursasyncRR    ASYNC    hurdg01      AUTO            horcm2          300           3600
```

Adding the TrueCopy/HUR replicated resources to a resource group

To add a TrueCopy replicated resource to a resource group, follow these steps:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration** → **Extended Resource Configuration** → **Extended Resource Group Configuration**.

   Depending on whether you are working with an existing resource group or creating a resource group, the TrueCopy Replicated Resources entry is displayed at the bottom of the page in SMIT. This entry is a pick list that shows the resource names that are created in the previous task.
3. Ensure that the volume groups that are selected on the Resource Group configuration display match the volume groups that are used in the TrueCopy/HUR Replicated Resource:
   - If you are changing an existing resource group, select **Change/Show Resource Group**.
   - If you are adding a resource group, select **Add a Resource Group**.
4. In the TrueCopy Replicated Resources field, press F4 for a list of the TrueCopy/HUR replicated resources that were previously added. Verify that this resource matches the volume group that is specified.

   Important: You cannot mix regular (non-replicated) volume groups and TrueCopy/HUR replicated volume groups in the same resource group.

   Press Enter.

In this scenario, we changed an existing resource group, emlecRG, for the Austin site and specifically chose a site relationship, also known as an Inter-site Management Policy of *Prefer Primary Site*. We added a resource group, valhallarg, for the Miami site and chose to use the same site relationship. We also added the additional nodes from each site. We configured both to failover locally within a site and failover between sites. If a site failure occurs, the node falls over to the remote site standby node, but never to the remote production node.

Example 11-17 shows the relevant resource group information.

Example 11-17 Resource groups for the TrueCopy/HUR replicated resources

```
Resource Group Name                                 emlecRG
Participating Node Name(s)                          jessica bina maddi
Startup Policy                                      Online On Home Node Only
Fallover Policy                                     Fallover To Next Priority Node
Fallback Policy                                     Never Fallback
Site Relationship                                   Prefer Primary Site
Node Priority
```

```
Service IP Label                                service_1
Volume Groups                                   truesyncvg
Hitachi TrueCopy Replicated Resources           truelee

Resource Group Name                             valhallaRG
Participating Node Name(s)                      krod maddi bina
Startup Policy                                  Online On Home Node Only
Fallover Policy                                 Fallover To Next Priority Node
Fallback Policy                                 Never Fallback
Site Relationship                               Prefer Primary Site
Node Priority
Service IP Label                                service_2
Volume Groups                                   ursasyncvg
Hitachi TrueCopy Replicated Resources           ursasyncRR
```

Verifying the TrueCopy/HUR configuration

Before you synchronize the new cluster configuration, verify the TrueCopy/HUR configuration:

1. To verify the configuration, run the following command:

   ```
   /usr/es/sbin/cluster/tc/utils/cl_verify_tc_config
   ```

2. Correct any configuration errors that are shown.

   If you see error messages such as those shown in Figure 11-18, usually these types of messages indicate that the **raidscan** command was not run or was run incorrectly. See step 3 on page 545 in “Creating volume groups and file systems on replicated disks” on page 543.

3. Run the script again.

```
cl_verify_tc_config: ERROR - Disk hdisk38 added to VG truesyncvg does not match any hdisk in Device group htcdg01.
cl_verify_tc_config: ERROR - Disk hdisk39 added to VG truesyncvg does not match any hdisk in Device group htcdg01.
cl_verify_tc_config: ERROR - Volume Group truesyncvg in RG emlecRG has no hdisk from Device Group htcdg01.
cl_verify_tc_config: ERROR - Device Group htcdg01 added to TC truelee does not match any VG defined in RG emlecRG
cl_verify_tc_config: ERROR - TC truelee added to RG emlecRG does not match  any VG defined in RG emlecRG.
cl_verify_tc_config: ERROR - Volume Group truesyncvg inluded in RG emlecRG is a non-replicated VG.


Errors found verifying the HACMP TRUECOPY®/HUR configuration.  Status=4
~
```

Figure 11-18 Error messages that are found during TrueCopy/HUR replicated resource verification

Synchronizing the cluster configuration

You must verify the PowerHA SystemMirror Enterprise Edition cluster and the TrueCopy/HUR configuration before you can synchronize the cluster. To propagate the new TrueCopy/HUR configuration information and the additional resource group that were created across the cluster, follow these steps:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select **Extended Configuration** → **Extended Verification and Synchronization**.
3. In the Verify Synchronize or Both field select *Synchronize*. In the Automatically correct errors found during verification field select *No*. Press Enter.

   The output is displayed in the SMIT Command Status window.

11.5 Failover testing

This section explains the basic failover testing of the TrueCopy/HUR replicated resources locally within the site and across sites. You must carefully plan the testing of the site cluster failover because it requires more time to manipulate the secondary target LUNs at the recovery site. Also when testing the asynchronous replication, because of the nature of asynchronous replication, testing can also impact the data.

These scenarios do not entail performing a redundancy test with the IP networks. Instead, you configure redundant IP or non-IP communication paths to avoid isolation of the sites. The loss of all the communication paths between sites leads to a partitioned state of the cluster and to data divergence between sites if the replication links are also unavailable.

Another specific failure scenario is the loss of the replication paths between the storage subsystems while the cluster is running on both sites. To avoid this situation, configure redundant communication links for TrueCopy/HUR replication. You must manually recover the status of the pairs after the storage links are operational again.

Important: PowerHA SystemMirror Enterprise Edition does not trap SNMP notification events for TrueCopy/HUR storage. If a TrueCopy link goes down when the cluster is up and the link is repaired later, you must manually resynchronize the pairs.

This section explains how to perform the following tests for each site and resource group:

- Graceful site failover for the Austin site
- Rolling site failure of the Austin site
- Site re-integration for the Austin site
- Graceful site failover for the Miami site
- Rolling site failure of the Miami site
- Site reintegration for the Miami site

Each test, except for the last reintegration test, begins in the same initial state of each site that hosts its own production resource group on the primary node as shown in Example 11-18.

Example 11-18 Beginning of test cluster resource group states

```
clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                Node
-----------------------------------------------------------------------------
emlecRG        ONLINE                     jessica@Austin
               OFFLINE                    bina@Austin
               ONLINE SECONDARY           maddi@Miami
valhallaRG     ONLINE                     krod@Miami
               OFFLINE                    maddi@Miami
               ONLINE SECONDARY           bina@Austin
```

Before each test, we start copying data from another file system to the replicated file systems. After each test, we verify that the site service IP address is online and new data is in the file systems. We also had a script that inserts the current time and date into a file on each file system. Because of the small amounts of I/O in our environment, we were unable to determine whether we lost any data in the asynchronous replication.

11.5.1 Graceful site failover for the Austin site

Performing a controlled move of a production environment across sites is a basic test to ensure that the remote site can bring the production environment online. However, this task is done only during initial implementation testing or during a planned production outage of the site. You perform the graceful failover operation between sites by performing a resource group move.

In a true maintenance scenario, you most likely perform this task by stopping the cluster on the local standby node first. Then, you stop the cluster on the production node by using the *Move Resource Group*. You perform the following operations during this move:

- Releasing the primary online instance of emlecRG at the Austin site
  - Runs application server stop
  - Unmounts the file systems
  - Varies off the volume group
  - Removes the service IP address
- Releasing the secondary online instance of emlecRG at the Miami site.
- Acquire the `emlecRG` resource group in the secondary online state at Austin site.
- Acquire the `emlecRG` resource group in the online primary state at the Miami site.

To move the resource group by using SMIT, follow these steps:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **System Management (C-SPOC)** → **Resource Groups and Applications** → **Move a Resource Group to Another Node / Site** → **Move Resource Groups to Another Site**.

3. In the Move a Resource Group to Another Node / Site panel (Figure 11-19), select the ONLINE instance of the `emlecRG` resource group to be moved.

```
                Move a Resource Group to Another Node / Site

Move cursor to desired item and press Enter.
  +---------------------------------------------------------------------------+
  |                         Select Resource Group(s)                          |
  |                                                                           |
  | Move cursor to desired item and press Enter. Use arrow keys to scroll.    |
  |                                                                           |
  |   #                                                                       |
  |   # Resource Group                 State                Node(s) / Site    |
  |   #                                                                       |
  |   emlecRG                          ONLINE               jessica / Austi  |
  |   emlecRG                          ONLINE SECONDARY     maddi / Miami    |
  |   valhallarg                       ONLINE               krod / Miami     |
  |                                                                           |
  |   #                                                                       |
  |   # Resource groups in node or site collocation configuration:            |
  |   # Resource Group(s)                             State    Node / Site    |
  |   #                                                                       |
  |                                                                           |
  | F1=Help                F2=Refresh             F3=Cancel                   |
  | F8=Image               F10=Exit               Enter=Do                    |
F1| /=Find                 n=Find Next                                        |
F9+---------------------------------------------------------------------------+
```

Figure 11-19 Moving the Austin resource group across to site Miami

4. In the Select a Destination Site panel, select the Miami site as shown in Figure 11-20.

```
  +---------------------------------------------------------------------------+
  |                        Select a Destination Site                          |
  |                                                                           |
  | Move cursor to desired item and press Enter.                              |
  |                                                                           |
  |   # *Denotes Originally Configured Primary Site                           |
  |   Miami                                                                   |
  |                                                                           |
  | F1=Help                F2=Refresh             F3=Cancel                   |
  | F8=Image               F10=Exit               Enter=Do                    |
F1| /=Find                 n=Find Next                                        |
F9+---------------------------------------------------------------------------+
```

Figure 11-20 Selecting the site for resource group move

5. Verify the information in the final menu and press Enter.

   Upon completion of the move, `emlecRG` is online on the `maddi` node at the Miami site as shown in Example 11-19.

Example 11-19 Resource group status after a move to the Miami site

```
root@maddi# clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
emlecRG        ONLINE SECONDARY             jessica@Austin
               OFFLINE                      bina@Austin
               ONLINE                       maddi@Miami

valhallarg     ONLINE                       krod@Miami
               OFFLINE                      maddi@Miami
               OFFLINE                      bina@Austin
```

6. Repeat the resource group move to move it back to its original primary site and node to return to the original starting state.

Attention: In our environment, after the first resource group move between sites, we were unable to move the resource group back without leaving the pick list for the destination site empty. However, we were able to move it back by node, instead of by site. Later in our testing, the by-site option started working, but it moved it to the standby node at the primary site instead of the original primary node. If you encounter similar problems, contact IBM support.

11.5.2 Rolling site failure of the Austin site

In this scenario, you perform a rolling site failure of the Austin site by performing the following tasks:

1. Halt the primary production node `jessica` at the Austin site.
2. Verify that the resource group `emlecRG` is acquired locally by the `bina` node.
3. Halt the bina node to produce a site down.
4. Verify that the resource group `emlecRG` is acquired remotely by the maddi node.

To begin, all four nodes are active in the cluster and the resource groups are online on the primary node as shown in Example 11-18 on page 550.

1. On the `jessica` node, run the **`reboot -q`** command. The bina node acquires the `emlecRG` resource group as shown in Example 11-20.

Example 11-20 Local node failover within the Austin site

```
root@bina: clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
emlecRG        OFFLINE                      jessica@Austin
               ONLINE                       bina@Austin
               OFFLINE                      maddi@Miami
```

```
valhallarg      ONLINE                      krod@Miami
                OFFLINE                     maddi@Miami
                ONLINE SECONDARY            bina@Austin
```

2. Run the **pairdisplay** command (as shown in Example 11-21) to verify that the pairs are still established because the volume group is still active on the primary site.

Example 11-21 Pairdisplay status after a local site failover

```
root@bina: pairdisplay -g htcdg01 -IH2 -fe
Group   PairVol(L/R) (Port#,TID, LU),Seq#,LDEV#.P/S,Status,Fence,Seq#,P-LDEV# M CTG JID AP
htcdg01 htcd01(L)   (CL1-E-0, 0, 10)45306   272.P-VOL PAIR NEVER ,35764   268 -   -   -  1
htcdg01 htcd01(R)   (CL1-B-0, 0, 28)35764   268.S-VOL PAIR NEVER ,-----   272 -   -   -  -
htcdg01 htcd02(L)   (CL1-E-0, 0, 11)45306   273.P-VOL PAIR NEVER ,35764   269 -   -   -  1
htcdg01 htcd02(R)   (CL1-B-0, 0, 29)35764   269.S-VOL PAIR NEVER ,-----   273 -   -   -  -
```

3. Upon cluster stabilization, run the **reboot -q** command on the bina node. The maddi node at the Miami site acquires the emlecRG resource group as shown in Example 11-22.

Example 11-22 Hard failover between sites

```
root@maddi: clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
emlecRG        OFFLINE                     jessica@Austin
               OFFLINE                     bina@Austin
               ONLINE                      maddi@Miami

valhallarg     ONLINE                      krod@Miami
               OFFLINE                     maddi@Miami
               OFFLINE                     bina@Austin
```

4. Verify that the replicated pairs are now in the suspended state from the command line as shown in Example 11-23.

Example 11-23 Pairdisplay status after a hard site failover

```
root@maddi: pairdisplay -g htcdg01 -IH2 -fe
Group   PairVol(L/R) (Port#,TID, LU),Seq#,LDEV#.P/S,Status,Fence,Seq#,P-LDEV# M CTG JID AP
htcdg01 htcd01(L)   (CL1-B-0, 0,  28)35764   268.S-VOL SSUS NEVER ,-----   272 W   -   -  1
htcdg01 htcd01(R)   (CL1-E-0, 0,  10)45306   272.P-VOL PSUS NEVER ,35764   268 -   -   -  1
htcdg01 htcd02(L)   (CL1-B-0, 0,  29)35764   269.S-VOL SSUS NEVER ,-----   273 W   -   -  1
htcdg01 htcd02(R)   (CL1-E-0, 0,  11)45306   273.P-VOL PSUS NEVER ,35764   269 -   -   -  1
```

You can also verify that the replicated pairs are in the suspended state by using the Storage Navigator (Figure 11-21).

Important: Although our testing resulted in a site_down event, we never lost access to the primary storage subsystem. In a true site failure, including loss of storage, re-establish the replicated pairs, and synchronize them before you move them back to the primary site. If you must change the storage LUNs, modify the `horcm.conf` file, and use the same device group and device names. You do not have to change the cluster resource configuration.

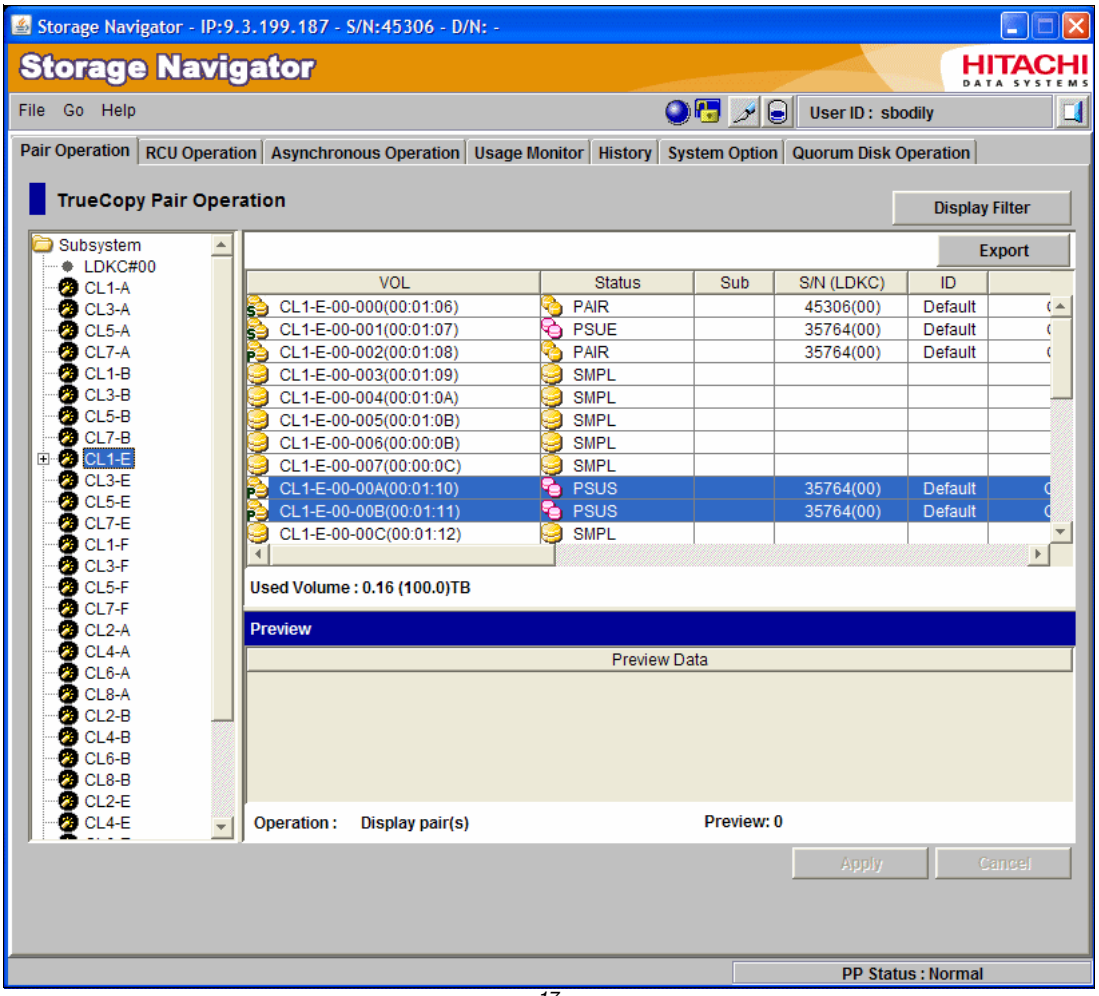


Figure 11-21 Pairs suspended after a site failover[17]

11.5.3 Site re-integration for the Austin site

In this scenario, we restart both cluster nodes at the Austin site by using the **`smitty clstart`** command. Upon startup of the primary node `jessica`, the `emlecRG` resource group is automatically gracefully moved back to and returns to the original starting point as shown in Example 11-18 on page 550.

[17] Courtesy of Hitachi Data Systems

Important: The resource group settings of the *Inter-site Management Policy,* also known as the *site relationship*, dictate the behavior of what occurs upon reintegration of the primary node. Because we chose *Prefer Primary Site*, the automatic fallback occurred.

Initially we are unable to restart the cluster on the `jessica` node because of verification errors at startup, which are similar to the errors shown in Figure 11-18 on page 549. Of the two possible reasons for these errors, one reason is that we failed to include starting the `horcm` instance on bootup. The second is reason is that we also had to remap the copy protected device groups by running the **`raidscan`** command again.

Important: Always ensure that the `horcm` instance is running before you rejoin a node into the cluster. In some cases, if all instances, cluster nodes, or both are down, you might need to run the **`raidscan`** command again.

11.5.4 Graceful site failover for the Miami site

This move scenario starts from the states that are shown in Example 11-18 on page 550. You repeat the steps from the previous three sections, one section at a time. However, these steps are performed to test the asynchronous replication of the Miami site.

The following tasks are performed during this move:

1. Release the primary online instance of `valhallaRG` at the Miami site.
   - Runs the application server stop.
   - Unmounts the file systems
   - Varies off the volume group
   - Removes the service IP address
2. Release the secondary online instance of `valhallaRG` at the Austin site.
3. Acquire `valhallaRG` in the secondary online state at the Miami site.
4. Acquire `valhallaRG` in the online primary state at the Austin site.

Perform the resource group move by using SMIT as follows:

1. From the command line, type the **`smitty hacmp`** command.
2. In SMIT, select the path **System Management (C-SPOC) → Resource Groups and Applications → Move a Resource Group to Another Node / Site → Move Resource Groups to Another Site**.
3. Select the ONLINE instance of `valhallaRG` to be moved.
4. Select the Austin site from the menu.
5. Verify the information in the final menu and press Enter.

   Upon completion of the move, the `valhallaRG` resource group is online on the bina node at the Austin site. The resource group is online secondary on the local production `krod` node at the Miami site as shown in Example 11-24.

Example 11-24 Resource group status after moving to the Austin site

```
root@bina: clRGinfo
Group Name     Group State                 Node
-----------------------------------------------------------------------------
emlecRG        ONLINE                      jessica@Austin
               OFFLINE                     bina@Austin
               ONLINE SECONDARY            maddi@Miami
```

```
valhallarg     ONLINE SECONDARY            krod@Miami
               OFFLINE                     maddi@Miami
               ONLINE                      bina@Austin
```

6. Repeat these steps to move a resource group back to the original primary `krod` node at the Miami site.

Attention: In our environment, after the first resource group move between sites, we were unable to move the resource group back without leaving the pick list for the destination site empty. However, we were able to move it back by node, instead of by site. Later in our testing, the by-site option started working, but it moved it to the standby node at the primary site instead of the original primary node. If you encounter similar problems, contact IBM support.

11.5.5 Rolling site failure of the Miami site

In this scenario, you perform a rolling site failure of the Miami site by performing the following tasks:

1. Halt primary production node krod at site Miami
2. Verify resource group valhallaRG is acquired locally by node maddi
3. Halt node maddi to produce a site down
4. Verify resource group valhallaRG is acquired remotely by node bina

To begin, all four nodes are active in the cluster, and the resource groups are online on the primary node as shown in Example 11-18 on page 550. Follow these steps:

1. On the `krod` node, run the **`reboot -q`** command. The `maddi` node brings the `valhallaRG` resource group online, and the remote bina node maintains the online secondary status as shown in Example 11-25. This time the failover time was noticeably longer, specifically in the fsck portion. The longer amount of time is most likely a symptom of the asynchronous replication.

Example 11-25 Local node fallover within the Miami site

```
root@maddi: clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                 Node
-----------------------------------------------------------------------------
emlecRG        ONLINE                      jessica@Austin
               OFFLINE                     bina@Austin
               ONLINE SECONDARY            maddi@Miami

valhallarg     OFFLINE                     krod@Miami
               ONLINE                      maddi@Miami
               ONLINE SECONDARY            bina@Austin
```

2. Run the **pairdisplay** command as shown in Example 11-26 to verify that the pairs are still established because the volume group is still active on the primary site.

Example 11-26 Status using the pairdisplay command after the local Miami site fallover

```
root@maddi: pairdisplay -fd -g hurdg01 -IH2 -CLI
Group   PairVol L/R  Device_File     Seq# LDEV# P/S Status Fence Seq# P-LDEV# M
hurdg01 hurd01 L     hdisk40        35764   270 P-VOL PAIR ASYNC  45306   274 -
hurdg01 hurd01 R     hdisk40        45306   274 S-VOL PAIR ASYNC      -   270 -
hurdg01 hurd02 L     hdisk41        35764   271 P-VOL PAIR ASYNC  45306   275 -
hurdg01 hurd02 R     hdisk41        45306   275 S-VOL PAIR ASYNC      -   271 -
```

3. Upon cluster stabilization, run the **reboot -q** command on the maddi node. The bina node at the Austin sites acquires the valhallaRG resource group as shown in Example 11-27.

Example 11-27 Hard failover from Miami site to Austin site

```
root@bina: clRGinfo
-----------------------------------------------------------------------------
Group Name     Group State                  Node
-----------------------------------------------------------------------------
emlecRG        ONLINE                       jessica@Austin
               OFFLINE                      bina@Austin
               OFFLINE                      maddi@Miami

valhallarg     OFFLINE                      krod@Miami
               OFFLINE                      maddi@Miami
               ONLINE                       bina@Austin
```

Important: Although our testing resulted in a site_down event, we never lost access to the primary storage subsystem. In a true site failure, including loss of storage, re-establish the replicated pairs, and synchronize them before you move them back to the primary site. If you must change the storage LUNs, modify the horcm.conf file, and use the same device group and device names. You do not have to change the cluster resource configuration.

11.5.6 Site reintegration for the Miami site

In this scenario, we restart both cluster nodes at the Miami site by using the **smitty clstart** command. Upon startup of the primary node krod, the valhallaRG resource group is automatically gracefully moved back to and returns to the original starting point as shown in Example 11-18 on page 550.

Important: The resource group settings of the *Inter-site Management Policy*, also known as the site relationship, dictate the behavior of what occurs upon reintegration of the primary node. Because we chose *Prefer Primary Site* policy, the automatic fallback occurred.

Initially we are unable to restart the cluster on the jessica node because of verification errors at startup, which are similar to the errors shown in Figure 11-18 on page 549. Of the two possible reasons for these errors, the first reason is that we failed to include starting the horcm instance on bootup. The second is reason is that we also had to remap the copy protected device groups by running the **raidscan** command again.

Important: Always ensure that the `horcm` instance is running before you rejoin a node to the cluster. In some cases, if all instances, cluster nodes, or both are down, you might need to run the **`raidscan`** command again.

11.6 LVM administration of TrueCopy/HUR replicated pairs

This section explains common scenarios for adding more storage to an existing replicated environment by using Hitachi TrueCopy/HUR. In this scenario, you work only with the Austin site and the `emlecRG` resource group in a TrueCopy synchronous replication. Overall the steps are the same for both types of replication. The difference is the initial pair creation. You perform the following tasks:

- Adding LUN pairs to an existing volume group
- Adding a new logical volume
- Increasing the size of an existing file system
- Adding a LUN pair to a new volume group

Important: This section does not explain how to dynamically expand a volume through Hitachi Logical Unit Size Expansion (LUSE) because this option is not supported.

11.6.1 Adding LUN pairs to an existing volume group

In this task, you assign a new LUN to each site as you did in 11.4.1, “Assigning LUNs to the hosts (host groups)” on page 525. Table 11-2 shows a summary of the LUNs that are used. Before you continue, the LUNs must already be established in a paired relationship, and the LUNs or hdisk must be available on the appropriate cluster nodes.

Table 11-2 Summary of the LUNs implemented

Austin - Hitachi USPV - 45306		**Miami - Hitachi USPVM - 35764**	
Port	CL1-E	Port	CL-1B
CU	01	CU	01
LUN	000E	LUN	001B
LDEV	01:14	LDEV	01:1F
jessica hdisk#	hdisk42	krod hdisk#	hdisk42
bina hdisk#	hdisk42	maddi hdisk#	hdisk42

Then, follow the same steps from of defining new LUNs:

1. Run the **`cfgmgr`** command on the primary node `jessica`.
2. Assign the PVID on the `jessica` node.

   ```
   chdev -l hdisk42 -a pv=yes
   ```

3. Run the **`pairsplit`** command on the replicated LUNs.
4. Run the **`cfgmgr`** command on each of the remaining three nodes.
5. Verify that the PVID shows up on each node by using the **`lspv`** command.
6. Run the **`pairresync`** command on the replicated LUNs.

7. Shut down the `horcm2` instance on each node:

   `/HORCM/usr/bin/horcmshutdown.sh 2`

8. Edit the `/etc/horcm2.conf` file on each node as appropriate for each site:
   - The `krod` and `maddi` nodes on the Miami site added the following new line:

     `htcdg01 htcd03 35764 01:1F`

   - The `jessica` and `bina` nodes on the Austin site added the following new line:

     `htcdg01 htcd03 45306 01:14`

9. Restart `horcm2` instance on each node:

   `/HORCM/usr/bin/horcmstart.sh 2`

10. Map the devices and device group on any node:

    `lsdev -Cc disk|grep hdisk|/HORCM/usr/bin/raidscan -IH2 -find inst`

    We ran this command on the `jessica` node.

11. Verify that the `htcgd01` device group pairs are now showing the new pairs, which consist of `hdisk42` on each system as shown in Example 11-28.

Example 11-28 New LUN pairs in the htcgd01 device group

```
root@jessica: pairdisplay -fd -g htcdg01 -IH2 -CLI
Group   PairVol L/R  Device_File     Seq# LDEV# P/S Status Fence Seq# P-LDEV# M
htcdg01 htcd01 L     hdisk38        45306   272 P-VOL PAIR NEVER  35764   268 -
htcdg01 htcd01 R     hdisk38        35764   268 S-VOL PAIR NEVER      -   272 -
htcdg01 htcd02 L     hdisk39        45306   273 P-VOL PAIR NEVER  35764   269 -
htcdg01 htcd02 R     hdisk39        35764   269 S-VOL PAIR NEVER      -   273 -
htcdg01 htcd03 L     hdisk42        45306   276 P-VOL PAIR NEVER  35764   287 -
htcdg01 htcd03 R     hdisk42        35764   287 S-VOL PAIR NEVER      -   276 -
```

You are now ready to use C-SPOC to add the new disk into the volume group:

Important: You cannot use C-SPOC for the following LVM operations to configure nodes at the remote site that contain the target volume:

- Creating a volume group
- Operations that require nodes at the target site to write to the target volumes

  For example, changing the file system size, changing the mount point, or adding LVM mirrors cause an error message in C-SPOC. However, nodes on the same site as the source volumes can successfully perform these tasks. The changes are then propagated to the other site by using a lazy update.

For C-SPOC operations to work on all other LVM operations, perform all C-SPOC operations with the (TrueCopy/HUR) volume pairs in the Synchronized or Consistent states or the cluster ACTIVE on all nodes.

1. From the command line, type the **`smitty cl_admin`** command.
2. In SMIT, select the path **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Add a Volume to a Volume Group**.
3. Select the volume group *truesyncvg* from the menu.

4. Select **hdisk42** as shown in Figure 11-22.

```
                    Set Characteristics of a Volume Group

Move cursor to desired item and press Enter.

  Add a Volume to a Volume Group
  Change/Show characteristics of a Volume Group
  Remove a Volume from a Volume Group
  +--------------------------------------------------------------------------+
  |                          Physical Volume Names                           |
  |                                                                          |
  | Move cursor to desired item and press Enter.                             |
  |                                                                          |
  |   000a621aaf47ce83 ( hdisk2 on nodes bina,jessica )                      |
  |   000a621aaf47ce83 ( hdisk3 on nodes krod,maddi )                        |
  |   000cf1da43e72fc2 ( hdisk5 on nodes bina,jessica )                      |
  |   000cf1da43e72fc2 ( hdisk6 on nodes krod,maddi )                        |
  |   00cb14ce74090ef3 ( hdisk42 on all selected nodes )                     |
  |   00cb14ceb0f5bd25 ( hdisk4 on nodes bina,jessica )                      |
  |   00cb14ceb0f5bd25 ( hdisk14 on nodes krod,maddi )                       |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
  | F8=Image                F10=Exit                Enter=Do                 |
F1| /=Find                  n=Find Next                                      |
F9+--------------------------------------------------------------------------+
```

Figure 11-22 Selecting a disk to add to the volume group

5. Verify the menu information, as shown in Figure 11-23, and press Enter.

```
                        Add a Volume to a Volume Group

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

                                                      [Entry Fields]
  VOLUME GROUP name                                  truesyncvg
  Resource Group Name                                emlecRG
  Node List                                          bina,jessica,krod,mad>
  Reference node                                     bina
  VOLUME names                                       hdisk42
```

Figure 11-23 Adding a volume to a volume group

The krod node does not need the volume group because it is not a member of the resource group. However, we started with all four nodes seeing all volume groups and decided to leave the configuration that way. This way, we have more flexibility later if we need to change the cluster configuration to allow the krod node to take over as a last resort.

Upon completion of the C-SPOC operation, all four nodes now have the new disk as a member of the volume group as shown in Example 11-29.

Example 11-29 New disk added to the volume group on all nodes

```
root@jessica: lspv |grep truesyncvg
hdisk38         00cb14ce564c3f44                    truesyncvg      active
hdisk39         00cb14ce564c40fb                    truesyncvg      active
hdisk42         00cb14ce74090ef3                    truesyncvg      active

root@bina: lspv |grep truesyncvg
hdisk38         00cb14ce564c3f44                    truesyncvg
hdisk39         00cb14ce564c40fb                    truesyncvg
hdisk42         00cb14ce74090ef3                    truesyncvg

root@krod: lspv |grep truesyncvg
hdisk38         00cb14ce564c3f44                    truesyncvg
hdisk39         00cb14ce564c40fb                    truesyncvg
hdisk42         00cb14ce74090ef3                    truesyncvg

root@maddi: lspv |grep truesyncvg
hdisk38         00cb14ce564c3f44                    truesyncvg
hdisk39         00cb14ce564c40fb                    truesyncvg

hdisk42         00cb14ce74090ef3                    truesyncvg
```

We do not need to synchronize the cluster because all of these changes are made to an existing volume group. However, you might want to run the **cl_verify_tc_config** command to verify the resources that are replicated correctly.

11.6.2 Adding a new logical volume

To perform this task, again you use C-SPOC, which updates the local nodes within the site. For the remote site, when a failover occurs, the lazy update process updates the volume group information as needed. This process also adds a bit of extra time to the failover time.

To add a new logical volume:

1. From the command line, type the **smitty cl_admin** command.
2. In SMIT, select the path **System Management (C-SPOC)** → **Storage** → **Logical Volumes** → **Add a Logical Volume**.
3. Select the **truesyncvg** volume group from the menu.

4. Choose the newly added disk **hdisk42** as shown in Figure 11-24.

```
                                Logical Volumes

Move cursor to desired item and press Enter.

  List All Logical Volumes by Volume Group
  Add a Logical Volume
  Show Characteristics of a Logical Volume
  Set Characteristics of a Logical Volume
  +--------------------------------------------------------------------------+
  |                          Physical Volume Names                           |
  |                                                                          |
  | Move cursor to desired item and press F7.                                |
  |     ONE OR MORE items can be selected.                                   |
  | Press Enter AFTER making all selections.                                 |
  |                                                                          |
  |   Auto-select                                                            |
  |   jessica hdisk38                                                        |
  |   jessica hdisk39                                                        |
  |   jessica hdisk42                                                        |
  |                                                                          |
  | F1=Help                 F2=Refresh              F3=Cancel                |
  | F7=Select               F8=Image                F10=Exit                 |
F1| Enter=Do                /=Find                  n=Find Next              |
F9+--------------------------------------------------------------------------+
```

Figure 11-24 Selecting a disk for new logical volume creation

5. Complete the information in the final menu and press Enter.

   We added a new logical volume, named `micah`, which consists of 50 logical partitions (LPARs) and selected a type of **raw**. We accepted the default values for all other fields as shown in Figure 11-25.

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

[TOP]                                                 [Entry Fields]
  Resource Group Name                                 emlecRG
  VOLUME GROUP name                                   truesyncvg
  Node List                                           bina,jessica,krod,mad>
  Reference node                                      jessica
* Number of LOGICAL PARTITIONS                       [50]                    #
  PHYSICAL VOLUME names                               hdisk42
  Logical volume NAME                                [micah]
  Logical volume TYPE                                [raw]                   +
  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                    1                      +
```

Figure 11-25 Defining a new logical volume

6. Upon completion of the C-SPOC operation, verify that the new logical was created locally on the `jessica` node as shown Example 11-30.

Example 11-30 Newly created logical volume

```
root@jessica: lsvg -l truesyncvg
truesyncvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
oreolv              jfs2       125     125     1    closed/syncd  /oreofs
majorlv             jfs2       125     125     1    closed/syncd  /majorfs
truefsloglv         jfs2log    1       1       1    closed/syncd  N/A
micah               raw        50      50      1    closed/syncd  N/A
```

11.6.3 Increasing the size of an existing file system

To perform this task, again you use C-SPOC, which updates the local nodes within the site. For the remote site, when a failover occurs, the lazy update process updates the volume group information as needed. This process also adds a bit of extra time to the failover time.

To increase the size of an existing file system:

1. From the command line, type the **`smitty cl_admin`** command.
2. In SMIT, select the path **System Management (C-SPOC)** → **Storage** → **File Systems** → **Change / Show Characteristics of a File System**.
3. Select the **oreofs** file system from the menu.
4. Complete the information in the final menu as desired and press Enter.

   In this scenario, we roughly tripled the size of the file system from 500 MB (125 LPARs), as shown in Example 11-30, to 1536 MB as shown in Figure 11-26.

```
        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                                 truesyncvg
  Resource Group Name                                emlecRG
* Node Names                                         krod,maddi,bina,jessi>

* File system name                                   /oreofs
  NEW mount point                                   [/oreofs]               /
  SIZE of file system
         Unit Size                                   M                      +
         Number of units                            [1536]                   #
  Mount GROUP                                       []
  Mount AUTOMATICALLY at system restart?             no                     +
  PERMISSIONS                                        read/write             +
  Mount OPTIONS                                     []
```

Figure 11-26 Changing the file system size

5. Upon completion of the C-SPOC operation, verify the new file system size locally on the `jessica` node as shown in Example 11-31.

Example 11-31 Newly increased file system size

```
root@jessica: lsvg -l truesyncvg
truesyncvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
oreolv              jfs2       384     384     1    closed/syncd  /oreofs
majorlv             jfs2       125     125     1    closed/syncd  /majorfs
truefsloglv         jfs2log    1       1       1    closed/syncd  N/A
micah               raw        50      50      1    closed/syncd  N/A
```

You do not need to synchronize the cluster because all of these changes are made to an existing volume group. However, you might want to make sure that the replicated resources verify correctly. Use the **`cl_verify_tc_config`** command first to isolate the replicated resources specifically.

Testing failover after making the LVM changes

Because you do not know if the cluster is going to work when needed, repeat the steps from 11.5.2, “Rolling site failure of the Austin site” on page 553. The new logical volume `micah` and the additional space on `/oreofs` show up on each node. However, there is a noticeable difference in the total time that is involved during the site failover when the lazy update was performed to update the volume group changes.

11.6.4 Adding a LUN pair to a new volume group

The steps for adding a volume are the same as the steps in 11.6.1, “Adding LUN pairs to an existing volume group” on page 559. The differences are that you are creating a volume group, which is required to add a volume group into a resource group. For completeness, the initial steps are documented here along with an overview of the new LUNs to be used:

1. Run the **`cfgmgr`** command on the primary node `jessica`.
2. Assign the PVID on the `jessica` node:

   ```
   chdev -l hdisk43 -a pv=yes
   ```

3. Run the **`pairsplit`** command on the replicated LUNs.
4. Run the **`cfgmgr`** command on each of the remaining three nodes.
5. Verify that the PVID shows up on each node by using the **`lspv`** command.
6. Run the **`pairresync`** command on the replicated LUNs.
7. Shut down the `horcm2` instance on each node:

   ```
   /HORCM/usr/bin/horcmshutdown.sh 2
   ```

8. Edit the `/etc/horcm2.conf` file on each node as appropriate for each site:
   - On the Miami site, the `krod` and `maddi` nodes added the following new line:

     ```
     htcdg01      htcd04       45306     00:20
     ```

   - On the Austin site, the `jessica` and `bina` nodes added the following new line:

     ```
     htcdg01      htcd04       35764     00:0A
     ```

9. Restart the `horcm2` instance on each node:

   ```
   /HORCM/usr/bin/horcmstart.sh 2
   ```

10. Map the devices and device group on any node. We ran the **raidscan** command on the jessica node. See Table 11-3 for more configuration details.

```
lsdev -Cc disk|grep hdisk|/HORCM/usr/bin/raidscan -IH2 -find inst
```

Table 11-3 Details on the Austin and Miami LUNs

Austin - Hitachi USPV - 45306		**Miami - Hitachi USPVM - 35764**	
Port	CL1-E	Port	CL-1B
CU	00	CU	00
LUN	000F	LUN	0021
LDEV	00:20	LDEV	00:0A
jessica hdisk#	hdisk43	krod hdisk#	hdisk43
bina hdisk#	hdisk43	maddi hdisk#	hdisk43

11. Verify that the htcgd01 device group pairs are now showing the new pairs that consist of hdisk42 on each system as shown in Example 11-32.

Example 11-32 New LUN pairs add to htcgd01 device group

```
root@jessica: pairdisplay -fd -g htcdg01 -IH2 -CLI
Group   PairVol L/R  Device_File     Seq# LDEV# P/S Status Fence Seq# P-LDEV# M
htcdg01 htcd01 L     hdisk38        45306   272 P-VOL PAIR NEVER  35764   268 -
htcdg01 htcd01 R     hdisk38        35764   268 S-VOL PAIR NEVER      -   272 -
htcdg01 htcd02 L     hdisk39        45306   273 P-VOL PAIR NEVER  35764   269 -
htcdg01 htcd02 R     hdisk39        35764   269 S-VOL PAIR NEVER      -   273 -
htcdg01 htcd04 L     hdisk43        45306    32 P-VOL PAIR NEVER  35764    10 -
htcdg01 htcd04 R     hdisk43        35764    10 S-VOL PAIR NEVER      -    32 -
```

You are now ready to use C-SPOC to create a volume group:

1. From the command line, type the **smitty cl_admin** command.
2. In SMIT, select the path **System Management (C-SPOC)** → **Storage** → **Volume Groups** → **Create a Volume to a Volume Group**.

3. In the Node Names panel, select the specific nodes. We chose all four as shown in Figure 11-27.

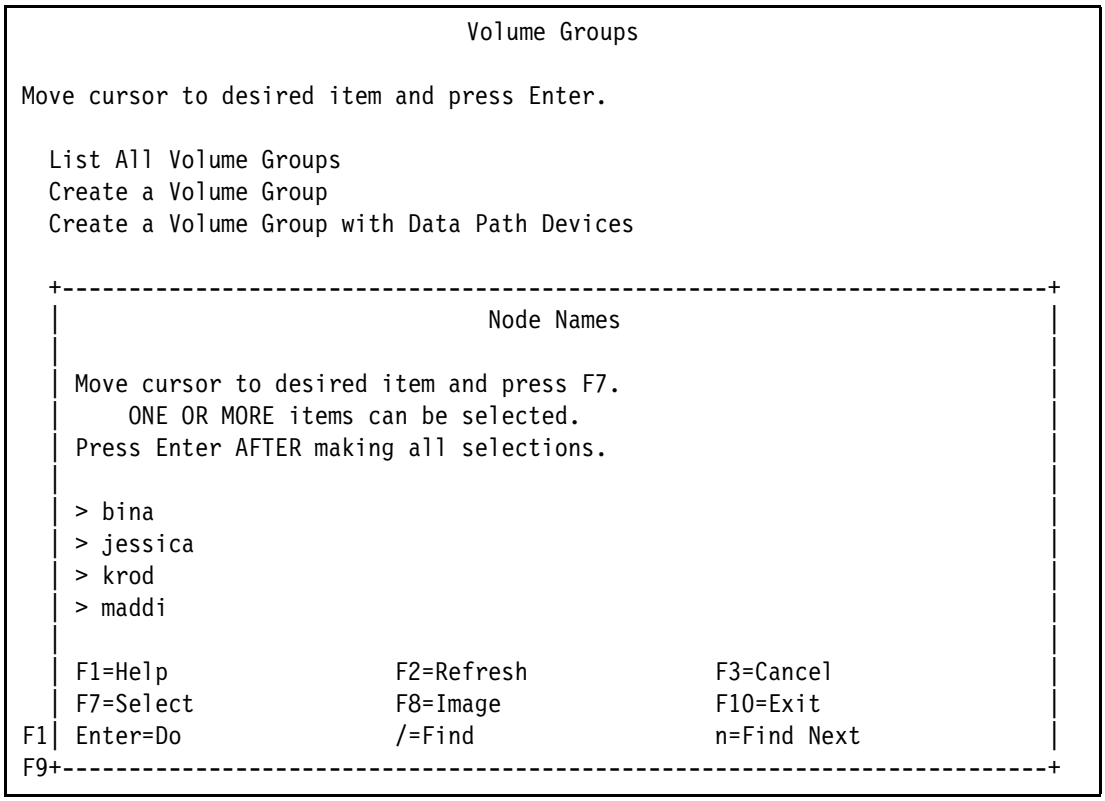


Figure 11-27 Selecting a volume group node

4. In the Physical Volume Names panel (Figure 11-28), select **hdisk43**.

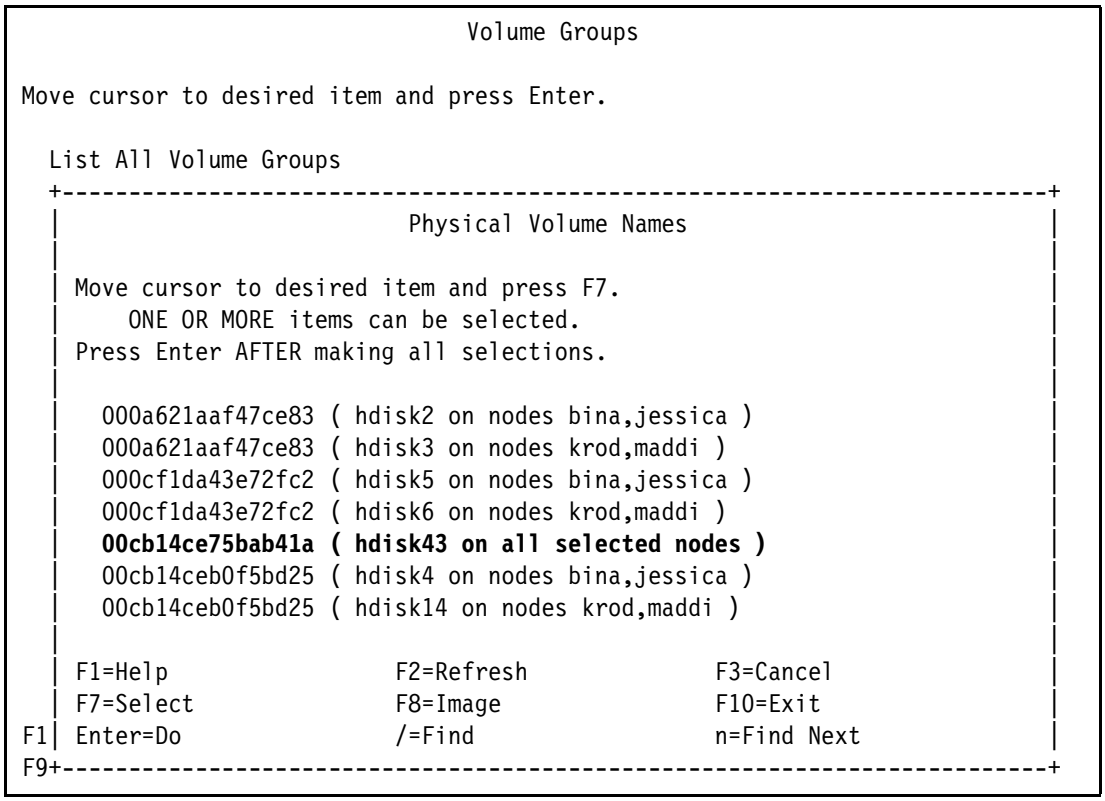


Figure 11-28 Selecting an hdisk for a new volume group

5. In the Volume Group Type panel, select the volume group type. We chose **Scalable** as shown in Figure 11-29.

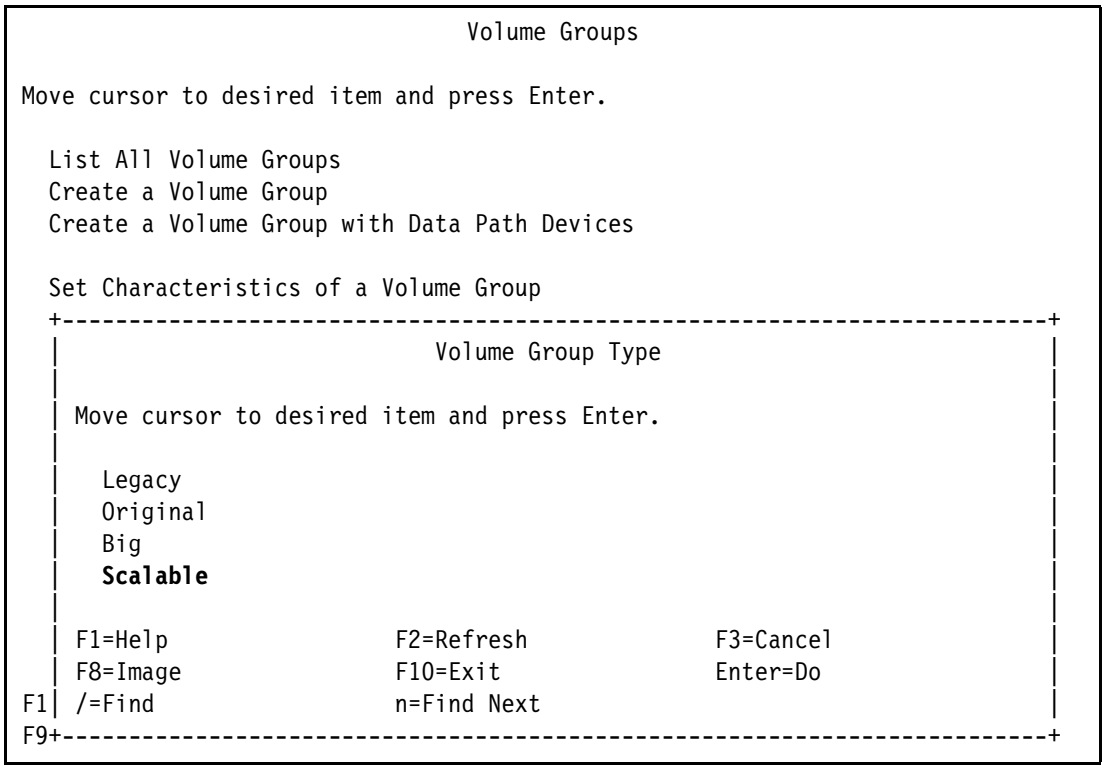


Figure 11-29 Selecting the volume group type for a new volume group

6. In the Create a Scalable Volume Group panel, select the resource group. We chose **emlecRG** as shown in Figure 11-30.

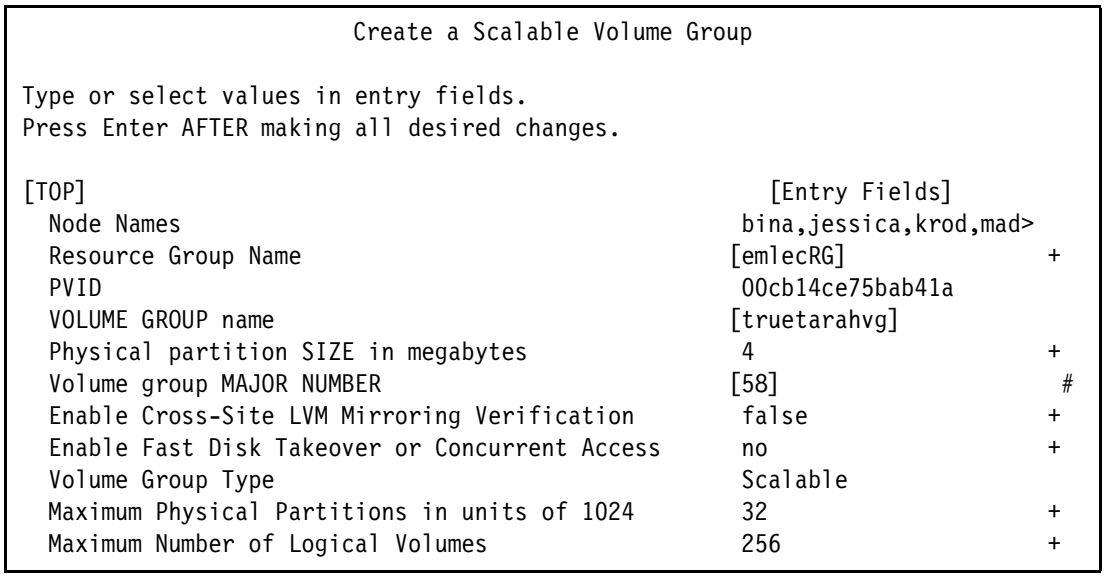


Figure 11-30 Create a volume group final C-SPOC SMIT menu

7. Choose a volume group name. We chose **truetarahvg**. Press Enter.

8. Verify that the volume group is successfully created, which we do on all four nodes as shown in Example 11-33.

Example 11-33 Newly created volume group on all nodes

```
root@jessica: lspv |grep truetarahvg
hdisk43         00cb14ce75bab41a                    truetarahvg

root@bina: lspv |grep truetarahvg
hdisk43         00cb14ce75bab41a                    truetarahvg

root@krod: lspv |grep truetarahvg
hdisk43         00cb14ce75bab41a                    truetarahvg

root@maddi: lspv |grep truetarahvg
hdisk43         00cb14ce75bab41a                    truetarahvg
```

When you create the volume group, the volume group is automatically added to the resource group as shown in Example 11-34. However, we do not have to change the resource group any further, because the new disk and device are added to the same device group and TrueCopy/HUR replicated resource.

Example 11-34 Newly added volume group also added to the resource group

```
Resource Group Name                             emlecRG
Participating Node Name(s)                      jessica bina maddi
Startup Policy                                  Online On Home Node Only
Fallover Policy                                 Fallover To Next Priority Node
Fallback Policy                                 Never Fallback
Site Relationship                               Prefer Primary Site
Node Priority
Service IP Label                                service_1
Volume Groups                                   truesyncvg truetarahvg
Hitachi TrueCopy Replicated Resources           truelee
```

9. Repeat the steps in 11.6.2, “Adding a new logical volume” on page 562, to create a logical volume, named `tarahlv` on the newly created volume group `truetarahvg`. Example 11-35 shows the new logical volume.

Example 11-35 New logical volume on newly added volume group

```
root@jessica: lsvg -l truetarahvg
truetarahvg:
LV NAME             TYPE       LPs     PPs     PVs  LV STATE      MOUNT POINT
tarahlv             raw        25      25      1    closed/syncd  N/A
```

10. Manually run the **`cl_verify_tc_config`** command to verify that the new addition of the replicated resources is complete.

Important: During our testing, we encountered a defect after the second volume group was added to the resource group. The **cl_verify_tc_config** command produced the following error messages:

```
cl_verify_tc_config: ERROR - Disk hdisk38 included in Device Group htcdg01 does
not match any hdisk in Volume Group truetarahvg.
cl_verify_tc_config: ERROR - Disk hdisk39 included in Device Group htcdg01 does
not match any hdisk in Volume Group truetarahvg.
cl_verify_tc_config: ERROR - Disk hdisk42 included in Device Group htcdg01 does
not match any hdisk in Volume Group truetarahvg.
Errors found verifying the HACMP TRUECOPY/HUR configuration.  Status=3
```

These results incorrectly imply a one to one relationship between the device group/replicated resource and the volume group, which is not intended. To work around this problem, ensure that the cluster is down, do a forced synchronization, and then start the cluster but ignore the verification errors. Usually when performing both a forced synchronization and then starting the cluster, do not ignore errors. Contact IBM support to see whether a fix is available.

Synchronize the resource group change to include the new volume that you just added. Usually you can perform this task within a running cluster. However, because of the defect that is mentioned in the previous Important box, we had to have the cluster down to synchronize it. To perform this task:

1. From the command line, type the **smitty hacmp** command.
2. In SMIT, select the path **Extended Configuration** → **Extended Verification and Synchronization and Verification**.
3. In the HACMP Verification and Synchronization display (Figure 11-31), for Force synchronization if verification fails, select **Yes**.

```
                    HACMP Verification and Synchronization

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

                                                      [Entry Fields]
* Verify, Synchronize or Both                        [Both]                  +
* Automatically correct errors found during          [No]                    +
  verification?

* Force synchronization if verification fails?       [Yes]                   +
* Verify changes only?                               [No]                    +
* Logging                                            [Standard]              +

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
```

Figure 11-31 Extended Verification and Synchronization SMIT menu

4. Verify the information is correct, and press Enter. Upon completion, the cluster configuration is in sync and can now be tested.
5. Repeat the steps for a rolling system failure as explained in 11.5.2, “Rolling site failure of the Austin site” on page 553. In this scenario, the tests are successful.

Testing failover after adding a volume group

Because you do not know whether the cluster is going to work when needed, repeat the steps of a rolling site failure as explained in 11.5.2, “Rolling site failure of the Austin site” on page 553. The new volume group `truetarahvg` and new logical volume `tarahlv` are displayed on each node. However, there is a noticeable difference in total time that is involved during the site failover when the lazy update is performed to update the volume group changes.

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 publications

For information about ordering these publications, see “Help from IBM” on page 574. Some of the documents referenced here may be available in softcopy only.

- *AIX 5L Performance Tools Handbook*, SG24-6039
- *IBM PowerVM Live Partition Mobility*, SG24-7460
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *Implementing the IBM System Storage SAN Volume Controller V5.1*, SG24-6423
- *Introduction to Storage Area Networks and System Networking*, SG24-5470
- *NIM from A to Z in AIX 5L*, SG24-7296
- *PowerHA for AIX Cookbook*, SG24-7739
- *RS/6000 SP System Performance Tuning Update*, SG24-5340
- *Tivoli Storage Manager Version 5.1 Technical Guide*, SG24-6554

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:

- *EMC PowerPath for AIX Version 5.3 Installation and Administration Guide,* P/N 300-008-341
- *EMC Solutions Enabler Symmetrix SRDF Family CLI Version 7.0 Product Guide,* P/N 300-000-877
- *EMC Solutions Enabler Version 7.0 Installation Guide,* P/N 300-008-918
- *EMC Symmetrix Remote Data Facility (SRDF) Connectivity Guide,* P/N 300-003-885
- *EMC Symmetrix Remote Data Facility (SRDF) Product Guide,* P/N 300-001-165
- *HACMP for AIX 6.1 Administration Guide,* SC23-4862
- *HACMP for AIX 6.1 Concepts and Facilities Guide,* SC23-4864
- *HACMP for AIX 6.1 Geographic LVM: Planning and Administration Guide,* SA23-1338
- *HACMP for AIX 6.1 Installation Guide,* SC23-5209
- *HACMP for AIX 6.1 Metro Mirror: Planning and Administration Guide,* SC23-4863
- *HACMP for AIX 6.1 Planning Guide*, SC23-4861

Online resources

These websites are also relevant as further information sources:

- IBM PowerHA SystemMirror for AIX

  http://www.ibm.com/systems/power/software/availability/aix/index.html
- IBM PowerHA High Availability wiki

  http://www.ibm.com/developerworks/wikis/display/WikiPtype/High%20Availability
- Implementation Services for Power Systems for PowerHA for AIX

  http://www-935.ibm.com/services/us/index.wss/offering/its/a1000032
- Online PowerHA manuals (still called HACMP.)

  http://http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.hacmp.doc/hacmpbooks.html
- GEO to GLVM Migration white paper

  http://www.ibm.com/systems/resources/systems_p_os_aix_whitepapers_pdf_aix_glvm.pdf
- IBM PowerHA SystemMirror for AIX page

  http://http://www.ibm.com/systems/power/software/availability/aix/index.html
- IBM PowerHA High Availability Wiki

  http://en.wikipedia.org/wiki/IBM_High_Availability_Cluster_Multiprocessing
- Yahoo PowerHA User Forum

  http://tech.groups.yahoo.com/group/hacmp/
- External site where you can download PowerHA PTFs

  http://www.software.ibm.com/webapp/set2/sas/f/hacmp/home.html
- Disaster Recovery Solutions for System p Servers and AIX 5L

  http://www.ibm.com/systems/resources/systems_p_software_whitepapers_disaster_recovery.pdf
- Documentation for the PowerHA Enterprise Edition is supplied in HTML and PDF formats

  http://publib.boulder.ibm.com/infocenter/aix/v6r1/index.jsp?topic=/com.ibm.aix.doc/doc/base/hacmp.htm

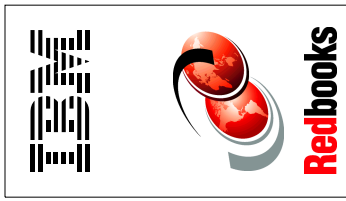
Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services



Exploiting IBM PowerHA SystemMirror V6.1 for AIX Enterprise Edition

Exploiting IBM PowerHA SystemMirror V6.1 for AIX Enterprise Edition

Highlights the benefits of deploying PowerHA Enterprise Edition

Includes multiple implementation scenarios

Describes networking planning and design

This IBM Redbooks publication positions the IBM PowerHA SystemMirror V6.1 for AIX Enterprise Edition as the cluster management solution for high availability that enables near-continuous application service and minimizes the impact of planned and unplanned outages.

This publication explains that the primary goal of this high-availability solution is to recover operations at a remote location after a system or data center failure, establish or strengthen a business recovery plan, and provide separate recovery location. The IBM PowerHA Enterprise Edition is targeted at multisite, high-availability disaster recovery.

The objective of this publication is to help new and existing PowerHA customers understand how to plan to accomplish a successful installation and configuration of the PowerHA Enterprise Edition.

This publication emphasizes the IBM Power Systems strategy to not only deliver more advanced functional capabilities for business resiliency, but also to enhance product usability and robustness through deep integration with AIX and affiliated software stack technologies. PowerHA SystemMirror is architected, developed, integrated, tested, and supported by IBM top to bottom.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7841-01 ISBN 0738437999